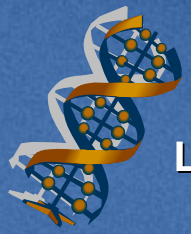
Earth Sciences



Life Sciences

Computer and Information Sciences

e-Science and Cyberinfrastructure

Social Sciences

Tony Hey
Corporate Vice President
Technical Computing
Microsoft Corporation

New Materials, Technologies and Processes

Multidisciplinary Research

Licklider's Vision

"Lick had this concept – all of the stuff linked together throughout the world, that you can use a remote computer, get data from a remote computer, or use lots of computers in your job"

Larry Roberts – Principal Architect of the ARPANET

Physics and the Web

- Tim Berners-Lee developed the Web at CERN as a tool for exchanging information between the partners in physics collaborations
- The first Web Site in the USA was a link to the SLAC library catalogue
- It was the international particle physics community who first embraced the Web
- ‘Killer’ application for the Internet
- Transformed modern world – academia, business and leisure

Beyond the Web?

- Scientists developing collaboration technologies that go far beyond the capabilities of the Web
  - To use remote computing resources
  - To integrate, federate and analyse information from many disparate, distributed, data resources
  - To access and control remote experimental equipment
- Capability to access, move, manipulate and mine data is the central requirement of these new collaborative science applications
  - Data held in file or database repositories
  - Data generated by accelerator or telescopes
  - Data gathered from mobile sensor networks

What is e-Science?

‘e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it’

John Taylor

Director General of Research Councils

UK, Office of Science and Technology

The e-Science Vision

- e-Science is about multidisciplinary science and the technologies to support such distributed, collaborative scientific research
  - Many areas of science are in danger of being overwhelmed by a ‘data deluge’ from new high-throughput devices, sensor networks, satellite surveys …
  - Areas such as bioinformatics, genomics, drug design, engineering, healthcare … require collaboration between different domain experts
- ‘e-Science’ is a shorthand for a set of technologies to support collaborative networked science

e-Science – Vision and Reality

Vision

- Oceanographic sensors - Project Neptune
  - Joint US-Canadian proposal

Reality

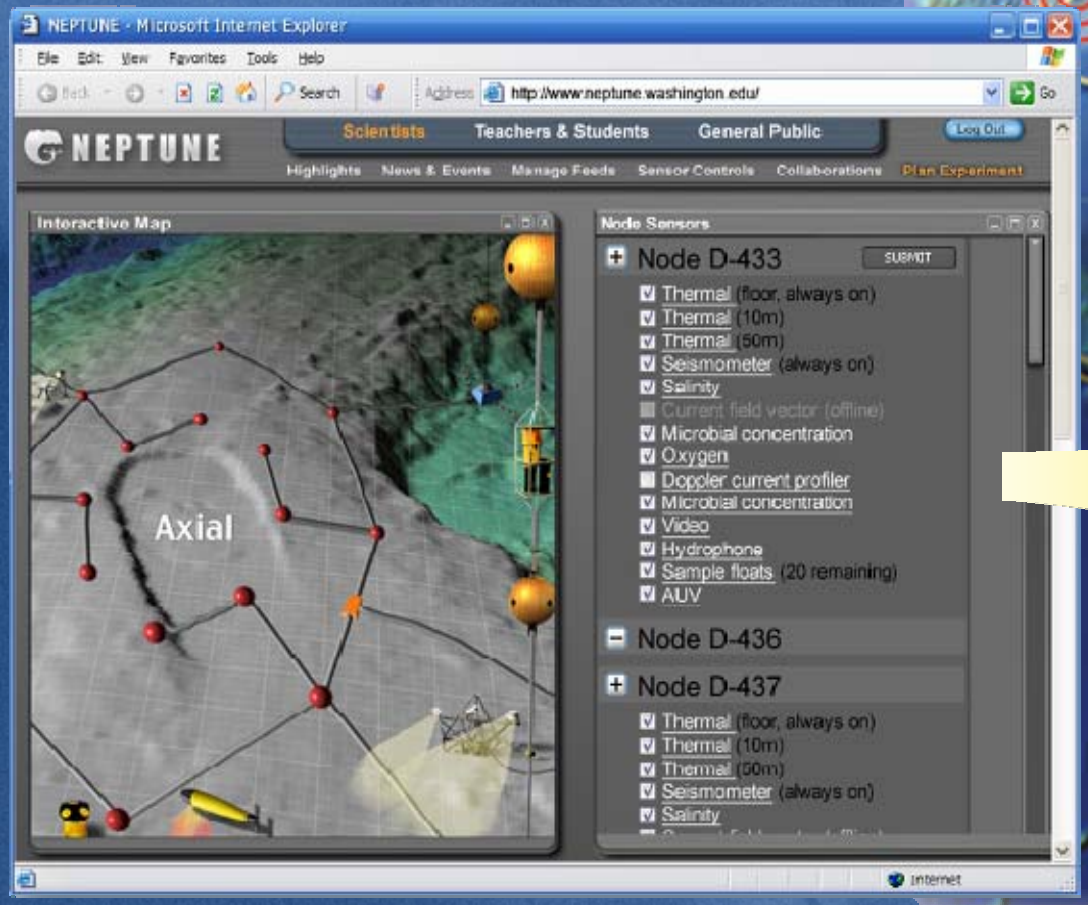
- Chemistry – The Comb-e-Chem Project
  - Annotation, Remote Facilities and e-Publishing

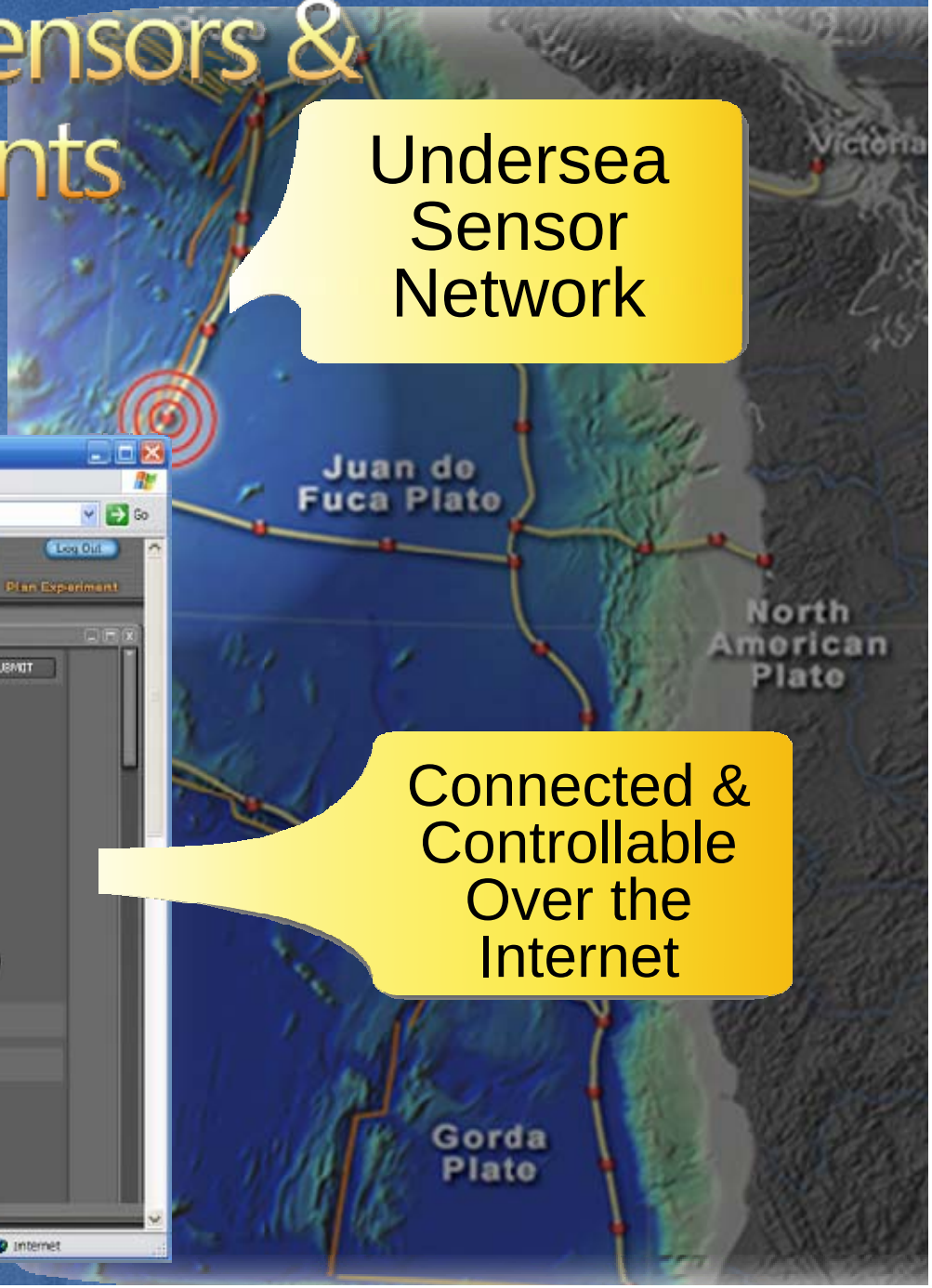
Vision For Scientific Workflow

Example: Project NEPTUNE

http://www.neptune.washington.edu/

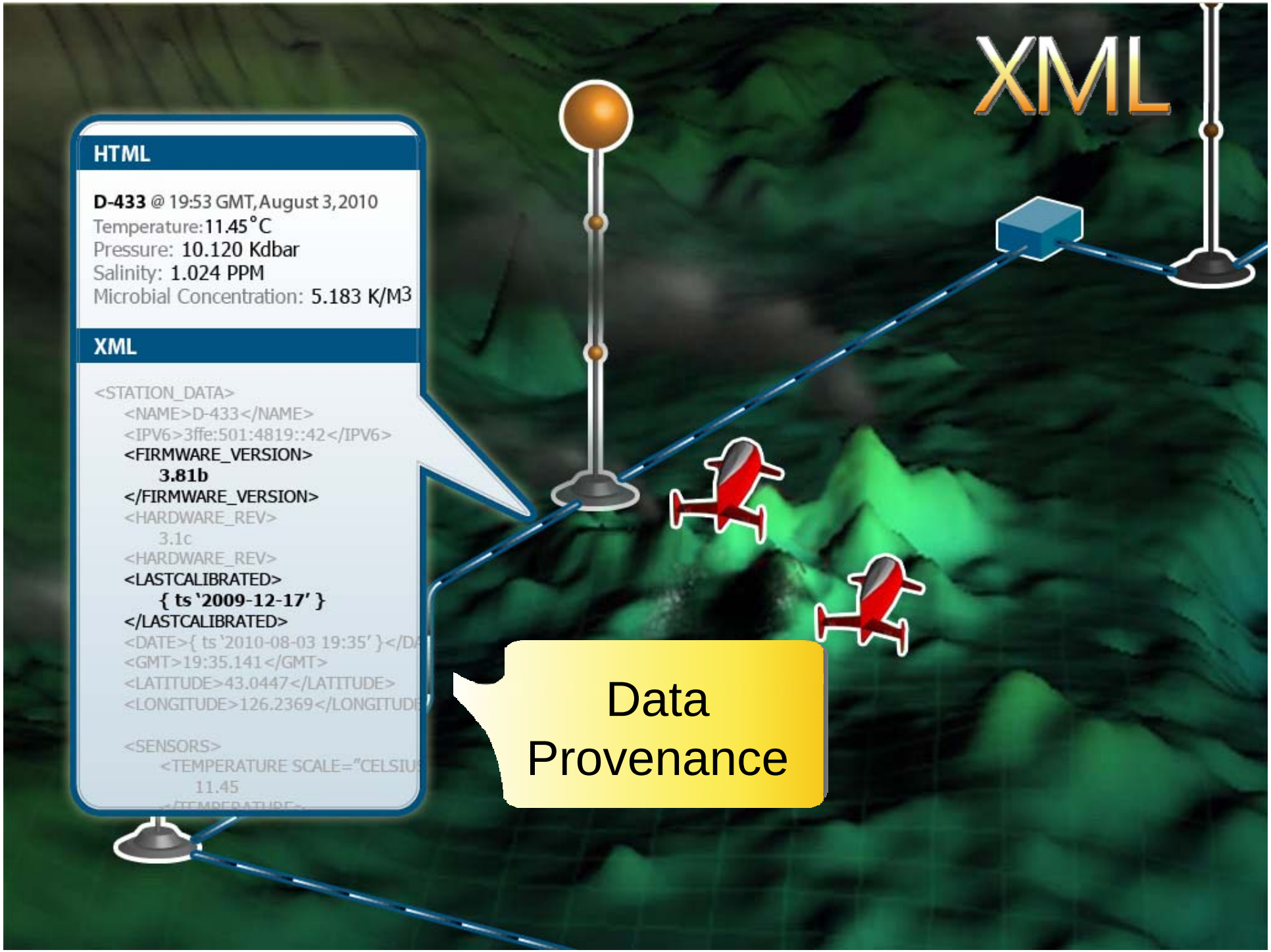
Programmable Sensors & Remote Instruments



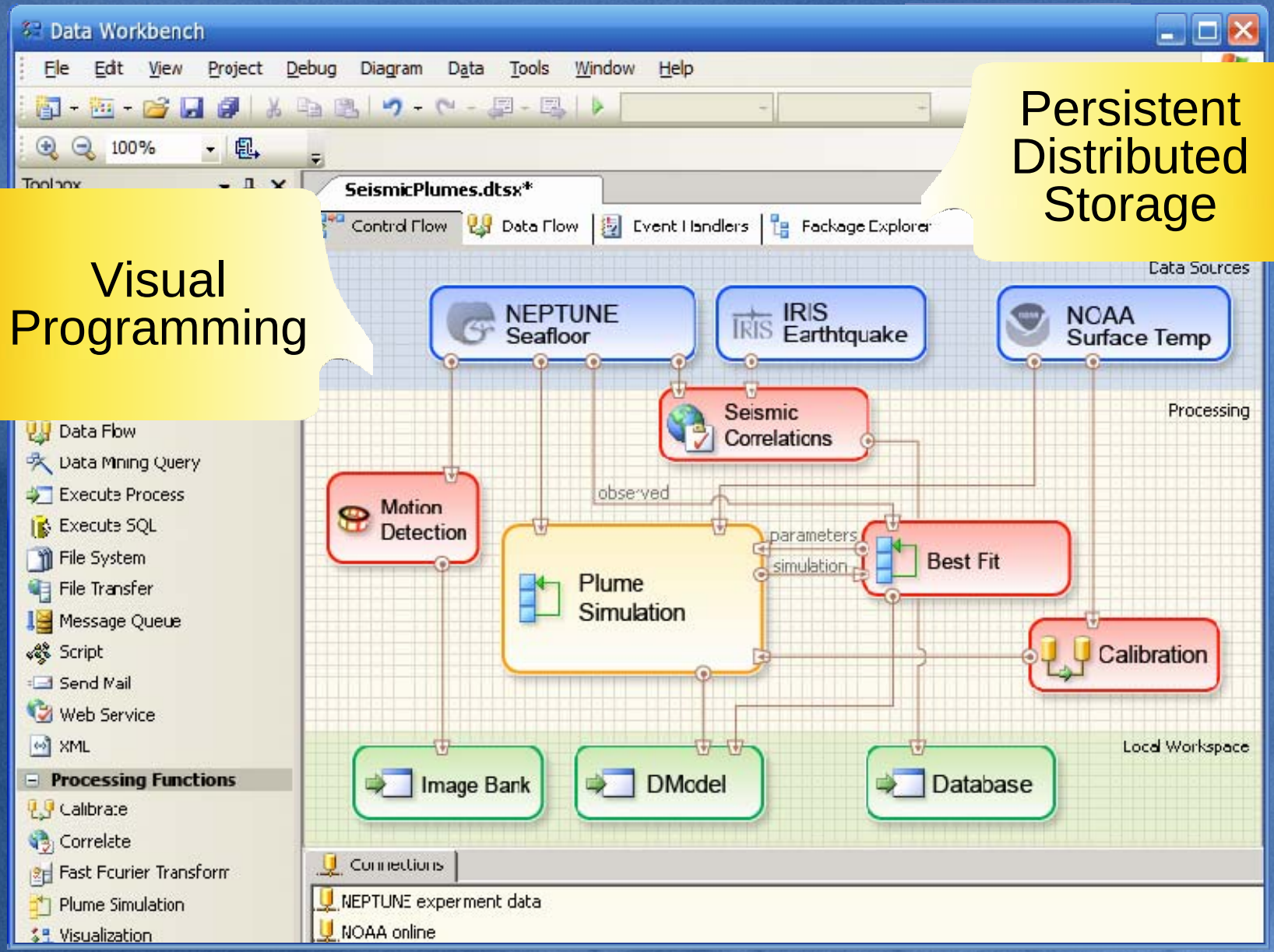


Undersea Sensor Network

Connected & Controllable Over the Internet


XML
HTML
D-433 @ 19:53 GMT, August 3, 2010
Temperature: 11.45°C
Pressure: 10.120 Kdbar
Salinity: 1.024 PPM
Microbial Concentration: 5.183 K/M3
XML
<STATION_DATA>
<NAME>D-433</NAME>
<IPV6>3ffe:501:4819::42</IPV6>
<FIRMWARE_VERSION>
3.81b
</FIRMWARE_VERSION>
<HARDWARE_REV>
3.1c
<HARDWARE_REV>
<LASTCALIBRATED>
{ ts '2009-12-17' }
</LASTCALIBRATED>
<DATE>{ ts '2010-08-03 19:35' }</DA
<GMT>19:35.141</GMT>
<LATITUDE>43.0447</LATITUDE>
<LONGITUDE>126.2369</LONGITUDE
<SENSORS>
<TEMPERATURE SCALE="CELSIU
11.45
Data
Provenance

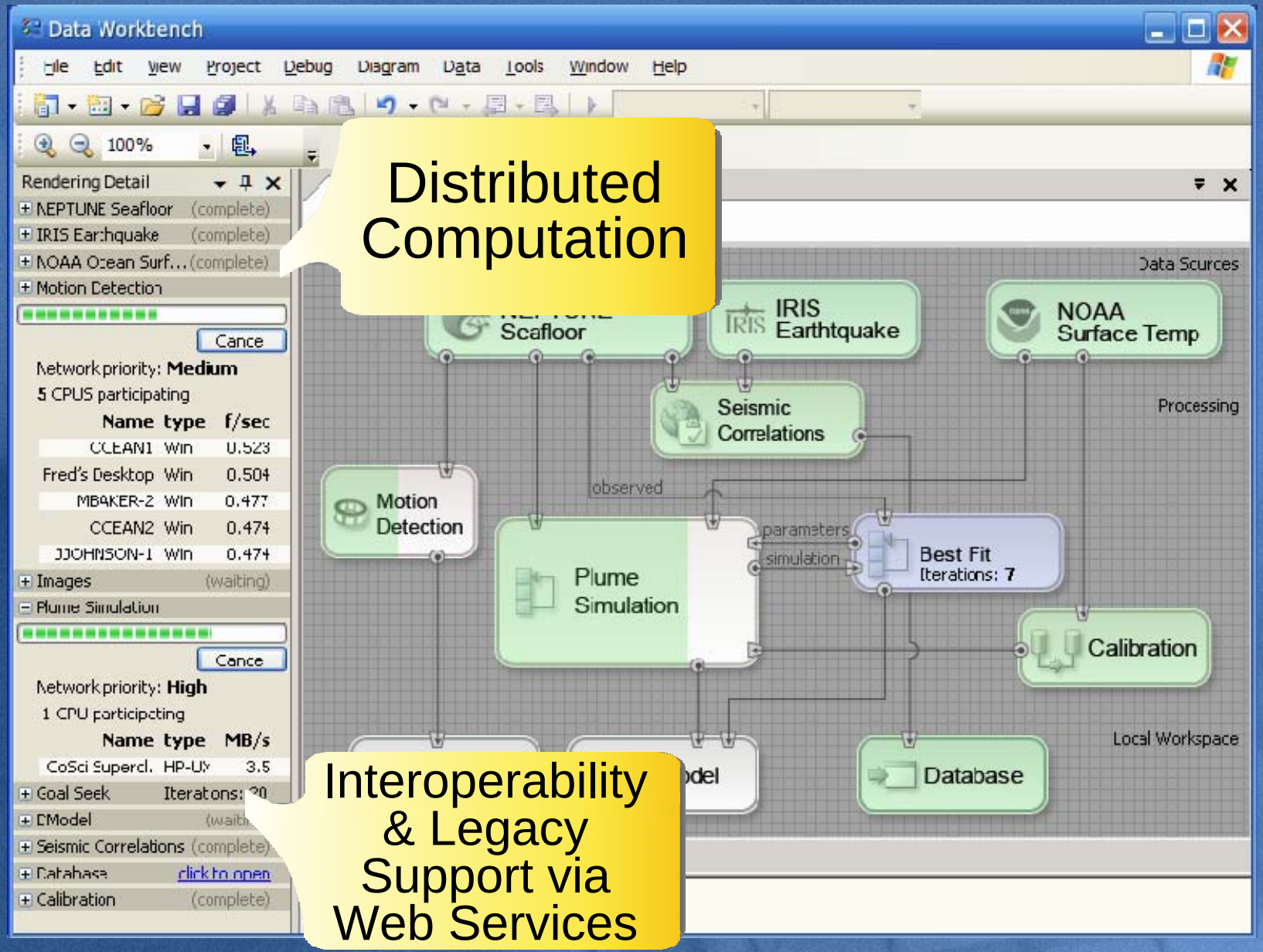
Data Workbench



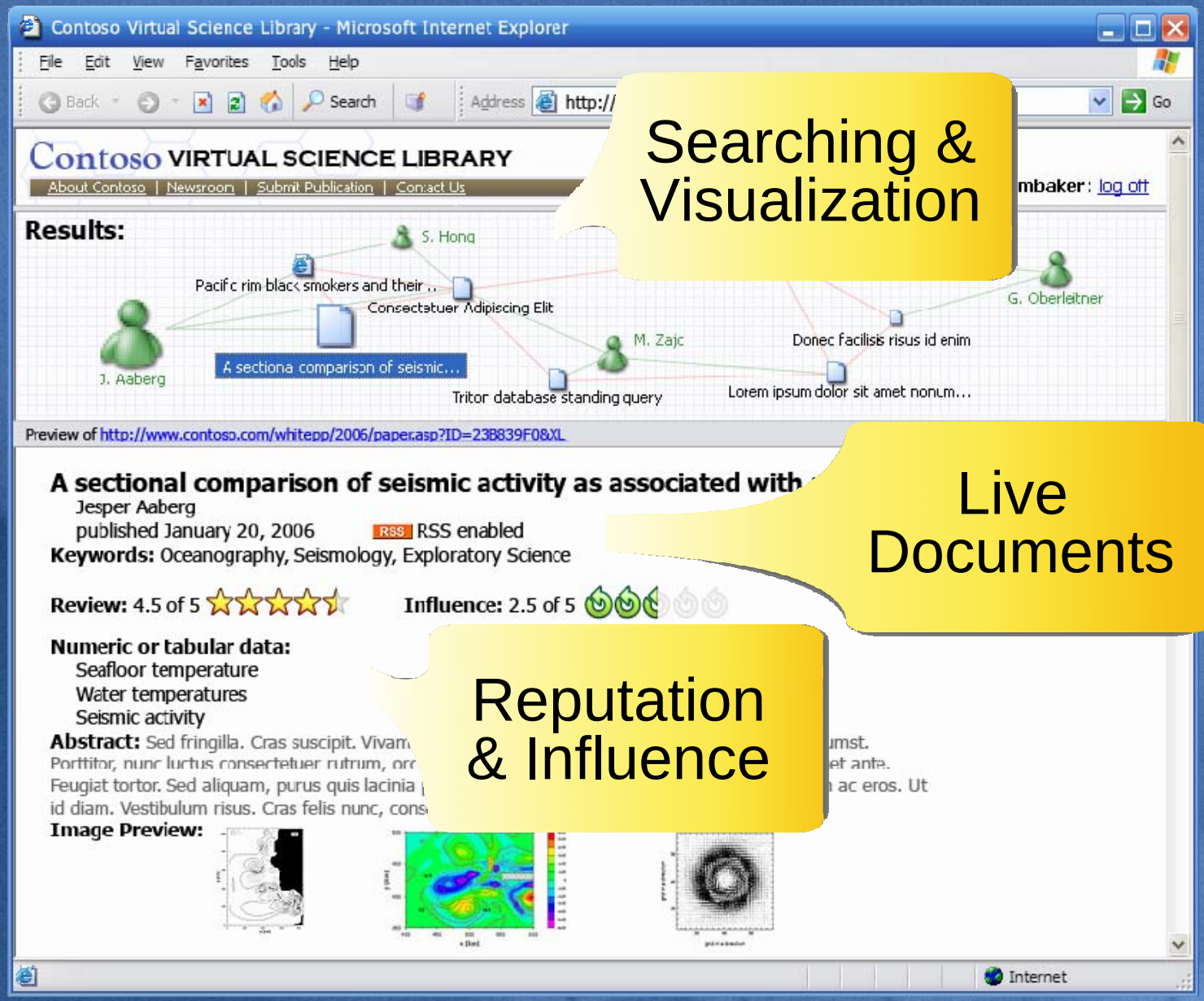
Visual Programming

Persistent Distributed Storage

Data Workbench



Research

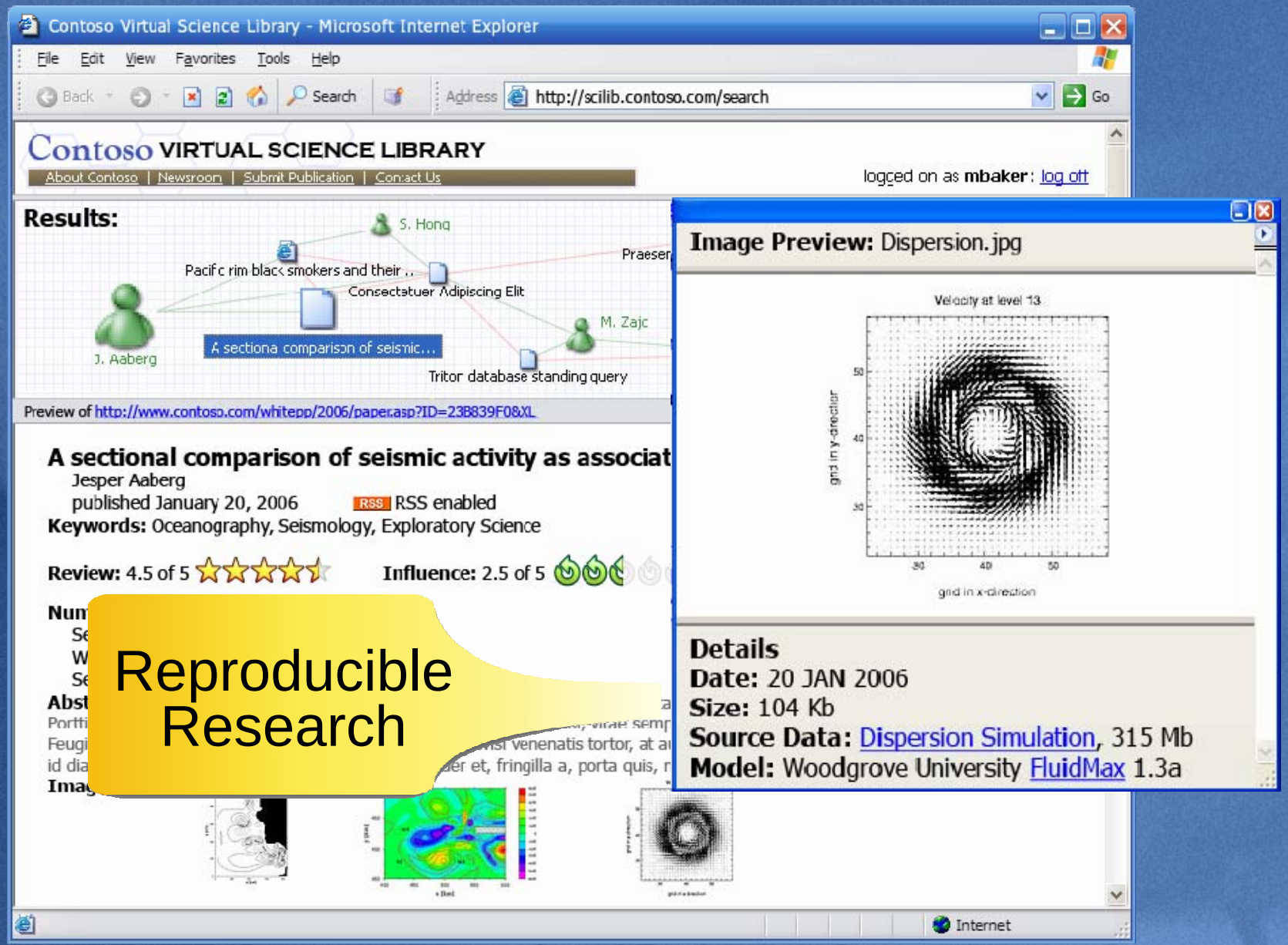


Searching & Visualization

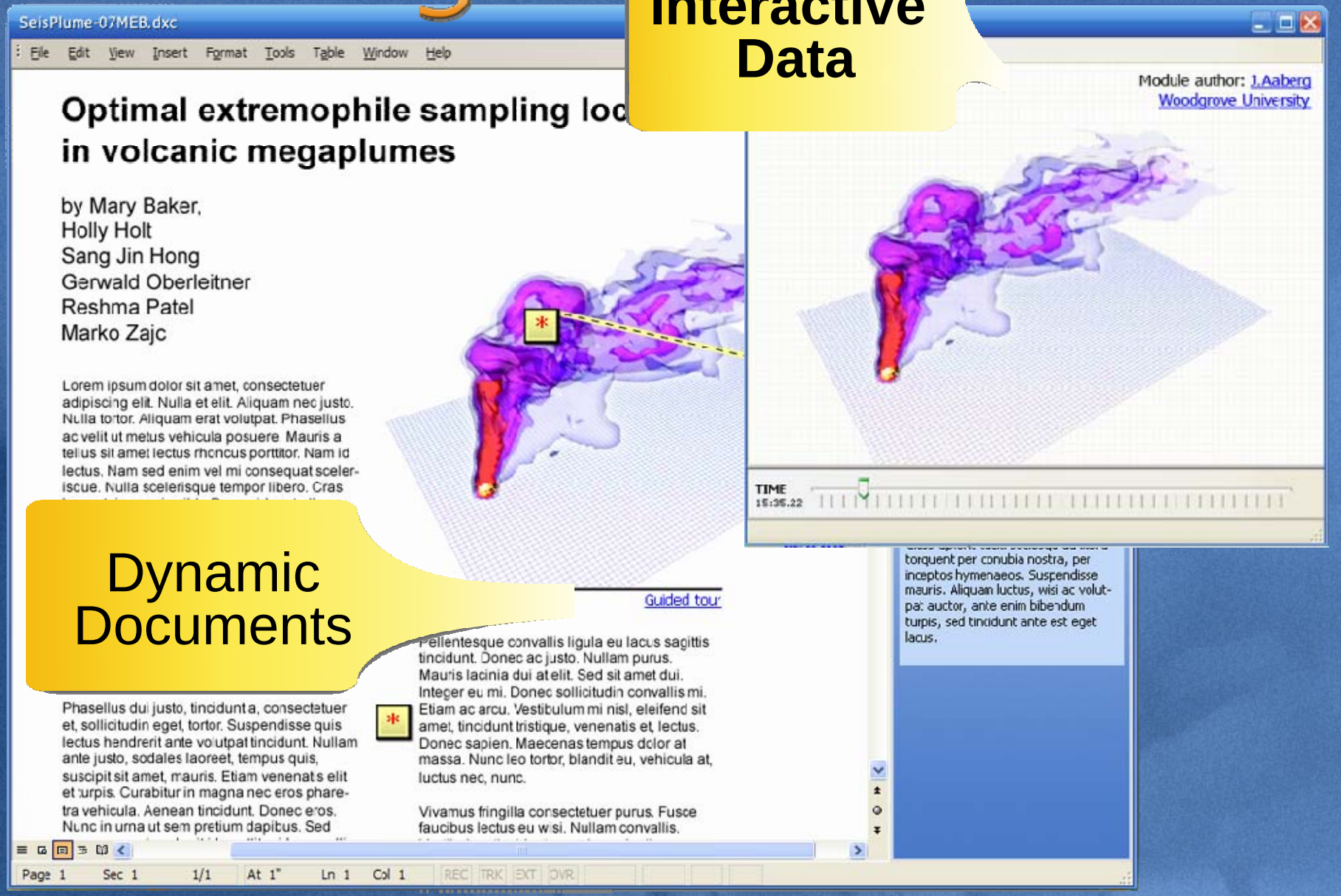
Live Documents

Reputation & Influence

Research


Contoso Virtual Science Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Address http://scilib.contoso.com/search Go
Contoso VIRTUAL SCIENCE LIBRARY
About Contoso | Newsroom | Submit Publication | Contact Us
logged on as mbaker: log off
Results:
S. Hong
Pacific rim black smokers and their ..
Consectetuer Adipiscing Elit
A sectiona comparison of seismic...
J. Aaberg
M. Zajc
Tritor database standing query
Preview of http://www.contoso.com/whitepp/2006/paper.asp?ID=23B839F0&XL
A sectional comparison of seismic activity as associat
Jesper Aaberg
published January 20, 2006
RSS RSS enabled
Keywords: Oceanography, Seismology, Exploratory Science
Review: 4.5 of 5
Influence: 2.5 of 5
Reproducible Research
Image Preview: Dispersion.jpg
Velocity at level 13
grid in y-direction
grid in x-direction
Details
Date: 20 JAN 2006
Size: 104 Kb
Source Data: Dispersion Simulation, 315 Mb
Model: Woodgrove University FluidMax 1.3a
Internet

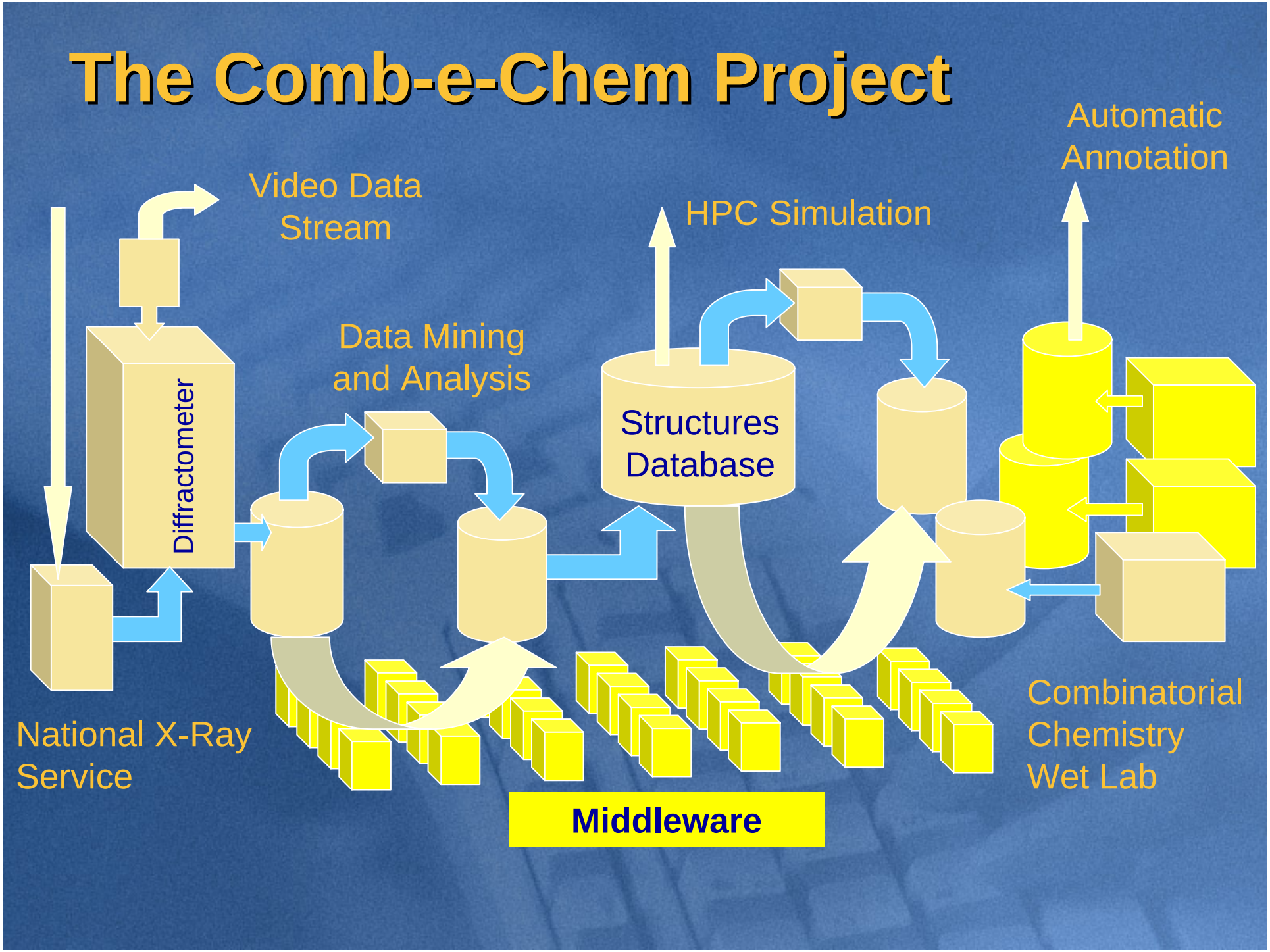
Publishing



Interactive Data

Dynamic Documents

The Comb-e-Chem Project



National Crystallographic Service

Send sample material to NCS service

Collaborate in e-Lab experiment and obtain structure

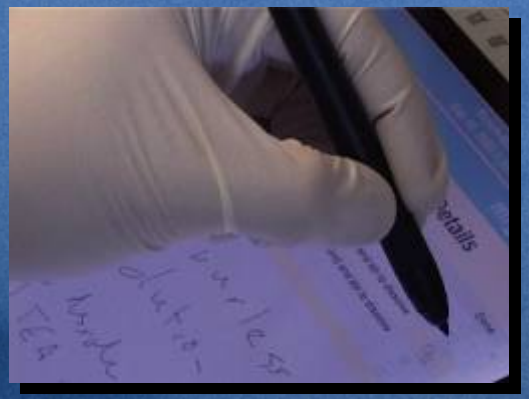
Search materials database and predict properties using Grid computations

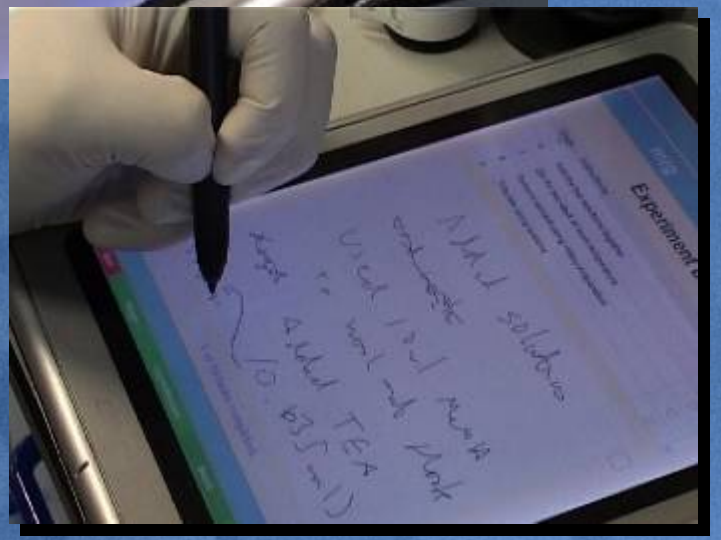
Download full data on materials of interest

X-Ray e-Laboratory

Structures Database

Computation Service

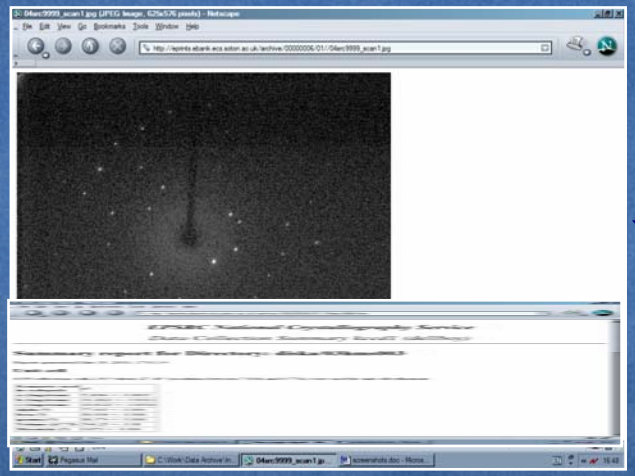




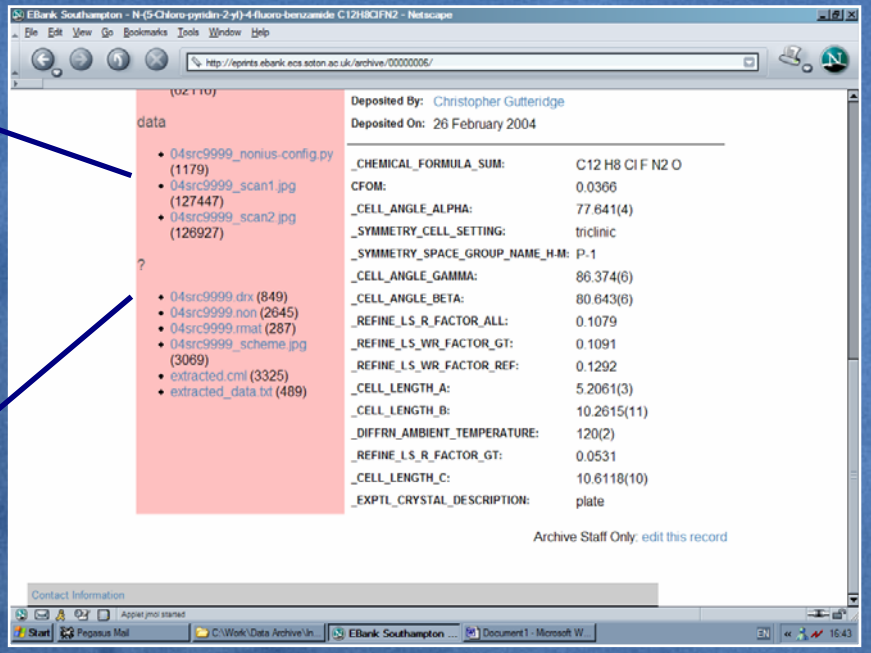
A digital lab book replacement that chemists were able to use, and liked

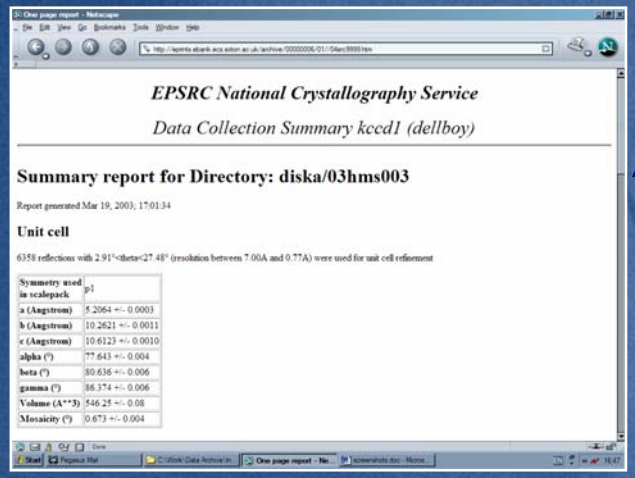
Monitoring laboratory experiments using a broker delivered over GPRS on a PDA

Crystallographic e-Prints



Direct Access to Raw Data from scientific papers





Raw data sets can be very large - stored at UK National Datastore using SRB software

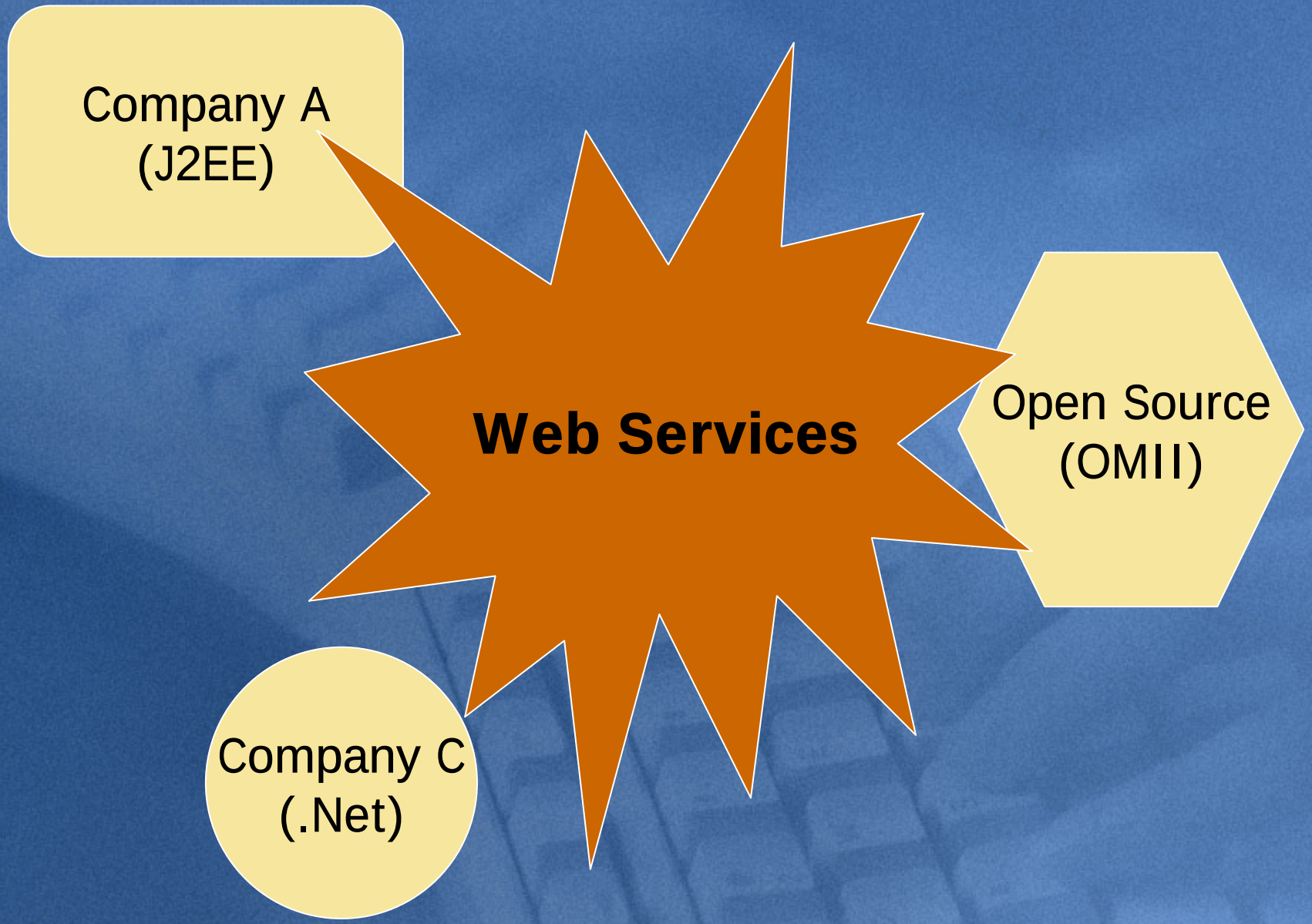
Support for e-Science

- Cyberinfrastructure and e-Infrastructure
  - In the US, Europe and Asia there is a common vision for the 'cyberinfrastructure' required to support the e-Science revolution
  - Set of Middleware Services supported on top of high bandwidth academic research networks
- Similar to vision of the Grid as a set of services that allows scientists – and industry – to routinely set up 'Virtual Organizations' for their research – or business
  - Many companies emphasize computing cycle aspect of Grids
  - The 'Microsoft Grid' vision is more about data management than about compute clusters

Six Key Elements for a Global Cyberinfrastructure for e-Science

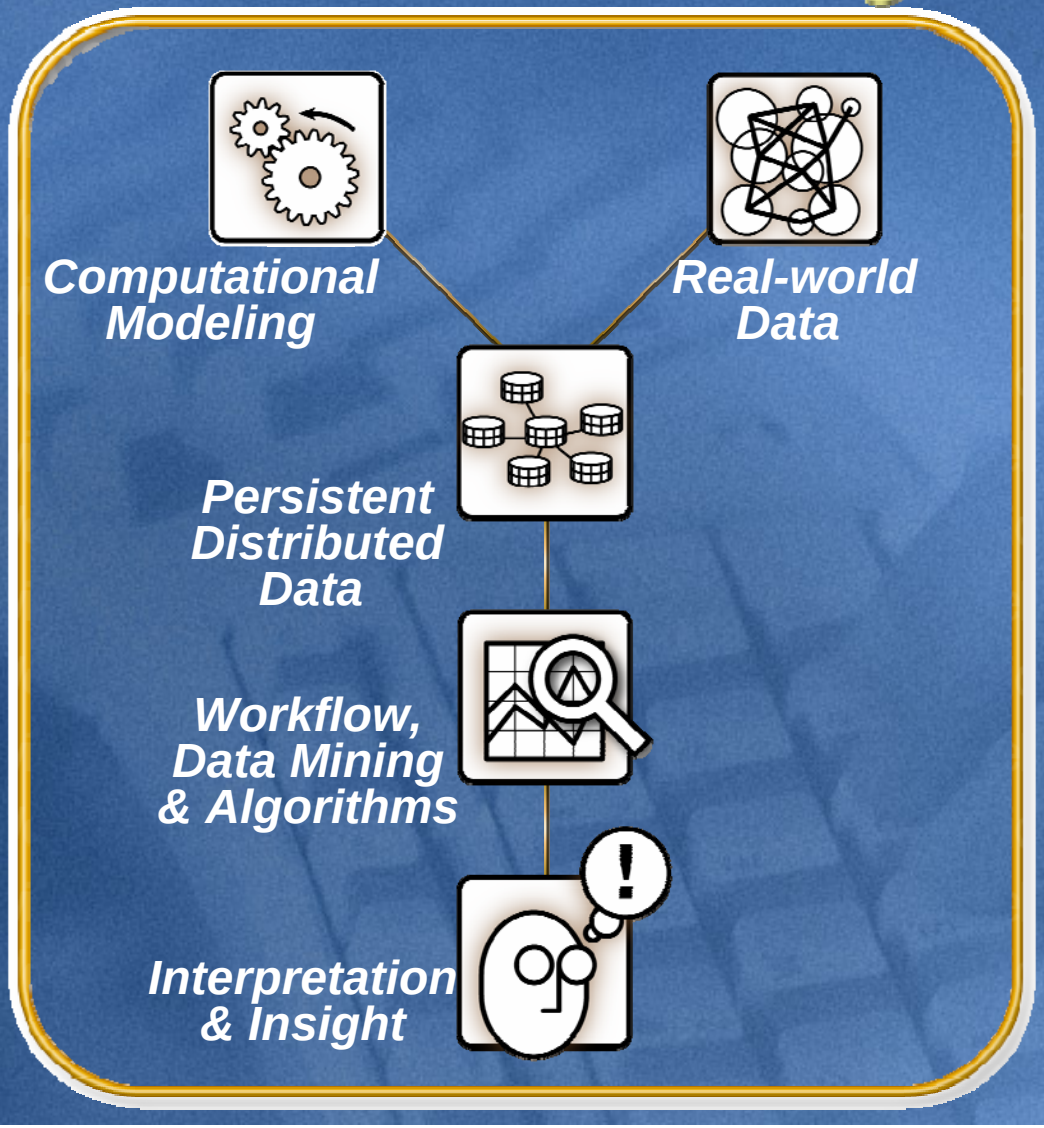
1. High bandwidth Research Networks
2. Internationally agreed AAA Infrastructure
3. Development Centers for Open Standard Grid Middleware
4. Technologies and standards for Data Provenance, Curation and Preservation
5. Open access to Data and Publications via Interoperable Repositories
6. Discovery Services and Collaborative Tools

The Web Services ‘Magic Bullet’



Technical Computing

Reduced Time To Insight



Technical Computing in Microsoft

- Radical Computing
  - Research in potential breakthrough technologies
- Advanced Computing for Science and Engineering
  - Application of new algorithms, tools and technologies to scientific and engineering problems
- High Performance Computing
  - Application of high performance clusters and database technologies to industrial applications

Radical Computing

- The end of Moore's Law as we know it
  - Number of transistors on a chip will continue to increase
  - No significant increase in Clock speed
- Remember Amdahl's Law
  - If application is 90% parallel, maximum speed-up that can be gained from parallelism is at most 10X
- Future of silicon chips
  - "100's of cores on a chip in 2015" (Justin Rattner, Intel)
  - "4 cores"/Tflop => 25 Tflops/chip

Radical Computing (continued)

- IT industry has been driven by increasing chip volumes and new applications
  - Multi-core chips for servers
  - Multi-core chips for clients?
- Challenge not only for Microsoft but for entire IT industry
  - New paradigms to exploit parallelism
  - What applications can exploit such on-chip parallelism?

Advanced Computing for Science and Engineering

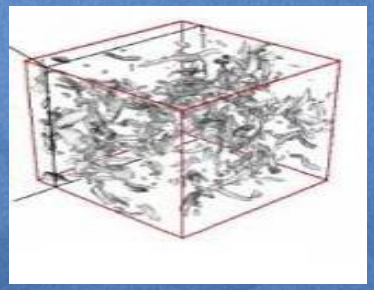
TOOLS	Workflow, Collaboration, Visualization, Data Mining
DATA	Acquisition, Storage, Annotation, Provenance, Curation, Preservation
CONTENT	Scholarly Communication, Institutional Repositories

New Science Paradigms

- **Thousand years ago:**
  Experimental Science
  - description of natural phenomena
- **Last few hundred years:**
  Theoretical Science
  - Newton's Laws, Maxwell's Equations …

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

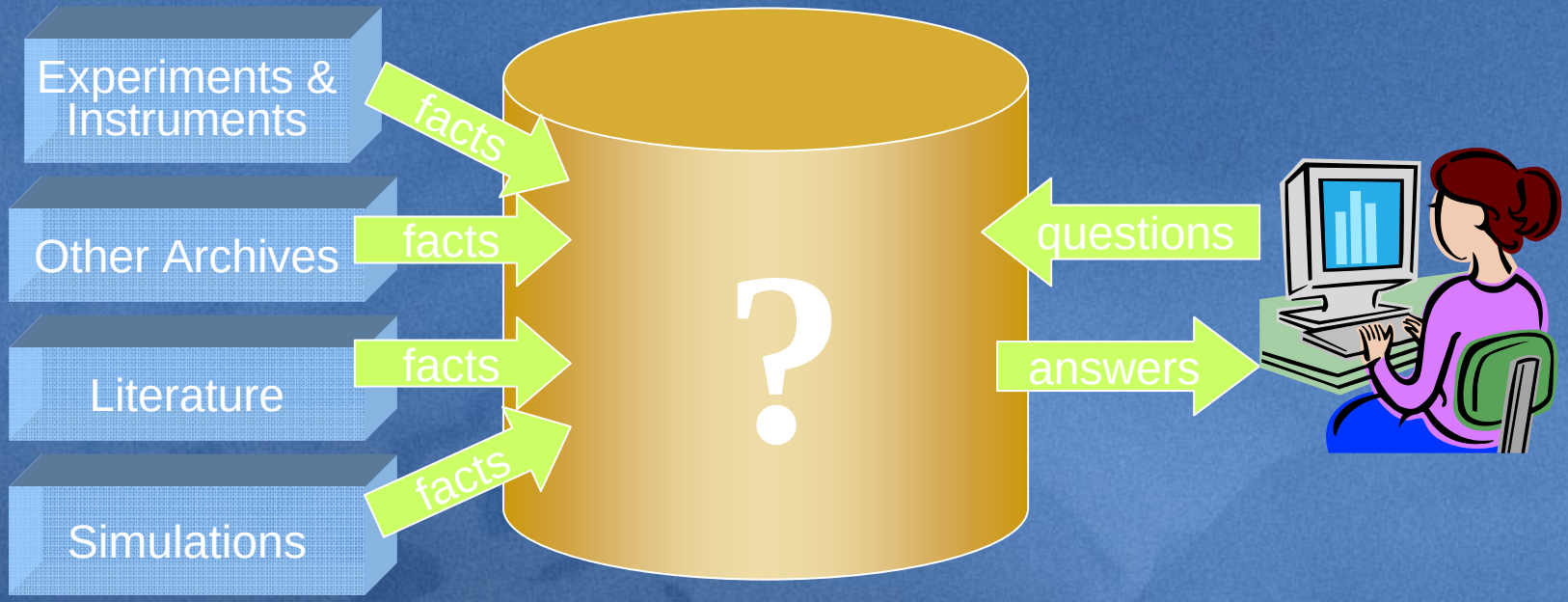
- **Last few decades**:
  Computational Science
  - simulation of complex phenomena



- **Today:**
  e-Science or Data-centric Science
  - unify theory, experiment, and simulation
  - using data exploration and data mining
  - **Data captured by instruments**
  - **Data generated by simulations**
  - **Processed by software**
  - **Scientist analyzes databases/files**

(With thanks to Jim Gray)

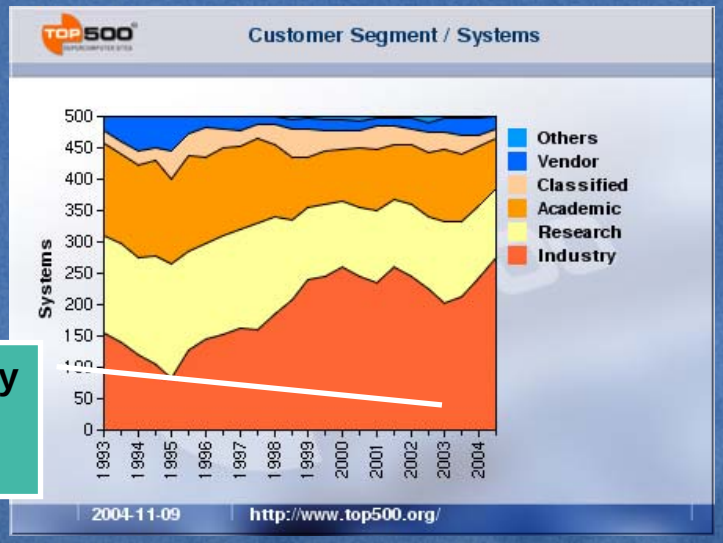
The Problem for the e-Scientist



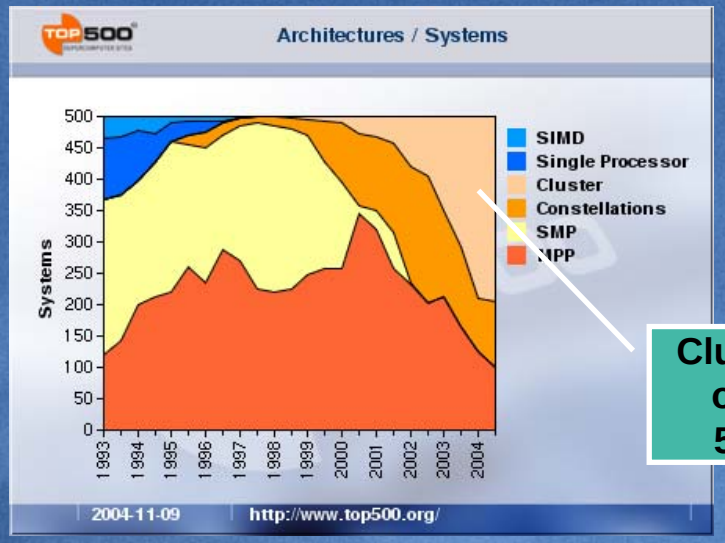
- Data ingest
- Managing a petabyte
- Common schema
- How to organize it?
- How to *re*organize it?
- How to coexist & cooperate with others?

- Data Query and Visualization tools
- Support/training
- Performance
  - Execute queries in a minute
  - Batch (big) query scheduling

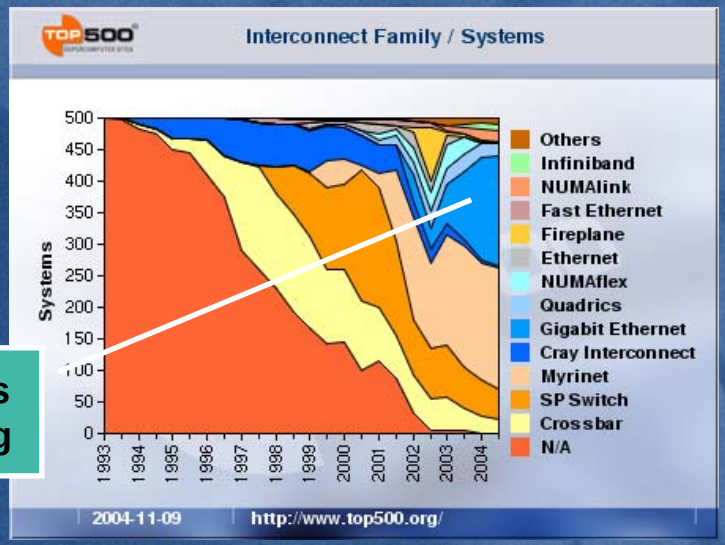
Top 500 Supercomputer Trends



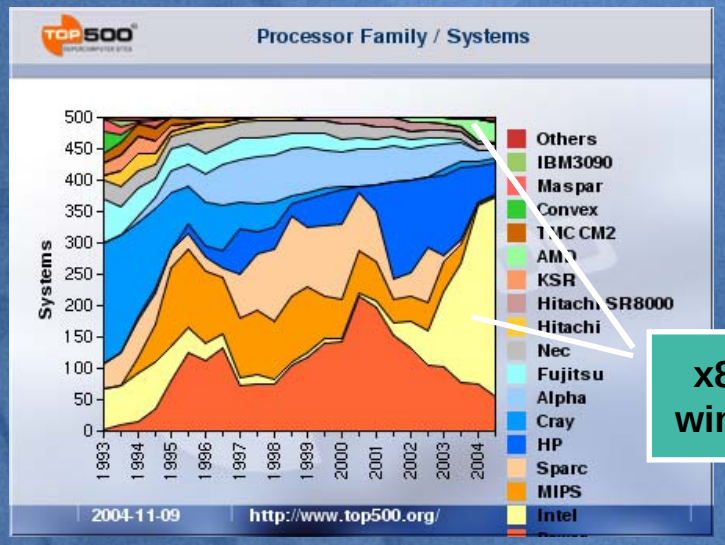
Industry usage rising



Clusters over 50%



GigE is gaining

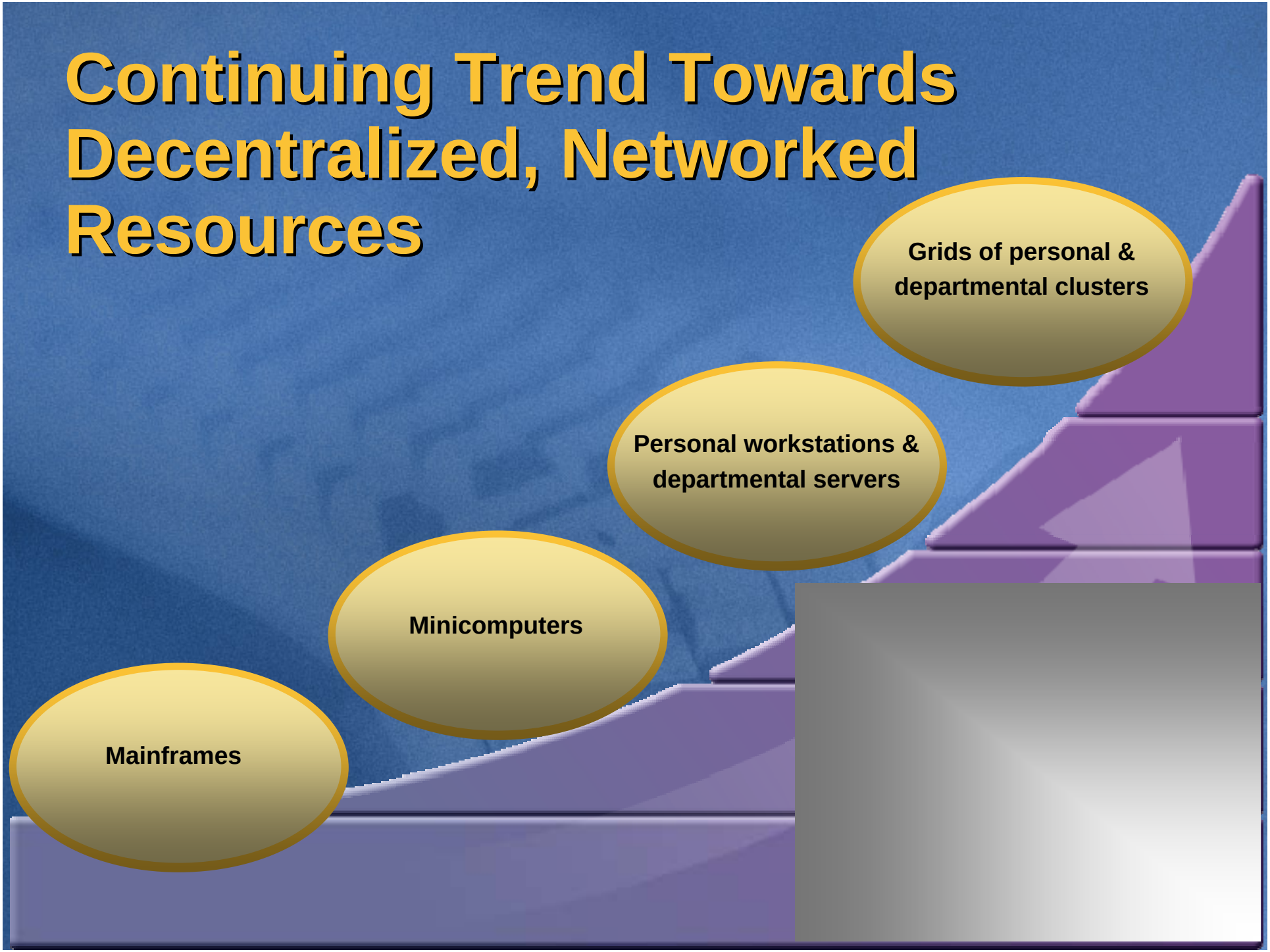


x86 is winning

Supercomputing Goes Personal

	1991	1998	2005
System	Cray Y-MP C916 	Sun HPC10000 	Shuttle @ NewEgg.com 
Architecture	16 x Vector 4GB, Bus	24 x 333MHz Ultra-SPARCII, 24GB, SBus	4 x 2.2GHz x64 4GB, GigE
OS	UNICOS	Solaris 2.5.1	Windows Server 2003 SP1
GFlops	~10	~10	~10
Top500 #	1	500	N/A
Price	$40,000,000	$1,000,000 (40x drop)	< $4,000 (250x drop)
Customers	Government Labs	Large Enterprises	Every Engineer & Scientist
Applications	Classified, Climate, Physics Research	Manufacturing, Energy, Finance, Telecom	Bioinformatics, Materials Sciences, Digital Media

Continuing Trend Towards Decentralized, Networked Resources



Berlin Declaration 2003

- ‘To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection’
- Defines open access contributions as including:
  - ‘original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material’

NSF ‘Atkins’ Report on Cyberinfrastructure

- ‘the primary access to the latest findings in a growing number of fields is through the Web, then through classic preprints and conferences, and lastly through refereed archival papers’
- ‘archives containing hundreds or thousands of terabytes of data will be affordable and necessary for archiving scientific and engineering information’

Microsoft Strategy for e-Science

Microsoft intends to work with both the scientific and library communities:

- to define open standard and/or interoperable high-level services, work flows and tools
- to assist the community in developing open scholarly communication and interoperable repositories

Acknowledgements

With special thanks to Geoffrey Fox, Jeremy Frey, Brad Gillespie, Jim Gray and Marvin Theimer