Got Data? New Roles for Libraries in Shaping 21st Century Research

ALCTS President's Program, June 2010

Dr. Francine Berman

Vice President for Research, Rensselaer Polytechnic Institute

Co-Chair, Blue Ribbon Task Force for Sustainable Digital Preservation and Access

The Digital World

DATA.GOV

E-Government

Microsoft

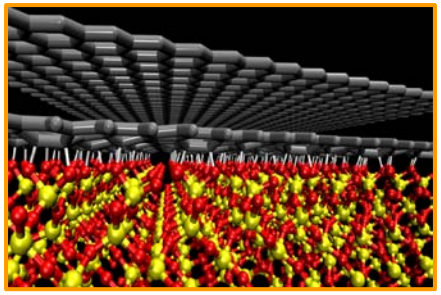
Google

E-Business

Research and Education

iTunes

Digital Entertainment



Communication and Information

Rensselaer

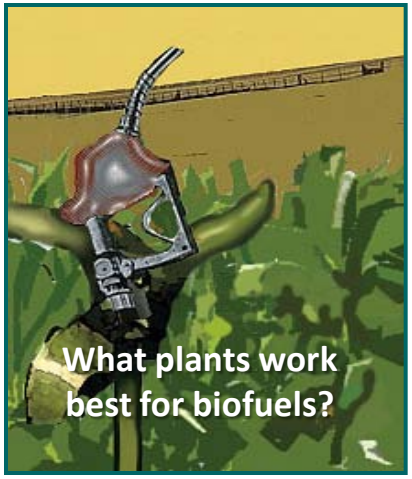
why not change the world? SM

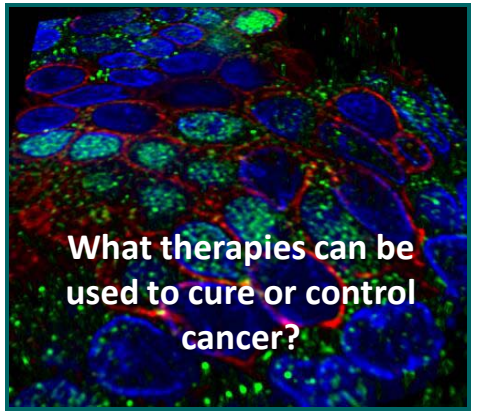
Fran Berman

Science and Technology Needed to Address Modern Challenges in Research, Education, Practice

"Science is more essential for our prosperity, our security, our health, our environment, and our quality of life than it has ever been before."

President Barack Obama





Rensselaer

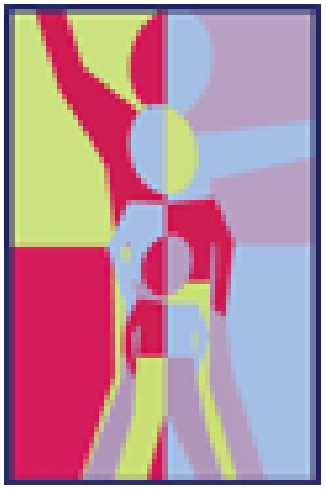
why not change the world? SM

Fran Berman

Research Today

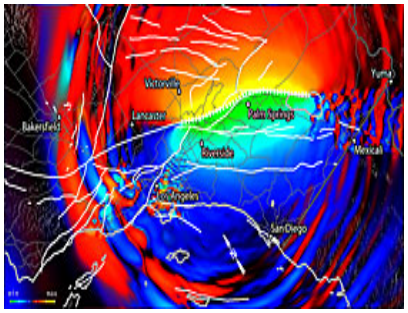
Which has the greatest impact – nature or nurture?

PSID: longitudinal data on 8000 families over 40 years



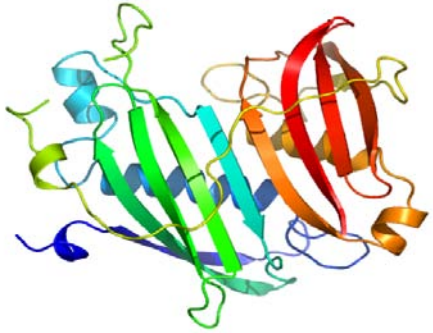
How does disease spread?

PDB: World wide reference collection of protein structure information



What is the impact of a large-scale earthquake on the Southern San Andreas Fault?

Digital data from Southern California Earthquake Center simulations used for disaster planning and building requirements



Are current stresses on this bridge dangerous?

Terabridge data set: Structure sensor data for real-time data mining, event detection, decision support and alert dissemination

Where are the brown dwarfs?

NVO: Data from 50+ astronomical sky surveys and large-scale telescopes.

Rensselaer

why not change the world? SM

Fran Berman

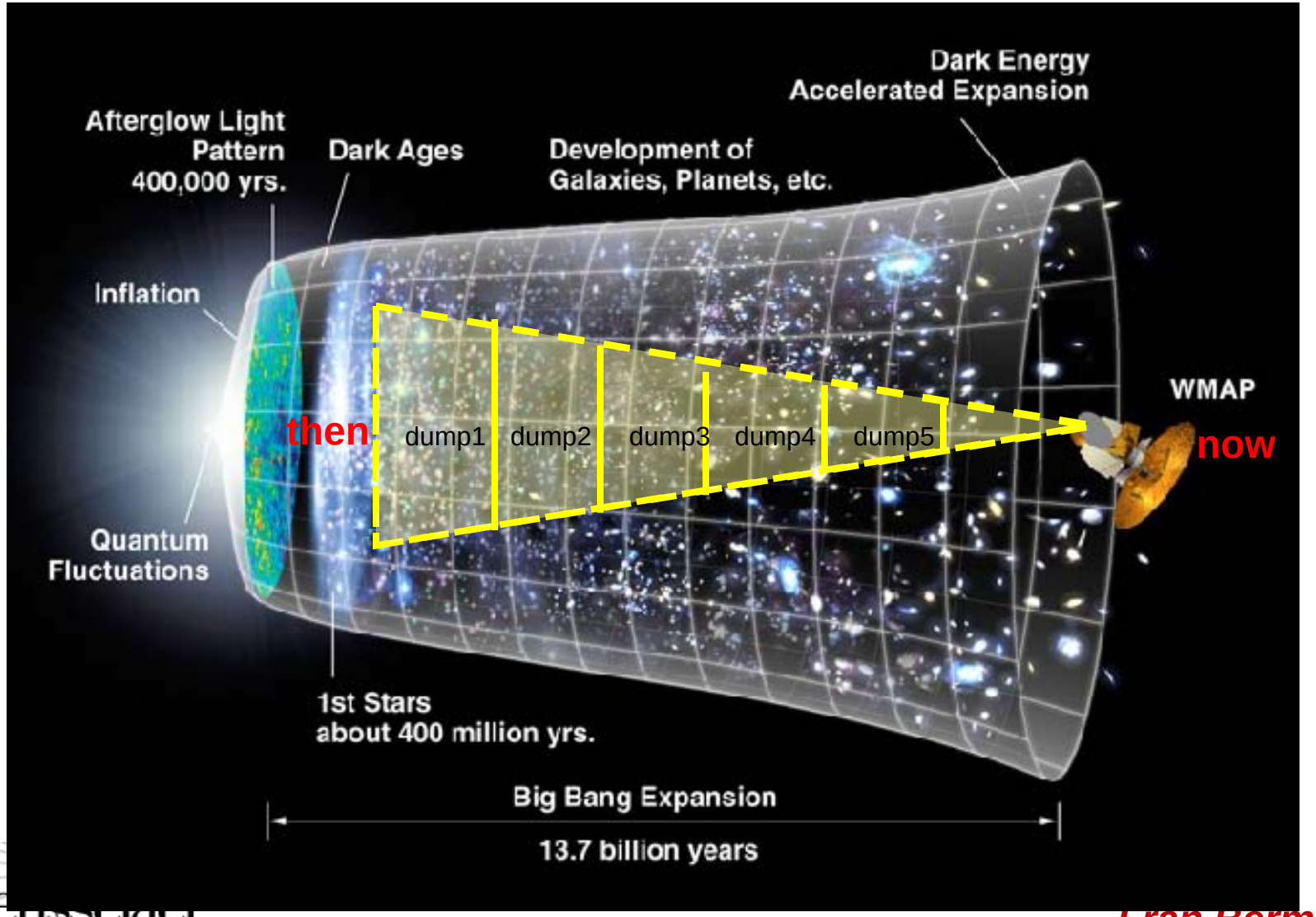
Today's Presentation

- **Digital Research Data -- Evolving the Universe after the Big Bang**
- **Supporting Digital Research Data**
- **Preserving Digital Data Over the Long Term**
- **Economics and Digital Preservation**

Digital Research Data

Research and Data: Evolving the Universe from the "Big Bang"

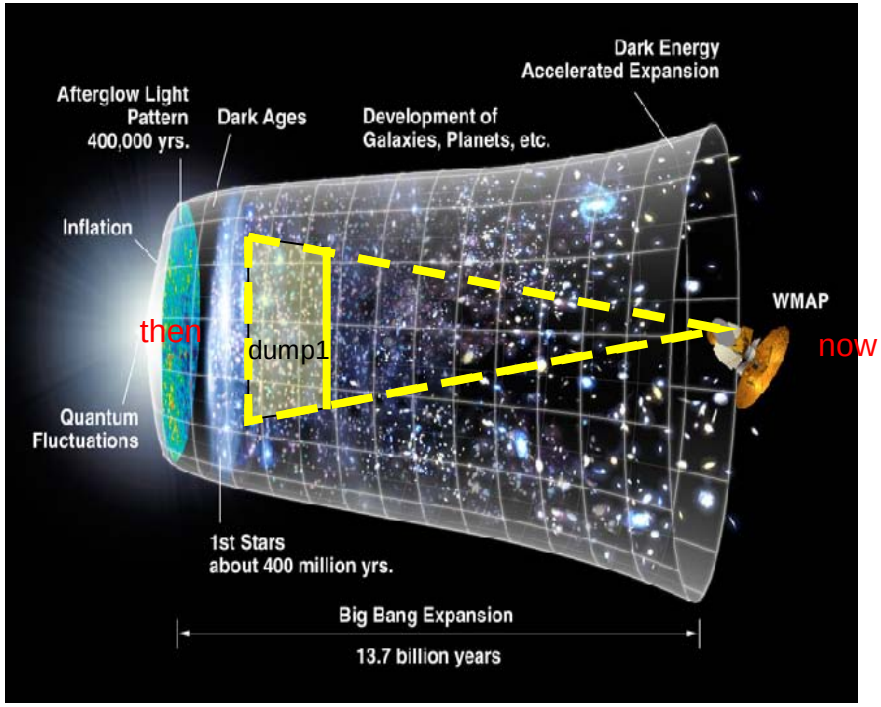
Composing simulation outputs from different timeframes builds up light-cone volume



Slide modified from Mike Norman, SDSC

After the "Big Bang" – the Universe's First Billion Years

- **ENZO** simulates the first billion years of cosmic evolution after the "Big Bang"
- Key period which represents
  - A tumultuous period of intense star formation *throughout the universe*
  - Synthesis of the first heavy elements in massive stars
  - Supernovae, gamma-ray bursts, seed black holes, and the corresponding growth of supermassive black holes and the birth of quasars
  - Assembly of first galaxies



Rensselaer
why not change the world?

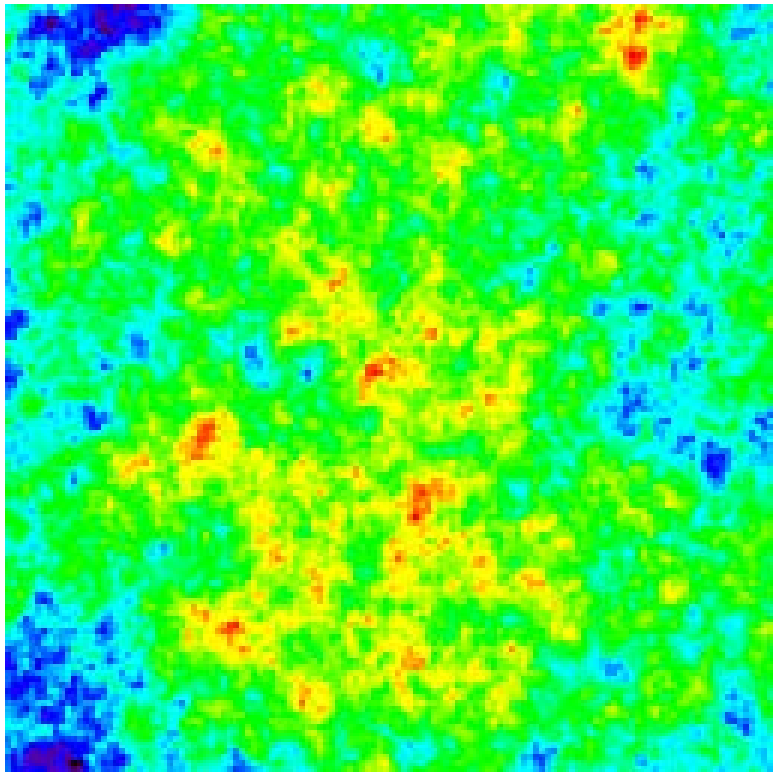
Slide modified from Mike Norman, SDSC

Fran Berman

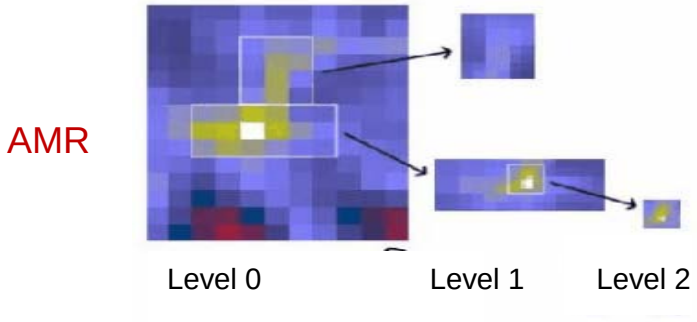
ENZO Simulations

What ENZO does:

- Calculates the growth of cosmic structure from seed perturbations to form stars, galaxies, and galaxy clusters, including simulation of
  - *Dark matter*
  - *Ordinary matter (atoms)*
  - *Self-gravity*
  - *Cosmic expansion*
- Uses adaptive mesh refinement (AMR) to provide high spatial resolution in 3D
  - The Santa Fe light cone simulation generated over 350,000 grids at 7 levels of refinement
  - **Effective resolution = $65{,}536^3$**



Formation of a galaxy cluster



Greater Simulation Accuracy Requires More Computing and Generates More Data

ENZO at Petascale

(10^15 calculations per second)

- Self-consistent **radiation**-hydro simulations of structural, chemical, and radiative evolution of the universe simulates from first stars to first galaxies

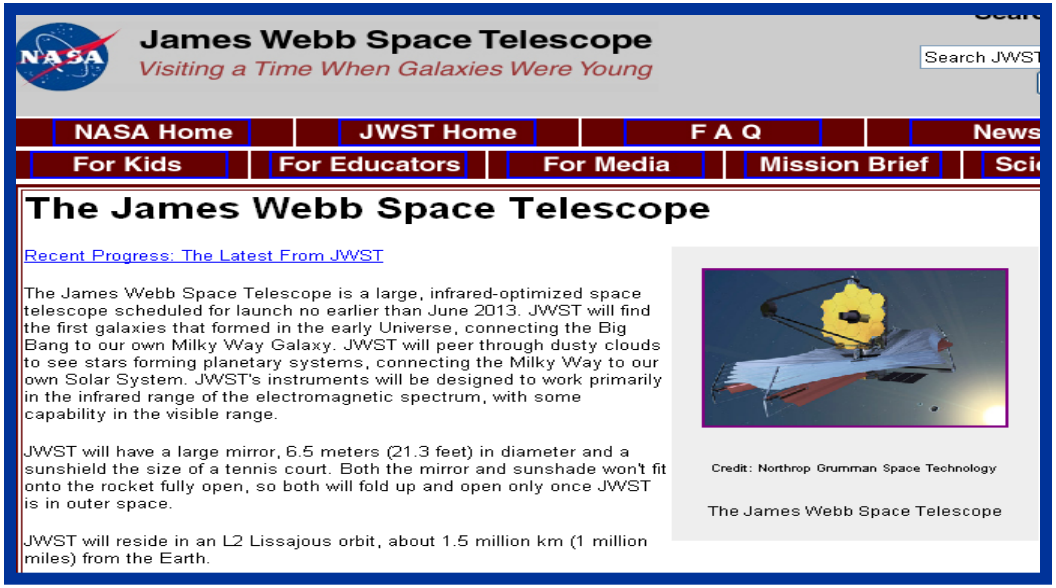
Computer Science challenges

- Parallelizing the grid hierarchy metadata for millions of subgrids distributed across 10s of thousands of cores
- Efficient dynamic load balancing of the numerical computations, taking memory hierarchy and latencies into account
- Efficient parallel “packed AMR” I/O for 100 TB data dumps
- Inline data analysis/viz. to reduce I/O

Slide modified from Mike Norman, image by Robert Harkness

Verifying Theory with Observation

- **James Webb Space Telescope**, coming in 2013 will probe the first billion years of the universe – providing observations of unprecedented depth and breadth
- **Simulation data** will enable tight integration of observation and theory, and will enable simulations to approach realistic complexity
- Analysis of **petascale data sets** will be essential for validating model



Rensselaer why r

Slide modified from Mike Norman, image by Robert Harkness

Fran Berman

Supporting Digital Research Data

Rensselaer

why not change the world? SM

Fran Berman

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

Create

Data creation / capture / gathering from

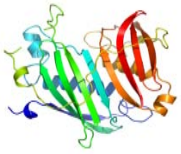
- laboratory experiments
- fieldwork
- surveys
- devices
- media
- simulation output ...

Edit

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

Use / Reuse

- Analyze
- Mine
- Model
- Derive additional data
- Visualize
- Input to instruments / computers / devices ....



Publish

- Disseminate
- Create portals / data collections / databases
- Associate with literature ....

Preserve / Destroy

- Store / preserve
- Store / replicate / preserve
- Store / ignore
- Destroy ....

Rensselaer

why not change the world? SM

Information adapted from Chris Rusbridge and Liz Lyon

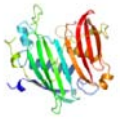
Fran Berman

Data Cyberinfrastructure: Access and services enable researchers to get the most out of their data

Coordinated systems make innovation the challenge rather than IT

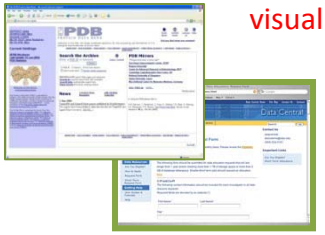
modeling

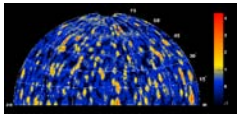
analysis



simulation

visualization

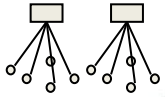




File systems, Database systems,
Collection Management
Data Integration, etc.

Many Data Sources

instruments



Sensor-nets

computers

- Data Access
- Data Use
- Data Management
- Data Storage
- Data Preservation

Services Are Critical for Use

- *Data visualization*
- *Portal creation and collection publication*
- *Data analysis*
- *Data mining*
- *Data hosting*
- *Preservation services*
- *Domain-specific tools*
  - *Biology Workbench*
  - *Montage (astronomy mosaicking)*
  - *Kepler (Workflow management)*
- *Data anonymisation, etc.*

Rensselaer

why not change the world? SM

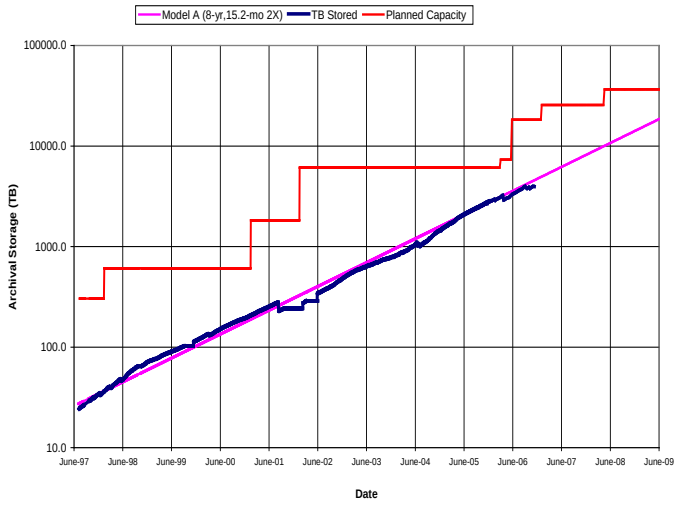
Fran Berman

Reliable Data Cyberinfrastructure Incurs Real Costs

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

Resources and Resource Refresh



SDSC Data Storage Growth

- *Most valuable data must be replicated*
- *SDSC research collections doubled every 15 months.*
- *SDSC storage is 36+ PB*

Digital Research Data: One Size Does Not Fit All

- ***RETENTION TIMEFRAME:***
  Short-term (few months, years) to long-term (decades, centuries, …)
- ***SIZE / SCALE:***
  Small-scale (GBs) to large-scale (PBs)
- ***PREPARATION:***
  Well-tended (metadata, cleaned and filtered) to poorly tended (flat files, insufficient metadata)
- ***POLICY / REGULATION RESTRICTIONS:***
  Subject to more restrictive policy and regulation (HIPAA) vs. subject to less restrictive policy and regulation (OMB)
- ***LIFE CYCLE PLANNING:***
  Has a data management and sustainability plan (PDB, PSID, NVO) vs. ad hoc approach
- ***STANDARDIZATION:***
  Organized using community standards vs. ad hoc or home-grown

Opportunity for Greater Synergy between Modern Researcher Needs and Traditional Library Strengths

- Research community characterized by culture of innovation
  - Periodic new starts
  - Experimentation
  - Customized solutions to ill-defined problems
  - Collaboration and competition

- Researchers need help with things Librarians are good at
  - Developing reliable management, preservation and use environments
  - Proper curation and annotation
  - Navigating policy, regulation, intellectual property
  - Collaboration (partnership to share resources, create economies of scale, etc.)
  - Sustainability

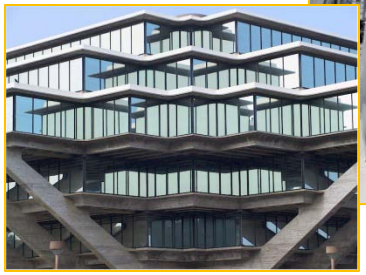
The "Local" Digital Research Data Repository: Emerging Role for University Libraries

- Researchers are increasingly required to retain the digital products of their research, University libraries can play a new role as local stewards of digital research data.

- A "Preservation Stimulus" may be needed to make this realistically viable on a broad scale.

Data Mgt. Plans

Data.gov



Digital "stacks"

Rensselaer
why not change the world? SM

Fran Berman

Preserving Digital Research Data Over the Long Term

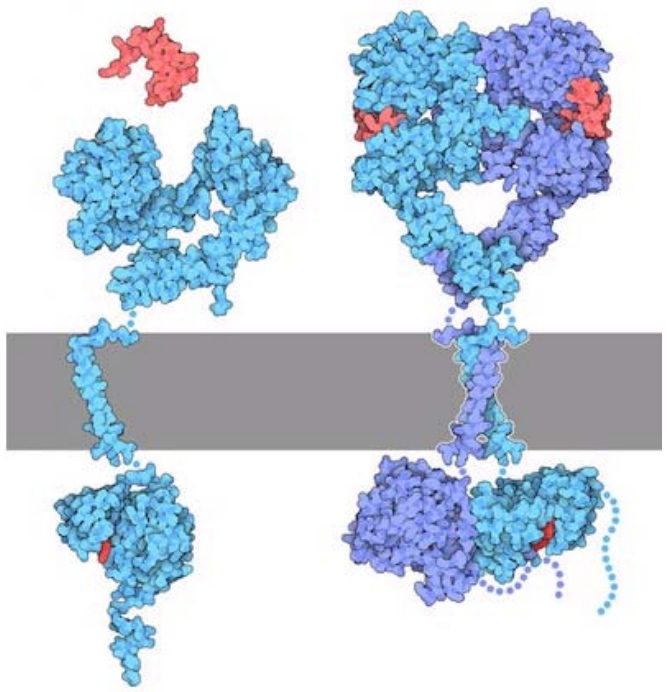
Rensselaer

why not change the world? SM

Fran Berman

Life Sciences Data

- **The Protein Data Bank**
  - worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids.
- PDB represents $80 billion + investment in research resulting in PDB structures
- PDB supported by funds from NSF, NIGMS, DOE, NLM, NCI, NCRR, NIBIB, NINDS, NIDDK.

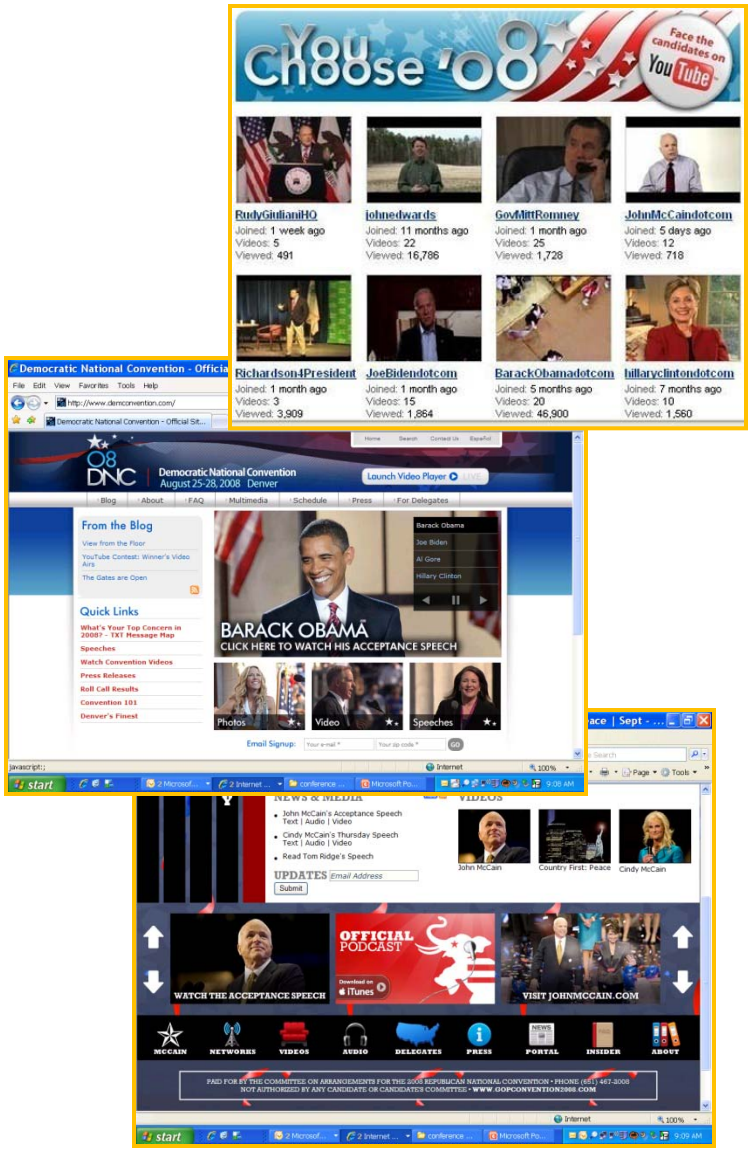


June Molecule of the Month: **Epidermal Growth Factor**

"The cells in your body constantly communicate with each other, negotiating the transport and use of resources and deciding when to grow, when to rest, and when to die. Often, these messages are carried by small proteins, such as epidermal growth factor (EGF), shown here in red from PDB entry 1egf. EGF is a message telling cells that they have permission to grow. … "

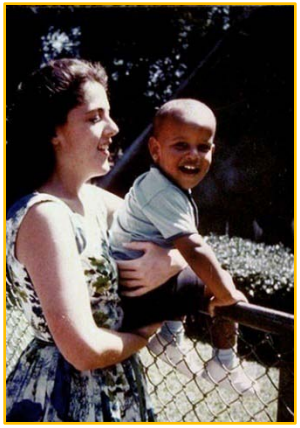
Historical Data

- **The 2008 Cyber-election**
  - Fundraising via website
  - YouTube videos of the candidates and conventions
  - Blogs as vehicles for discussing issues
  - On-line organizing
- Digital data from historic 2008 cyber-election will be valuable for **decades+ to come**



Cultural Data

- **Historical photographs**



Rensselaer

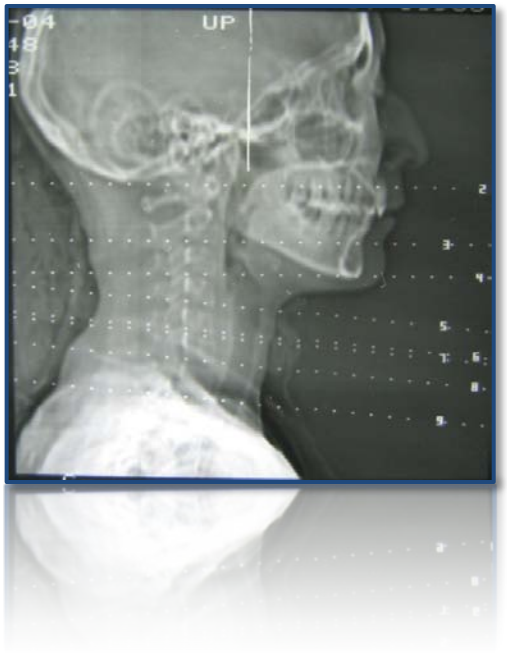
Fran Berman

Some images courtesy of David Minor and the Library of Congress

Access to Information Tomorrow Requires Preservation Today

- **Digital Access and Preservation** is a technical, management, policy, regulatory, social, and economic problem

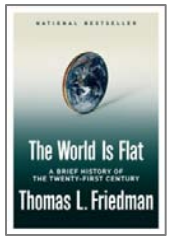
- Key issues to resolve:
  1. **What should we preserve?**
  2. **Who is responsible** for digital information?
  3. **Who pays** for digital information and its supporting cyberinfrastructure?



What Should We Save?
Saving Everything Isn't An Option …

U.S. Library of Congress manages over 300 TB of digital data,

YouTube: 6M videos in 2006 = 600 TB

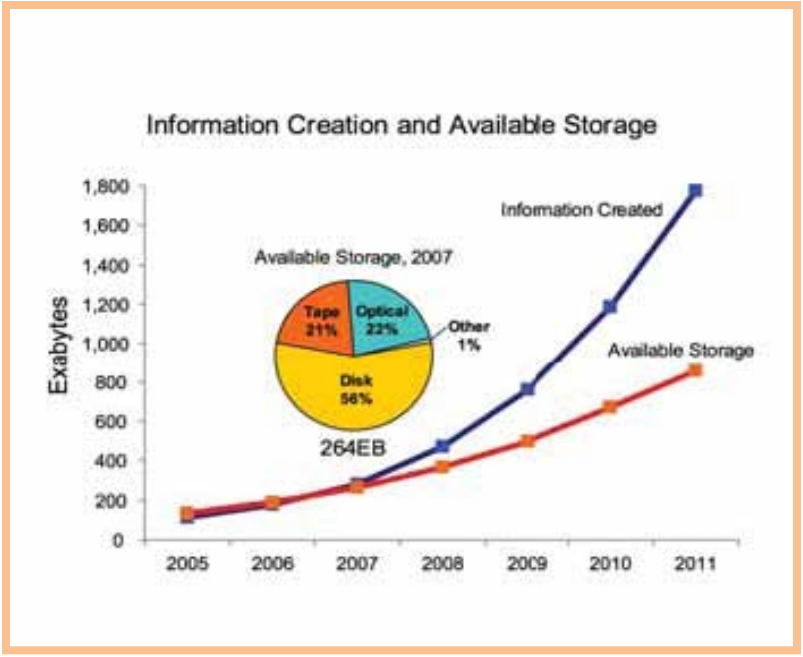


1 novel = 1 MB

SDSC Tape Archives = 36+ PB

- **2007 was the "crossover year"** where the amount of digital information exceeded the amount of available storage (~264 exabytes)
- *By 2023, the amount of digital data will exceed Avogadro's number. (6.02 X 10^23, the number of atoms in 12 grams of carbon).*

Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}
Zetta	10^{21}



Rensselaer

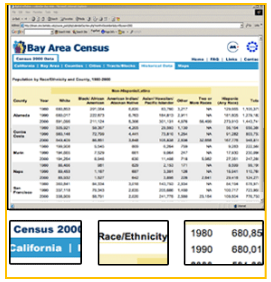
Fran Berman

Source: *"The Diverse and Exploding Digital Universe" IDC Whitepaper, March 2008*

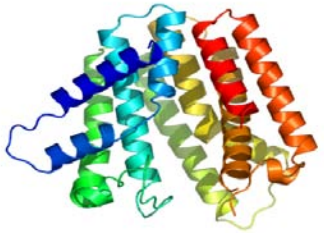
What do We Want to Save?

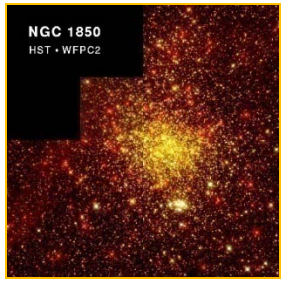
Data we* want to keep over the long-term:

- **We = "Society"**
  - Official and historically valuable data (Census information, presidential emails, Shoah Collection, etc.)
- **We = Research Community**
  - Protein Data Bank, National Virtual Observatory, etc.
- **We = Me**
  - My medical record, my Quicken data, digital family photos, etc.



Sarbanes-Oxley
Financial and Accounting Disclosure Information





Rensselaer
why not change the world? SM

Fran Berman

What do We Have to Save?

- **HIPAA** applies to health information created or maintained by health care providers
- **Sarbanes-Oxley** regulations apply to all U.S. public company boards, management, and public accounting firms.
- **OMB** regulations apply to federally funded research data (NIH, NSF, DOE, etc.)

Crime and Punishment

Regulations	Retention Requirement	Penalty
HIPAA	Retain patient data for 6 years	$250K fine and up to 10 years in prison
Sarbanes-Oxley	Auditors must retain relevant data for at least 7 years	Fines to $5M and 20 years in prison
Gramm-Leach-Baily	Ensure confidentiality of customer financial information	Up to $500K and 10 years in prison
SEC 17a	Broker data retention for 3-6 years. Some require longer retention	Variable based on violation
OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)	"a three year period is the minimum amount of time that research data should be kept by the grantee"	Penalty structure unclear, likely fines?

Table information partly based on "Data Retention – More Value, Less Filling", John Murphy, http://www.tdan.com/view-articles/5222

Fran Berman

Who Will Pay?
Economics and Digital Preservation

Rensselaer
why not change the world? SM

Fran Berman

Responsibility and Economics: Blue Ribbon Task Force on Sustainable Digital Preservation and Access

brtf.sdsc.edu

BRTF Charge:

1. Conduct a comprehensive **analysis** of sustainable digital preservation
2. Identify and evaluate **best practices**
3. Make specific **recommendations for action**
4. Articulate **next steps** for further work

Fran Berman

Rensselaer
why not change the world? SM

BRTF Participants

Blue Ribbon Task Force:

- Paul Ayris, University College London
- Fran Berman, SDSC/UCSD
- Bob Chadduck, NARA Liaison
- Sayeed Choudhury, Johns Hopkins University
- Elizabeth Cohen, AMPAS/Stanford
- Paul Courant, University of Michigan
- Lee Dirks, Microsoft
- Amy Friedlander, CLIR
- Chris Greer, NITRD Liaison
- Vijay Gurbaxani, UC Irvine
- Anita Jones, University of Virginia
- Ann Kerr, Consultant
- Brian Lavoie, OCLC
- Cliff Lynch, CNI
- Dan Rubinfeld, UC Berkeley
- Chris Rusbridge, DCC
- Roger Schonfeld, Ithaka
- Abby Smith, Consultant
- Anne Van Camp, Smithsonian

Sponsoring Agencies/Institutions:

- National Science Foundation
- Mellon Foundation
- Library of Congress
- National Archives and Records Administration
- CLIR
- NITRD
- JISC
- Member institutions

Specific Responsibilities

- Fran Berman / co-Chair
- Amy Friedlander / First Report Editor
- Ann Kerr / January Panel Rapporteur
- Brian Lavoie / co-Chair
- Susan Rathbun / Task Force Support
- Abby Smith / Second Report Editor
- Jan Zverina / Communications Lead
- Lucy Nowell / NSF Program Officer
- Don Waters / Mellon Program Officer
- Laura Campbell, Martha Anderson / LC representatives

What is required to support digital information over the long term?

Economic sustainability for digital information* requires

- **Recognition of the benefits** of long-term access and preservation
- **Incentives** for decision-makers to act
- **Means of selecting "valued" information** for long-term preservation
- Mechanisms to support **ongoing, efficient allocation of resources**
- Appropriate **organization and governance** of preservation and access activities

Rensselaer
why not change the world? SM

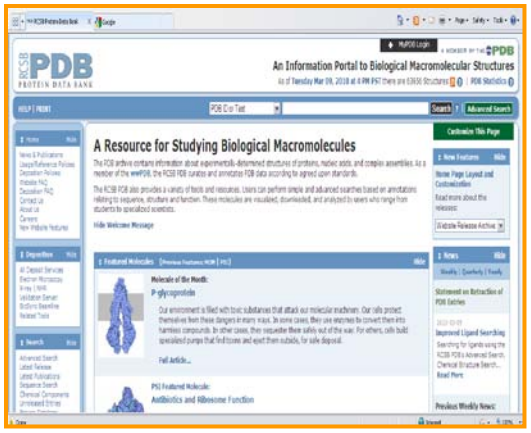
** From Blue Ribbon Task Force Interim Report*

Fran Berman

Who's Paying the Bills?

- The "free rider" *non*-solution: **"Let X do it"** where **X** is:
  - The Government
  - The Libraries
  - The Archivists
  - Google, Microsoft, etc.
  - Data users
  - Data owners
  - Data creators, etc.

How do we currently support access to digital information?



Federal grants

Advertisements

Subscription

Pay per service

Donations, etc.

The Stakeholder Problem

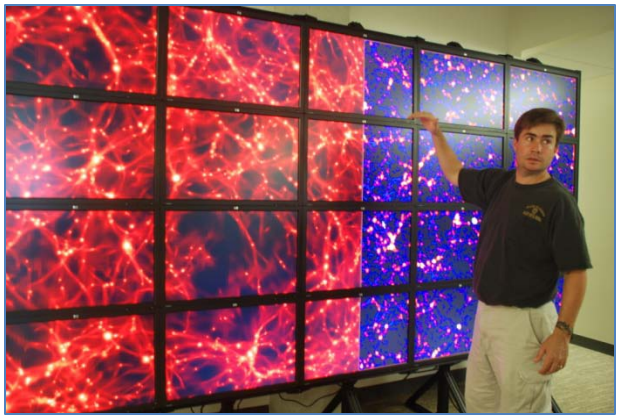
- **Many Stakeholders in digital preservation …**
  - **Stakeholders who benefit** from use of the preserved asset
  - **Stakeholders who select** what to preserve
  - **Stakeholders who own** the asset
  - **Stakeholders who preserve** the asset
  - **Stakeholders who pay**
- ***The greater the alignment between key stakeholder groups, the better the prospects for sustainable preservation***

4 Common Stakeholder Scenarios

- *Research data*
- *Scholarly discourse*
- *Commercially-owned Cultural content*
- *Collectively-produced web content*

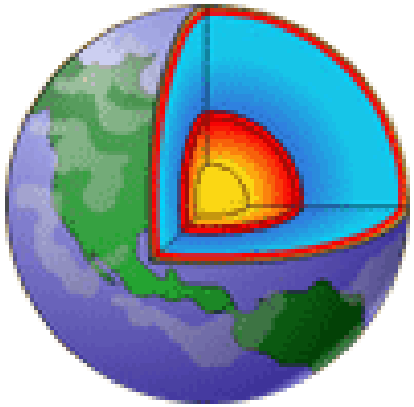
Research Data

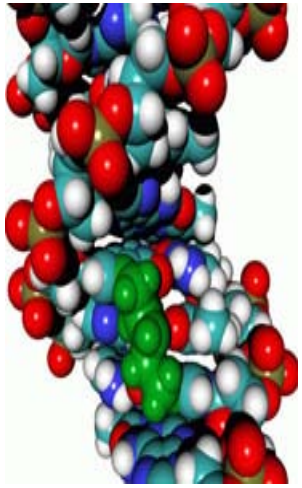
- ***Stakeholders who benefit:*** the greater research community
- ***Stakeholders who select:*** Often the individuals who generate the data
- ***Stakeholders who own:*** Often the data generators
- ***Stakeholders who preserve:*** Often the data generators and their proxies
- ***Stakeholders who pay:*** Federal agencies, institutions



Needed actions involve

- *the development of federal agency policies that mandate the stewardship of important research data*
- *the identification of viable support options for third-party archives (e.g. university libraries) to host valuable research data*



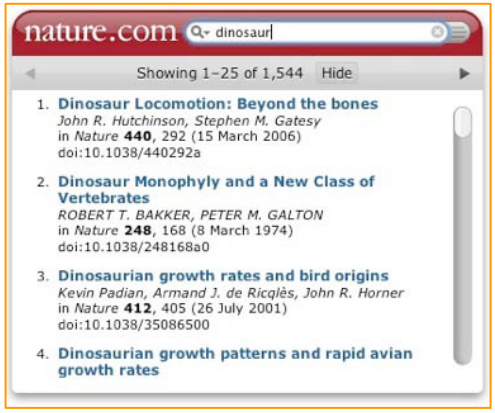


Scholarly Discourse

- ***Stakeholders who benefit:*** the greater research and learning community
- ***Stakeholders who select:*** Publishers, based on community review
- ***Stakeholders who own:*** Publishers generally own rights
- ***Stakeholders who preserve:*** Publishers and third-party entities
- ***Stakeholders who pay:*** Publishers, libraries, and third-party entities

Needed actions involve

- Clarification (with respect to licensing, ownership, rights, etc.) of the responsibilities of publishers, third-party archives, and scholars
- Granting of non-excusive rights to content by scholars to enable decentralization of publishing and preservation.



Commercially Owned Cultural Content

- ***Stakeholders who benefit:*** the general public, cultural historians
- ***Stakeholders who select:*** Studios, third-party organizations
- ***Stakeholders who own:*** Studios, third-party organizations
- ***Stakeholders who preserve:*** Institutional and individual repositories, third-party organizations, etc.
- ***Stakeholders who pay:*** Studios, professional organizations, private owners, custodial organizations, etc.

Needed actions involve

- *Alignment of requirements for copyright deposit with the requirements of digital preservation and access*
- *Development and involvement of organizations that can ensure secure handoffs of cultural materials from private owners to economically viable public preservers*

Collectively-produced Web Content

- ***Stakeholders who benefit:*** the general public, cultural historians, etc.
- ***Stakeholders who select:*** Often the entities that preserve the data
- ***Stakeholders who own:*** Often unclear
- ***Stakeholders who preserve:*** Third parties interested in preservation of cultural assets
- ***Stakeholders who pay:*** Third parties interested in the preservation of cultural assets

Needed actions involve

- *the development of appropriate licensing and regulations that permit third-parties to preserve web content*
- *the development of incentives for host sites or third parties to preserve*

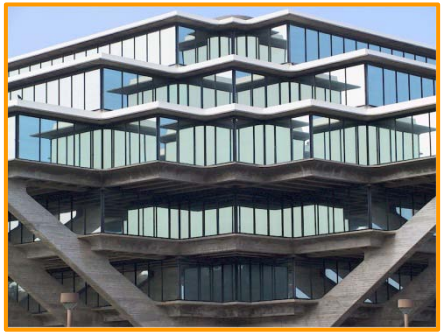
An Action Agenda for Trusted International, National and Public Institutions

1. Create **mechanisms for public-private partnerships** to align distinct groups: Convene stakeholders, sponsor cooperation and collaboration, etc.

2. **Convene expert communities to address the selection and preservation** needs of valuable materials for which there is no stewardship (Web materials, digital orphans).

3. Act expeditiously to **reform national and international copyright legislation** to address digital preservation needs.

4. **Create financial incentives** to encourage private entities to preserve digital materials on the public behalf.



Fran Berman

An Action Agenda for Funders and Sponsors of Data Creation

1. **Create preservation mandates** when possible
2. **Invest in building / seeding stewardship capacity** throughout the system.
   - **Fund the modeling and testing** of domain-specific preservation strategies
3. **Provide leadership in training and education** for 21st century preservation, including domain expertise and core competencies in STEM. Promote digital preservation skills.

An Action Agenda for Organizations and Individuals

Organizations

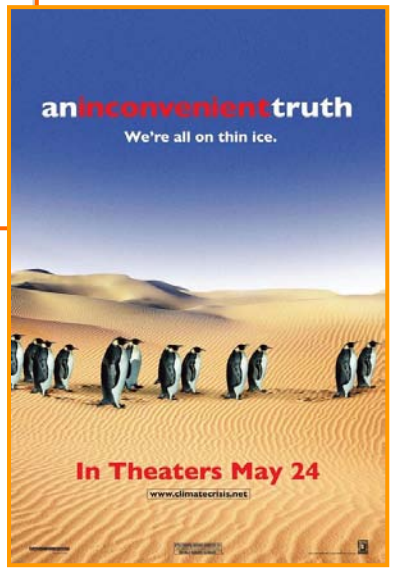
1. **Fund internal preservation and access activities as core infrastructure.**
2. Create economies of scope and economies of scale by partnering with related organizations and industry professional associations.
3. Develop preservation strategies that reflect technical, policy, and workforce best practices

Individuals

1. **Provide nonexclusive rights to preserve and distribute created content.**
2. Partner with preservation experts *throughout your data's lifecycle* to ensure that data is ready to hand off in a form that will be useful over the long term.
3. Pro-actively participate in professional societies and relevant organizations to create stewardship best practices and selection priorities.

Our responsibility: Making the Case

- **To Decision Makers:**
  - What are liabilities and the opportunity costs of *not* acting?
  - What specific actions need to be made a priority *now*?
- **To the General Public:**
  - Does your dry cleaner know what digital preservation is?

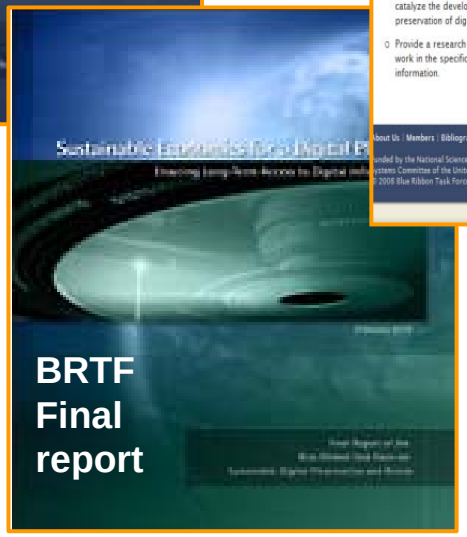


Rensselaer
why not change the world? SM

Fran Berman

Thank You

www.rpi.edu



brtf.sdsc.edu

Rensselaer
why not change the world? SM

Fran Berman