100 Years of Digital Data

Georgia Institute of Technology

Dr. Francine Berman
Director, San Diego Supercomputer Center
Professor and High Performance Computing Endowed Chair, UC San Diego
Digital Data Drives the Information Age
How much Digital Data is there?

- 5 exabytes of digital information produced in 2003
- 161 exabytes of digital information produced in 2006
  - 25% of the 2006 digital universe is born digital (digital pictures, keystrokes, phone calls, etc.)
  - 75% is replicated (emails forwarded, backed up transaction records, movies in DVD format)
- 1 zettabyte aggregate digital information projected for 2010

<table>
<thead>
<tr>
<th>Scale</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kilo</td>
<td>$10^3$</td>
</tr>
<tr>
<td>Mega</td>
<td>$10^6$</td>
</tr>
<tr>
<td>Giga</td>
<td>$10^9$</td>
</tr>
<tr>
<td>Tera</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>Peta</td>
<td>$10^{15}$</td>
</tr>
<tr>
<td>Exa</td>
<td>$10^{18}$</td>
</tr>
<tr>
<td>Zetta</td>
<td>$10^{21}$</td>
</tr>
</tbody>
</table>

SDSC HPSS tape archive = 25+ PetaBytes

iPod (up to 20K songs) = 80 GB

1 novel = 1 MegaByte

U.S. Library of Congress manages 295 TB of digital data, 230 TB of which is “born digital”

Data Drives Research and Education

Data at multiple scales in the Biosciences

Data from multiple sources in the Geosciences

Where should we drill for oil?
What is the Impact of Global Warming?
How are the continents shifting?

Users

Data Access and Use

Data Integration

Disciplinary Databases

Anatomy

Physiology

Cell Biology

Proteomics

Genomics

Medicinal Chemistry

What genes are associated with cancer?
What parts of the brain are responsible for Alzheimers?

Data Integration

Complex “multiple-worlds” mediation

Geologic Map
Geo-Chemical
Geo-Physical
Geo-Chronologic
Foliation Map

Images courtesy of Mark Miller and Chaitan Baru
Data a Fundamental Component of Cyberinfrastructure

- Cyberinfrastructure is the organized aggregate of technologies enabling access and coordination of information technology resources to facilitate science, engineering, and societal goals.
  - Data
  - Computation
  - Communication
  - Visualization
  - Scientific Instruments
  - Expertise, etc.

Published in 2003, NSF Blue Ribbon Panel (Atkins) Report provided a compelling and comprehensive vision of an integrated Cyberinfrastructure
Today’s Presentation

• Revolutionizing astronomy through digital data

• Data Cyberinfrastructure Today – Designing and developing infrastructure to enable today’s data-oriented applications

• Challenges in Building and Delivering Data Cyberinfrastructure
Astronomy in the Last Century

1908  First reported detection of a magnetic field in any astronomical object -- a sunspot. Discovered by George Hale using telescope at Mount Wilson Solar Observatory.

1916  Einstein introduces General Theory of Relativity

1916  Hubble shows that galaxies exist outside the Milky Way Galaxy

1930  Discovery of Pluto by Clyde Tombaugh (following work begun by Percival Lowell). Pluto’s existence hypothesized based on anomalies in the orbits of Neptune and Uranus.

1957  Sputnik (first human-made satellite) launched, marking beginning of the “Space Age”

1958  NASA created (from former National Advisory Committee for Aeronautics) and other US Govt organizations

1965  Penzias and Wilson discover cosmic fossil radiation, providing direct evidence of the Big Bang Theory

1990  Hubble Space Telescope put into orbit

Etc., etc., etc.
“The Universe is now being explored systematically, in a panchromatic way, over a range of spatial and temporal scales that lead to a more complete, and less biased understanding of its constituents, their evolution, their origins, and the physical processes governing them.”

Towards a National Virtual Observatory
The Virtual Observatory

- Premise: most observatory data is (or could be) online
- So, the Internet is the world’s best telescope:
  - It has data on every part of the sky
  - In every measured spectral band: optical, x-ray, radio..
  - It’s as deep as the best instruments
  - It is up when you are up
  - The “seeing” is always great
  - It’s a smart telescope:
    - links objects and data to literature on them
- Software has became a major expense
  - Share, standardize, reuse..
The National Virtual Observatory

- NVO combines data from sky surveys and over 50 ground and space-based telescopes and instruments to create a comprehensive picture of the heavens

- HOW NVO Works
  - Raw data comes from large-scale synoptic telescopes. Scientists “clean” data, convert data from temporal to spatial, indexing over both dimensions
  - NVO data available to the public without restriction after 1 year by community agreement
  - NVO databases distributed and mirrored at multiple sites
The Sloan Digital Sky Survey provides a 3D map of a million galaxies and quasars covering more than a quarter of the sky.

The 2 Micron All Sky Survey (2MASS) provides direct answers to questions on the large-scale structure of the Milky Way and the Local Universe.

The Palomar Oschin telescope provides a catalogue of the entire northern sky in blue, red and near-infrared colors.

SDSC’s NVO collection is nearly 100 TB and has grown over 5-fold since 2002.

The USNO-B all-sky catalogue was obtained from various sky surveys during the last 50 years. USNO provides all-sky coverage and 85% accuracy for distinguishing stars from non-stellar objects.

2MASS gathers data from a northern facility in Arizona and a southern facility in Chile.

Photometric data from Mt. Stromlo observatory in Australia on several million stars gathered since 1992 to explore constitution of dark matter in the halo of the Milky Way.
Data-Driven Astronomy

- **Looking for**
  - *Needles in haystacks* – the Higgs particle
  - *Haystacks* -- Dark matter, Dark energy

- **Statistical analysis often deals with**
  - Creating uniform samples
  - Data filtering
  - Assembling relevant subsets
  - Censoring bad data
  - “Likelihood” calculations
  - Hypothesis testing, etc.

- **Traditionally these are performed on files, most of these tasks are much better done inside a database**
Making Discoveries Using the NVO

Scientists at Johns Hopkins, Caltech and other institutions confirmed the discovery of a new brown dwarf. Search time on 5,000,000 files went from months to minutes using NVO database tools and technologies.

Brown dwarfs are often called the “missing link” in the study of star formations. They are considered small, cool “failed stars”.
Evolving the Universe from the “Big Bang”

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

*Slide modified from Mike Norman*
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 = 65,536³**
ENZO at Petascale

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

- Technical 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

Verifying Theory with Observation

- James Webb Space Telescope (JWST), coming in 2013 will probe the first billion years of the universe – providing observations of unprecedented depth and breadth

- Data will enable tight integration of observation and theory, and will enable simulations to approach realistic complexity

- Petascale computing and scientific data management essential for achieving new results
Building and Delivering infrastructure for Data-oriented Applications
Today’s Data-oriented Applications Span the Spectrum

Designing Infrastructure for Data:

Data and High Performance Computing

Data and Cyberinfrastructure Services

Support for management and preservation of data of community value

Fran Berman
Data and High Performance Computing

• For many applications, development of “balanced systems” needed to support applications which are both data-intensive and compute-intensive. Codes for which
  • Grid platforms not a strong option
  • Data must be local to computation
  • I/O rates exceed WAN capabilities
  • Continuous and frequent I/O is latency intolerant

• Scalability is key
  • Need high-bandwidth and large-capacity local parallel file systems, archival storage
Data and HPC: What you see is what you’ve measured

FLOPS alone are not enough.

Appropriate benchmarks needed to rank/bring visibility to more balanced machines critical for today’s applications.

Three systems using the same processor and number of processors.

- AMD Opteron 64 processors 2.2 GHz
- Difference is in way the processors are interconnected

HPC Challenge benchmarks measure different machine characteristics

- Linpack and matrix multiply are computationally intensive
- PTRANS (matrix transpose), RandomAccess, bandwidth/latency tests and other tests begin to reflect stress on memory system

Information courtesy of Jack Dongarra
# An Integrated Resource Environment Needed to Support Data-Oriented Applications

## SDSC High Performance Computing Systems

- **DataStar**
  - 15.6 TFLOPS Power 4+ system
  - 7.125 TB total memory
  - Up to 4 GBps I/O to disk
  - 115 TB GPFS filesystem

- **Blue Gene Data**
  - First academic IBM Blue Gene system
  - 17.1 TF
  - 1.5 TB total memory
  - 3 racks, each with 2,048 PowerPC processors and 128 I/O nodes

- **TeraGrid Cluster**
  - 524 Itanium2 IA-64 processors
  - 2 TB total memory
  - Also 16 2-way data I/O nodes

  [http://www.sdsc.edu/user_services/](http://www.sdsc.edu/user_services/)

## SDSC Data Collections, Archival and Storage Systems

- 2.4 PB Storage-area Network (SAN)
- 25 PB StorageTek/IBM tape library
- HPSS and SAM-QFS archival systems
- DB2, Oracle, MySQL
- Storage Resource Broker
- Supporting servers: IBM 32-way p690s, 72-CPU SunFire 15K, etc.

  [Support for community data collections and databases](http://www.sdsc.edu/)

  [Data management, mining, analysis, and preservation](http://www.sdsc.edu/)

## SDSC Science and Technology Staff, Software, Services

- Data-oriented Community SW, toolkits, portals, codes
- DataCentral national hosting repository
- Chronopolis services (w/ UCSDL)
- Data User Services
- Application/Community Collaborations
- Education and Training

  [http://www.sdsc.edu/](http://www.sdsc.edu/)
Data Services – What do Users Want?

- How do I make sure that my data will be there when I want it?
- How should I display my data?
- How can I combine my data with my colleague’s data?
- How should I organize my data?
- What are the trends and what is the noise in my data?
- How can I make my data accessible to my collaborators?
- My data is confidential; how do I make sure that it is seen/used only by the right people?
- How do I make sure that my data will be there when I want it?
Services: Integrated Environment Is Key

- File systems, Database systems, Collection Management, Data Integration, etc.
- Many Data Sources
- Data Access
- Data Use
- Data Management
- Data Storage
- Integrated Infrastructure
- Many Data Sources
- Modeling
- Analysis
- Simulation
- Visualization
- File systems, Database systems, Collection Management, Data Integration, etc.
- Many Data Sources
- Data Access
- Data Use
- Data Management
- Data Storage
- Integrated Infrastructure
- Many Data Sources
- Modeling
- Analysis
- Simulation
- Visualization
- Database selection and schema design
- 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 visualization
- Data anonymization, etc.
SDSC DataCentral: National Data Hosting Facilities

- Broad program to support research and community data collections and databases

- DataCentral services include:
  - Public Data Collections and Database Hosting
  - Long-term storage and preservation (tape and disk)
  - Remote data management and access (SRB, portals)
  - Data Analysis, Visualization and Data Mining
  - Professional, qualified 24/7 support

- DataCentral resources include:
  - 1 PB On-line disk
  - 25 PB StorageTek tape library capacity
  - 540 TB Storage-area Network (SAN)
  - DB2, Oracle, MySQL
  - Storage Resource Broker
  - Gpfs-WAN with 700 TB
## DataCentral Allocated Collections include

<table>
<thead>
<tr>
<th>Seismology</th>
<th>3D Ground Motion Collection for the LA Basin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atmospheric Sciences</td>
<td>Sciences50 year Downscaling of Global Analysis over California Region</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>NEXRAD Data in Hydrometeorology and Hydrology</td>
</tr>
<tr>
<td>Elementary Particle Physics</td>
<td>AMANDA data</td>
</tr>
<tr>
<td>Biology</td>
<td>AfCS Molecule Pages</td>
</tr>
<tr>
<td>Biomedical Neuroscience</td>
<td>BIRN</td>
</tr>
<tr>
<td>Networking</td>
<td>Backbone Header Traces</td>
</tr>
<tr>
<td>Networking</td>
<td>Backscatter Data</td>
</tr>
<tr>
<td>Biology</td>
<td>Bee Behavior</td>
</tr>
<tr>
<td>Biology</td>
<td>Biocyc (SRI)</td>
</tr>
<tr>
<td>Art</td>
<td>CS landscape Database</td>
</tr>
<tr>
<td>Geology</td>
<td>Chronos</td>
</tr>
<tr>
<td>Biology</td>
<td>CKAAPS</td>
</tr>
<tr>
<td>Biology</td>
<td>DigEmbryo</td>
</tr>
<tr>
<td>Earth Science Education</td>
<td>ERESE</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>UCI ESMF</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>EarthRef.org</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>ERDA</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>ERR</td>
</tr>
<tr>
<td>Biology</td>
<td>Encyclopedia of Life</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>Geosciences</td>
<td>GEON</td>
</tr>
<tr>
<td>Geosciences</td>
<td>GEON-LIDAR</td>
</tr>
<tr>
<td>Geochemistry</td>
<td>Kd</td>
</tr>
<tr>
<td>Biology</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>Geochemistry</td>
<td>GERM</td>
</tr>
<tr>
<td>Networking</td>
<td>HPWREN</td>
</tr>
<tr>
<td>Ecology</td>
<td>HyperLter</td>
</tr>
<tr>
<td>Networking</td>
<td>IMDC</td>
</tr>
<tr>
<td>Biology</td>
<td>Interpro Mirror</td>
</tr>
<tr>
<td>Biology</td>
<td>JCSG Data</td>
</tr>
<tr>
<td>Government</td>
<td>Library of Congress Data</td>
</tr>
<tr>
<td>Geophysics</td>
<td>Magnetics Information Consortium data</td>
</tr>
<tr>
<td>Education</td>
<td>UC Merced Japanese Art Collections</td>
</tr>
<tr>
<td>Geochemistry</td>
<td>NAVDAT</td>
</tr>
<tr>
<td>Earthquake Engineering</td>
<td>NEESIT data</td>
</tr>
<tr>
<td>Education</td>
<td>NSDL</td>
</tr>
<tr>
<td>Astronomy</td>
<td>NVO</td>
</tr>
<tr>
<td>Government</td>
<td>NARA</td>
</tr>
<tr>
<td>Anthropology</td>
<td>GAPP</td>
</tr>
<tr>
<td>Neurobiology</td>
<td>Salk data</td>
</tr>
<tr>
<td>Seismology</td>
<td>SCEC TeraShake</td>
</tr>
<tr>
<td>Seismology</td>
<td>SCEC CyberShake</td>
</tr>
<tr>
<td>Oceanography</td>
<td>SIO Explorer</td>
</tr>
<tr>
<td>Networking</td>
<td>Skitter</td>
</tr>
<tr>
<td>Astronomy</td>
<td>Sloan Digital Sky Survey</td>
</tr>
<tr>
<td>Geology</td>
<td>Sensitive Species Map Server</td>
</tr>
<tr>
<td>Biology</td>
<td>SD and Tijuana Watershed data</td>
</tr>
<tr>
<td>Oceanography</td>
<td>Seamount Catalogue</td>
</tr>
<tr>
<td>Oceanography</td>
<td>Seamounts Online</td>
</tr>
<tr>
<td>Biodiversity</td>
<td>WhyWhere</td>
</tr>
<tr>
<td>Ocean Sciences</td>
<td>Southeastern Coastal Ocean Observing and Prediction Data</td>
</tr>
<tr>
<td>Structural Engineering</td>
<td>TeraBridge</td>
</tr>
<tr>
<td>Various</td>
<td>TeraGrid data collections</td>
</tr>
<tr>
<td>Biology</td>
<td>Transporter Classification Database</td>
</tr>
<tr>
<td>Biology</td>
<td>TreeBase</td>
</tr>
<tr>
<td>Art</td>
<td>Tsunami Data</td>
</tr>
<tr>
<td>Education</td>
<td>ArtStor</td>
</tr>
<tr>
<td>Biology</td>
<td>Yeast regulatory network</td>
</tr>
<tr>
<td>Biology</td>
<td>Apoptosis Database</td>
</tr>
<tr>
<td>Cosmology</td>
<td>LUSciD</td>
</tr>
</tbody>
</table>
Data Visualization

SCEC Earthquake simulations

Visualization of Cancer Tumors

Prokudin–Gorskii historical images

Information and images courtesy of Amit Chourasia, SCEC, Steve Cutchin, Moores Cancer Center, David Minor, U.S. Library of Congress
Building Successful Cyberinfrastructure
Infrastructure Should be Non-memorable

• Good infrastructure should be
• Predictable
• Pervasive
• Cost-effective
• Easy-to-use
• Reliable
• Unsurprising

What's required to build and provide useful, usable, and capable data Cyberinfrastructure?
“Good” Data Cyberinfrastructure incorporates the “ilities”

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

<table>
<thead>
<tr>
<th>Entity at risk</th>
<th>What can go wrong</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>File</td>
<td>Corrupted media, disk failure</td>
<td>1 year</td>
</tr>
<tr>
<td>Tape</td>
<td>+ Simultaneous failure of 2 copies</td>
<td>5 years</td>
</tr>
<tr>
<td>System</td>
<td>+ Systemic errors in vendor SW, or malicious user, or operator error that deletes multiple copies</td>
<td>15 years</td>
</tr>
<tr>
<td>Archive</td>
<td>+ Natural disaster, obsolescence of standards</td>
<td>50 - 100 years</td>
</tr>
</tbody>
</table>

Data Reliability: What can go wrong

Predictable performance critical for user planning and optimization

Information courtesy of Reagan Moore, Jennifer Schopf
Good Data Infrastructure Incurs Real Costs

Capacity Costs

- Most valuable data must be replicated
- SDSC research collections have been doubling every 15 months.
- **SDSC storage** is 25 PB and counting. Data is from supercomputer simulations, digital library collections, etc.

Capability Costs

- **Reliability** increased by up-to-date and robust hardware and software for
  - Replication (disk, tape, geographically)
  - Backups, updates, syncing
  - Audit trails
  - Verification through checksums, physical media, network transfers, copies, etc.
- **Data professionals needed** to facilitate
  - Infrastructure maintenance
  - Long-term planning
  - Restoration, and recovery
  - Access, analysis, preservation, and other services
  - Reporting, documentation, etc.

Information courtesy of Richard Moore
**Economic Sustainability**

- **Data Preservation** is the Grand Challenge for economically sustainable Cyberinfrastructure

- Good preservation requires **continuous support**

- **Key questions:**
  - What should we save?
  - Who should save it?
  - How should we save it?
  - Who should have access to it?
  - Who should pay for it?
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 Mom’s 80th birthday, etc.
Who Pays? The “Free Rider” Non-Solution

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

- Creative partnerships needed to provide preservation solutions with
  - Trusted stewards
  - Feasible costs for users
  - Sustainable costs for infrastructure
  - Very low risk for data loss, etc.
A Framework for Digital Stewardship and Preservation

Digital Data Collections
- Reference, nationally/internationally important, irreplaceable data collections
- Key research and community data collections
- Personal data collections

Repositories / Facilities
- National / internationally scale repositories, libraries, archives
- "Regional" scale libraries and targeted data archives / centers
- Private repositories

The Data Pyramid

Increasing Value
Increasing Trust
Increasing responsibility, increasing risk
Increasing stability
Increasing infrastructure

Fran Berman
Who Pays? Multiple Solutions

The Data Pyramid

National, International Scale

“Regional” Scale

Local Scale

Who Pays? Multiple Solutions

SAN DIEGO SUPERCOMPUTER CENTER

Fran Berman
The Data Problem is About to Get Worse

- **2007 is the “crossover year”** where the amount of digital information is greater than the amount of available storage

- **Increasing requirements for data retention** in private, public, academic sectors render data infrastructure and preservation policies more critical
  - Agency requirements for data management and preservation, Sarbaanes-Oxley, etc.

- **Data Center systems / functionalities needed include**
  - Data Management & Movement
  - Metadata Catalog
  - Client Interfaces
  - Administration and policies
  - Authentication
  - Trust management, etc.

100 Years of Digital Data

- Digital data is the natural resource of the Information Age
  - Fragile—dependent on rapidly changing technologies
  - Valuable data lost often cannot be covered

- 100 Years represents
  - Dozens of new generations of technologies
  - 100’s of new data standards and formats in many communities
  - Thousands+ of new valued collections
  - Millions of potential users with as yet unknown information needs and workflows

What can we do?

- Plan for preservation throughout all stages of the digital life cycle – from creation to stewardship and beyond
- Ensure that IT policies include requirements that support preservation
- Ensure that IT budgets at all levels include appropriate costs for adequate levels of preservation
- Coordinate creative solutions throughout the data pyramid to spread responsibility and cost
Many Thanks

Chaitan Baru, Jennifer Schopf, Brian Lavoie, Brian Schottlaender, Mark Miller, Alex Szalay, Reagan Moore, Authors of the IDC Report, Ben Tolo, Richard Moore, Mike Norman, David Minor, Amit Chourasia, Jack Dongarra, Natasha Balac, U.S. Library of Congress, Moores Cancer Center, National Archives and Records Administration, NSF, Southern California Earthquake Center, Chris Greer, Steve Cutchin, UCAR, NVO, NASA, and many others

www.sdsc.edu

berman@sdsc.edu