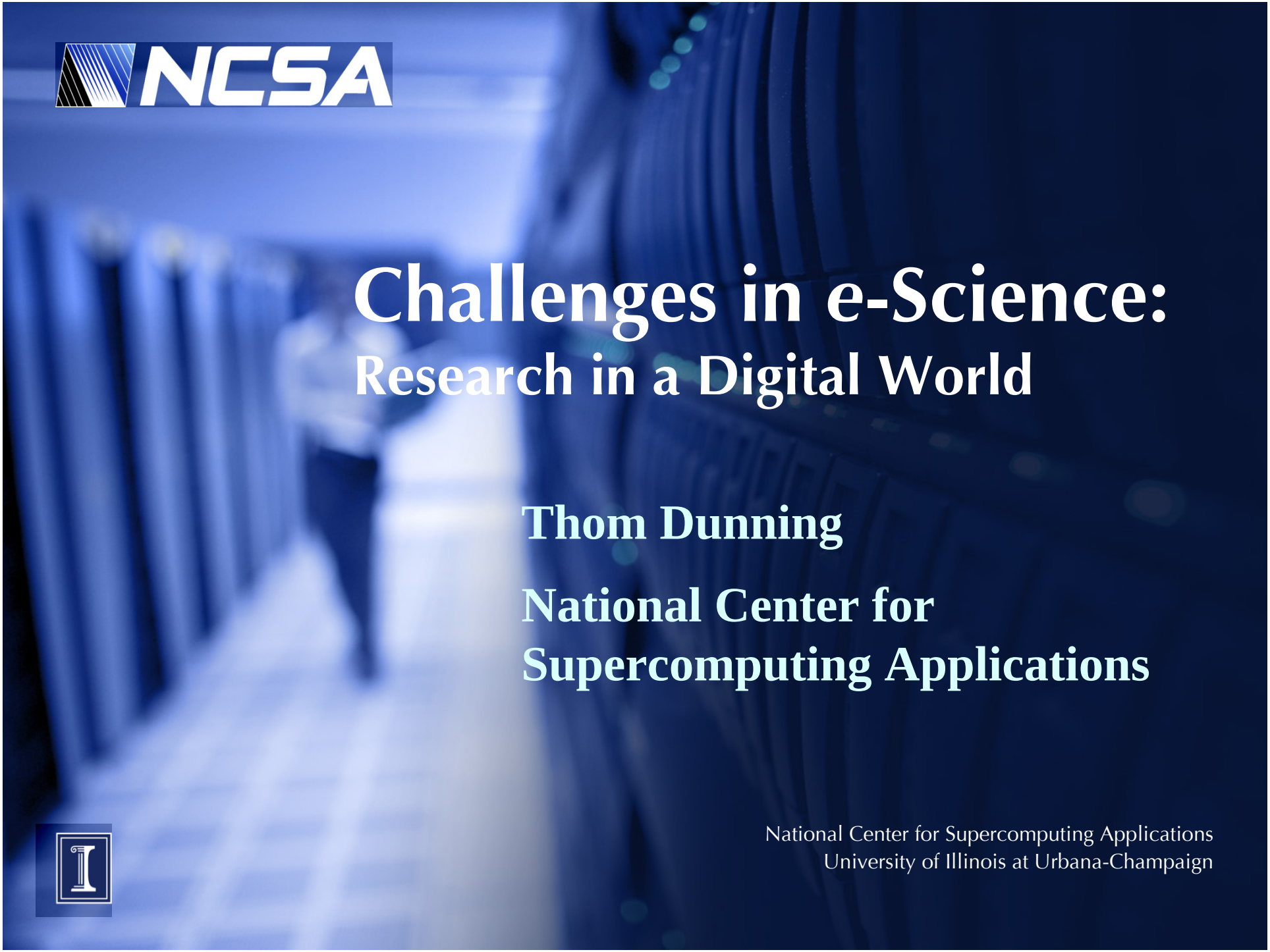


Challenges in e-Science:
Research in a Digital World

Thom Dunning

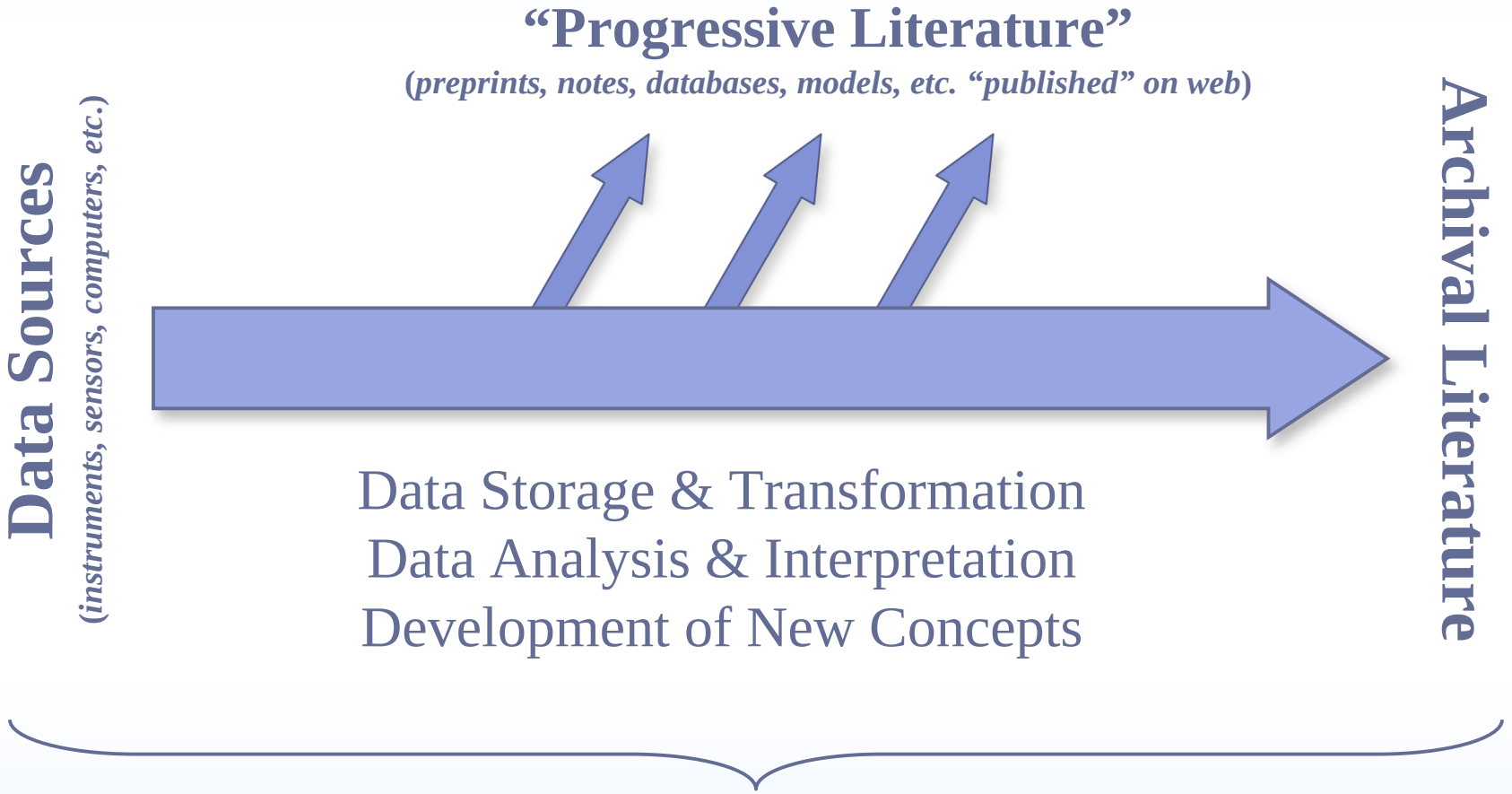
National Center for Supercomputing Applications

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

li-brary, ***n.*** a collection of books, journals, reference materials, films, recorded music, etc., organized systematically and kept for research or borrowing.

Oxford English Dictionary

Research Process

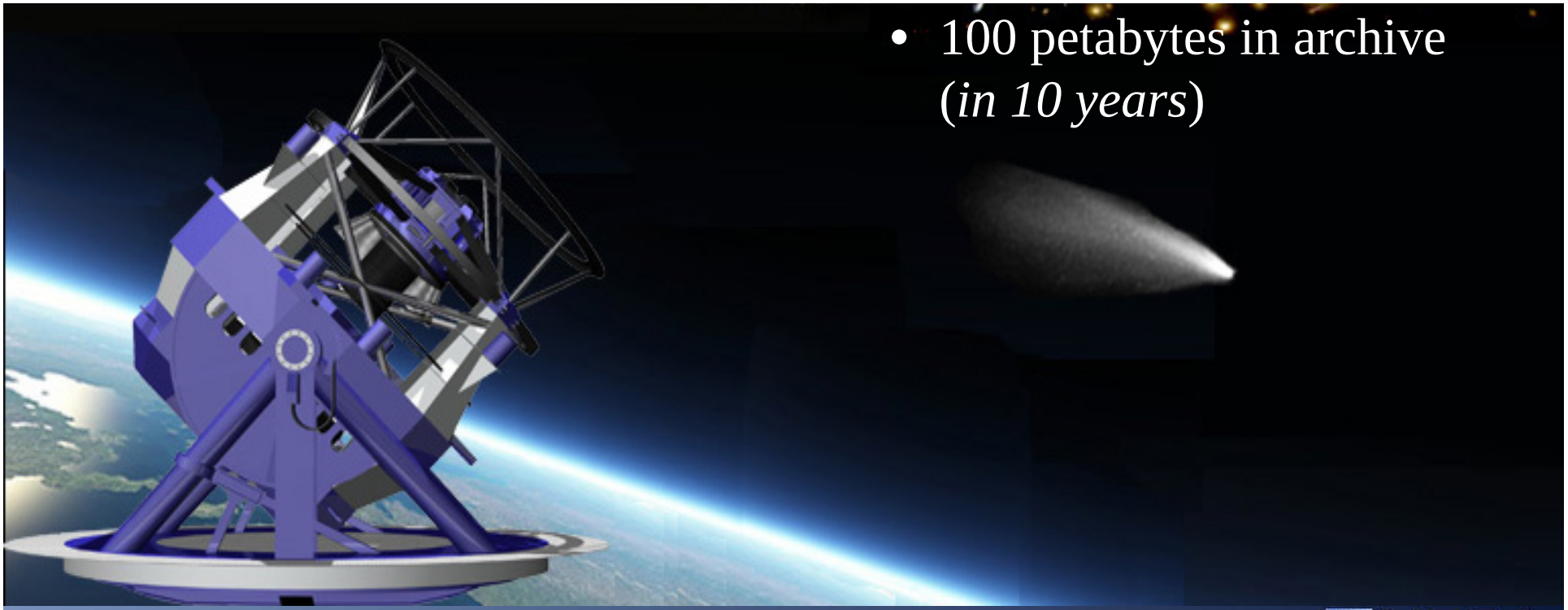


e-Science: All Digital

Data Sources

Astronomical Observatories

- **SDSS**
  - Map 1/4-th of sky
  - 16 terabytes in archive (DR6)
- **LSST**
  - Map 1/3-rd of sky *per night*
  - 15-20 terabytes *per night*
  - 100 petabytes in archive (*in 10 years*)



Data Sources

Large Synoptic Survey Telescope (LSST)

- **New Telescope**
  - Located in Chile (El Peñon) with first light in 2013
  - 8.4-m Mirror with 3 Gigapixel camera
  - Image available sky every 3 days
- **Science Missions**
  - Nature of dark energy and accelerating universe
  - Comprehensive census of solar system objects, create galactic map
  - Explore transients and variable objects
- **Data Sets**
  - 15-20 terabytes per night
  - 50-100 petabytes in 10 years

Data Sources

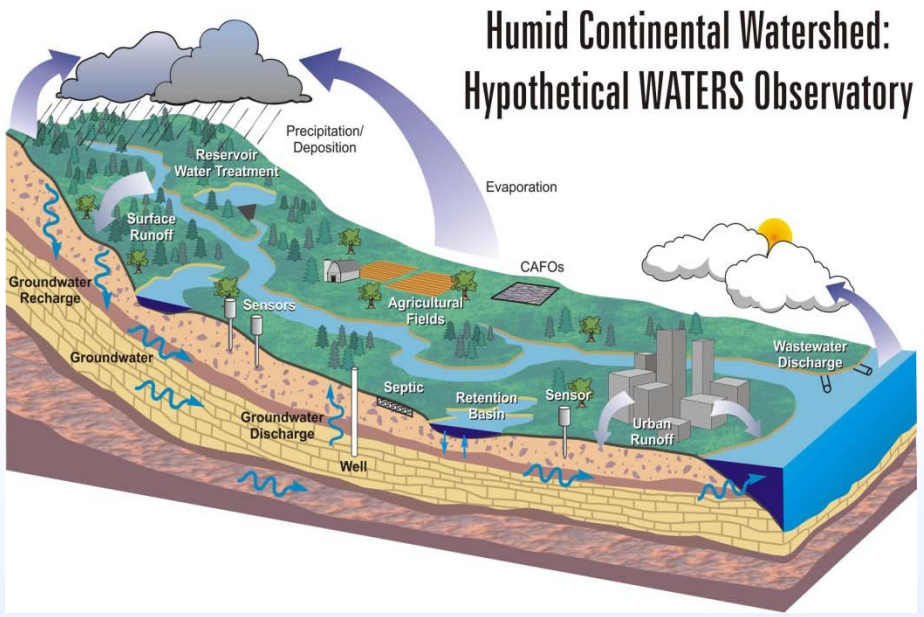
Environmental Observatories

- **Sensors and Sensor Networks**
  - Intensively instrumented sites shared by research community
- **Cyberinfrastructure**
  - Sites, data, computers and researchers connected by high bandwidth networks

neon

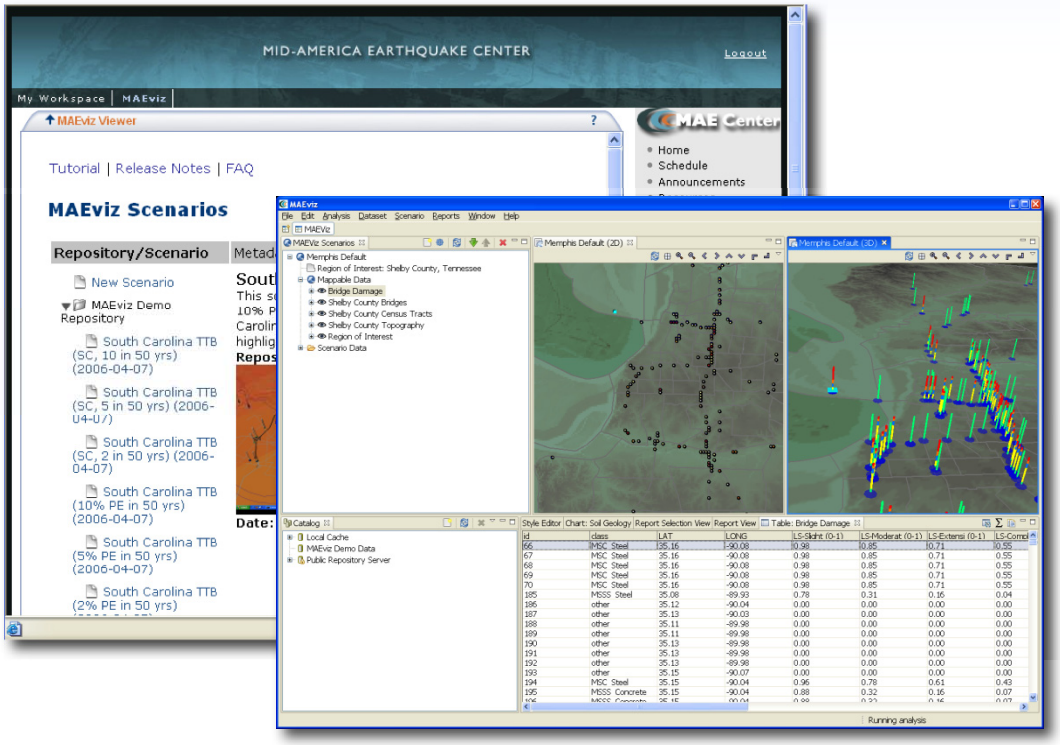
WATERS NETWORK

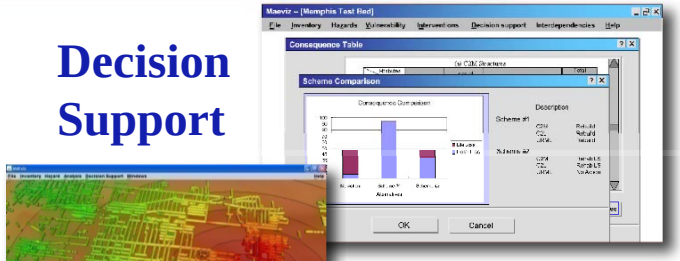
CONSORTIUM FOR Ocean Leadership
DISCOVERY • UNDERSTANDING • ACTION



Data Sources

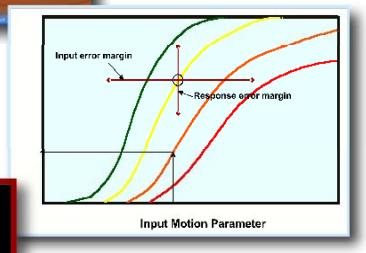
Cyberenvironment for Earthquake Engineering



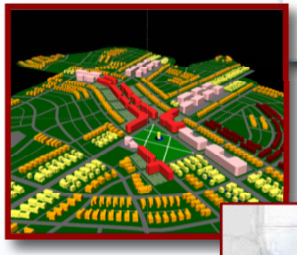


Decision Support

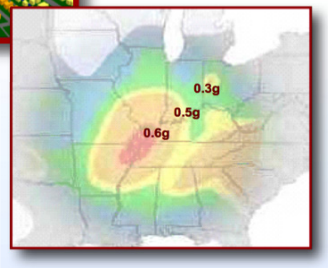
Damage Prediction



Fragility Models



Inventory Selection



Hazard Definition

- **Mid-America Earthquake Center**
  - ***Consequence-based risk management for seismic events***
    - Portal-based collaboration environment
    - Distributed data/metadata sources
    - Multi-disciplinary collaboration

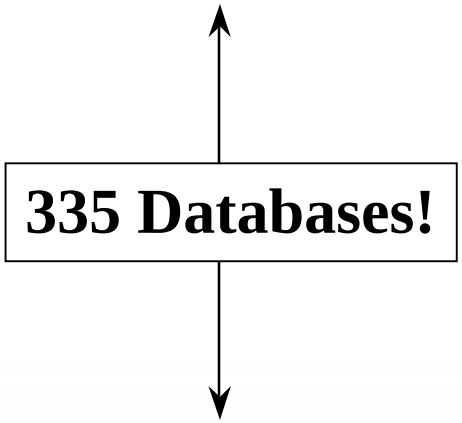
Progressive Literature

Molecular Biology Databases 2002

Nucleic Acids Research, 2002, Vol. 30, No. 1

Table 1. Molecular Biology Database Collection

Major Public Sequence Repositories		
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp	All known nucleotide and protein sequences …
Varied Biomedical Content		
…		
VirOligo	http://viroligo.okstate.edu	Virus-specific oligonucleotides for PCR and …



Progressive Literature

Molecular Biology Databases 2008

Nucleic Acids Research, 2002, Vol. 30, No. 1

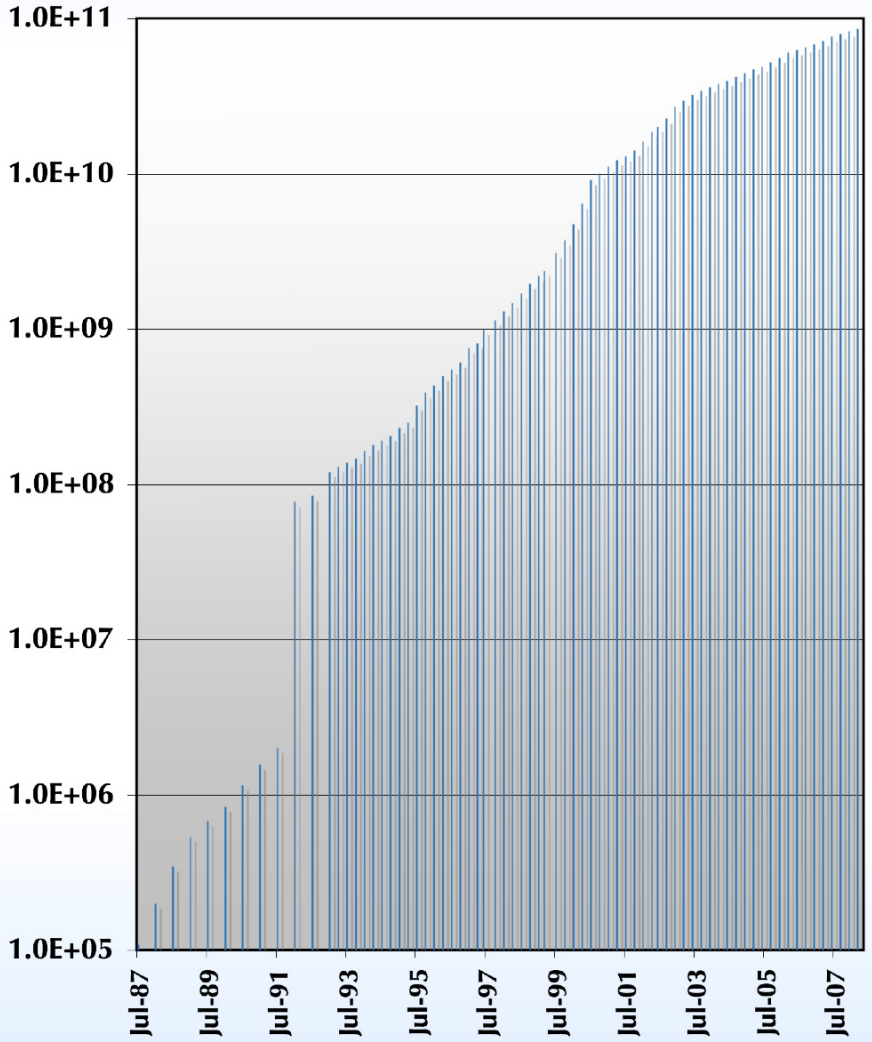
Table 1. Molecular Biology Database Collection

Major Public Sequence Repositories		
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp	All known nucleotide and protein sequences
…		
1078 Databases!		
Varied Biomedical Content		
…		
VirOligo	http://viroligo.okstate.edu	Virus-specific oligonucleotides for PCR and
…		

Progressive Literature

Growth of GenBank

- **Nucleotide Sequence Dataset**
  - Exponential growth with doubling time of 6-12 months
  - 182 billion base pairs from 300,000 organisms
  - 4192 organisms completely sequenced
  - Raw data rapidly approaching 1 terabyte, processed data 10's terabytes



Archival Literature

A Missed Opportunity

Manipulate and analyze molecular structures

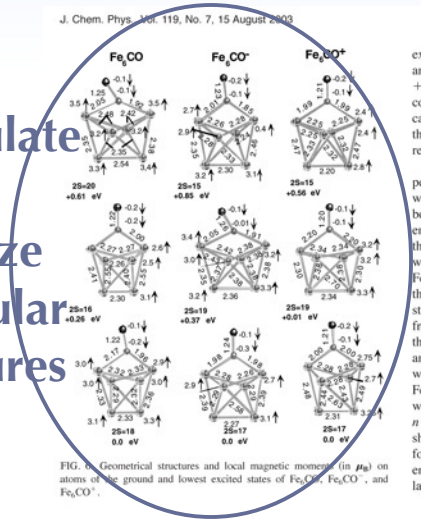


FIG. 6. Geometrical structures and local magnetic moments (in μ_B) on atoms of the ground and lowest excited states of Fe_6CO, Fe_6CO^-, and Fe_6CO^+.

paring Fe_n–CO and Fe_{n-1}–Fe bond strengths, one sees that the latter energies are about 1 eV larger, as might be expected.

G. Catalytic ability of iron clusters

The energetics of the Boudouard-type disproportionation reactions $Fe_nCO+CO \rightarrow Fe_nC+CO_2$, $Fe_nCO^-+CO \rightarrow Fe_nC^-+CO_2$, and $Fe_nCO^++CO \rightarrow Fe_nC^++CO_2$, for $n=1-6$, are presented in Table VIII. A plus sign indicates the reaction is endothermic. As is seen, the effectiveness of iron clusters grows rapidly with n, and already Fe_4 shows a slight exothermicity which remains nearly the same for larger Fe_5 and Fe_6. The cationic channel $Fe_6CO^++CO \rightarrow Fe_6C^++CO_2$ has the highest exothermicity among all the channels considered, while anionic channels are less favorable because the electron affinities of Fe_nCO clusters are larger those of Fe_nC (see Table VI), thus stabilizing the reactants relative to the products.

Since the Fe_n–CO binding energies are relatively independent of the cluster size, the change in reaction energy with cluster size must be due to the change in the Fe_nC–O bond, which are summarized in Table IX. The Fe_nC–O bond energies are much smaller than that of free CO (11.18 eV at the BPW91/6-311+G* level which is in good agreement with the experimental value[63] of 11.09 eV). Unlike the Fe_n–CO bond energies, which are relatively independent of the cluster size, there is a decrease in the Fe_nC–O bond strength from $n=1$ to $n=4$, and then it is slowly varying from $n=4$ to $n=6$. Another way to look at the weakening of the Fe_nC–O bond is the strengthening of the Fe_n–C bonds, and there are experimental results for the Fe_n–C^+ species, which are summarized in Table IX along with our computed Fe_n^+–C. The computed results are in qualitative agreement with experiment. While the computed results only extend to $n=6$, the experiment extends to $n=15$, and experiment shows that the Fe_n–C^+ bond energies increase only slightly for $n=6$ to $n=15$. Therefore we expect the Fe_nC–O bond energies and hence the Fe_nCO+CO reaction energies for the larger clusters to be similar to those for $n=4$ to $n=6$.

H. Barrier heights for the $Fe_nCO+CO \rightarrow Fe_nC+CO_2$ reactions

While the reaction energies are of interest, the barrier height is more critical in evaluating the reaction rates, and therefore we have determined the transition state for the $FeCO+CO \rightarrow FeC+CO_2$ and $Fe_4CO+CO \rightarrow Fe_4C+CO_2$ reactions. Figure 7 shows the geometries of the transition states found along with energetics of the corresponding reaction channels. The FeC···OCO transition state is a planar $^3A''$ state, while transition state of Fe_4C···OCO ($2S=13$)

TABLE IV. Computed vibrational frequencies (in cm^{-1}) and intensities (in km/mol) of ground-state neutral and charged Fe_2CO, Fe_4CO and their ions.

	Fe_2CO^a	Fe_2CO^b	Fe_2CO^-	Fe_2CO^+	Fe_4CO	Fe_4CO^-	Fe_4CO^+
ω_1	49[1.0]	46[1.2]	69[1.9]	43[0.7]	37[0.5]	35[0.7]	38[0.4]
ω_2	55[2.3]	52[2.5]	71[0.4]	70[2.5]	65 [0.0]	134[1.3]	64[3.1]
ω_3	189[0.5]	189[0.7]	208[3.6]	182[1.3]	119[0.1]	135[1.7]	151[0.1]
ω_4	219[9.0]	217[8.9]	237[0.1]	198[4.5]	140 [2.5]	172[0.0]	159[1.0]
ω_5	284[8.5]	282[9.3]	315[0.1]	303[4.2]	210[0.0]	223[1.6]	217[6.3]
ω_6	336[0.2]	333[0.2]	343[0.5]	312[0.1]	245[3.3]	225[0.1]	220[2.0]
ω_7	365[0.4]	365[0.4]	350[1.7]	346[3.6]	255[0.2]	264[0.0]	250[1.8]
ω_8	481[0.2]	483[0.2]	422[2.9]	436[13.2]	312 [1.3]	340[1.8]	283[4.1]
ω_9	1805[828]	1809[834]	1653[753]	1975[707]	317 [2.5]	348[0.0]	322[0.9]
ω_{10}					360[1.7]	358[8.5]	359[3.9]
ω_{11}					402[0.2]	422[3.5]	443[8.6]
ω_{12}					1753[683]	1674[778]	1907[775]

[a]Standard thresholds.
[b]Tighter thresholds, i.e., ultrafine and OPT=TIGHT.

TABLE 1. Recommended enthalpies of formation of C1 and C2 hydrocarbons from several commonly cited sources. Uncertainties (if given) are those of the cited source. We have selected the values of Gurvich *et al.* [1991GVA] (in bold, see also text in Secs. 3.1, 4.1, 5.1, and 6.1)

Reference	$\Delta_f H°[CH_4$ (g), 298.15 K] (kJ mol^{-1})	$\Delta_f H°[C_2H_2$ (g), 298.15 K] (kJ mol^{-1})	$\Delta_f H°[C_2H_4$ (g), 298.15 K] (kJ mol^{-1})	$\Delta_f H°[C_2H_6$ (g), 298.15 K] (kJ mol^{-1})
[2001B]	−74.60	227.4	52.3	−83.85
[2001DIP]	−74.52	228.2	52.51	−83.82
[1998C]	−74.873 ± 0.34[a]	226.73 ± 0.79[b]	52.467 ± 0.29[b]	—
[1994FKM][c]	−74.5	228.2	52.5	−83.8
[1992ABC]	−74.81	228.0	52.2	−84.0
[1991GVA] (selected values)	**−74.6 ± 0.3[d]**	**227.4 ± 0.8[e]**	**52.4 ± 0.5[0][e]**	**−84.0 ± 0.4[0][e]**
[1986PNK]	−74.40 ± 0.40	228.20 ± 0.70	52.5 ± 0.4	−83.80 ± 0.40
[1985TRC]	−74.475[c]	228.2[f]	52.51[g]	−83.85[c]
[1982PRS]	−74.48 ± 0.42	—	—	−83.85 ± 0.09
[1975CZ]	—	—	52.51 ± 0.63	—
[1970CP]	−74.85 ± 0.29	227.36 ± 0.79	52.09 ± 0.42	−84.68 ± 0.50
[1969SWS]	−74.85	226.73	52.45	−84.68

[a]Evaluation date 1961.
[b]Evaluation date 1965.
[c]This is a compilation of the data sheets of the TRC Tables, as detailed in footnotes c, d, and e.
[d]Evaluation date uncertain, value is unchanged from previous edition, [1979G].
[e]Data sheet 1010, Evaluation date 1981, Ref. [1981C].
[f]Data sheet 3040, Evaluation date 1993, Ref. [1993KWD].
[g]Data sheet 2500, Evaluation date 1981, Ref. [1981C].

where P_{sat}, T, and B are the saturated vapor pressure, temperature in Kelvin, and second virial coefficients, respectively. The vapor pressures and second virial coefficients were taken from the DIPPR Tables [2001DIP]. Where possible we have checked our estimates with previous calculations by Majer and Svoboda, [1985MS], although they do not report values for all compounds of present interest. Our calculated values and those of Majer and Svoboda are plotted versus the normal boiling point in Fig. 1. The values for chloroethene and trichloroethene calculated from the DIPPR data appear to be incorrect and were not used. Additional details can be found in the evaluations and at the NIST Kinetics Database website [2001KIN].

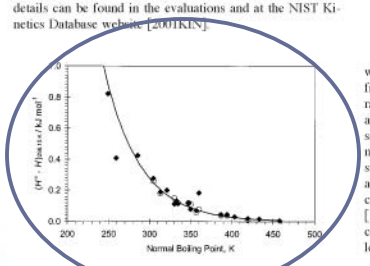


FIG. 1. The enthalpic departure function [from Eq. (1)] at 298.15 K vs the normal boiling point T_b for the C1 and C2 chlorinated hydrocarbons. Filled symbols are values calculated by us taking data on the saturated vapor pressure and second virial coefficients from [2001DIP]. Open symbols are from [1985MS]. Two values calculated from the DIPPR data, those of chloroethene and trichloroethene, do not fall on the curve indicated by the other points. We were unable to determine an obvious reason for this, but these data were not used. The empirical fit to the data is given by $(H°-H)_{298} = 250.35\exp(-0.227\,T_b)$.

A second issue has to do with the extrapolation of values of $\Delta_{vap}H$ at a particular temperature to the temperature of interest. There are numerous methodologies [1987RPP] for doing this that require knowledge of the critical pressure and temperature of the relevant species. Such data are not always available and we examine an alternative approach applicable to the limited range of compounds and temperatures considered herein. The general thermodynamic relation is:

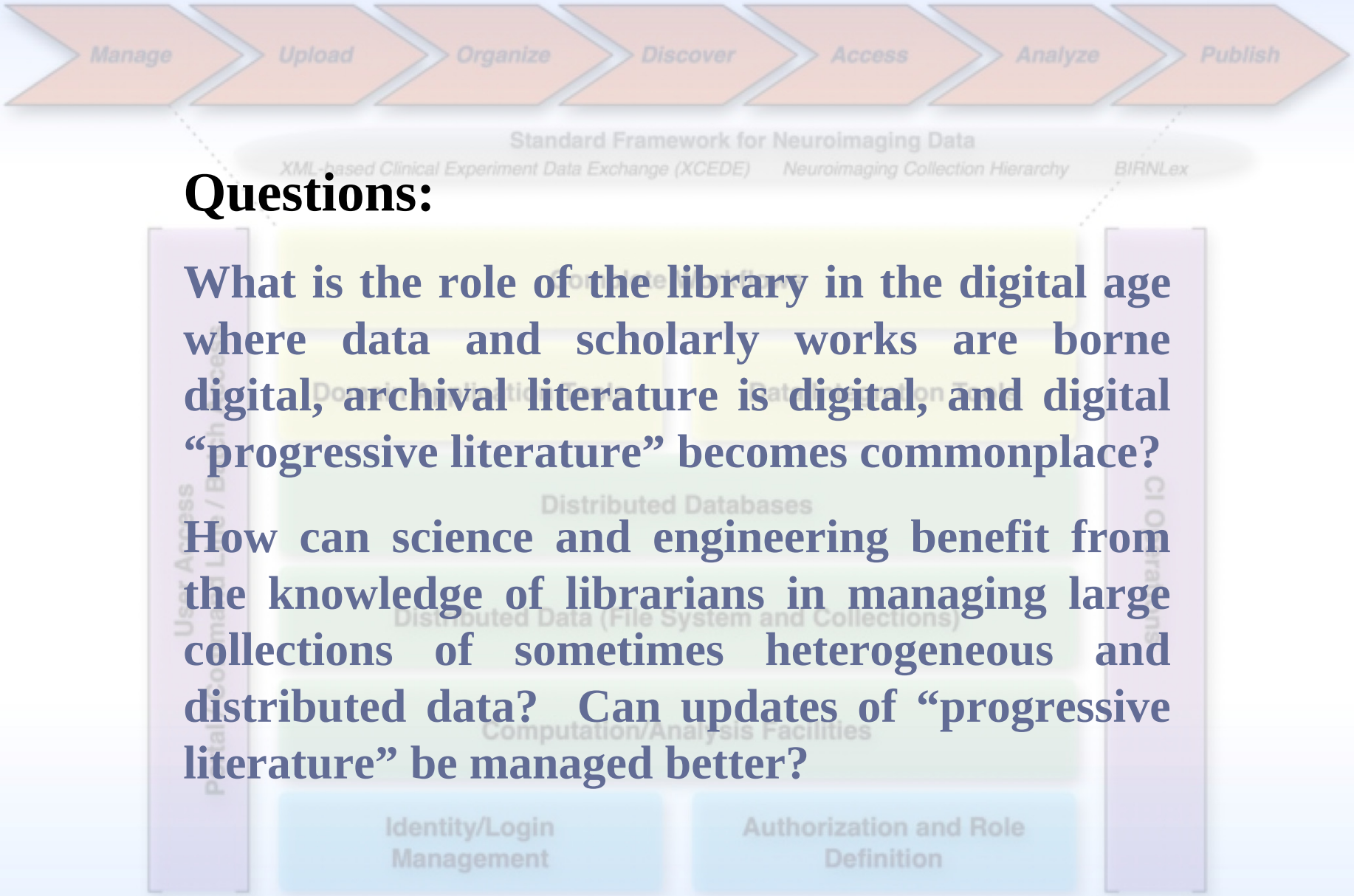
$$\Delta_{vap}H(T_2)=\Delta_{vap}H(T_1)+\int_{T_1}^{T_2}\Delta_{vap}C_p\,dT,$$

where $\Delta_{vap}C_p$ is the change in the heat capacity in going from the condensed [illegible] gas. Over the moderate ranges of temperature typically encountered, $\Delta_{vap}C_p$ is usually approximately constant for a given molecule (*vide infra*, see Fig. 8 in Section 6.9). Its value is sometimes taken to be near $\Delta_{vap}C_p = -$[illegible] $J\,K^{-1}\,mol^{-1}$ [illegible] assumed to be independent of the chemical structure. In actuality there are no compelling reasons for this quantity to be constant across a [illegible] Shikos *et al.*, [1993CHH] for example, [illegible] data on a variety of compounds and concluded that $\Delta_{vap}C_p$ increased with molecular size.

In a related approach, for the chlorinated hydrocarbons we have correlated this property with the normal boiling points of the compounds. Figure 2 shows clearly that the value of $\Delta_{vap}C_p$ increases with the normal boiling point of the species. A good straight line is obtained for the chloroalkanes with an intercept of very close to zero. The intercept can be rationalized since $\Delta_{vap}C_p$ should be related to the intermolecular forces in the condensed phase and those forces must

Extract data from figures

+ animation and more



Questions:

What is the role of the library in the digital age where data and scholarly works are borne digital, archival literature is digital, and digital "progressive literature" becomes commonplace?

How can science and engineering benefit from the knowledge of librarians in managing large collections of sometimes heterogeneous and distributed data? Can updates of "progressive literature" be managed better?

Blue Waters:
Brave new world to explore

University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications, IBM, and Great Lakes Consortium

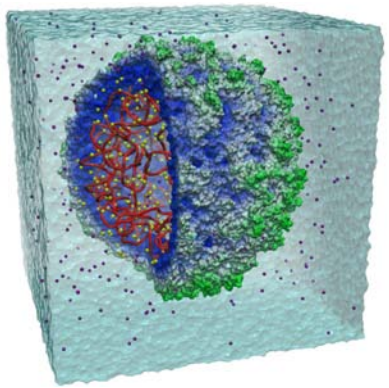
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

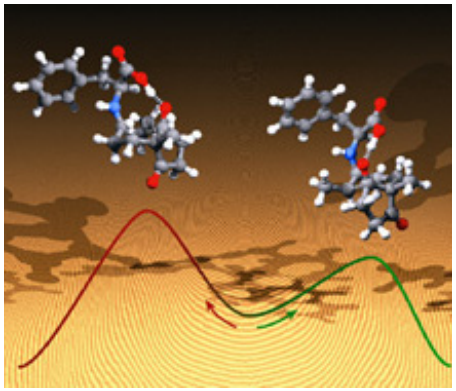
NSF Solicitation and Award

- **NSF Solicitation**
  - Request for computing system capable of *sustained performance approaching a petaflops* (10^{15} floating point operations per second) on real applications that consume *large amounts of memory* and/or that work with *very large data sets* (NSF 06-573)
- **NSF Award**
  - Award to UIUC/NCSA announced August 8, 2007
  - Cooperative Agreement signed on September 28, 2007
  - Project documentation
    - PEP, SOWs, Configuration Management Plan, Risk Mitigation Plan, *etc*. (> 1000 pages)
    - Submitted April 4, 2008

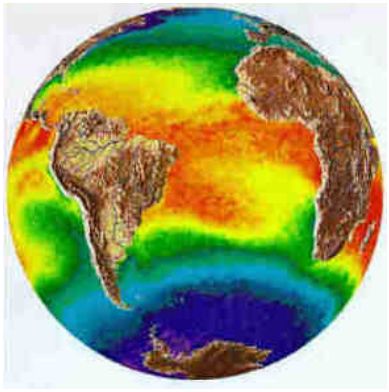
Science @ Petascale

Molecular Science

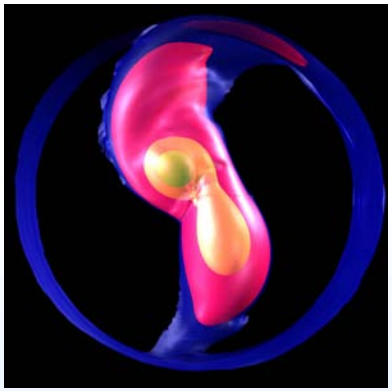




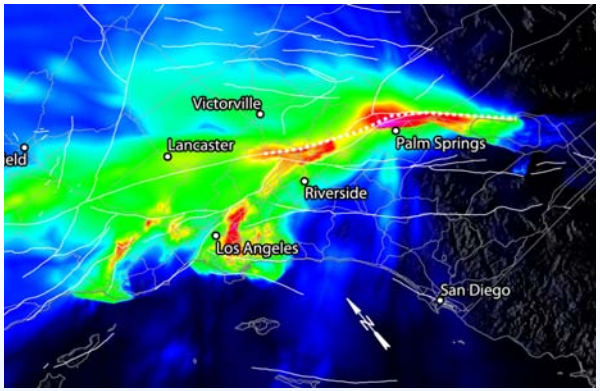
Weather & Climate Forecasting



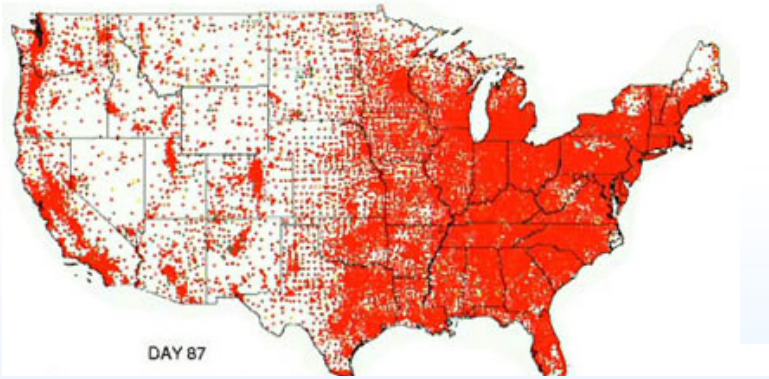
Astronomy



Earth Science



Health



Blue Waters Computing System

System Attribute	Abe	Blue Waters
Vendor	Dell	IBM
Processor	Intel Xeon 5300	IBM Power7
Peak Performance (PF)	0.09	
Sustained Performance (PF)	0.005	≥1
Number of Cores/Chip	4	
Number of Processor Cores	9,600	>200,000
Amount of Memory (PB)	0.0144	>0.8
Amount of Disk Storage (PB)	0.1	>10
Amount of Archival Storage (PB)	5	>500
External Bandwidth (Gbps)	40	100-400

Great Lakes Consortium for Petascale Computation

Goal: Facilitate the widespread and effective use of petascale computing to address frontier research questions in science, technology and engineering at research, educational and industrial organizations across the Great Lakes region and nation.

Charter Members

Argonne National Laboratory
Fermi National Accelerator Laboratory
Illinois Math and Science Academy
Illinois Wesleyan University
Indiana University*
Iowa State University
Illinois Mathematics and Science Academy
Krell Institute, Inc.
Louisiana State University
Michigan State University*
Northwestern University*
Parkland Community College
Pennsylvania State University*
Purdue University*
The Ohio State University*
Shiloh Community Unit School District #1
Shodor Education Foundation, Inc.
Southeastern Universities Research Association
University of Chicago*
University of Illinois at Chicago*
University of Illinois at Urbana-Champaign*
University of Iowa*
University of Michigan*
University of Minnesota*
University of North Carolina–Chapel Hill
University of Wisconsin–Madison*
Wayne City High School

** CIC Universities*

Virtual School of Comp Science & Engineering

- **Members of Virtual School**
  - University of Minnesota, University of Wisconsin, University of Iowa, Iowa State University, University of Illinois, Northwestern University, University of Chicago, Indiana University, Pennsylvania State University, Purdue University, University of Michigan, Michigan State University, The Ohio State University, Louisiana State University (CCT)
- **Activities of Virtual School**
  - Enhance existing graduate courses, new courses for petascale computing
  - Summer schools, workshops and seminars to introduce graduate students to opportunities and challenges in petascale computing
    - 1st Summer School – Accelerators in Science & Engineering Applications: GPUs and Multicores (August 18-22)
  - "Best practices" for graduate programs in computational science and engineering
- **Virtual School Leadership**
  - Sharon Glotzer, University of Michigan
  - Thom Dunning, University of Illinois at Urbana-Champaign

Petascale Computing Facility

Partners

EYP MCF/
Gensler
IBM
Yahoo!

- **Modern Data Center**
  - 90,000+ ft^2 total
  - 20,000 ft^2 machine room
- **Efficiency**
  - LEED certified (goal: silver)
  - Efficient cooling system

Blue Waters Team

PI & Co-PIs

Dunning
Director

Pennington
Deputy Director

Hwu
Hardware

Snir
Software

Seidel
Applications

Task Leads

Broeren
Consortium

Butler
Storage

Cox
Visualization

Glotzer
Virtual School

Iyer
Reliability

Kale
Apps Simulations

Melchi
Facilities

Giles
Industry

Olson
Proj. Mgmt.

Panoff
Education

Shoop
Networking

Towns
Ops Transition

Question:

What is the role of the library in the era of petascale computing? How can libraries benefit from petascale computing and its derivatives?

How can petascale computing benefit from librarians' knowledge of managing large quantities of complex data—the output of many petascale simulations?

Questions?