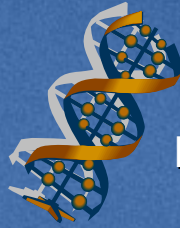




**Earth
Sciences**



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

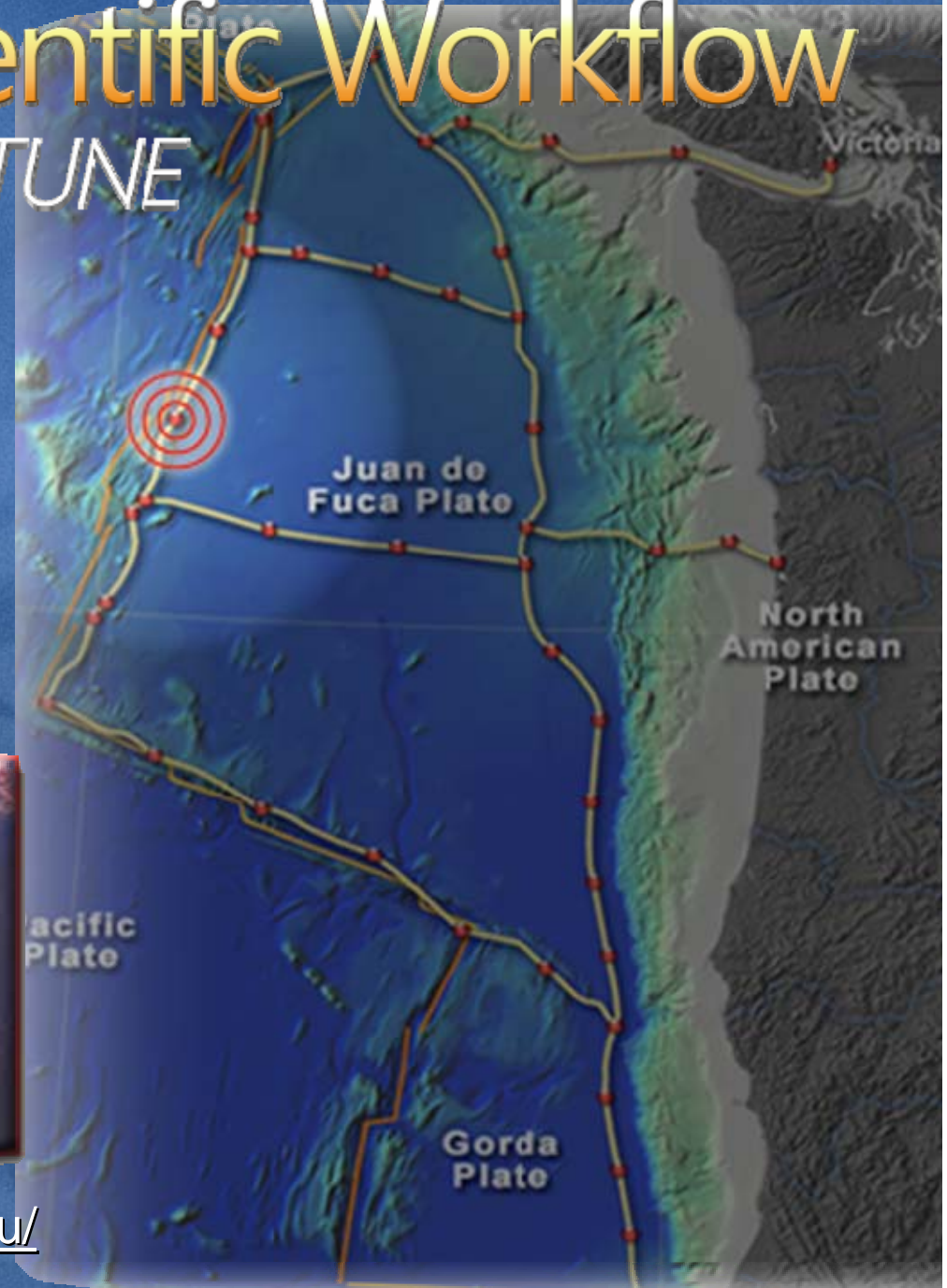
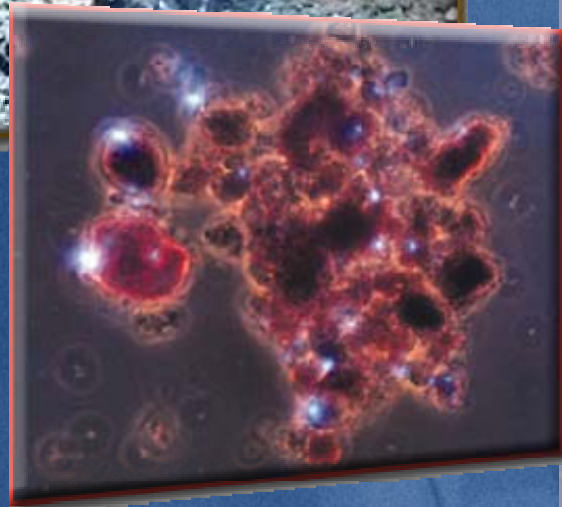
- ◆ Oceanographic sensors - Project Neptune
 - Joint US-Canadian proposal

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

The screenshot shows the NEPTUNE web interface in a Microsoft Internet Explorer browser window. The address bar displays <http://www.neptune.washington.edu/>. The page features a navigation menu with options for Scientists, Teachers & Students, and General Public. Below the menu is a main content area with an 'Interactive Map' on the left and a 'Node Sensors' panel on the right. The map shows a network of red nodes connected by lines, with labels for 'Axial', 'Juan de Fuca Plate', and 'Gorda Plate'. The 'Node Sensors' panel lists various sensors for three nodes: Node D-433, Node D-436, and Node D-437. Each node's sensors are listed with checkboxes, some of which are checked.

Node	Sensor	Status
Node D-433	Thermal (floor, always on)	<input checked="" type="checkbox"/>
	Thermal (10m)	<input checked="" type="checkbox"/>
	Thermal (50m)	<input checked="" type="checkbox"/>
	Seismometer (always on)	<input checked="" type="checkbox"/>
	Salinity	<input checked="" type="checkbox"/>
	Current field vector (offline)	<input type="checkbox"/>
	Microbial concentration	<input checked="" type="checkbox"/>
	Oxygen	<input checked="" type="checkbox"/>
	Doppler current profiler	<input type="checkbox"/>
	Microbial concentration	<input checked="" type="checkbox"/>
Node D-436	Video	<input checked="" type="checkbox"/>
	Hydrophone	<input checked="" type="checkbox"/>
	Sample floats (20 remaining)	<input checked="" type="checkbox"/>
	AUV	<input checked="" type="checkbox"/>
Node D-437	Thermal (floor, always on)	<input checked="" type="checkbox"/>
	Thermal (10m)	<input checked="" type="checkbox"/>
	Thermal (50m)	<input checked="" type="checkbox"/>
	Seismometer (always on)	<input checked="" type="checkbox"/>
	Salinity	<input checked="" type="checkbox"/>

Connected & Controllable Over the Internet

XML

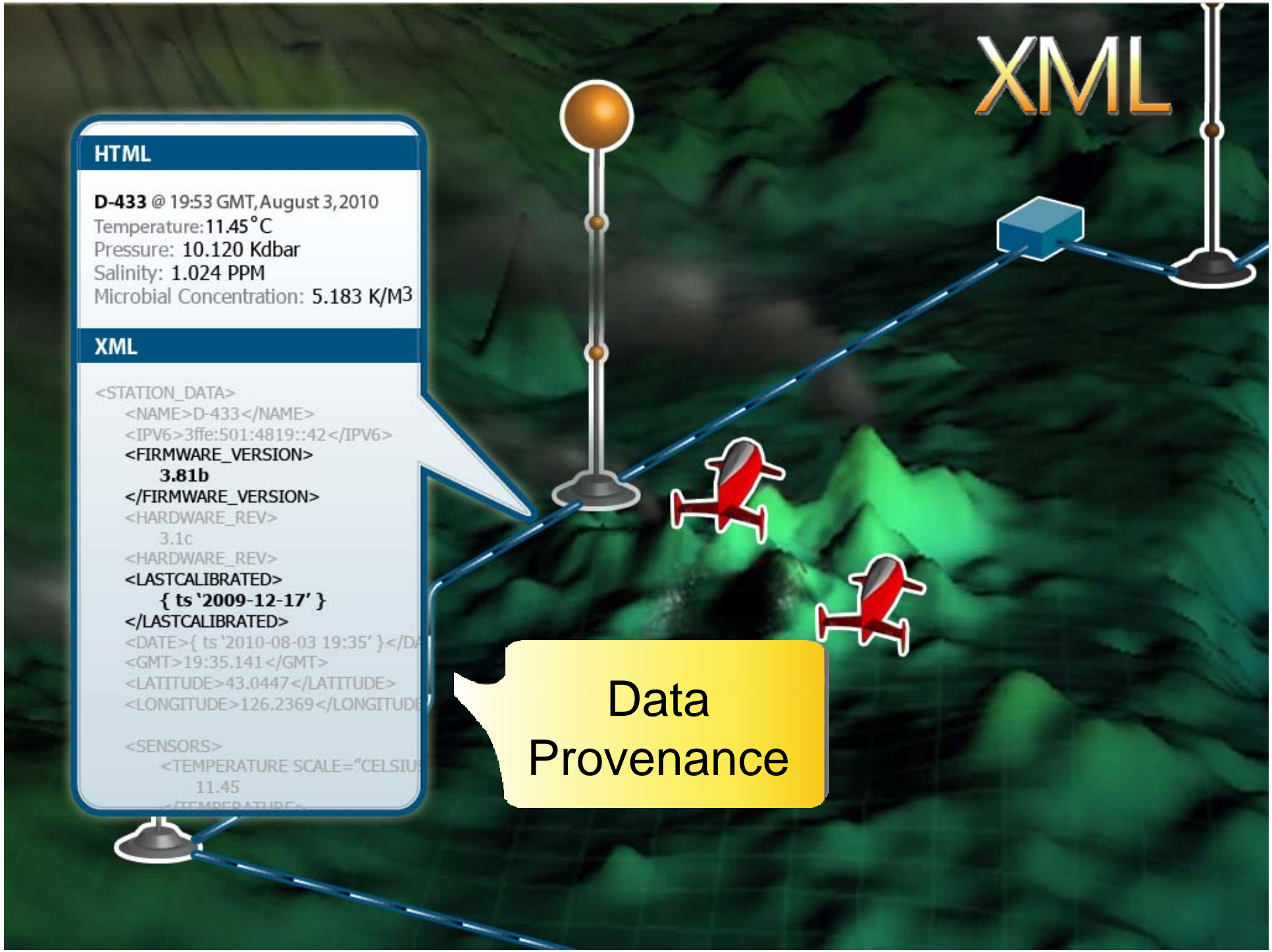
HTML

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

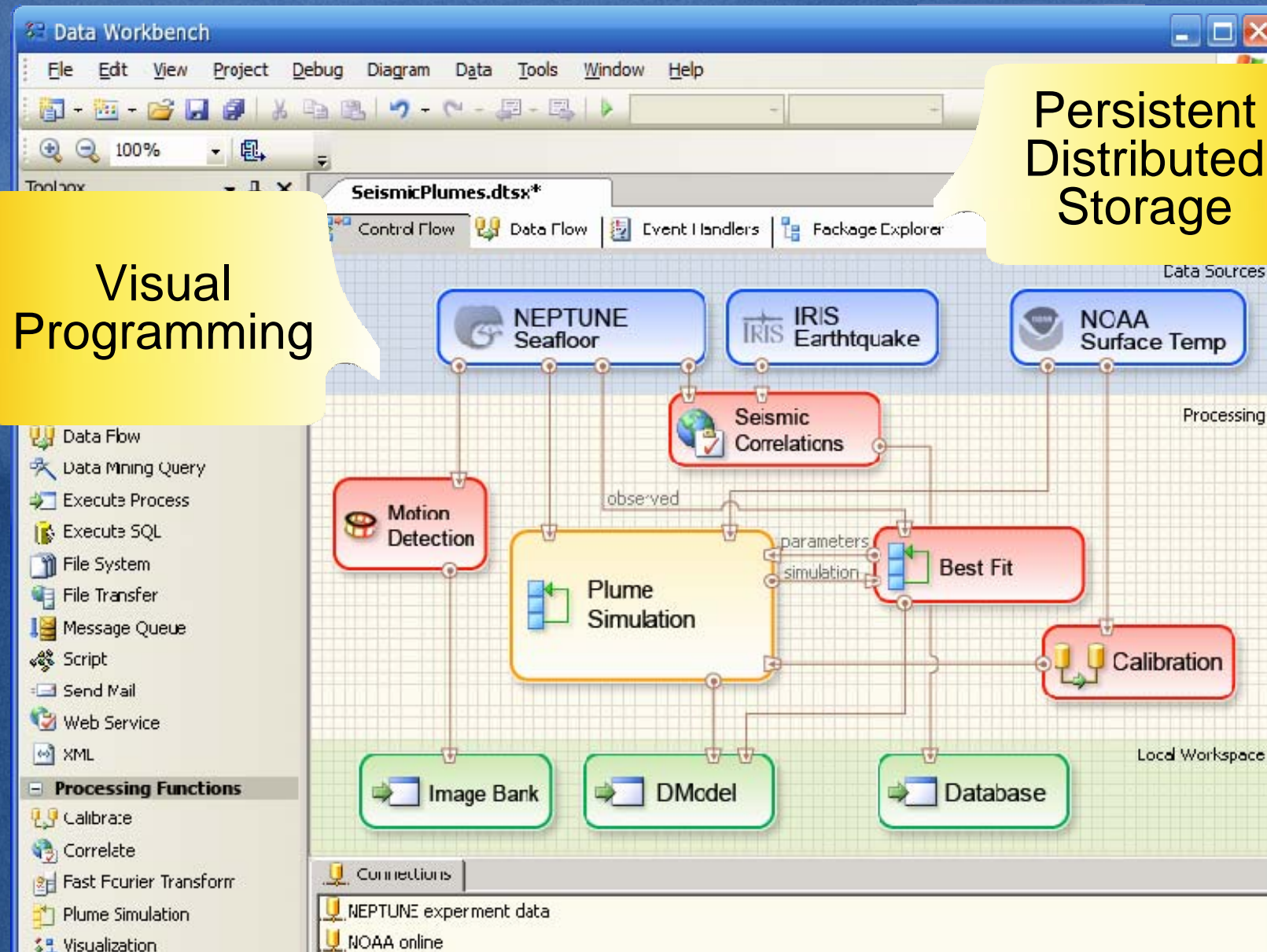
XML

```
<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' }</DATE>  
<GMT>19:35.141</GMT>  
<LATITUDE>43.0447</LATITUDE>  
<LONGITUDE>126.2369</LONGITUDE>  
</STATION_DATA>  
  
<SENSORS>  
<TEMPERATURE SCALE="CELSIUS">  
  11.45  
</TEMPERATURE>  
</SENSORS>
```

Data Provenance



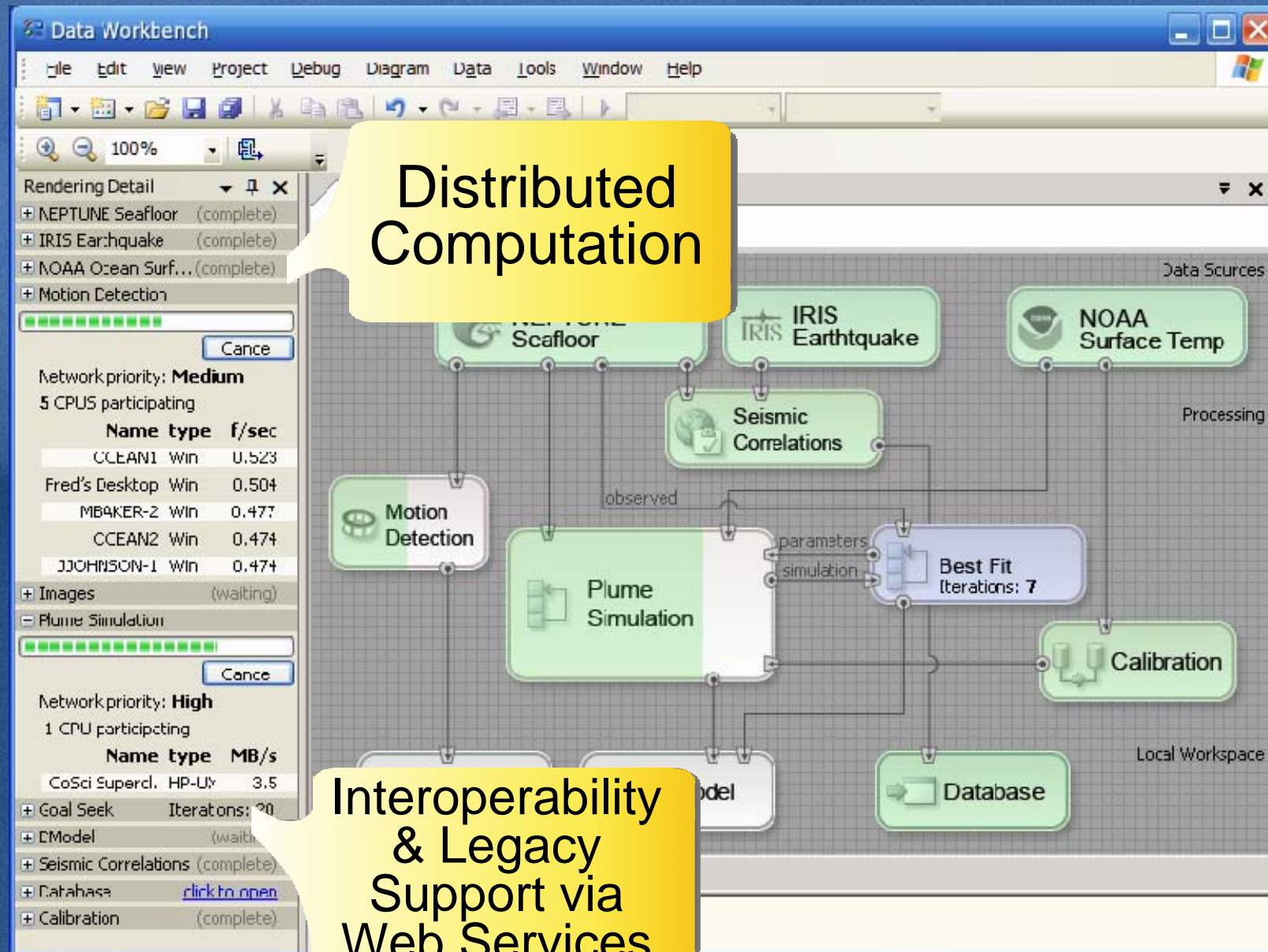
Data Workbench



Visual Programming

Persistent Distributed Storage

Data Workbench



Distributed Computation

Interoperability & Legacy Support via Web Services

Research

The screenshot shows a Microsoft Internet Explorer window titled "Contoso Virtual Science Library". The address bar shows a URL starting with "http://". The page content includes a navigation menu with "About Contoso", "Newsroom", "Submit Publication", and "Contact Us". A "Results:" section displays a network graph with nodes representing authors (J. Aaberg, S. Hong, M. Zajc, G. Oberleitner) and documents. A yellow callout bubble points to the graph with the text "Searching & Visualization". Below the graph, a preview of a paper titled "A sectional comparison of seismic activity as associated with" by Jesper Aaberg is shown. A yellow callout bubble points to the paper title with the text "Live Documents". The paper's metadata includes a review score of 4.5 of 5 stars and an influence score of 2.5 of 5. A yellow callout bubble points to these scores with the text "Reputation & Influence". The abstract and image preview are also visible at the bottom of the page.

Searching & Visualization

Live Documents

Reputation & Influence

Contoso VIRTUAL SCIENCE LIBRARY
About Contoso | Newsroom | Submit Publication | Contact Us

Results:

Pacific rim black smokers and their ...
Consectatur Adipiscing Elit
Donec facilis risus id enim
Tritor database standing query
Lorem ipsum dolor sit amet nonum...

J. Aaberg
S. Hong
M. Zajc
G. Oberleitner

Preview of <http://www.contoso.com/whitepp/2006/paper.asp?ID=238839F0&XL>

A sectional comparison of seismic activity as associated with
Jesper Aaberg
published January 20, 2006 RSS enabled
Keywords: Oceanography, Seismology, Exploratory Science

Review: 4.5 of 5 **Influence:** 2.5 of 5

Numeric or tabular data:
Seafloor temperature
Water temperatures
Seismic activity

Abstract: Sed fringilla. Cras suscipit. Vivam...
Porttitor, nunc luctus consectetur rutrum, or...
Feugiat tortor. Sed aliquam, purus quis lacinia...
id diam. Vestibulum risus. Cras felis nunc, cons...

Image Preview:

Internet

Research

Contoso Virtual Science Library - Microsoft Internet Explorer

File Edit View Favorites Tools Help

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:

Preview of <http://www.contoso.com/whitepp/2006/paper.asp?ID=238839F0&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 🌱🌱🌱🌱🌱

Details
Date: 20 JAN 2006
Size: 104 Kb
Source Data: [Dispersion Simulation](#), 315 Mb
Model: Woodgrove University [FluidMax](#) 1.3a

Image Preview: Dispersion.jpg

Velocity at level 13

grid in y-direction

grid in x-direction

Internet

Reproducible
Research

Publishing

Interactive
Data

SeisPlume-07MEB.dxc

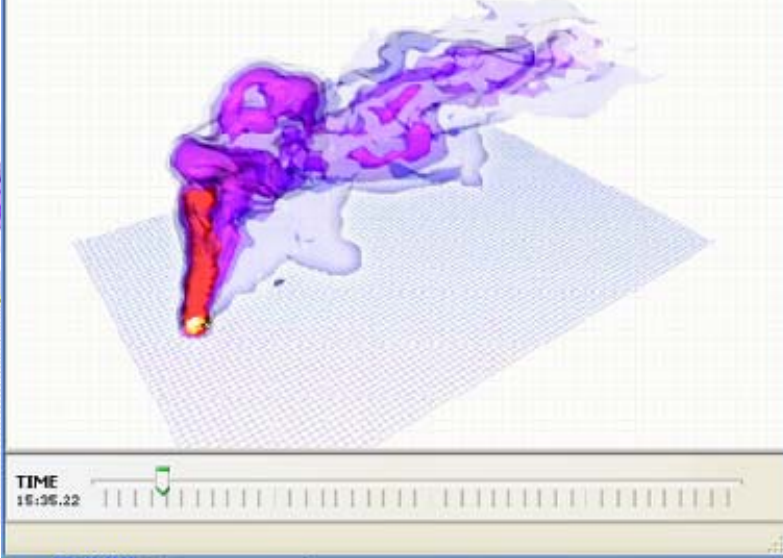
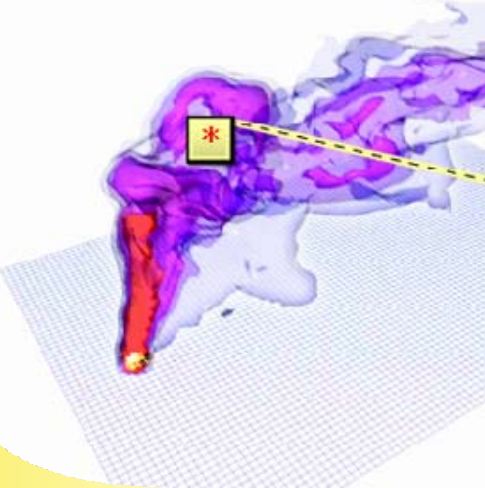
File Edit View Insert Format Tools Table Window Help

Optimal extremophile sampling locations in volcanic megaplumes

by Mary Baker,
Holly Holt
Sang Jin Hong
Gerwald Oberleitner
Reshma Patel
Marko Zajc

Module author: [J. Aaberg](#)
[Woodgrove University](#)

>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla et elit. Aliquam nec justo. Nulla tortor. Aliquam erat volutpat. Phasellus ac velit ut metus vehicula posuere. Mauris a tellus sit amet lectus rhoncus porttitor. Nam id lectus. Nam sed enim vel mi consequat scelerisque. Nulla scelerisque tempor libero. Cras



TIME 15:35:22

[Guided tour](#)

Phellentesque convallis ligula eu lacus sagittis tincidunt. Donec ac justo. Nullam purus. Mauris lacinia dui at elit. Sed sit amet dui. Integer eu mi. Donec sollicitudin convallis mi. Etiam ac arcu. Vestibulum mi nisi, eleifend sit amet, tincidunt tristique, venenatis et lectus. Donec sapien. Maecenas tempus dclor at massa. Nunc leo tortor, blandit eu, vehicula at, luctus nec, nunc.

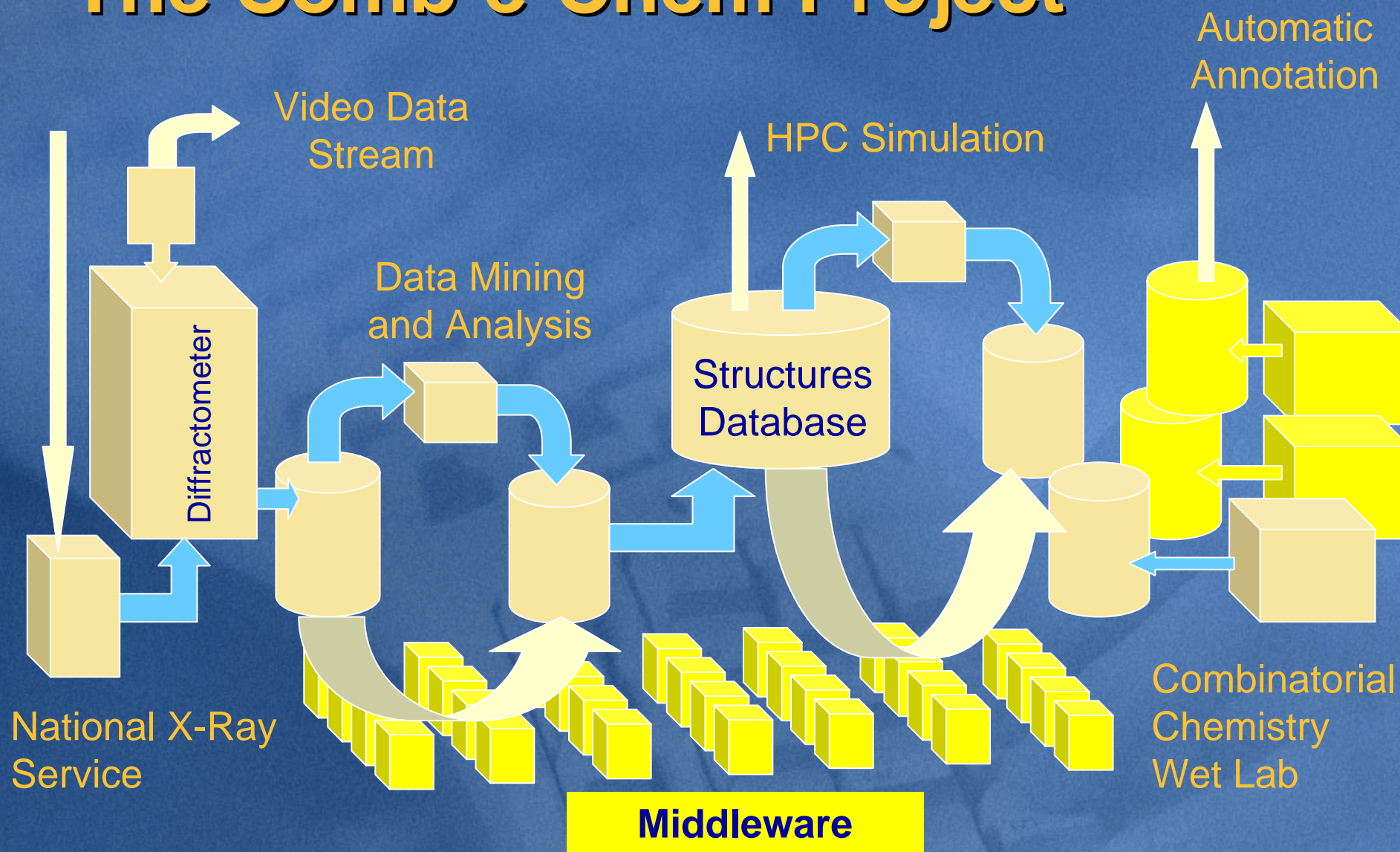
Vivamus fringilla consectetur purus. Fusce faucibus lectus eu wisi. Nullam convallis.

torquent per conubia nostra, per inceptos hymenaeos. Suspendisse mauris. Aliquam luctus, wisi ac volutpat: auctor, ante enim bibendum turpis, sed tincidunt ante est eget lacus.

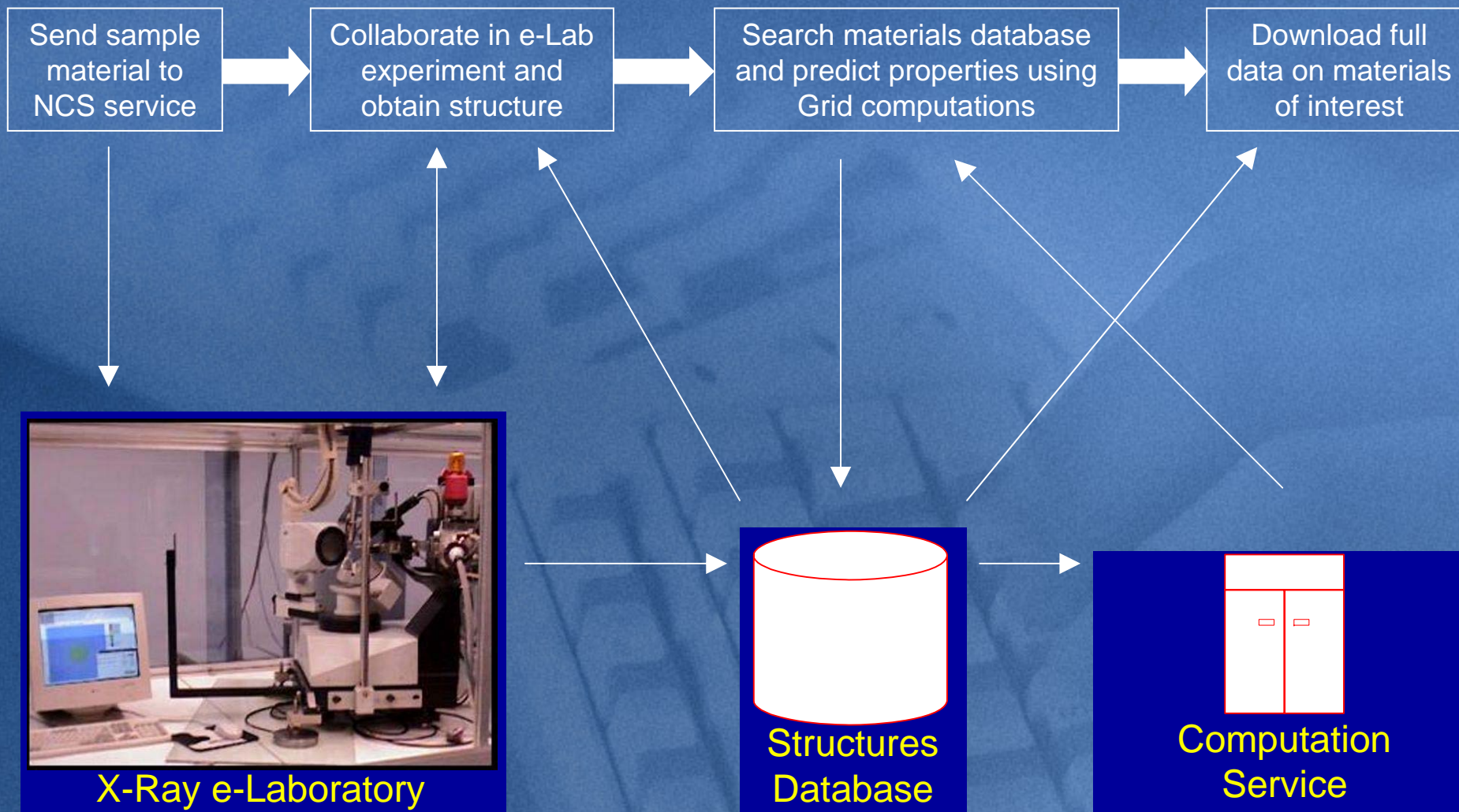
Page 1 Sec 1 1/1 At 1" Ln 1 Col 1 REC TRK EXT DVR

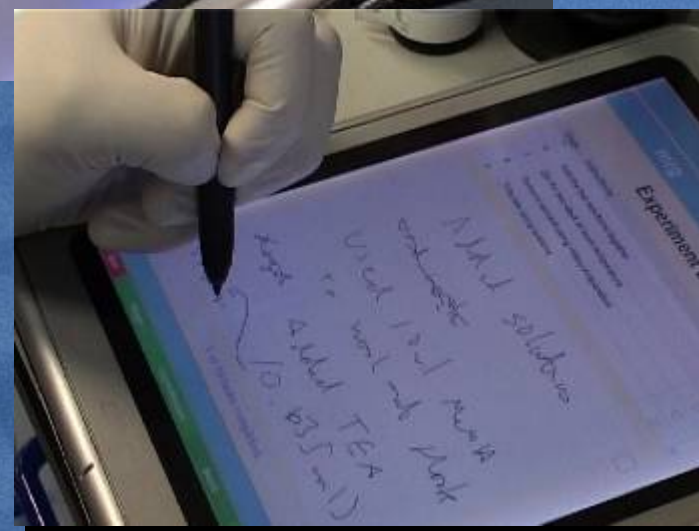
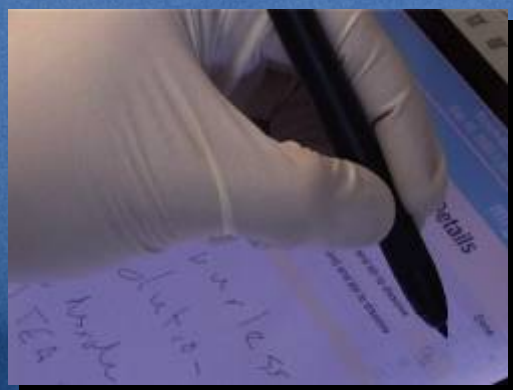
Dynamic
Documents

The Comb-e-Chem Project



National Crystallographic Service





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

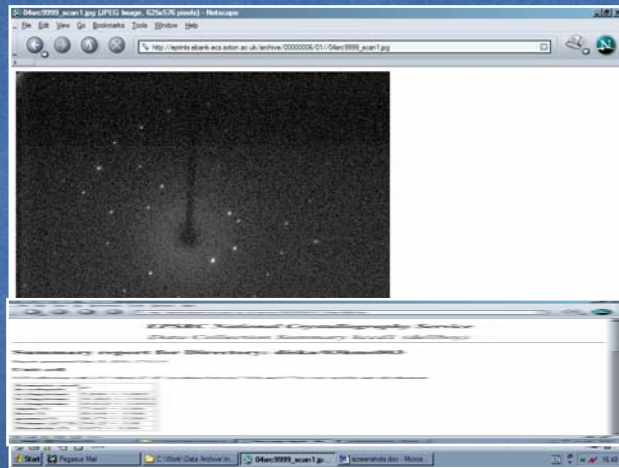


Monitoring laboratory experiments using a broker delivered over GPRS on a PDA



Crystallographic e-Prints

Direct Access to Raw Data from scientific papers

A screenshot of a web browser window showing a list of data files and crystallographic parameters. The browser's address bar shows a URL from the EPrints archive. The page is divided into two columns. The left column lists data files, and the right column lists parameters.

data

- 04src9999_nonius-config.py (1179)
- 04src9999_scan1.jpg (127447)
- 04src9999_scan2.jpg (126927)

Deposited By: Christopher Guttenidge
Deposited On: 26 February 2004

_CHEMICAL_FORMULA_SUM:	C12 H8 Cl F N2 O
CFOM:	0.0366
_CELL_ANGLE_ALPHA:	77.641(4)
_SYMMETRY_CELL_SETTING:	triclinic
_SYMMETRY_SPACE_GROUP_NAME_H_M:	P-1
_CELL_ANGLE_GAMMA:	86.374(6)
_CELL_ANGLE_BETA:	80.643(6)
_REFINE_LS_R_FACTOR_ALL:	0.1079
_REFINE_LS_WR_FACTOR_GT:	0.1091
_REFINE_LS_WR_FACTOR_REF:	0.1292
_CELL_LENGTH_A:	5.2061(3)
_CELL_LENGTH_B:	10.2615(11)
_DIFFRN_AMBIENT_TEMPERATURE:	120(2)
_REFINE_LS_R_FACTOR_GT:	0.0531
_CELL_LENGTH_C:	10.6118(10)
EXPTL_CRYSTAL_DESCRIPTION:	plate

Archive Staff Only: edit this record

A screenshot of a web browser window showing a summary report for a directory. The browser's address bar shows a URL from the EPrints archive. The page is titled "EPSRC National Crystallography Service" and "Data Collection Summary keed1 (dellboy)".

Summary report for Directory: diska/03hms003

Report generated Mar 19, 2003, 17:01:34

Unit cell

6358 reflections with $2.91^\circ < \theta < 27.48^\circ$ (resolution between 7.00Å and 0.77Å) were used for unit cell refinement

Symmetry used in scalpack	P-1
a (Angstrom)	5.2064 +/- 0.0003
b (Angstrom)	10.2621 +/- 0.0011
c (Angstrom)	10.6123 +/- 0.0010
alpha (°)	77.643 +/- 0.004
beta (°)	80.636 +/- 0.006
gamma (°)	86.374 +/- 0.006
Volume (Å ³)	546.25 +/- 0.08
Mosaicity (°)	0.673 +/- 0.004

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'

Company A
(J2EE)

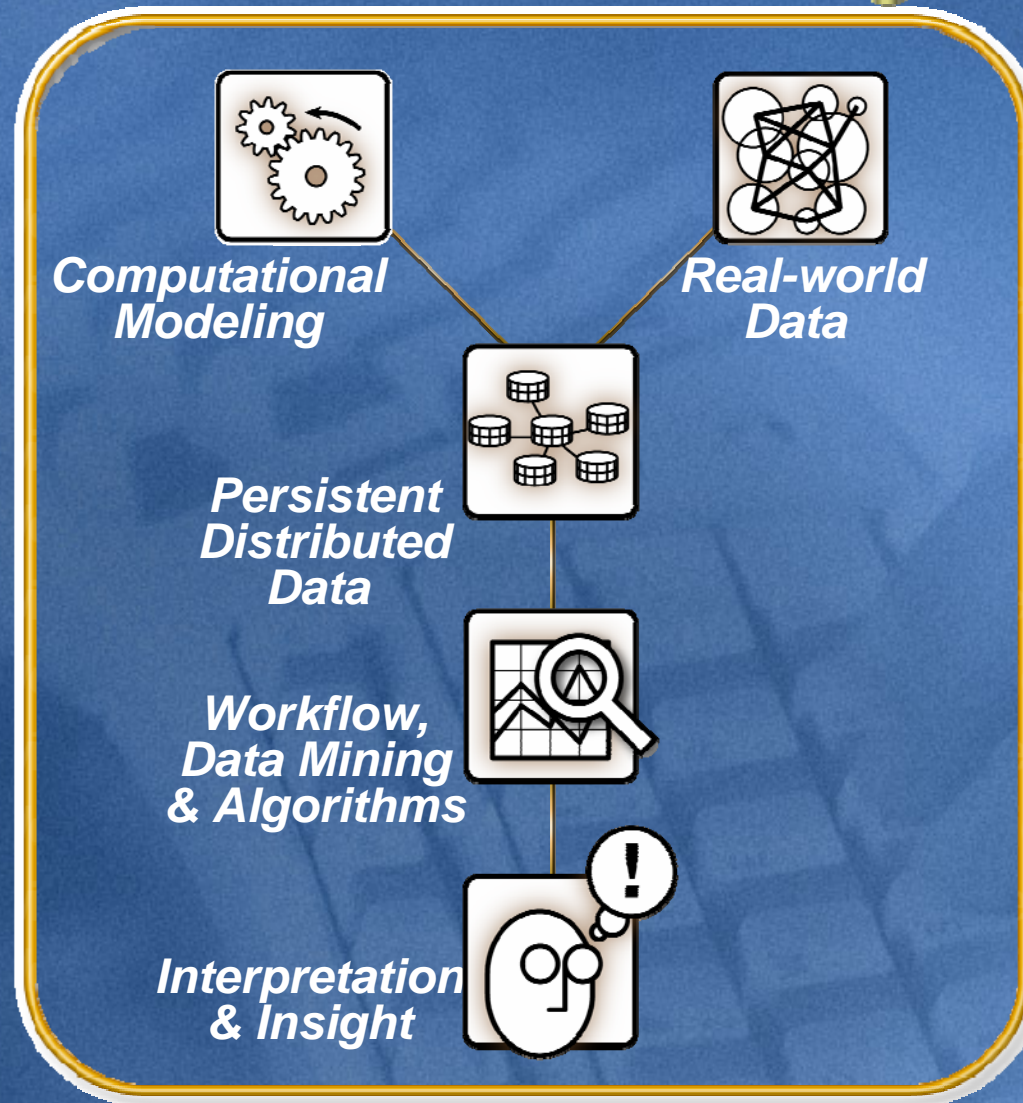
Web Services

Open Source
(OMII)

Company C
(.Net)

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

Bioinformatics

Energy Science

Earth Science

Engineering

■ ■ ■

TOOLS

Workflow, Collaboration, Visualization, Data Mining

DATA

Acquisition, Storage, Annotation, Provenance, Curation, Preservation

CONTENT

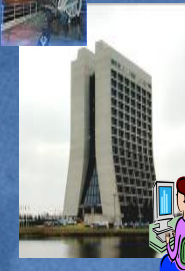
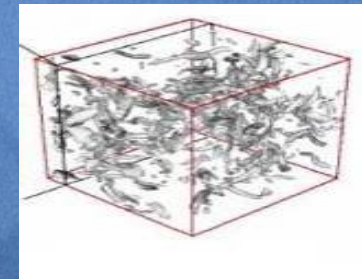
Scholarly Communication, Institutional Repositories

New Science Paradigms

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

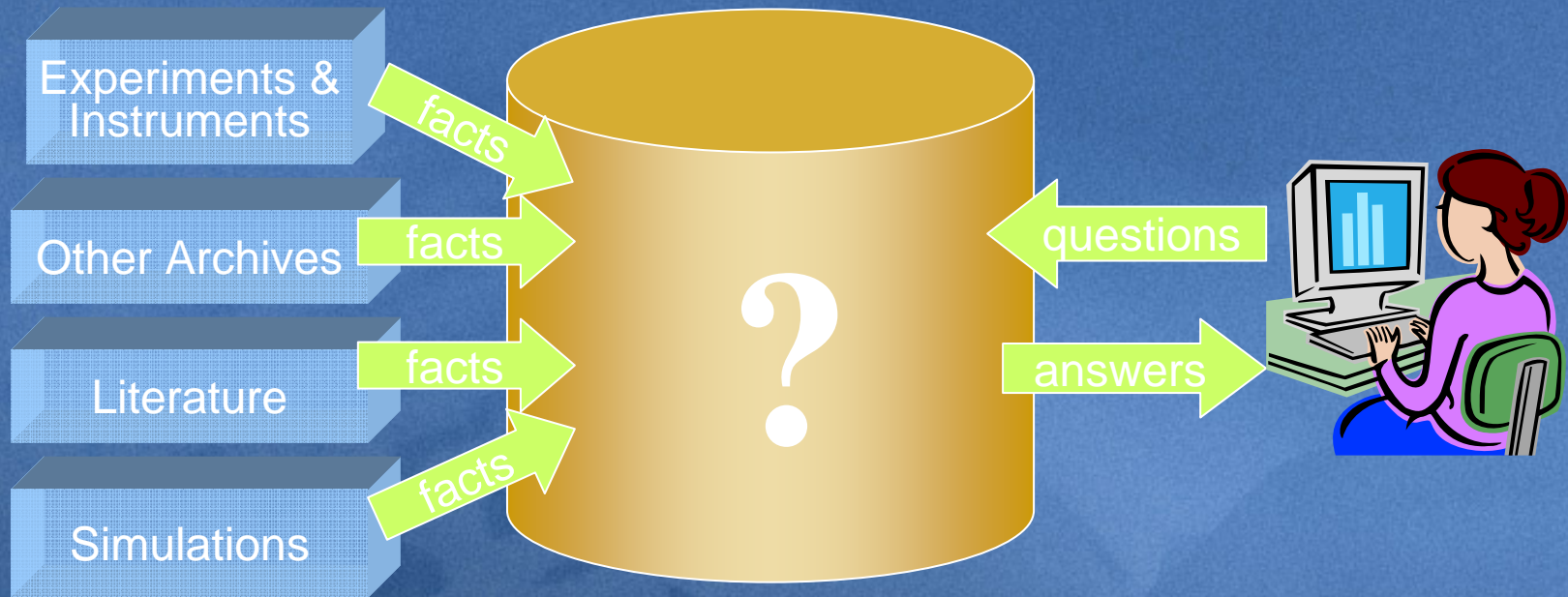


$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K \frac{c^2}{a^2}$$



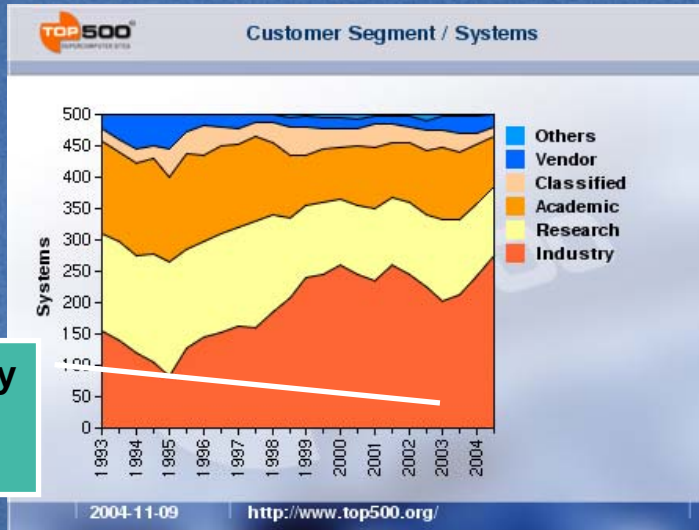
(With thanks to Jim Gray)

The Problem for the e-Scientist

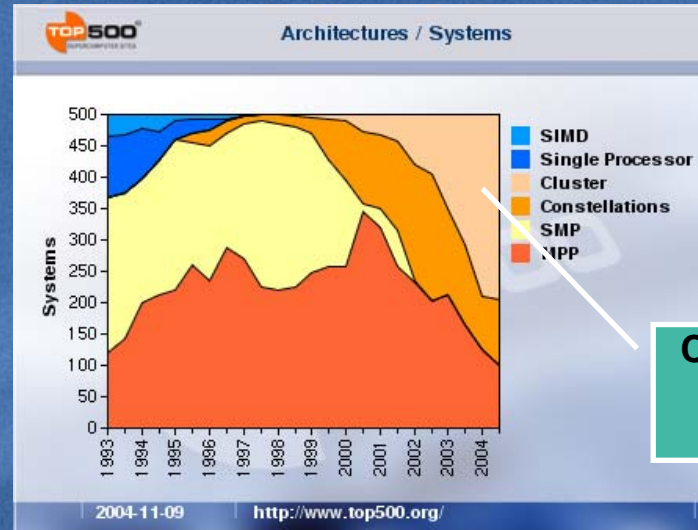


- ◆ Data ingest
- ◆ Managing a petabyte
- ◆ Common schema
- ◆ How to organize it?
- ◆ How to *reorganize* it?
- ◆ How to coexist & cooperate with others?
- ◆ Data Query and Visualization tools
- ◆ Support/training
- ◆ Performance
 - Execute queries in a minute
 - Batch (big) query scheduling

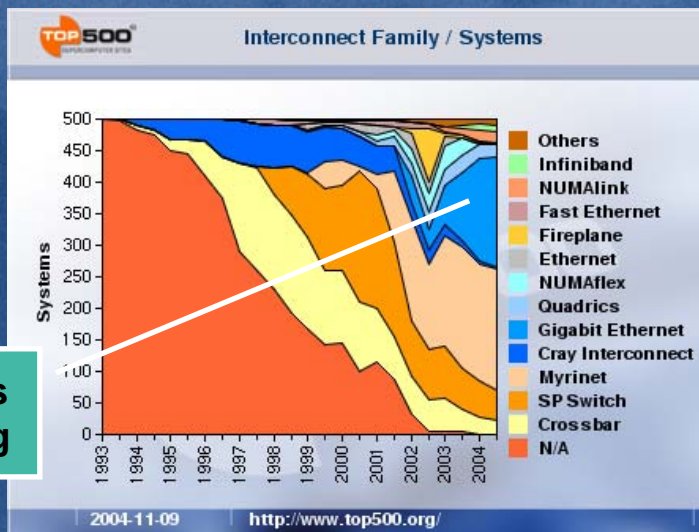
Top 500 Supercomputer Trends



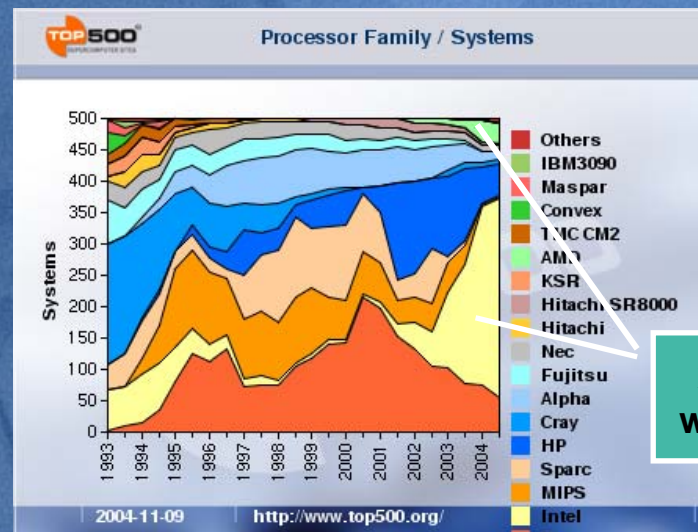
Industry usage rising



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

Grids of personal &
departmental clusters

Personal workstations &
departmental servers

Minicomputers

Mainframes



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