Laying the Groundwork for Success in the Information Age

Dr. Fran Berman

Vice President for Research
Professor of Computer Science
Rensselaer Polytechnic Institute

Rensselaer
why not change the world?SM

Ken Kennedy – Pioneer, Colleague, Inspiration, Friend

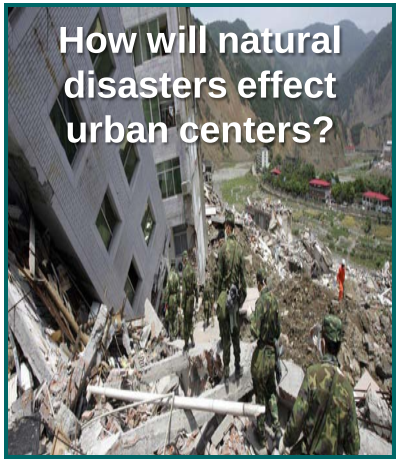
- Ken was a stellar example of leadership
  - Clear focus on, and prioritization of, what’s important
  - Effective, strategic, pragmatic, high-integrity, respectful of colleagues and collaborators at all levels
- Ken was focused on moving the community forward
  - Through contributions to computer science
  - Through the use of cyberinfrastructure to address major challenges in science and engineering
  - Through the next generation of scholars and leaders
  - Through service at the whole-discipline level

Creating a Successful Future:
Science and Engineering Drive Solutions to 21st Century Challenges

What is the potential impact of Global Warming?

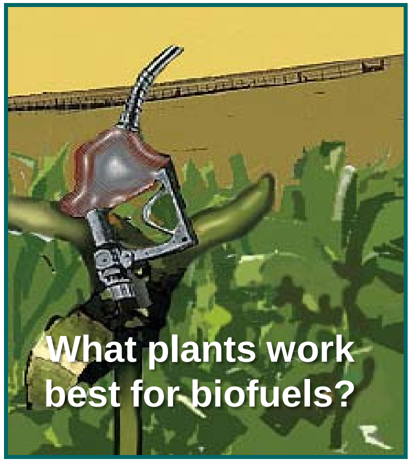
"Science is more essential for our prosperity, our security, our health, our environment, and our quality of life than it has ever been before."

President Barack Obama

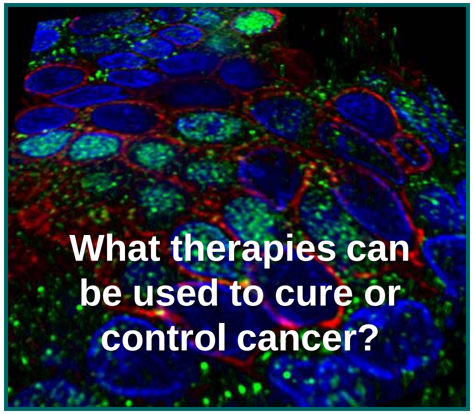


How will natural disasters effect urban centers?

Can we accurately predict market outcomes?



What plants work best for biofuels?



What therapies can be used to cure or control cancer?

Rensselaer

why not change the world? SM

Fran Berman

21st Century Challenges Require 21st Century Tools

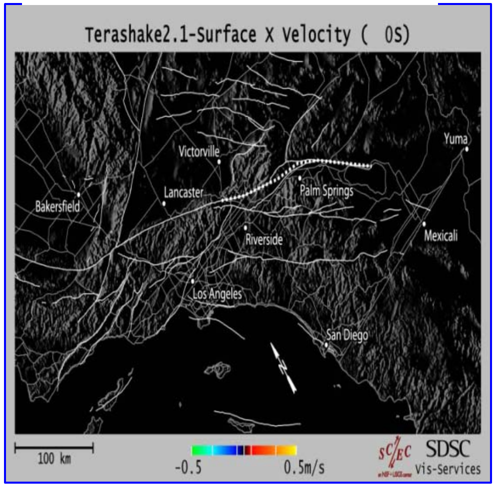
Cyberinfrastructure

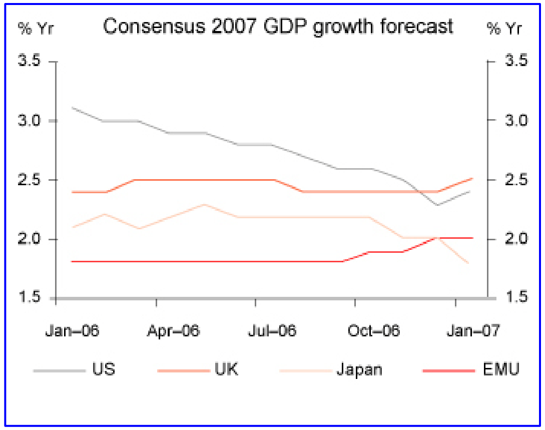
Sensors

“If infrastructure is required for an industrial economy, then we could say that **cyberinfrastructure is required for a knowledge economy.”**

The “Atkins Report”:
Revolutionizing Science and Engineering
Through Cyberinfrastructure, 2003

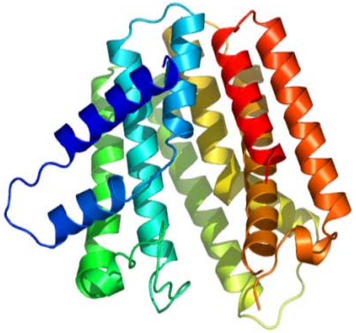
Visualization





Models

Computation



Data

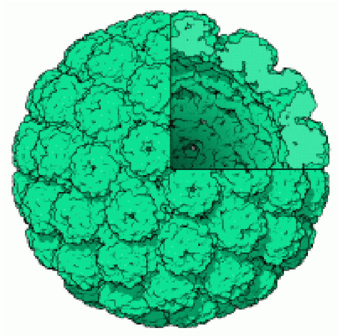
Rensselaer

why not change the world? SM

Fran Berman

Images and movies courtesy of Al Wallace/RPI, Amit Chourasia/SDSC, and JCSG

Data Cyberinfrastructure-Enabled Research

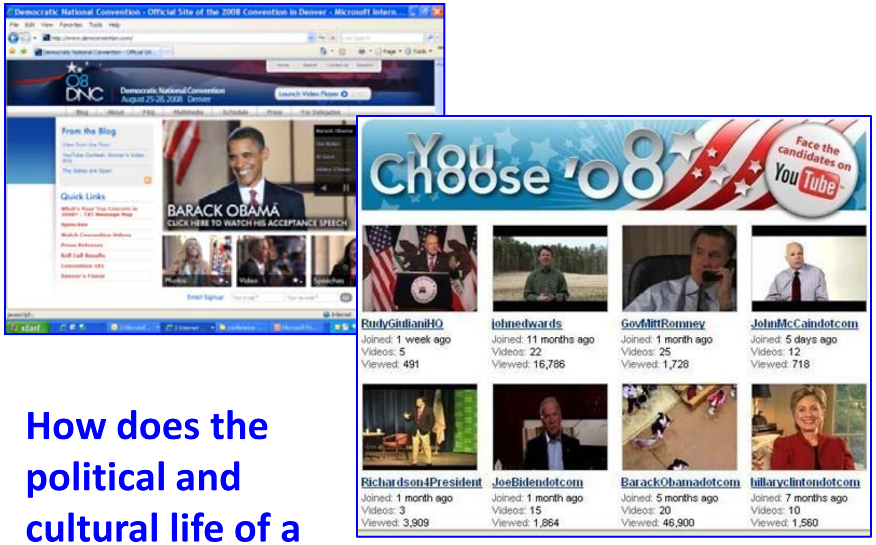


How does disease spread?

PDB: World wide reference collection of protein structure information

PROTEIN DATA BANK PDB

Rensselaer
why not change the world? SM


Democratic National Convention - Official Site of the 2008 Convention in Denver - Microsoft Intern...
08 DNC
Democratic National Convention
August 25-28, 2008 Denver
Launch Video Player
From the Blog
Quick Links
BARACK OBAMA
CLICK HERE TO WATCH HIS ACCEPTANCE SPEECH
Photos
Video
Speeches
Email Signup
start
YouChoose '08
Face the candidates on YouTube

RudyGiulianiHQ
Joined: 1 week ago
Videos: 5
Viewed: 491

johnedwards
Joined: 11 months ago
Videos: 22
Viewed: 16,786

GovMittRomney
Joined: 1 month ago
Videos: 25
Viewed: 1,728

JohnMcCaindotcom
Joined: 5 days ago
Videos: 12
Viewed: 718

Richardson4President
Joined: 1 month ago
Videos: 3
Viewed: 3,909

JoeBidendotcom
Joined: 1 month ago
Videos: 15
Viewed: 1,864

BarackObamadotcom
Joined: 5 months ago
Videos: 20
Viewed: 46,900

hillaryclintondotcom
Joined: 7 months ago
Videos: 10
Viewed: 1,560

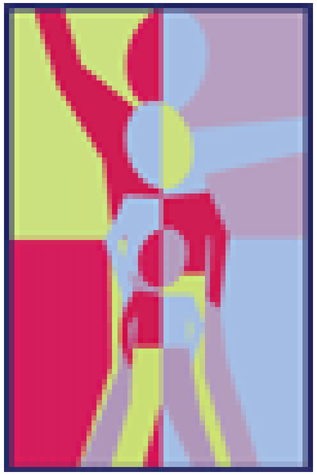
How does the political and cultural life of a society evolve?

The U.S. "cyber-election" of 2008

Life at the time of the Russian Revolution

Which has the greatest impact – nature or nurture?

Panel Study of Income Dynamics: longitudinal data on 8000 families over 40 years



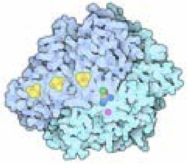
Fran Berman

Images and movies courtesy of Library of Congress, PDB, ICPSR

How Much Digital Information is There?

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

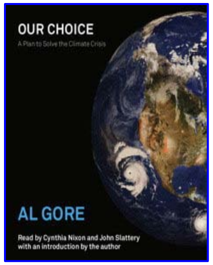
Google Earth =**71+ terabytes**



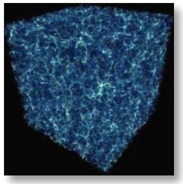
50,000 Protein Data Bank Structures = **35 terabytes**

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

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


OUR CHOICE
A Plan to Solve the Climate Crisis
AL GORE
Read by Cynthia Nixon and John Slattery with an introduction by the author

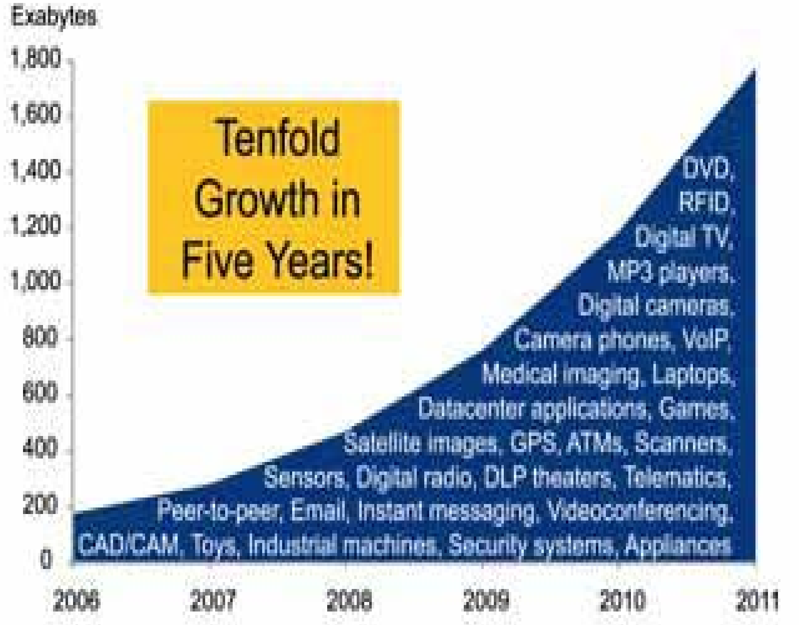
1 novel = **1 megabyte**



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

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*)*

Digital Information Created, Captured, Replicated Worldwide



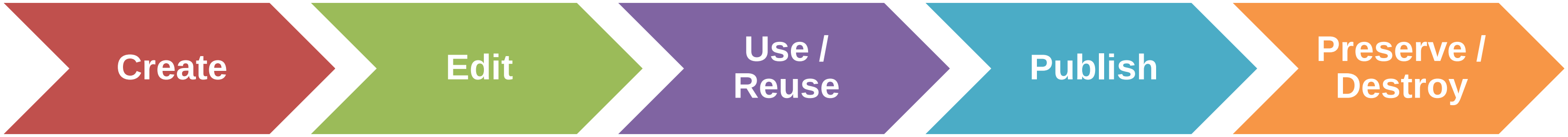
Rensselaer
why not change the world? SM

Fran Berman

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

Information from birth to death/immortality:

The Digital Data Life Cycle



Data creation / capture / gathering from

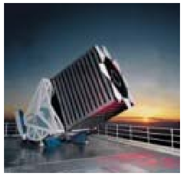
- laboratory experiments
- fieldwork
- surveys
- devices
- media
- simulation output ...

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

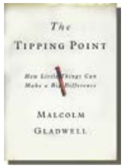
- Analyze
- Mine
- Model
- Derive additional data
- Visualize
- Input to instruments / computers / devices ....

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

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

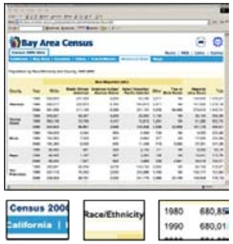


PDB


The Tipping Point

How Little Things Can Make a Big Difference

Malcolm Gladwell


Bay Area Census

Census 2000 California

Race/Ethnicity

1980 680,85

1990 680,01

Rensselaer

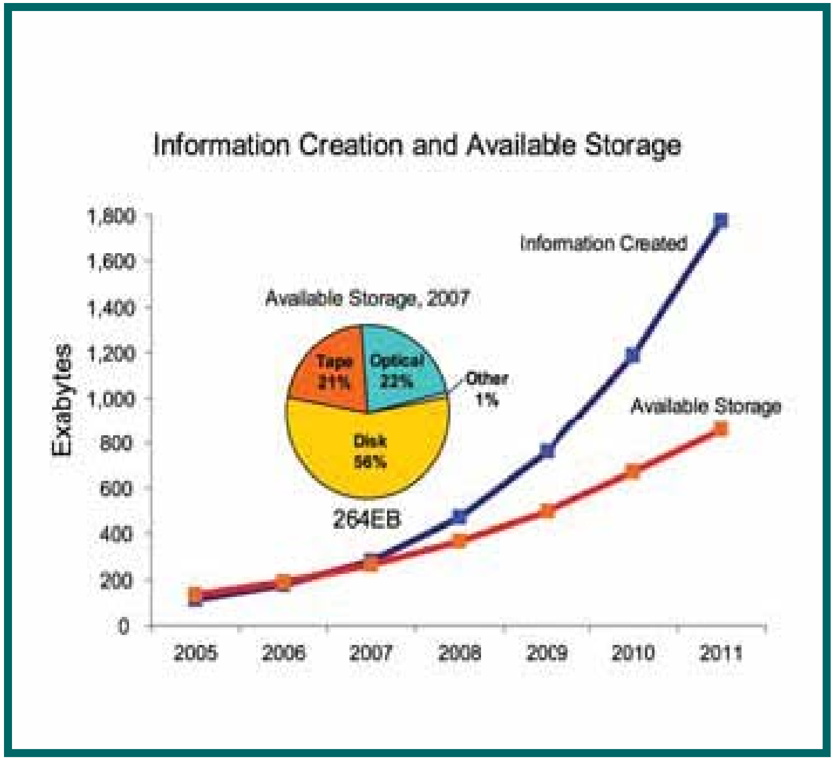
why not change the world?℠

Fran Berman

Information adapted from Chris Rusbridge and Liz Lyon

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 **proactive approach to information stewardship**



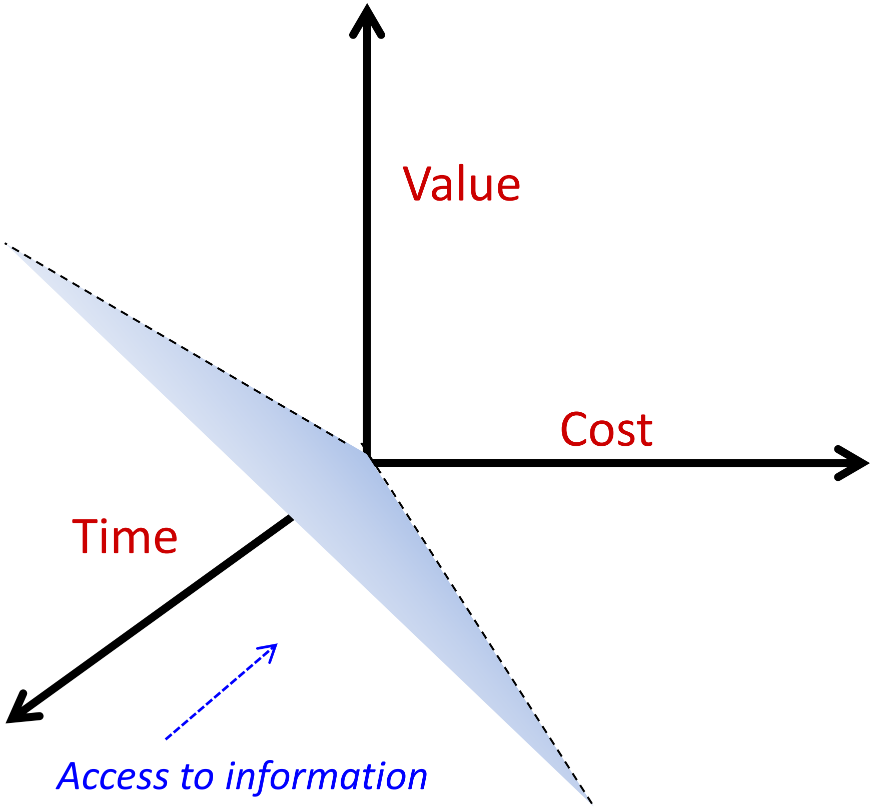
Rensselaer
why not change the world? SM

Fran Berman

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

Laying the groundwork for information stewardship: *value (to whom and how), regulation, economics*

- **Key Questions:**
  1) **What** should we save?
  2) **How** should we save it?
  3) **Who** should pay for it?



Access to information tomorrow requires preservation of information today

Rensselaer

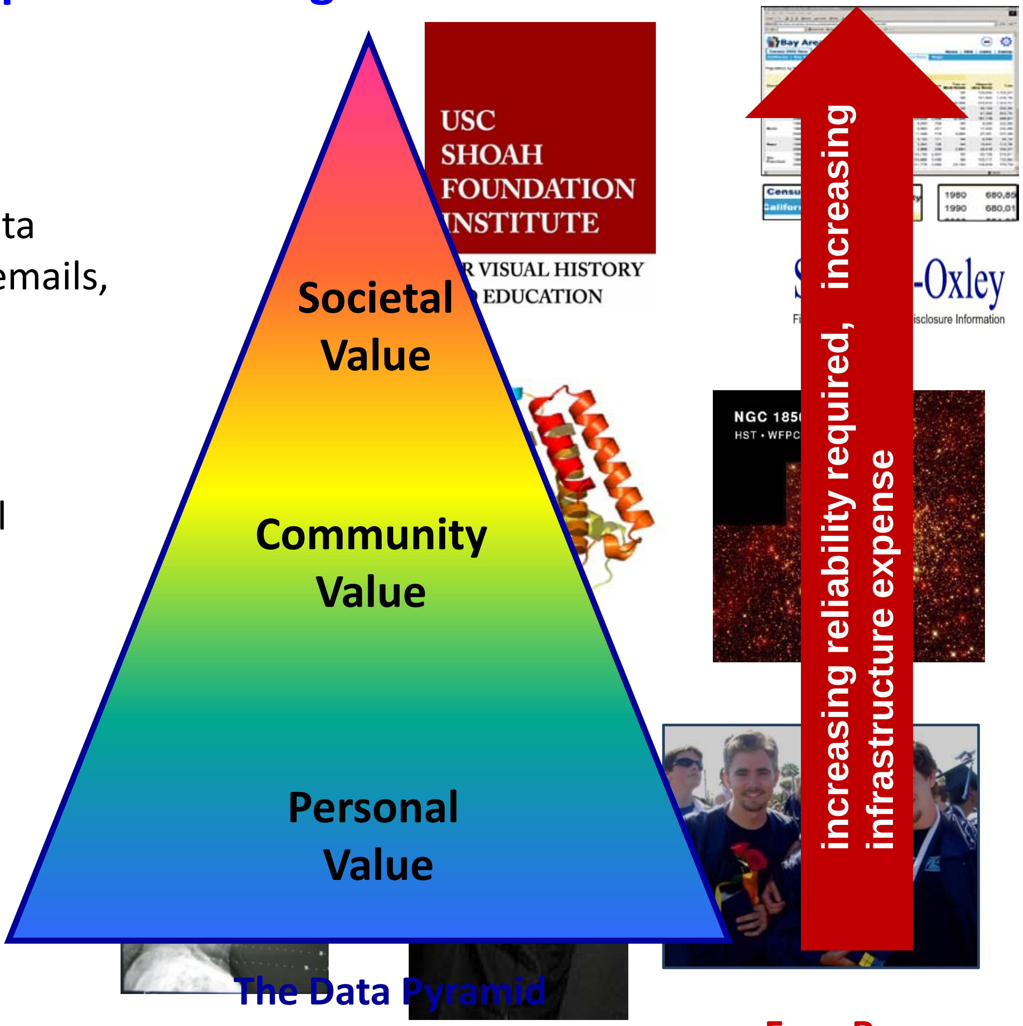
why not change the world?SM

Fran Berman

What Should We Save?

Digital information 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 kids’ graduations, etc.



The Data Pyramid

Rensselaer

why not change the world? SM

Fran Berman

Increasing Policy and Regulation Affecting Digital Information

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)

1. "Don't forget that **email and instant messaging are business records …**
4. Don't assume that the retention requirement …is …7 years. … **most lawyers that understand information retention agree that business records need to be kept indefinitely.**

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

Crime and Punishment

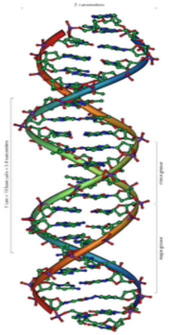
Regulations	Retention Requirement	Penalty
Sarbanes-Oxley	Auditors must retain relevant data for at least 7 years	Fines to $5M and 20 years in prison
HIPAA	Retain patient data for 6 years	$250K fine and up to 10 years in prison
Gramm-Leach-Baily	Ensure confidentiality of customer financial information	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?

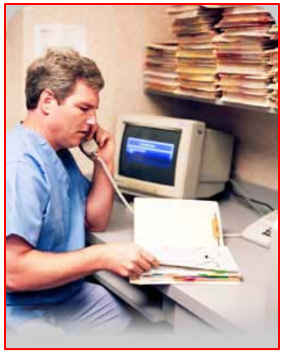
Rensselaer

why not change the world?SM

Fran Berman

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





Increasing Policy and Regulation Affecting Digital Information

Crime and Punishment

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.**

Regulations	Retention Requirement	Penalty
Sarbanes-Oxley	Auditors must retain relevant data for at least 7 years	Fines to $5M and 20 years in prison
HIPAA	Retain patient data for 6 years	$250K fine and up to 10 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?

…

Increasing Policy and Regulation Affecting Digital Information

- The U.S. Office of Management and Budget 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
Sarbanes-Oxley	Auditors must retain relevant data for at least 7 years	Fines to $5M and 20 years in prison
HIPAA	Retain patient data for 6 years	$250K fine and up to 10 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?

How Should We Save it?

Technology: Increasing activity around data storage and preservation technologies, programs, and services in both public, private, academic sectors

- DuraSpace, LOCKSS, Irods, Chronopolis, …
- Amazon, MS, Google, Apple, Flickr, Sun, etc.

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.

Rensselaer

why not change the world?℠

Fran Berman

“Good” Data Cyberinfrastructure …

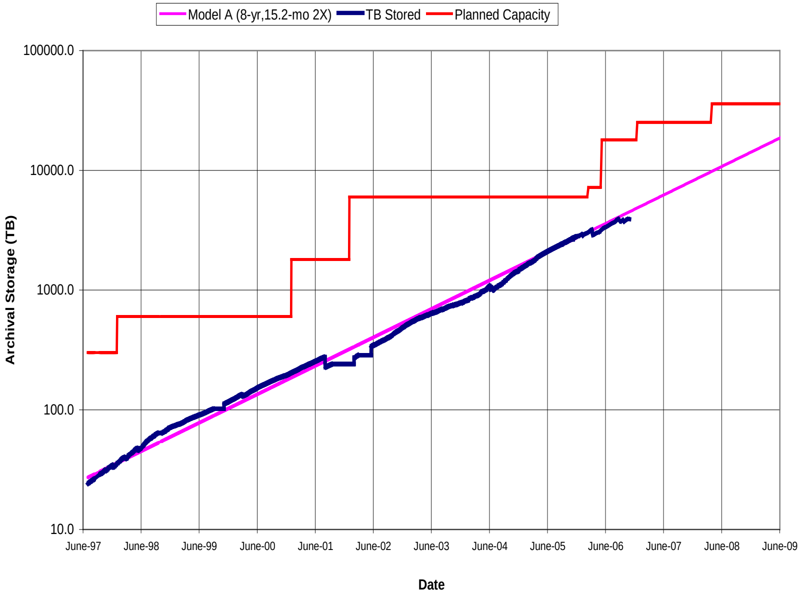
Incorporates the “ilities”:

- Scalability
- Interoperability
- Reliability
- Capability
- Sustainability
- Predictability
- Accessibility
- Responsibility
- Accountability
- …

Entity at risk	What can go wrong	Frequency
File	Corrupted media, disk failure	1 year
System	+ Systemic errors in vendor SW, or malicious user, or operator error that deletes multiple copies	15 years
Archive	+ Natural disaster, obsolescence of standards	50 - 100 years

Rensselaer

why not change the world? SM



Incurs real costs:

- Additional media for replication (disk, tape, geographically)
- Backup power systems
- Audit, reporting, access control systems
- Analysis, mining, other services
- Infrastructure maintenance
- Labor

Fran Berman

Information courtesy of Richard Moore, Reagan Moore

Support Models for Data Cyberinfrastructure Different from Supercomputing

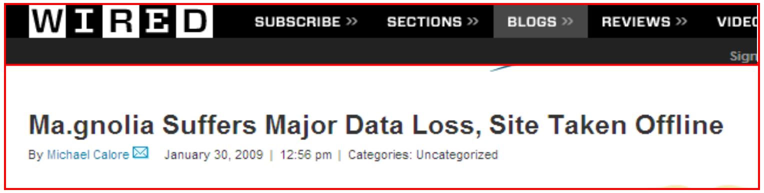
	Supercomputers	Archival Storage Systems
Metrics of Success	**High Performance;** good ranking on the Top500 list; application impact	**High reliability;** Minimal data loss and damage
Next Generation Systems	**Growth in capability/capacity** key: Compatibility of systems not required although there should be application transition paths	**Smooth migration for data** key: Preservation collections must migrate to new media without loss of data or disruption to users
Funding Model	Serial **"one time" funding** for each new HPC resource possible	**No gaps.** Funding must be available for continuous support of data collections

Digits

Technology News and Insights

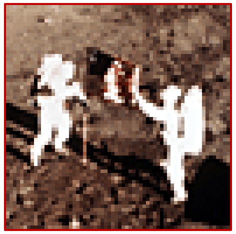
NOVEMBER 13, 2009, 5:48 PM ET

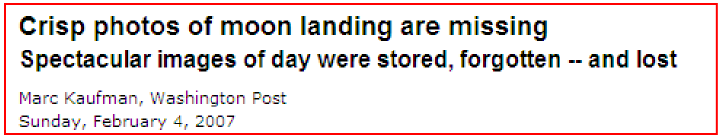
Faster Supercomputers: Your Tax Dollars at Work



Ma.gnolia Suffers Major Data Loss, Site Taken Offline

By Michael Calore | January 30, 2009 | 12:56 pm | Categories: Uncategorized





Crisp photos of moon landing are missing

Spectacular images of day were stored, forgotten -- and lost

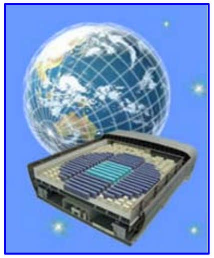
Marc Kaufman, Washington Post

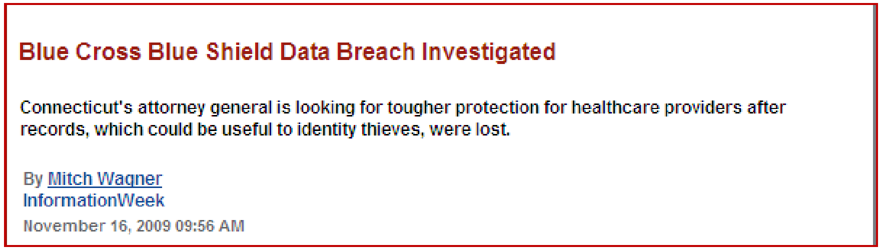
Sunday, February 4, 2007

Supercomputer to boost research output

Posted Mon Nov 16, 2009 12:36pm AEDT
Updated Mon Nov 16, 2009 3:16pm AEDT

The most powerful supercomputer in the country is now online at the Australian National University (ANU) in Canberra.





Blue Cross Blue Shield Data Breach Investigated

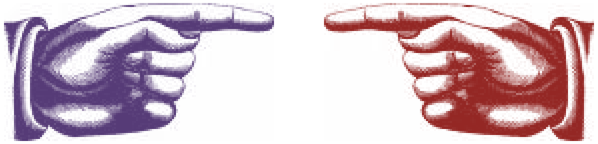
Connecticut's attorney general is looking for tougher protection for healthcare providers after records, which could be useful to identity thieves, were lost.

By Mitch Wagner
InformationWeek
November 16, 2009 09:56 AM

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, Microsoft, etc.
  - Data users
  - Data owners
  - Data creators, etc.



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

- Feasible costs for providers and users
- Very low risk for data loss
- Adequate access controls and management structure, etc.

Rensselaer

why not change the world?SM

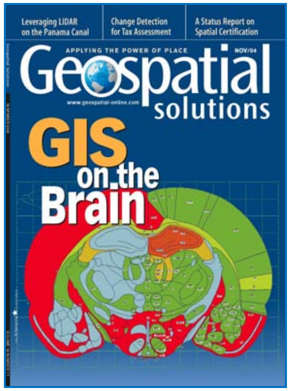
Fran Berman

Multiple Economic Models Possible for Digital Preservation and Access

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


Leveraging LIDAR on the Panama Canal
Change Detection for Tax Assessment
A Status Report on Spatial Certification
APPLYING THE POWER OF PLACE
NOV/04
Geospatial solutions
www.geospatial-online.com
GIS on the Brain

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 (brtf.sdsc.edu)

Setting the Stage for Cost-Effective Sustainability:
Blue Ribbon Task Force to Provide Actionable Recommendations for Digital Preservation and Access

Blue Ribbon Task Force on Sustainable Digital Preservation and Access Final Report (out in Jan 2010)

- **Key digital preservation scenarios:**
  - **Research data**
  - **Scholarly discourse and publications**
  - **Blogs/Collectively-created content**
  - **Music/Movies/Commercially-owned cultural content**
- Set of **economic models** that provide alternative ways of addressing sustainable digital preservation

*(First year) BRTF **Interim Report** available at Task Force website:* ***brtf.sdsc.edu***

- **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

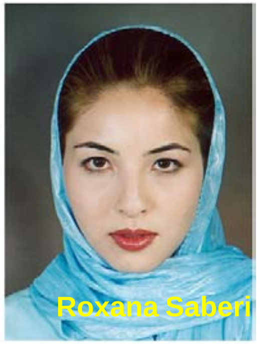
Ken Kennedy – Pioneer, Colleague, Inspiration, Friend

- Ken was a stellar example of leadership
  - Clear focus on, and prioritization of, what's important
  - Effective, strategic, pragmatic, high-integrity, respectful of colleagues and collaborators at all levels
- Ken was focused on moving the community forward
  - Through contributions to computer science
  - Through the application of computer science to address major challenges in science and engineering
  - Through the next generation of scholars and leaders
  - Through service at the whole-discipline level

Tomorrow's Leaders

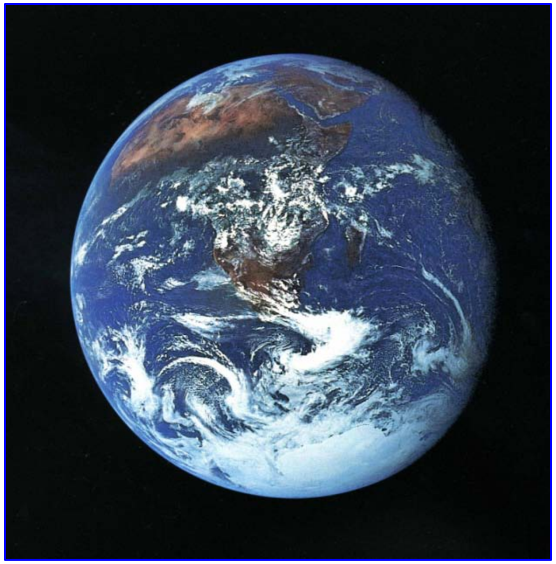
Most of the tomorrow's leaders in science, technology, commerce, politics, art, etc. leaders of tomorrow's are students today

- 20 years ago or less …
  - President **Barack 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



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", "rights" and other attributes of real-world professional life



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

Rensselaer

why not change the world? SM

Fran Berman

Call to Action to the Computer Science Community: *We have the Power to Lay the Groundwork for Future Success*

- **Power of asking the question**
  - "How many women and under-represented minorities PIs and co-PIs are associated with your Department/School/Institution?"
- **Power of creating explicit goals and metrics of success**
  - "We will devote more than 3 percent of our GDP to research and development."
- **Power of recognition and encouragement**
  - Public recognition of our success, nomination our outstanding students and colleagues for awards, prizes, recognitions, prestigious memberships, etc.
- **Power of policy, resource allocation, and prioritization**
  - We can use the resources under our control strategically, and to help drive a more successful future

Thank You

- Special Thanks to the Ken Kennedy Award Committee, ACM, IEEE, Jan Cuny, my family, and the extraordinary colleagues and students I’ve come to know through the GrADS and VGrADS projects that we shared with Ken.

GrADS

Grid Application Development Software Project

VGrADS

Virtual Grid Application Development Software Project