Creating Technology for a Successful Future: A Prologue

Dr. Fran Berman

Vice President for Research

Rensselaer Polytechnic Institute

Rensselaer

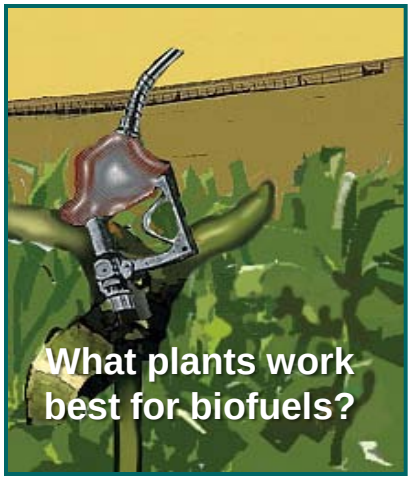
why not change the world? SM

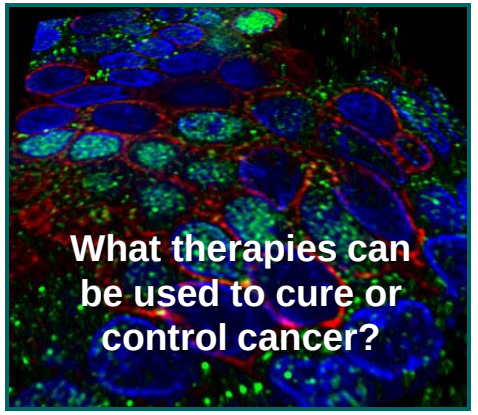
Creating a Successful Future:

Science, Engineering, and Technology Matter

"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

Foundation for a Better World

Computers for the Third World

Mary Lou Jepsen,
founding CTO,
One Laptop Per Child

Mobility for amputees

Van Phillips,
inventor of the *C-leg*

Robots for dangerous situations

Helen Greiner, co-founder, *iRobot*

Mobile technologies for increasing information available on population health

Joel Selanikio and
Rose Donna co-
founders, *DataDyne*

Rensselaer
why not change the world? SM

Fran Berman

New and Expanded Opportunities Now to Make an Impact

- Greater focus on Science, Engineering, Technology, Education in all branches of Govt.
  - Increased Opportunity for science and technology community to guide national-scale solutions for societal problems
- Increased engagement of OSTP in Administration priorities
- President's Council of Advisors on Science and Technology (**PCAST**) tasked with addressing most pressing national priorities (health, energy, climate, manufacturing, etc.)

PCAST 2009

OSTP S&T Budget Priorities:

- Build on American Recovery and Reinvestment Act (**ARRA**) and focus on S&T strategies that can be used to drive economic recovery
- Focus on science and technology to address sustainable **energy**, health-focused **biomedical research**, critical and cyber-**infrastructure**, and advanced capabilities in **space**
- Improve STEM **education**

Rensselaer

why not change the world? SM

Fran Berman

21st Century Solutions
Driven by 21st Century Tools

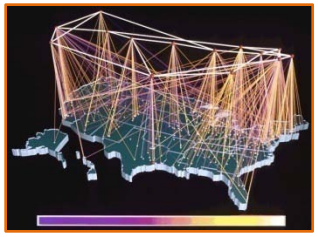
New materials

Sensors

Mobile devices

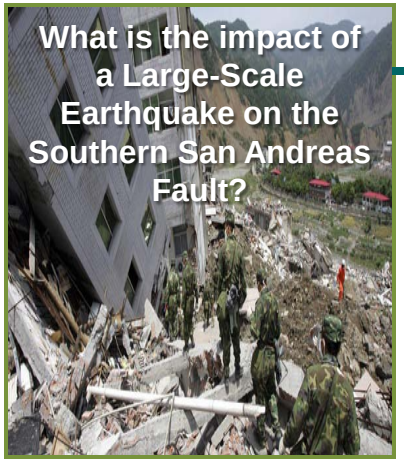
Scientific Instruments

Computers, Networks, Data



Rensselaer
why not change the world? SM

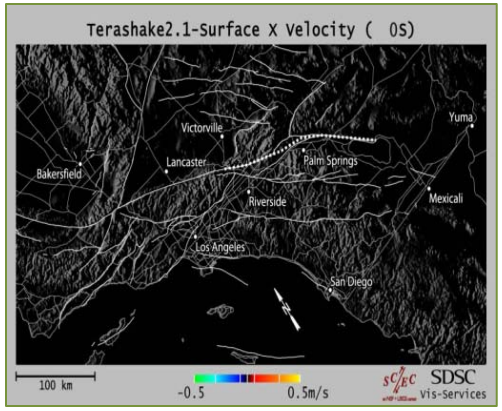
From Challenge to Solution: Earthquake Impact Prediction



Computer model used to predict seismic activity, parameterized by Southern California sensor data

High performance computer and large-scale data storage needed to run high-resolution model

Scientific visualizations of seismic predictions require additional computation

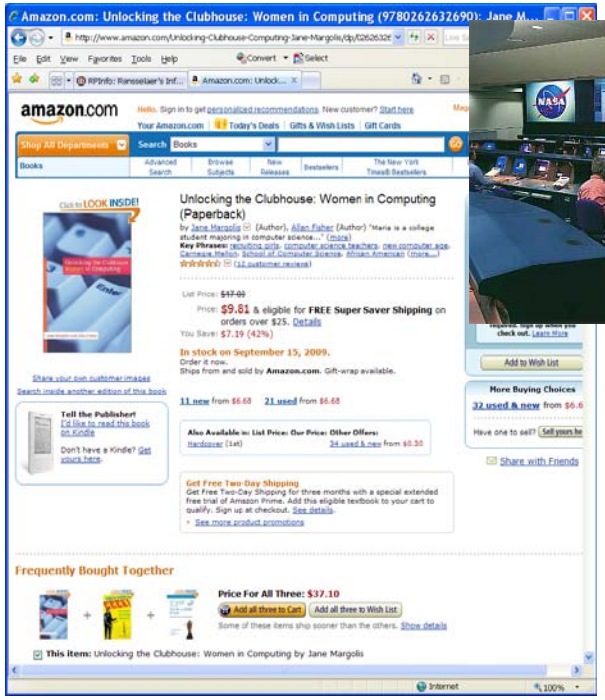


Rensselaer
why not change the world? SM

Fran Berman

Some images courtesy of Amit Chourasia

Tomorrow's discoveries require even more capability, functionality, capacity



Data →
Information →
Knowledge

Boundaries between domains are blurring, bringing new opportunities for innovation, agility, and synergy

Cyber-
Bio-
physical
systems

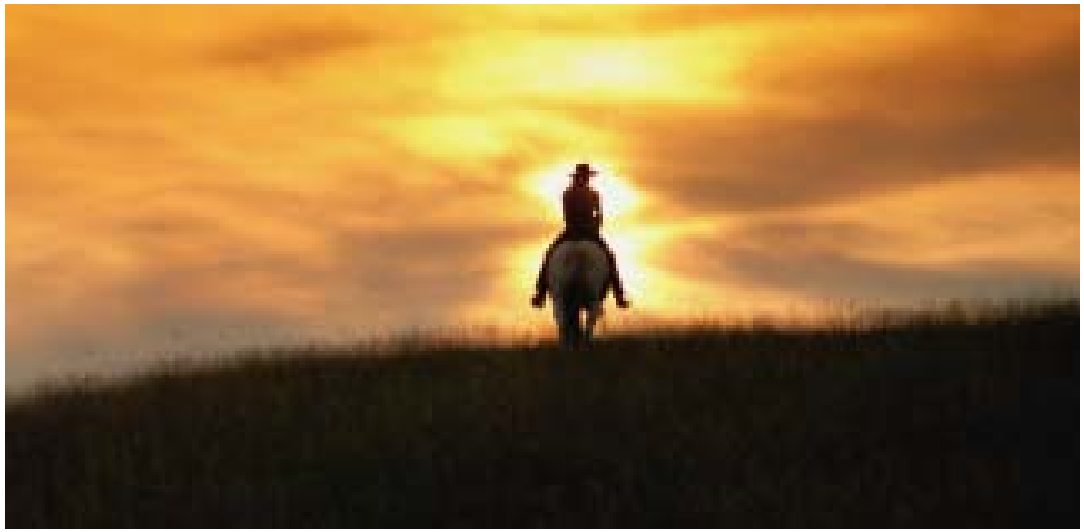
Social networks,
Mass communication

Rensselaer
why not change the world? SM

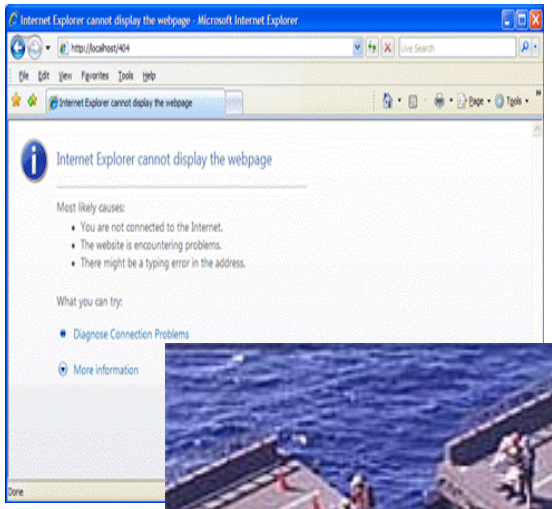
Fran Berman

Are We There Yet?

- Hopes are high that we will meet our ambitious and critical goals and technical aspirations



- What is required to be successful?
- Why do our efforts oftentimes fall short?



Rensselaer
why not change the world? SM

Fran Berman

Creating a Strong Foundation for Future Success

1. Creating a strong foundation for science, engineering, and technology ***efforts*** in the Information Age

2. Creating a strong foundation for future science, engineering, and technology ***leadership***

Rensselaer
why not change the world? SM

Fran Berman

The Information Age

*"The **Information Age** [is] characterized by the ability of individuals to transfer information freely, and to have instant access to knowledge that would have been difficult or impossible to find previously."*

Wikipedia

- At the heart of successful efforts in the Information Age is digital information itself, and the ability to **access and preserve** that information as needed

- ***Focus:*** Digital information and the environment needed to support its availability, preservation and use.

Rensselaer

why not change the world? SM

The Digital World 1

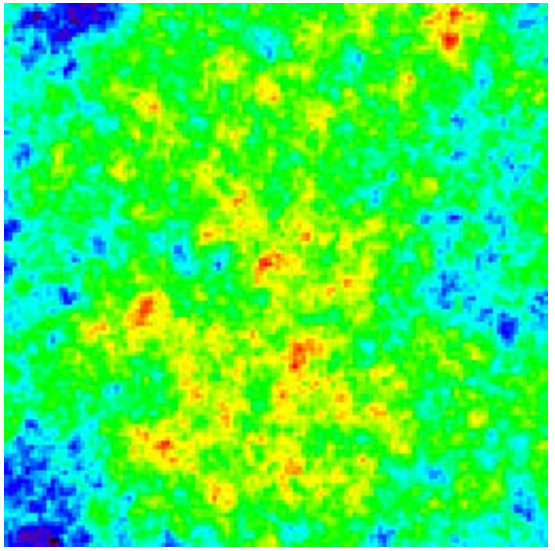
- **The 2008 Cyber-election**
  - Fundraising via website
  - YouTube videos of the candidates and conventions
  - Blogs as vehicles for discussing issues
  - On-line organizing
- Digital data from historic 2008 cyber-election is valuable for **decades+ to come**

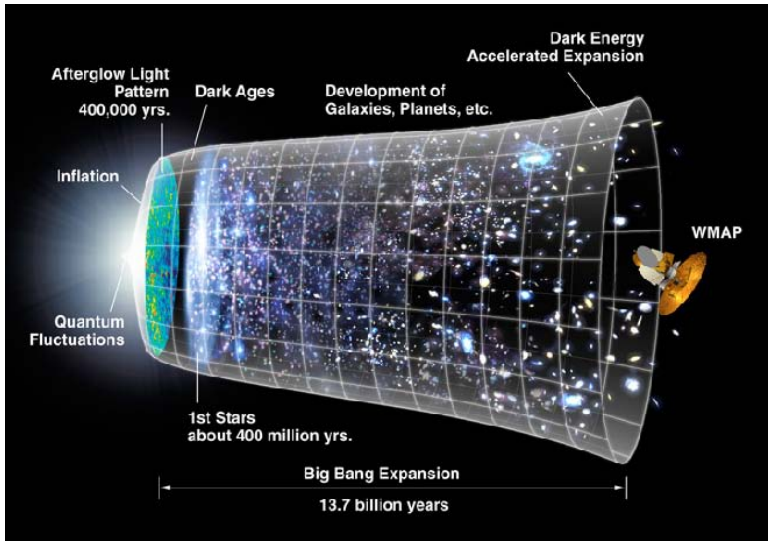
Rensselaer
why not change the world? SM

Fran Berman

The Digital World 2

- **Science: The First Billion Years After the Big Bang**
  - 400 TB of data produced from ENZO astrophysics simulations
  - Data will be mined and analyzed, of great value for **several years** after computation
- Simulation results illustrate growth of stars, galaxies, and galaxy clusters, dark matter, etc. after the Big Bang
- Large-scale simulations "refreshed" as resources become available





Rensselaer
why not change the world? SM

Fran Berman

Images courtesy of Mike Norman

The Digital World 3

- **Family Photographs**

Rensselaer

why not change the world? SM

Fran Berman

Some images courtesy of the Library of Congress

How Much Digital Data is There?

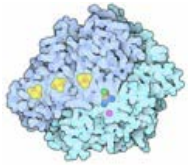
U.S. Library of Congress manages **295+ terabytes** of digital data, 230+ of which are "born digital"

SDSC Tape Archives = **36+ petabytes capacity**

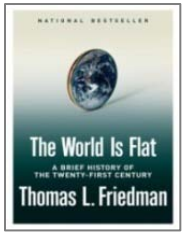
By 2023, the amount of digital data will exceed **Avogadro's number**. (6.02 X 10^23 = number of atoms in 12 grams of carbon)

Google Earth =**71+ terabytes**

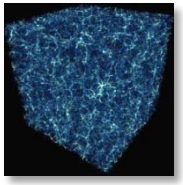
Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}
Zetta	10^{21}



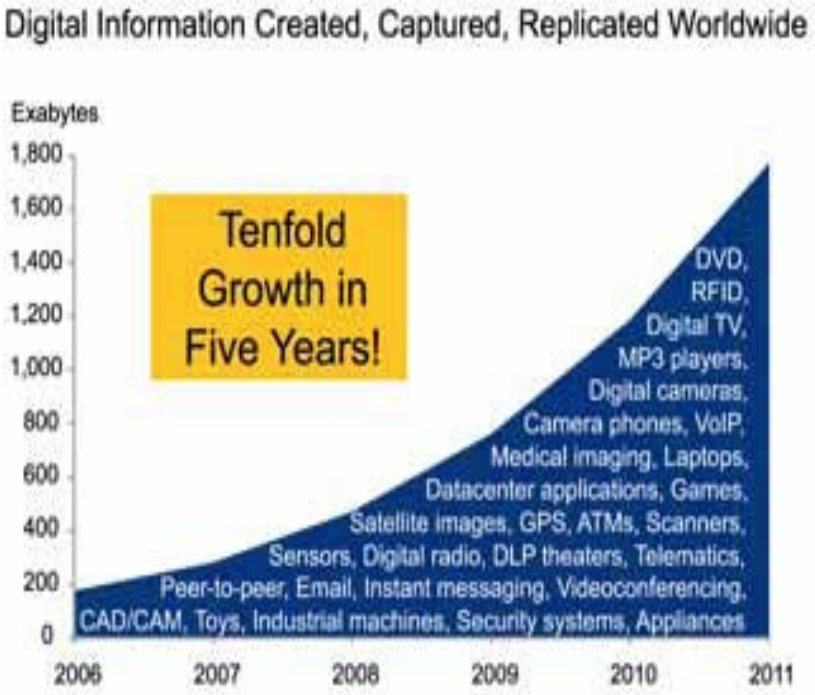
50,000 Protein Data Bank Structures = **35 terabytes**



1 novel = **1 megabyte**



Stored data from ENZO cosmological simulations = **500 terabytes**



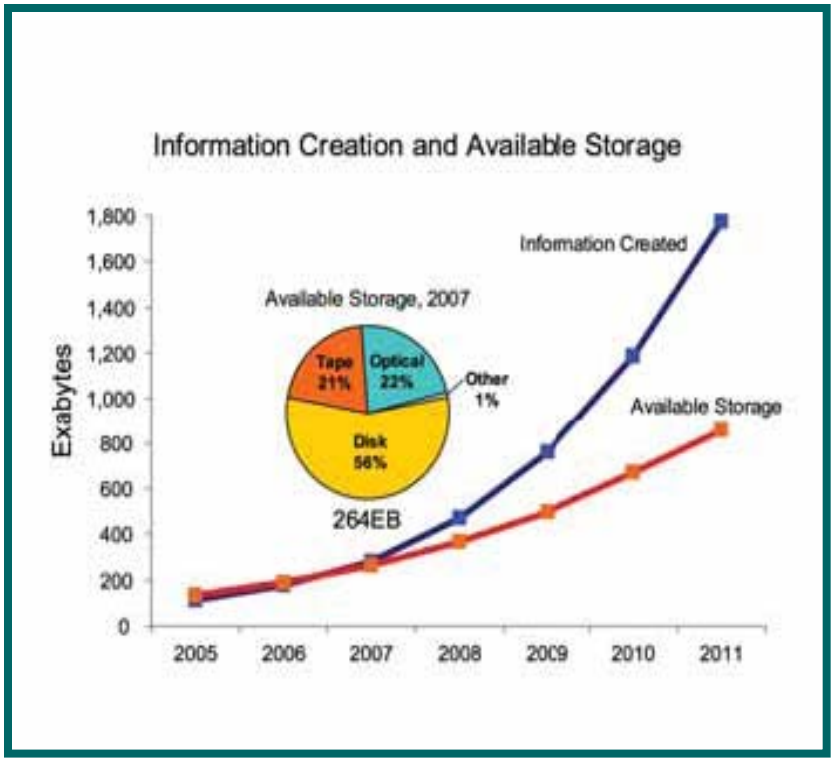
Rensselaer
why not change the world?℠

Fran Berman

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

Running Out of Room

- We may be generating unimaginable amounts of data, **but we can't save it all.**
- **2007 was the "crossover year"** where the amount of digital information became greater than the amount of available storage
- Importance of digital data and the need to make choices mandates a **more thoughtful approach to data stewardship** in the Information Age



Rensselaer

why not change the world? SM

Fran Berman

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

The Data Life Cycle

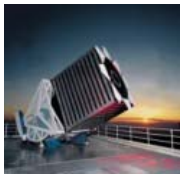
Create

Data creation / capture / gathering from

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

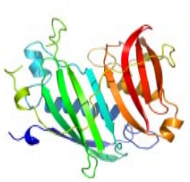
Edit

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



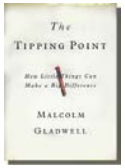
Use / Reuse

- 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 ....

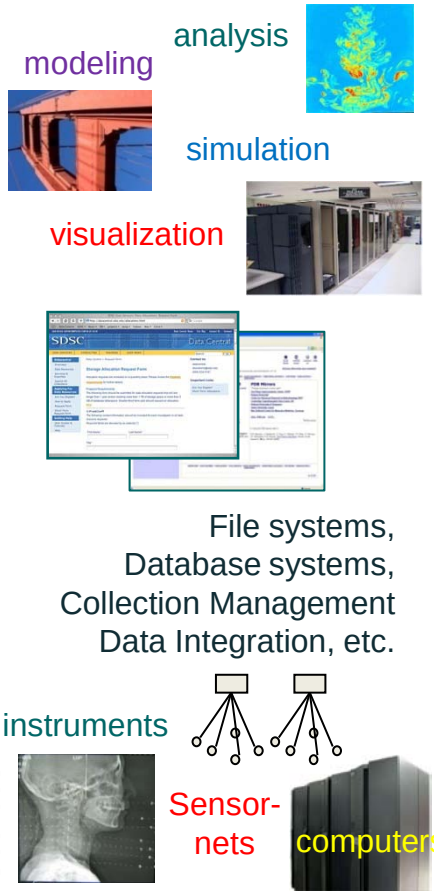
Rensselaer

why not change the world? SM

Fran Berman

Information adapted from Chris Rusbridge and Liz Lyon,

Coordinated Information Technologies Needed to Support the Data Life Cycle



Data Access

Data Use

Data Management

Data Storage

SDSC Integrated Data Cyberinfrastructure

Services

- Database selection and schema design
- Portal creation and collection publication
- Data mining
- Storage services
- Preservation services
- Domain-specific tools
  - Biology Workbench
  - Montage (astronomy mosaicking)
  - Kepler (Workflow management)
- Data visualization, etc.

Rensselaer
why not change the world? SM

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

Fran Berman

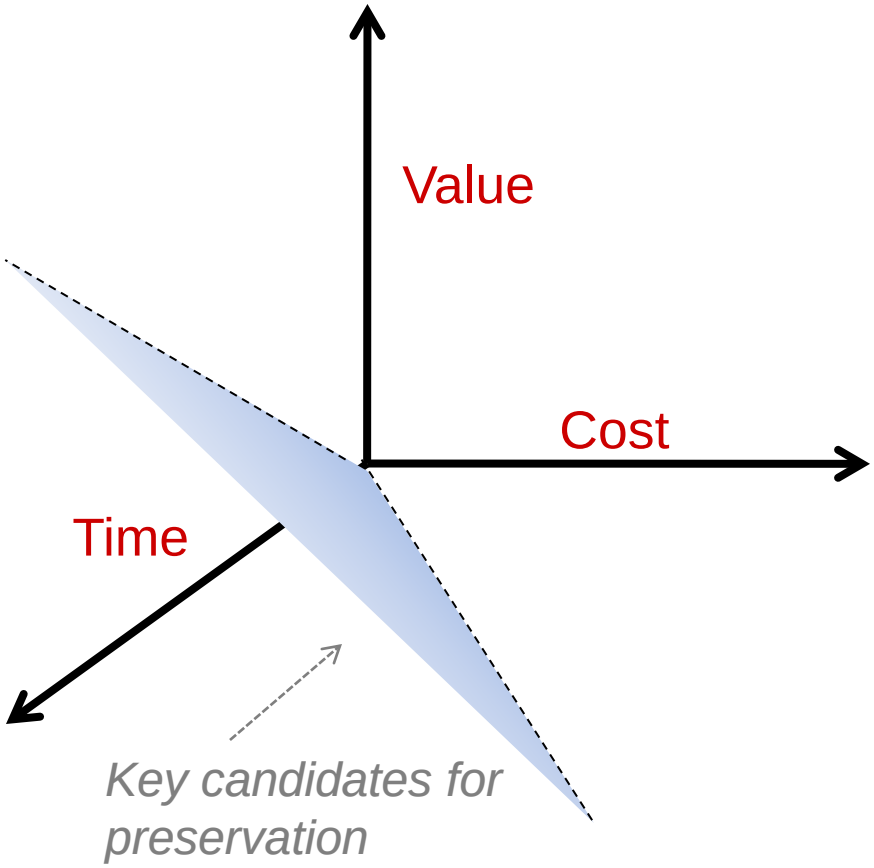
Access to Data Tomorrow Requires Preservation of Data Today

- **Key Questions:**

  1) What should we save? – ***value, policy, regulation***

  2) How should we save it? – ***technology, best practice***

  3) Who should pay for it? -- ***economics***

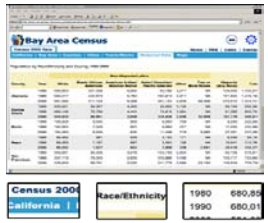


What Should We Save?

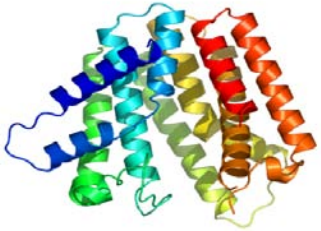
Data we* want to keep over the long-term:

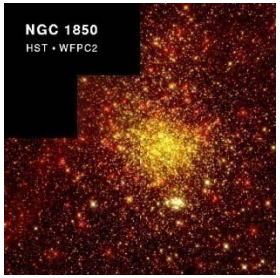
- **We = "Society"**
  - Official and historically valuable data (Census information, presidential emails, Shoah Collection, etc.)
- **We = Research Community**
  - Protein Data Bank, National Virtual Observatory, etc.
- **We = Me**
  - My medical record, my Quicken data, digital photos of my daughter's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

USC SHOAH FOUNDATION INSTITUTE
FOR VISUAL HISTORY AND EDUCATION



Sarbanes-Oxley
Financial and Accounting Disclosure Information





Rensselaer
why not change the world? SM

Fran Berman

Business Regulation Requiring Data Preservation

Sarbanes-Oxley (Public Accounting Reform and Investor Protection Act of 2002)

- *Applies to all U.S. public company boards, management, and public accounting firms*
- **Includes electronic records** (correspondence, work papers, memoranda, etc.) that are created, sent, or received in connection with an audit or a review)
  - Section 103: "Board must require registered public accounting firms to "prepare, and maintain for a period of **not less than 7 years,** audit work papers, and other information related to any audit report, in sufficient detail to support the conclusions reached in that report."
  - Section 802: "any accountant who conducts an audit of an issuer of securities to which section 10(a) of the SEC …applies, shall maintain all audit or review work papers for a period of **5 years** from the end of the fiscal period in which the audit or review was concluded."

1. "Don't forget that **email and instant messaging are business records** …
4. Don't assume that the retention requirement …is …7 years. There are a lot of variables depending on the industry, type of organization and type of information. … **most lawyers that understand information retention agree that business records need to be kept indefinitely.**
10. Don't assume that just because you have access to archived information that you're going to be able to restore it within a reasonable amount of time…"

Kevin Beaver, "Thirteen Data Retention Mistakes to Avoid"
http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91_gci1186910,00.html

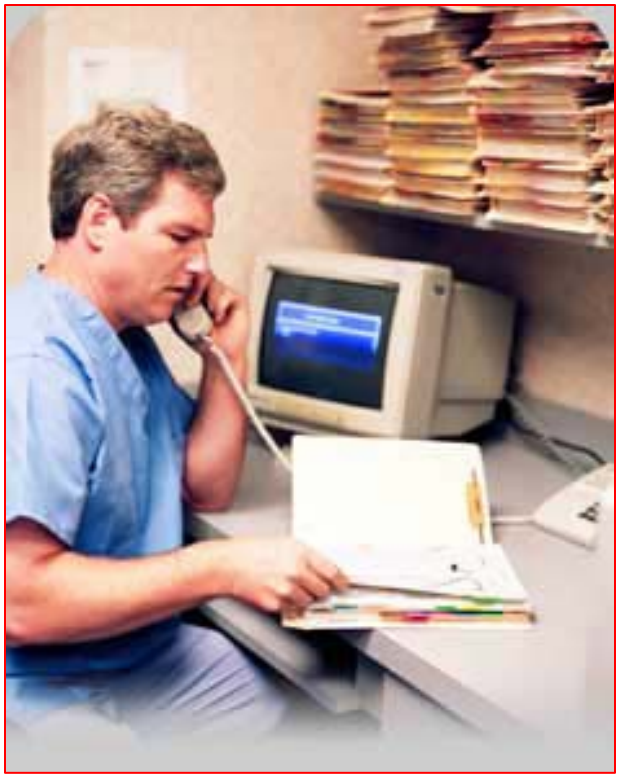
Rensselaer
why not change the world? SM

Fran Berman

Health Regulation Requiring Data Preservation

HIPAA (Health Insurance Portability and Accountability Act)

- *Applies to health information created or maintained by health care providers "who engage in certain **electronic transactions**, health plans, and health care clearinghouses" [www.hipaa.org]*
- Title II: Requires HHS to create rules and standards for the use and dissemination of health care information
- Healthcare providers must retain healthcare records for a period of **not less than 6 years.**



Rensselaer
why not change the world? SM

Increasing Policy and Regulation Affecting Research Community

- OMB requires that **federally funded research data,** supporting documentation, scientific notebooks, financial records, etc. be maintained **by the grantee for 3+ years**

- University libraries, federal agencies, institutional repositories ***not currently prepared*** to address the economic, technological, legal and social issues associated with widespread compliance of data retention policies

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

How Should We Save It?

Technology: Increasing activity around data storage and preservation technologies, programs, and services

- **Academic sector:**
  - ***Chronopolis*** *(*preservation grid*);* ***IRODS*** (rule-based distributed data management), ***Fedora*** (digital object repository system), ***D-Space*** (digital asset management), ***LOCKSS*** (peer-to-peer digital preservation infrastructure), etc.
- **Private sector:** *Amazon, MS, Google, Apple, Flickr, Sun, etc.*
- **Public Agencies/Institutions:** Library of Congress, NARA, NSF, NIH, DOE, Museums, Libraries, universities, state governments, etc.

However, there is no technology magic bullet …

Preserving digital data 100+ years will involve

- Tens-hundreds of new generations of technologies
- Thousands+ of new data standards and formats
- Millions+ of new valued collections
- Billions+ of potential users with as yet unknown information needs and workflows

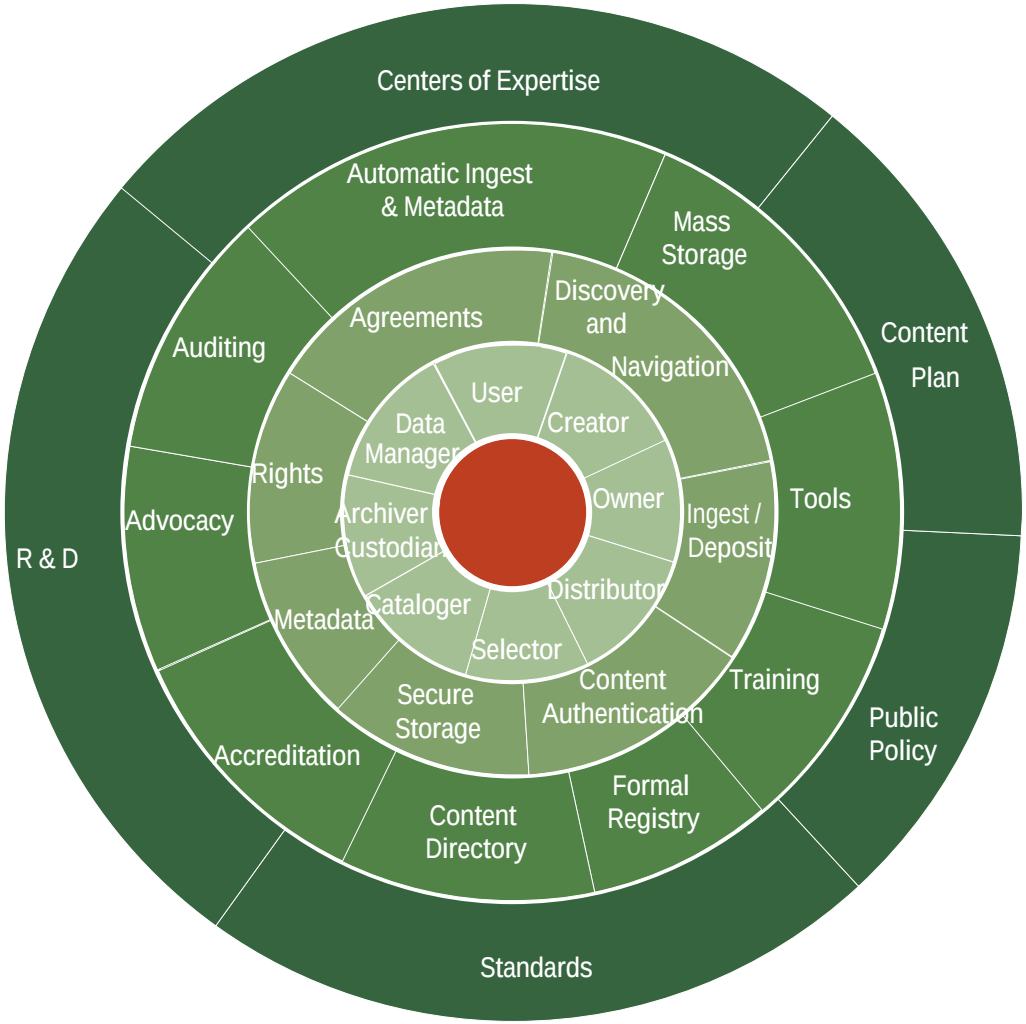
Rensselaer
why not change the world? SM

Fran Berman

Librarians' Perspective: People, Planning, Protections Critical Focus for the Preservation Environment

A Sample View of the
Library of Congress
Stewardship Network
for Humanities and Sciences

- NETWORK ENVIRONMENT
- NETWORK SERVICES
- ORGANIZATIONAL FUNCTIONS
- ROLES
- NETWORK MANAGEMENT



Rensselaer

Fran Berman

Figure courtesy of Martha Anderson, Library of Congress

Research and Education User's Perspective: Key Questions Focus on Outcomes rather than Technology

How do I make sure that my data will be there when I want it?

How can I combine my data with my colleague's data?

How should I organize my data?

How should I display my data?

What are the trends and what is the noise in my data?

My data is confidential; how do I make sure that it is seen/used only by the right people?

How can I make my data accessible to my collaborators?

Rensselaer

why not change the world? SM

Fran Berman

Current Best Practices in Digital Preservation

- **Replication** – make multiple copies and store some off-site
- **Heterogeneity** – more bio-diverse solutions tolerate greater error
- Associate **metadata** with data to aid access, management, search
- **Plan ahead** for smooth transition of data to new generations of media
- Align necessary level of **"trust"** with **reliability, infrastructure**
- Include **data costs** as part of the IT bill
- Pay attention to **security**
- Know the appropriate **regulations, policies, and penalties** that pertain to your data

Why are 3 copies used as best practice?

- Approach comes from Lamport, Shostak, and Pease's solution to the *Byzantine General's Problem*
  - Method for agreement on a battle plan for a group of Byzantine generals communicating only by messenger
  - Analogous to reliable computer systems with malfunctioning components
- *Solution:* When generals can send unforgeable signed messages to one another, the minimum number required for agreement is 3.

Who Should Pay?
The "Free Rider" Non-Solution

- Inadequate/unrealistic approach: **"Let X do it"** where **X** is:
  - The Government
  - The Libraries
  - The Archivists
  - Google, Yahoo, Microsoft, etc.
  - Data users
  - Data owners
  - Data creators, etc.

Creative partnerships needed *to provide preservation solutions for digital data in the public interest, overseen by trusted stewards, wtih*

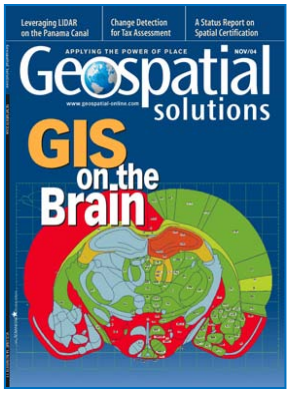
- *Feasible costs for users*
- *Sustainable costs for infrastructure*
- *Appropriate access*
- *Very low risk for data loss, etc.*

Multiple Economic Models Possible to Support Sustainable Digital Access and Preservation

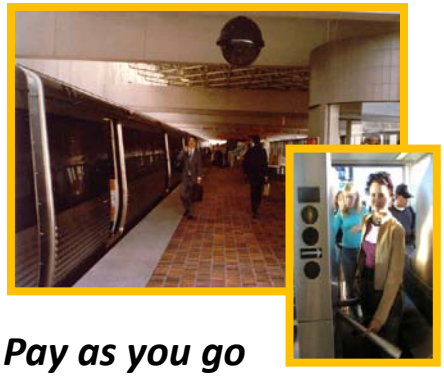
Key requirements for Sustainable Digital Preservation

- **Recognition** of the benefits of preservation *from decision makers*
- **Systemic incentives** to implement preservation efforts ("carrots and sticks")
- **Ongoing funding** for preservation resources
- **Appropriate organization and governance** of preservation activities.

Subscription



Advertisement



Pay as you go

Institutional subsidy

Rensselaer

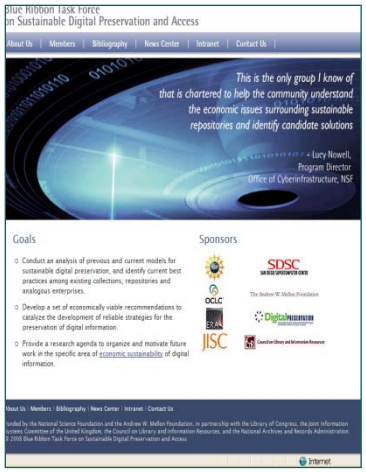
why not change the world? SM

Fran Berman

Requirements courtesy of Blue Ribbon Task Force on Sustainable Digital Preservation and Access

Setting the Stage for Cost-Effective Sustainability:
Blue Ribbon Task Force on Sustainable Digital Preservation and Access

Blue Ribbon Task Force on Sustainable Digital Preservation and Access x



Download of BRTF reports onTask Force website: ***brtf.sdsc.edu***

- Focus of investigations:
  - General **cost framework**: key cost categories of digital preservation
  - Set of **economic models** which provide alternative ways of addressing sustainable digital preservation
    - Pros, cons, costs, trade-offs of each
    - List real world conditions for which each model is best suited.
  - **Actionable recommendations**: "If your digital preservation context is X, you should consider using model Y for sustainable digital access and preservation."

Rensselaer
why not change the world? SM

Fran Berman

Creating a Strong Foundation for Future Success

1. Creating a strong foundation for science, engineering, and technology ***efforts*** in the Information Age

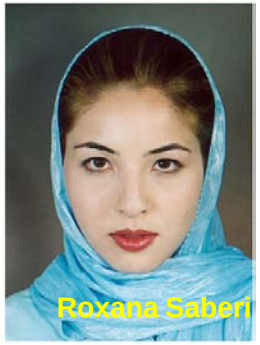
2. Creating a strong foundation for future science, engineering, and technology ***leadership***

Rensselaer
why not change the world? SM

Tomorrow's Leaders

Most of the tomorrow's leaders in science, technology, commerce, politics, art, etc. leaders of tomorrow's are students today

- **20 years ago or less …**
  - President **Barak Obama** graduated from Law School
  - Pulitzer Prize winner **Jhumpa Lahiri** graduated from College
  - Teach for America Founder **Wendy Kopp** was working on her Senior Thesis
  - Journalist **Roxana Saberi** was in Junior High School
  - Facebook Creator **Mark Zuckerberg** was in kindergarten

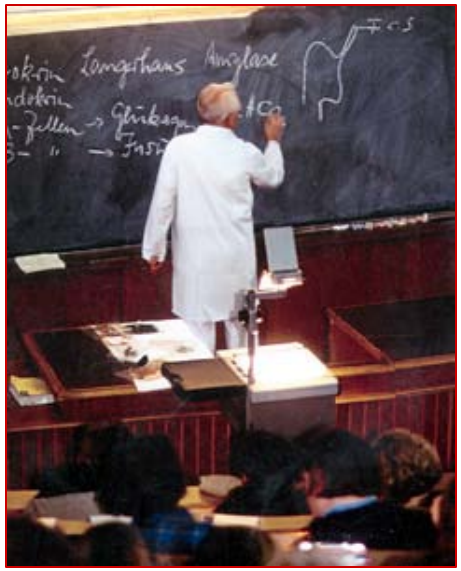


Rensselaer
why not change the world? SM

Fran Berman

Our Responsibility: Prepare today's students for a world of unprecedented complexity

- There's no "answer key" in real life
- Today's students need experience with
  - Challenging problems
  - Modern instruments and up-to-date technologies
  - Failure
  - International cultures
  - The "business", "political", "policy" and other attributes of real-world professional life



Our educational institutions must prepare students for the "outside" world they will encounter when they graduate

Rensselaer
why not change the world? SM

We Have the Power to Lay the Foundation for the Next Generation's Success

- **Power of asking the question**
  - "How many women and under-represented minorities PIs and co-Pis are associated with your Center?"
- **Power of creating explicit goals and metrics of success**
  - "I'd like to see more students doing research involving RPI's unique instruments and assets."
- **Power of recognition and encouragement**
  - Recognize success publicly, nominate our outstanding students and colleagues for awards, prizes, recognitions, etc.
- **Power of policy, resource allocation, and prioritization**
  - Use the resources under your control strategically and to help drive a more successful future

Rensselaer
why not change the world? SM

Fran Berman

We Can Lay the Foundation for Future Success

Rensselaer

why not change the world? SM

Fran Berman

Thank You

Rensselaer

why not change the world? SM

Fran Berman