# eScience, Semantic Computing and the Cloud

## Towards a Smart Cyberinfrastructure

**Tony Hey**

Corporate Vice President

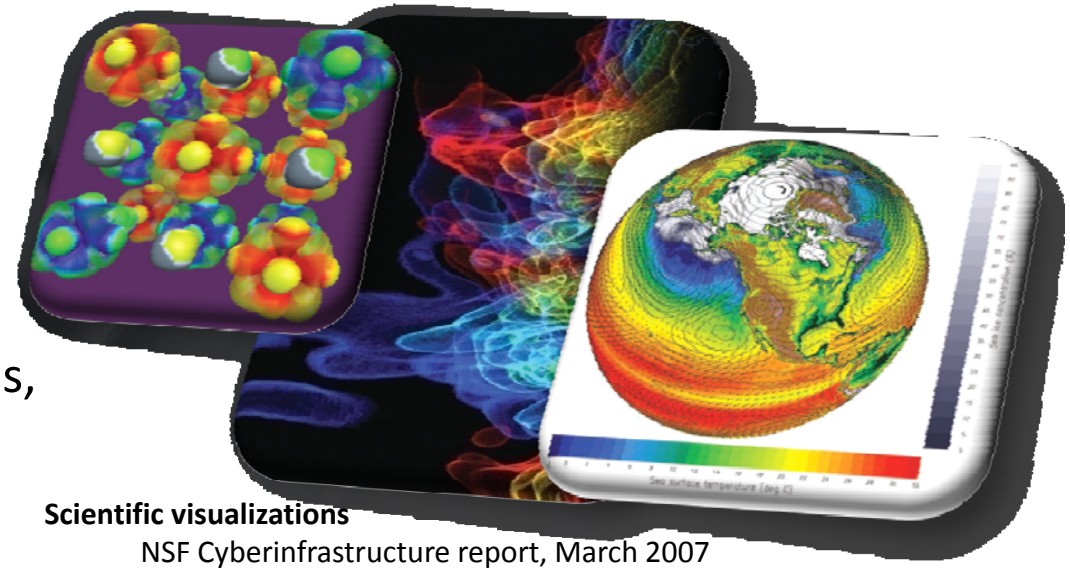
Microsoft Research

# eScience

# 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.

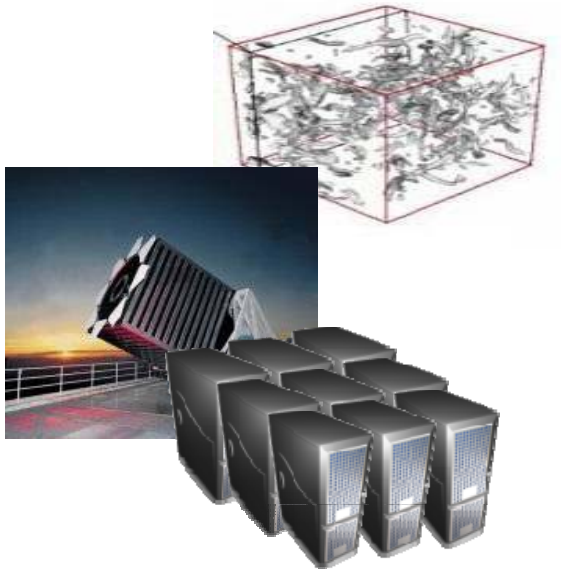


**Scientific visualizations**
NSF Cyberinfrastructure report, March 2007

# 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

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



(With thanks to Jim Gray)

# Today

## 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

# Open Collaboration

**Open access**

**Open source**

**Open data**

**"In order to help catalyze and facilitate the growth of advanced CI, a critical component is the adoption of open access policy for data, publications and software."**

NSF Advisory Committee on Cyberinfrastructure (ACCI)


**http://www.microsoft.com/interop/**

- Microsoft Interoperability Principles
  - Open Connections to Microsoft Products
  - Support for Standards
  - Data Portability
  - Open Engagement

# Today…

Computers are great **tools** for


huge amounts of **data**

For example, Google and Microsoft both have copies of the Web for indexing purposes

# Tomorrow…

Computers will still be great **tools** for


huge amounts of **data**

We would like computers to also help with the **automatic**

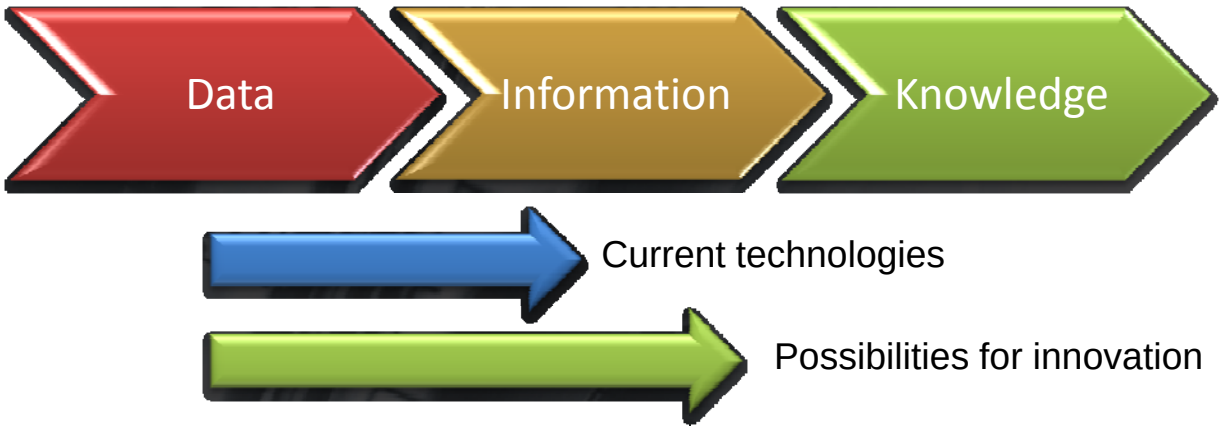

of the world's **information**

# 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?
http:/www.hpl.hp.com/personal/marbut/isTheSemanticWebHype.pdf

# 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

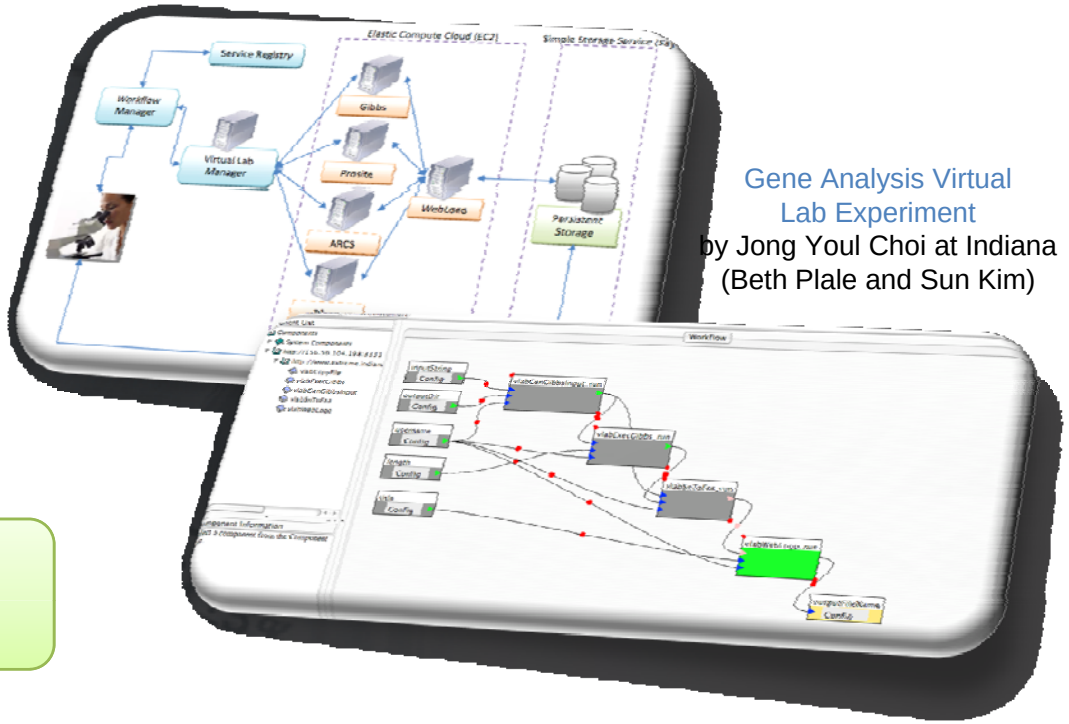
## SimpleDB

- Structured data

## Simple Queue Service

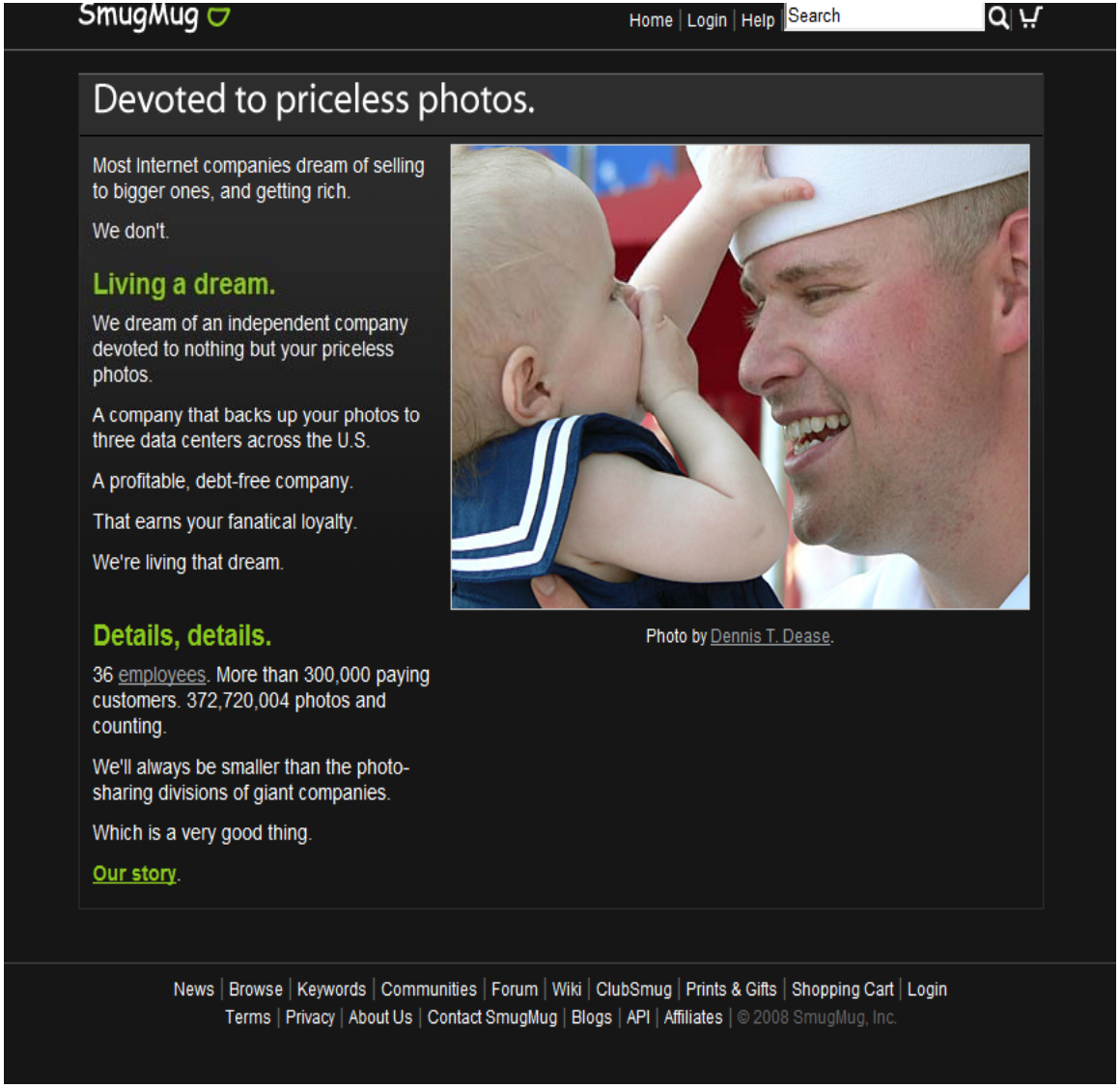
- Scalable message queuing

Standards-based REST and SOAP Web Service interfaces




Gene Analysis Virtual Lab Experiment
by Jong Youl Choi at Indiana
(Beth Plale and Sun Kim)

# www.smugmug.com




# Microsoft Cloud Services

- Exchange
- Live ID
- Xbox Live
- SQL Server Data Services
- Office Live Workspaces
- Windows Live
- Live Mesh
- .NET Online

➢ 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

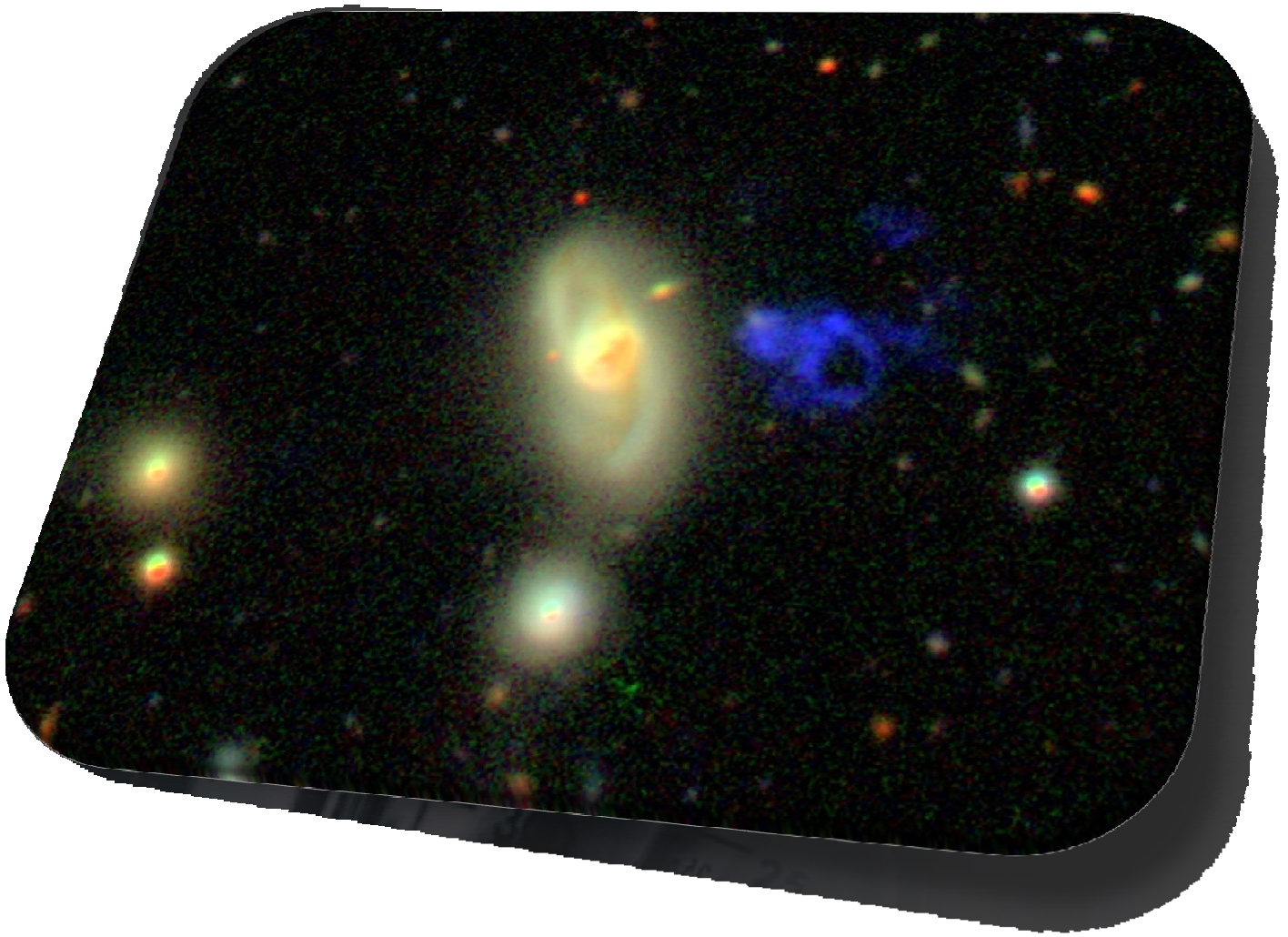

➢ **World's most used astronomy facility for last 2 years**

# 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 …**

# Hanny's Voorwerp



# World Wide Telescope


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

Participants

- Alyssa Goodman; Harvard University
- Alex Szalay; Johns Hopkins University
- Curtis Wong, Jonathan Fay; Microsoft Research

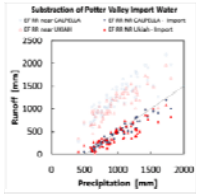
Goals

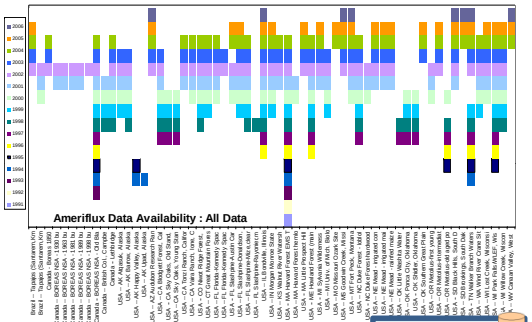
- Integration of data sets and one-click contextual access
- Easy access and use
- In just over a little more than two months, a million users have downloaded, installed and launched the application (2,206,497 unique sessions)

We invite you to experience it!
www.worldwidetelescope.org

# Berkeley Water Center

Understanding regional hydrology



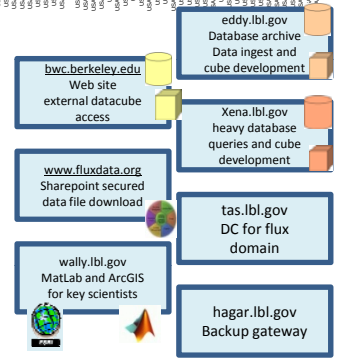




## Project Organization

- 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

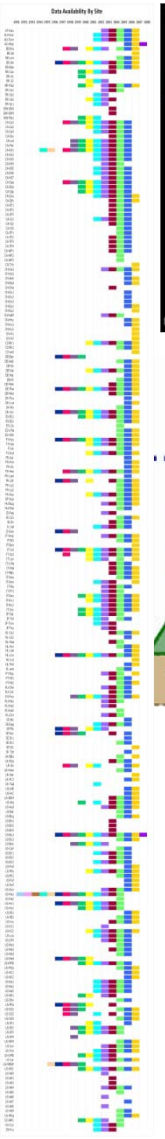


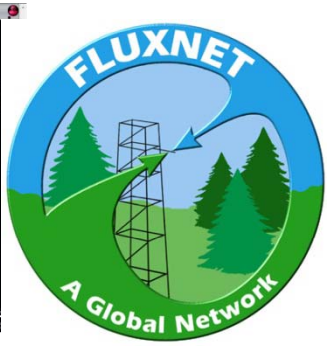

## Proof Points

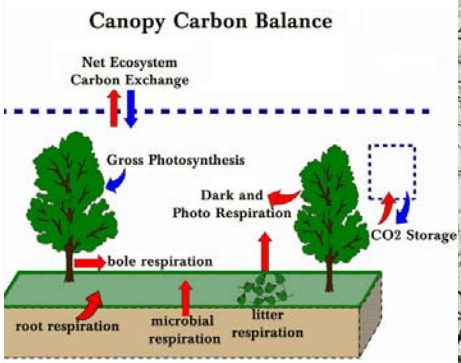
- Environmental Data Server, www.fluxdata.org (SharePoint), serves **921 site years** of carbon-climate field data from 160+ field teams to 60+ paper writing teams (800M values)
- Multiple projects now **leveraging** same SQL Server database and data cube approach
- CUAHSI consortium: **100 universities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

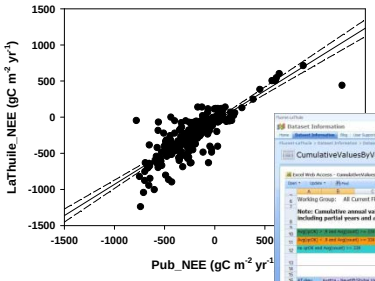
# Carbo-Climate Synthesis (BWC Dennis Baldocchi et al)

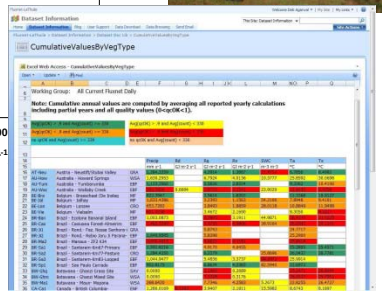
- What is the role of photosynthesis in global warming?
  - Measurements of CO2 in the atmosphere show 16-20% less than emissions estimates predict
  - The difference is either due to plants or ocean absorption.
- Communal field science – each investigator acts independently.
- Cross site studies and 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





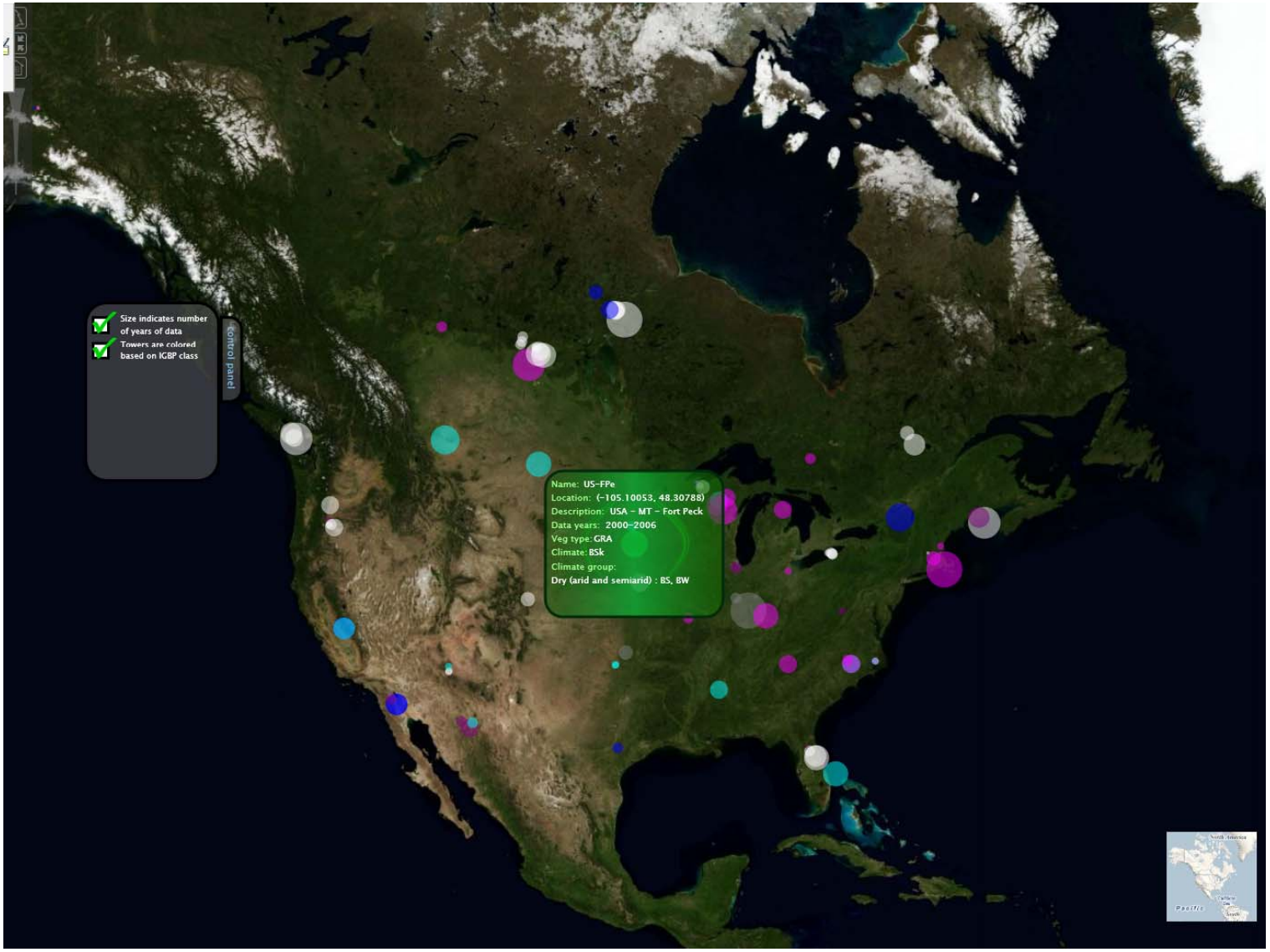



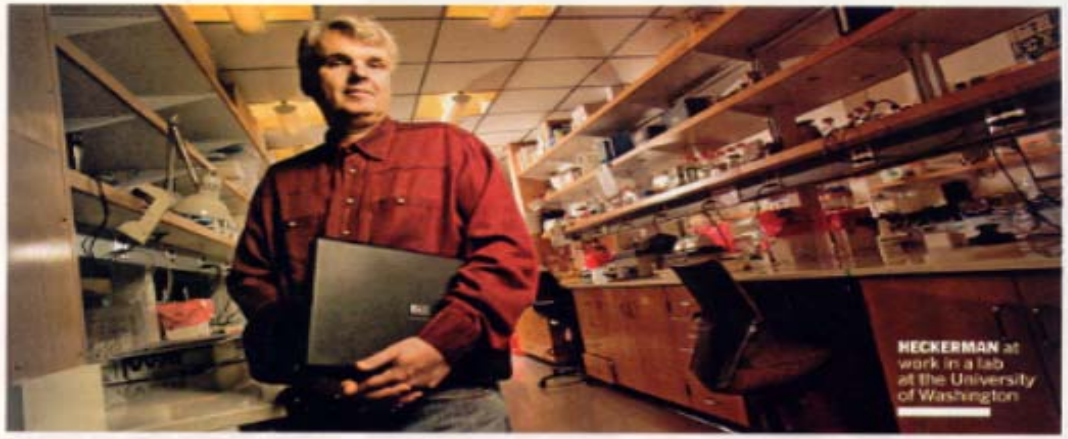






# Mashup of Ameriflux Sites





HECKERMAN at work in a lab at the University of Washington

# 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

CUT-RATE PAINKILLERS! Unclaimed riches in Nigeria!! Most of us quickly identify such e-mail messages as spam. But how would you teach that skill to a machine? David Heckerman needed to know. Early this decade, Heckerman was leading a spam-blocking team at Microsoft Research. To build their tool, team members meticulously mapped out thousands of signals that a message might be junk. An e-mail featuring "Viagra," for example, was a good bet to be spam—but things got complicated in a hurry.

If spammers saw that "Viagra" messages were getting zapped, they switched to Viagra, or Vi agra. It was almost as if spam, like a living thing, were mutating.

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, fastest-mutating conundrums: 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.

**Similar mutations may crop up in computer and medical viruses**

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 mountains of 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 nook 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 Bette Korber, an HIV researcher at Los Alamos National Laboratories. The key quarry they all face is the virus itself, which is proving wilier 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

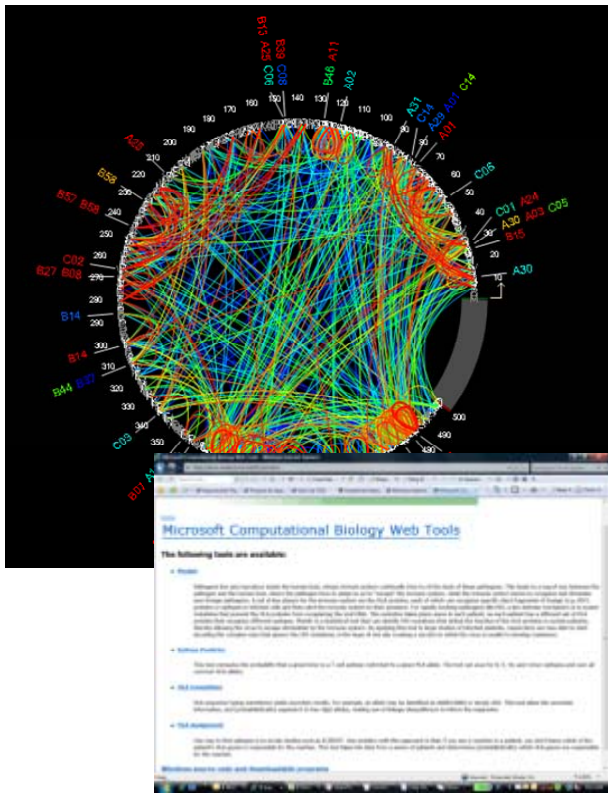
JACK CHATMAN



# Computational Biology Web Tools

Better vaccine design through improved understanding of HIV evolution

PARTNERS AIDS RESEARCH CENTER




## 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

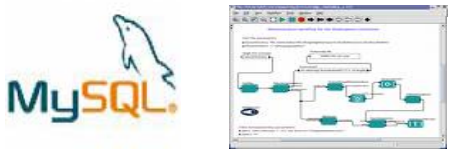
- **Discovered epitope decoys** that could have predicted recent failure of Merck vaccine
- **Patent filed** on new method for learning graphical models from data
- Algorithms and medical results **published in Science and Nature Medicine**
- MSR Computational Biology **Tools published** (Source on CodePlex)

# Supporting researchers worldwide

Adding Semantics to Software Tools

# 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











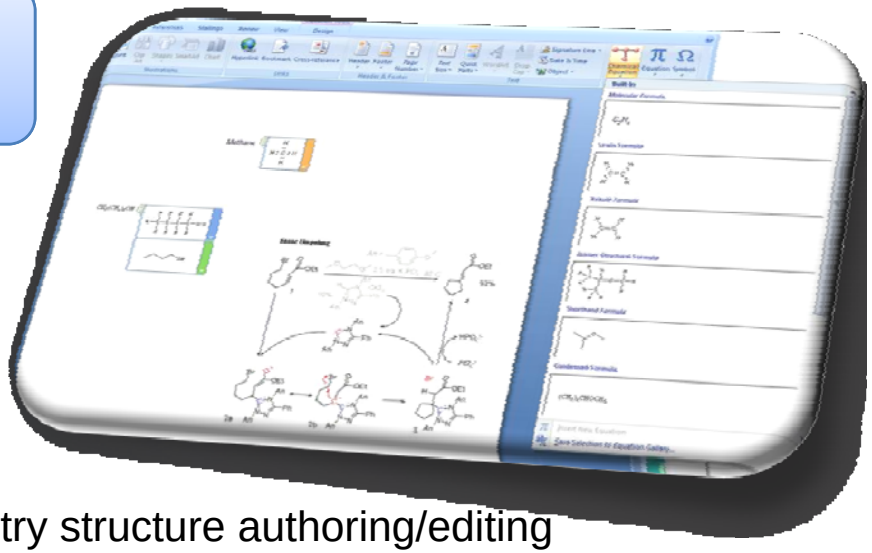




***Microsoft has technologies that can offer end-to-end support***

Data Acquisition & Modeling | Collaboration | Analysis | **Disseminate & Share** | Archiving

# Chemistry Drawing for Office



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

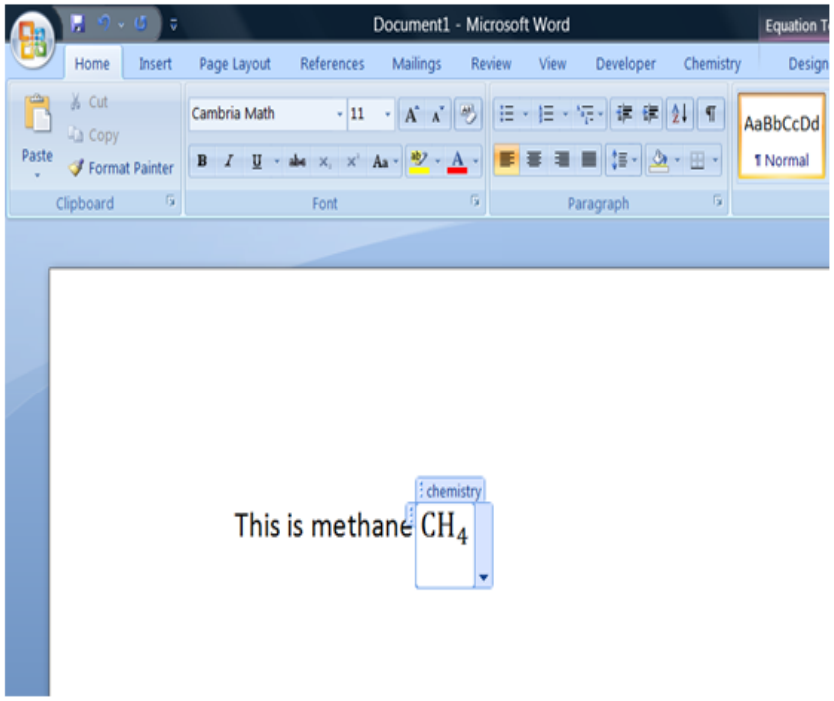
## Goals

- 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

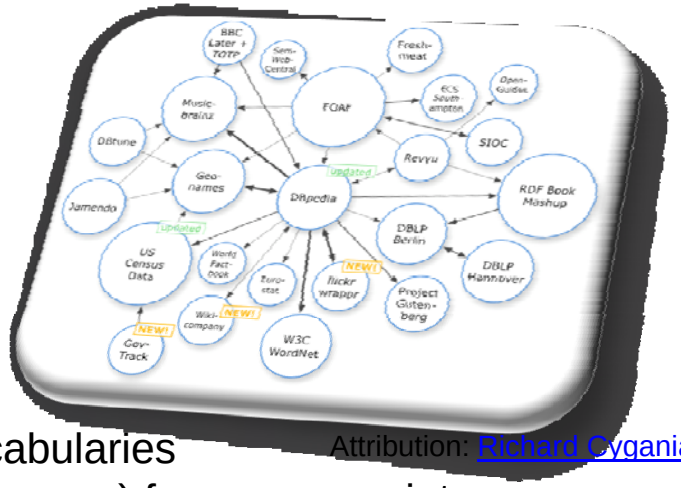
# A “Chemistry Zone” in a Word document and the CML representation (in pseudo-XML) stored inside the OOXML document




```
<cml ...>
  <molecule ...>
    <atomArray>
      <atom elementType="C" ... />
      <atom elementType="H" ... />
      ...
    </atomArray>
    <bondArray>
      <bond ... />
      <bond ... />
      ...
    </bondArray>
  </molecule>
</cml>
```

# Semantic Annotations in Word


- Phil Bourne and Lynn Fink, UCSD


Attribution: Richard Cyganiak

## 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

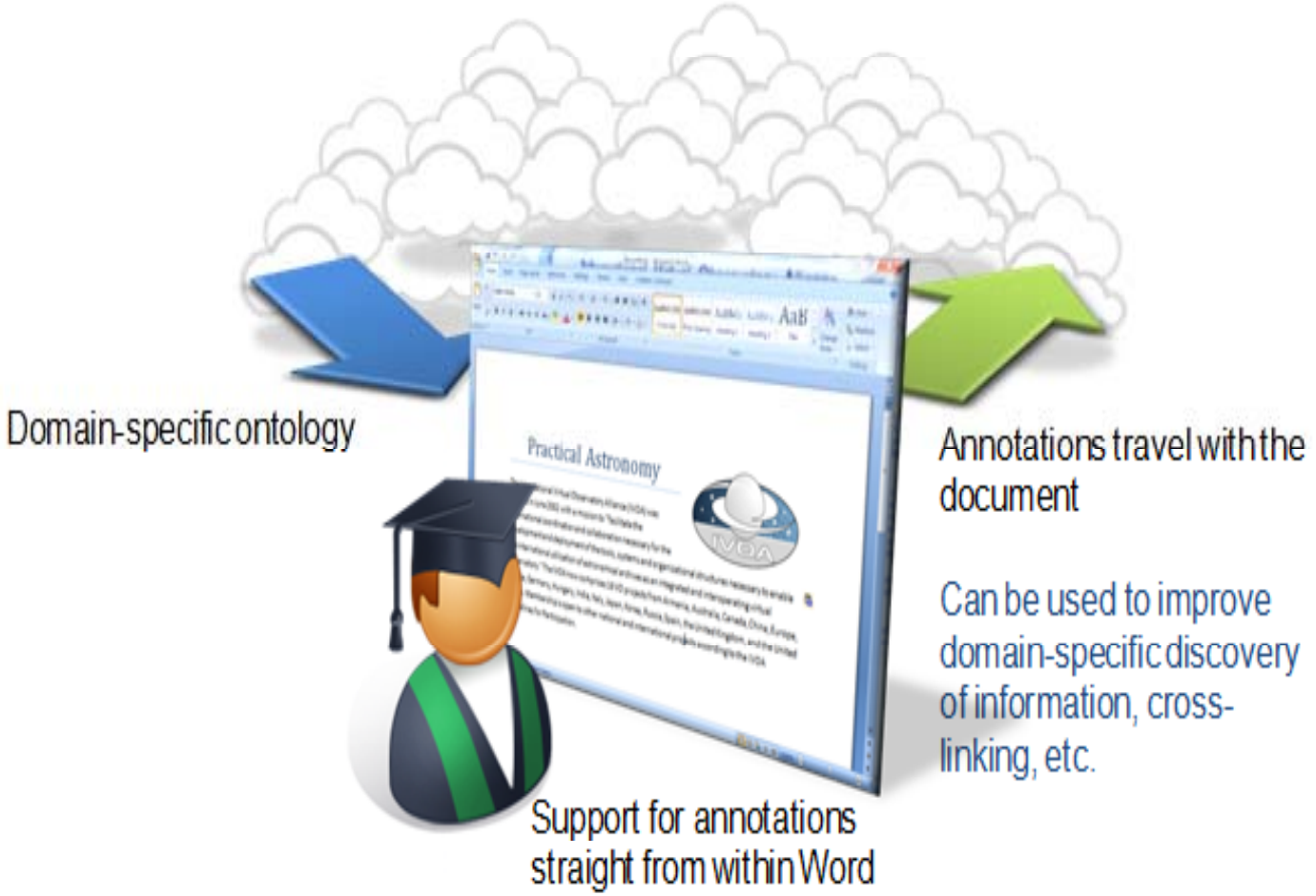
## Scenario

• Authors do not need to be aware of the use of semantic technologies

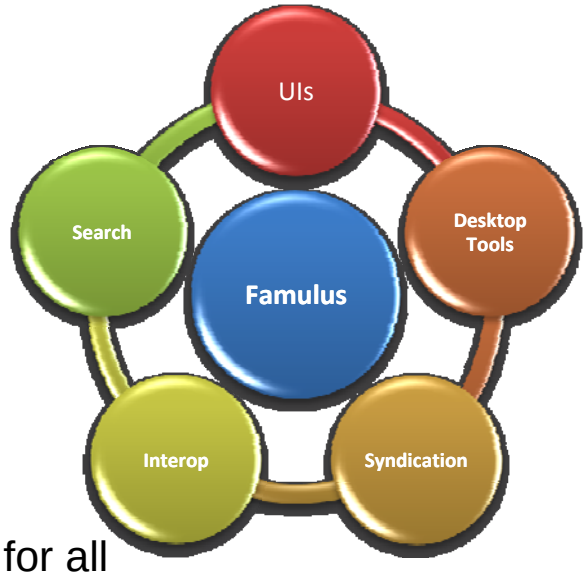
• A domain-specific ontology is downloaded and made available from within Microsoft Word 2007

• Authors can record their intention, the meaning of the terms they use based on their community's agreed vocabulary

# Semantic annotations in Word




# Research Output Repository

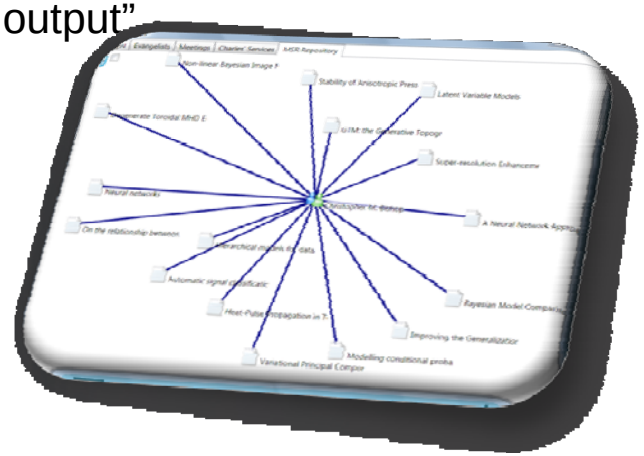



A platform for building services and tools for research output repositories

- 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

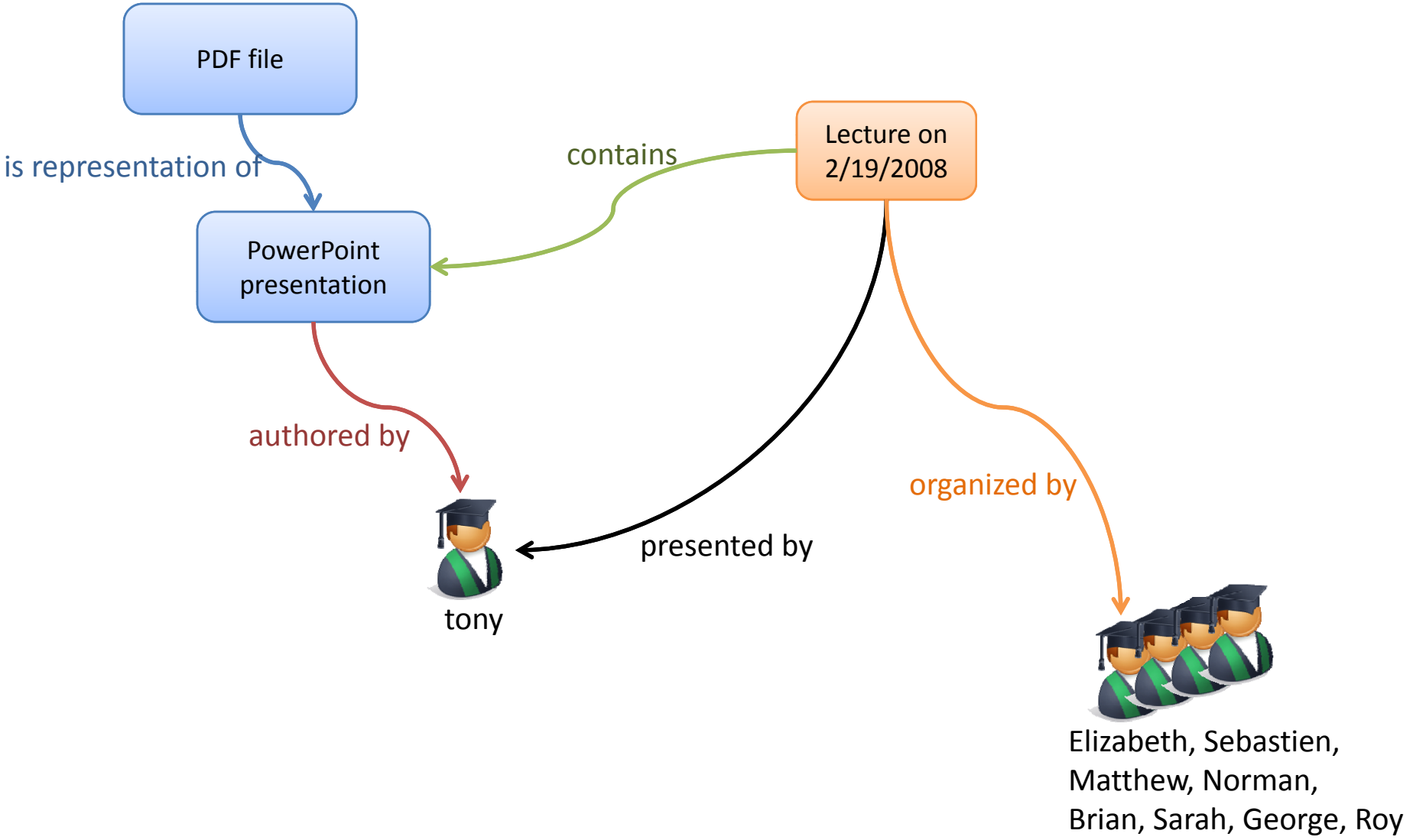
- A Semantic Computing platform
- A hybrid between a relational database and a triple store

**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

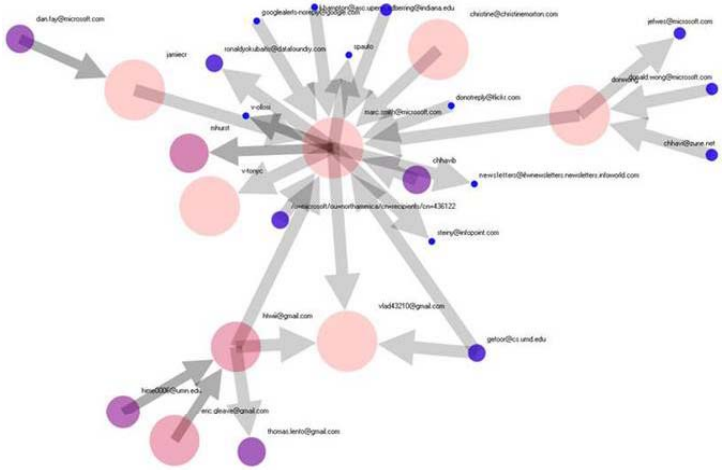
**Research Output Repository Platform**
-Maintain a balance
-Try to model the frequently used entities in our app domain
-Try to capture the frequently used relationships
-Allow for extensibility (Relationships, Properties)

# Research Output Repository Platform




# .NET Map

Network Analysis Visualization

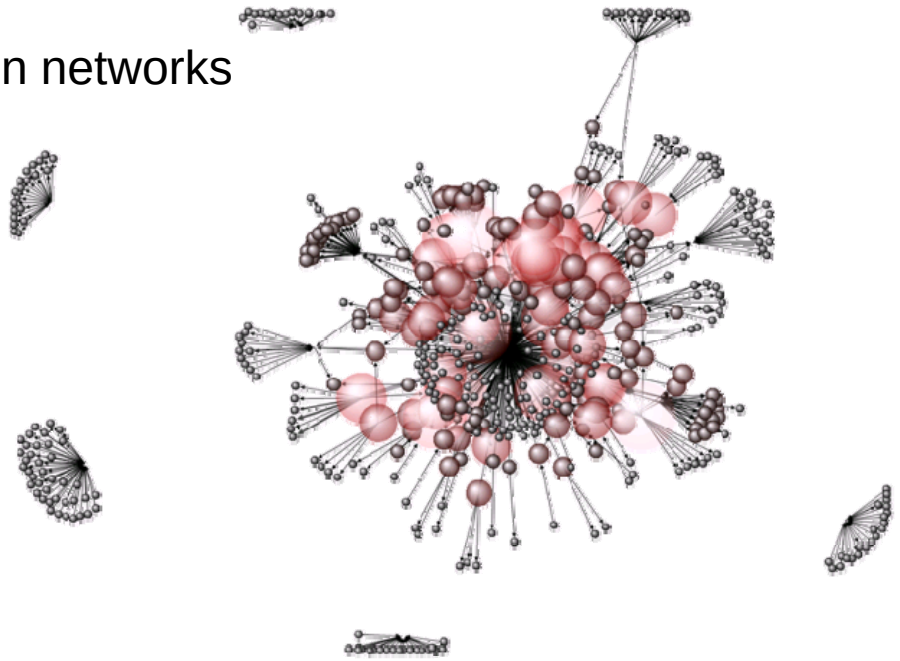


## Project Organization

- Marc Smith, Senior Research Sociologist (MSR)

## Goals

- Research in the visualization of interaction networks
- Support for directed graphs
- Relationship analysis



## Proof Points

- Standalone tools on Windows
- Available as an Excel 2007 plugin

# eScience and Semantic Computing meet the Cloud

The cyberinfrastructure for the next generation of researchers

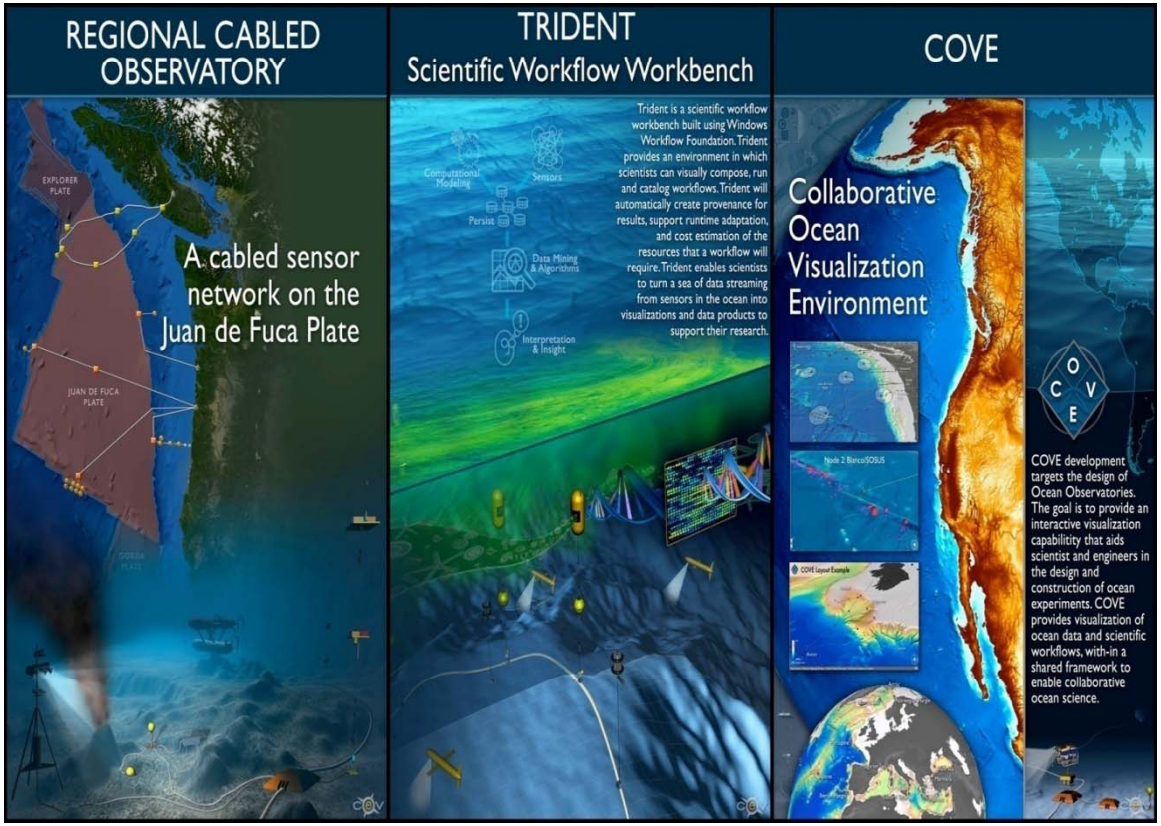
# The Future: Software plus Services for Science?

- 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
  - No resource sharing across different research labs
  - High storage costs
  - Low resource utilization
  - Excess capacity
  - High costs of reliably keeping machines up-to-date
  - Need less support for developers, system operators

# Trident – Scientific Workbench

*Workflow for Ocean Observatories, part of an "oceanographer's workbench"*

*Jim Gray*

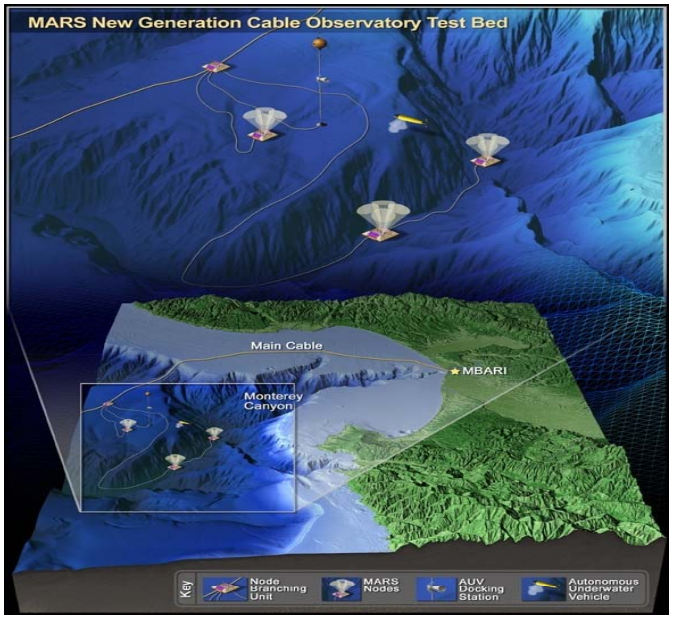



http://www.microsoft.com/mscorp/tc/trident.mspx

Data Acquisition & Modeling | Collaboration | **Analysis** | Disseminate & Share | Archiving

## Trident Scientific Workflow Workbench

Univ. of Washington and Monterey Bay Aquarium Research Institute




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
- Focusing on cleaning, analysis, re-gridding, interpolation
- Support real time, on-demand visualizations
- Custom activities and workflow libraries for authoring
- Visual programming accessible via a browser
- Trial Cloud Services for science

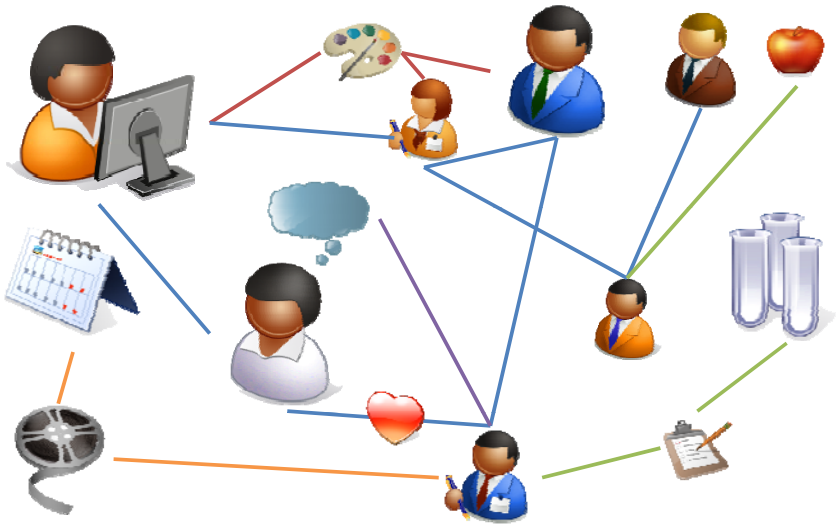
### Proof Points

- A **scientific workflow workbench** for a number of science projects, reusable workflows, automatic provenance capture.
- **Demonstrate scientific use** of Windows WF, HPCS, SQL Server and Cloud Service SSDS

# Towards a smart cyberinfrastructure?

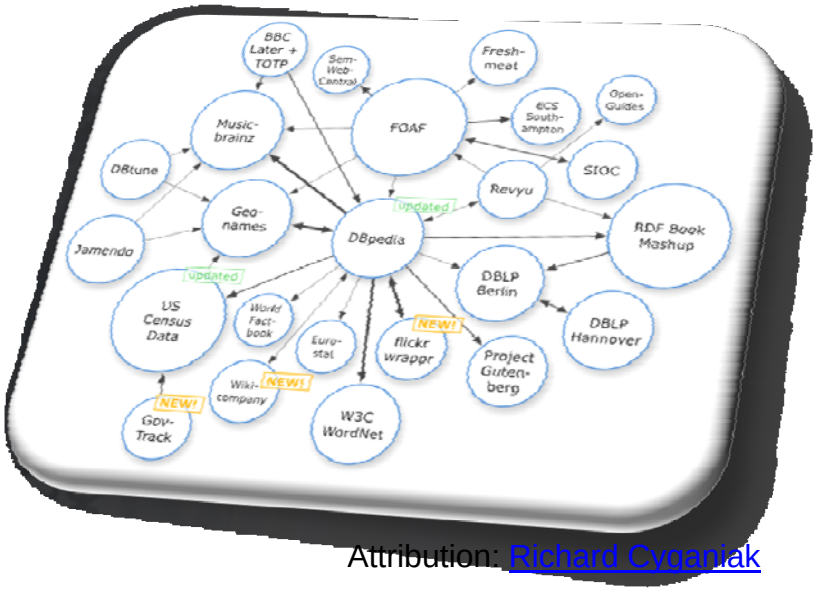
- Collective intelligence
  - If last.fm can recommend what song to broadcast to me based on what my friends are listening to, the cyberinfrastructure of the future should recommend articles of potential interest based on what the experts in the field that I respect are reading?
  - Examples are emerging but the process is presently manual (Connotea, BioMedCentral Faculty of 1000 ...)
- Semantic Computing
  - 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 inter-connected through machine-interpretable information (e.g. paper X **is about** star Y)
- Social networks are a special case of ‘data meshes’

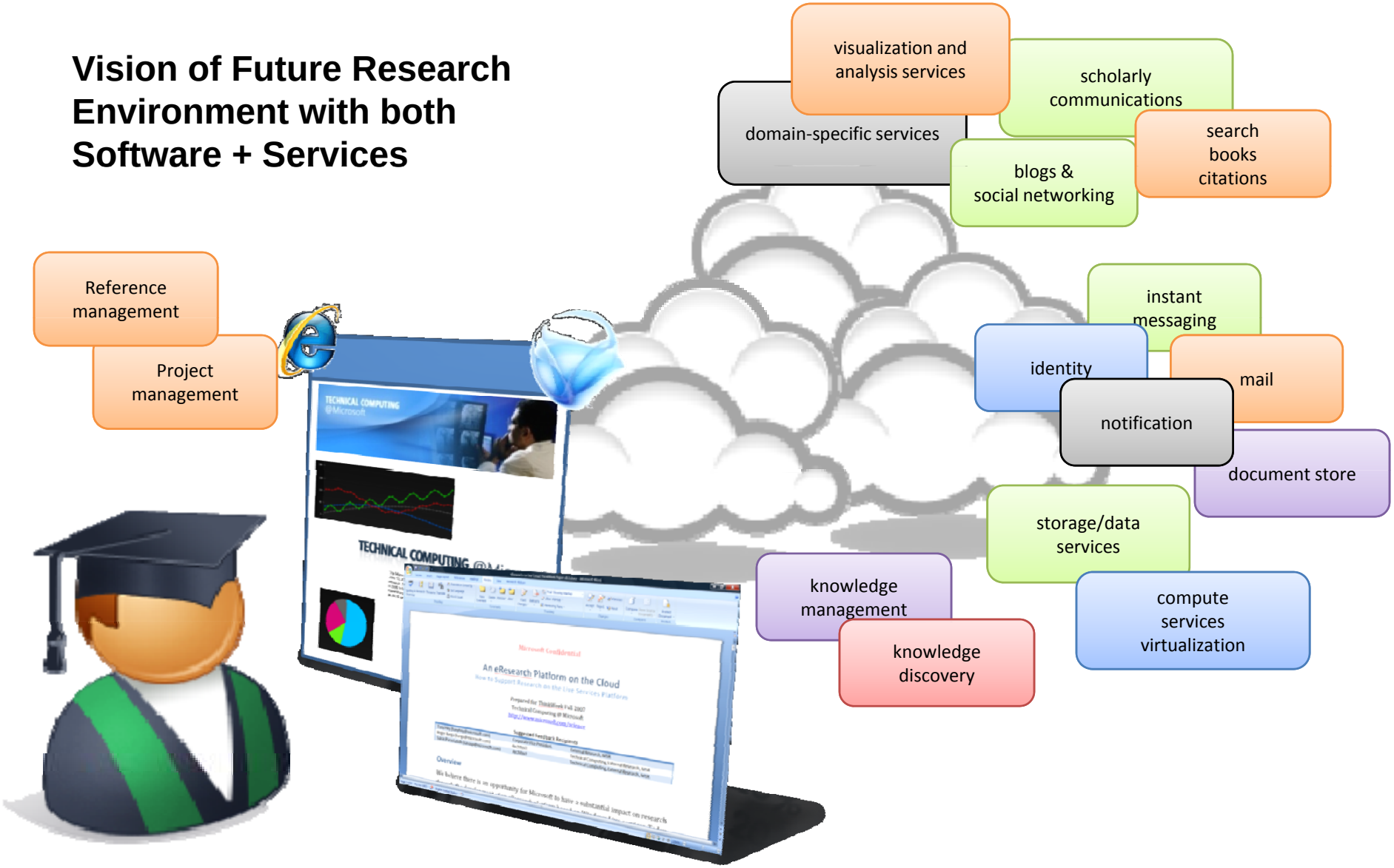
- **Important/key considerations**
  - Formats or “well-known” representations of data/information
  - Pervasive access protocols are key (e.g. HTTP)
  - Data/information is uniquely identified (e.g. URIs)
  - Links/associations between data/information




Attribution: Richard Cyganiak

# …and stored/processed/analyzed in the cloud

**Vision of Future Research Environment with both Software + Services**



# 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 for some more details of Microsoft's activities in Scientific and Technical Computing

Microsoft®

*Your potential. Our passion.™*