# The Fourth Paradigm: Data-Intensive Scientific Discovery

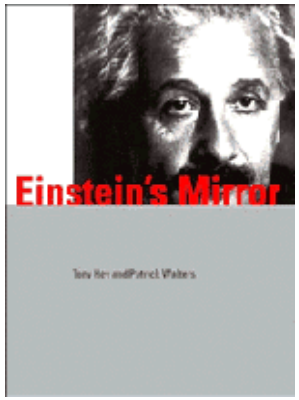
Tony Hey

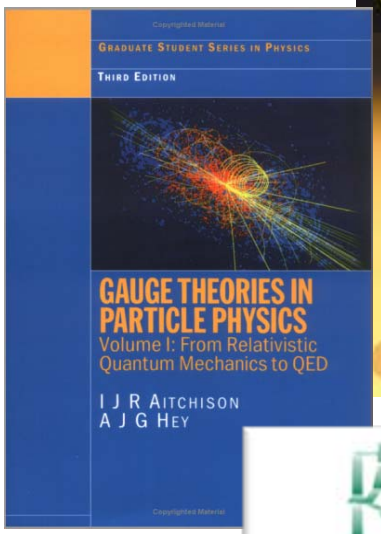
Corporate Vice President

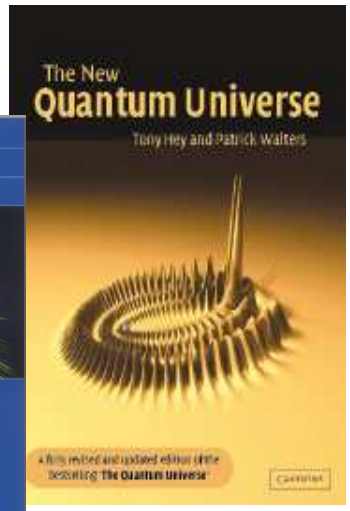
Microsoft External Research

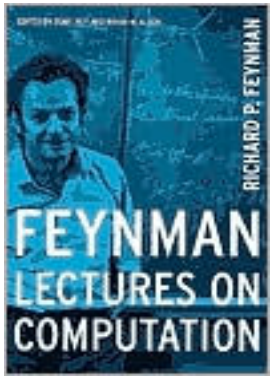
# Tony Hey – An Introduction

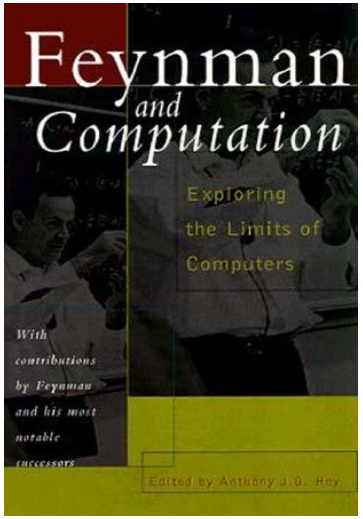

UNIVERSITY OF Southampton



MPI
Message Passing Interface Standard

The Royal Academy of Engineering

BCS


Commander of the British Empire



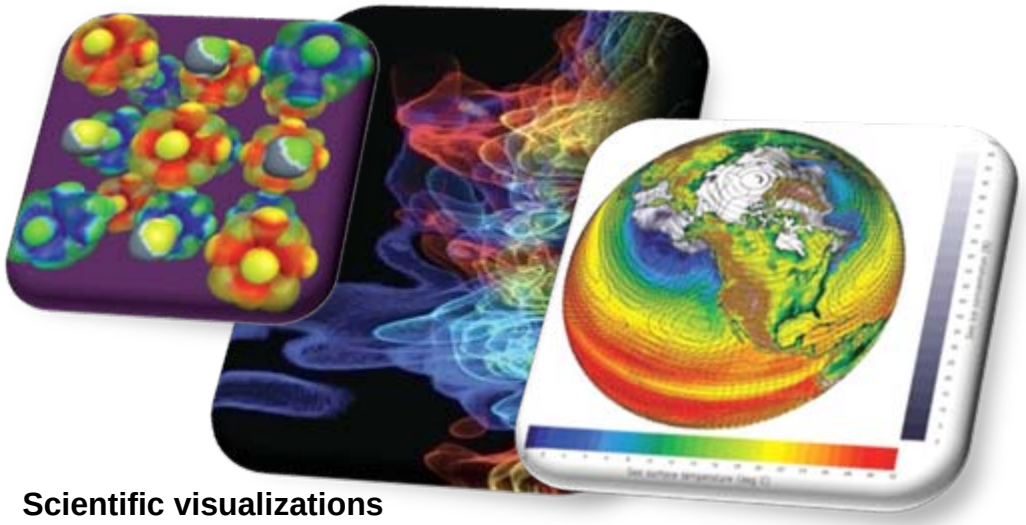
# The Fourth Paradigm

# 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



# Emergence of a Fourth Research Paradigm

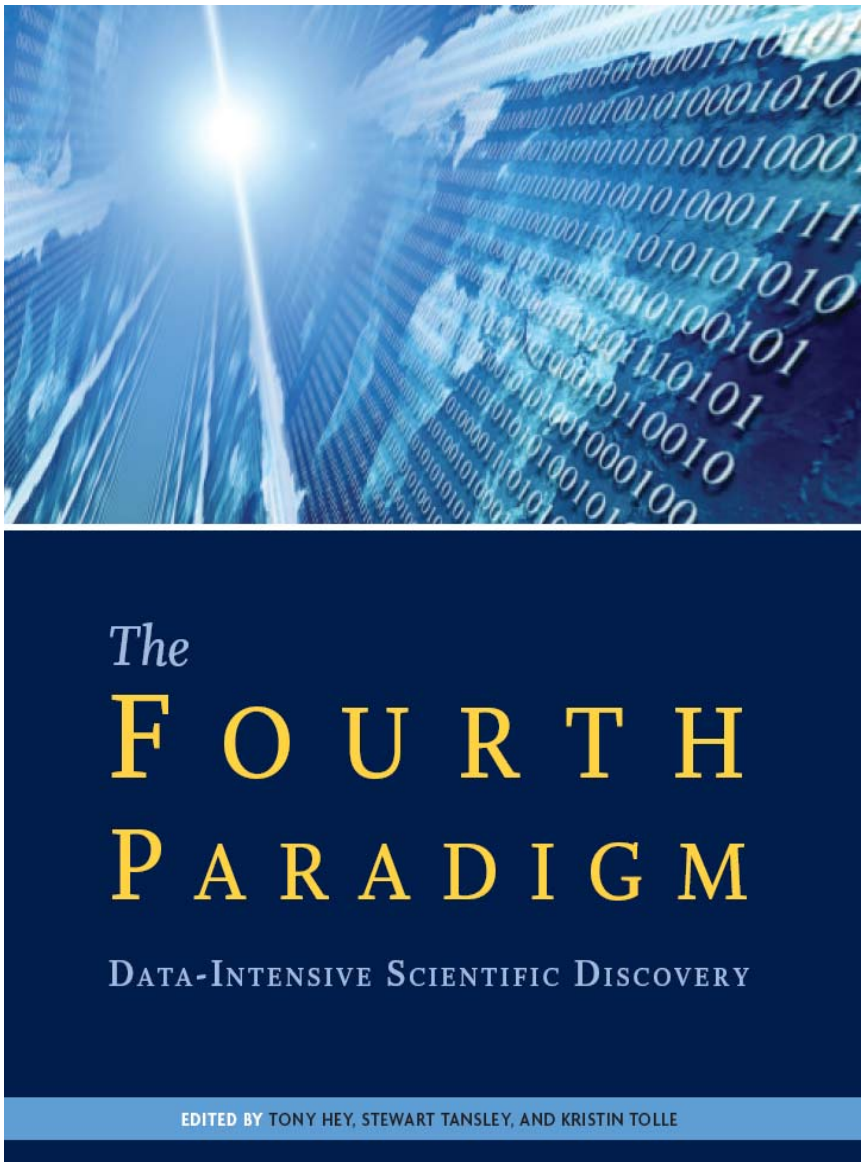
1. Thousand years ago – **Experimental Science**
   - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   - Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   - 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

*Astronomy has been one of the first disciplines to embrace data-intensive science with the Virtual Observatory (VO), enabling highly efficient access to data and analysis tools at a centralized site. The image shows the Pleiades star cluster form the Digitized Sky Survey combined with an image of the moon, synthesized within the WorldWide Telescope service.*

*Science must move from data to information to knowledge*



With thanks to Jim Gray







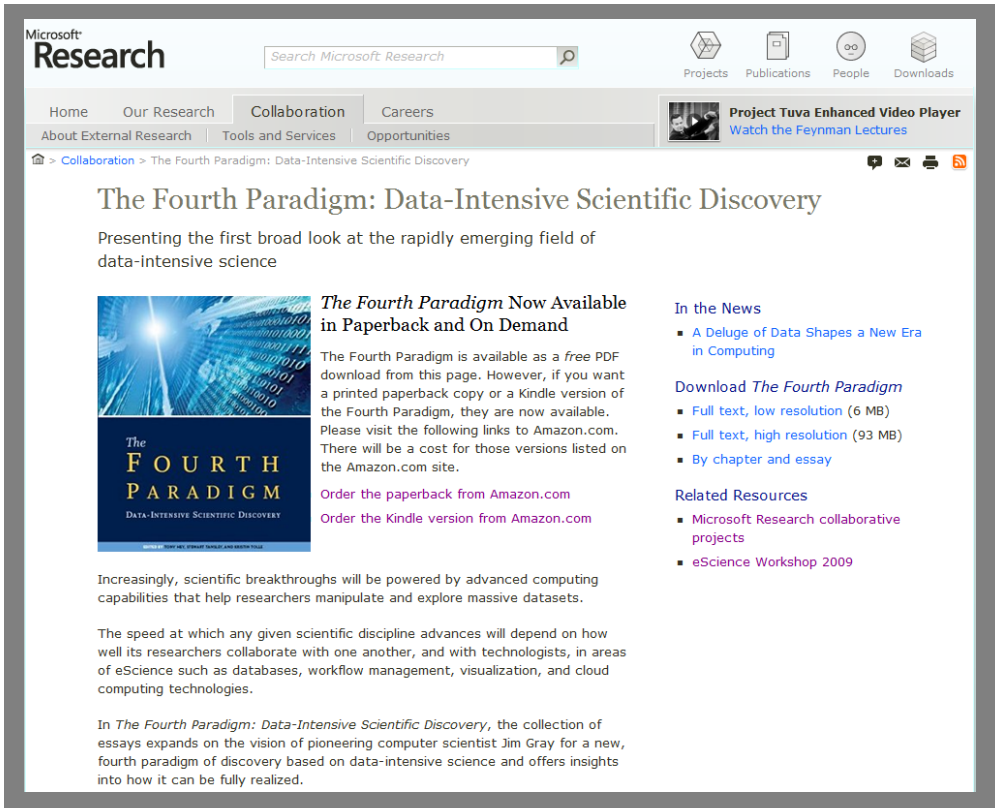
# Free PDF Download
# Amazon Kindle version; Paperback print on demand

http://research.microsoft.com/fourthparadigm/

- "The impact of Jim Gray's thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science."
- — **Bill Gates,** Chairman, Microsoft Corporation

- "One of the greatest challenges for 21st-century science is how we respond to this new era of data-intensive science. This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working."
- — **Douglas Kell**, University of Manchester

- "The contributing authors in this volume have done an extraordinary job of helping to refine an understanding of this new paradigm from a variety of disciplinary perspectives."
- — **Gordon Bell**, Microsoft Research



# Jim Gray's Call to Action

Listed 7 key areas for action by Funding Agencies:

1. Fund both development and support of software tools
2. Invest at all levels of the finding 'pyramid'
3. Fund development of 'generic' Laboratory Information Management Systems
4. Fund research into scientific data management, data analysis, data visualization, new algorithms and tools



# Jim Gray's Call to Action (continued)

Remaining three key areas for action relate to the future of Scholarly Communication and Libraries:

5. Establish Digital Libraries that support the other sciences like the NLM does for Medicine
6. Fund development of new authoring tools and publication models
7. Explore development of digital data libraries that contain scientific data (not just the metadata) and support integration with published literature



# Developing a Sustainable e-Infrastructure

# Accelerating time to insight with Advanced Research Tools and Services


Our goal is to accelerate research by collaborating with academic communities to use advanced computer science research technologies

Aim to help scientists spend less time on IT issues and more time on science by creating open tools and services based on Microsoft platforms and productivity software



# Data Acquisition and Modeling

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

**The Swiss Experiment**

•Powerful Software Improves Environmental Forecasting

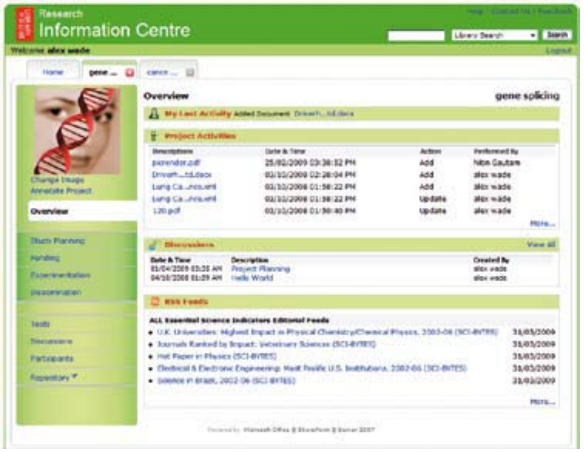
•*Environmental scientists face many challenges in monitoring and understanding our planet's changing climate. Through an international collaboration called the Swiss Experiment, environmental scientists and computer science experts are deploying advanced sensor networks and data management tools to improve environmental monitoring and forecasting.*

**Life Under Your Feet**

•*Researchers at The Johns Hopkins University are deploying large arrays of wireless soil sensors in a variety of environmental settings, including a park, an urban forest and a wetland. The networks enable scientists to monitor ecological changes on an unprecedented scale and offer insights into hydrology, greenhouse gases and the activity of organisms in the soil.*

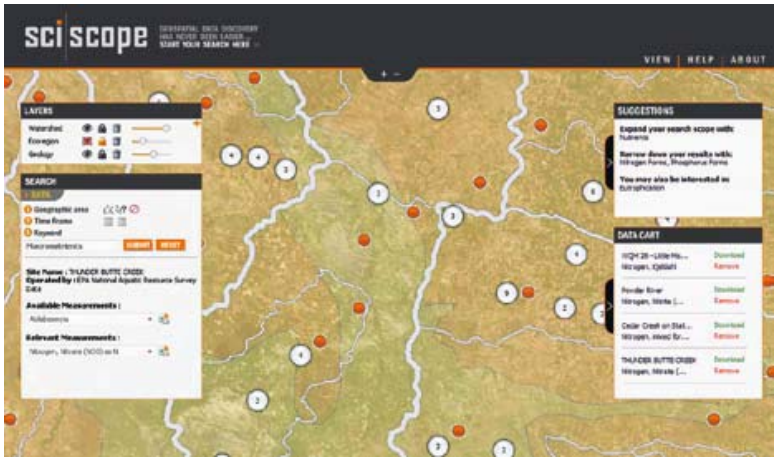
# Collaboration and Visualization

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




**Research Information Center**

*Collaboration and information sharing among researchers are among the most important but challenging aspects of scientific research. In recent years, scientists have begun using "virtual research environments" to exchange information with colleagues in specific areas of study. Microsoft Research and The British Library are teaming up to build the Research Information Centre.*



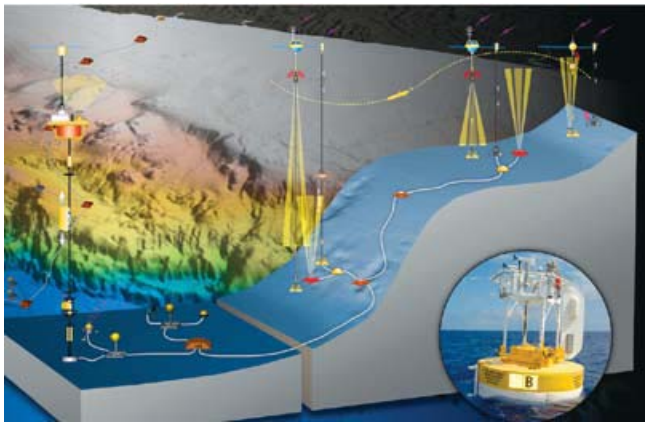

**SciScope -- Speeds Data Retrieval from Multiple Repositories**

*For environmental scientists and engineers, finding and retrieving relevant data can be a daunting and tedious task. Microsoft Research is developing an online search engine called SciScope that enables researchers to search multiple data repositories simultaneously and retrieve information in a consistent format.*



# Analysis and Data Mining

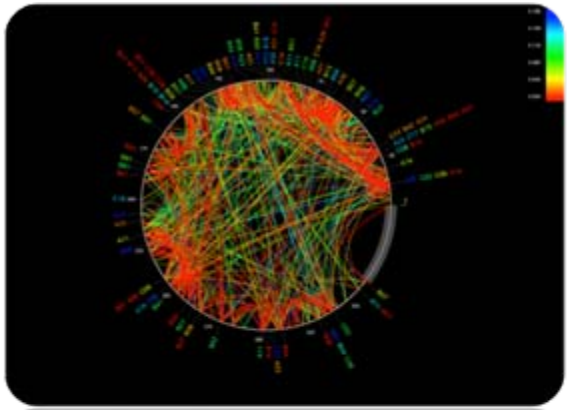
Data Acquisition and Modeling | Collaboration and Visualization | **Analysis and Data Mining** | Disseminate and Share | Archiving and Preservation



**Trident**

•A Scientific Workflow Workbench Brings Clarity to Data

•*Scientists at the University of Washington are working with Microsoft External Research to demonstrate how marrying visualization and workflow technologies can allow researchers to better manage, evaluate and interact with even the most complex scientific datasets.*



**PhyloD**

•Statistical tool used to analyze DNA of HIV from large studies of infected patients

•Typical job, 10 – 20 CPU hours with extreme jobs requiring 1K – 2K CPU hours

- Very CPU efficient
- Requires a large number of test runs for a given job (1 – 10M tests)
- Highly compressed data per job ( ~100 KB per job)

# Disseminate and Share

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

## Chem4Word

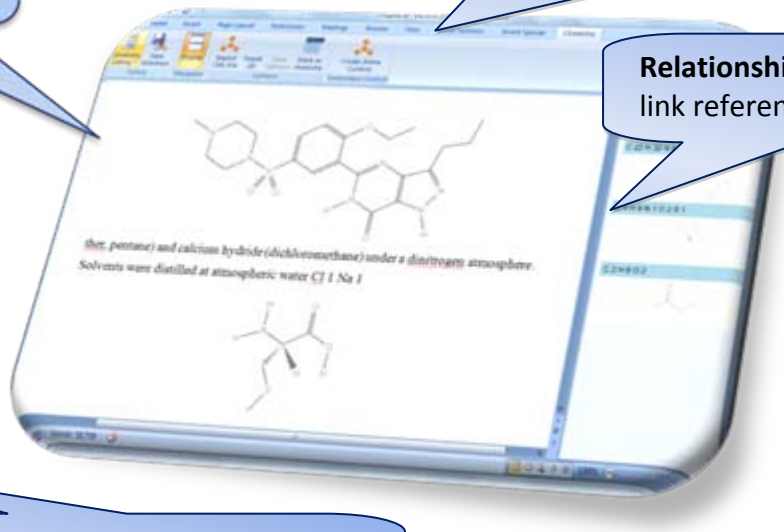
- Chemistry Drawing in Word
- Created in collaboration with University of Cambridge; Peter Murray-Rust, et.al.

**Intent:** Recognizes chemical dictionary and ontology terms

**Author/edit 1D and 2D chemistry. Change chemical layout styles.**

**Relationships:** Navigate and link referenced chemistry

**Data:** Semantics stored in Chemistry Markup Language



```
<?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="-1.9073486455691087E-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>
```

**Intelligence:** Verifies validity of authored chemistry



# Disseminate and Share

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

## Ontology Plug-In for Word

science commons

- John Wilbanks

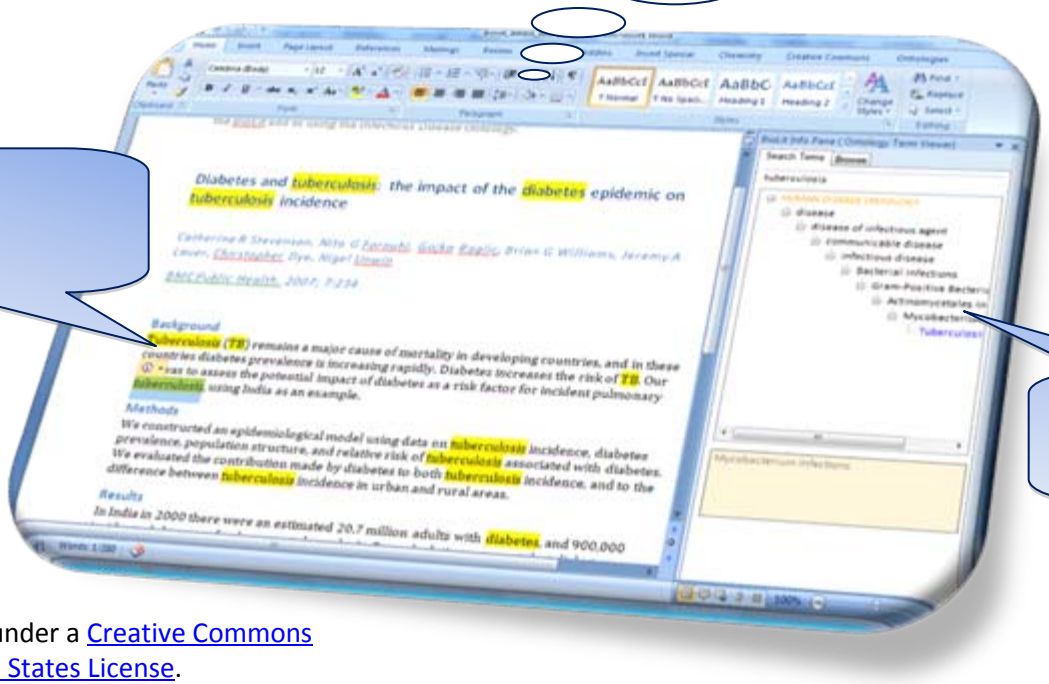
THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

**Services:** Ontology download web service

University of California San Diego

- Phil Bourne
- Lynn Fink

**Intent:** Term recognition & disambiguation




**Relationships:** Ontology browser

# Archiving and Preservation

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

Default web UI with CSS support and custom ASP.Net controls

## Zentity

Native support for RSS, OAI-PMH, OAI-ORE, AtomPub and SWORD

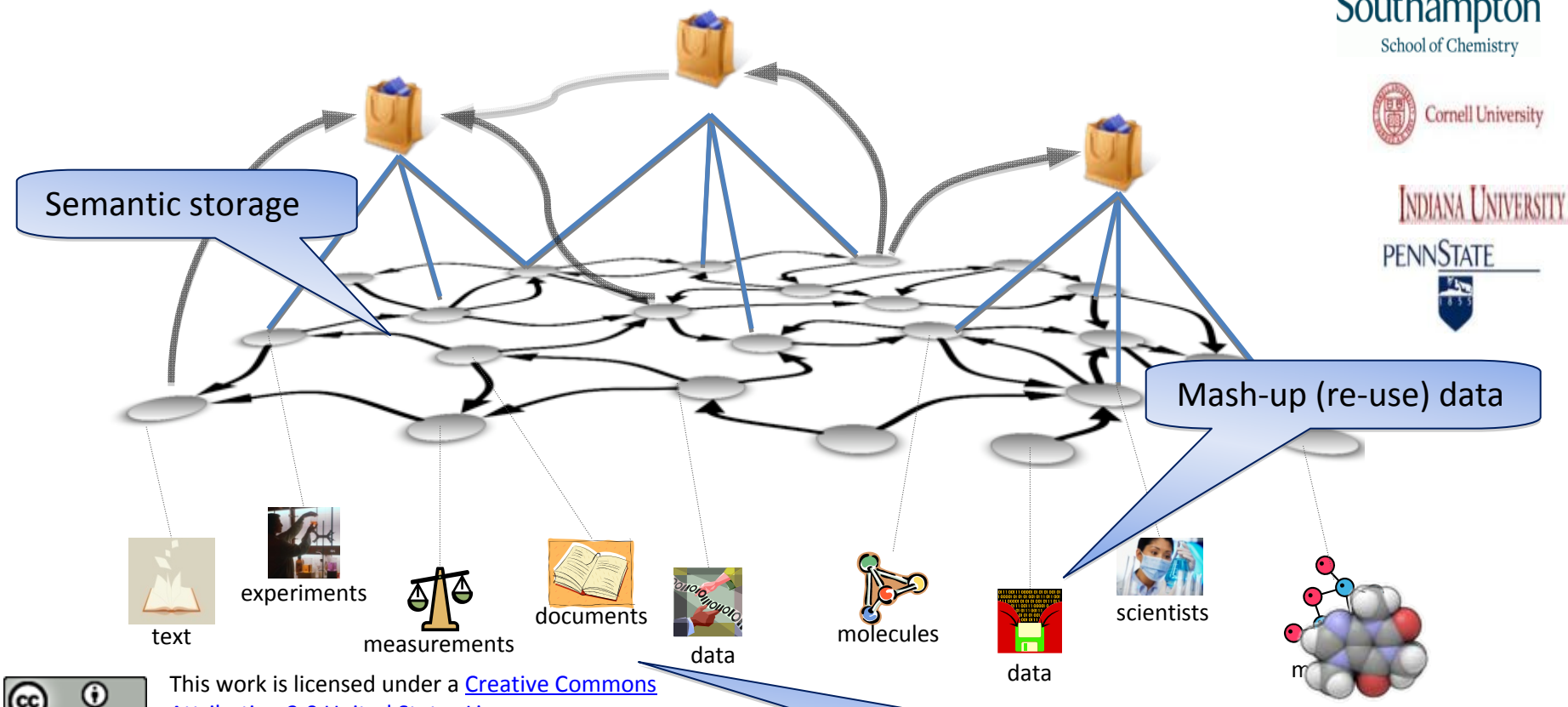

Flexible data model enables many scenarios and can be easily extended over time

A semantic computing platform to store and expose relationships between digital assets

# Archiving and Preservation

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

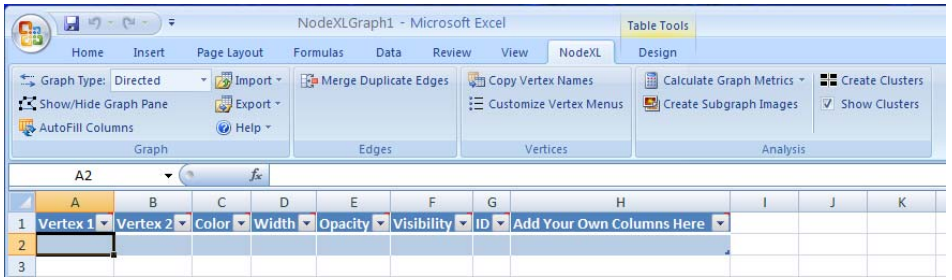
**oreChem** – the Chemical Semantic Web



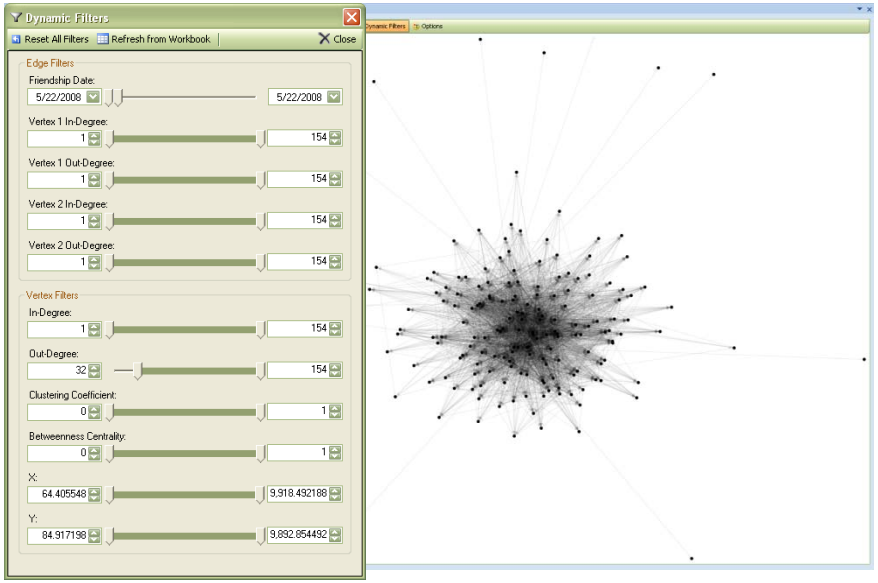
# Node XL

## Network analysis and visualization tool

- Network analysis is of growing importance in academic, commercial, and Internet social media contexts
- Existing Social Network Tools are challenging for many novice users
- Tools like Excel are widely used
- Leveraging a spreadsheet as a host for Social Network Analysis lowers barriers to network data analysis and display




Leverage spreadsheet for storage of edge and vertex data




Apply dynamic filters to the data

# Creative Commons Add-in for Office 2007

**Intent**: Insert Creative Commons licenses from within Office 2007

**Services**: Integrates with Creative Commons Web API to create new licenses


**Relationships**: license information stored as RDF XML within the document OOXML

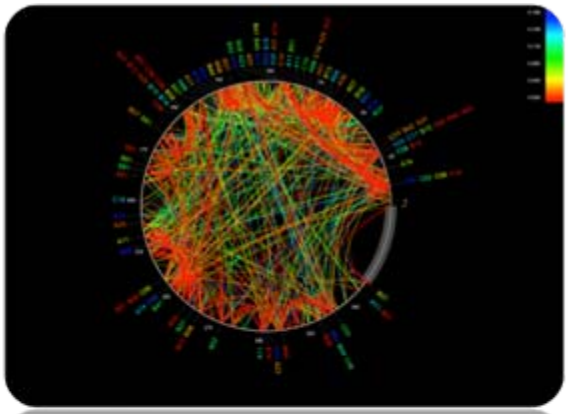


http://ccaddin2007.codeplex.com

# The Future Research e-Infrastructure: Client + Cloud

# PhyloD as an Azure Cloud Service

- Statistical tool used to analyze DNA of HIV from large studies of infected patients
- PhyloD was developed by Microsoft Research and has been highly impactful
- Small but important group of researchers
  - 100's of HIV and HepC researchers actively use it
  - 1000's of research communities rely on these results



Cover of PLoS Biology
November 2008

- Typical job, 10 – 20 CPU hours with extreme jobs requiring 1K – 2K CPU hours
  - Very CPU efficient
  - Requires a large number of test runs for a given job (1 – 10M tests)
  - Highly compressed data per job ( ~100 KB per job)

- **PhyloD now ported as Windows Azure Cloud Service**
- **Cloud enables agile deployment of scalable scientific services**

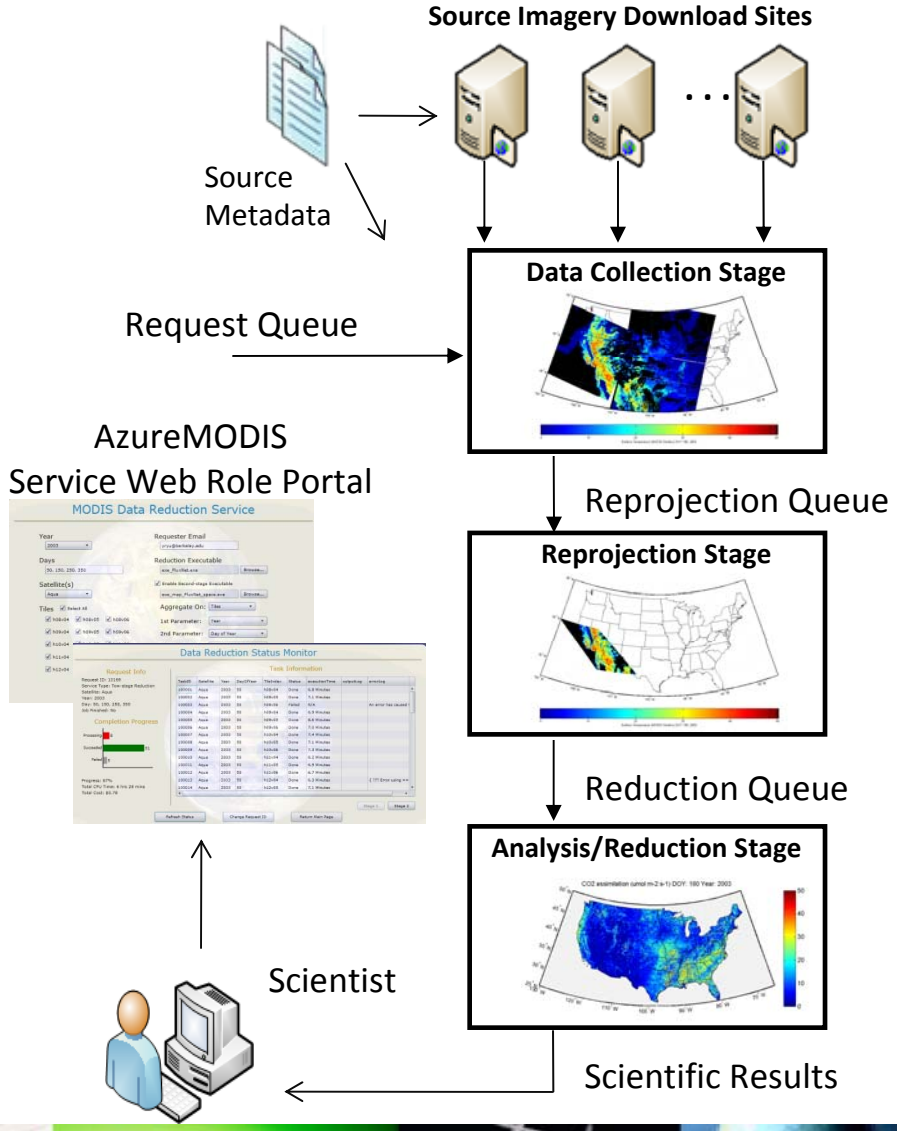


Courtesy of Roger Barga

# AzureMODIS – Azure Service for Remote Sensing Geoscience

- Science pipeline for download, initial processing and reduction of satellite imagery. Developed by MSR, UVa, UCB.
- Dramatically lowers resource and complexity barriers to use satellite imagery for terrestrial hydrology and geoscience.
  - Common imagery location determination and upload from diverse sources
  - Common reprojection and harmonization to produce science-ready imagery with the same length, time and quality attributes
  - Optional scientist-provided reduction algorithm (.NET, Java, or MatLab)
  - On-demand scalability beyond local desktop or cluster
- In use now to compute 10 year continental scale water balance for North America. Per year:
  - 500 GB (~60K files) upload of 9 different source imagery products from 15 different locations
  - 400 GB reprojected harmonized imagery consuming ~3500 cpu hours
  - 5 GB reduced science result leveraging reported field data aggregates consuming ~60 cpu hour
- Additional science requests pending
  - Expanding above to Europe
  - Additional source imagery products and formats

Catharine van Ingen (MSR), Jie Li, Marty Humphreys (UVA), Youngryel Ryu (UCB), Deb Agarwal (BWC/LBL)

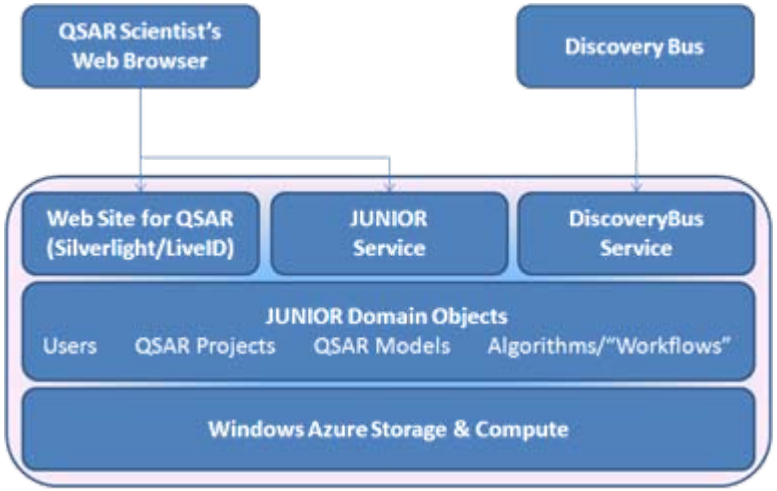


# Project JUNIOR

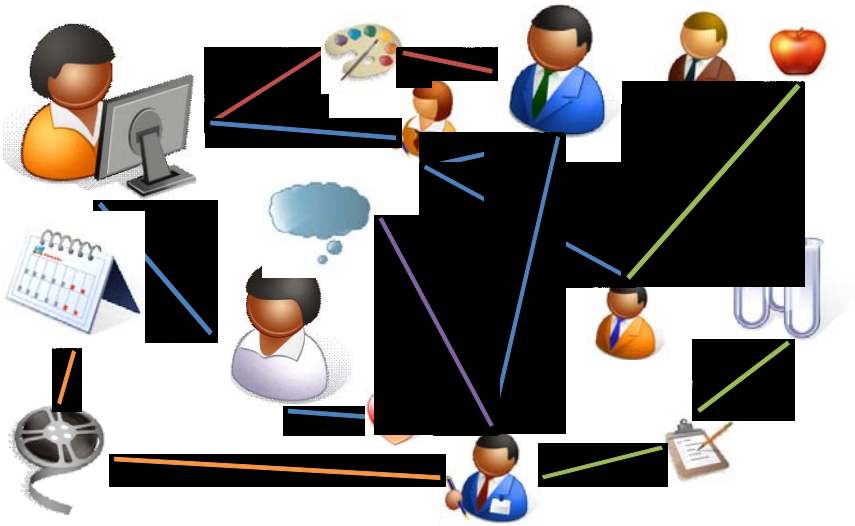
## Demonstrating the Value of Azure Cloud Services for Science

- Led by Newcastle University, UK (Paul Watson), project supported by ER
  - Investigating applicability of commercial clouds for scientific research
  - Build a working prototype for use-cases in chemo-informatics
  - Uses Microsoft technologies to build science-related services (Windows Azure, Silverlight…)


- Built initial proof-of-concept
  - Silverlight UI for basic Quantitative Structure-Analysis Relationship (QSAR) modeling
  - Demonstrated ability to scale QSAR computations in Windows Azure

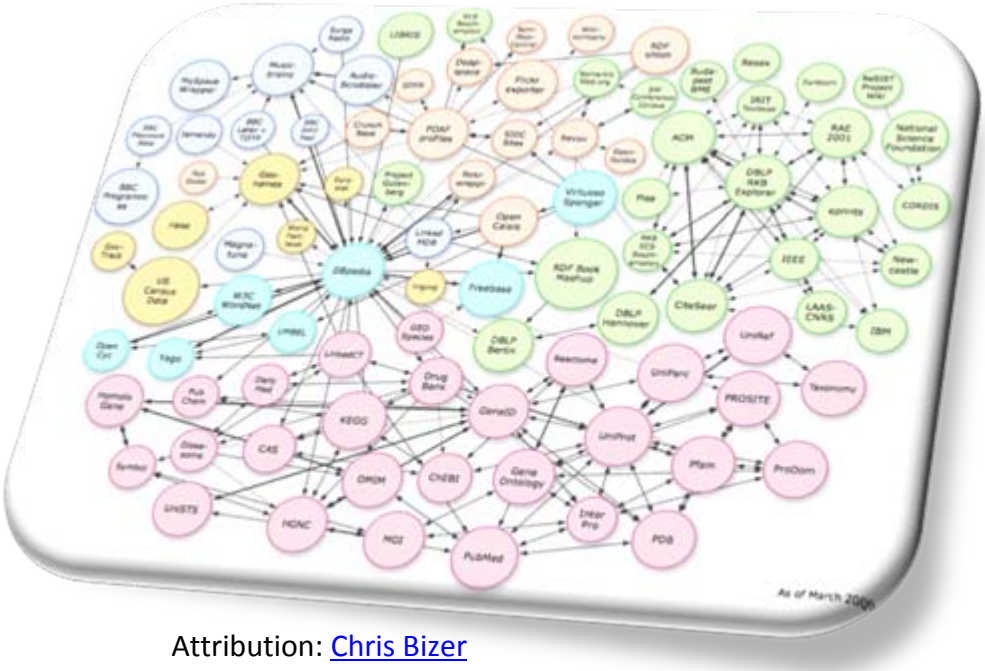


# Moving to a world where all data is linked …



- 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’

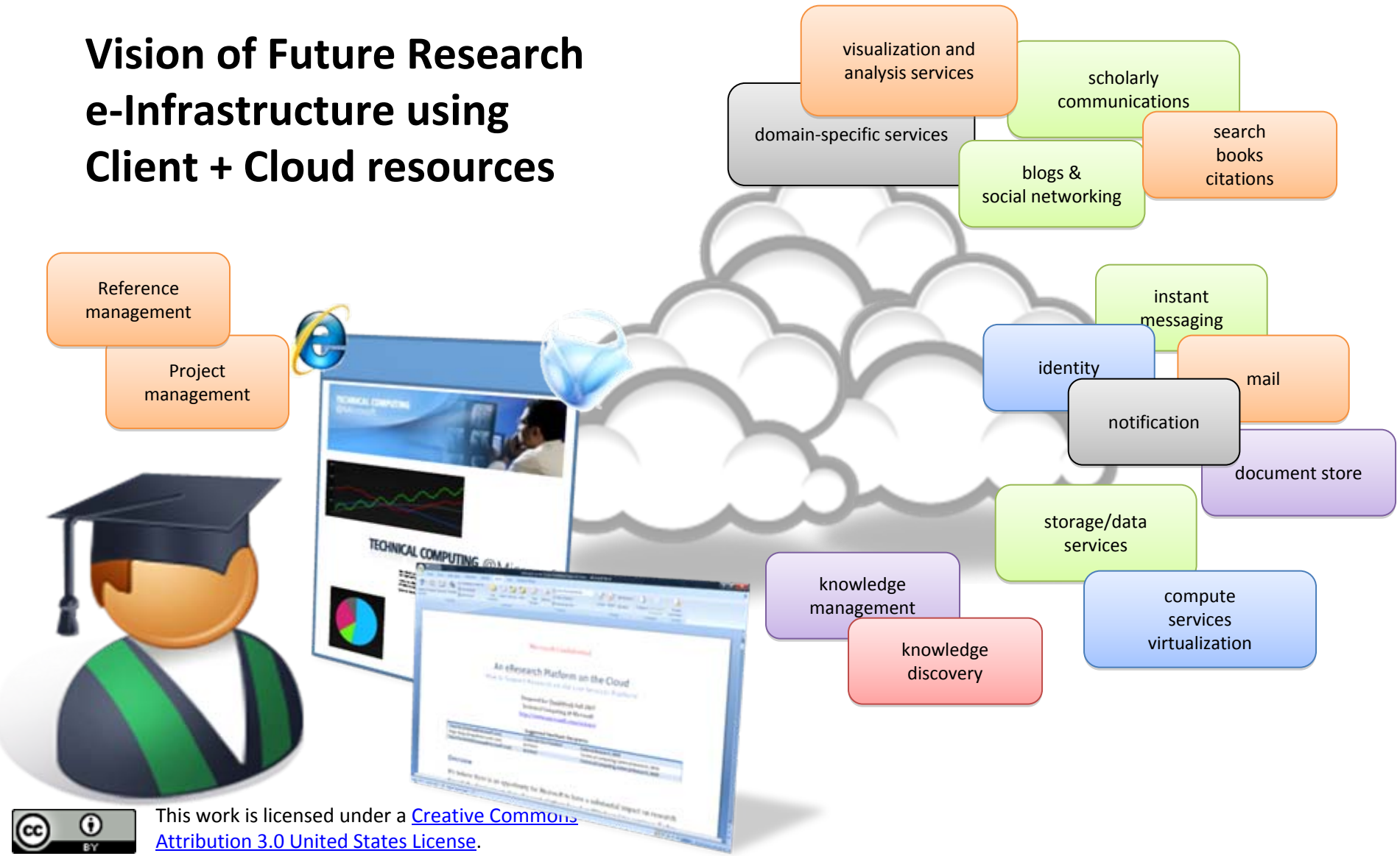
- A knowledge ecosystem:
  - A richer authoring experience
  - An ecosystem of services
  - Semantic storage
  - Open, Collaborative, Interoperable, and Automatic



Attribution: Chris Bizer

# … and can be stored/analyzed in the Cloud

## Vision of Future Research e-Infrastructure using Client + Cloud resources



# Where to download the tools

## research.microsoft.com/en-us/collaboration/tools

The site contains access and downloads of relevant open tools and resources for the worldwide academic research community. Examples of our open tools and services:

**Plug-ins for Office**
Ontology Add-in for Word
Article Authoring Add-in for Word
Chem4Word – Chemistry Drawing in Word

**Microsoft Biology Foundation MBF**
Enables and accelerates fundamental advances in biology

**F#**
Collaboration with the academic and research community on F#'s typed functional and object-oriented programming on the .NET platform

**Software Engineering Tools**
Spec#: Program verifier for C# extended with design by contract
VCC: Program verifier for Concurrent C
PEX: automatic unit testing tool for .NET
CHESS: Unit testing tools for concurrent Win32 executable and .NET

# Resources

- Microsoft Research
  - http://research.microsoft.com
  - Microsoft Research downloads: http://research.microsoft.com/research/downloads
- Microsoft External Research
  - http://research.microsoft.com/externalresearch
- Science at Microsoft
  - http://www.microsoft.com/science
- CodePlex
  - http://www.codeplex.com
- The Faculty Connection
  - http://www.microsoft.com/education/facultyconnection
- MSDN Academic Alliance
  - http://msdn.microsoft.com/en-us/academic



Microsoft®

*Your potential. Our passion.™*