eScience, Semantic Computing and the Cloud
Towards a Smart Cyberinfrastructure for eScience

Tony Hey
Corporate Vice President
Microsoft Research
A Data Deluge in Science

- Data collection
  - Sensor networks, satellite surveys, high throughput laboratory instruments, observation devices, supercomputers, LHC ...

- Data processing, analysis, visualization
  - Legacy codes, workflows, data mining, indexing, searching, graphics ...

- Archiving
  - Digital repositories, libraries, preservation, ...

SensorMap
Functionality: Map navigation
Data: sensor-generated temperature, video camera feed, traffic feeds, etc.

Scientific visualizations
NSF Cyberinfrastructure report, March 2007
Emergence of a New Research Paradigm?

- 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 – **eScience or Data-centric Science**
  - Unify theory, experiment, and simulation
  - Using data exploration and data mining
    - Data captured by instruments
    - Data generated by simulations
    - Data generated by sensor networks
  - Scientists overwhelmed with data
  - Computer Science and IT companies have technologies that will help

(With thanks to Jim Gray)
Web users...

- Generate content on the Web
  - Blogs, wikis, podcasts, videocasts, etc.
- Form communities
  - Social networks, virtual worlds
- Interact, collaborate, share
  - Instant messaging, web forums, content sites
- Consume information and services
  - Search, annotate, syndicate

Scientists...

- Annotate, share, discover data
  - Custom, standalone tools
- Conferences, Journals
  - Publication process is long, subscriptions, discoverability issues
- Collaborate on projects, exchange ideas
  - Email, F2F meetings, video-conferences
- Use workflow tools to compose services
  - Domain-specific services/tools
Data can be easily produced

http://ecrystals.chem.soton.ac.uk
Thanks to Jeremy Frey
Data and services can be easily composed

Taverna Workflow
Compose services from the Web

SensorMap
Functionality: Map navigation
Data: sensor-generated temperature, video camera feed, traffic feeds, etc.
Data is easily accessible

With thanks to Catharine van Ingen
Data is easily shareable

Sloan Digital Sky Server/SkyServer
http://cas.sdss.org/dr5/en/
Today...

Computers are great tools for storing, computing, managing, and indexing huge amounts of data.

For example, Google and Microsoft both have copies of the Web for indexing purposes.
We would like computers to also help with the automatic processing of the world's information.

Computers will still be great tools for managing huge amounts of data.
Semantic Computing
Need for Semantic Computing?

- Semantic computing combines concepts and technologies that
  - Enable data modeling
  - Capture relationships
  - Allow communities to define ontologies
  - Exploit machine learning

➤ Will empower computers to reason about the data
Semantic Computing

• Some efforts are driven by the traditional “knowledge engineering” community
  – Engaged in building well-controlled ontologies
  – Important for domain-specific vocabularies with data formats and relationships specific to a community
  – Model does not easily scale to the Internet

• Some efforts are driven by the Web 2.0 community
  – Focus on the pervasiveness of Web protocols/standards
  – Emphasis on microformats (small, flexible, embeddable structures)
  – Exploit evolving and ever-expanding vocabularies such as folksonomies and tag clouds
Semantic Web as the platform?

Mark Butler (2003) Is the semantic web hype?
eScience and Semantic Computing in action
myGrid

- Semantic relationships between different data
- Semantic descriptions of services
- Annotations
- Provenance
- Repositories
- Ontologies
**myGrid: Semantic Web Technologies**

- **myGrid** built on Web Services, Workflows AND Semantic Web technologies

- Semantic Web technologies are used to:
  - Find appropriate services during workflow design
  - Find similar workflows for reuse and repurposing
  - Record the process and outcome of an experiment, in context
    - the experimental provenance
• Goal
  – Apply the capabilities of the AKT and MIAS IRCs to collaborative medical problem solving in the domain of breast cancer screening and diagnosis.

• Focus
  – Provide support to the Multi-Disciplinary Meetings (MDMs) that take place between various medical practitioners of different expertise, in coming to a collaborative diagnosis and plan of action in symptomatic focal breast disease.

• Services and technologies
  – Ontology Services
  – Annotation and Enrichment Services
  – Reasoning using services and GRID-services
MIAKT

Concept editing
MIAKT

Semantic associations

Annotations
SWAN Project

• Semantic Web Applications of Neuromedicine
• Project led by Tim Clark and the IIC
Cloud Computing
Rationale for Cloud computing

• Outsourcing of IT infrastructure
• Minimize costs
  – Large cloud/utility computing provides can have relatively very small ownership and operational costs due to the huge scale of deployment and automation
• Small businesses have access to large scale resources
  – The acquisition, operation, and maintenance costs would have been prohibiting
Example: Amazon Web Services

Simple Storage Service (S3)
- storage for the Internet
- Simple Web Services interface to store and retrieve any amount of data from anywhere on the Web

Elastic Compute Cloud (EC2)
- Compute on demand
- Virtualization
- Integration with S3

SimpleDB
- Structured data

Simple Queue Service
- Scalable message queuing

Standards-based REST and SOAP Web Service interfaces
Microsoft Cloud Services

- Exchange Server hosting
- Live@EDU
- BizTalk Services
- Mail (Live Mail, Hotmail)
- Identity (Live ID)
- Dataflow (PopFly)
- Xbox Live
- SQL Server Data Services
- Office Live Workspaces
- Windows Live

Many more coming
eScience and Cloud Computing in action
The SkyServer Project
Jim Gray (MSR) and Alex Szalay (JHU)

• The Sloan Digital Sky Survey (SDSS): The “Cosmic Genome Project”
  – 5 color images of ¼ of the sky
  – Pictures of 300 million celestial objects
  – Distances to the closest 1 million galaxies

• Built the public archive for the SDSS
• Interesting challenge in digital publishing
  – Have to publish first in order to analyze
Public Use of the SkyServer

- Posterchild in 21st century data publishing
  - 380 million web hits in 6 years
  - 930,000 distinct users vs 10,000 astronomers
  - 1600 refereed papers!
  - Delivered 50,000 hours of lectures to high schools
  - Delivered 100B rows of data

- World’s most used astronomy facility for last 2 years
GalaxyZoo

• Goal of 1 million visual galaxy classifications by the public
• Enormous publicity (CNN, Times, Washington Post, BBC)
• 100,000 people participating, blogs, poems ...

• Application is like Amazon’s ‘Mechanical Turk’ Web Service that allows users to search for photographs ...
World Wide Telescope

Seamless Rich Social Media Virtual Sky
Web application for science and education

Project organization
• Alyssa Goodman (Harvard)
• Alex Szalay (JHU)
• Curtis Wong, Jonathan Fay (MSR)

Project Goals
• Science- Seamless integration of data sets and one click contextual access
• Education- Easy as Powerpoint

Soon: http://www.worldwidetelescope.org/
WWT Architecture

• Web 2.0 browsing environment for large distributed image and information data sets with integrated rich media authoring and annotation sharing
• Integrated easy to use rich social media authoring environment
• Geospatial Tiled Multiresolution Image Browser - distributed massive image and data sets
• Links to deep web information source

Web info sources
- SIMBAD Astronomy Databases
- Published papers Astrophysics Database
- Wikipedia
- Others *

Large tiled image datasets
- Sloan Digital Sky Survey
- Hubble Space Telescope
- Chandra X-Ray Telescope
- Spitzer Infrared Space Telescope
- Digital Palomar All Sky Survey*
- Virtual Earth Surveys
- NASA Planetary Image Surveys*

Web image sources
- Multiple data sources
  By object, wavelength, format, etc. needed

Community created rich media narrated tours
Produced by astronomers, astrophysicists & educators from: Harvard, Space Telescope, Chandra, Spitzer, Adler, NOAO, Hayden, Johns Hopkins, Cal Tech et al.
Berkeley Water Center

Understanding regional hydrology

**Project Organization**
- Jim Hunt, Dennis Baldocchi, UC Berkeley
- Deb Agarwal, Lawrence Berkeley Laboratory
- Catharine van Ingen, MSR

**Goals**
- Enable rapid scientific data browsing for availability and applicability
- Enable environmental science via data synthesis from multiple sources

**Proof Points**
- Environmental Data Server, www.fluxdata.org (SharePoint), serves **921 site years** of carbon-climate field data from 160+ field teams to 60+ paper writing teams (800M values)
- Multiple projects now **leveraging** same SQL Server database and data cube approach
- CUAHSI consortium: **100 universities collaborating** on hydrology
Carbo-Climate Synthesis (BWC Dennis Baldocchi et al)

- What is the role of photosynthesis in global warming?
  - Measurements of CO2 in the atmosphere show 16-20% less than emissions estimates predict
  - The difference is either due to plants or ocean absorption.
- Communal field science – each investigator acts independently.
- Cross site studies and integration with modeling increasingly important.
- Sharepoint site [www.fluxnet.org](http://www.fluxnet.org)
  - 921 site-years of data from 240 sites around the world; 80+ site-years now being added
  - 60+ paper writing teams
  - American data subset is public and served more widely
  - Summary data products greatly simplify initial data discovery
Mashup of Ameriflux Sites
Digital Watersheds (BWC, James Hunt)

- Russian River watershed challenges: forestry, farming, urbanization, gravel mining, and fish habitat restoration.
  - Can we understand historic and on-going changes using only publically available data sources such as USGS, NOAA, Sonoma Ecology Center, etc?

- Early studies examined overall water balance and changes in suspended sediment
  - scientific data “mashups” are leading to useful results.

- Recent engagement with National Marine Fisheries and USBR expanding this to other watersheds across Northern California

- Sharing technology with CUAHSI (100 universities)
  “We see water through a fish eye lens”

http://bwc.berkeley.edu
http://www-esd.lbl.gov/BWC/California
http://www.cuahsi.org/
Supporting researchers worldwide

A Software + Services vision
Research Pipeline

- Data Acquisition and Modeling
  - Data capture from source, cleaning, storage, etc.
  - SQL Server, SSIS, Windows WF
- Support Collaboration
  - Allow researchers to work together, share context, facilitate interactions
  - SharePoint Server, One Note 2007 (shared)
- Data Analysis, Modeling, and Visualization
  - Mining techniques (OLAP, cubes) and visual analytics
  - SQL Analysis Services, BI, Excel, Optima, SILK (MSR-A)
- Disseminate and Share Research Outputs
  - Publish, Present, Blog, Review and Rate
  - Word, PowerPoint
- Archiving
  - Published literature, reference data, curated data, etc.
  - SQL Server

Microsoft has technologies that can offer end-to-end support
Math in Word 2007

\[ E = mc^2 \]

\[ e^{-ti\theta} \]

\[ x = \lim_{n \to \infty} \left( 1 + \frac{1}{n} \right)^n \]

\[ y = \int \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

\[ x + y = \sum_{n}^{5} z + 1 \]
Scientific workflow workbench to automate the data processing pipelines of the world’s first plate-scale undersea observatory

Goals
- From raw data to useable data products (visualizations)
- Focusing on cleaning, analysis, regridding, interpolation
- Support real time, on-demand visualizations
- Custom activities and workflow libraries for authoring
- Visual programming accessible via a browser

Proof Points
- A scientific workflow workbench for a number of science projects, reusable workflows, automatic provenance capture.
- Demonstrate scientific use of Windows WF and SQL Server
- Demo at TechFest 2008, in collaboration with the UW Neptune team.
Using Spam Blockers To Target HIV, Too
A Microsoft researcher and his team make a surprising new assault on the AIDS epidemic

BY STEPHEN BAKER
AND JAY GREENE

This parallel between spam and biology resonated for Heckerman, a physician as well as a PhD in computer science. It didn't take him long to realize that his spam-blocking tool could extend far beyond junk e-mail, into the realm of life science. In 2003, he surprised colleagues in Redmond, Wash., by refocusing the spam-blocking technology on one of the world's deadliest, fast-evolving consumers: HIV, the virus that leads to AIDS.

Heckerman was plunging into medicine—and carrying Microsoft with him. When he brought his plan to Bill Gates, the company chairman "got really excited," Heckerman says. Well versed on HIV from his philanthropy work, Gates lined up Heckerman with AIDS researchers at Massachusetts General Hospital, the University of Washington, and elsewhere.

Since then, the 50-year-old Heckerman and two colleagues have created their own biology niche at Microsoft, where they build HIV-detecting software. These are research tools to spot infected cells and correlate the viral mutations with the individual's genetic profile.

Heckerman's team runs mountainous data through enormous clusters of 320 computers, operating in parallel. Thanks to smarter algorithms and more powerful machines, they're sifting through the data 480 times faster than a year ago. In June, the team released its first batch of tools for free on the Internet.

"A new industry for the behemoth to conquer? Not exactly. Heckerman's tool in Redmond represents just one small node in a global AIDS research effort marked largely by cooperation. "The Microsoft group has a different perspective and a good statistical background," says Bethe Korber, an HIV researcher at Los Alamos National Laboratories. The key query they all face is the virus itself, which is proving wilder than any of Microsoft's corporate foes. While Heckerman has high hopes that his tools will lead to vaccines that can be tested on humans within three years, his research
Better vaccine design through improved understanding of HIV evolution

Project Organization
• Bruce Walker & Zabrina Brumme, Mass General
• Philip Goulder, Oxford
• Richard Harrigan, University of British Columbia
• David Heckerman, Jonathan Carlson and Carl Kadie, MSR

Goals
• Use machine learning and visualization tools developed at Microsoft, which require HPC, to build maps of within-individual evolution of the HIV virus

Proof Points
• Discovered epitope decoys that could have predicted recent failure of Merck vaccine
• Patent filed on new method for learning graphical models from data
• Algorithms and medical results published in Science and Nature Medicine
• MSR Computational Biology Tools published (Source on CodePlex)
PhyloD: Leveraging phylogeny for associations studies

- Problem: Find associations between genotypes and phenotypes (e.g., genetic causes of disease) using data from a set of “individuals” (humans, HIV viruses, etc.)

- Previous solutions ignore phylogenetic structure (i.e., assume the data to be IID). This solution leverages phylogenetic structure of the individuals.

- Applications:
  - Identify effects of immune pressure on HIV evolution (papers in *Science* and *Nature Medicine*)
  - Inferring protein structure (in collaboration with David Baker)
  - Genome-wide association studies (personalized medicine)
Chemistry Drawing for Office

- Peter Murray Rust, Univ. of Cambridge
- Murray Sargent, Office
- Geraldine Wade, Advanced Reading Technologies

Goals
- Support students/researchers in simple chemistry structure authoring/editing
- Enable ecosystem of tools around lifecycle of chemistry-related scholarly works
- Support the Chemistry Markup Language
- Proof of concept plug-in

Execution
- MSR Developer to work on the proof of concept
- Post-doc in Cambridge to use plug-in and give feedback and move their chemistry tools to .NET and Office
- Advanced Reading Technologies to create necessary glyphs
Semantic Annotations in Word

• Phil Bourne and Lynn Fink, UCSD

Goals
• Semantic mark-up using ontologies and controlled vocabularies
• Facilitate/automate referencing to PDB (and other resources) from manuscript
• Conversion of manuscript to NLM DTD for direct submission to publisher

Scenario
• Authors do not need to be aware of the use of semantic technologies
• A domain-specific ontology is downloaded and made available from within Microsoft Word 2007
• Authors can record their intention, the meaning of the terms they use based on their community’s agreed vocabulary
NLM DTD plug in

XML

nlmx

Journal Template
A platform for building services and tools for research output repositories
• Papers, Videos, Presentations, Lectures, References, Data, Code, etc.
• Relationships between stored entities

Goals
• Support the MSR publishing and dissemination platform for all researcher outputs
• Enable a tools and services ecosystem for “research output” repositories on MS technologies

Execution
• Support Eprints and Dspace front ends
• Deployment within MSR early Q2
• Release to the community late Q2
• Built on SQL Server 2008 + Entity Framework
Research Output Repository Platform

- A Semantic Computing platform
- A hybrid between a relational database and a triple store

**Triple stores**
- Evolution friendly
- Poor performance
- No need to model everything in advance
- Semantic interpretation at the application level

**Relational schema**
- Evolution not so easy
- Great opportunities for optimization
- Model everything in advance

**Research Output Repository Platform**
- Maintain a balance
- Try to model the frequently used entities in our app domain
- Try to capture the frequently used relationships
- Allow for extensibility (Relationships, Properties)
Research Output Repository Platform

PDF file

PowerPoint presentation

authored by tony

Lecture on 2/19/2008

contains

presented by

organized by Elizabeth, Sebastien, Matthew, Norman, Brian, Sarah, George, Roy
A Digital Dark Age?

The National Archives in Microsoft deal
National Archive project to avert digital dark age

Mavis Kennedy

The National Archives and Microsoft have announced a partnership to preserve what was described as a “digital dark age” and unlock millions of unarchived computer files.

The archives, which contain a vast collection of documents, are attempting to address the problem of how to digitize and preserve the archives. The archives are estimated to hold over 5,000 terabytes of data.

The partnership involves the software giant, Microsoft, and the National Archives. The partnership aims to create a cloud-based platform for managing and preserving the archives.

Mavis Kennedy, head of the archives, said: “This is a crucial first step for us, and for society as a whole. We want our personal records to survive, and we expect our personal records to be protected, but someone with long-term care needs should not lose access to their documents.”
**Organization**
- High-profile EU Commission Project, €14M for 4 years
- Consortium of 5 national libraries, 4 national archives, 4 universities and 4 industry partners

**Goals**
- Preservation of Office Documents based on OpenXML
- Deliver converters for MS Office binary formats
- Funded open source project for ODF to/from OpenXML converter
- Deliver Preservation Toolkit

**PLANETS**
Tools and methods for sustainable long-term preservation of digital objects
eScience and Semantic Computing meet the Cloud

The cyberinfrastructure for the next generation of researchers
The Future: Software plus Services for Science?

• Expect scientific research environments will follow similar trends to the commercial sector
  – Leverage computing and data storage in the cloud
  – Scientists already experimenting with Amazon S3 and EC2 services, with mixed results;

• For many of the same reasons
  – Siloed research teams, no resource sharing across labs
  – High storage costs
  – Low resource utilization
  – Excess capacity
  – High costs of reliably keeping machines up-to-date
  – Little support for developers, system operators
A smart cyberinfrastructure

• Collective intelligence
  – If last.fm can recommend what song to broadcast to me based on what my friends are listening to, why cannot the cyberinfrastructure of the future recommend articles of potential interest based on what the experts in the field that I respect are reading?
  – Already examples emerging but the process is manual (Connotea, BioMedCentral Faculty of 1000 ...)

• Automatic correlation of scientific data
• Smart composition of services and functionality
• Cloud computing to aggregate, process, analyze and visualize data
A world where all data is linked...

• Data/information is interconnected through machine-interpretable information (e.g. paper X is about star Y)
• Social networks are a special case of ‘data networks’

• Important/key considerations
  – Formats or “well-known” representations of data/information
  – Pervasive access protocols are key (e.g. HTTP)
  – Data/information is uniquely identified (e.g. URIs)
  – Links/associations between data/information

Attribution: Richard Cyganiak
...and stored/processed/analyzed in the cloud

Vision of Future Research Environment with both Software + Services

- Reference management
- Project management
- Visualization and analysis services
- Scholarly communications
- Domain-specific services
- Blogs & social networking
- Search books citations
- Instant messaging
- Identity
- Notification
- Mail
- Document store
- Storage/data services
- Knowledge management
- Compute services virtualization
- Knowledge discovery
**Acknowledgements**

- The ideas presented here were developed with input from many colleagues in the community and at Microsoft Research:
  - Thanks are due to David De Roure, Jeremy Frey, Carole Goble, Peter Murray-Rust, Alan Rector, Nigel Shadbolt and Alex Szalay
  - And special thanks to Roger Barga, Savas Parastatidis and Evelyne Viegas at Microsoft Research who have tried to educate me ...
- See [www.microsoft.com/science](http://www.microsoft.com/science) for some more details of Microsoft’s activities in Scientific and Technical Computing
Microsoft

Your potential. Our passion.