eScience
eInfrastructure and Semantic Computing

Tony Hey

Corporate Vice President

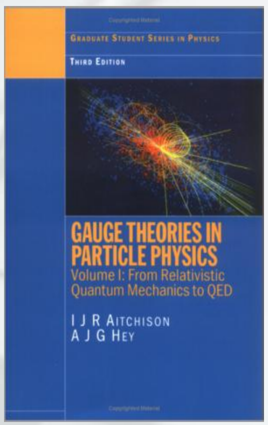
Microsoft Research

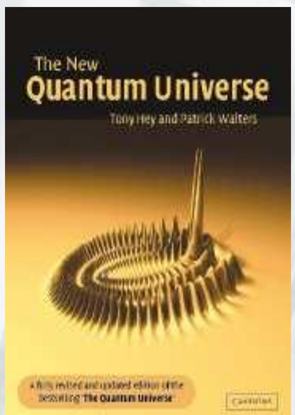
Tony Hey – My Background

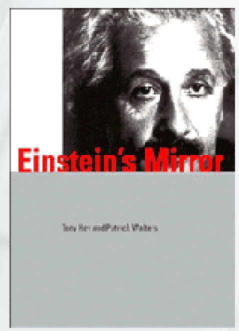
UNIVERSITY OF Southampton

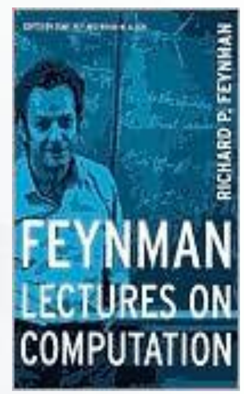
MPI
Message Passing Interface Standard

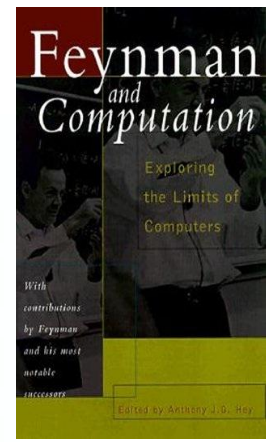
e-Science











The Royal Academy of Engineering

BCS

The

eScience 1.0

- In 2001, distributed computing technologies for eScience were in transition
  – Distributed authentication
  – CORBA and Web Services
- Over-emphasis on computation rather than data
  – Computational Grids difficult to use and too complex
  – Most communities do not want to install 100,000’s of lines of code before they can do anything
  – Grid standards not supported by industry

eScience 2.0

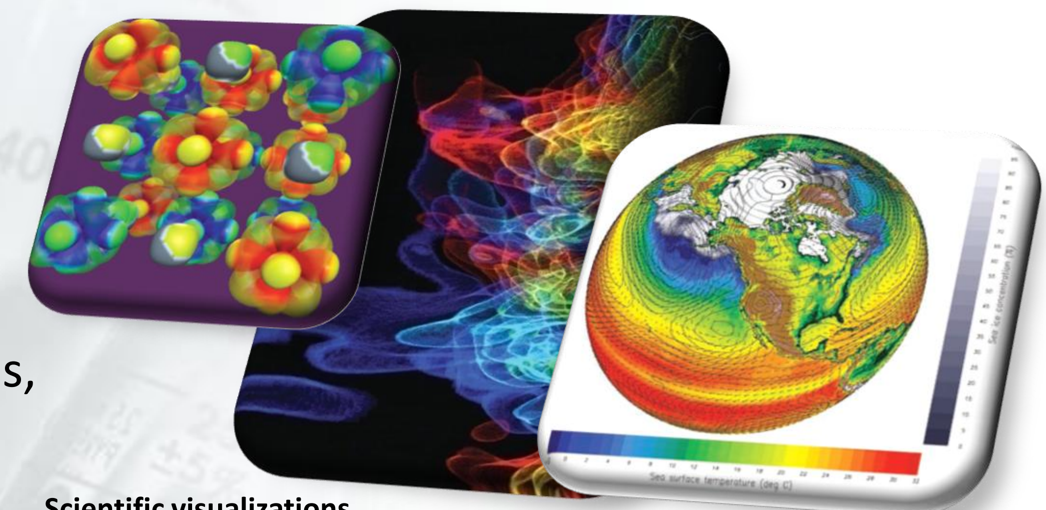
- Use Web 2.0 and the Web as a Platform
  - Simple protocols supported by industry
  - Blogs, Wikis, RSS feeds, Tagging, Mash-ups …
- Challenge for Computer Science community and the IT industry to deliver powerful and easy-to-use tools and technologies to support Data-Intensive research
  - Interoperability and open standards
  - Collaborative and multidisciplinary
  - Parallelism and Multicore
  - Client + Cloud: Software + Services

The Fourth Paradigm: Data-Intensive Science

A Digital Data Deluge in Research

- 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

The ‘Cosmic Genome’ Project

- The Sloan Digital Sky Survey is the first major astronomical survey project:
  - 5 color images of ¼ of the sky
  - Pictures of 300 million celestial objects
  - Distances to the closest 1 million galaxies
- Jim Gray from Microsoft Research worked with astronomer Alex Szalay to build the public ‘SkyServer’ archive for the survey

➢ New model of scientific publishing
  - Have to publish the data <u>before</u> astronomers publish their analysis

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**

World Wide Telescope

www.worldwidetelescope.org

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

- Integration of data sets and one-click contextual access
- Easy access and use

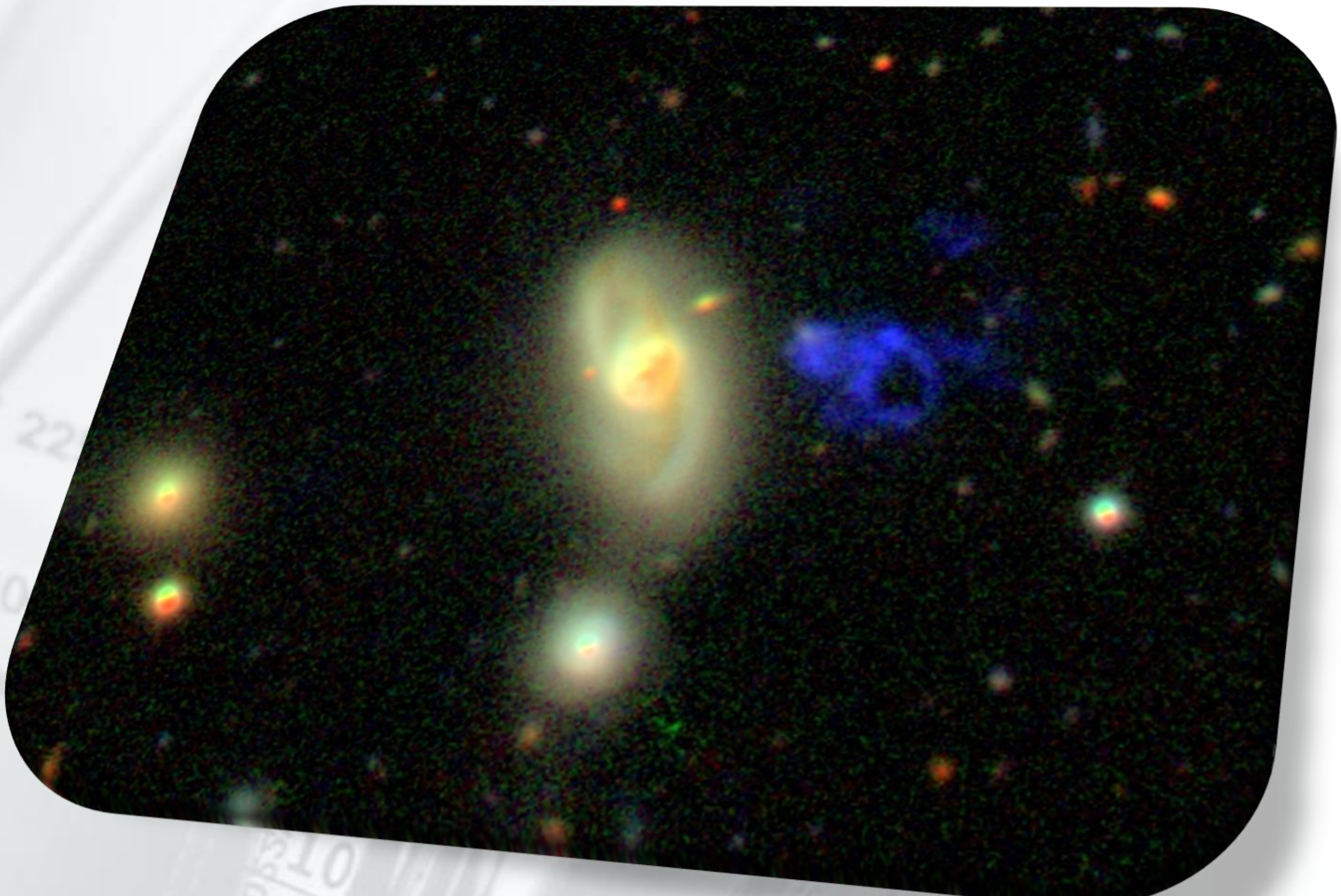
- As of 1/23/2009: 1,606,950 unique users (someone that has downloaded, installed, and successfully used WWT)
- There have been 4,089,898 sessions for an average of 2.55 sessions per user
- The average number of new users that have installed and used WWT has been 3,773 per day

Citizen Science: GalaxyZoo

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

➢ **Allows general public to search for photographs and classify different types of galaxies**

Hanny van Arkle's Voorwerp



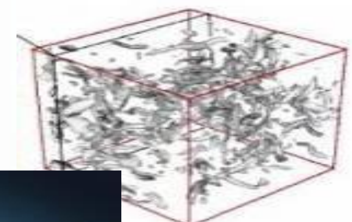
Emergence of a Fourth Research Paradigm

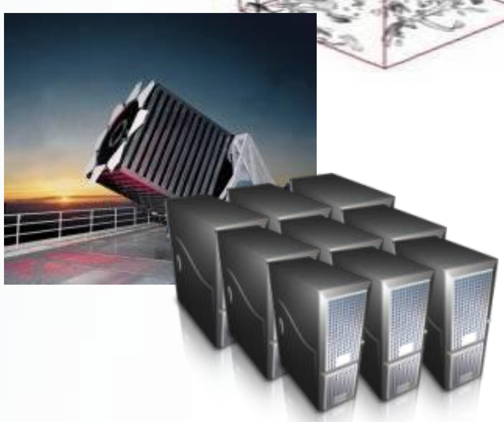
1. Thousand years ago – **Experimental Science**
   – Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   – Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   – Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   – Scientists overwhelmed with data sets from many different sources
      - Data captured by instruments
      - Data generated by simulations
      - Data generated by sensor networks

➢ eScience is the set of tools and technologies to support data federation and collaboration
   - For analysis and data mining
   - For data visualization and exploration
   - For scholarly communication and dissemination

(With thanks to Jim Gray)

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





eResearch and 'Small Data' Research Fields

eResearch

- In Australia use 'eResearch' in preference to eScience
- Social science and the humanities increasingly face many of the same challenges

'Small Data' Research Fields

- Fields such as chemistry, condensed matter physics or ecology do not have Petabytes but still major data management challenges
- 'Born digital' data in files, spreadsheets or databases stored on hard drives, digital notebooks, Web sites, blogs and wikis
- Management, curation and archiving of these digital increasingly burdensome

External Collaborative Research: Worldwide "Themes"

Core Computer Science

Earth, Energy and Environment

Education & Scholarly Communications

Health & Wellbeing

Advanced Research Tools and Services

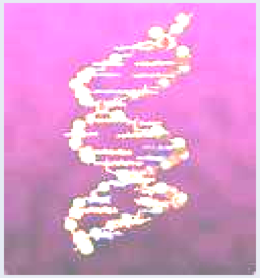
External Research Global Themes

Core Computer Science

Earth, Energy & Environment

Education & Scholarly Communications

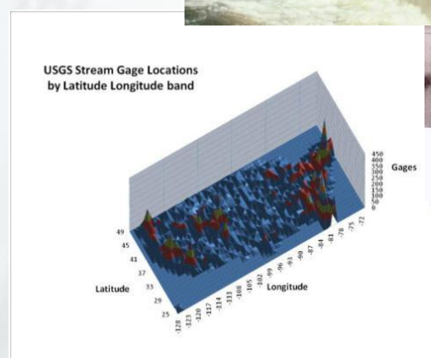
Health & Wellbeing

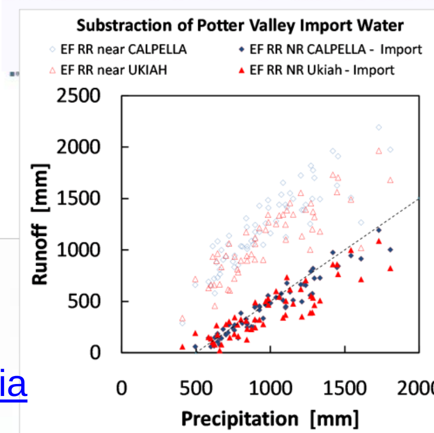


- **Visualizing and Experiencing E^3 Data + Information:** Provide a unique experience to reduce time to insight and knowledge through visualizing data and information
- **Accessible Data:** Ensure E^3 data (remote and local sensing) is easily accessible and consumable in the scientists domain

Digital Watersheds (BWC, James Hunt)

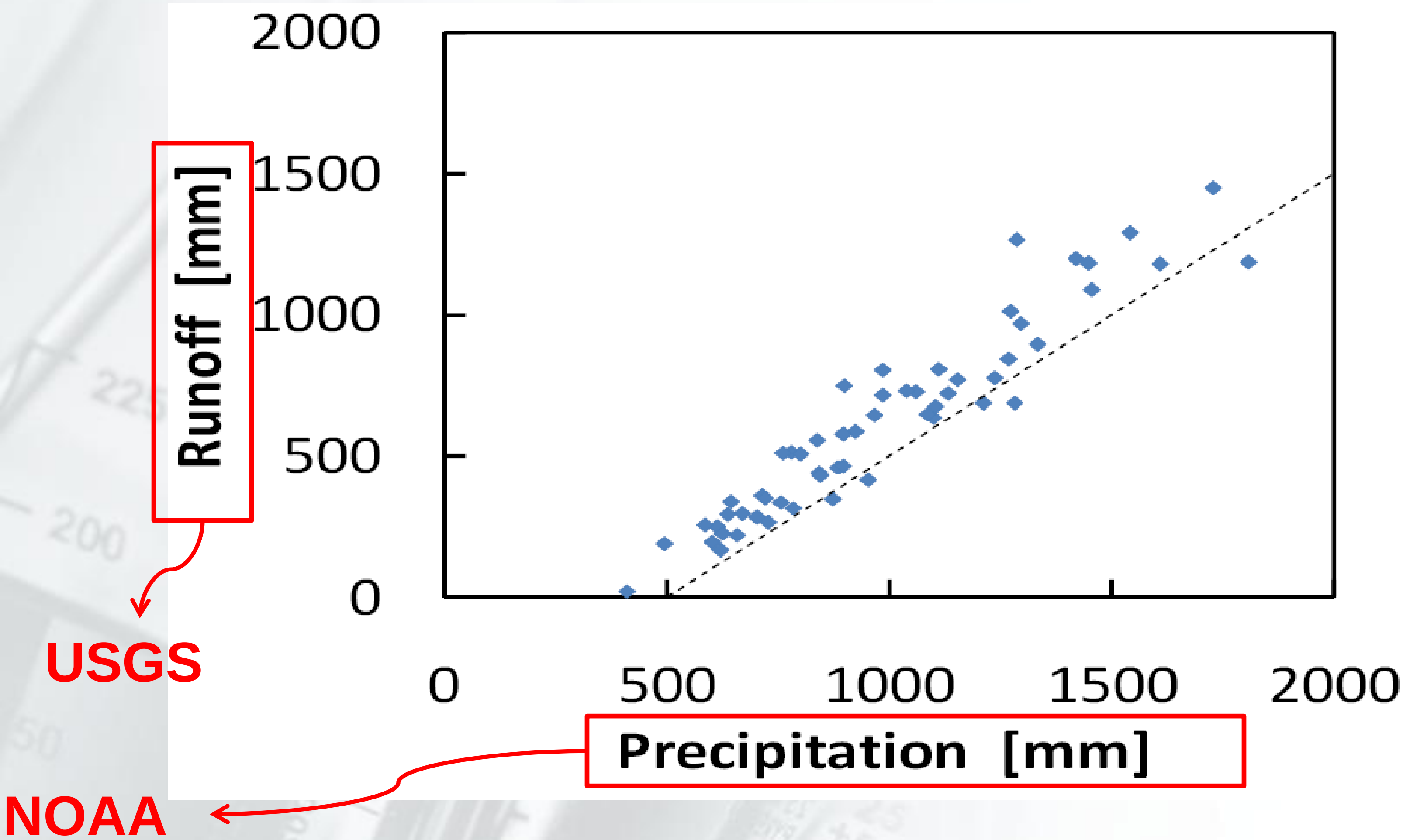
- Russian River watershed challenges: forestry, farming, urbanization, gravel mining, and fish habitat restoration.
  - Can we understand historic and on-going changes using only publically available data sources such as USGS, NOAA, Sonoma Ecology Center, etc?
- Early studies examined overall water balance and changes in suspended sediment
  - scientific data “mashups” are leading to new and useful results.
- Recent engagement with National Marine Fisheries and USBR expanding this to other watersheds across Northern http://www-esd.lbl.gov/BWC/ California
- Sharing technology with CUAHSI (100 universities)

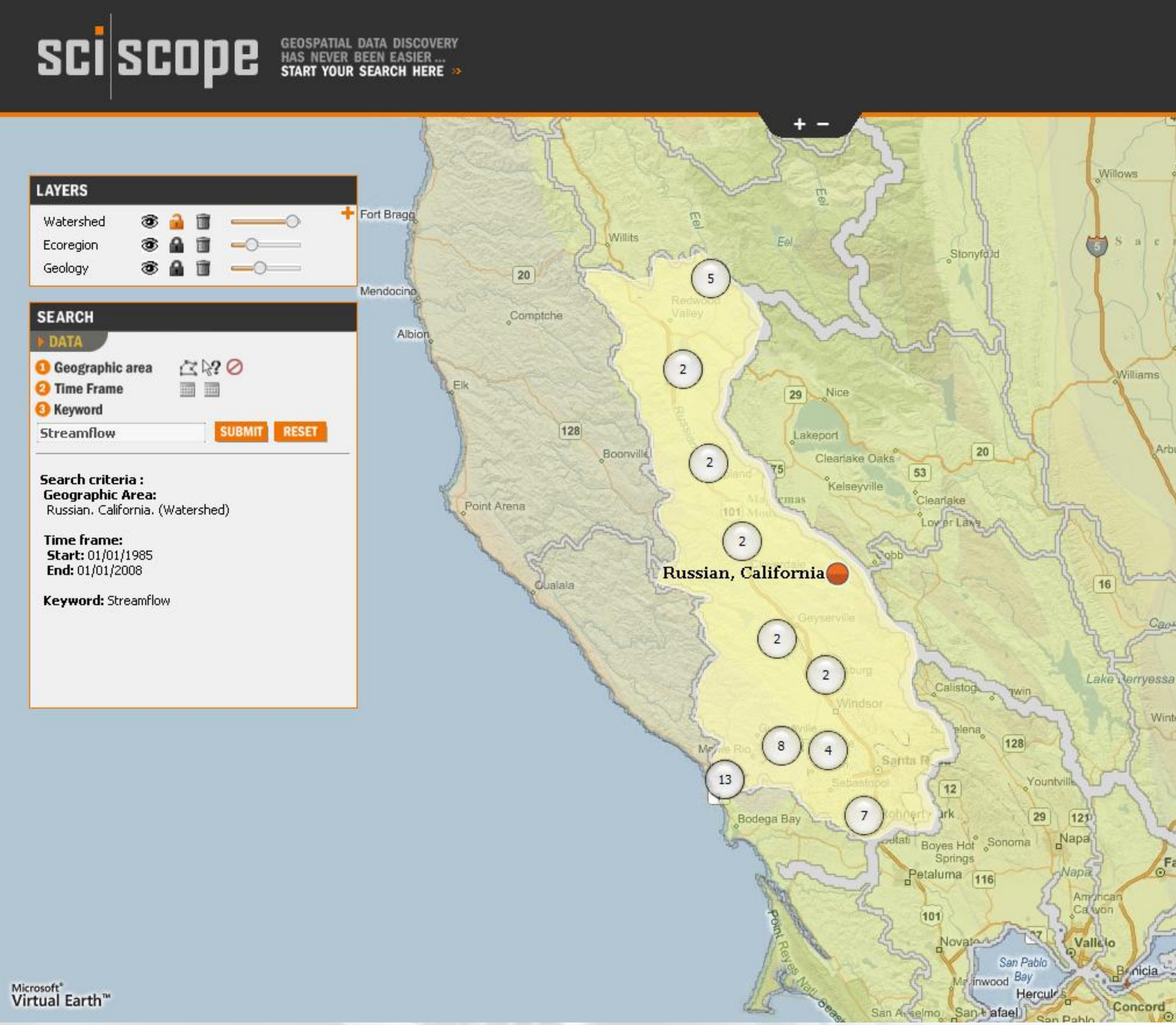


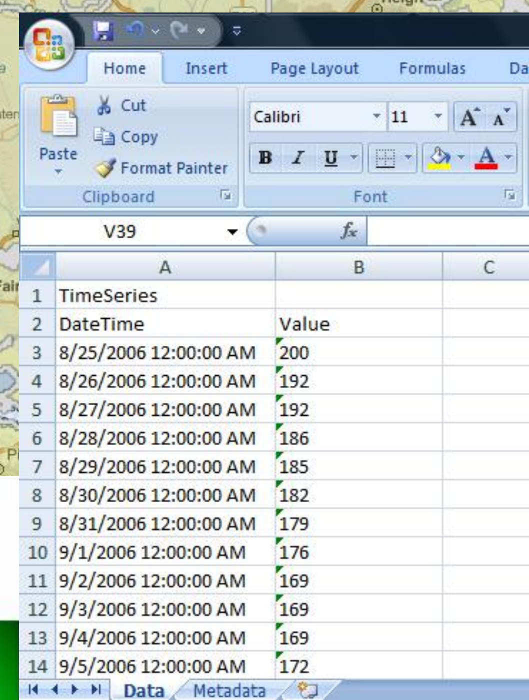


http://bwc.berkeley.edu
http://www-esd.lbl.gov/BWC/California
http://www.cuahsi.org/

Data from a variety of sources



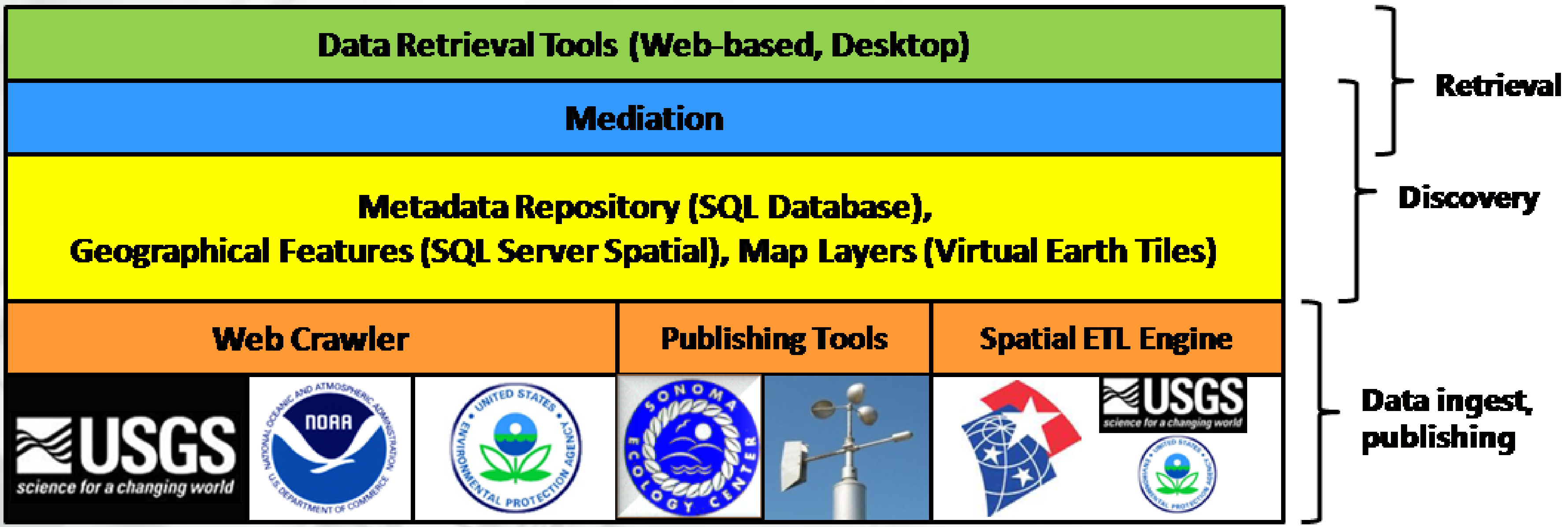




Automated download

- Add to favorites
- Download data

SciScope Stack



Knowledge base

- Relationships are stored as RDF triples in a relational database

  ‘Escherichia coli’ = ‘E. coli’
  ‘E. coli’ is-a ‘Indicator Organism’

  ‘Nitrogen’ is-a ‘Macronutrient’
  ‘Macronutrient’ is-a ‘Nutrient’
  ‘Hypoxia’ isMeasuredUsing ‘DissolvedOxygen’
  ‘Hypoxia’ isRelatedTo ‘Eutrophication’

- Supports transitive, symmetric and inverse properties
- Inferred statements are pre-computed

External Research Global Themes

Core Computer Science

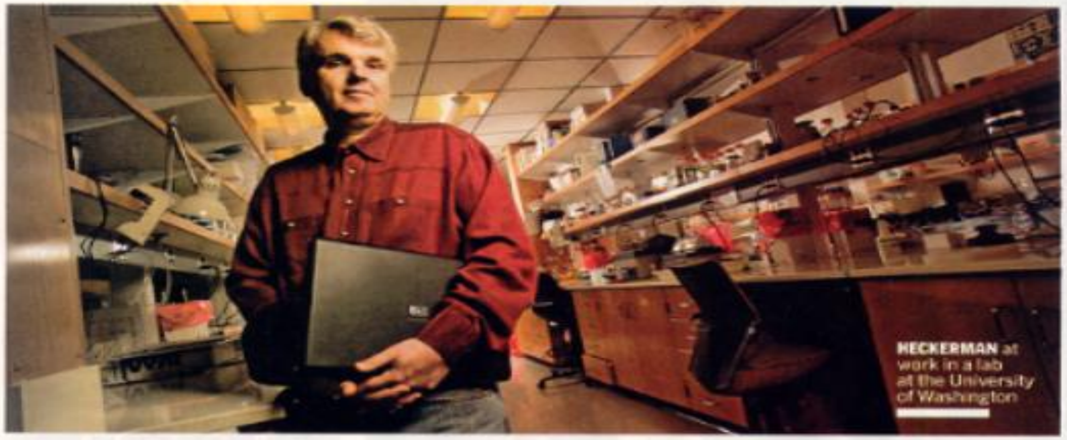
Earth, Energy & Environment

Education & Scholarly Communications

Health & Wellbeing

- **Devices, Sensors and Mobility**: Cellphone as a platform for healthcare;

  Proof points for the value of new modes of interaction with health data
- **Genomics in Healthcare:** research and tools in genomics



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.

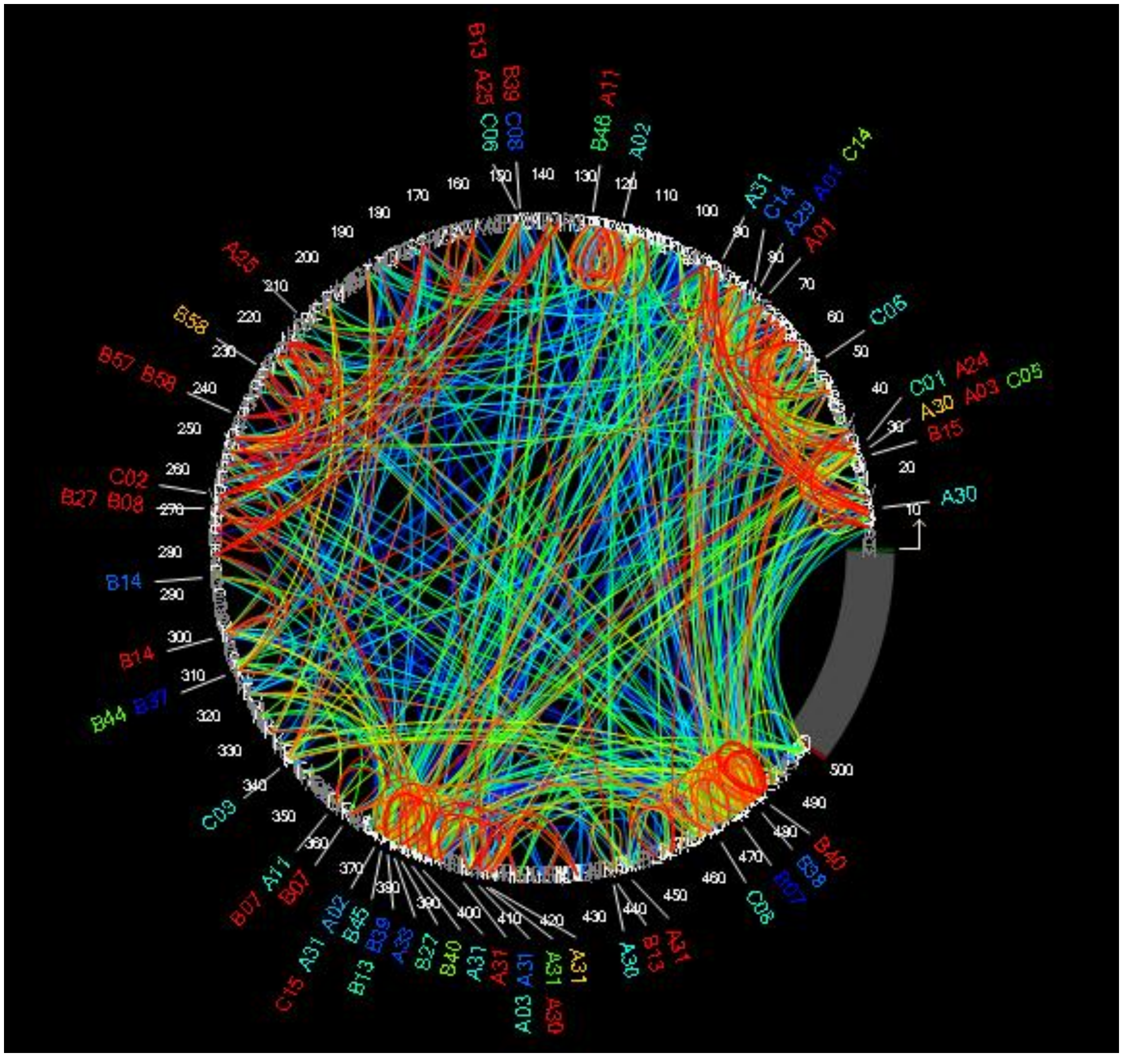
Similar mutations may crop up in computer and medical viruses

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 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

We do this in part by tracking the evolution of HIV inside an individual using advanced machine-learning algorithms



Results to date …

- **Discovered 'decoy epitopes'** that could have predicted recent failure of Merck vaccine
  - ➢ Verified hypothesis on Merck data
- **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)

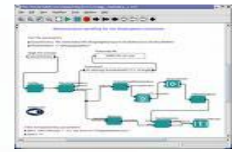
eScience: Supporting researchers

Adding Semantics to Software Tools

Research Pipeline

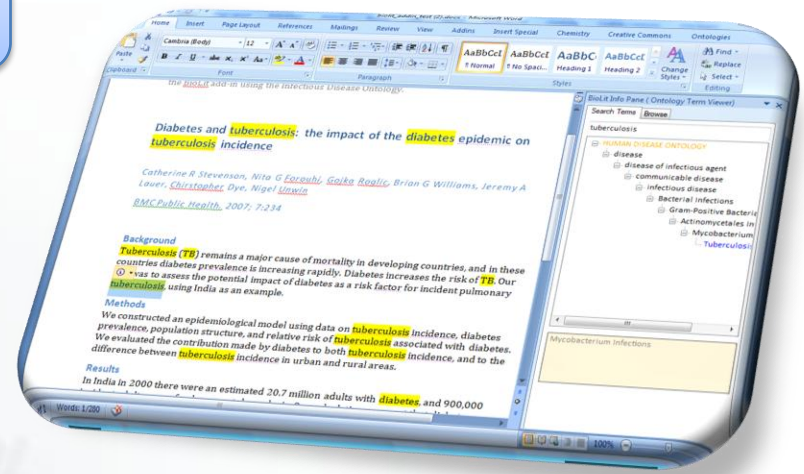
Data Acquisition and Modeling → Collaboration and Visualization → Analysis and Data Mining → Disseminate and Share → Archiving and Preservation

- 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



Ontology Add-in for Word

- Phil Bourne and Lynn Fink, UCSD



Goals

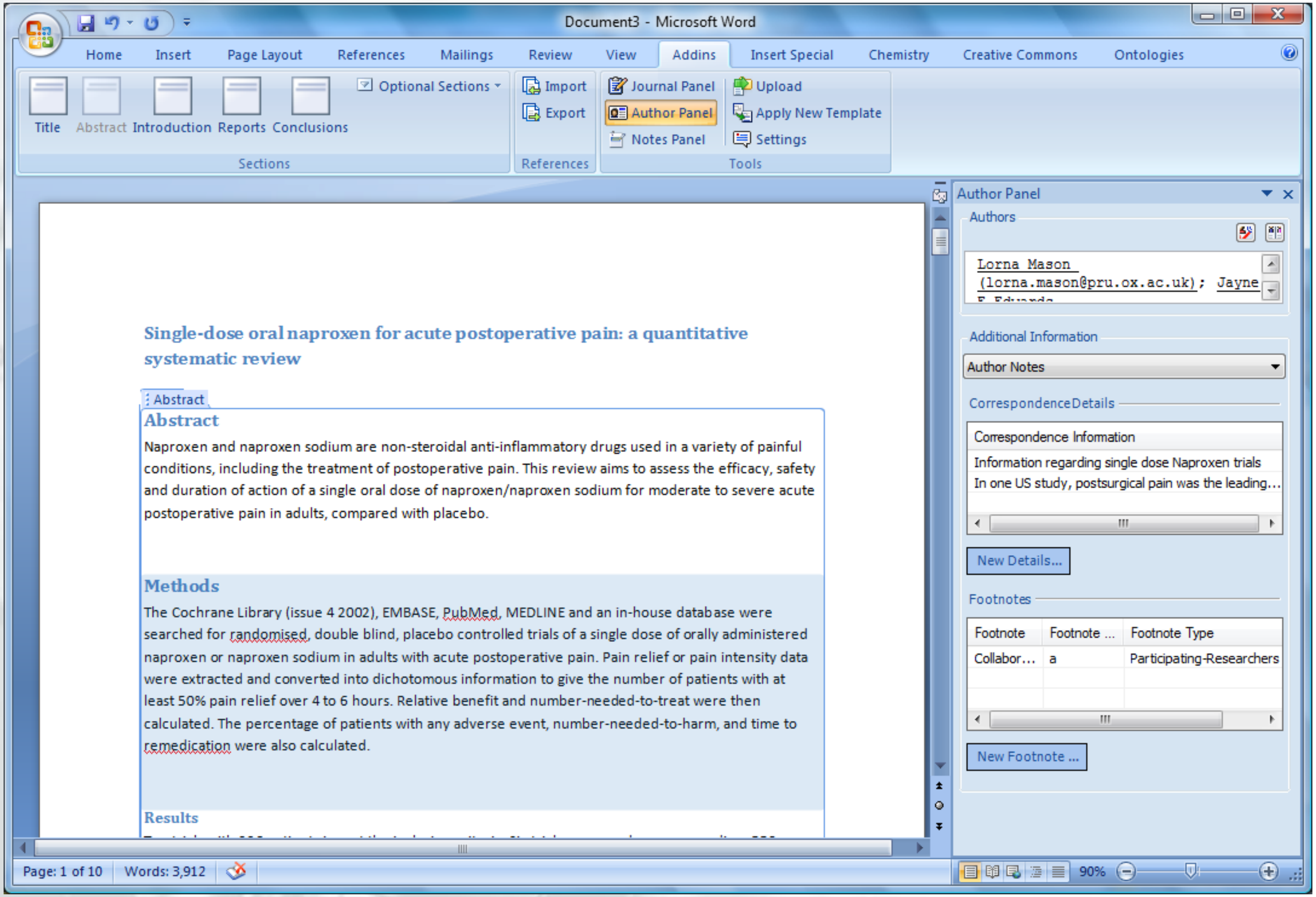
- Semantic mark-up using ontologies and controlled vocabularies
- Facilitate/automate referencing to PDB (and other resources) from manuscript
- Conversion of manuscript to NLM DTD for direct submission to publisher

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

Data Acquisition & Modeling | Collaboration and Visualization | Analysis and Data Mining | Disseminate & Share | Archiving and Preservation

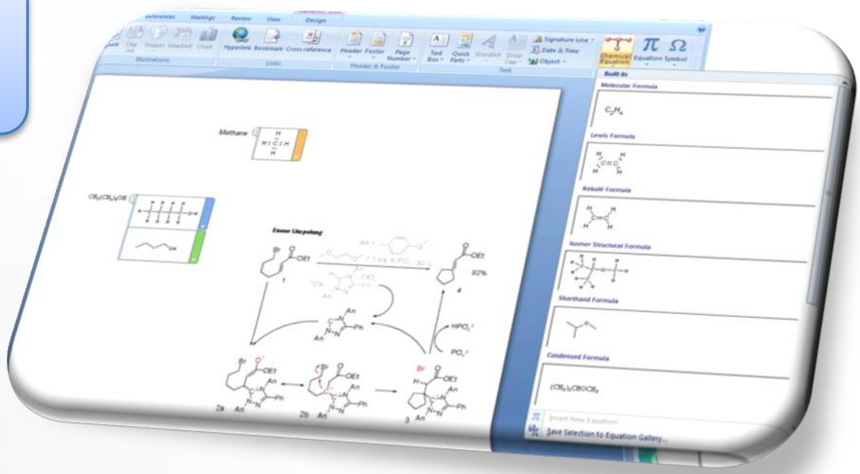
Article Authoring Add-in for Word 2007



Data Acquisition and Modeling | Collaboration and Visualization | Analysis and Data Mining | Disseminate and Share | Archiving and Preservation

Chem4Word
Chemistry Drawing in Word

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

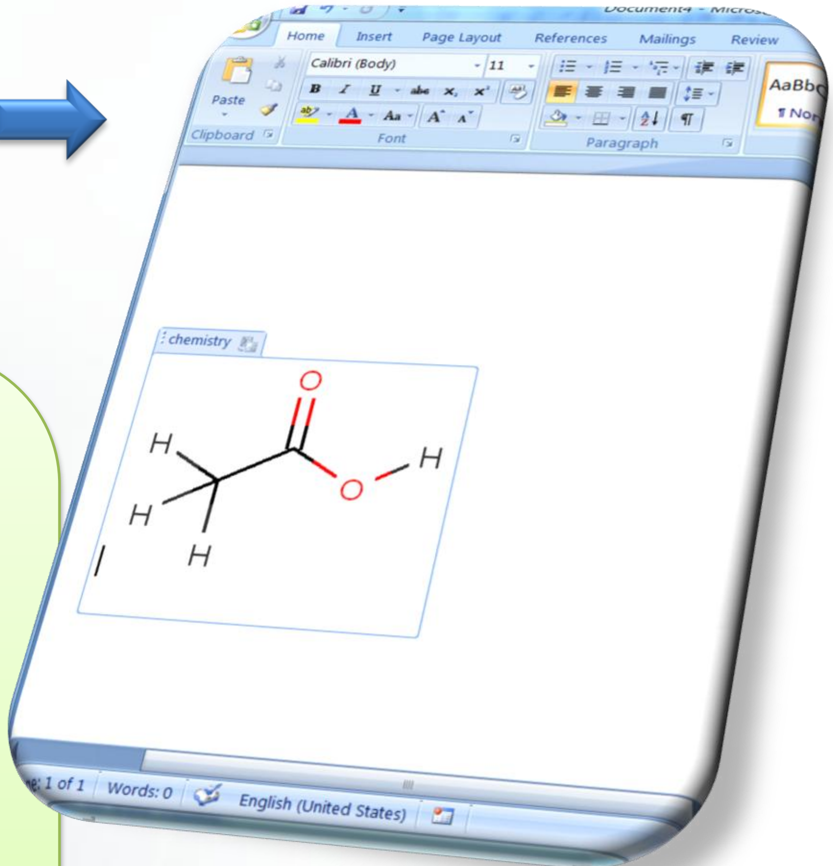


Goals
- Support students/researchers in simple chemistry structure authoring/editing
- Enable ecosystem of tools around lifecycle of chemistry-related scholarly works
- Store chemical semantics in Chemistry Markup Language
- Extensible via built-in CML dictionaries, Chemical Styles, and web-services
- Release to community as open source project

Execution
- Coordinated development in Cambridge and in Redmond
- Post-doc in Cambridge to use plug-in and give feedback and migrate existing chemical intelligence tools to .NET
- Advanced Reading Technologies to create necessary glyphs

Chem4Word Example

Molecule added in Word*



CML stored in DOCX container


```xml
<?xml version="1.0" ?>
<cml version="3" convention="org-synth-report" xmlns="http://www.xml-cml.org/schema">
 <molecule id="m1">
  <atomArray>
   <atom id="a1" elementType="C" x2="-2.9149999618530273" y2="0.7699999809265137" />
   <atom id="a2" elementType="C" x2="-1.5813208400249916" y2="1.5399999809265137" />
   <atom id="a3" elementType="O" x2="-0.24764171819695613" y2="0.7699999809265134" />
   <atom id="a4" elementType="O" x2="-1.5813208400249912" y2="3.0799999809265137" />
   <atom id="a5" elementType="H" x2="-4.248679083681063" y2="1.5399999809265137" />
   <atom id="a6" elementType="H" x2="-2.914999961853028" y2="-0.7700000190734864" />
   <atom id="a7" elementType="H" x2="-4.248679083681063" y2="-1.907348645691087E-8" />
   <atom id="a8" elementType="H" x2="1.0860374036310796" y2="1.5399999809265132" />
  </atomArray>
  <bondArray>
   <bond atomRefs2="a1 a2" order="1" />
   <bond atomRefs2="a2 a3" order="1" />
   <bond atomRefs2="a2 a4" order="2" />
   <bond atomRefs2="a1 a5" order="1" />
   <bond atomRefs2="a1 a6" order="1" />
   <bond atomRefs2="a1 a7" order="1" />
   <bond atomRefs2="a3 a8" order="1" />
  </bondArray>
 </molecule>
</cml>
```


* This is a screenshot from a very early prototype. We are actively working on improving the quality of the rendering.

Data Acquisition and Modeling | Collaboration and Visualization | Analysis and Data Mining | Disseminate and Share | Archiving and Preservation

PLANETS

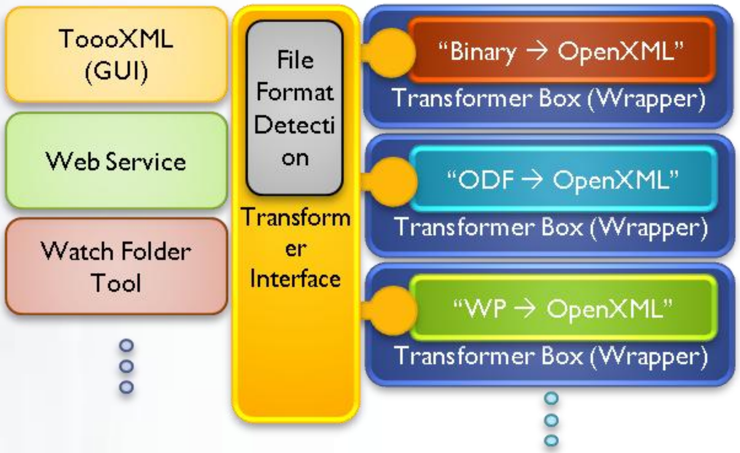
Tools and methods for sustainable long-term preservation of digital objects

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



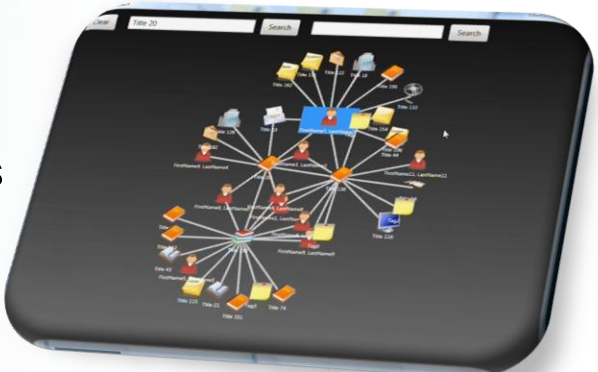
Research Output Repository Platform

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

- Utilizing OAI-ORE, SWORD, and other community protocols
- In development, deployment within MSR in early Q4
- Release to the community in late Q4
- Built on SQL Server 2008 + Entity Framework
  - Using WPF and Silverlight for UI

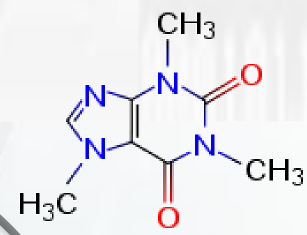
oreChem – *The Chemical Semantic Web*

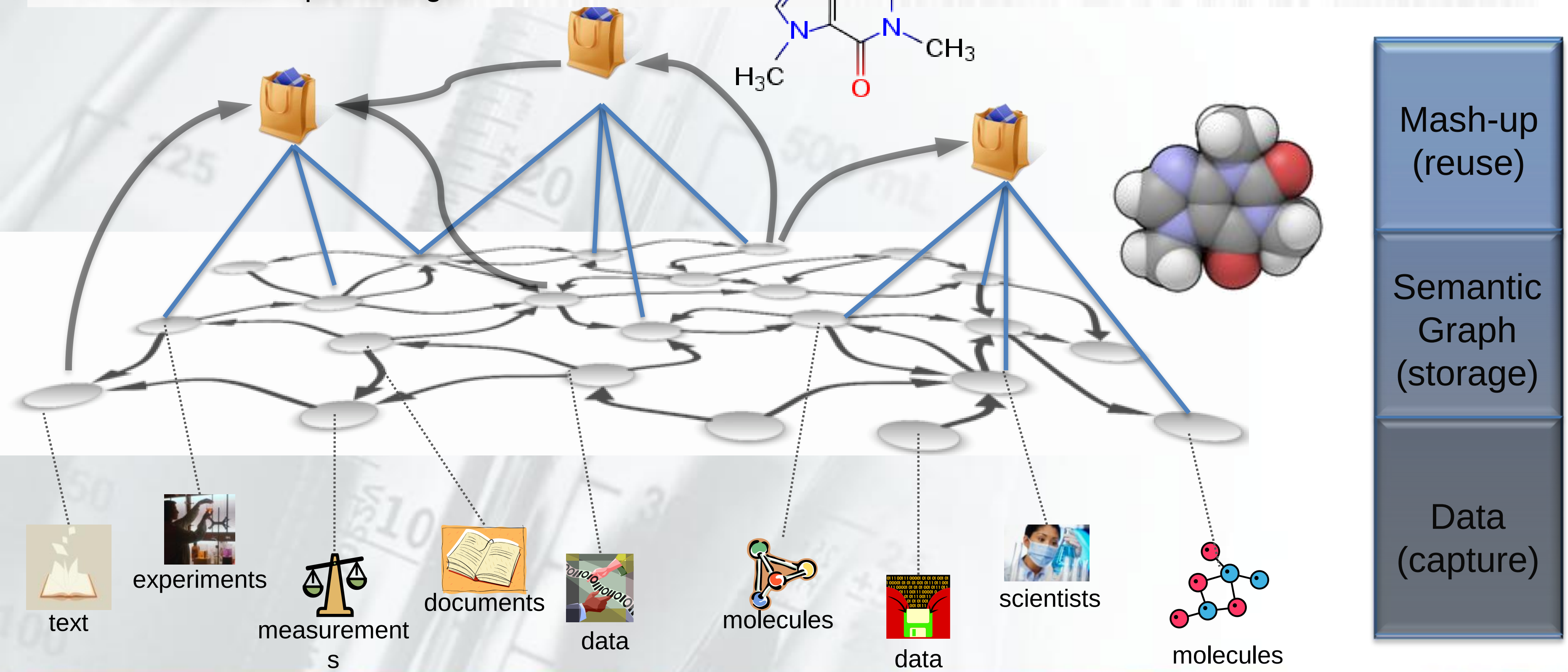
supported by Microsoft

- Large collaboration project focusing on interoperability
- At-source capture of chemistry data
- Chemical structure search
- Compound object authoring
- Retrospective harvesting of chemistry data
- Reuse through common ORE data model
- Semantic authoring
- Virtualized triple storage

Participating Institutions:

- University of Cambridge
- Cornell University
- Indiana University
- Penn State University
- University of Queensland
- University of Southampton





Client + Cloud Computing for Science

www.smugmug.com

SmugMug

Home | Login | Help | Search

Devoted to priceless photos.

Most Internet companies dream of selling to bigger ones, and getting rich.

We don't.

Living a dream.

We dream of an independent company devoted to nothing but your priceless photos.

A company that backs up your photos to three data centers across the U.S.

A profitable, debt-free company.

That earns your fanatical loyalty.

We're living that dream.

Photo by Dennis T. Dease.

Details, details.

36 employees. More than 300,000 paying customers. 372,720,004 photos and counting.

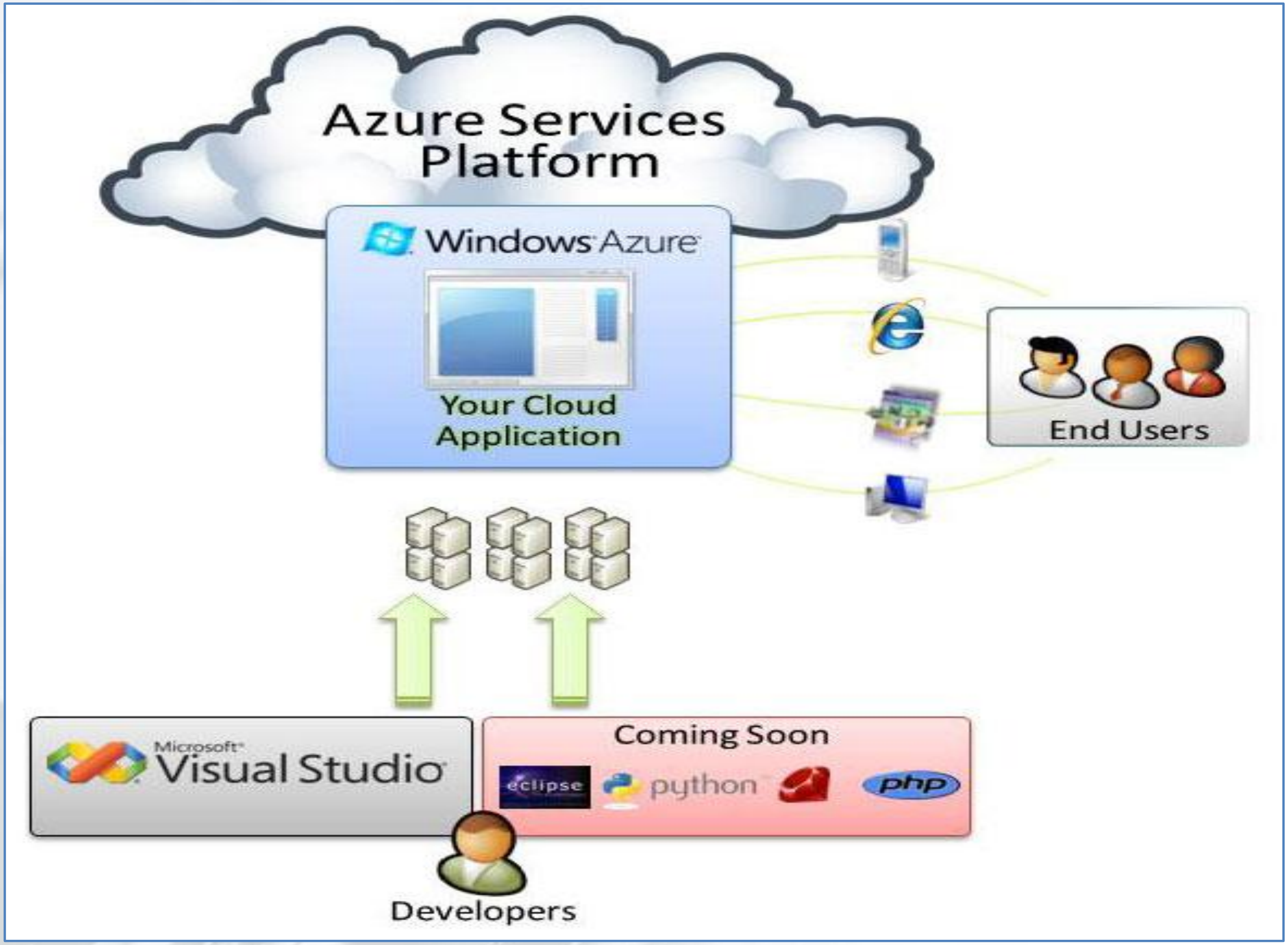
We'll always be smaller than the photo-sharing divisions of giant companies.

Which is a very good thing.

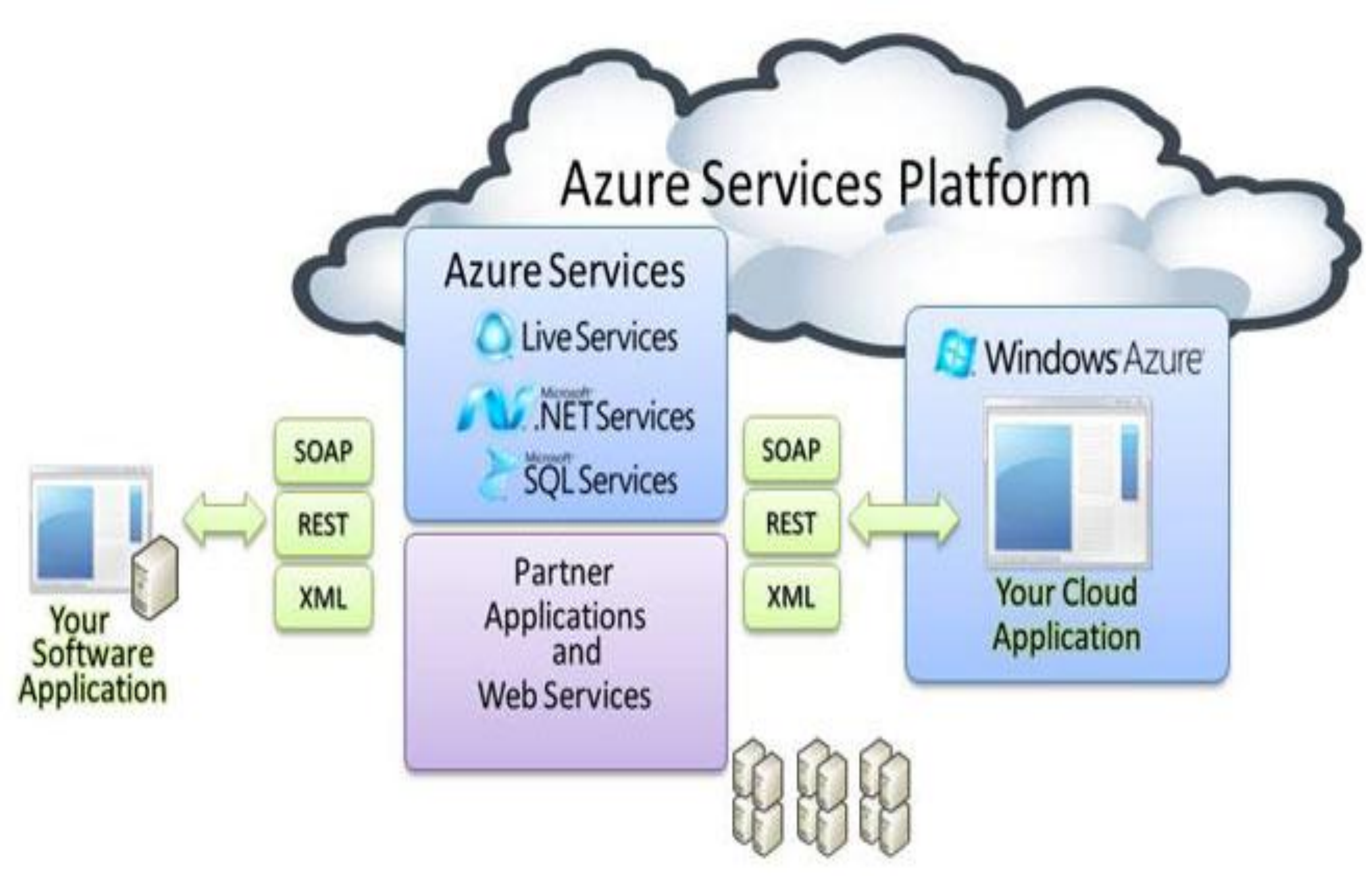
Our story.

News | Browse | Keywords | Communities | Forum | Wiki | ClubSmug | Prints & Gifts | Shopping Cart | Login
Terms | Privacy | About Us | Contact SmugMug | Blogs | API | Affiliates |

Windows Azure
An Operating System for the Cloud (1)



Windows Azure
An Operating System for the Cloud (2)



Two Cloud Science Examples

- Virtual Research Environments
- Oceanography Work Bench

Researchers Information Centre

British Library for Research

A one stop solution for carrying out **research studies** in planned & phased manner and **networking** with fellow community members

Plan The Research
Search for study ideas, plan the study, and apply for funding.

Network
Connect with fellow researchers for sharing ideas, resources etc.

Experiment
Use online tools to achieve faster results.

Publish
Disseminate the study results for the public.

Existing RIC Members

Username:

Password:

Remember Me

Login

Forgot your ID or Password?

New to RIC?

Sign Up

Currently in beta evaluation, directed by The British Library.

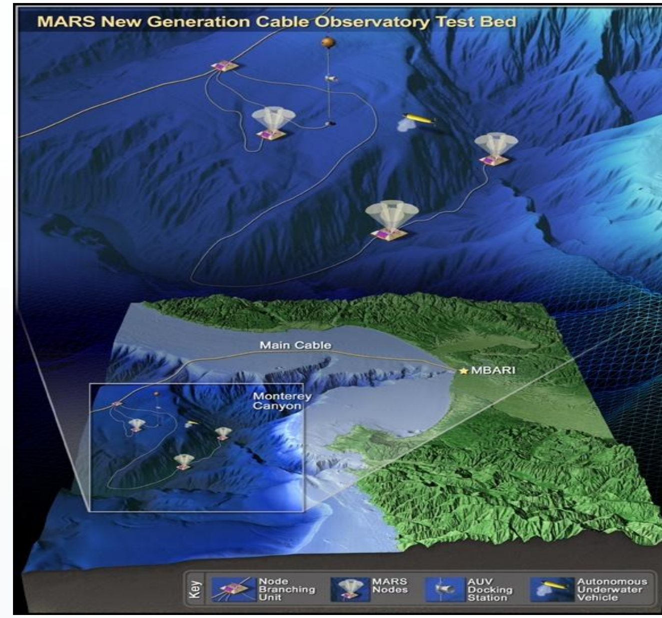
Microsoft Online Services

- Exchange, Sharepoint, Live Meeting as on Online Service
- No need to build your own infrastructure or maintain and manage servers
- Moving forward, science-related services could move easily to the Cloud (e.g. RIC with British Library)

http://www.microsoft.com/online/

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

Microsoft SQL Services

- "Hosted" SQL Server functionality
- Structured data, structured queries
- On-demand scalability
- Service-Level Agreements
  - High availability, performance, fault-tolerance
- Programmability
  - An easy-to-use programming API (SOAP and REST)

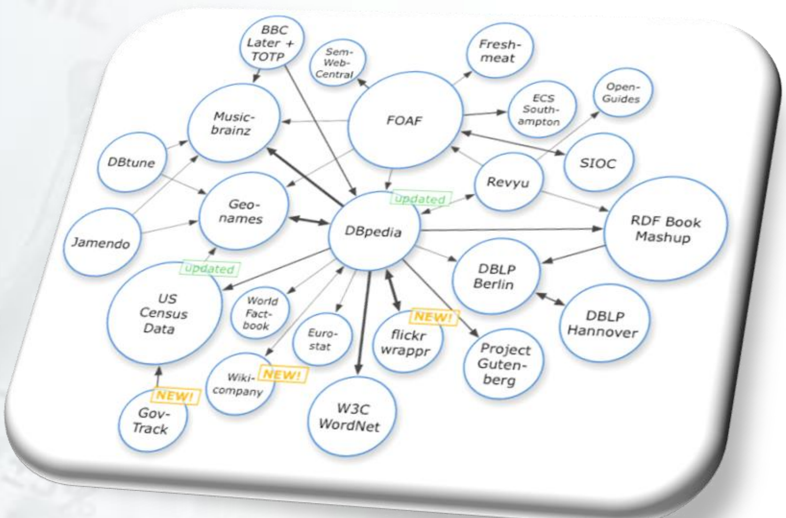
Microsoft® SQL Server® Data Services
Your Data, Any Place, Any Time

http://www.microsoft.com/sql/dataservices/

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’

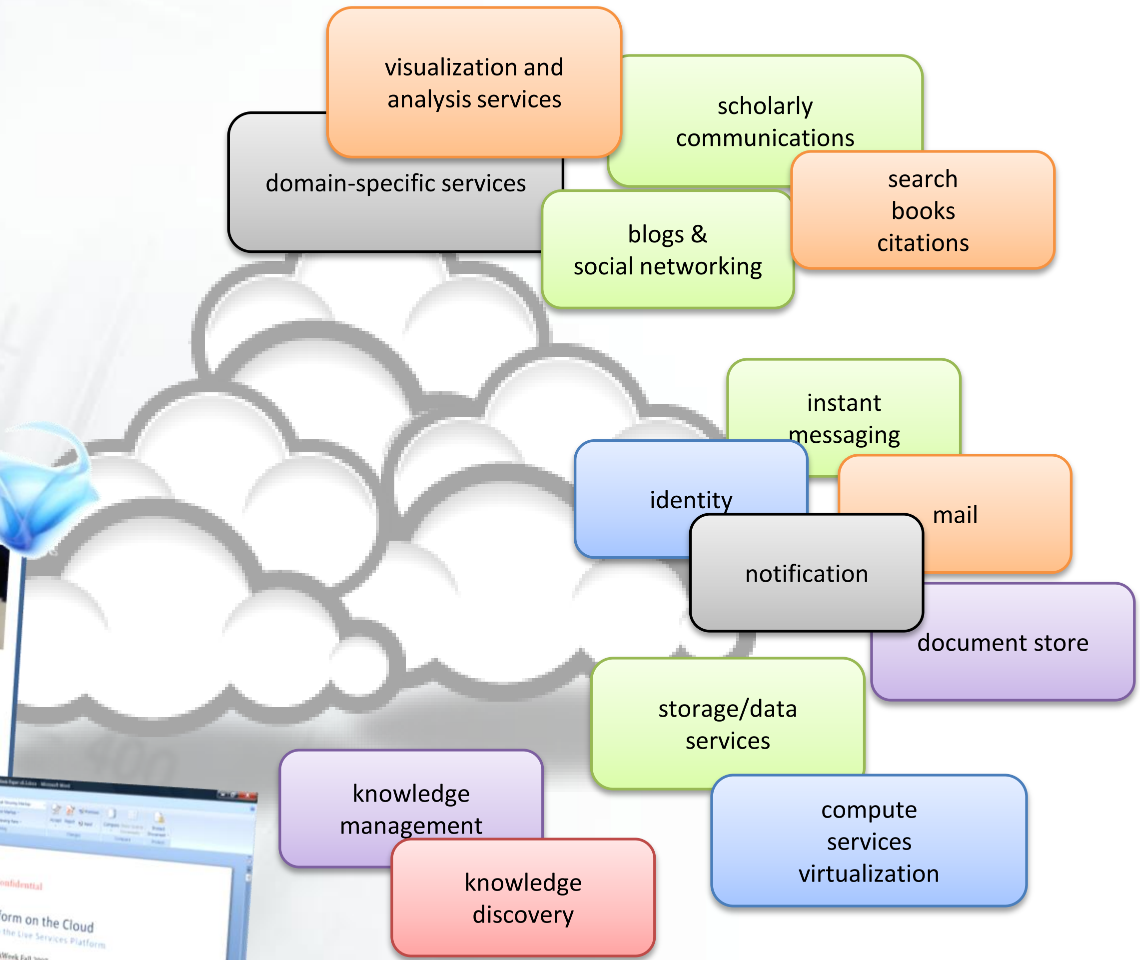
- **Important/key considerations**
  - Formats or “well-known” representations of data/information
  - Pervasive access protocols are key (e.g. HTTP)
  - Data/information is uniquely identified (e.g. URIs)
  - Links/associations between data/information

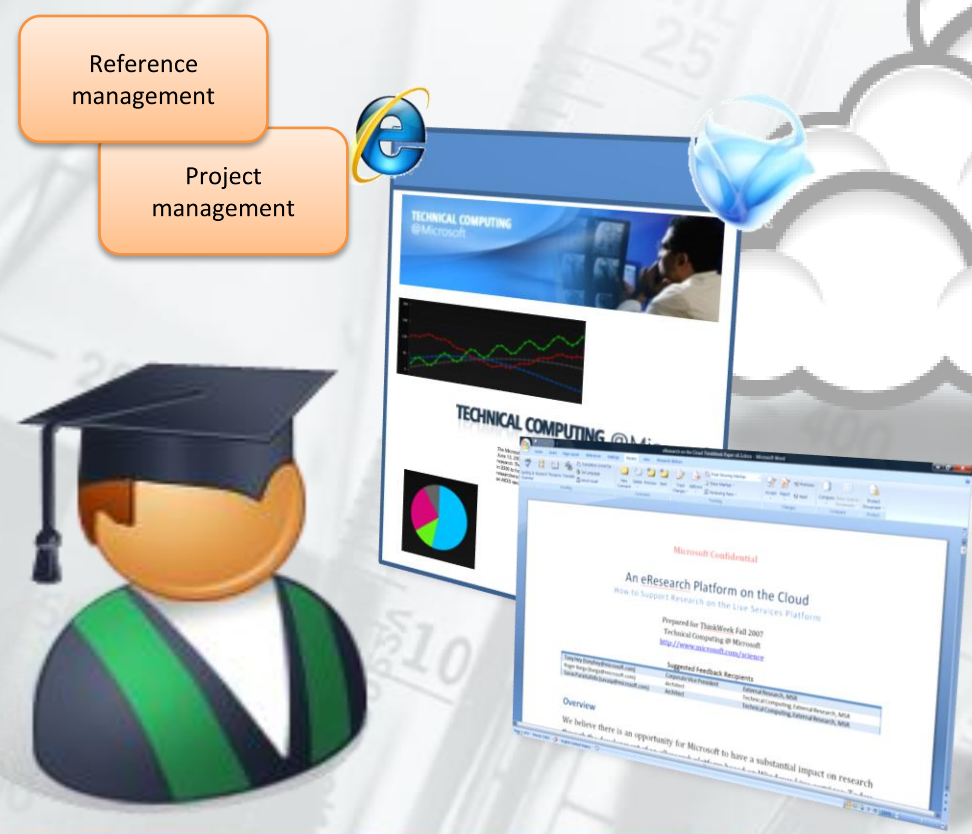


Attribution: [Richard Cyganiak]

…and stored/processed/analyzed in the cloud

Vision of Future Research Environment with both Software + Services





Advanced Research and Tools

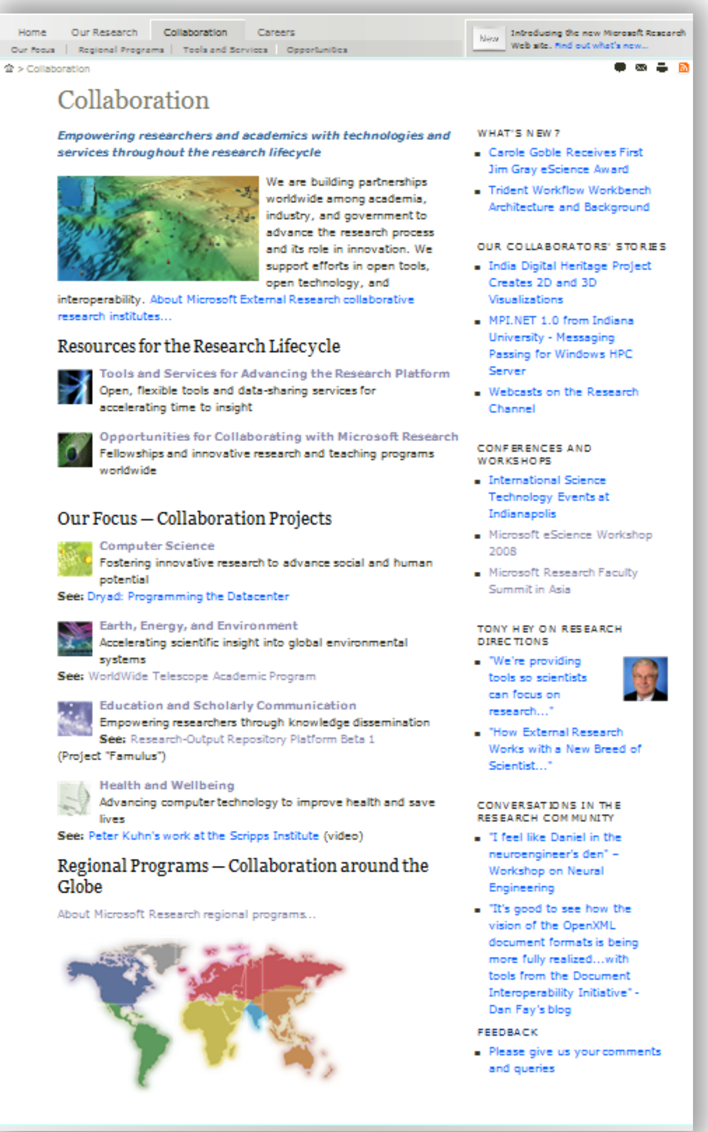
- Accelerate research by collaborating with academic communities to create open tools and services based on Microsoft platforms and productivity software.
- Build open software solutions in collaboration with the research community
- Help scientists spend more time on their research and less time on IT issues.

- **Tools and Services for Research Collaboration**
  - Tools for Computational Science
  - Database, HPC, and Cloud Services for Scientists
  - Managing Scientific Workflows
  - Microsoft Office Addins for Scientists
  - Publication and Conference Management for Scientists
  - Software Tools for Computational Biology
  - Tools for Web Search Research

Further Information and Resources

http://research.microsoft.com

- The site contains access and downloads of relevant tools and resources for the worldwide academic research community. A small set of examples include:
  - **Research Output Repository**: building blocks, tools, and services for developers who are tasked with creating and maintaining an organization's repository ecosystem. http://research.microsoft.com/en-us/downloads/48e60ac1-a95a-4163-a23d-28a914007743/
  - **Tools and Services for Research Collaboration**: http://research.microsoft.com/en-us/collaboration/tools/default.aspx
  - **WorldWide Telescope Academic Program**: http://research.microsoft.com/en-us/collaboration/wwt-ap.aspx



Microsoft®

Your potential. Our passion.™