Got Data? Building a Data Ecosystem to Drive 21st Century Research

Dr. Francine Berman

Chair, Research Data Alliance / US

Hamilton Distinguished Professor in Computer Science, Rensselaer Polytechnic Institute

Rensselaer
why not change the world? ℠

Fran Berman

The Information Age:
Data-enabled Everything
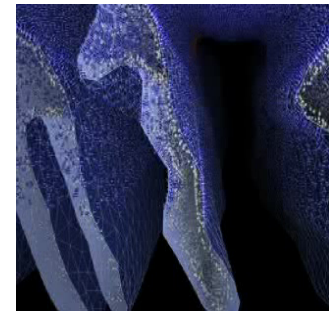
Physical Infrastructure

Communication / Community

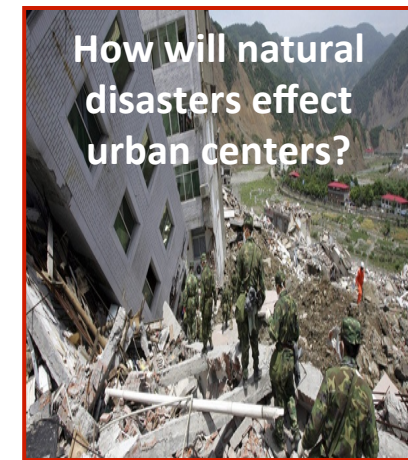**Commerce**

Education

**Entertainment**

Health

Research

Rensselaer
why not change the world? ℠

Fran Berman

Digital research data driving our ability to address science and societal challenges

What is the potential impact of Global Warming?

How do we ensure a future pipeline of innovation?

How will natural disasters effect urban centers?

How do we harness the power of community?

How can we create a more sustainable environment?

What therapies can be used to cure or control cancer?

Rensselaer
why not change the world? SM

Fran Berman

Today's Presentation

- **Data-driven Research**

- **Supporting Digital Research Data – Research Cyberinfrastructure**

- **Supporting the Research Data Ecosystem – Who's Paying the Data Bill?**

- **It takes a Village: Community Efforts to Build Coordinated Global Data Infrastructure**

Rensselaer
why not change the world? SM

Fran Berman

Data-Driven Research

Fran Berman
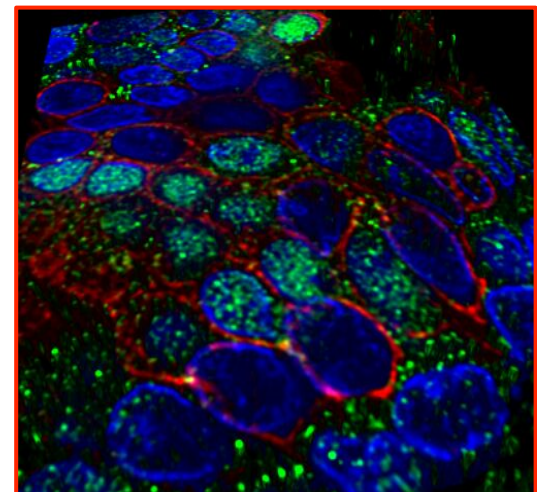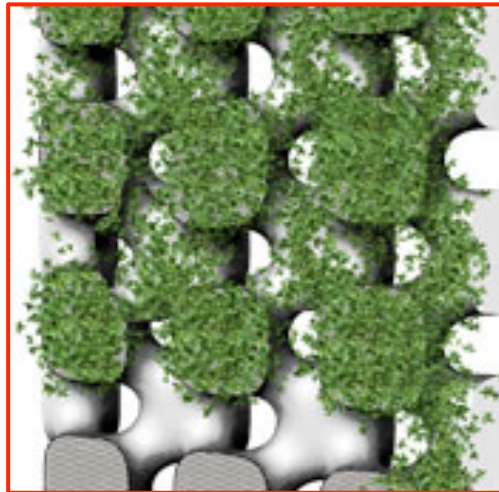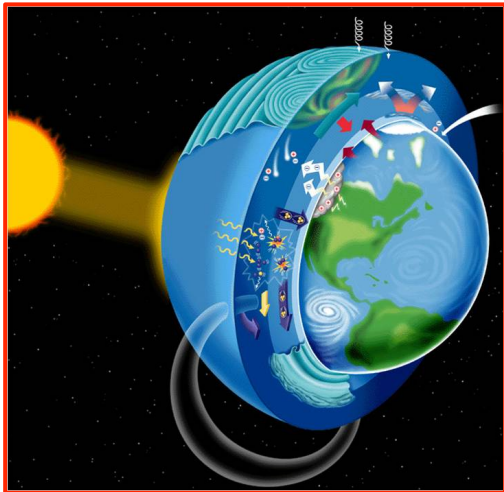
Data-Driven Geoscience – What is the impact of a large-scale earthquake on the San Andreas Fault?

- Earthquake simulations provide critical scientific information enabling better
  - *Estimation of seismic risk*
  - *Emergency preparation, response and planning*
  - *Design of next generation of earthquake-resistant structures*

- Simulations combine large-scale data collections, high-resolution models, supercomputer runs

Major Earthquakes on the San Andreas Fault, 1680-present

1906
M 7.8

1857
M 7.8

1680
M 7.7

?
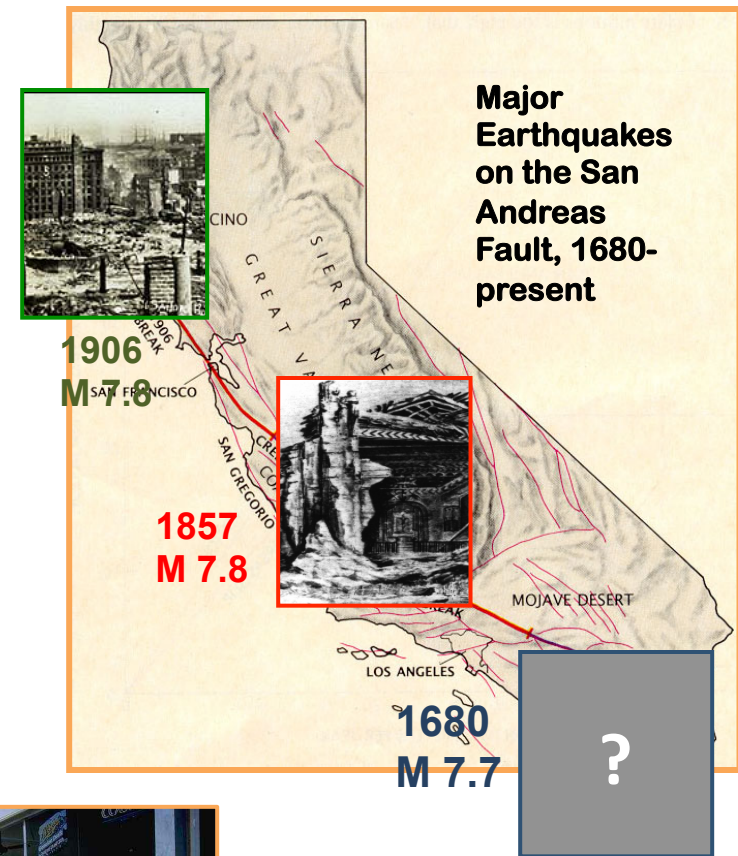
6.7 M earthquake in Northridge California, 1964, earthquake brought estimated $20B damage

Fran Berman

Rensselaer
why not change the world? SM

Earthquake Simulation

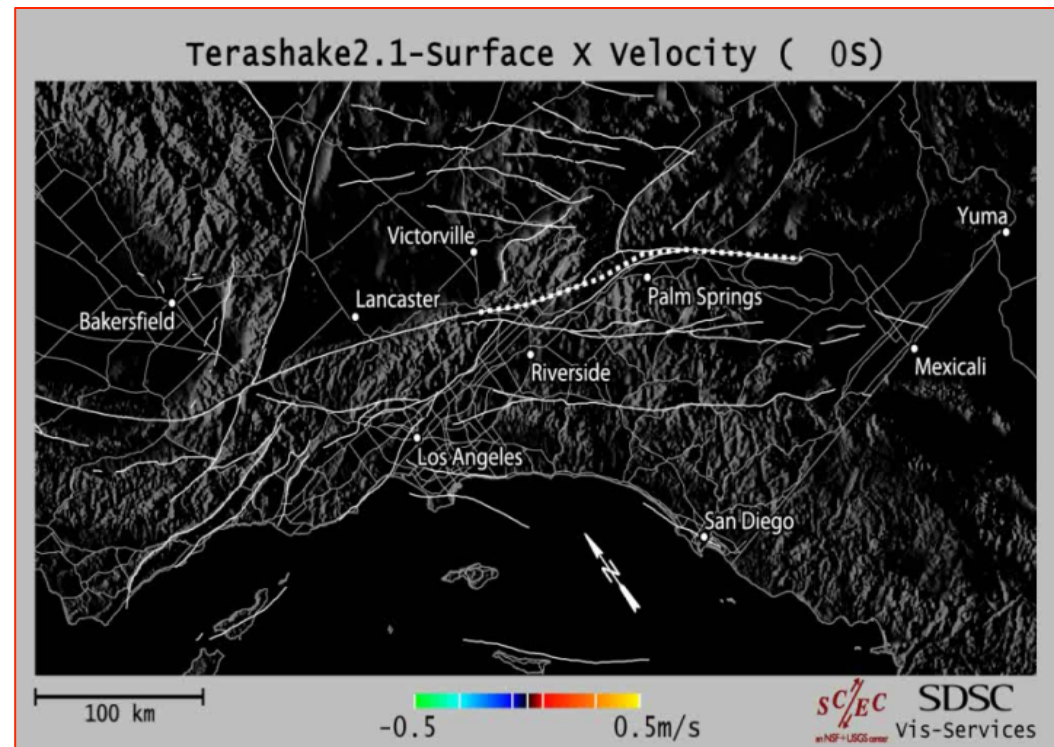Terashake2.1-Surface X Velocity (0S)

- **Input for supercomputer simulation:**

  – Model of 7.7 earthquake targeted to lower San Andreas fault

 – 10 years of sensor data on southern CA terrain

- **Post simulation**

  – Additional computation (80,000+ CPU hours) used for visualization of seismic wave propagation and analysis

  – Derived data products (velocity magnitude, displacement vector field, cumulative peak maps, statistics, etc.) included in SCEC digital library (>168 TB).

- **Results used for better science and a safer society (better building codes, more informed disaster response planning)**
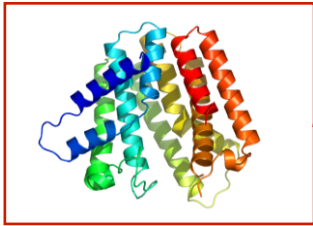
Visualization courtesy of Amit Chourasia, SDSC; Model from SCEC

Fran Berman

Rensselaer
why not change the world? SM

Better Prediction Accuracy Involves "Bigger" Data

Estimated figures for simulated 240 second period, 100 hour run-time	TeraShake domain (600x300x80 km^3)	PetaShake domain (800x400x100 km^3)
Fault system interaction	NO	YES
Inner Scale	200m	25m
Resolution of terrain grid	1.8 billion mesh points	2.0 trillion mesh points
Magnitude of Earthquake	7.7	8.1
Time steps	20,000 (.012 sec/step)	160,000 (.0015 sec/step)
Surface data	1.1 TB	1.2 PB
Volume data	43 TB	4.9 PB

Fran Berman

Rensselaer
why not change the world? ℠
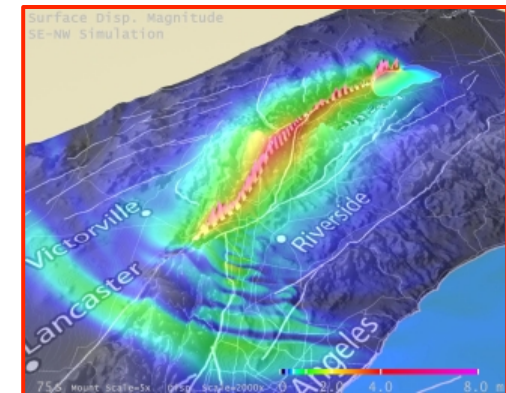
It's Not Just About Size – Data-enabled research applications span the spectrum

Analysis with Protein Data Bank structures

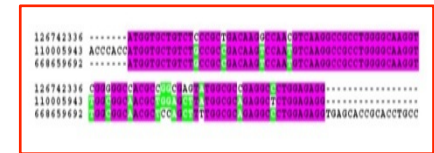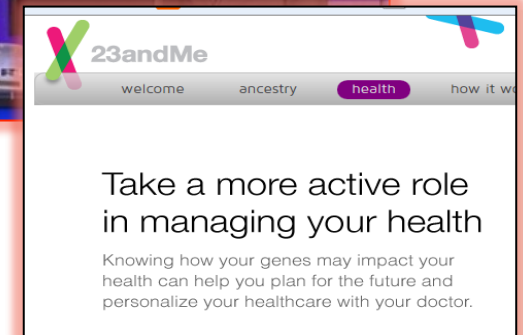*Cosmology simulations*

MORE DATA NEEDED ↑

Data needed beyond the local environment	More data and compute needed than local environment
Compute and data capacity in local environment sufficient	More compute needed than local environment

MORE COMPUTE NEEDED →
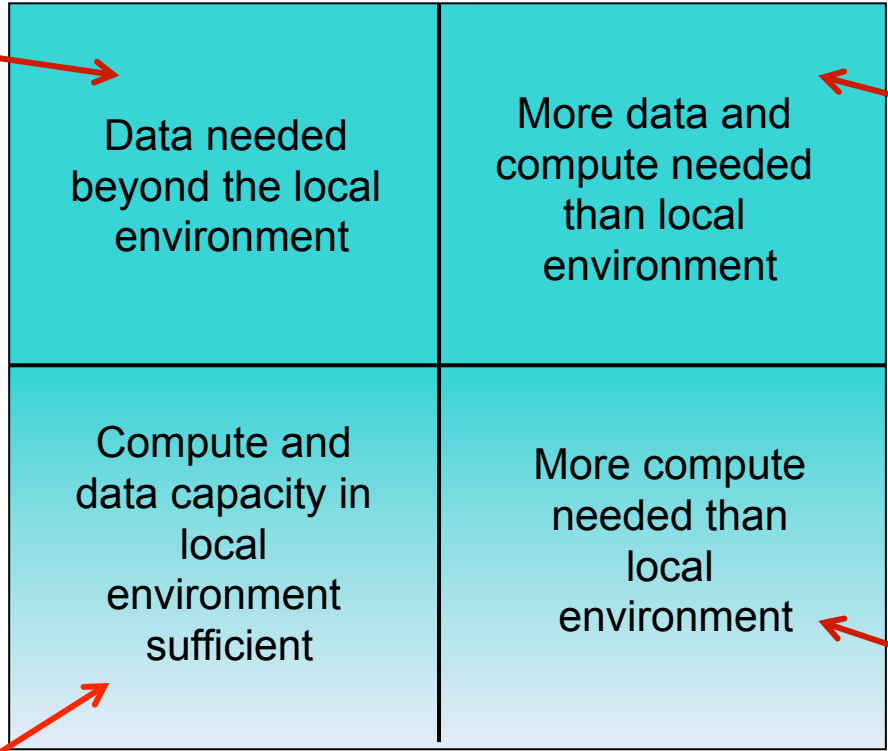
EurAsia male:female death rate study
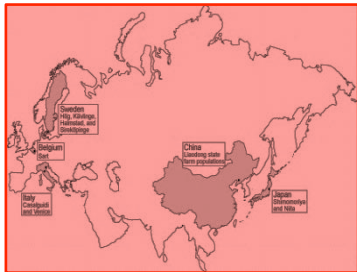
CIPRES Science Gateway Phylogeny and Ontology

Rensselaer
why not change the world? SM

Fran Berman

Moving Forward: Data-Driven Innovation Accelerated by Emergent Technologies

Exascale computing → more compute and data **at all tiers** in the Branscomb Pyramid.

- Exa-scale
- Petascale, Terascale
- Small-scale Devices

Take a more active role in managing your health

Knowing how your genes may impact your health can help you plan for the future and personalize your healthcare with your doctor.
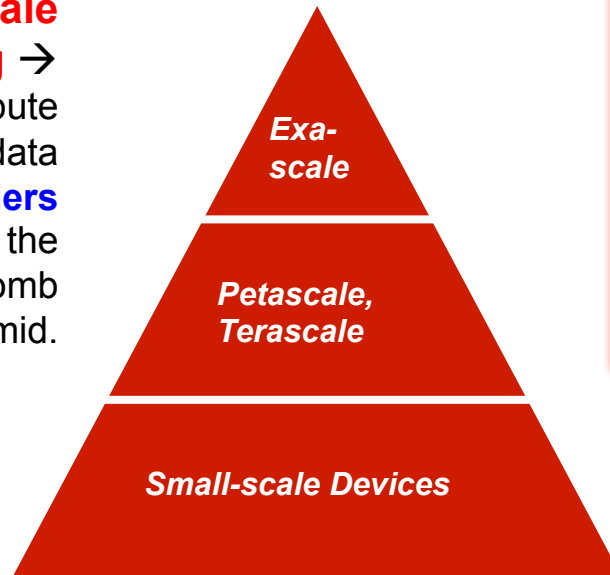
Smart Devices, Sensor Networks → More data-enabled devices and approaches drive **crowd-sourced, real-time, and other aggregation applications**

Information-Driven Analysis → X-informatics and X-analytics enable new targets for **data-driven research and decision-making** models

Rensselaer
why not change the world? ℠

Fran Berman

Data-Driven Innovation Transforming Academic Institutions and Communities

- **Communities coming together to**

  - Create coordinated research infrastructure

  - Disseminate useful information

  - Democratize education

  - Communicate and collaborate

  - Accelerate research collaboration and discoveries

Open access -- "Democratization" of participation

The T-shaped University – transformational potential of MOOCs on higher education

Data sharing – Acceleration of new results

Supporting Digital Research Data --
Research Cyberinfrastructure

Rensselaer
why not change the world? ℠

Fran Berman
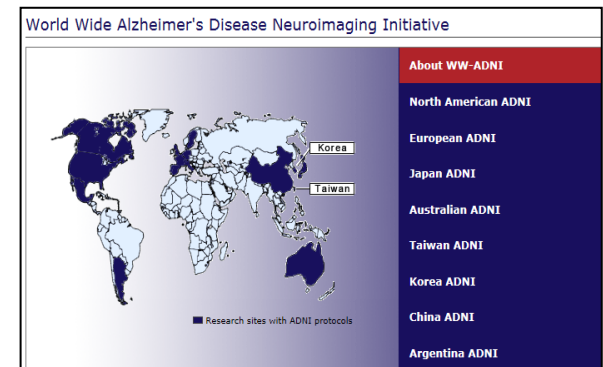
The Digital Research Data Life Cycle:
Data from birth to death / immortality

Create → **Edit** → **Use / Reuse** → **Publish, Disseminate** → **Preserve / Destroy**

Data creation / capture / gathering from
- laboratory experiments
- fieldwork
- surveys
- devices
- simulation output ...

- **Organize**
- **Annotate**
- **Clean**
- **Filter ....**

- **Analyze**
- **Mine**
- **Model**
- **Derive additional data**
- **Visualize**
- **Input to instruments / computers / devices ....**
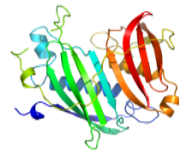
- **Disseminate**
- **Create portals / data collections / databases**
- **Couple with literature ....**
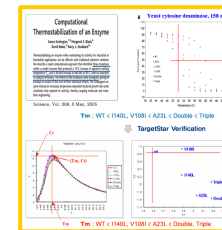
- **Store / preserve**
- **Store / replicate / preserve**
- **Store / ignore**
- **Destroy ....**

RENSSELAER
why not change the world? ℠

Information adapted from Chris Rusbridge and Liz Lyon

Fran Berman

Cyberinfrastructure, Tools, and Services:
Critical Support for Data-driven Research

visualization

modeling
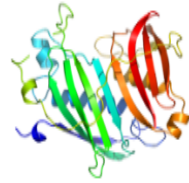
analysis
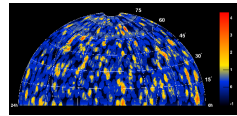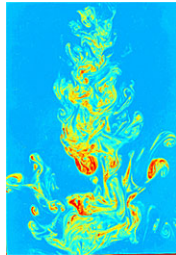
simulation

Interoperability drives research innovation

instruments
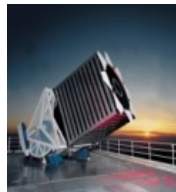
Sensors and Sensor-nets

computers

Data Use and Re-use

Data Access

Data Sharing

Data Storage

Data Management

Data Services
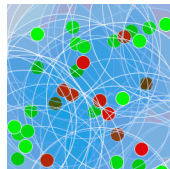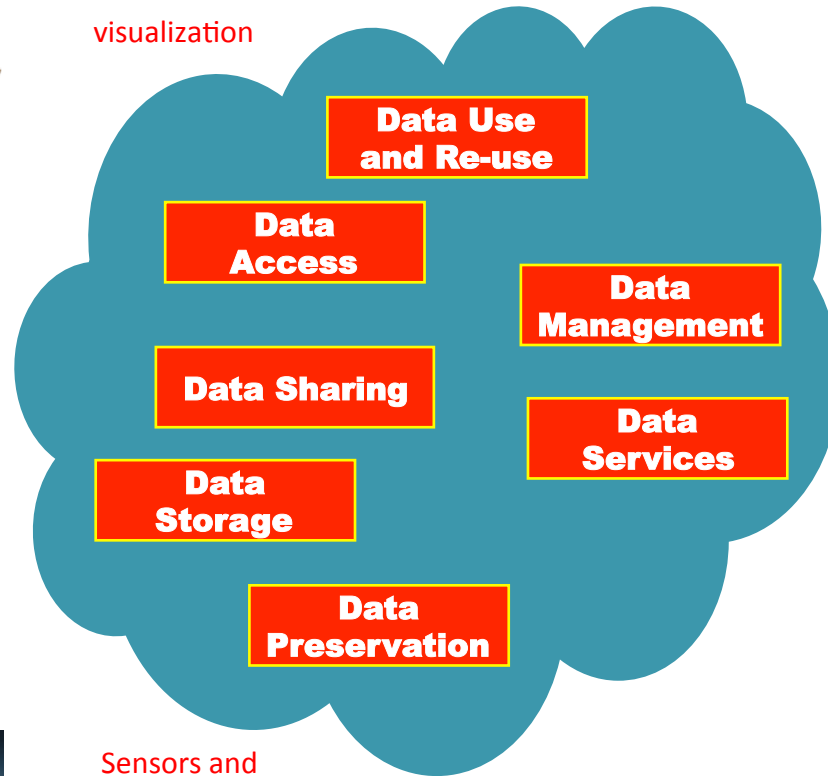
Data Preservation

Services drive data use and re-use

- Data mining
- Data Analytics
- Data Visualization
- Data Preservation
- Statistical packages
- Domain-specific tools
- Data de-identification
- Data hosting
- Collection Management …

RENSSELAER

why not change the world? SM

Fran Berman

Increasing Expectations of Stewardship
Access, management, preservation driving greater need for enabling Infrastructures

Costs of Research Data Stewardship Challenging for Academic Institutions

Data Central
datacentral.sdsc.edu

SDSC, circa 2006

Costs include

- Maintenance and upkeep
- Software tools and packages
- Utilities (power, cooling)
- Space
- Networking
- Security and failover systems
- People (expertise, help, infrastructure management, development)
- Training, documentation
- Monitoring, auditing
- Reporting costs
- Costs of compliance with regulation, policy, etc. …

Resources and Resource Refresh

Legend: — Model A (8-yr,15.2-mo 2X) — TB Stored — Planned Capacity

Y-axis: Archival Storage (TB)
X-axis: Date — June-97, June-98, June-99, June-00, June-01, June-02, June-03, June-04, June-05, June-06, June-07, June-08, June-09
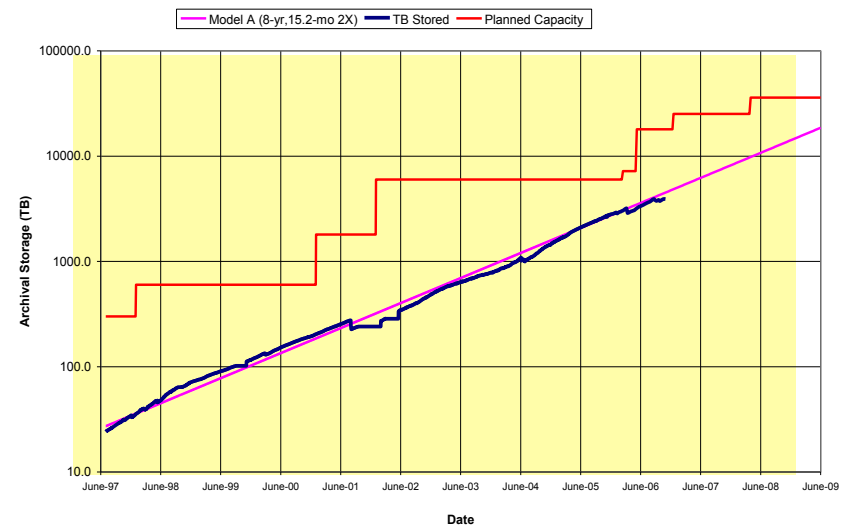
SDSC Data Storage Growth '97-'09

- *Most valuable data replicated*
- *As research collections increase, storage capacity must stay ahead of demand*

RENSSELAER
why not change the world? ℠

Information courtesy of Richard Moore and Natasha Balac, SDSC

Fran Berman

Supporting the Research Data Ecosystem – Who's Paying the Data Bill?

Economic

innovation

infrastructure

Fran Berman
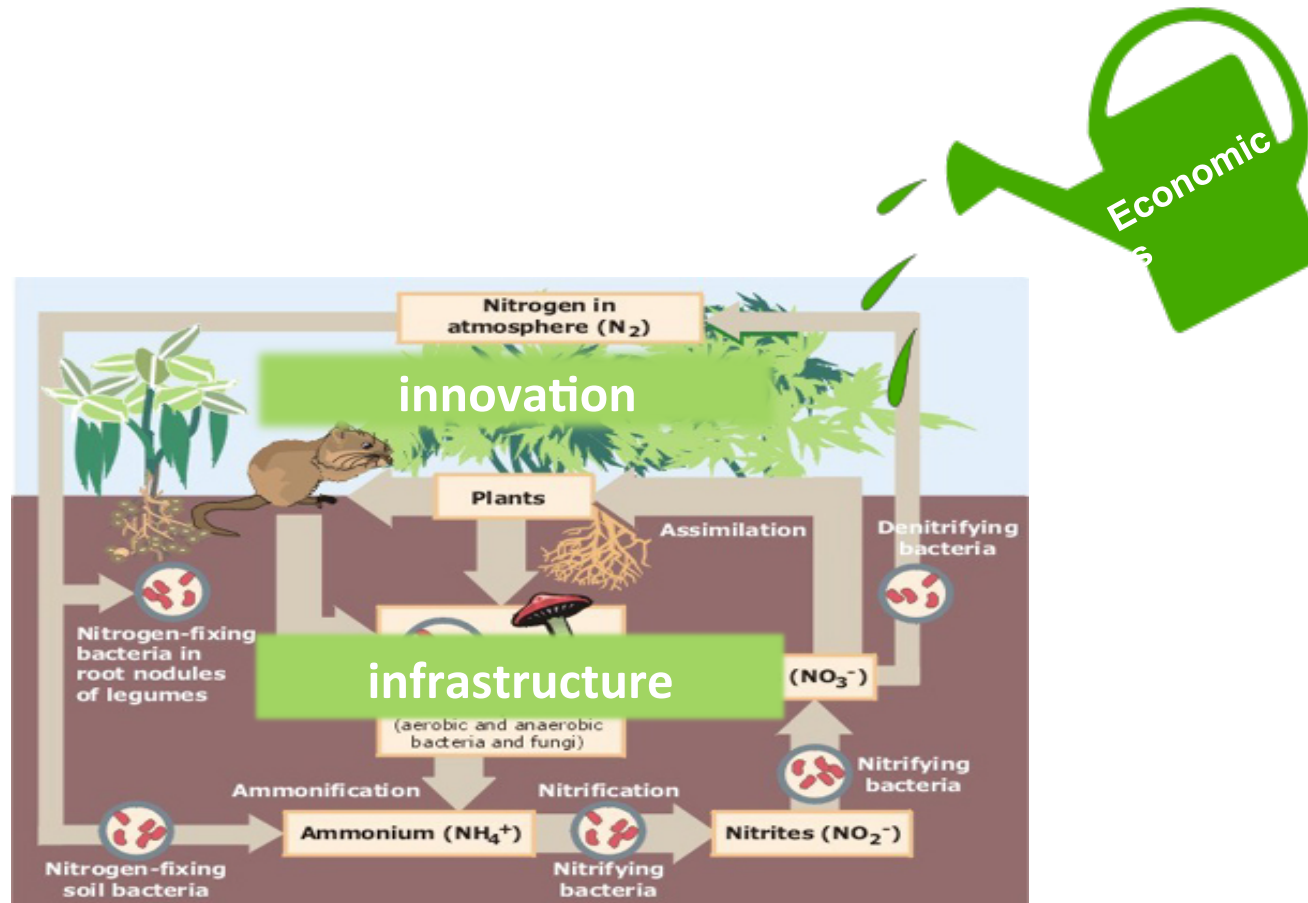
Enabling Research Data Cyberinfrastructure Requires Viable Economic Support

It's not just about size …

Data costs increase with usage, management requirements, perceived value

Greater costs at the extremes …

RENSSELAER
why not change the world? SM

Fran Berman
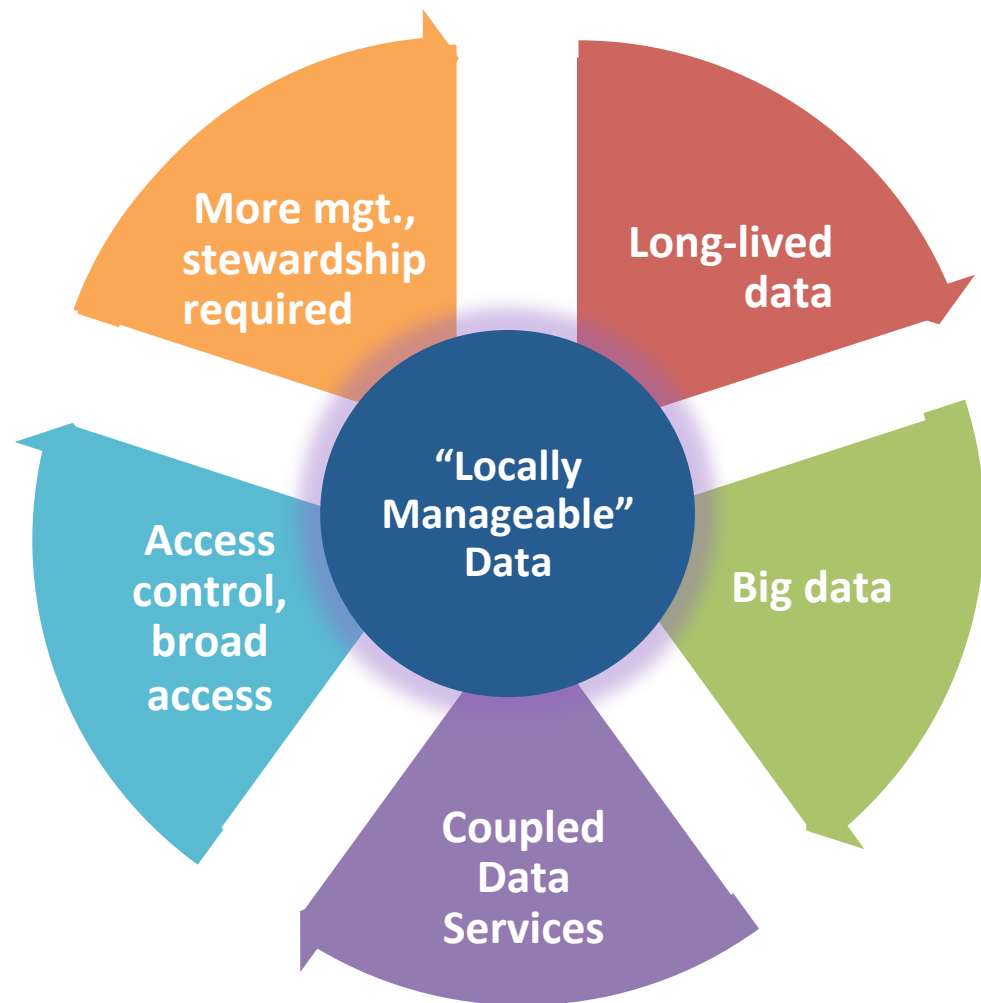
Who Pays the Data Bill?

Societal Value

Community Value

Individual Value

Data Pyramid

Fran Berman
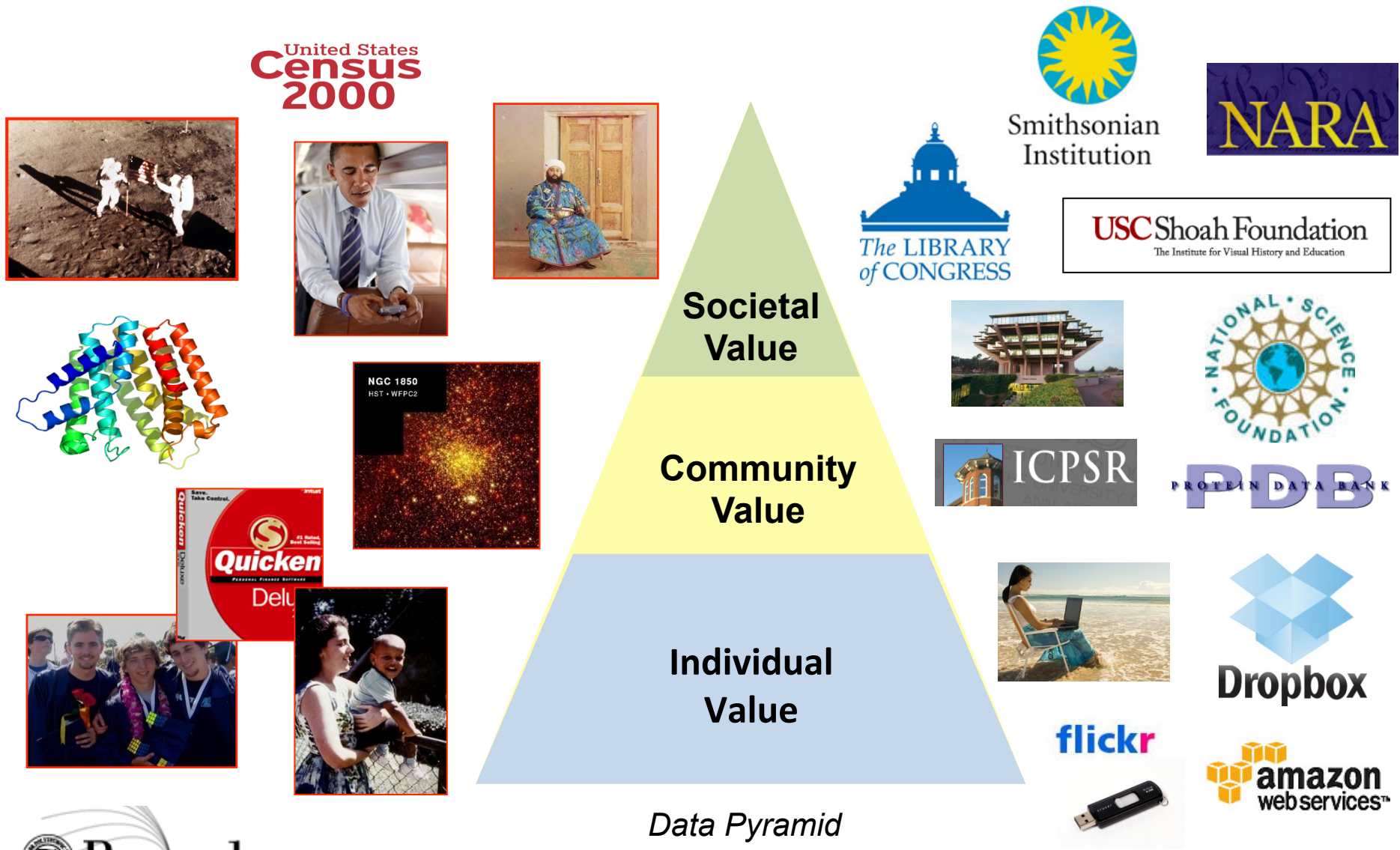
Data Value drives infrastructure, economic support models

Societal Value

Community Value

Individual Value

Data Pyramid

Publicly and privately hosted, **primarily publicly funded**
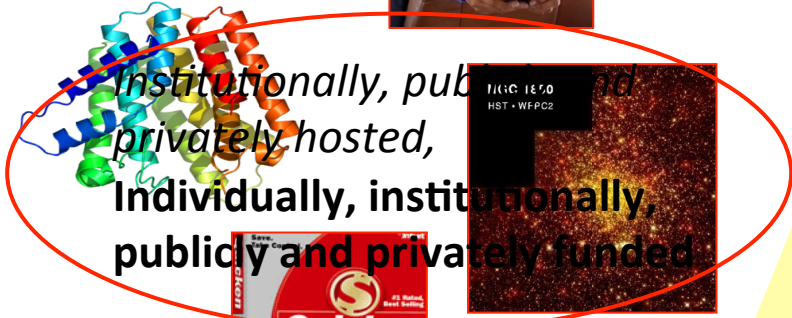
Institutionally, publicly and privately hosted, **Individually, institutionally, publicly and privately funded**
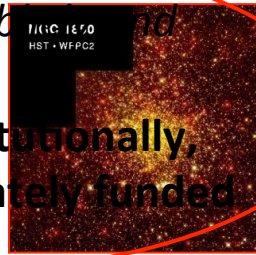
Individually and commercially hosted, **primarily individually and privately funded**

Greater public value

Greater access and preservation requirements

More infrastructure and management

Fran Berman

How can we better support research data of community value?

- *The greater the alignment between key stakeholder groups, the better the prospect for sustainable preservation*

- **Many stakeholders in digital access and preservation**
  - **Stakeholders who select** what to preserve
  - **Stakeholders who have rights** to the asset
  - **Stakeholders who benefit** from use of the preserved asset
  - **Stakeholders who preserve** the asset
  - **Stakeholders who pay**
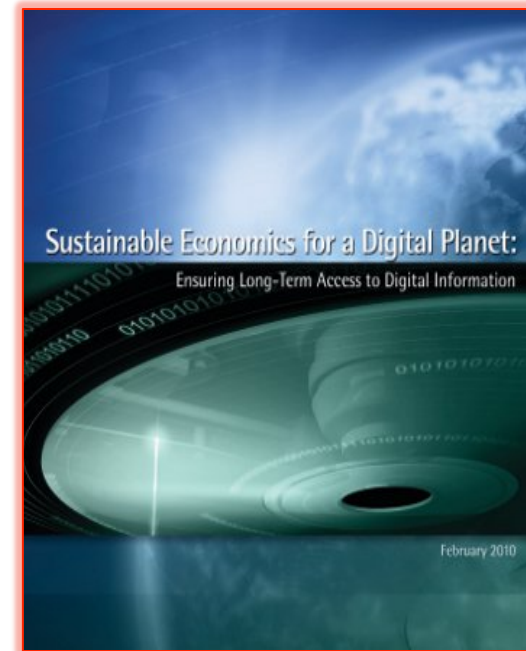
*More discussion in the Final Report of the **Blue Ribbon Task Force on Sustainable Digital Preservation and Access** (brtf.sdsc.edu)*

Rensselaer
why not change the world? SM
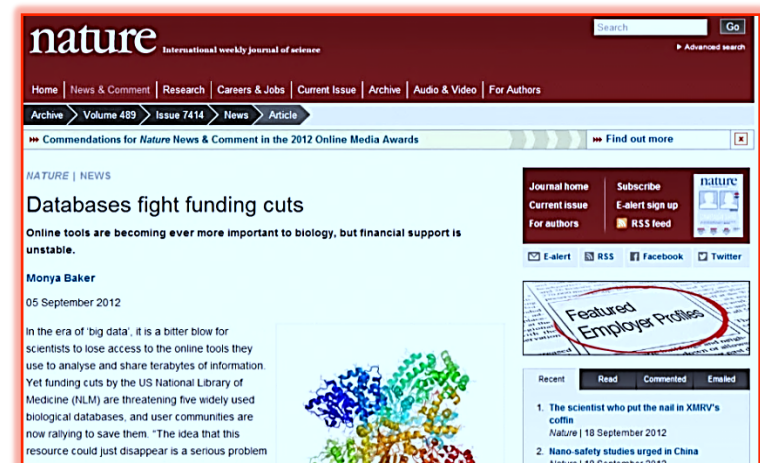
Paying for Data
Economic Models at play

Donations, etc.
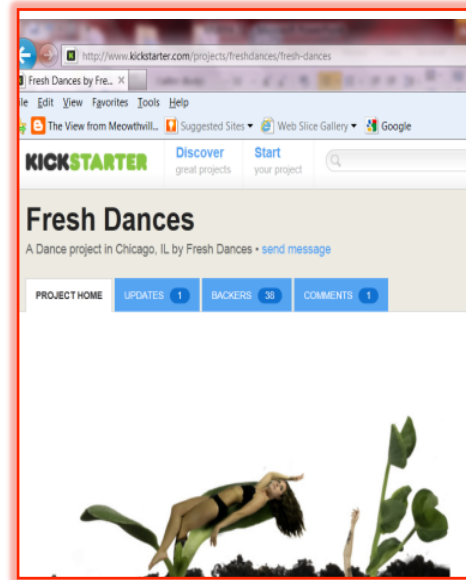
Crowd-sourcing

Subscription

Federal Grants

Pay per service

Advertisement

Rensselaer
why not change the world?℠
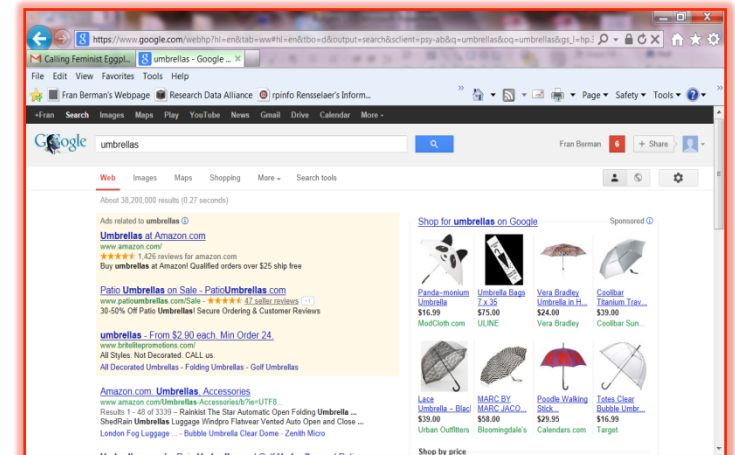
Economic sustainability – No one sector will do it all

WIRED
GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION

OVER 40 MAGAZINES. All in one place. All for only $14.99 a month. DOWNLOAD THE APP and get access now

SCIENCE

Google to Host Terabytes of Open-Source Science Data
BY ALEXIS MADRIGAL 01.18.08   11:23 AM

WIRED
GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION

OVER 75 MAGAZINES. All in one app. All FREE for 30 days. DOWNLOAD THE APP and get access now!

SCIENCE

Google Shutters Its Science Data Service
BY ALEXIS MADRIGAL 12.18.08   2:30 PM

Data at Risk Inventory

Tide Gauge Data 3

Tide Gauge Data 2

Growth, fruit yield and Azadirachtin content of seeds from Azadirachta indica in Nicaragua (thesis)

LILI-PAD: Long-term Investigations Library - Pathogens of Amphibians Database

Whooping Crane Captive Breeding Records Rescue

Organization & Development of a spatially referenced database for Agassiz's Desert Tortoise: a

Archiving Legacy Waterbird Nesting Data to Inform Adaptive Management of Wetland Habitat

Recovery of 62 DoD and State agency data sets

Colorado Fossil Inventory Proposal

SageSTEP DataStore Metadata

1 2 3 4 Next Last

The Scientist » The Nutshell

Funding Cuts Threaten Big Data
Reduced support from the US National Library of Medicine threatens to shut down five po[...] biological databases.

By Jef Akst | September 5, 2012

0 Comments

In 2007, the US National Library of Medicine (NLM) shunted millions of dollars away from infrastructure grants, which supported widely used databases that house terabytes of biolo[...]

BioT[...]
The International Journal of Life Science Methods

User Name   Password   Log In   Forgotten y[...]

NSF tears away funding for TAIR
12/03/2009
Tracy Vence

The National Science Foundation intends to eliminate funding for the Arabidopsis Information Resource by 2013.

The Arabidopsis Information Resource (TAIR) ha[...] a go-to resource for plant biologists around the w[...] over a decade. In collaboration with the Arabidop[...] Biological Resource Center (ABRC), TAIR allows[...] researchers to search and analyze information f[...] database of genetic and molecular biology data

---

TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

News   Magazine   Multimedia   Subjects   Surveys   Careers

Advertisement
Human DNA sequencing data + Ingenuity Variant Analysis
Can you find the causal variant?
TAKE THE CHALLENGE AND WIN!
Ingenuity Systems

The Scientist » News & Opinion » Daily News

Library cuts threaten research
As journal cancellations sweep across the US, scientists worry about how they will affect research

By Megan Scudellari | September 28, 2010

13 Comments

Earlier this month, New Mexico State University (NMSU) library announced the cancellation of over 700 journal and database subscriptions, the result of a perfect storm of rising journal prices and a slashed

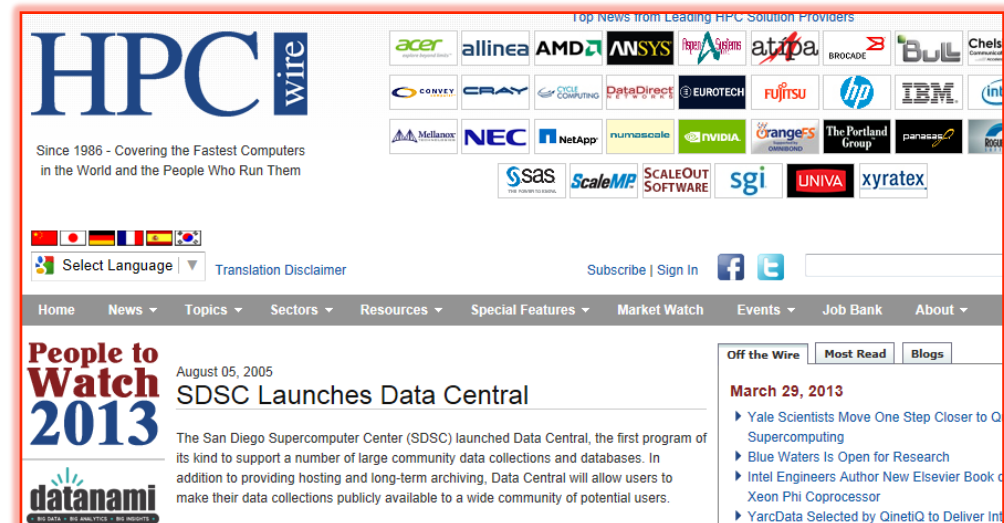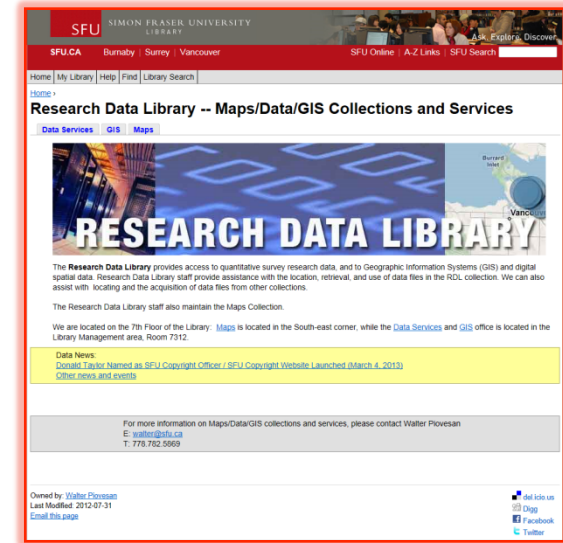## Rensselaer
why not change the world? ℠

Fran Berman

Potential Solutions in All Sectors

- **Four modest proposals …**

  1. **Facilitate private sector stewardship of public access research data**

     - *Encourage the private sector through tax incentives, philanthropy, and other means to provide stewardship for public access research data*

     - *Ensure stability and smooth transition for research data as companies move in and out of research data stewardship role*

Fran Berman

Potential Solutions in All Sectors

- **Four modest proposals …**

  1. Facilitate private sector stewardship of public access research data

  2. **Use public sector investment to provide an onramp to sustainable academic sector stewardship solutions**

Rensselaer
why not change the world? ℠

Potential Solutions in All Sectors

- **Four modest proposals …**

  1. Facilitate private sector stewardship of public access research data

  2. Use public sector investment to jumpstart sustainable academic sector stewardship solutions

  3. **Create and clarify public sector stewardship commitments for public access research data**

Which research data collections should be publicly hosted?

How long should public sector data collections be hosted?

How should funding be divided between research and infrastructure?

Who should host public sector research data?

Rensselaer

why not change the world? ℠

Fran Berman

Potential Solutions in All Sectors

- **Four modest proposals …**

  1. Facilitate private sector stewardship of public access research data

  2. Use public sector investment to jumpstart sustainable academic sector stewardship solutions

  3. Create and clarify public sector stewardship commitments for public access research data

  4. **Encourage research culture change to take advantage of what works in the private sector**

Fran Berman

Beyond the U.S.: Global Need for Data Infrastructure

Fran Berman

It takes a Village: Community Efforts to Build Coordinated Global Data Infrastructure

Image: http://ucsdnews.ucsd.edu/newsrel/supercomputer/02-08InternetUniverse.asp

It's Not Just Your Data, It's Other People's Data
Data Sharing and Interoperability critical for innovation

Data Sharing Infrastructure: Components needed throughout the Data Life Cycle

Create → Edit → Use / Reuse → Publish / Disseminate → Preserve / Destroy

- Stewardship and Management Practice
- Common Interoperability Framework
- Tools and Infrastructure that Promote Discoverability
- Sustainable economic model
- Digital Object Identifiers
- Framework for Comparative Analytics
- Access and Distribution Policy
- Preservation of Valued Collections
- Policy for Sharing Data
- Common Metadata standards

Fran Berman

Research Data Alliance (RDA):
New Community Organization formed to Accelerate Global Data Sharing and Exchange

- The Research Data Alliance (RDA) is a new international organization formed to implement / adopt infrastructure needed to accelerant research data **sharing and exchange worldwide**

- **12-18 month Working Groups (WG) will serve as drivers for** data sharing infrastructure and practice. WG deliverables to include

  – Adopted standards

  – Deployed infrastructure

  – Adopted policy

  – Implemented best practice, etc.

> **Simon Hodson** @simonhodson99
>
> Greer: Aim is to create an organisation that will not just envision a global data infrastructure, but implement & operate it #rdalaunch
>
> 4 DAYS AGO · REPLY · RETWEET · FAVORITE

> **John Erickson** @olyerickson
>
> Fran Berman: #RDA is not trying to boil the ocean #RDALaunch
>
> 4 DAYS AGO · REPLY · RETWEET · FAVORITE

Rensselaer
why not change the world?[SM]
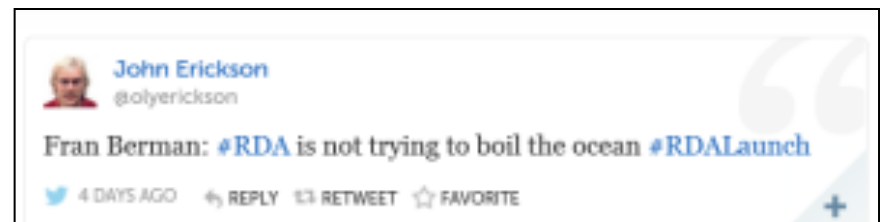
March 18-20, 2013
Gothenburg, Sweden
RDA goes live!

- Over 200 participants
- 31 countries
- 5 continents
- Public, private, academic sectors
- High-profile Govt. and Science speakers

RDA
RESEARCH DATA ALLIANCE

najla
@najlaoa

Jahanian #NSF: the #RDALaunch promises to accelerate pace of scientific discoveries. Data is new currency for science

4 DAYS AGO   REPLY   RETWEET   FAVORITE

ResearchDataAlliance
@resdatall

Another theme emerging: RDA as a "neutral" space across domains, organizations, nations, etc. #RDALaunch

2 DAYS AGO   REPLY   RETWEET   FAVORITE

ResearchDataAlliance
@resdatall

South Africa can't afford not to participate in RDA #RDALaunch

4 DAYS AGO   REPLY   RETWEET   FAVORITE

Rensselaer
why not change the world? SM

Left photo courtesy of Leif Laaksonen

Fran Berman

Research Data Alliance Approach:
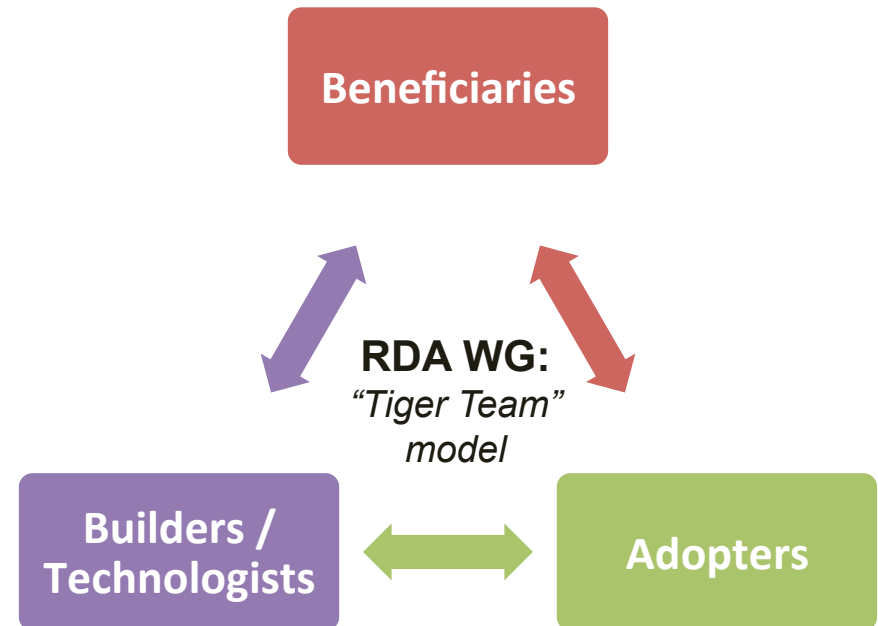## Focused efforts, key players, infrastructure adoption

RDA Working Group Deliverables are:

- **Focused pieces of adopted code, policy, infrastructure, standards, or best practices** that enable data to be shared and exchanged

- **"Harvestable" efforts** for which 12-18 months of work can eliminate a roadblock for a substantial community

- **Efforts that have substantive applicability** to "chunks" of the data community, but may not apply to everyone

- **Efforts for which working scientists and researchers can start today** while more long-term or far-reaching solutions are appropriately discussed in other venues
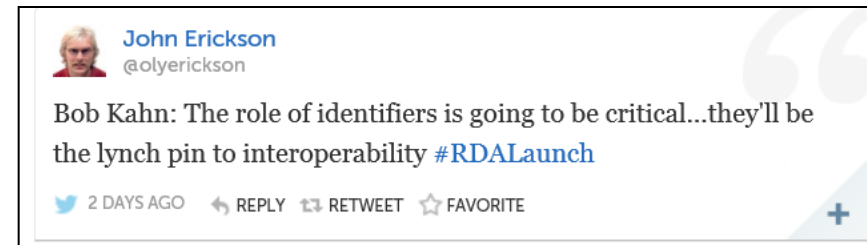
Organizational Sponsors:

- *NSF (RDA/US)*
- *European Commission (RDA/EU)*
- *Australian Govt. (ANDS)*

Beneficiaries

RDA WG:
"Tiger Team" model

Builders / Technologists

Adopters

**Rensselaer**

why not change the world? SM

Fran Berman

Many RDA Groups Forming

RDA Interest Group: Agricultural Data

- **Focus:** Share and create interoperability framework for data from agricultural research, in particular wheat

  - *Wheat data combines Germplasm data, genetic and phenotypic data, statistical data, bibliographical data, etc.*

- Group members from India, France, Greece, UN Food and Agricultural Organization
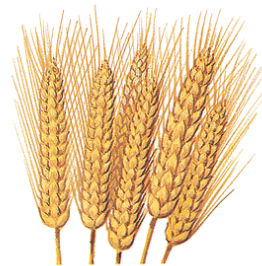
RDA Working Group: Persistent Identifier (PID) Information Types

- **Focus:** Develop framework for information types to be associated with unique identifiers, adopt framework at DKRZ (German supercomputing Center) and other sites

  - *Agreement about the information associated with PIDs would allow programmers to implement the same API independent of the PID type being used*

RDA 2013 Current Status

- **Next Plenary: September 16-18 at National Academies in Washington DC**

Growing the RDA community

– **New members: rd-alliance.org**

– **New Interest Groups** (discussion), **Working Groups** (deliverables), **Task Groups** (building the organization)

– Reaching out to other data community organizations to form **Affiliate Organizational Partners** to synergistically drive high impact community efforts

Rensselaer

why not change the world? SM

Fran Berman

Data-Driven Innovation Requires an Ecosystem
Impact dependent on effective development and integration of all components
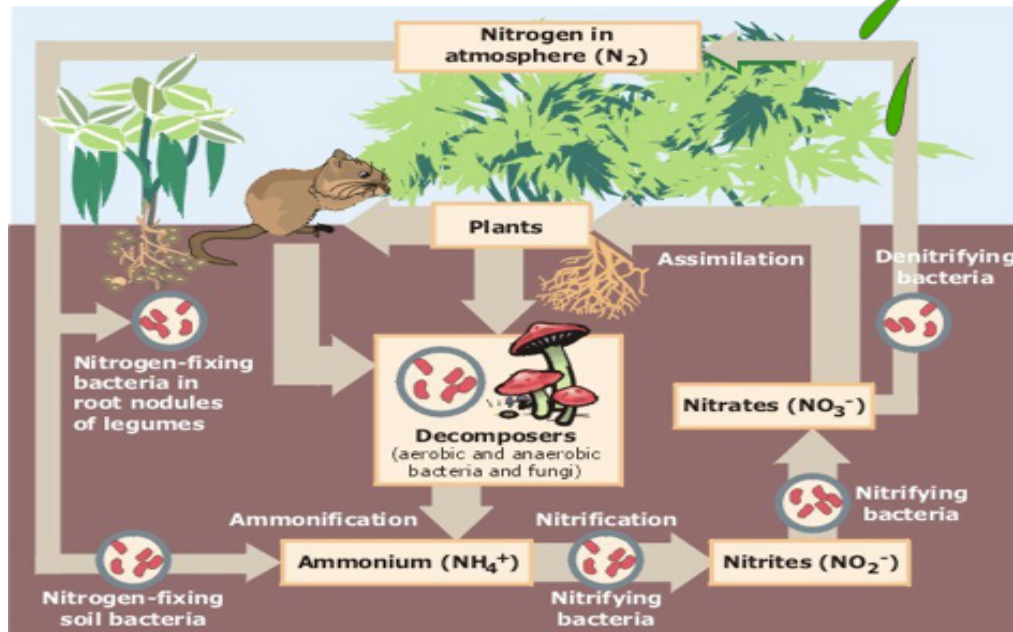
Innovation

Infrastructure
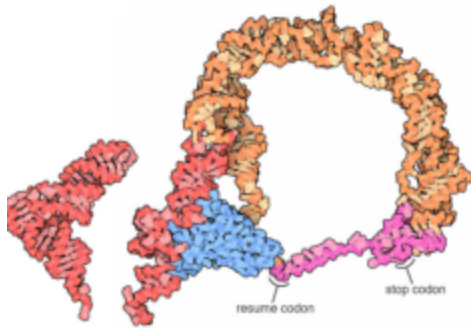
Economic support

Community effort

Policy

Practice

…

Fran Berman

Thank You

Happy 50th Birthday to Purdue CS!

Fran Berman

The Power of Many: Community infrastructure adoption and coordination drive innovation

- *"Just do it"* -- Focused efforts help communities drive tangible progress

Creation / adoption of **data sharing policies** have accelerated research innovation

Development and adoption of **parallel communication protocols** through the MPI Forum drove a generation of advances

THE LANCET Neurology

The Lancet Neurology, Volume 6, Issue 6, Pages 480 - 481, June 2007
doi:10.1016/S1474-4422(07)70118-7   Cite or Link Using DOI

Alzheimer's researchers open the doors to data sharing

James Butcher

Data from the US Alzheimer's Disease Neuroimaging Initiative will be made publicly available in a central database

Development of a **public access** to shared data collection enabling new results for Alzheimer's

I E T F®

Now 25 years old, the Internet Engineering Task Force's mission "to make the Internet work better" has resulted in key specifications of Internet **community standards** that support innovation

Fran Berman

Rensselaer

why not change the world? ℠

Community Momentum