Harnessing the Data Deluge

MURPA Presentation, March 2009

Dr. Francine Berman

Director, San Diego Supercomputer Center
Professor and High Performance Computing Endowed Chair,
UC San Diego
Data Drives 21st Century Research and Education

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

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

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

Where are the brown dwarfs?
NVO: Data from 50+ astronomical sky surveys and large-scale telescopes.

Fran Berman
Information Technologies Drive 21st Century Research and Education

- **Means to an end:** Cyberinfrastructure is the foundation for modern research and education

- **Cyberinfrastructure components:**
  - Digital data
  - Computers
  - Wireless and wireline networks
  - Personal digital devices
  - Scientific instruments
  - Storage
  - Software
  - Sensors
  - People …

Cyberinfrastructure: the organized aggregate of information technologies coordinated to address problems in science and society.

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

NSF Final Report of the Blue Ribbon Advisory Panel on Cyberinfrastructure
Data Cyberinfrastructure Needs Vary Over the Spectrum of Research Applications

- **Data-intensive applications**
  - Home, Lab, Campus, Desktop Applications
- **Data-intensive and Compute-intensive HPC applications**
  - Compute-intensive HPC Applications

- **Compute-intensive HPC applications**
  - Grid Applications
- **Network (more BW)**
  - Data Grid Applications
- **DATA (more BYTES)**
  - Storage of Data from the CERN Large Hadron Collider
  - Analysis with Protein Data Bank structures

- **NETWORK (more BW)**
  - SETI@Home, BOINC
  - Development of biofuels
  - Cosmology

- **COMPUTE (more FLOPS)**
  - Development of biofuels
  - SETI@home, BOINC
Today’s Presentation

- **Research and Data**
  *Data-Driven Cosmology*: Simulating the first billion years of the Universe after the Big Bang

- **Data Cyberinfrastructure**
  Developing a support system for data

- **Towards a Data Master Plan**
  Creating a framework to harness the data deluge
**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
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
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 = $65,536^3$**
Enzo Data Volumes

• 2048\(^3\) simulation (2008)
  – 8 gigazones \(\times\)
  – 16 fields/zone \(\times\)
  – 4 bytes/field
  – = 0.5 TB/output \(\times\)
  – 100 outputs/run
  – = 50 TB

• 4096\(^3\) simulation (2009)
  – 64 gigazones \(\times\)
  – 16 fields/zone \(\times\)
  – 4 bytes/field
  – = 4 TB/output \(\times\)
  – 50 outputs/run
  – = 200 TB
Greater Simulation Accuracy Requires More Computing and Generates More Data

ENZO at Petascale ($10^{15}$)

- 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
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

- 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
Data Cyberinfrastructure
How Much Data Is There?

- **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).

### Data Examples:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Abbreviation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kilo</td>
<td>K</td>
<td>10^3</td>
</tr>
<tr>
<td>Mega</td>
<td>M</td>
<td>10^6</td>
</tr>
<tr>
<td>Giga</td>
<td>G</td>
<td>10^9</td>
</tr>
<tr>
<td>Tera</td>
<td>T</td>
<td>10^12</td>
</tr>
<tr>
<td>Peta</td>
<td>P</td>
<td>10^15</td>
</tr>
<tr>
<td>Exa</td>
<td>E</td>
<td>10^16</td>
</tr>
<tr>
<td>Zetta</td>
<td>Z</td>
<td>10^21</td>
</tr>
</tbody>
</table>

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

- **SDSC Tape Archives = 25+ PB**

- **YouTube: 6M videos in 2006 = 600 TB**

Characteristics of Research Data Vary Widely

- Need for retention: **short-term** (days, months, few years) to **long-term** (decades, centuries, …)
- **Size:** Small-scale (GBs) to large-scale (PBs)
- **Attention:** Well-tended (metadata, community standards) to poorly tended (flat files, insufficient metadata)
- **Limitations:** Subject to more restrictive policy and regulation (HIPAA) vs. subject to less restrictive policy and regulation (OMB)
- **Planned stewardship:** data management and sustainability plan (PDB, PSID, NVO) vs. ad hoc approach
Data Cyberinfrastructure – Coordinated Resources and Services are Key

Many Data Sources

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

Computers

Sensor-nets

Instruments

Visualization

Analysis

Simulation

Modeling

Data Access

Data Use

Data Management

Data Storage

Coordination is Key

Useful Services

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

#### COMPUTATIONAL SYSTEMS

<table>
<thead>
<tr>
<th>Blue Gene (until Triton comes on-line)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 17.1 TF IBM system</td>
</tr>
<tr>
<td>• 1.5 TB total memory</td>
</tr>
<tr>
<td>• 3 racks, each with 2,048 PowerPC processors and 128 I/O nodes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TeraGrid Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 524 Itanium2 IA-64 processors</td>
</tr>
<tr>
<td>• 2 TB total memory</td>
</tr>
</tbody>
</table>

**Coming in Spring: Triton Resource**

| PSDAF data analysis supercomputer |
| Shared Condo cluster |
| Data Oasis high perf. disk |

#### DATA CYBERINFRASTRUCTURE

| 2.4 PB Storage-area Network (SAN) |
| 36 PB StorageTek/IBM tape library |
| HPSS and SAM-QFS archival systems |
| Oracle, MySQL |
| Storage Resource Broker, IRODS |
| Expertise in data mining, data curation, long-term preservation, etc. |

### CYBERINFRASTRUCTURE SERVICES AND COMMUNITY EXPERTISE

| Application collaborations |
| Education and training |
| Data storage and preservation |
| User Services |
| Computation, co-location, web, DB, other CI services |

**Data access, use, management, storage, and preservation**
New SDSC Resources to Support Data Science:

Triton Resource

**Data Oasis**
- 2 – 4 Petabytes
- Large-scale Storage for research data sets

**PSDAF**
- Designed for analysis of very large data sets

**Shared Resource Cluster**
- Condo-style cluster "starter" to create expandable UCSD computing resources.

**UCSD/UC Research Labs**
- Repository for valuable research databases to be administered with UCSD Libraries

**UCSD**
- Unique computer for extreme-scale data analysis
- Seed for UCSD's "home grown" supercomputer
## SDSC Resources for Computational and Data-driven Research

### Computational Systems

**Blue Gene (until Triton comes on-line)**
- 17.1 TF IBM system
- 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

**Coming in Spring: Triton Resource**
- PSDAF data analysis supercomputer
- Shared Condo cluster
- Data Oasis high perf. disk

### Data Cyberinfrastructure

- 2.4 PB Storage-area Network (SAN)
- 36 PB StorageTek/IBM tape library
- HPSS and SAM-QFS archival systems
- Oracle, MySQL
- Storage Resource Broker, IRODS
- Expertise in data mining, data curation, long-term preservation, etc.

### Cyberinfrastructure Services and Community Expertise

- Application collaborations
- Education and training
- Data storage and preservation
- User Services
- Computation, co-location, web, DB, other CI services

---

Fran Berman
What is Chronopolis?

Data Users

Data Providers
(User-focused repositories, libraries, etc.)

Data Grid supporting a Long-term Preservation Service

Data Migration to next generation technologies

Replication of data at multiple, geographically distinct sites

Trust Agreements between sites

Who is Chronopolis?

Chronopolis is being developed by a national consortium led by SDSC and the UCSD Libraries (UCSDL) and funded by the Library of Congress.

Initial Chronopolis nodes include:

- SDSC/UCSDL at UCSD
- University of Maryland Institute for Advanced Computer Studies (UMIACS)
- National Center for Atmospheric Research (NCAR) in Boulder, CO
Current Chronopolis Collections

Collections from Data Providers:

- **Inter-university Consortium of Political and Social Research** – preservation copy of all collections including 40 years of social science data and Census (8 TB)

- **California Digital Library** – political and government web crawls, Web-at-risk collection (5 TB)

- **SIO Explorer** – data from 50 years of research voyages (1 TB)

- **NCSU Libraries** – State and local geospatial data (6 TB)

[http://chronopolis.sdsc.edu](http://chronopolis.sdsc.edu)
Inside Chronopolis

- Sites linked by main staging grid where data is verified for integrity, and quarantined for security purposes.
- Collections independently pulled into each system.
- Manifest layer provides added security for database management and data integrity validation.
- Benefits
  - Each collection copy independently managed
  - Collections available from each site?
  - High reliability
Preservation is Part of the Larger Data Life Cycle

• In the next year, the amount of generated data will climb into the zettabytes

• To optimize the information available in this data, it is time to take a comprehensive and strategic approach to its storage, use, and preservation
Towards a Data Master Plan
Why we need a Comprehensive Approach to Data

• We need a Master Plan which relates multiple scenarios with appropriate underlying cyberinfrastructure, economic model, use case, policy / regulatory constraints, etc.

• No one-size-fits-all model for data. Degrees of freedom include
  – How data is used
  – Required reliability / performance of underlying infrastructure
  – Privacy / policy constraints
  – Who generates /acts as stewards
  – Desirable retention periods
  – Supporting economic models
  – Who values the data, etc.
Key Components of a Master Plan

- Framework for thinking about perceived data value
- Framework for thinking about stewardship and costs
- Approach for assessing data needs and storage solutions
- Quantifiable measures of success
What data is valuable?

To whom is it valuable?
Creating a Framework for Thinking About Data: Valued Data Collections

The Data Pyramid

- Societal Value
- Community Value
- Personal Value

United States Census 2000
ARTstor
PDB
R. Acknowledgment Berman, Jr.
Family Businesses
Quicklen Deluxe 2003
You Tube

Fran Berman
**Data of Long-term Cultural Value:**  
**Historical Data**

- The 2008 U.S. 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
What Level of Stewardship is Desirable / Needed?

- **“Gold standard”**
  - Preservation plan for a sufficient timeframe
  - Replication
  - Appropriate level metadata, monitoring, management, etc.
  - System compliance to appropriate policy, regulations

- **“Silver standard”**
  - Backup copy
  - Appropriate level metadata, monitoring, management, etc.
  - User/system compliance to appropriate policy, regulations

- **“Brass standard”**
  - Best effort storage
  - User-provided metadata, management, monitoring
  - User compliance to appropriate policy, regulations

The Data Pyramid

Increasing reliability, increasing expense
**Data Census: Assessing Data Needs and Storage Solutions**

- How much data is generated by federally funded research?

- How much of it must be retained by policy or regulation (OMB Circular A-110, Sarbanes-Oxley, HIPAA, etc.) and for how long?

- How much data is generated at various sizes -- <1GB, <1TB, <10 TB, <100 TB, … -- and what are their retention characteristics? …

- Where is the data funded by federal agencies stored?

- What is the level of reliability (e.g. “gold”, “silver”, “brass”) of the data hosting environment?

- How many copies of the data are retained?

- How long has the data been retained? What timeframe is planned for retention in the future?

- Who makes decisions about the data? …

Gauging the specifics of the problem will have an enormous impact on how we frame the solution.
Creating a Data Master Plan: What is Success?

- What is minimal data loss?
- What is high reliability?
- What characterizes a trustworthy repository?
- What characterizes a sustainable collection?
- Who decides how collections should be hosted / managed / preserved for stakeholders?
- Etc.
Creating a Data Master Plan

• Gather information about data generated, uses, needs, relevant policies, and environments via a Data Census

• Create a framework for quantifying key characteristics of data stewardship environments and stewardship costs

• Match data needs to stewardship and storage environments / solutions

• Develop quantifiable metrics of success to assess data cyberinfrastructure solutions
Harnessing the Data Deluge
A Cyberinfrastructure Grand Challenge

- Digital life cycle support creates key challenges for building and deploying Cyberinfrastructure
  - Approaches for sufficient levels of system reliability and security
  - Development of plans to smoothly migrate digital materials from one generation of technology to the next
  - Development of indexing / organizing / metadata structures and ontologies to express relationships between data
  - Development of management systems to support policy and regulation
  - Development of more advanced algorithms to support searching, mining, analysis, etc.
Thank you

www.sdsc.edu