eScience: Past, Present and Future

Tony Hey
Corporate Vice President
Microsoft Research

Microsoft Research Connections

e-Science: The Past

In the UK …

e-Science: A Definition

‘e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it.’

John Taylor
Director General of Research Councils
Office of Science and Technology
September 2000

UK e-Science Funding

1st Phase: 2001 –2004

- Application Projects
  - £74M
  - All areas of science and engineering
- Core Programme
  - £15M Research infrastructure
  - £20M Collaborative industrial projects

2nd Phase: 2003 –2006

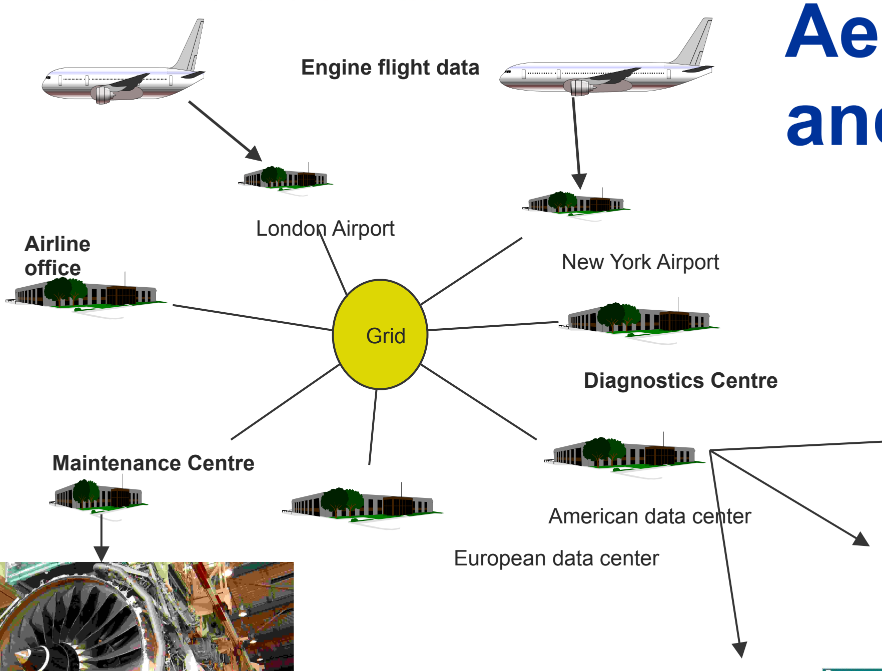
- Application Projects
  - £96M
  - All areas of science and engineering
- Core Programme
  - £16M Research Infrastructure
  - £11M Collaborative industrial projects

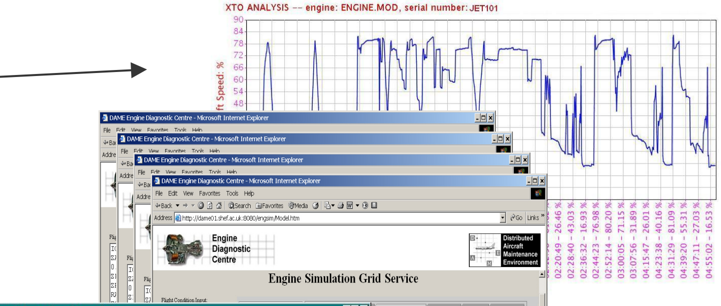
The UK e-Science Paradigm

- The **Integrative Biology Project** involved seven UK Universities led by Oxford and the University of Auckland in New Zealand
  - Models of electrical behaviour of heart cells developed by Denis Noble's team in Oxford
  - Mechanical models of beating heart developed by Peter Hunter's group in Auckland
- Researchers need robust middleware services to routinely build secure 'Virtual Organisations' to support an international "collaboratory"
  - ➢ Goal is to enable 'faster, better or different' research

Distributed Aircraft Maintenance Environment
DAME

DAME: Grid based tools and Infer-structure for Aero-Engine Diagnosis and Prognosis





DAME Engine Diagnostic Centre - Microsoft Internet Explorer

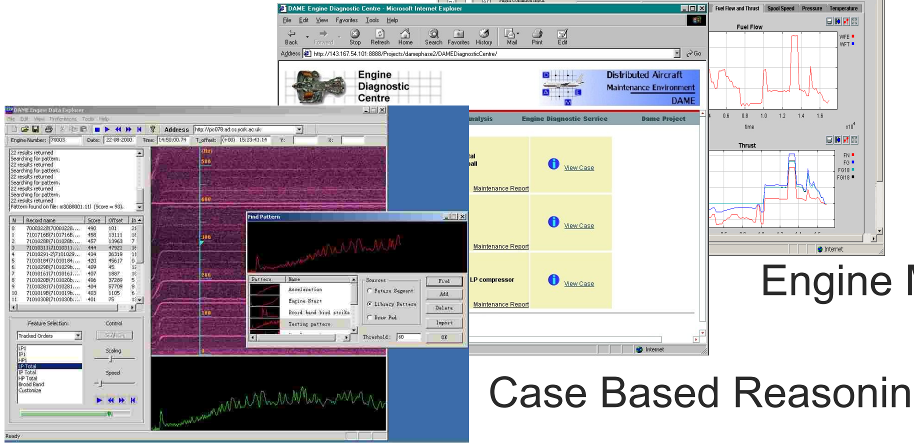
Address http://dame01.shef.ac.uk:8080/engsim/Model.htm

Engine Diagnostic Centre

Distributed Aircraft Maintenance Environment

Engine Simulation Grid Service

Flight Condition Input:



DAME Engine Diagnostic Centre - Microsoft Internet Explorer

Address http://143.167.54.101:8888/Projects/damephase2/DAMEDiagnosticCentre/

Engine Diagnostic Centre

Distributed Aircraft Maintenance Environment DAME

Engine Diagnostic Service | Dame Project

Maintenance Report | View Case

Maintenance Report | View Case

LP compressor | Maintenance Report | View Case

XTO

Engine Model

Case Based Reasoning

Signal Data Explorer

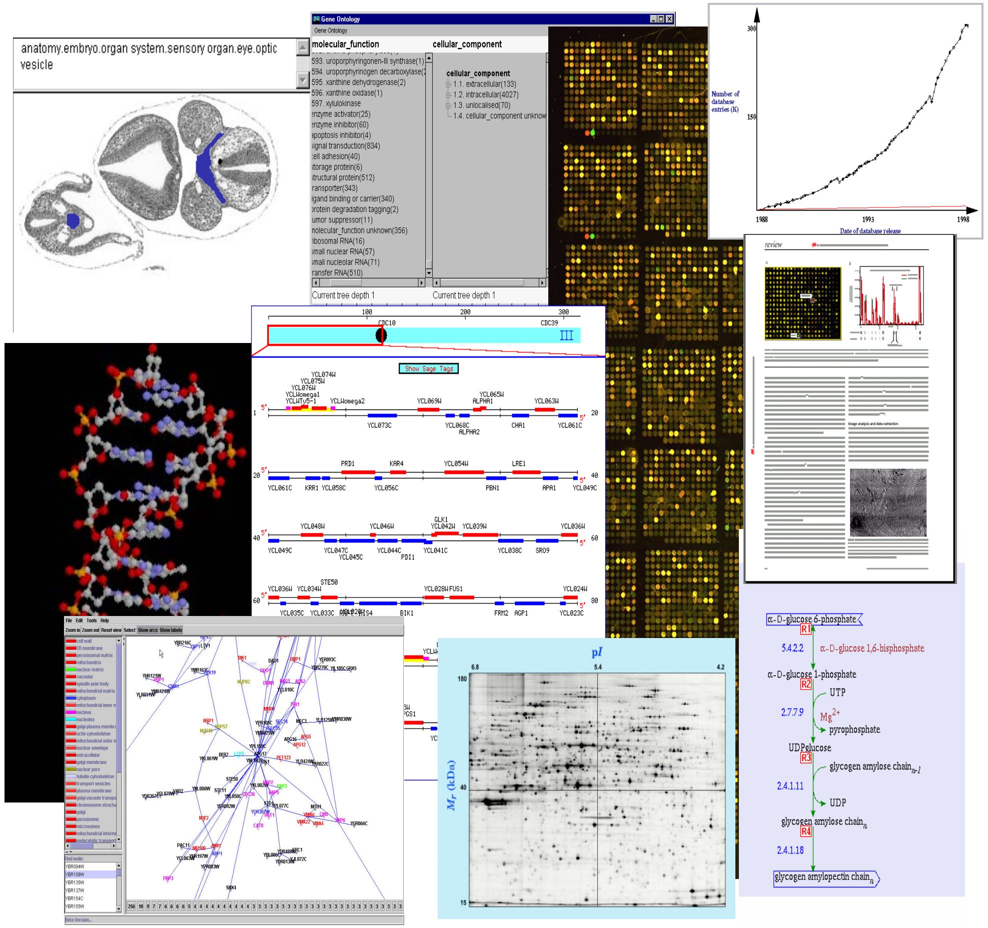
Companies:	Universities:
Rolls-Royce	York,
DS&S	Leeds,
Cybula	Sheffield, Oxford

PI: Jim Austin

Microsoft Research Connections

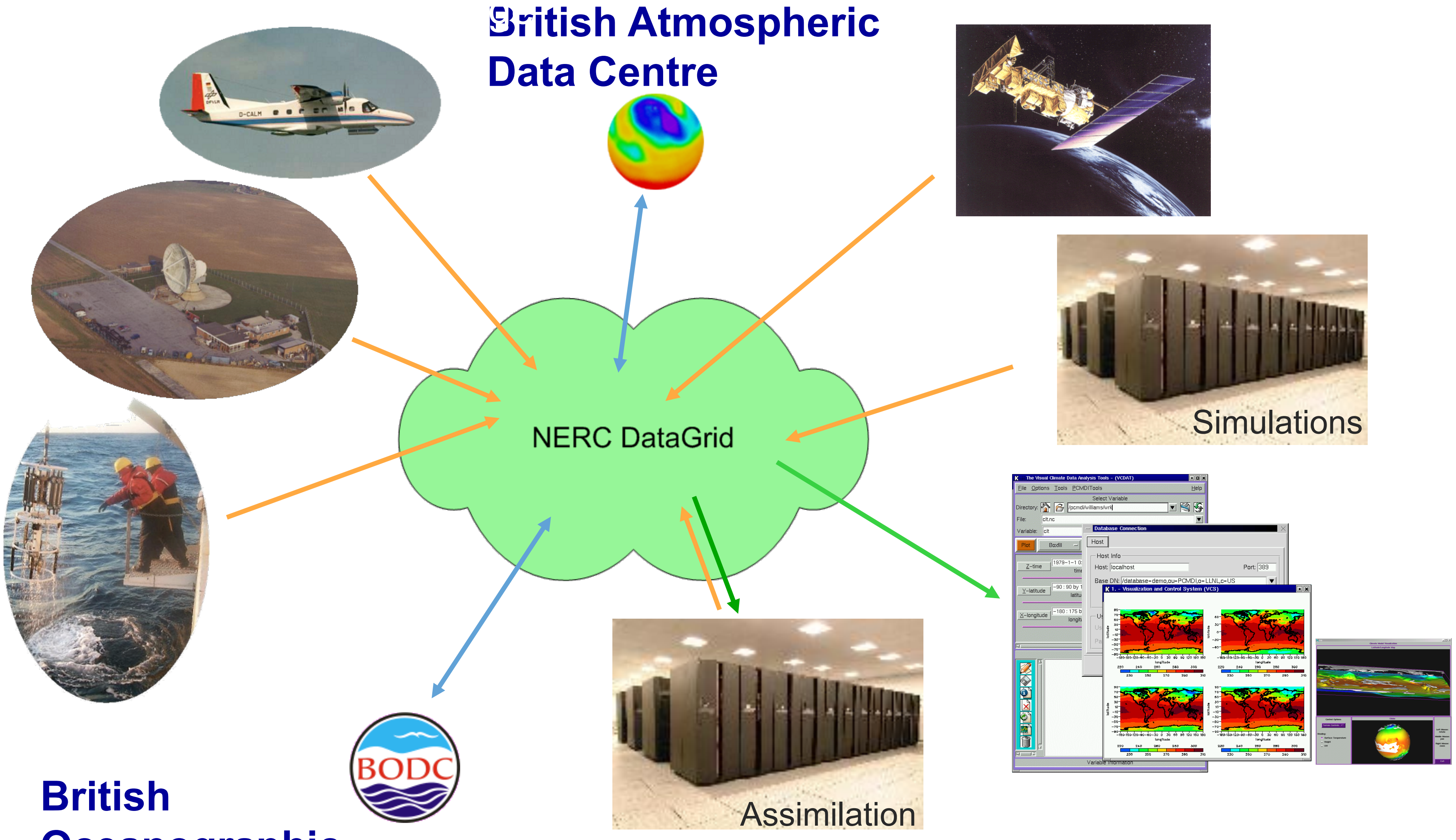
The myGrid Project

- Imminent ‘deluge’ of data
- Highly heterogeneous
- Highly complex and inter-related
- Convergence of data and literature archives



PI: Carole Goble

The NERC DataGrid Project



British Atmospheric Data Centre

NERC DataGrid

Simulations

Assimilation

The Visual Climate Data Analysis Tools - (VCDAT)

File Options Tools PCMDITools Help

Select Variable

Directory: /pcmdi/williams/wrk

File: clt.nc

Variable: clt

Plot Boxfill

Z-time 1979-1-1 0

Y-latitude -90 : 90 by 1

X-longitude -180 : 175 b

Database Connection

Host

Host Info

Host: localhost Port: 389

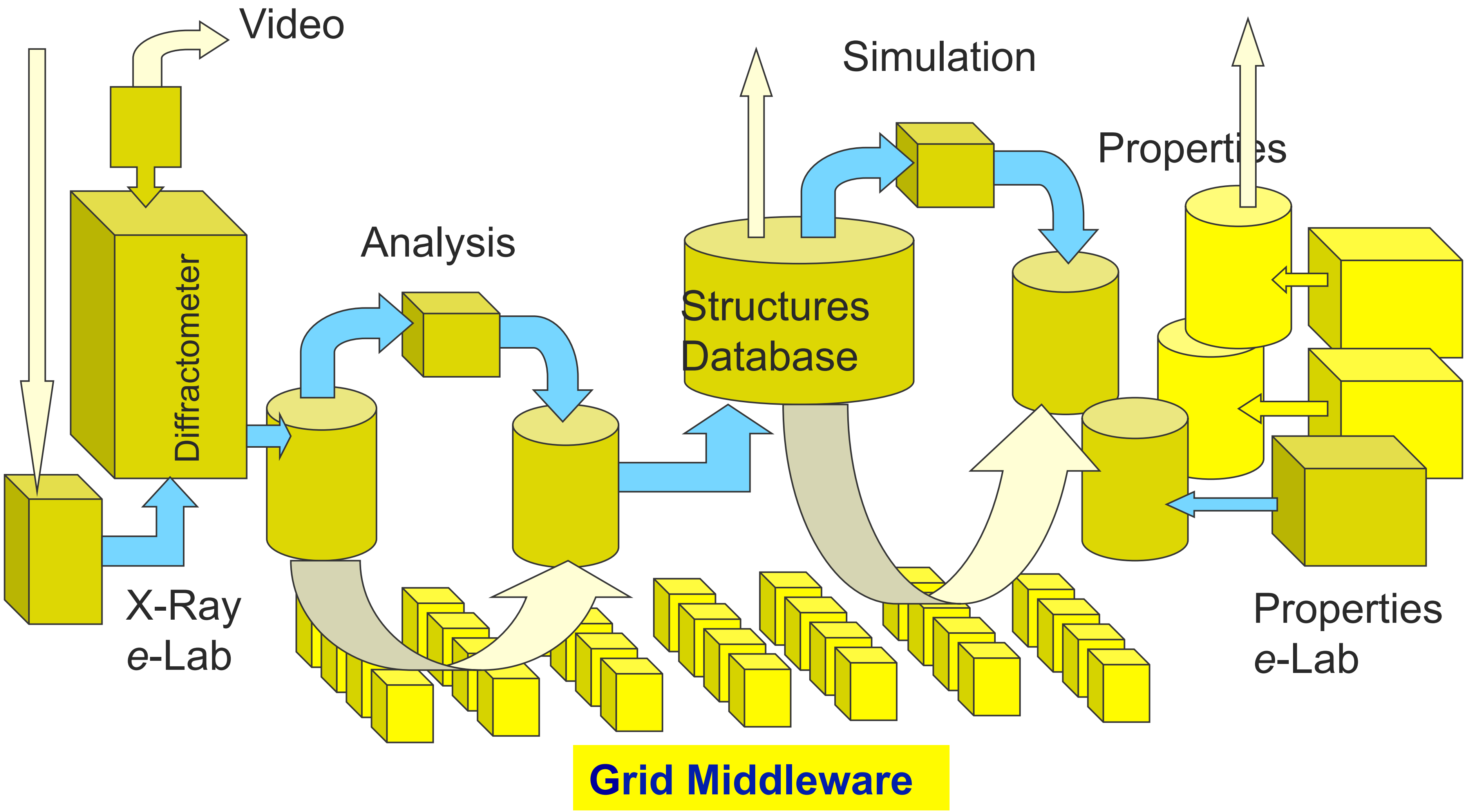
Base DN: /database=demo,ou=PCMDI,o=LLNL,c=US

1. - Visualization and Control System (VCS)

British Oceanographic Data Centre

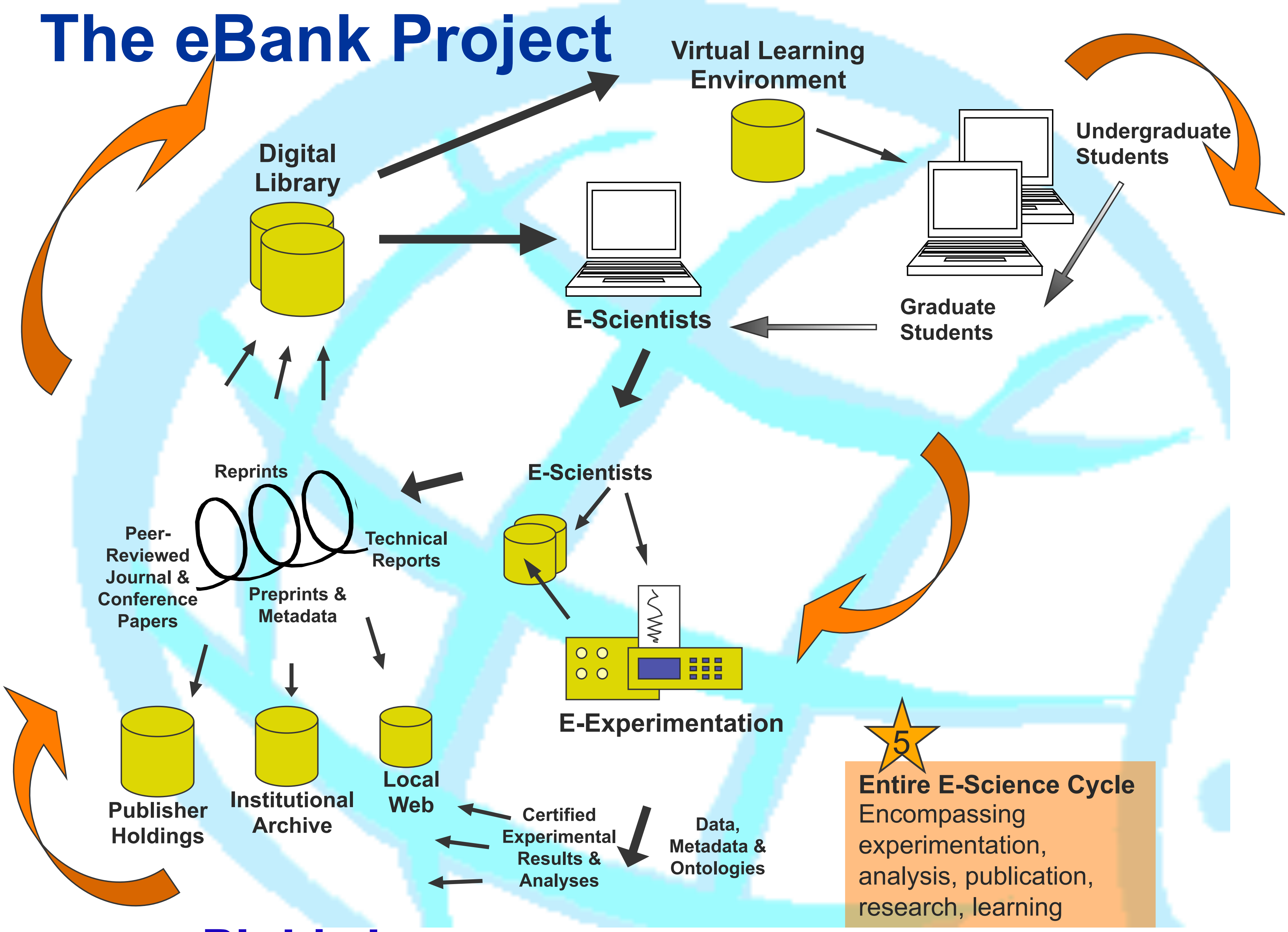
PI: Bryan Lawrence

The Comb-*e*-Chem Project



PI: Jeremy Frey

The eBank Project



Entire E-Science Cycle
Encompassing experimentation, analysis, publication, research, learning

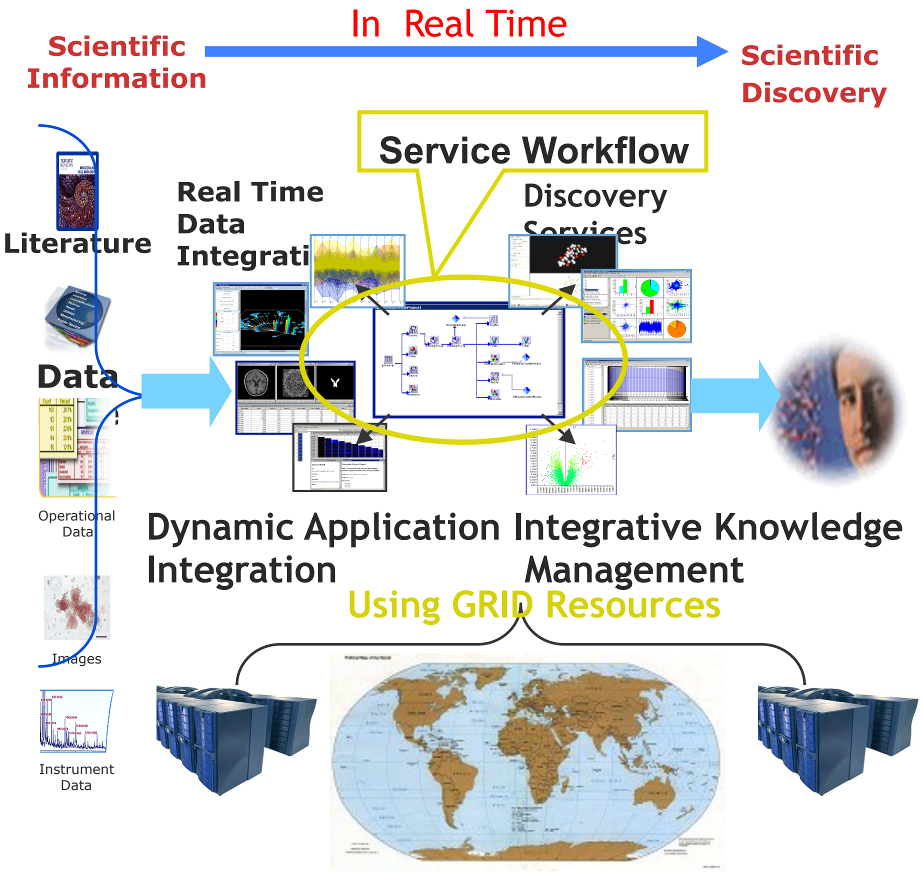
PI: Liz Lyons

Discovery Net

Unifying the World's Knowledge

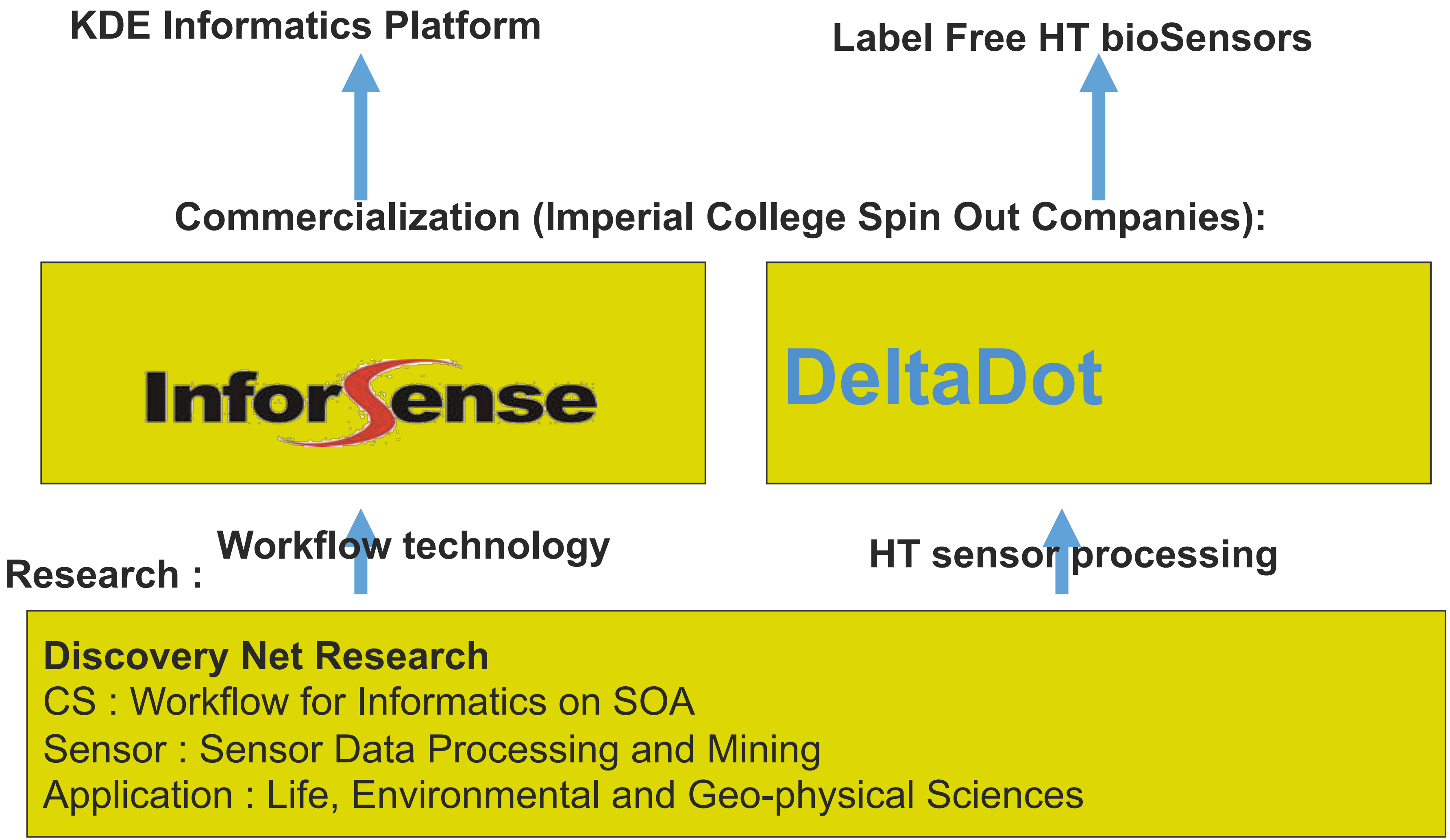
High Throughput Informatics

- Design, develop and implement an advanced infrastructure to support real-time processing, interpretation, integration, visualization and mining of vast amounts of time critical data generated by high throughput devices.
- Data mining, text mining
- Environmental monitoring, bioinformatics
- 2003: Discovery Net in Action: Fighting SARS in China
- 2002: Supercomputing 2002 Most Innovative Data Intensive Application Award
- 2002: KDD CUP 2002 Scientific Text Mining Awards



PI: Yike Guo

Discovery Net Commercialization



KDE Informatics Platform

Label Free HT bioSensors

Commercialization (Imperial College Spin Out Companies):

InforSense

DeltaDot

Workflow technology

HT sensor processing

Research :

Discovery Net Research
CS : Workflow for Informatics on SOA
Sensor : Sensor Data Processing and Mining
Application : Life, Environmental and Geo-physical Sciences

climate*prediction*.net

e-Science dti

A sophisticated scientific experiment, making use of the power of distributed computing

PI: Myles Allen

- To produce the most complete probability-based forecast for the climate of the 21st Century attempted to date
- To improve our understanding of current state-of-the-art climate models
- To engage the public in the climate change debate and to improve public understanding of the nature of uncertainty in climate prediction

climate*prediction*.net should give policy makers a better scientific basis for addressing one of the biggest potential global problems of the 21st century.

UGAMP

e SC

NATURAL ENVIRONMENT RESEARCH COUNCIL

Risk Management Solutions

R M S ™

The Open University

BOINC!

CLRC

Met Office

Tessella Scientific software solutions

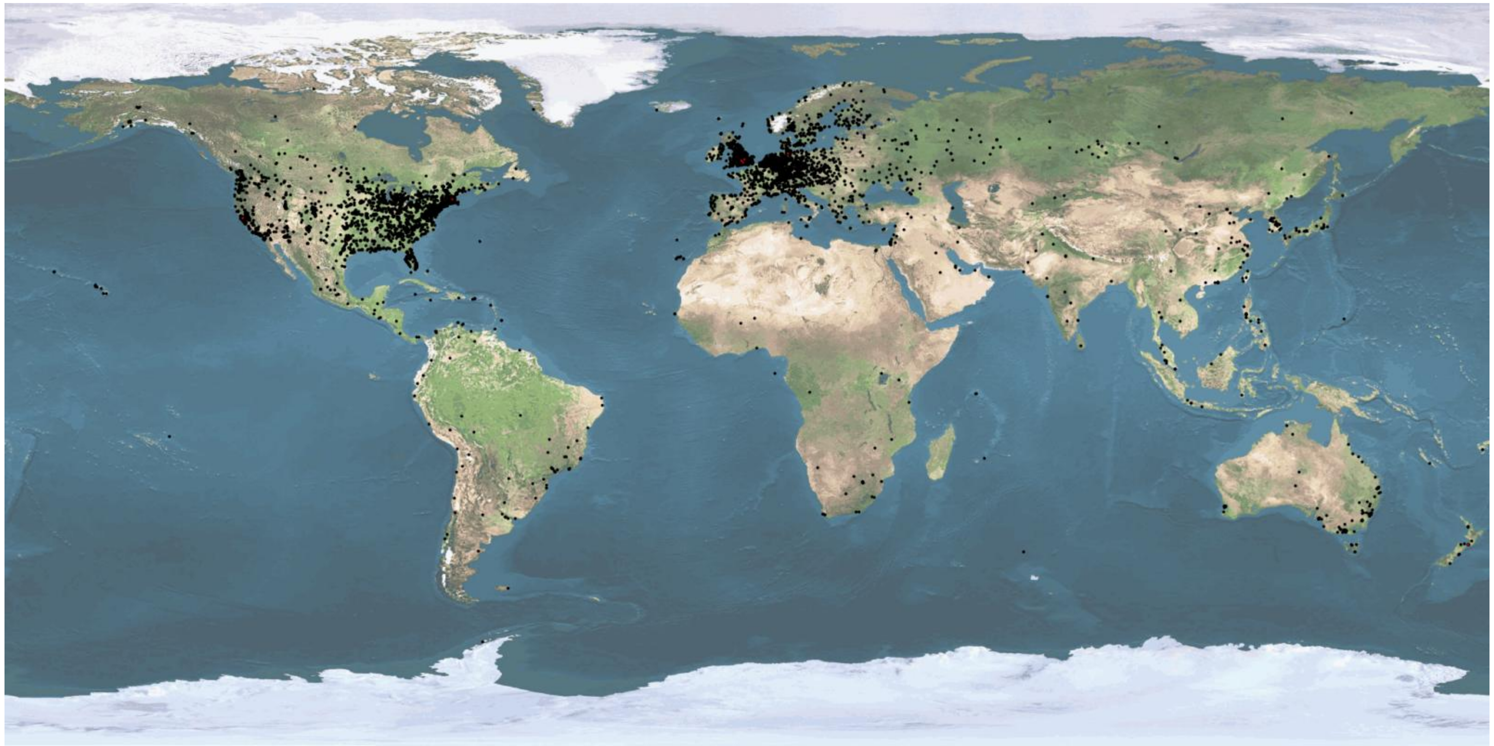
RESEARCH SYSTEMS A Kodak Company

Oxford Physics

UNIVERSITY OF OXFORD

Microsoft Research Connections

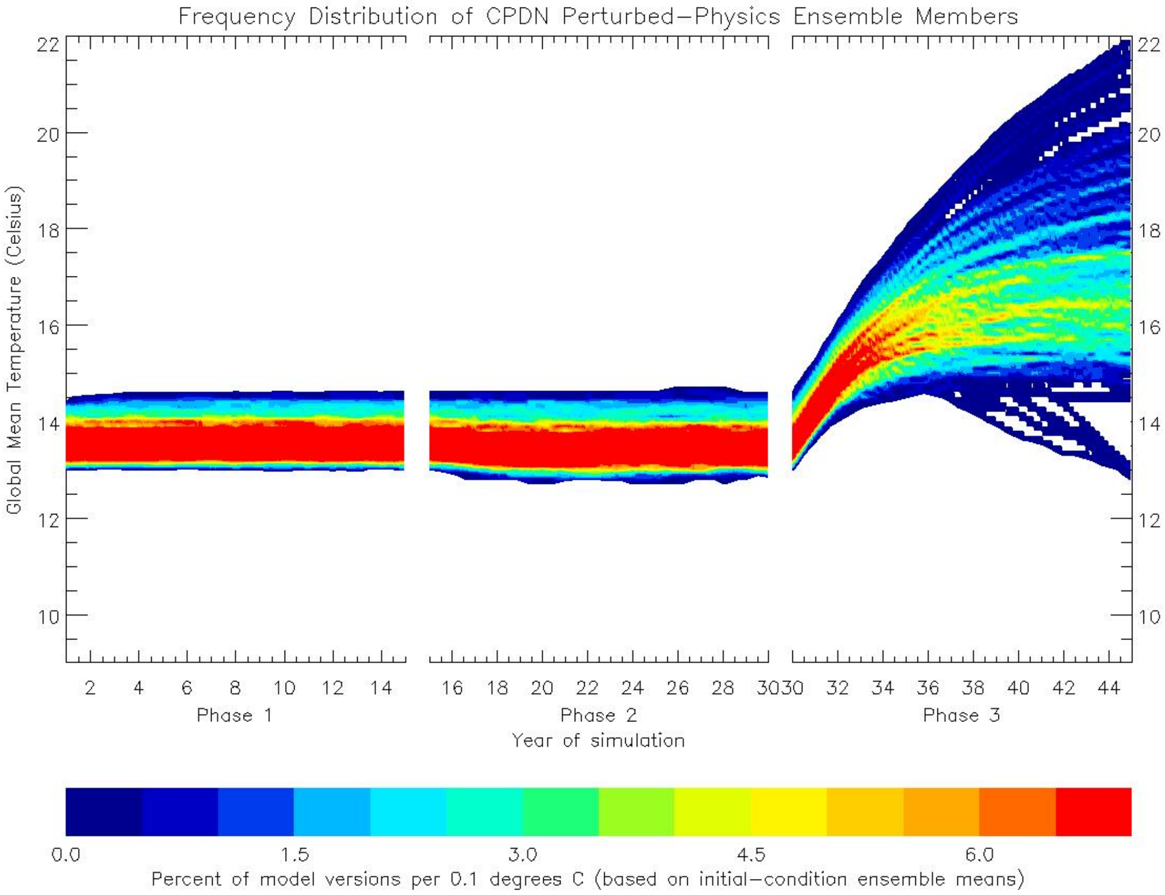
Achievements (December 04)



Since September 2003:
61,000 registered participants in **130** countries have…

Donated **5,000 years** of computer time
Completed **33,000 experiments**

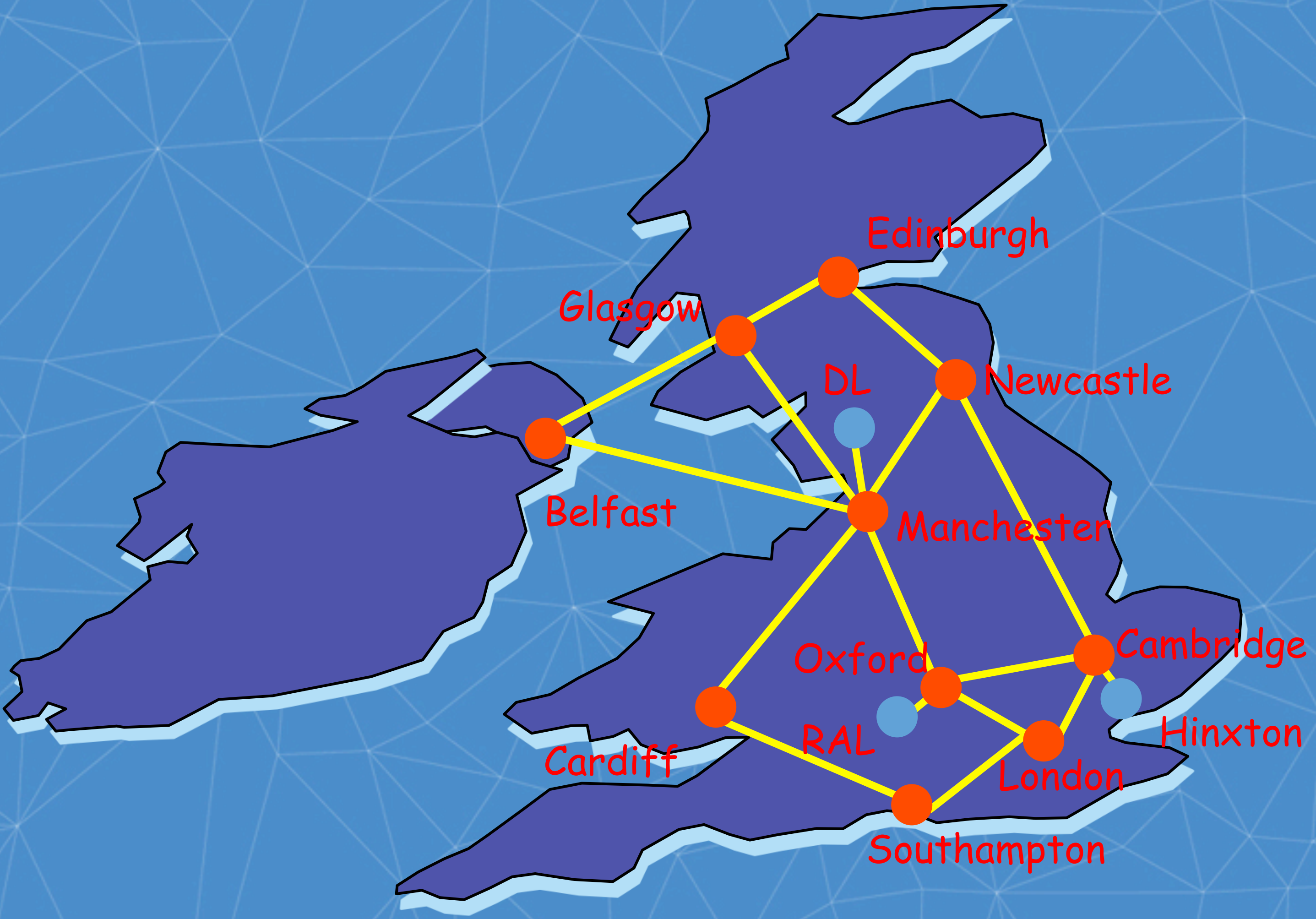
Results so far: the first steps towards a fully probability-based forecast



climate*prediction*.net

Microsoft Research Connections

UK e-Science Centres



Open Middleware Infrastructure Institute

The Three OMII Goals

- Set up Repository for WS-* generic Grid middleware for the UK e-Science community
- Capture generic middleware from UK e-Science Projects
- Commission middleware projects to fill specific ‘gaps’

➢ All supported by standard Software Engineering processes

Computer Science for e-Science

- 18 projects, 16 departments,
- 85% CS for e-Science projects in RAE 5*/5 deptartments
- 59 academics

HyOntUse
Secure location independent autonomic storage architectures
Grid-enabled numerical and symbolic services
MagikI: managing grids containing information and knowledge that are incomplete
A Semantic Firewall

Dynamic Ontologies: a Framework for Service Descriptions
Trusted Coordination in Dynamic Virtual Organisations
Service Level Agreement Based Scheduling Heuristics
A System for Publishing Scientific Data
PASOA: Provenance Aware Service Oriented Architecture
AMUSE: Autonomic Management of Ubiquitous Systems for e-Health
Dynamic Net Data: Theory and Experiment
Presenting Ontologies in Natural Language
Describing the Quality of Curated E-Science Information Resources
Pervasive Debugging
Inferring Quality of Service Properties for Grid Applications
Open Overlays: Component-Based Communications Support for the Grid
Virtual organisations

Total UK e-Science Funding 2001-2006

e-Science Initiative		£250M
Applications	£190M	
Core Programme	£30M	
DTI Industrial Projects	£30M	
High Performance Computing		£50M+
HPC	£50M	
University Systems	£??M	
Research Infrastructure		£250M
Network	£150M	
Content	£50M	
Research Support	£50M	

JISC Research Infrastructure Funding

- **National Academic Network**
  - SuperJANET4 plus MANs
  - ‘UKLight’ lambda connection
- **Content**
  - National Data Sets
  - Publications
- **Research Support**
  - Core Middleware (Authentication and Authorisation)
  - Digital Curation Centre
  - Virtual Research Environment

Authentication and Authorization Infrastructure

- Security Development Projects
  - Combine Shibboleth with PERMIS Authorization Services
  - Joint project with NSF Internet2 NMI project on Security Services for Virtual Organizations
- National Middleware Services
  - Deployment of National Authentication Framework based on Shibboleth
  - Support for both Digital Library and e-Science communities

DCC

Digital Curation Centre

- Actions needed to maintain and utilise digital data and research results over entire life-cycle
  - For current and future generations of users
- Digital Preservation
  - Long-run technological/legal accessibility and usability
- Data curation in science
  - Maintenance of body of trusted data to represent current state of knowledge in area of research
- Research in tools and technologies
  - Integration, annotation, provenance, metadata, security…..

Digital Preservation: The issues

- Long-term preservation
  - Preserving the bits for a long time (“digital objects”)
  - Preserving the interpretation (emulation/migration)
- Political/social
  - Appraisal - what to keep?
  - Responsibility - who should keep it?
  - Legal - can you keep it?
- Size
  - Storage of/access to Petabytes of regular data
  - Grid issues
- Finding and extracting metadata
  - Descriptions of digital objects

Data Publishing: The Background

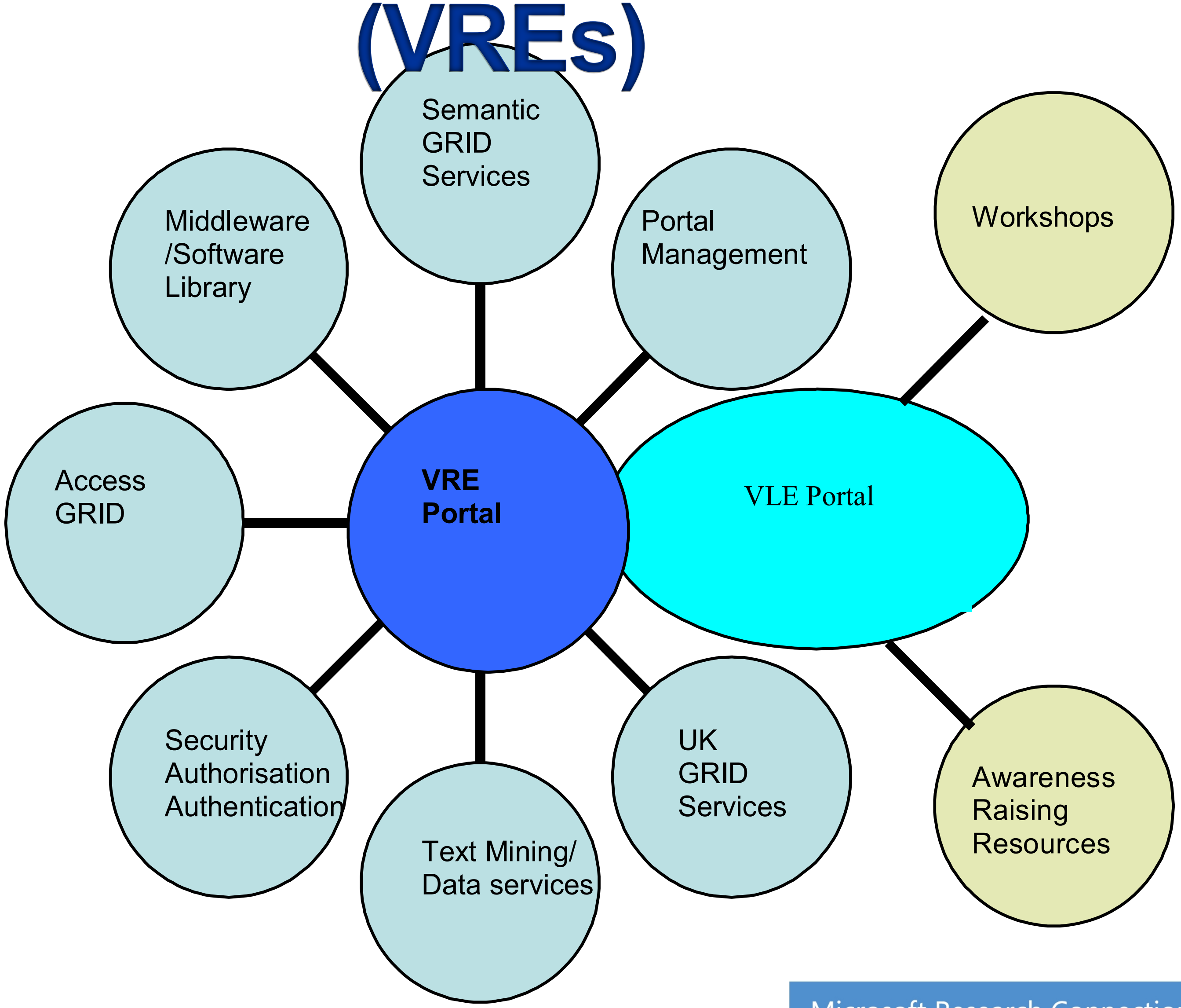
In some areas – notably biology – databases are replacing (paper) publications as a medium of communication

- These databases are built and maintained with a great deal of human effort
- They often do not contain source experimental data. Sometimes just annotations/metadata
- They borrow extensively from, and refer to, other databases
- You are now judged by your databases as well as your (paper) publications – but will they count in the UK RAE?
- Upwards of 1000 (public databases) in genetics

Data Publishing: The issues

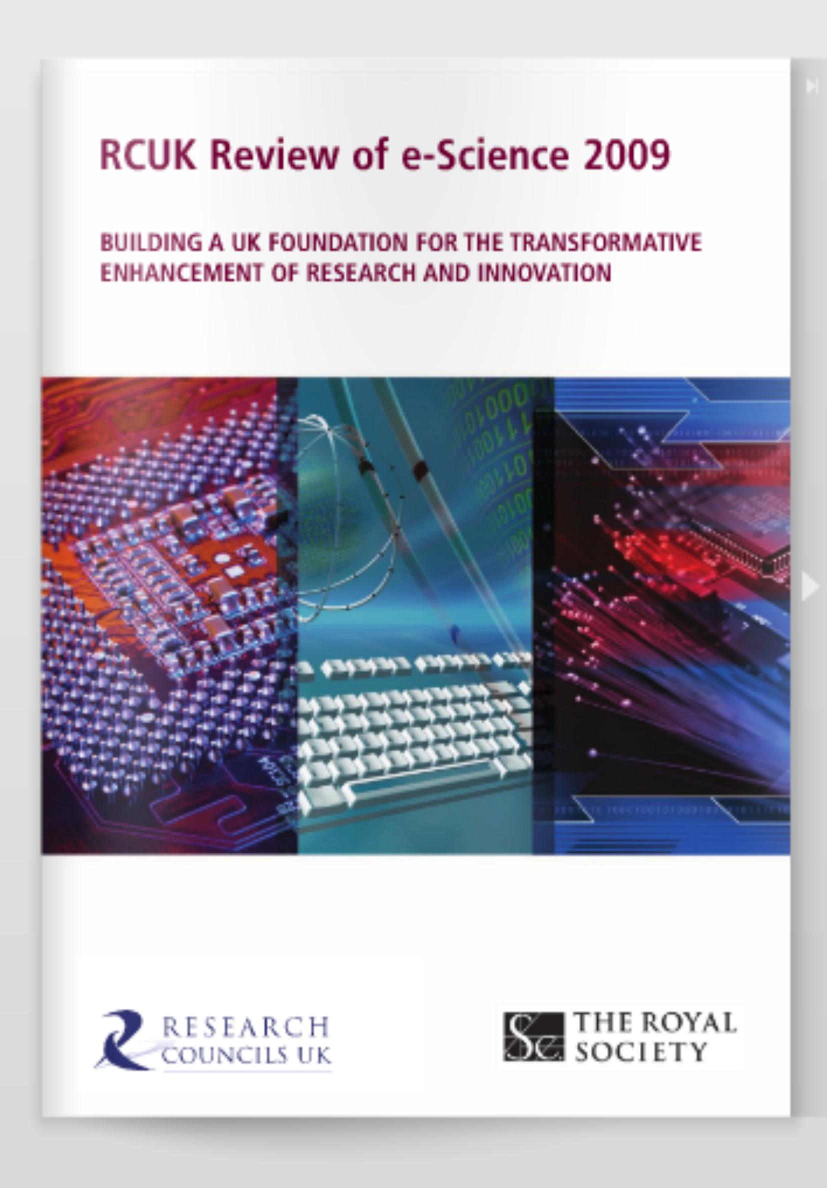
- Data integration
  - Tying together data from various sources
- Annotation
  - Adding comments/observations to existing data
  - Becoming a new form of communication
- Provenance
  - Where did this data come from?
- Exporting/publishing in agreed formats
  - To other program as well as people
- Security
  - Specifying/enforcing read/write access to *parts* of your data

JISC Programme for Virtual Research Environments (VREs)


Semantic GRID Services
Middleware /Software Library
Portal Management
Workshops
Access GRID
VRE Portal
VLE Portal
Security Authorisation Authentication
Text Mining/ Data services
UK GRID Services
Awareness Raising Resources

UK e-Science Program: Six Key Elements for a Global e-Infrastructure (2004)

1. High bandwidth Research Networks
2. Internationally agreed AAA Infrastructure
3. Development Centers for Open Software
4. **Technologies and standards for Data Provenance, Curation and Preservation**
5. **Open access to Data and Publications via Interoperable Repositories**
6. Discovery Services and Collaborative Tools



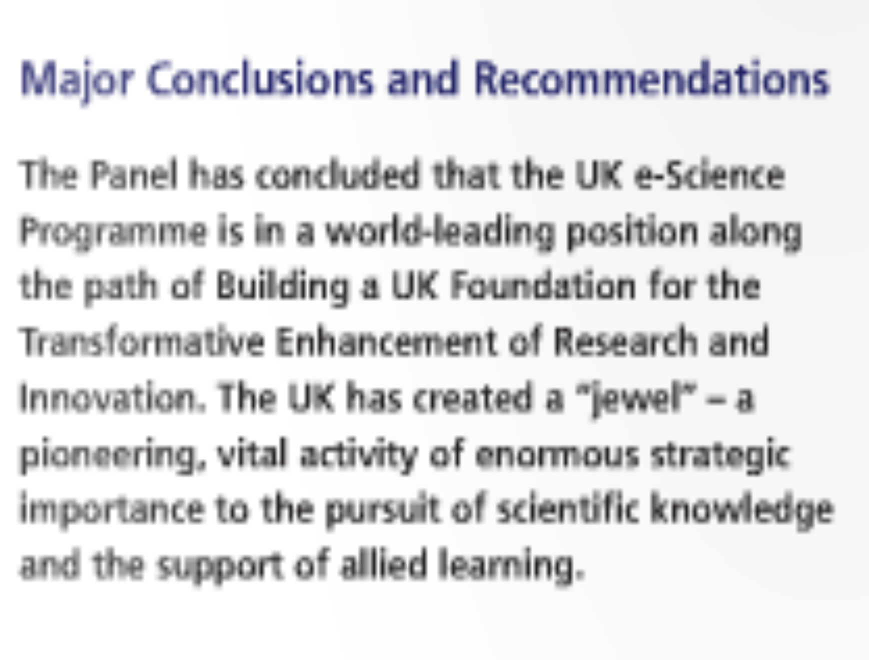
RCUK Review of e-Science 2009

BUILDING A UK FOUNDATION FOR THE TRANSFORMATIVE ENHANCEMENT OF RESEARCH AND INNOVATION

RESEARCH COUNCILS UK

THE ROYAL SOCIETY

e-Science



Major Conclusions and Recommendations

The Panel has concluded that the UK e-Science Programme is in a world-leading position along the path of Building a UK Foundation for the Transformative Enhancement of Research and Innovation. The UK has created a "jewel" – a pioneering, vital activity of enormous strategic importance to the pursuit of scientific knowledge and the support of allied learning.

Chair: Dan Atkins

http://www.epsrc.ac.uk/research/intrevs/escience/Pages/default.aspx

In the US …

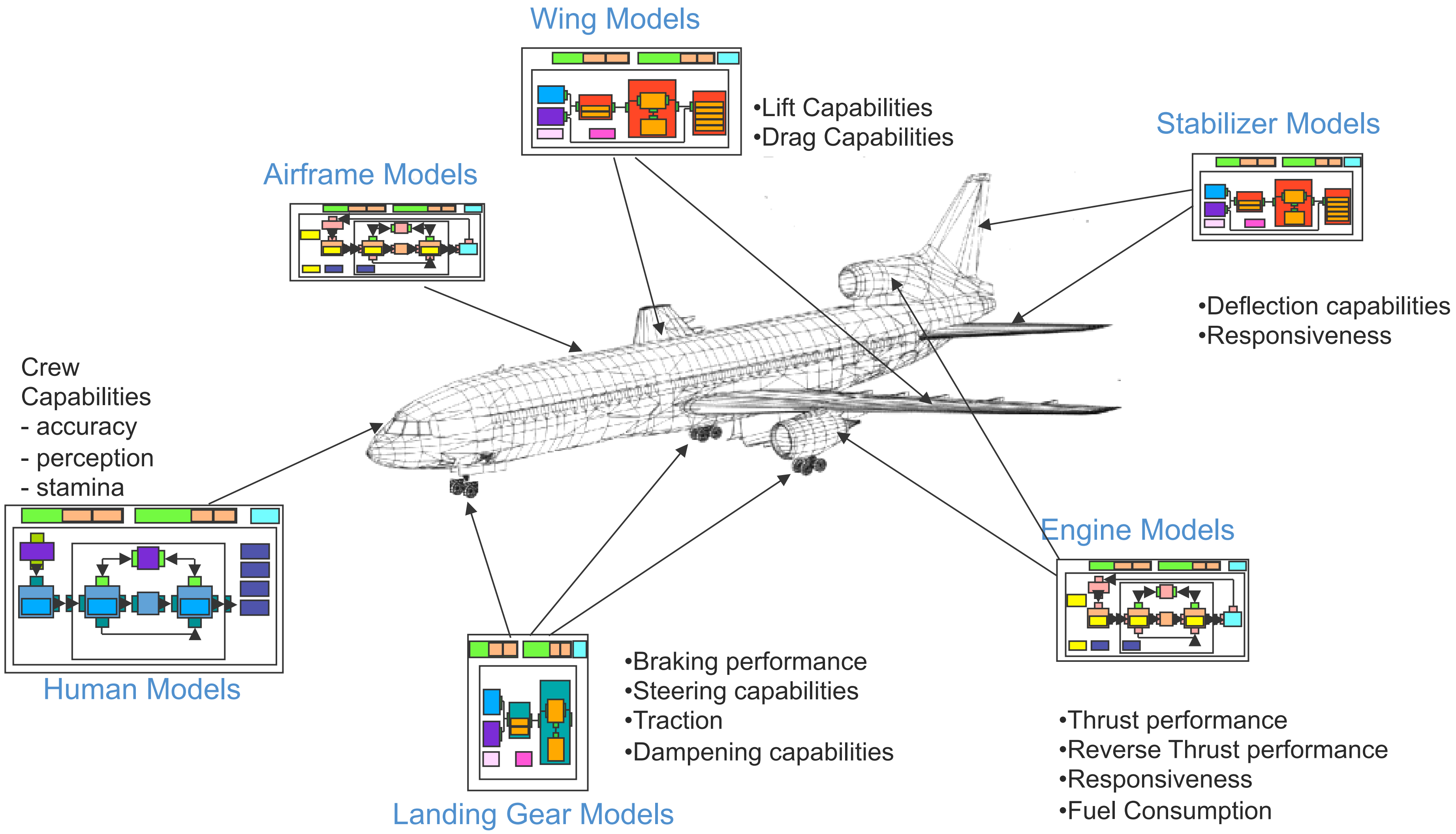
NASA’s Information Power Grid

Vision for the IPG was to promote a revolution in how NASA addresses large-scale science and engineering problems by providing persistent infrastructure for:

- “highly capable” computing and data management services that, on-demand, will locate and co-schedule the multi-Center resources needed to address large-scale and/or widely distributed problems
- the ancillary services that are needed to support the workflow management frameworks that coordinate the processes of distributed science and engineering problems

PIs: Bill Johnstone & Dennis Gannon

Vision for Multi-disciplinary Simulations



Wing Models

- Lift Capabilities
- Drag Capabilities

Stabilizer Models

- Deflection capabilities
- Responsiveness

Airframe Models

Crew Capabilities
- accuracy
- perception
- stamina

Human Models

Landing Gear Models

- Braking performance
- Steering capabilities
- Traction
- Dampening capabilities

Engine Models

- Thrust performance
- Reverse Thrust performance
- Responsiveness
- Fuel Consumption

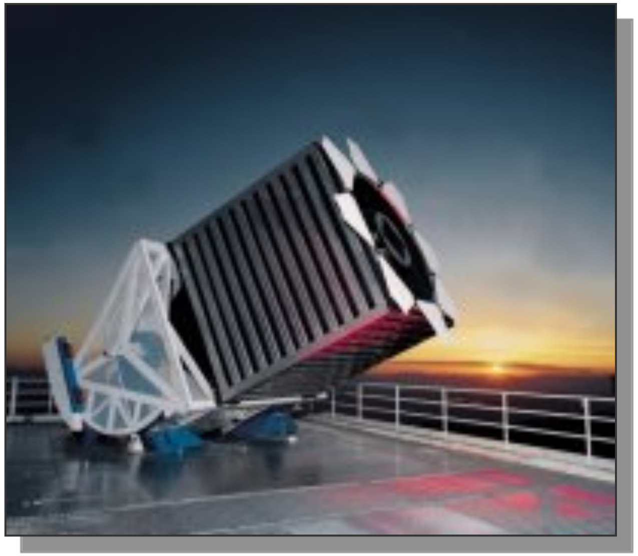
Whole system simulations are produced by coupling all of the sub-system simulations

The ‘Cosmic Genome Project’: The Sloan Digital Sky Survey

- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images
  - 40 TB of raw data => 120TB processed data
  - 5 TB catalogs => 35TB in the end
- Database and spectrograph built at JHU (SkyServer) by a team led by Alex Szalay and Jim Gray
- Started in 1992, finished in 2008

The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg

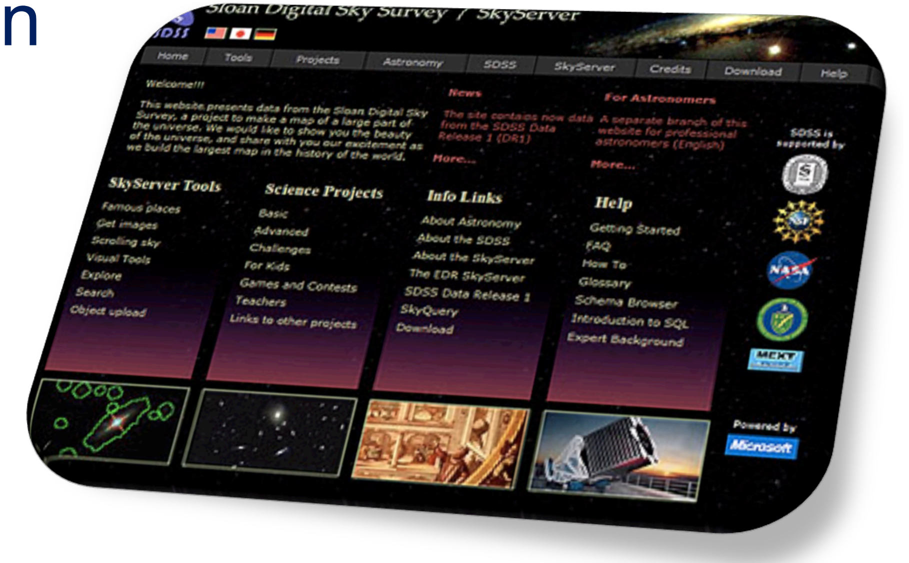
Sloan Foundation, NSF, DOE, NASA



Open Data: Public Use of the Sloan Data

Posterchild in 21st century data publishing

- Set up SkyServer web service
- Over 400 million web hits in 6 years
- About 1M distinct users vs 10,000 astronomers
- >1600 refereed papers!
- Delivered 50,000 hours of lectures to high schools



Welcome!!!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

The site contains new data from the SDSS Data Release 1 (DR1)

More...

For Astronomers

A separate branch of this website for professional astronomers (English)

More...

➢ New publishing paradigm: data is published <u>before</u> analysis by astronomers

eScience:
The Revolution Is Starting

Jim Gray & Alex Szalay
Microsoft & JHU

Presented @ Microsoft eScience Workshop

6-7 November 2005, Redmond, WA.

Emergence of a Fourth Research Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton’s Laws, Maxwell’s Equations…

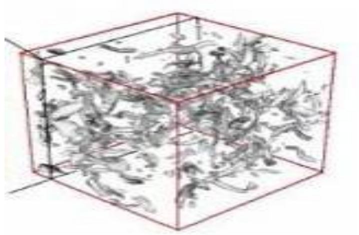
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Captured by instruments
  - Generated by simulations
  - Generated by sensor networks

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

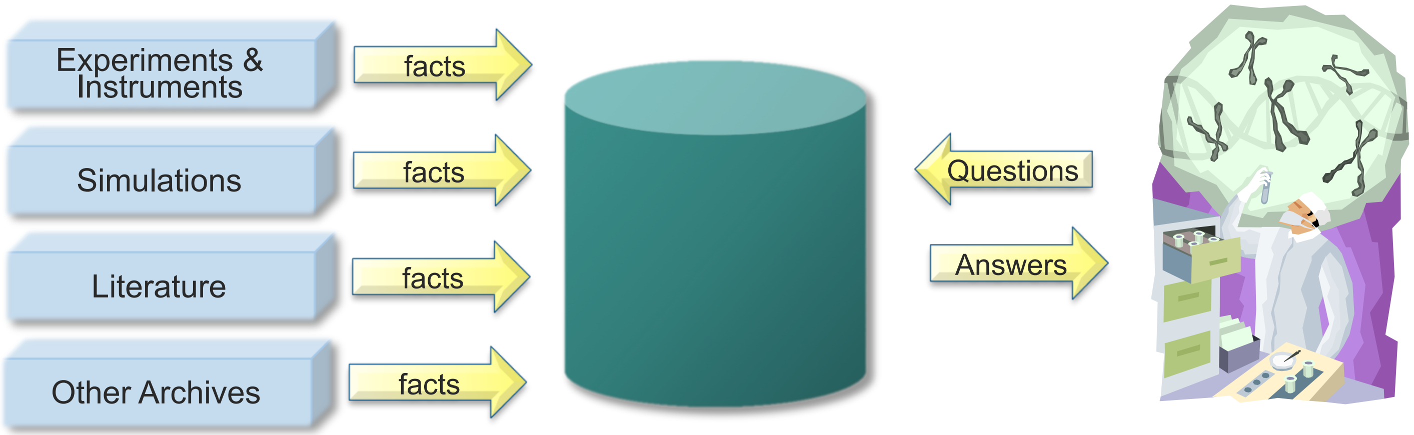


eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge

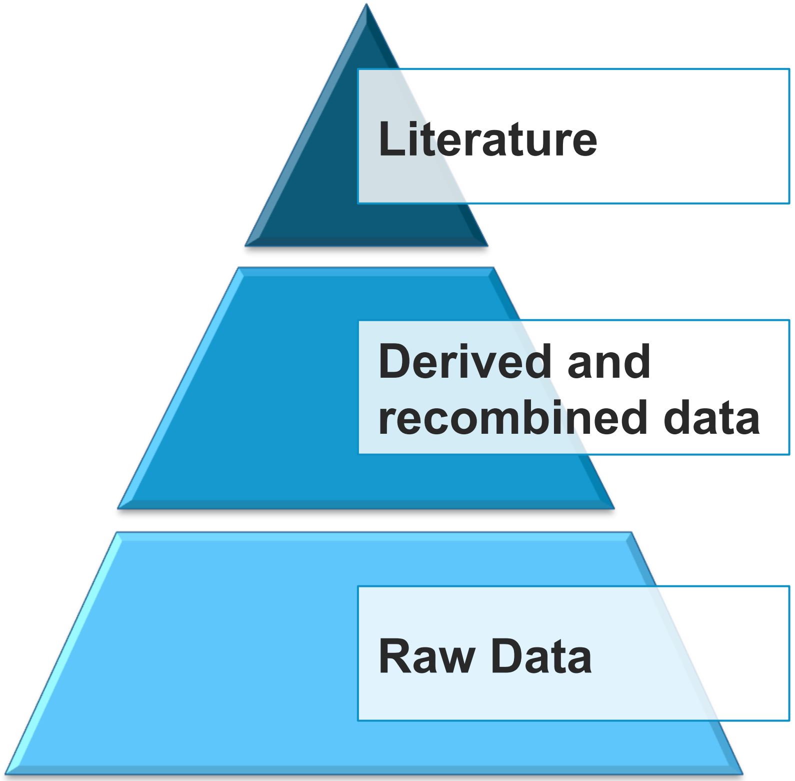


The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

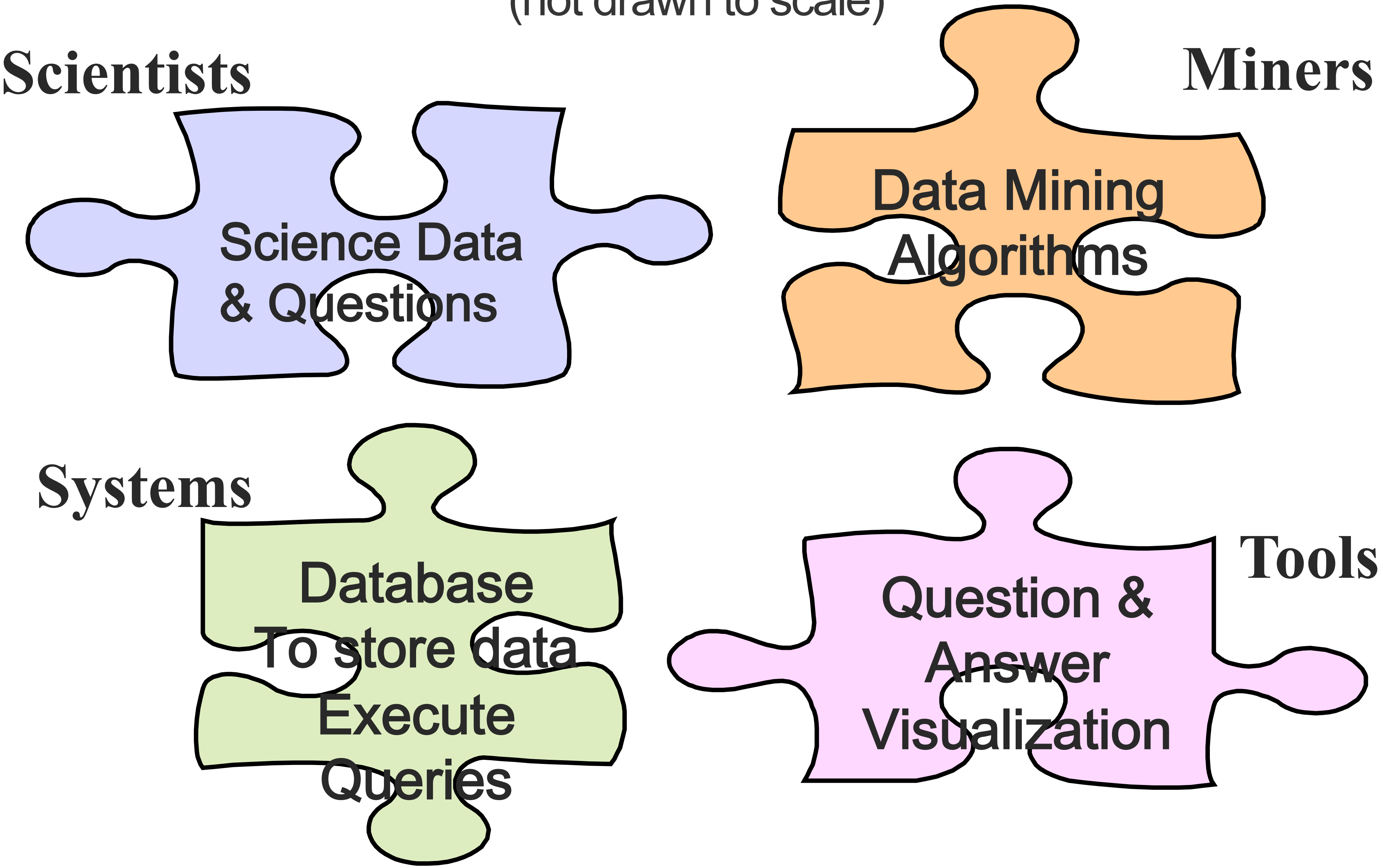
All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips – For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



What X-info Needs from Computer Science

(not drawn to scale)



Scientists

Science Data & Questions

Miners

Data Mining Algorithms

Systems

Database To store data Execute Queries

Tools

Question & Answer Visualization

Working Cross-Culture:
A Way to Engage With Domain Scientists

- Communicate in terms of scenarios
- Work on a problem that gives 100x benefit
  - Weeks/task vs hours/task
- Solve 20% of the problem
  - The other 80% will take decades
- Prototype
- Go from working-to-working: Always have
  - Something to show
  - Clear next steps
  - Clear goal
- Avoid death-by-collaboration-meetings

In Europe …

egee

Enabling Grids for E-sciencE

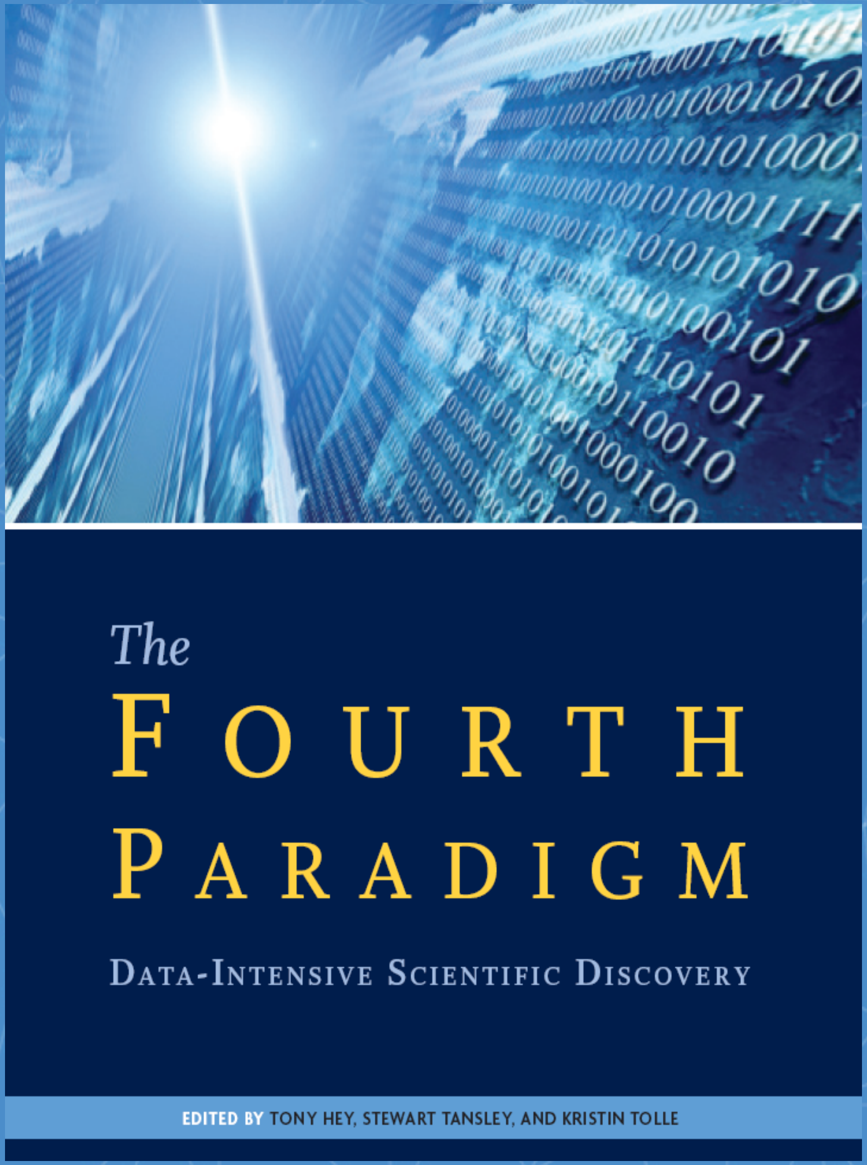
The project will primarily concentrate on three core areas:

- The first area is to build a consistent, robust and secure Grid network that will attract additional computing resources.
- The second area is to continuously improve and maintain the middleware in order to deliver a reliable service to users.
- The third area is to attract new users from industry as well as science and ensure they receive the high standard of training and support they need.

The EGEE Grid will be built on the EU Research Network GÉANT and exploit Grid expertise generated by many EU, national and international Grid projects to date.

PI: Fabrizio Gagliardi

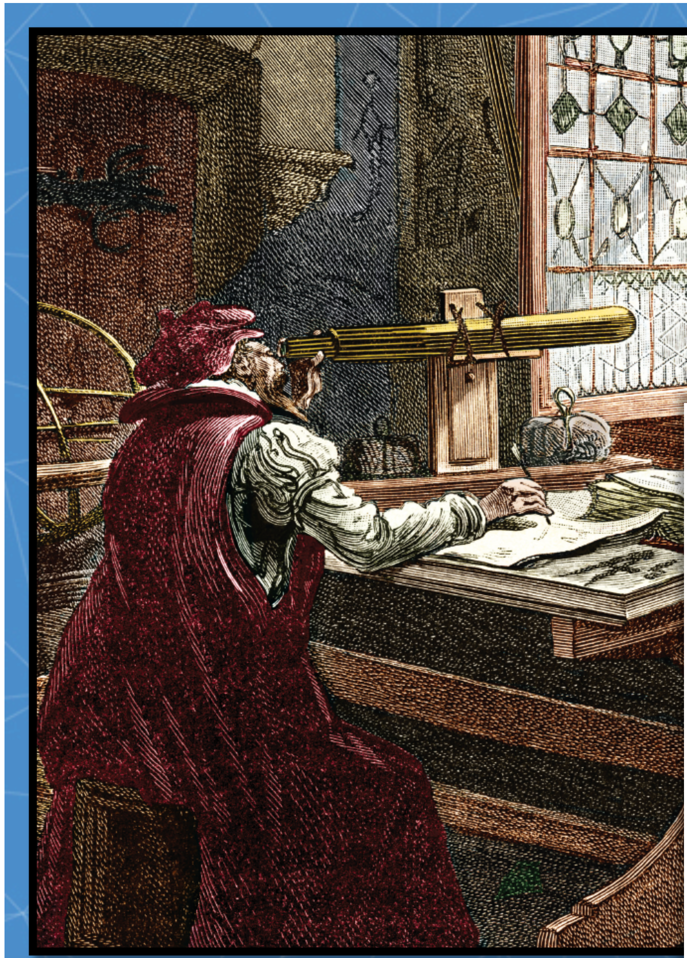
eScience:
The Present



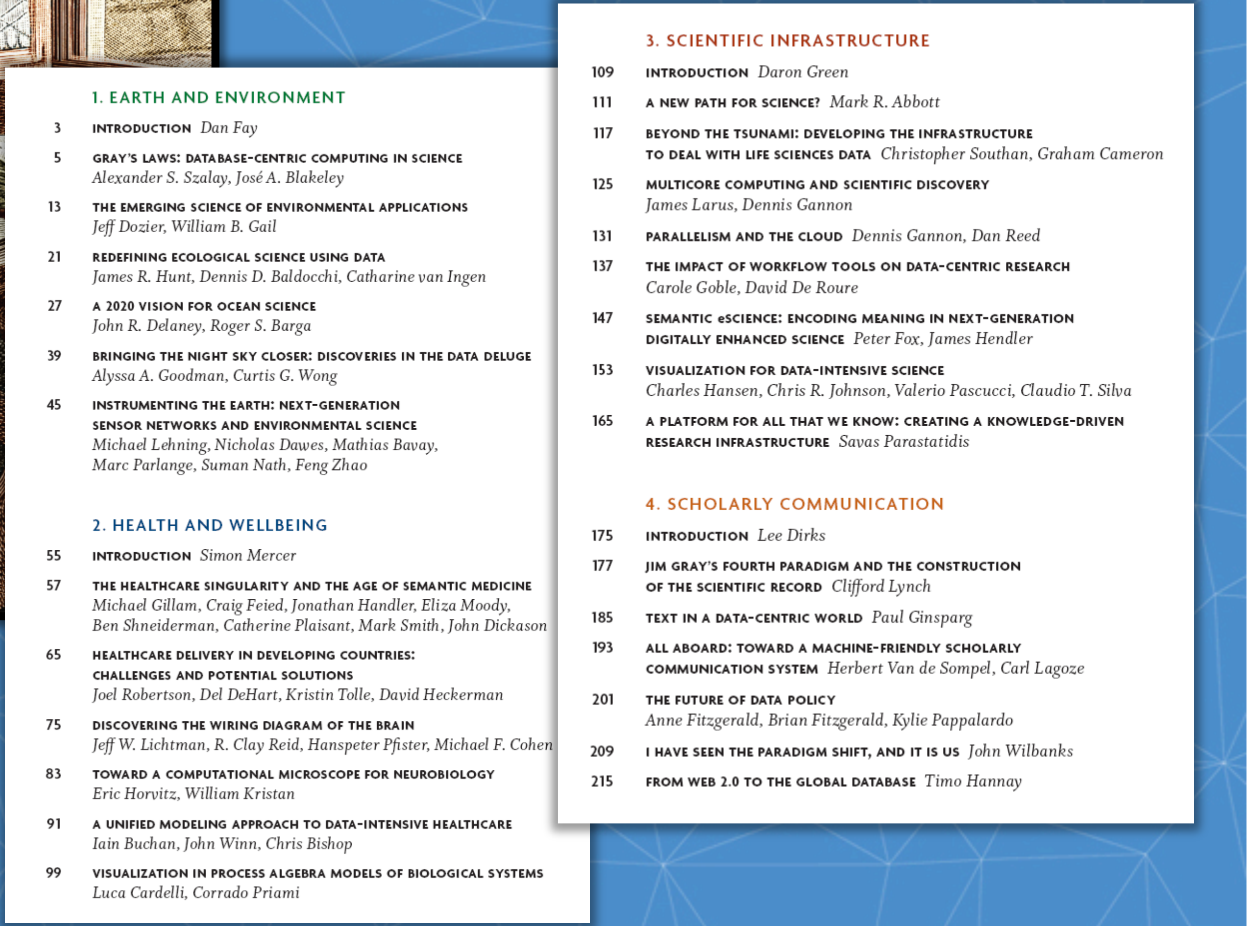
The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



An edited collection of 26 short technical essays, divided into 4 sections



1. EARTH AND ENVIRONMENT

2. HEALTH AND WELLBEING

3. SCIENTIFIC INFRASTRUCTURE

4. SCHOLARLY COMMUNICATION

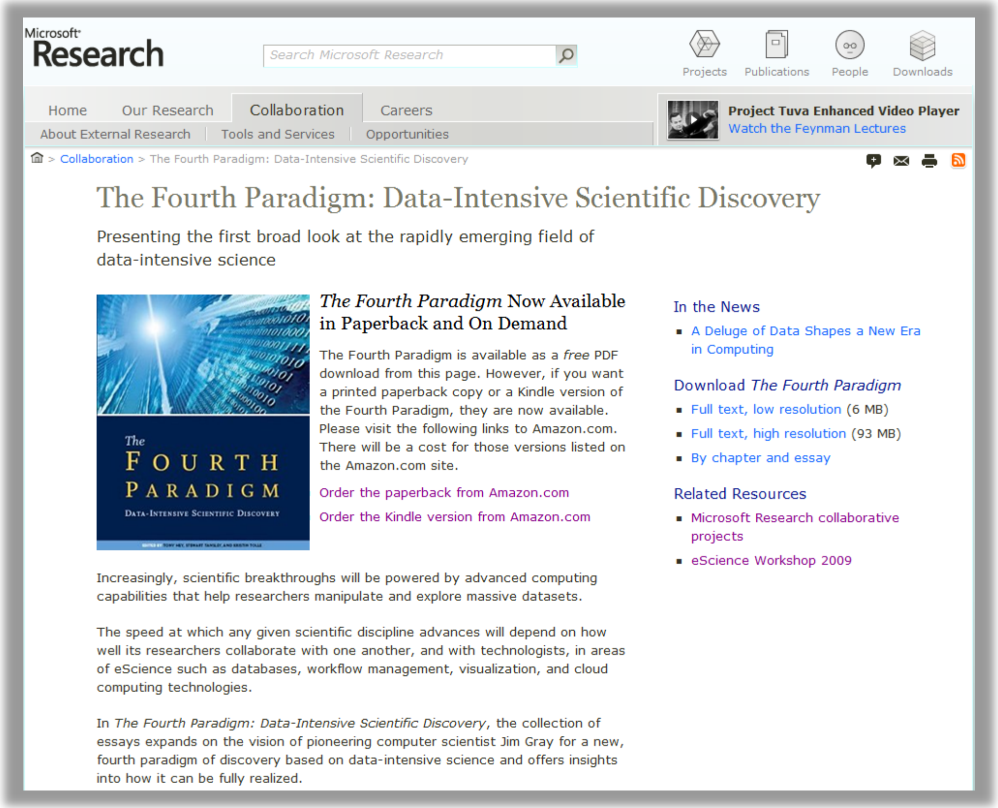
Free PDF Download; Amazon Kindle version; Paperback print-on-demand option

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


Microsoft Research
Search Microsoft Research
Projects
Publications
People
Downloads
Home
Our Research
Collaboration
Careers
About External Research
Tools and Services
Opportunities
Project Tuva Enhanced Video Player
Watch the Feynman Lectures
> Collaboration > The Fourth Paradigm: Data-Intensive Scientific Discovery
The FOURTH PARADIGM
Data-Intensive Scientific Discovery

The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science

The Fourth Paradigm Now Available in Paperback and On Demand

The Fourth Paradigm is available as a *free* PDF download from this page. However, if you want a printed paperback copy or a Kindle version of the Fourth Paradigm, they are now available. Please visit the following links to Amazon.com. There will be a cost for those versions listed on the Amazon.com site.

Order the paperback from Amazon.com

Order the Kindle version from Amazon.com

In the News

- A Deluge of Data Shapes a New Era in Computing

Download *The Fourth Paradigm*

- Full text, low resolution (6 MB)
- Full text, high resolution (93 MB)
- By chapter and essay

Related Resources

- Microsoft Research collaborative projects
- eScience Workshop 2009

Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Published under a Creative Commons License

ADNI ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

- Alzheimer’s Disease Neuroimaging Initiative (ADNI) launched in 2004 specifically to improve clinical trials by different centers agreeing to share data
  - Data fro the 14 different centers involved in the initiative be combined and compared
  - Data is typically made publicly available within a week of being collected
- Hundreds of scientists have made tens of thousands of downloads from the ADNI website

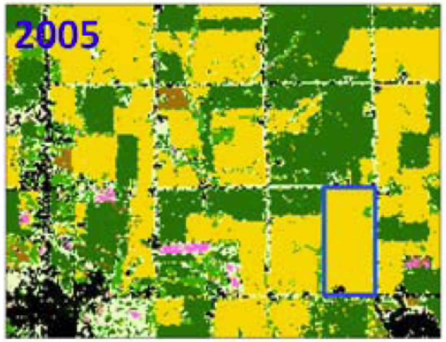
➢ Of several dozen papers that have so far been published using ADNI data, a significant number were authored by researchers who are not even directly funded by the project.

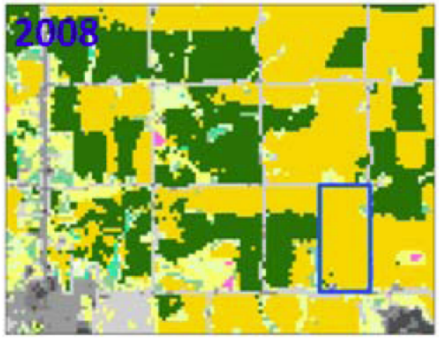
http://www.adni-info.org/

Calculating the Value of Information

Scientists at the U.S. Geological Survey (USGS)

- Developing an economic framework to measure what they call the “VOI” or **Value Of Information**
- Using storehouse of Land Use / Land Cover maps created from Landsat’s moderate resolution land imagery since the early 1970s.





USGS is aiming for a VOI calculation that can inform decisions that maximize agricultural production by:

- Reconciling groundwater pollution hazards with the region’s agricultural needs
- Thereby lowering mitigation and treatment costs necessary to avoid human health and other consequences of contaminated groundwater.

USGS
science for a changing world

ftp://ftpext.usgs.gov/

Funding Data Storage and Analysis

Historically, after a boating or aircraft accident at sea, the U.S. Coast Guard historically has relied on current charts and wind gauges to figure out where to hunt for survivors.

➡

Scientists have been collecting high frequency radar data that can remotely measure ocean surface waves and currents – it is now available to the USCG for rescue operations.

However, a large fraction of the data the Rutgers team collects has to be thrown out because there is no room to store it and no support within existing research projects to better curate and manage the data. **"I can get funding to put equipment into the ocean, but not to analyze that data on the back end,"**

Professor Oscar Schofield
Bio-Optical Oceanography

Citizen Scientists and Data Analysis

Galaxy Zoo activities give a useful indication of the latent appetite for scientific engagement in society. This is a collection of online astronomy projects which invite members of the public to assist in classifying galaxies.

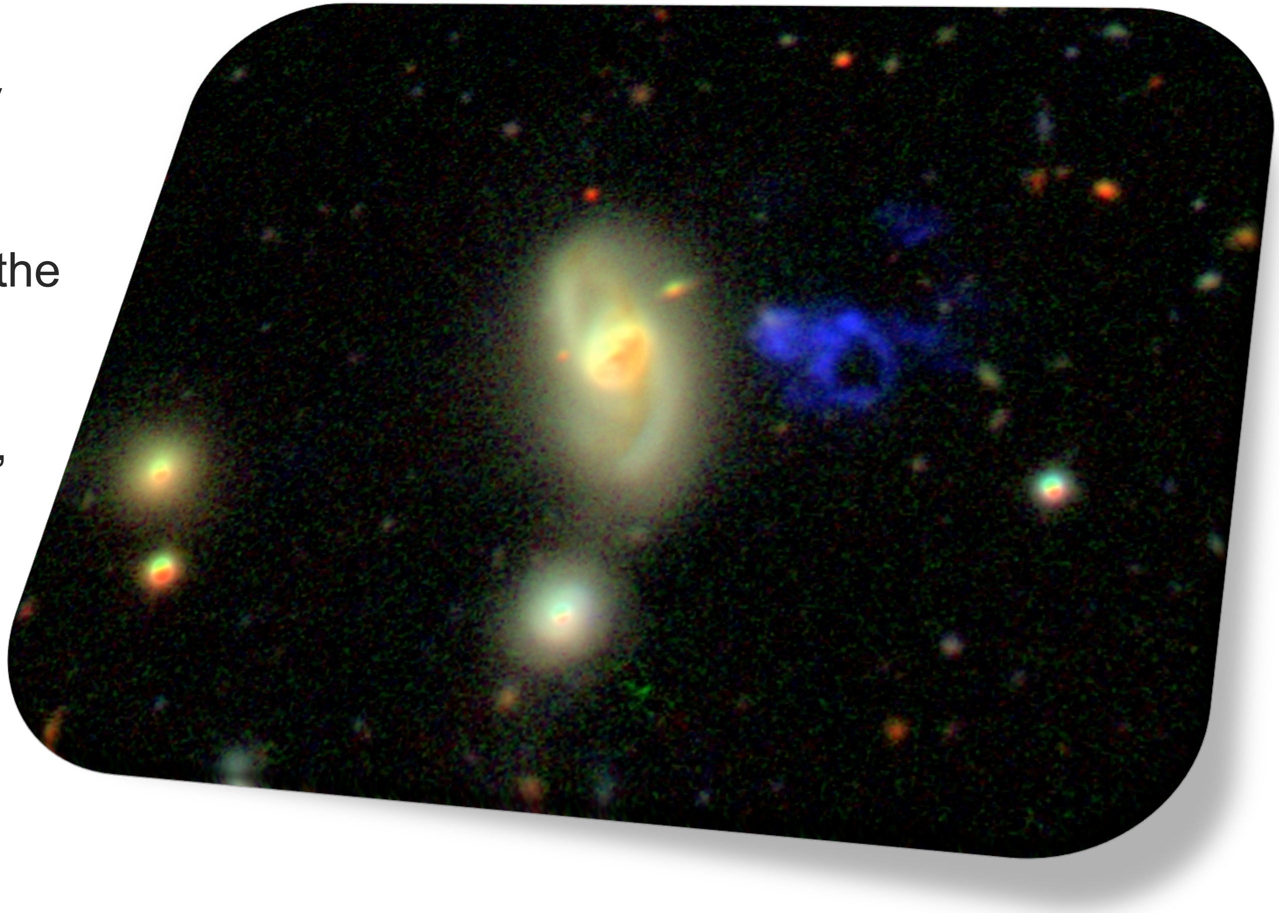
In the first year, **50 million classifications were made by 150,000 individuals in the general public** – it quickly became the world's largest database of galaxy shapes. The original project that it spawned Galaxy Zoo 2 in February 2009 to classify another 250,000 SDSS galaxies.

The project included unique scientific discoveries such as Hanny’s Voorwerp and ‘Green Pea’ galaxies.

GALAXY ZOO.org

Hanny van Arkle’s Voorwerp

Hanny Van Arkel, a Dutch schoolteacher and Galaxy Zoo volunteer, posted an image to the Galaxy Zoo forum and asked "What's the blue stuff below?" No one knew. The object became known as the "**Voorwerp**“, Dutch for "object".



Astro-informatics: A 21st Century Approach to Astronomy

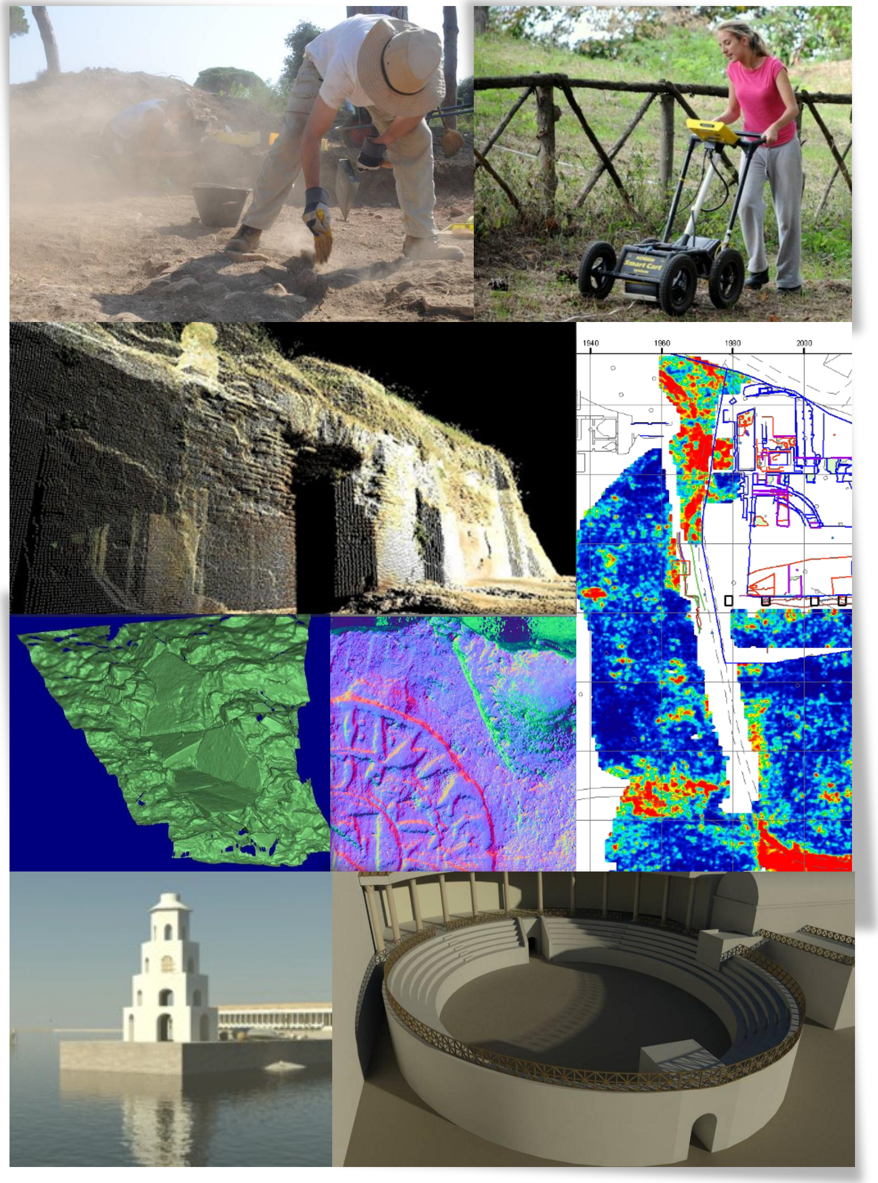
- “We recommend the formal creation, recognition, and support of a major new discipline, which we call Astro-informatics.
- Astro-informatics includes a set of naturally-related specialties including data organization, data description, astronomical classification taxonomies, astronomical concept ontologies, data mining, machine learning, visualization, and astro-statistics.
- We propose that astronomy now needs to integrate Astro-informatics as a formal sub-discipline within agency funding plans, university departments, research programs, graduate training, and undergraduate education.
- Now is the time for the recognition of Astro-informatics as an essential methodology of astronomical research.
- **The future of astronomy depends on it.”**

➢ **Position paper by George Djorgovski and others (September 2009)**

Keynote by Dan Fay, director of E3 at Microsoft Research Connections, on “The Rise of X-Informatics.

Archaeo-Informatics

- Archaeology is about piecing together the past
- Archaeologists must capture and organize artifacts and data
- Multiple sources, from excavation and hand sifting, to advanced geophysics and aerial surveying
- Context is everything
- Ultimately visualize and synthesize
- Advanced computational tools from data management and processing, to analysis and visualization
- Allow breakthroughs such as AHRC Portus Project, Rome



PI: Graeme Earl

eScience Research Group
Microsoft Research

Los Angeles

David Heckerman

Jonathan Carlson

Jennifer Listgarten

Henry Lin

Gabriel Margarido

Curtis Wong

Carl Kadie

Bob Davidson

Nebojsa Jojic

Redmond

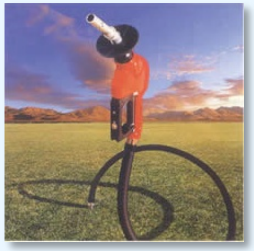
Machine Learning and eScience

Tackling societal challenges

Identifying genetic and environmental causes of disease

Fighting HIV/ AIDS

Increasing energy yield of sugar cane through genome assembly



World Wide Telescope

www.worldwidetelescope.org

Big data requires new types of data visualization tools

Collaborators:

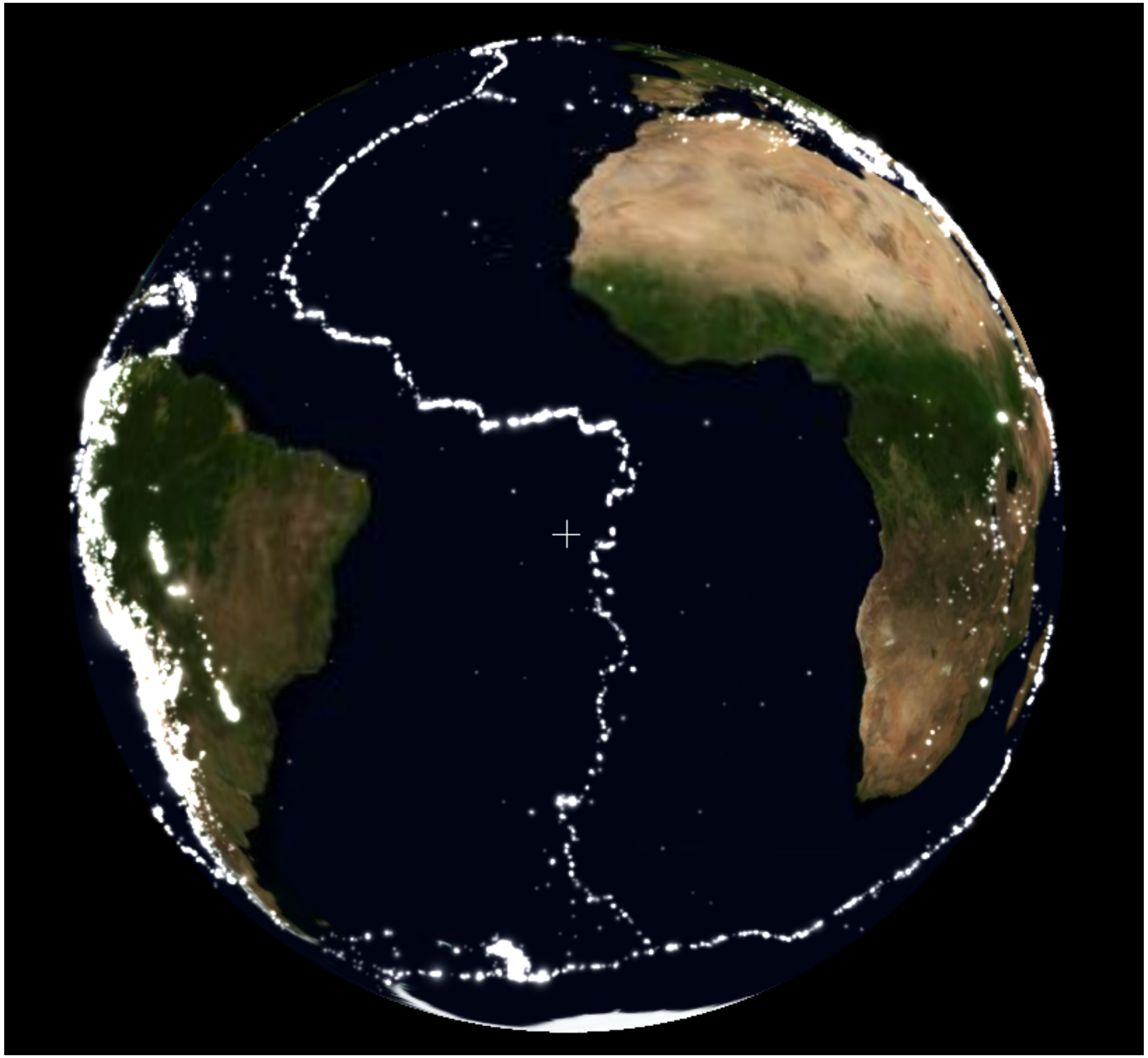
- Alyssa Goodman; Harvard University
- Alex Szalay; Johns Hopkins University
- Curtis Wong, Jonathan Fay; Microsoft Research

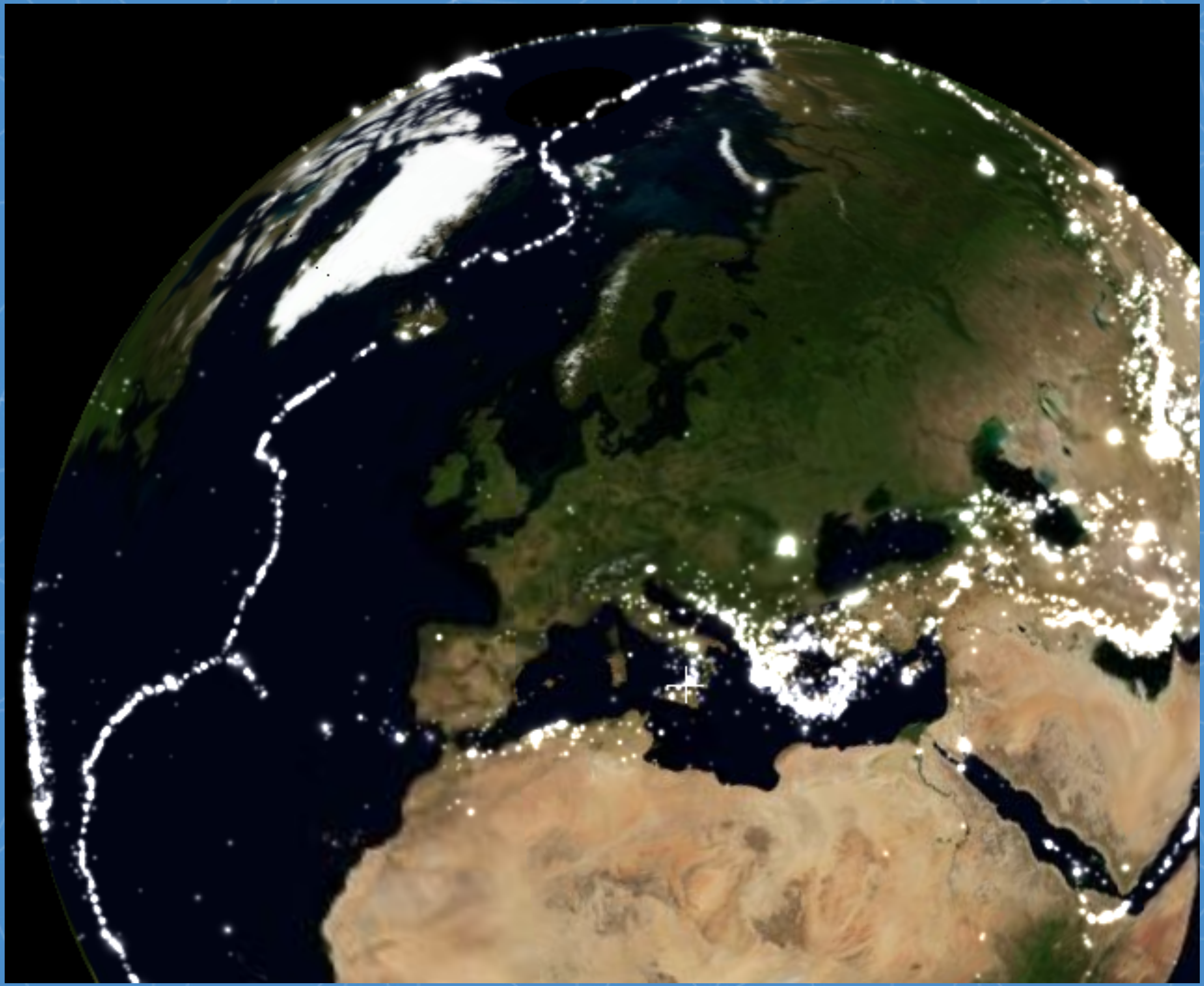
- Integration of data sets and one-click contextual access
- Easy access and use

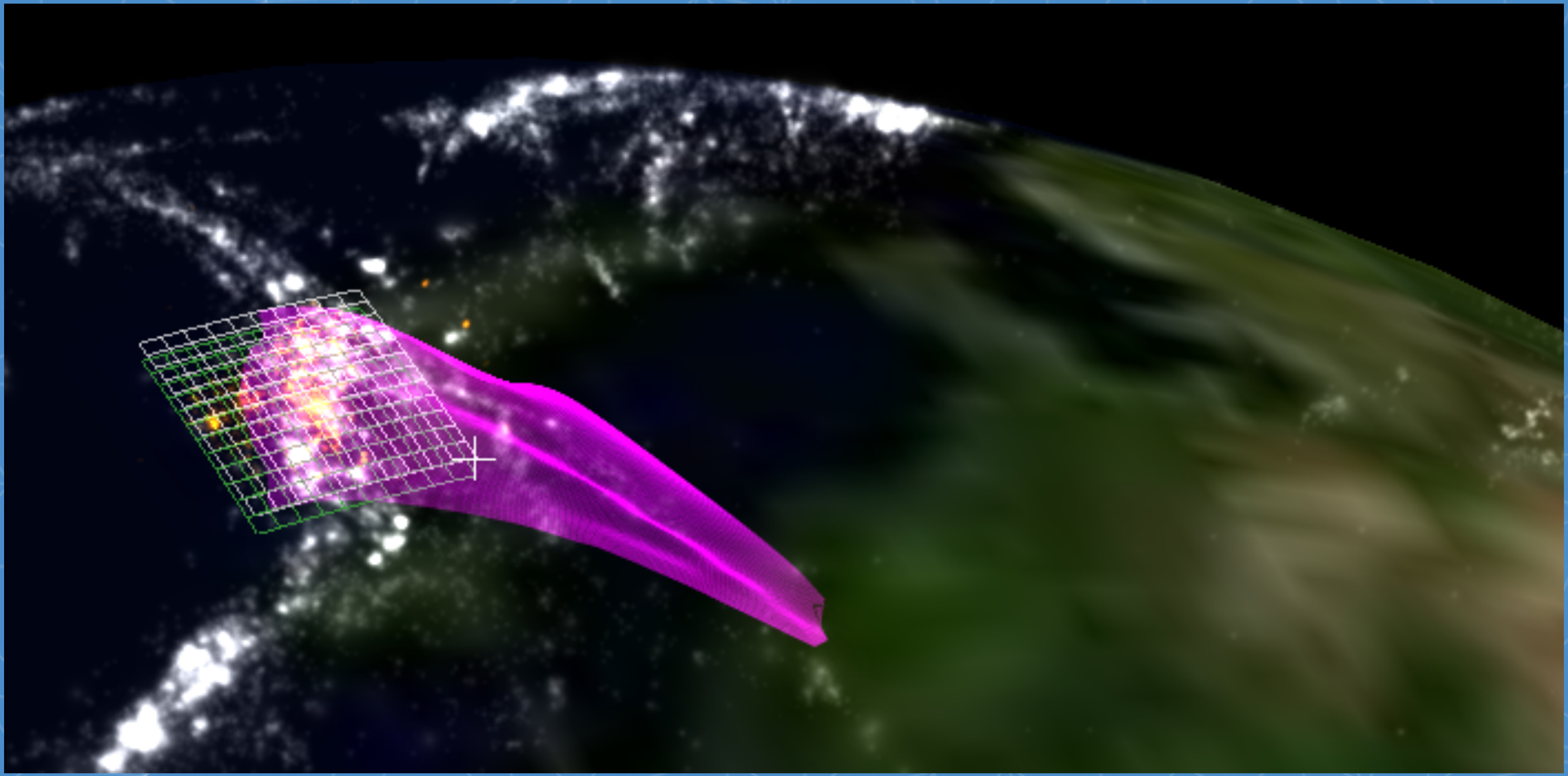
- As of May 2010, over 4M unique users
- Average number of WWT users is over 8K per day

See TED talk by Roy Gould and Curtis Wong
http://www.youtube.com/watch?v=NPu2j3JVmnw&feature=related

Layerscape





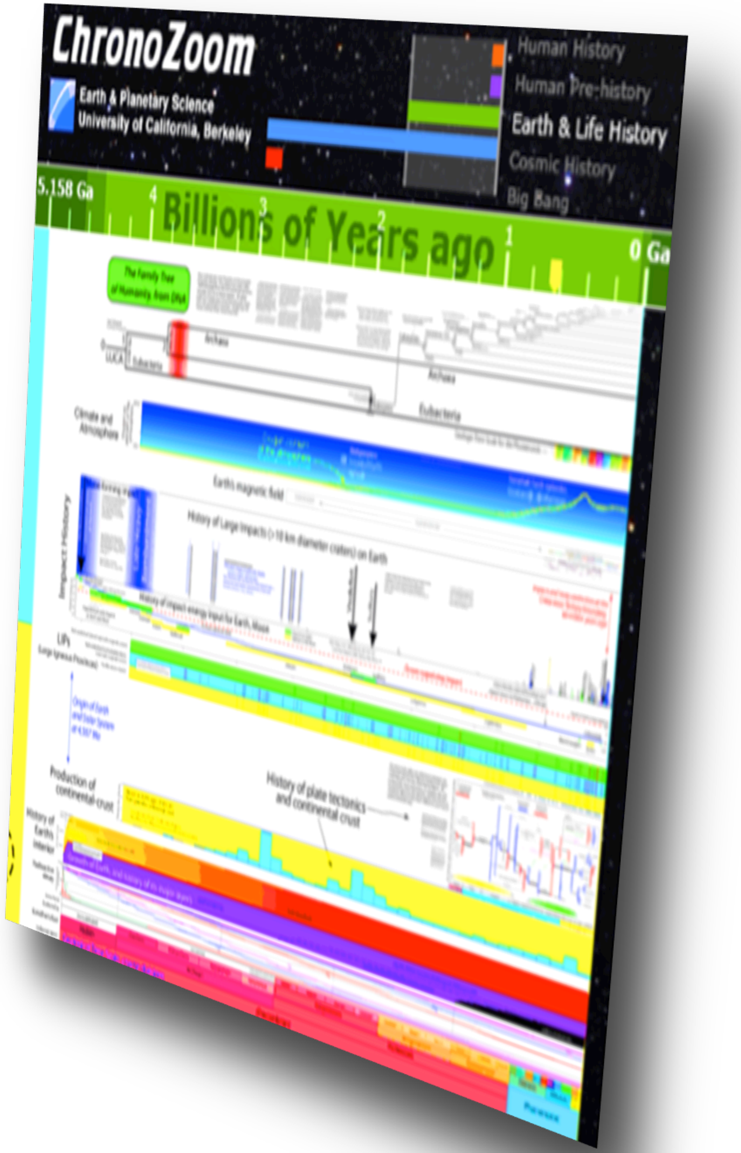


Tohoku events (shallow) and subduction slab

ChronoZoom – History in its broadest possible context …

The challenge: exploration of all known time series data with the ability to smoothly transition from billions of years down to individual nanoseconds…

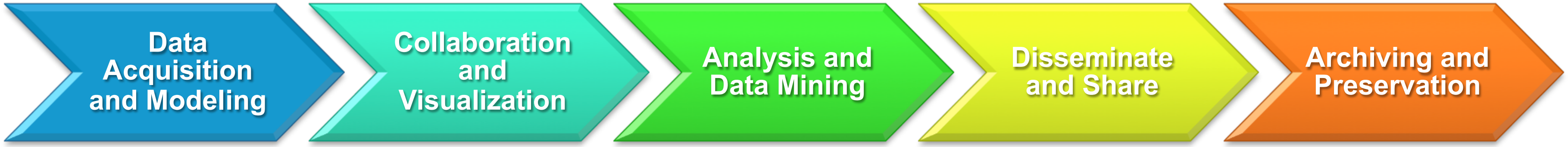
This is what Walter Alvarez, Professor of Earth and Planetary Science at University of Berkeley set out to do.



Our vision is to create an application that allows researchers to browse, overlay, and explore interdisciplinary data sources.

www.chronozoomtimescale.org

Supporting the Data Life Cycle



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, Clouds, etc.
- Relational and non-relational Databases, workflows, provenance …

Support Collaboration

- Allow researchers to work together, share context, facilitate interactions
- Collaboratories/Virtual Organizations

Data Analysis, Data Mining and Visualization

- Data Mining techniques (Machine Learning, OLAP)
- Visualization and visual analytics

Disseminate and Share Research Outputs

- Publish, Present, Blogs, Wikis …
- Review and Rate, social networks, tagging …

Archiving and Preservation

- Published literature, reference data, curated data, etc.
- Digital repositories, semantic computing

OUTERCURVE FOUNDATION

Enable the exchange of code and understanding among software companies and open source communities

Microsoft Research Connections Contributions

Project Trident: Toolset based on Windows Workflow Foundation that provides scientists' need for a flexible, powerful way to analyze large, diverse datasets.

Chemistry Add-in for Word: Chem4Word is an add-in for Microsoft Word that enables semantic authoring of chemical structures.

ConferenceXP: Platform for real-time collaboration that seamlessly connects people or groups over a network, providing high-quality, low-latency videoconferencing and a rich set of collaboration capabilities.

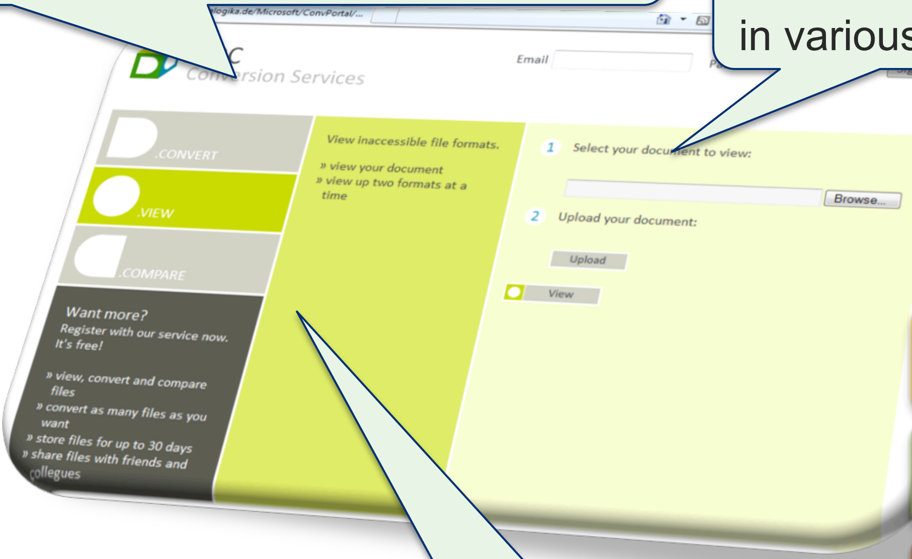
.NET Bio: This open-source platform features a library of commonly used bioinformatics functions plus applications built upon that framework, and can be extended by using any Microsoft .NET language.

http://www.outercurve.org/

Archiving and Preservation:
A Document Conversion Service

Convert Word Perfect and Word documents to OOXML, ODF and UOF

View documents in various formats



View inaccessible file formats.

» view your document
» view up two formats at a time

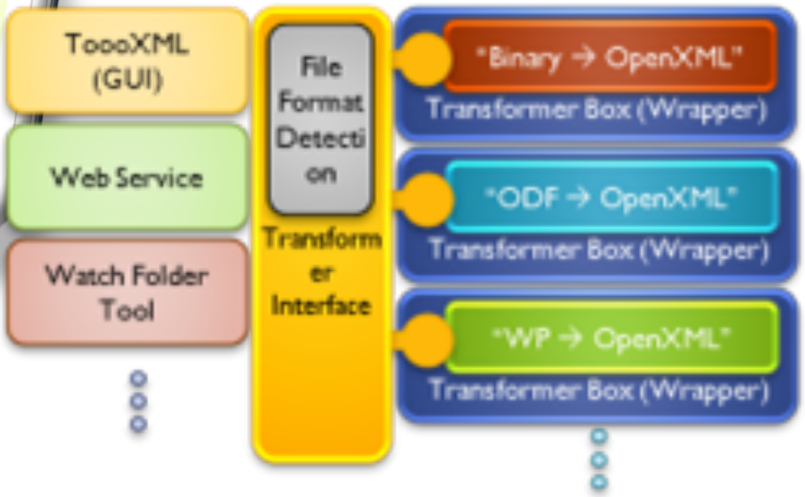
1 Select your document to view:

2 Upload your document:

Want more?
Register with our service now. It's free!

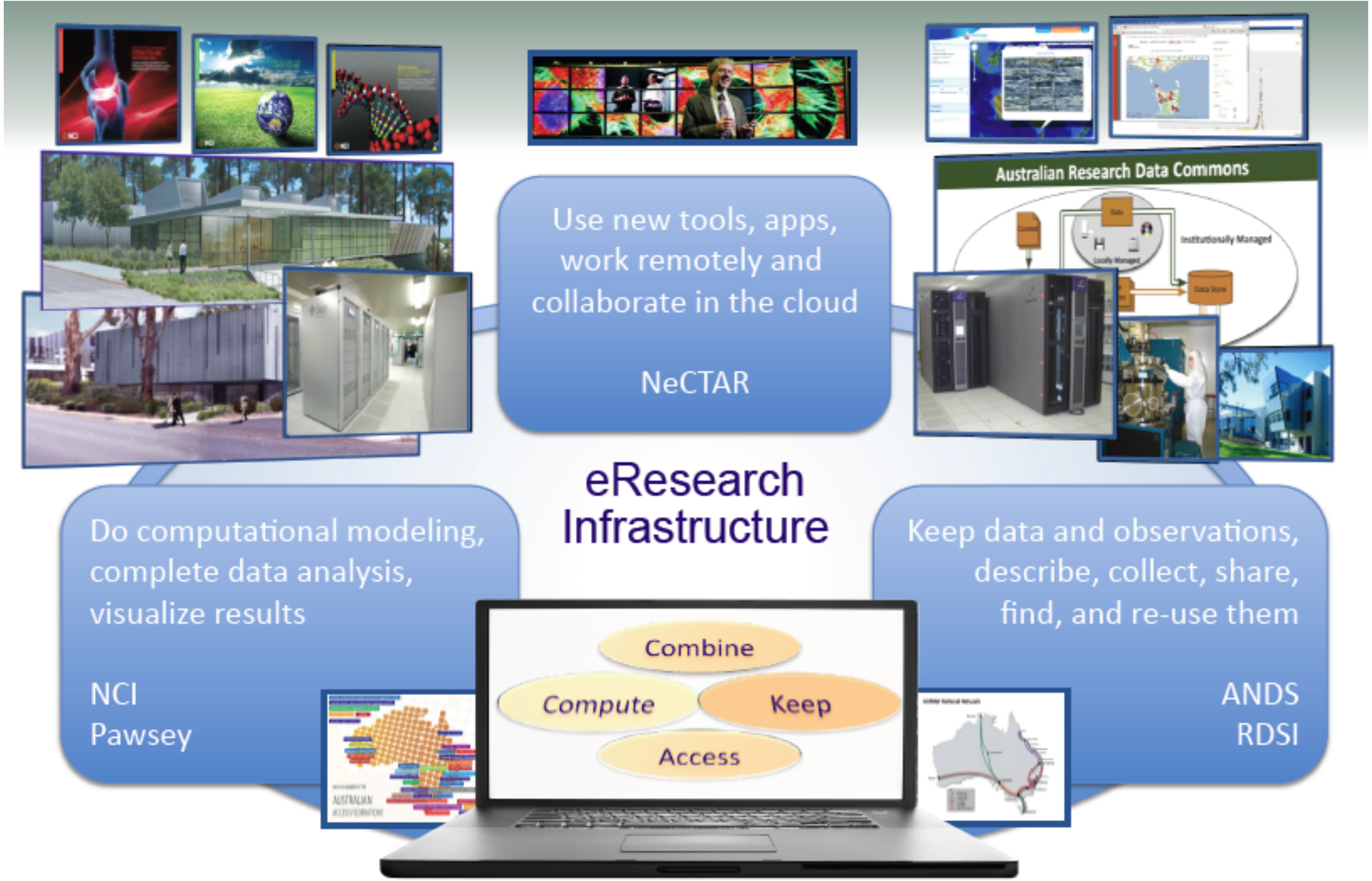
» view, convert and compare files
» convert as many files as you want
» store files for up to 30 days
» share files with friends and collegues

Compare original and converted documents



http://odf-converter.sourceforge.net/

Australian eResearch Program



Use new tools, apps, work remotely and collaborate in the cloud

NeCTAR

eResearch Infrastructure

Do computational modeling, complete data analysis, visualize results

NCI
Pawsey

Keep data and observations, describe, collect, share, find, and re-use them

ANDS
RDSI

Understand mechanisms impossible to observe or experiment with directly

Undertake novel research studies more extensive than ever before

Generate new theories using data at scales previously inconceivable

Swedish e-Science Research

- e-Science one of 20 strategic research areas funded by the Swedish government during 2010—2014
- SeRC: **KTH** – LiU – SU – KI, *30 MSEK/year*
- eSSENCE: **UU** – Umeå – Lund, *26 MSEK/ year*
- CheSC: Chalmers initiative
- Infrastructure funded separately through SNIC
  - HPC-centers PDC and NSC part of SeRC
  - UPPMAX, HPC2N, Lunarc part of eSSENCE
  - C3SE part of CheSC

UPPSALA UNIVERSITET

LUND UNIVERSITY

UMEÅ UNIVERSITET

KTH VETENSKAP OCH KONST

Linköpings universitet

Stockholms universitet

Karolinska Institutet

essence THE e-SCIENCE COLLABORATION

Microsoft Research Connections

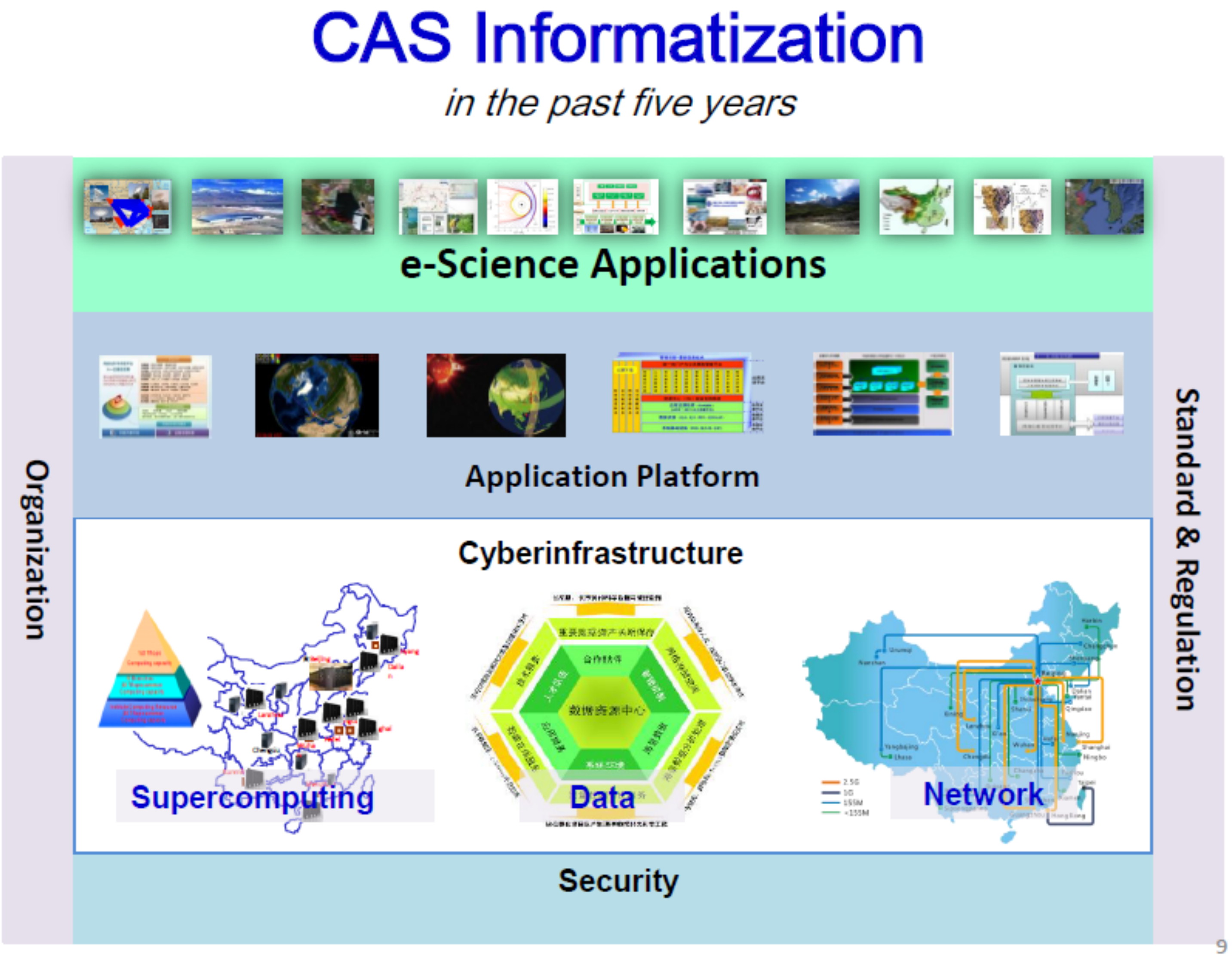
Netherlands eScience Center

Mission:

- The Netherlands eScience Center reinforces and accelerates multi-disciplinary and data-intensive research in the Netherlands by developing and applying eScience and by combining forces.
- The eScience Center breaks down the barriers between traditional disciplines and ICT technologies. In doing so, eScience is both a catalyst and "enabler".
- The eScience Center will ensure that scientific ICT – in all its forms – is deployed sustainably in support of the modern research process.
- The eScience Center develops relevant techniques, algorithms, models, and concepts that can be generalised and shared.
- The basic principles are to make research breakthroughs possible and to generate value for society.

A joint initiative of SURF and NWO

Chinese Academy of Sciences



CAS Informatization

in the past five years

e-Science Applications

Application Platform

Cyberinfrastructure

Supercomputing

Data

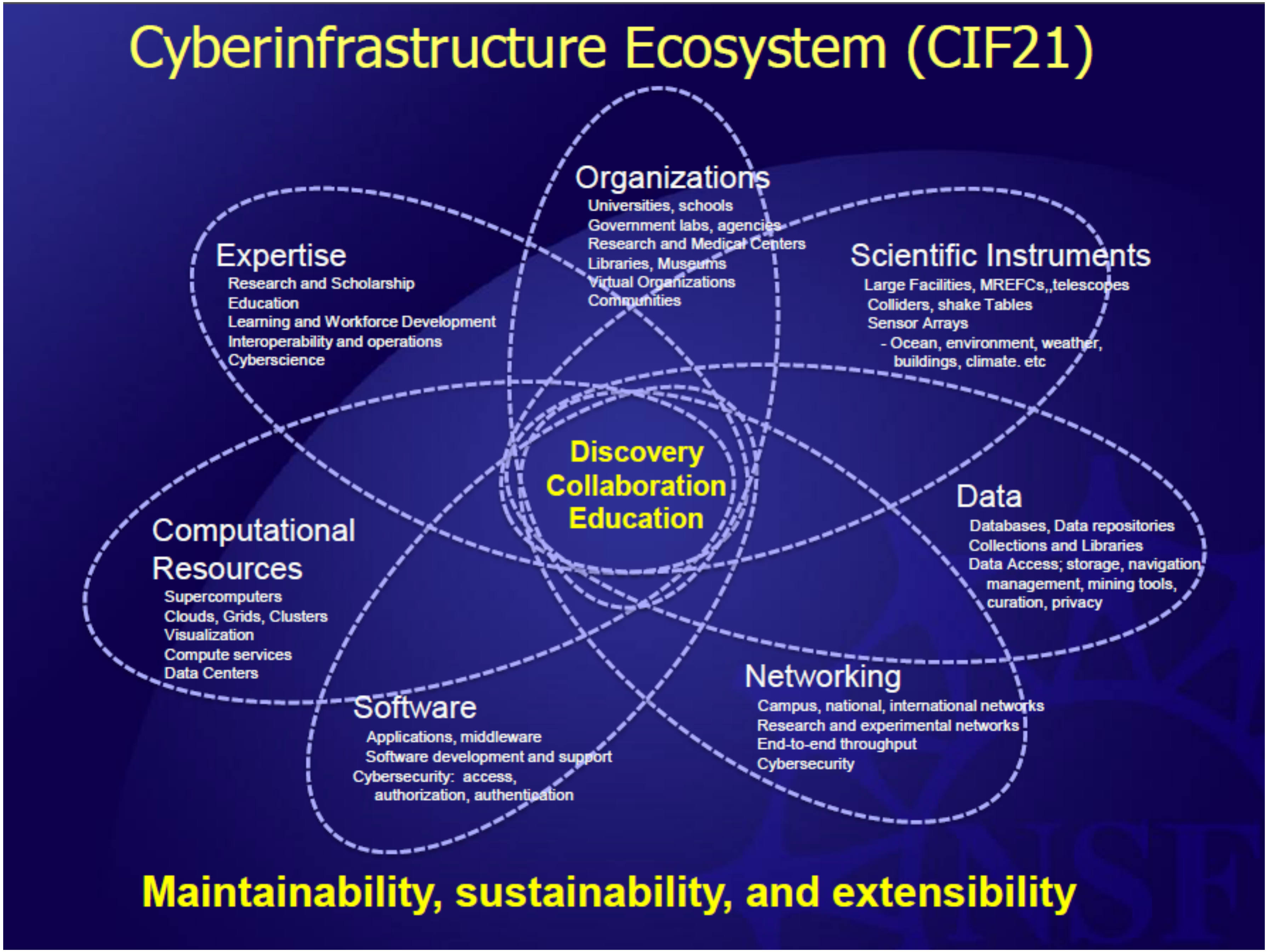
Network

Security

Organization

Standard & Regulation

9



Cyberinfrastructure Ecosystem (CIF21)

Organizations
Universities, schools
Government labs, agencies
Research and Medical Centers
Libraries, Museums
Virtual Organizations
Communities

Expertise
Research and Scholarship
Education
Learning and Workforce Development
Interoperability and operations
Cyberscience

Scientific Instruments
Large Facilities, MREFCs,,telescopes
Colliders, shake Tables
Sensor Arrays
- Ocean, environment, weather, buildings, climate. etc

Discovery
Collaboration
Education

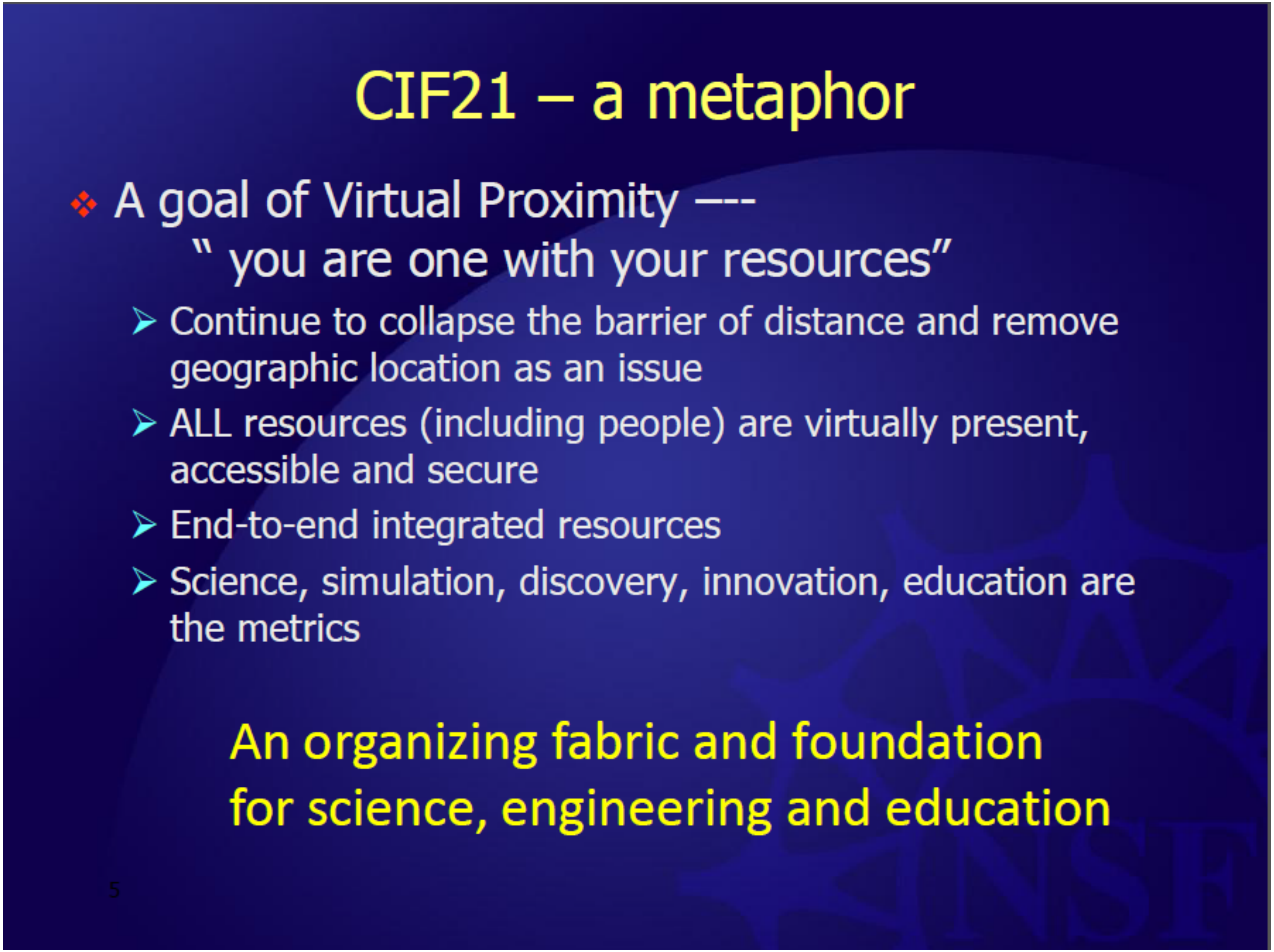
Computational Resources
Supercomputers
Clouds, Grids, Clusters
Visualization
Compute services
Data Centers

Data
Databases, Data repositories
Collections and Libraries
Data Access; storage, navigation
management, mining tools,
curation, privacy

Software
Applications, middleware
Software development and support
Cybersecurity: access,
authorization, authentication

Networking
Campus, national, international networks
Research and experimental networks
End-to-end throughput
Cybersecurity

Maintainability, sustainability, and extensibility

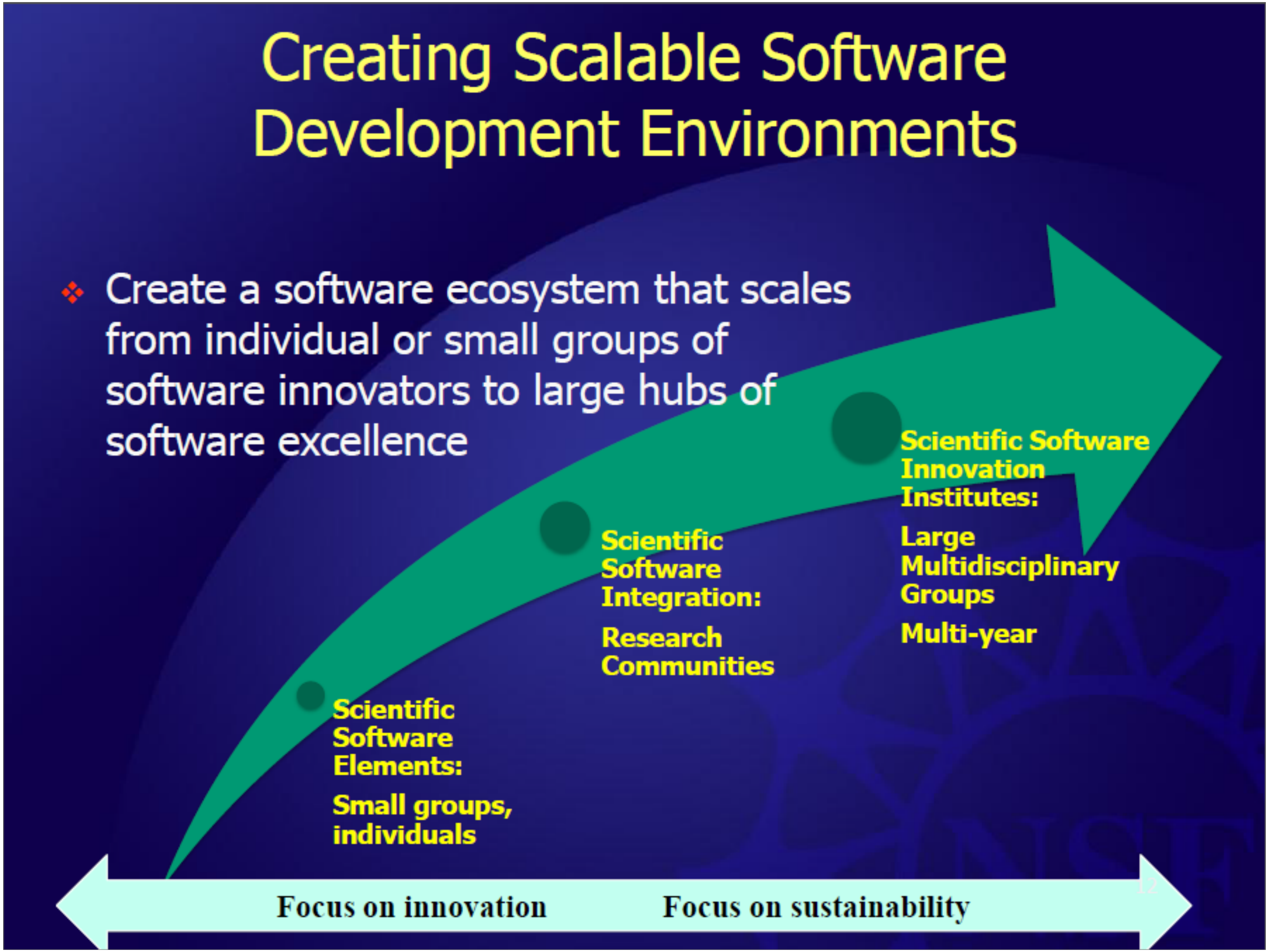


CIF21 – a metaphor

- A goal of Virtual Proximity ---
  " you are one with your resources"
  - Continue to collapse the barrier of distance and remove geographic location as an issue
  - ALL resources (including people) are virtually present, accessible and secure
  - End-to-end integrated resources
  - Science, simulation, discovery, innovation, education are the metrics

An organizing fabric and foundation
for science, engineering and education

5



Creating Scalable Software Development Environments

- Create a software ecosystem that scales from individual or small groups of software innovators to large hubs of software excellence

Scientific Software Elements:

Small groups, individuals

Scientific Software Integration:

Research Communities

Scientific Software Innovation Institutes:

Large Multidisciplinary Groups

Multi-year

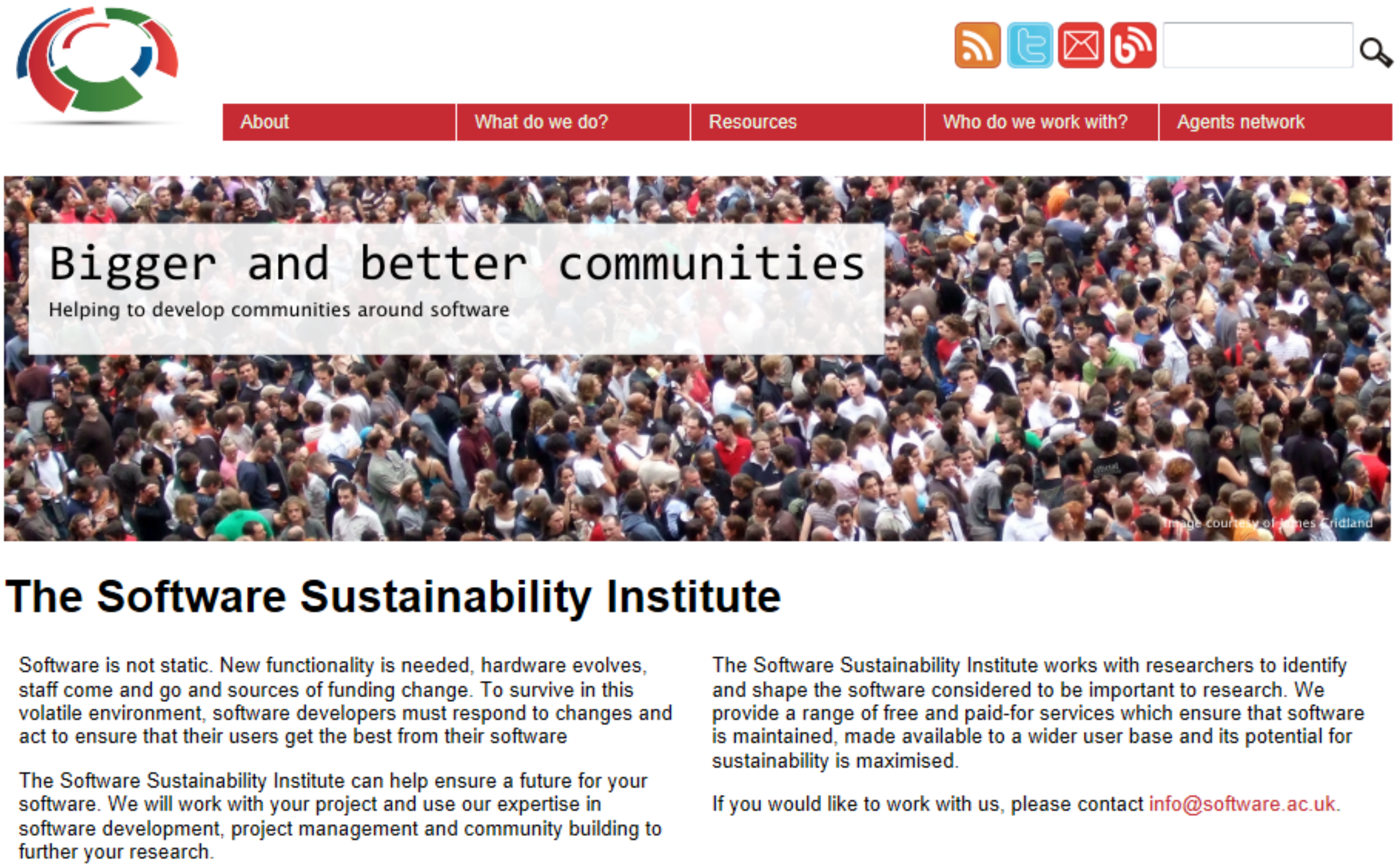
Focus on innovation **Focus on sustainability**

NSF Program Solicitation: Software Infrastructure for Sustained Innovation (SI^2)

The SI^2 program includes three classes of awards:

1. **Scientific Software Elements (SSE):** SSE awards target small groups that will create and deploy robust software elements for which there is a demonstrated need that will advance one or more significant areas of science and engineering.
2. **Scientific Software Integration (SSI):** SSI awards target larger, interdisciplinary teams organized around the development and application of common software infrastructure aimed at solving common research problems. SSI awards will result in sustainable community software frameworks serving a diverse community.
3. **Scientific Software Innovation Institutes (S2I2):** S2I2 awards will focus on the establishment of long-term hubs of excellence in software infrastructure and technologies that will serve a research community of substantial size and disciplinary breadth.

In UK, OMII → SSI



Bigger and better communities

Helping to develop communities around software

The Software Sustainability Institute

Software is not static. New functionality is needed, hardware evolves, staff come and go and sources of funding change. To survive in this volatile environment, software developers must respond to changes and act to ensure that their users get the best from their software

The Software Sustainability Institute can help ensure a future for your software. We will work with your project and use our expertise in software development, project management and community building to further your research.

The Software Sustainability Institute works with researchers to identify and shape the software considered to be important to research. We provide a range of free and paid-for services which ensure that software is maintained, made available to a wider user base and its potential for sustainability is maximised.

If you would like to work with us, please contact info@software.ac.uk.

NSF-OCI Task Force on Data and Visualization

National Science Foundation
Advisory Committee for Cyberinfrastructure
Task Force on Data and Visualization

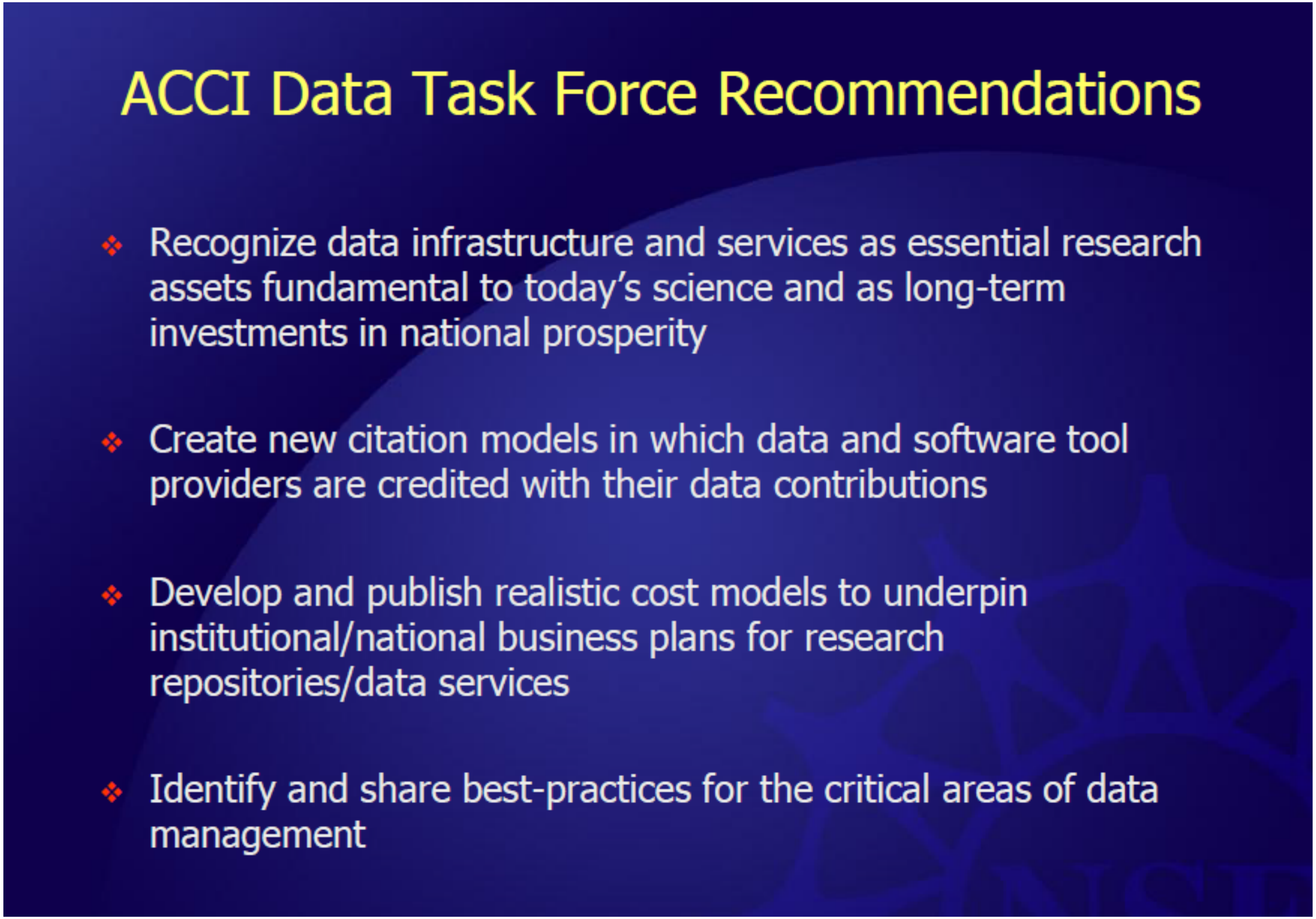
Final Report, March 2011

http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf

Advisory Committee on Cyberinfrastructure

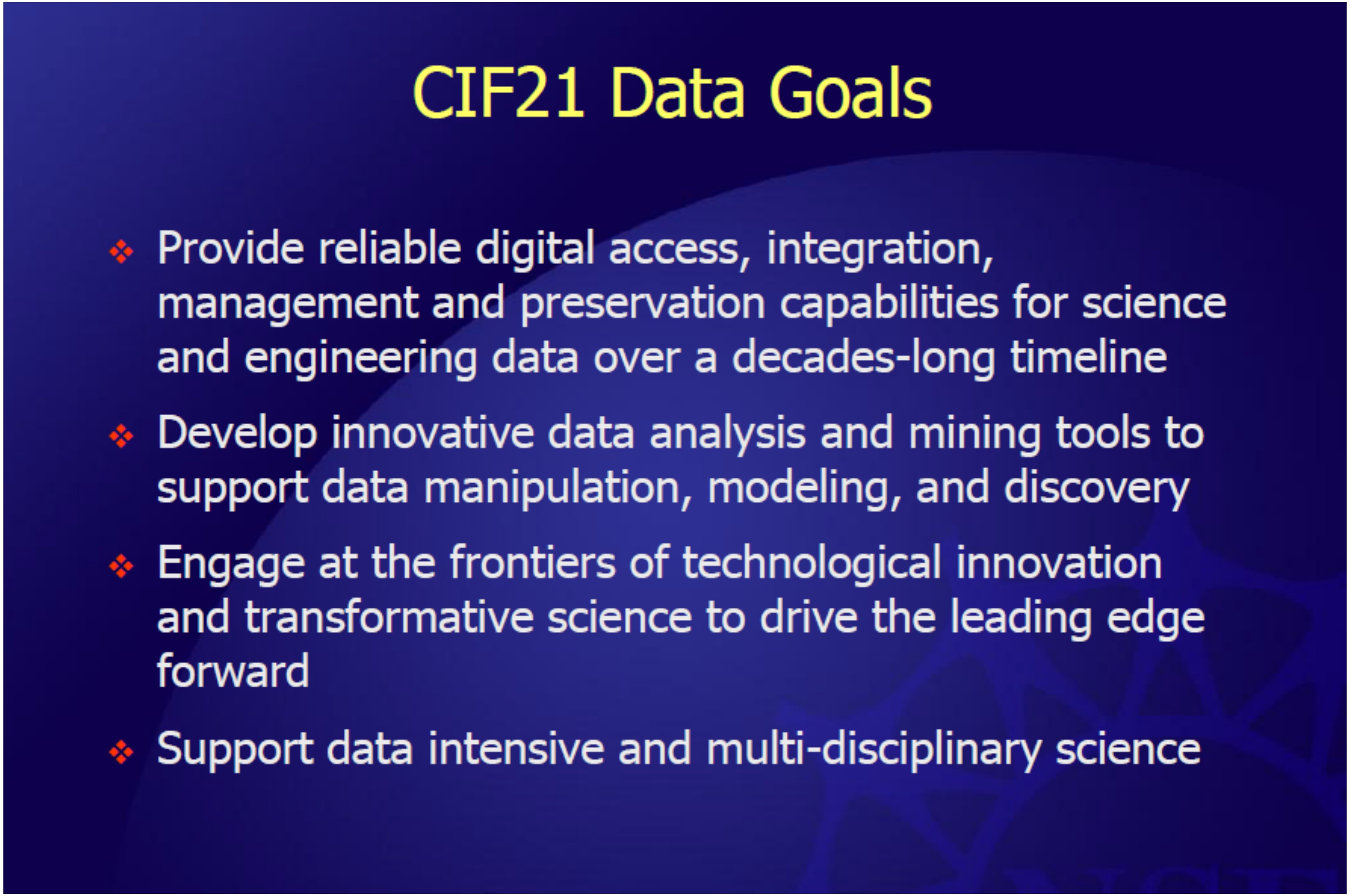
March 2011

Tony Hey, Co-Chair
Microsoft Corporation
Dan Atkins, Co-Chair
University of Michigan
Margaret Hedstrom
University of Michigan



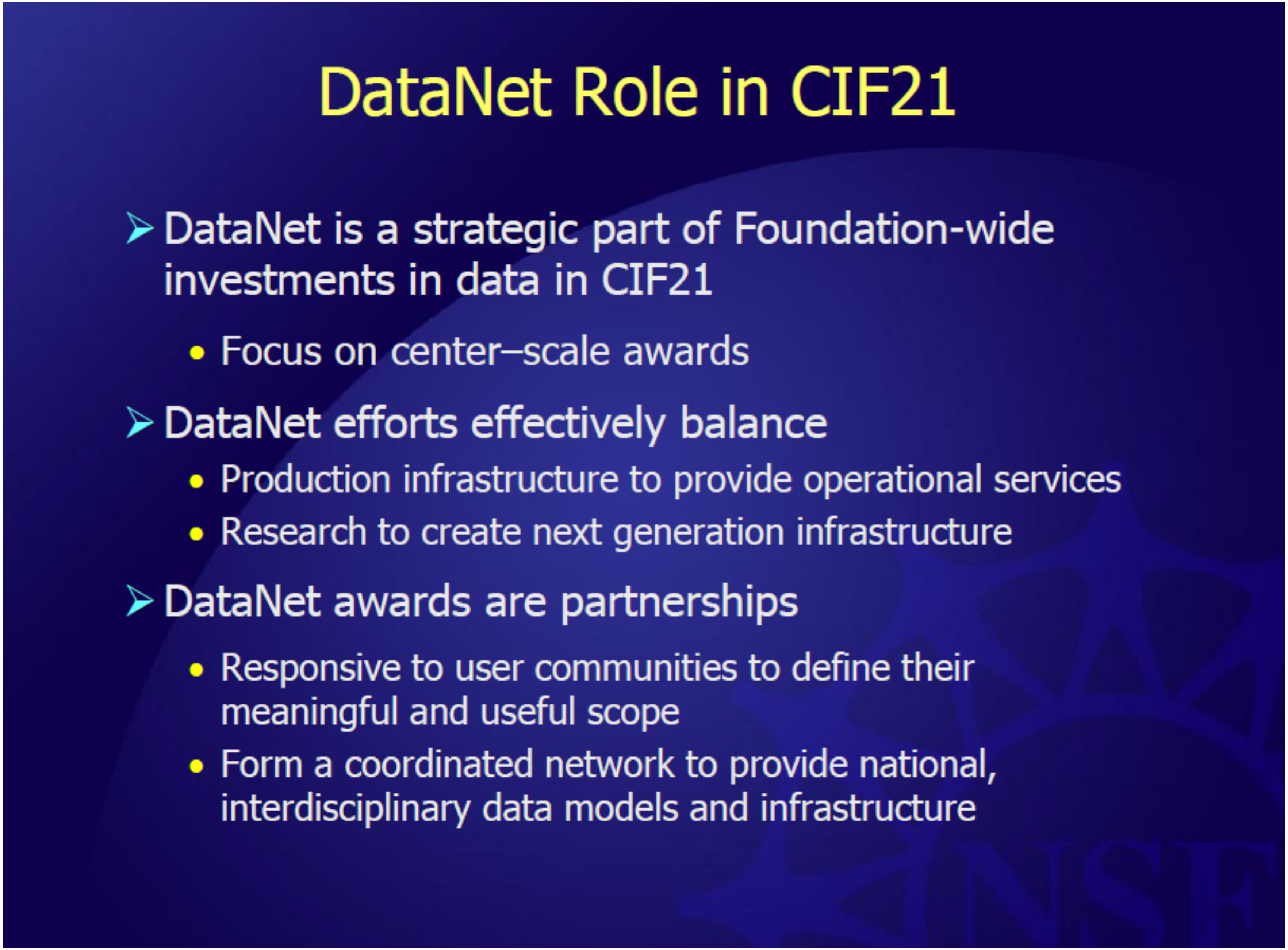
ACCI Data Task Force Recommendations

- Recognize data infrastructure and services as essential research assets fundamental to today's science and as long-term investments in national prosperity
- Create new citation models in which data and software tool providers are credited with their data contributions
- Develop and publish realistic cost models to underpin institutional/national business plans for research repositories/data services
- Identify and share best-practices for the critical areas of data management



CIF21 Data Goals

- Provide reliable digital access, integration, management and preservation capabilities for science and engineering data over a decades-long timeline
- Develop innovative data analysis and mining tools to support data manipulation, modeling, and discovery
- Engage at the frontiers of technological innovation and transformative science to drive the leading edge forward
- Support data intensive and multi-disciplinary science



DataNet Role in CIF21

- DataNet is a strategic part of Foundation-wide investments in data in CIF21
  - Focus on center–scale awards
- DataNet efforts effectively balance
  - Production infrastructure to provide operational services
  - Research to create next generation infrastructure
- DataNet awards are partnerships
  - Responsive to user communities to define their meaningful and useful scope
  - Form a coordinated network to provide national, interdisciplinary data models and infrastructure

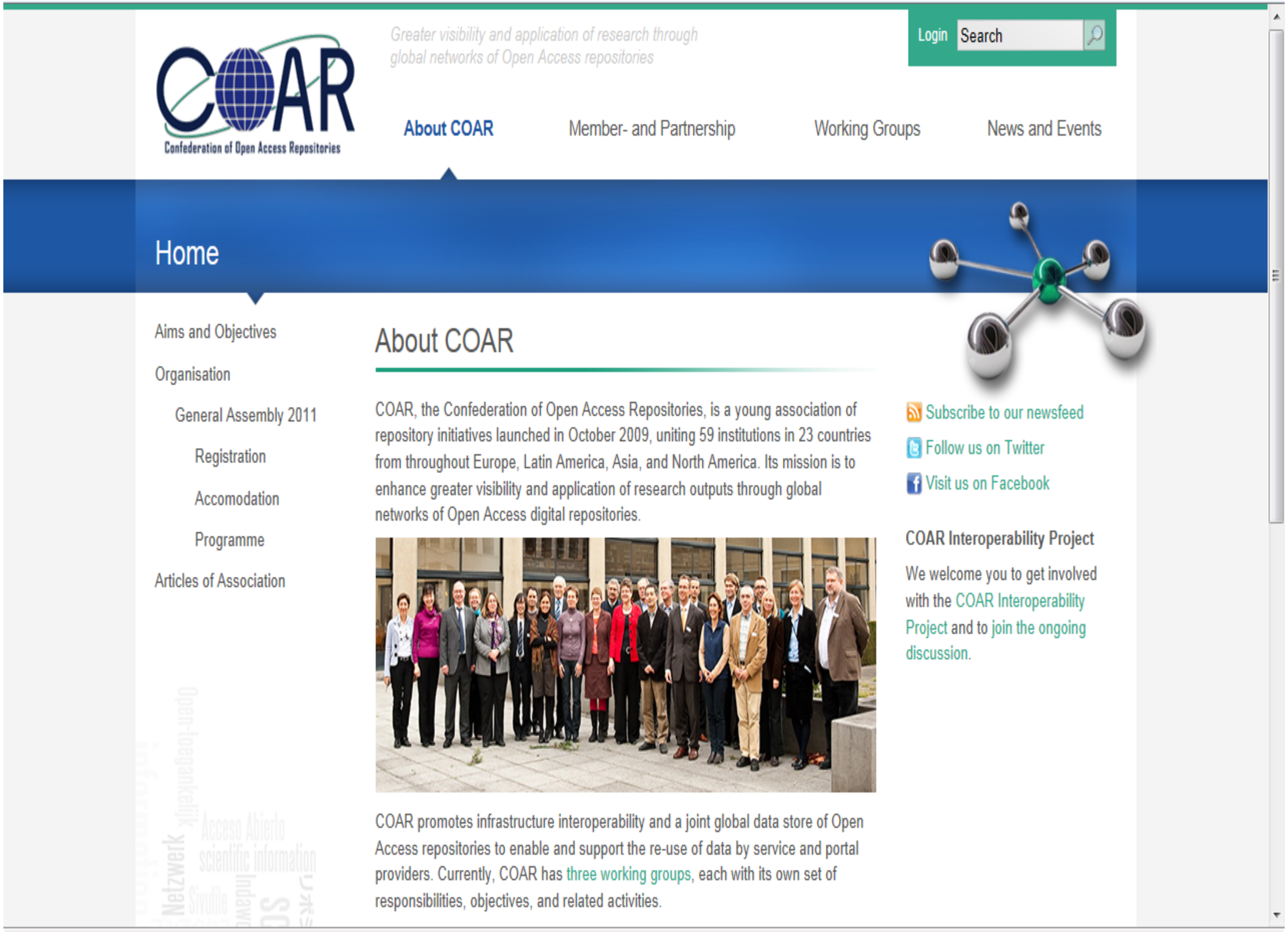
eScience:
The Future

Four Key Developments for the Future of eScience

- Open Access
- The Cloud
- Natural User Interfaces
- Semantic Computing

Open Access and Repositories

- University Research Repositories will contain full text versions of research papers and also ‘grey’ literature such as workshop papers, presentations, technical reports and theses
  - In the future repositories will also contain data, images and software
- With the advent of open access to both full text of papers and data, university research repositories will be an important part of the university’s reputation management strategy
  - COAR – Coalition of Open Access Repositories



Greater visibility and application of research through global networks of Open Access repositories

About COAR

COAR, the Confederation of Open Access Repositories, is a young association of repository initiatives launched in October 2009, uniting 59 institutions in 23 countries from throughout Europe, Latin America, Asia, and North America. Its mission is to enhance greater visibility and application of research outputs through global networks of Open Access digital repositories.

COAR promotes infrastructure interoperability and a joint global data store of Open Access repositories to enable and support the re-use of data by service and portal providers. Currently, COAR has three working groups, each with its own set of responsibilities, objectives, and related activities.

COAR Interoperability Project

We welcome you to get involved with the COAR Interoperability Project and to join the ongoing discussion.

Datacite and ORCID

DataCite

- International consortium to establish easier access to scientific research data
- Increase acceptance of research data as legitimate, citable contributions to the scientific record
- Support data archiving that will permit results to be verified and re-purposed for future study.

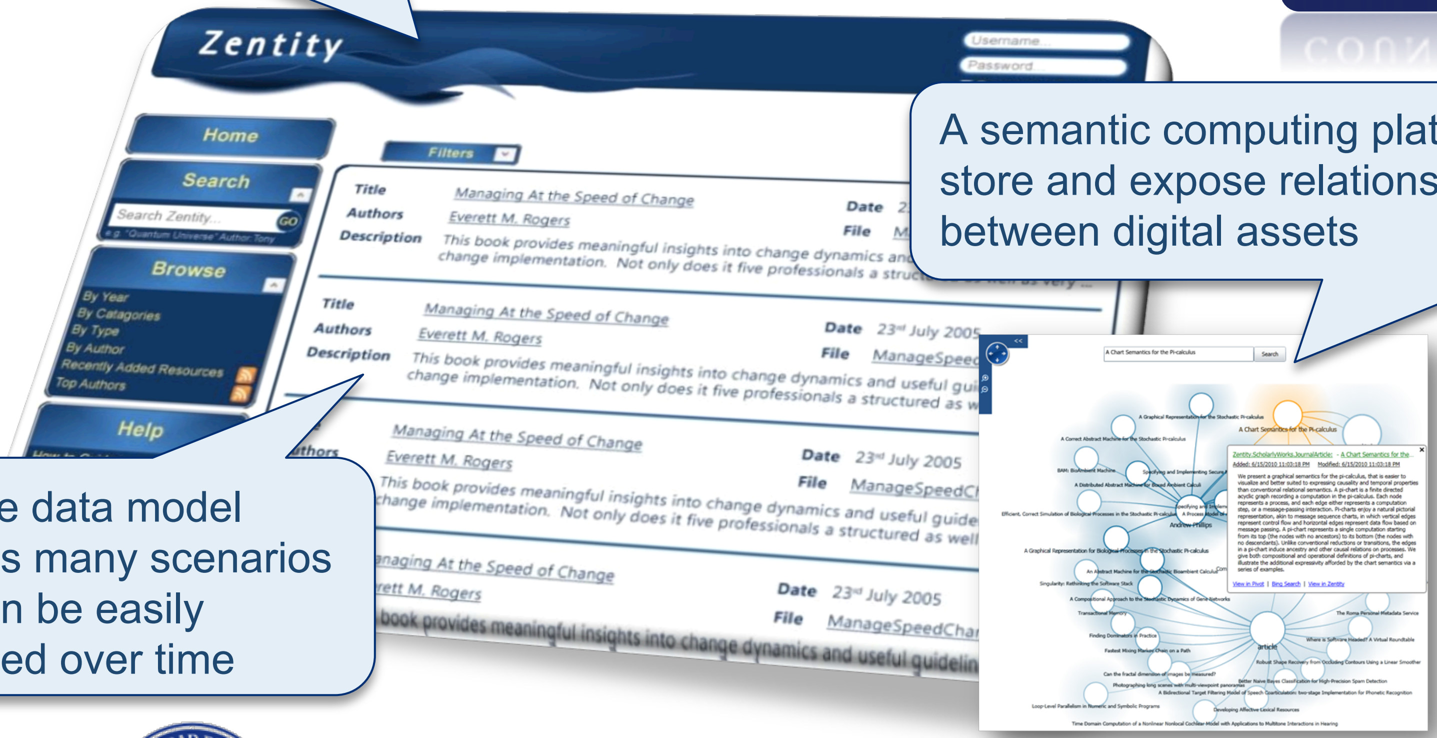
ORCID - Open Research & Contributor ID

- Aims to solve the author/contributor name ambiguity problem in scholarly communications
- Central registry of unique identifiers for individual researchers
- Open and transparent linking mechanism between ORCID and other current author ID schemes.
- Identifiers can be linked to the researcher’s output to enhance the scientific discovery process

Zentity: Semantically-enabled Repository Software

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

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



Title Managing At the Speed of Change
Authors Everett M. Rogers
Description This book provides meaningful insights into change dynamics and useful guidelines for change implementation. Not only does it five professionals a structured as well as very ...
Date 23rd July 2005
File ManageSpeedChange

Zentity.ScholarlyWorks.JournalArticle: - A Chart Semantics for the...

We present a graphical semantics for the pi-calculus, that is easier to visualize and better suited to expressing causality and temporal properties than conventional relational semantics.

View in Pivot | Bing Search | View in Zentity

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

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

UNIVERSIDAD DE BOGOTÁ
JORGE TADEO LOZANO

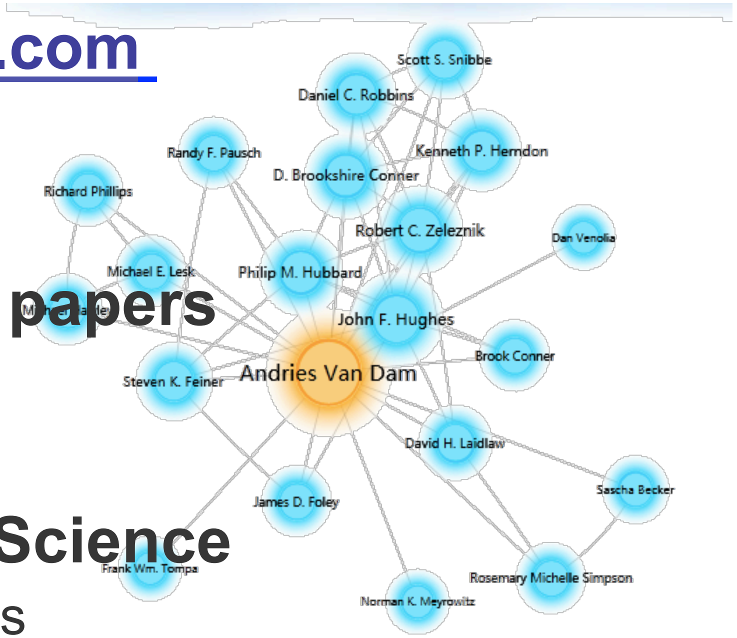
http://research.microsoft.com/zentity/

Microsoft Research Connections

Microsoft Research – Academic Research Beta

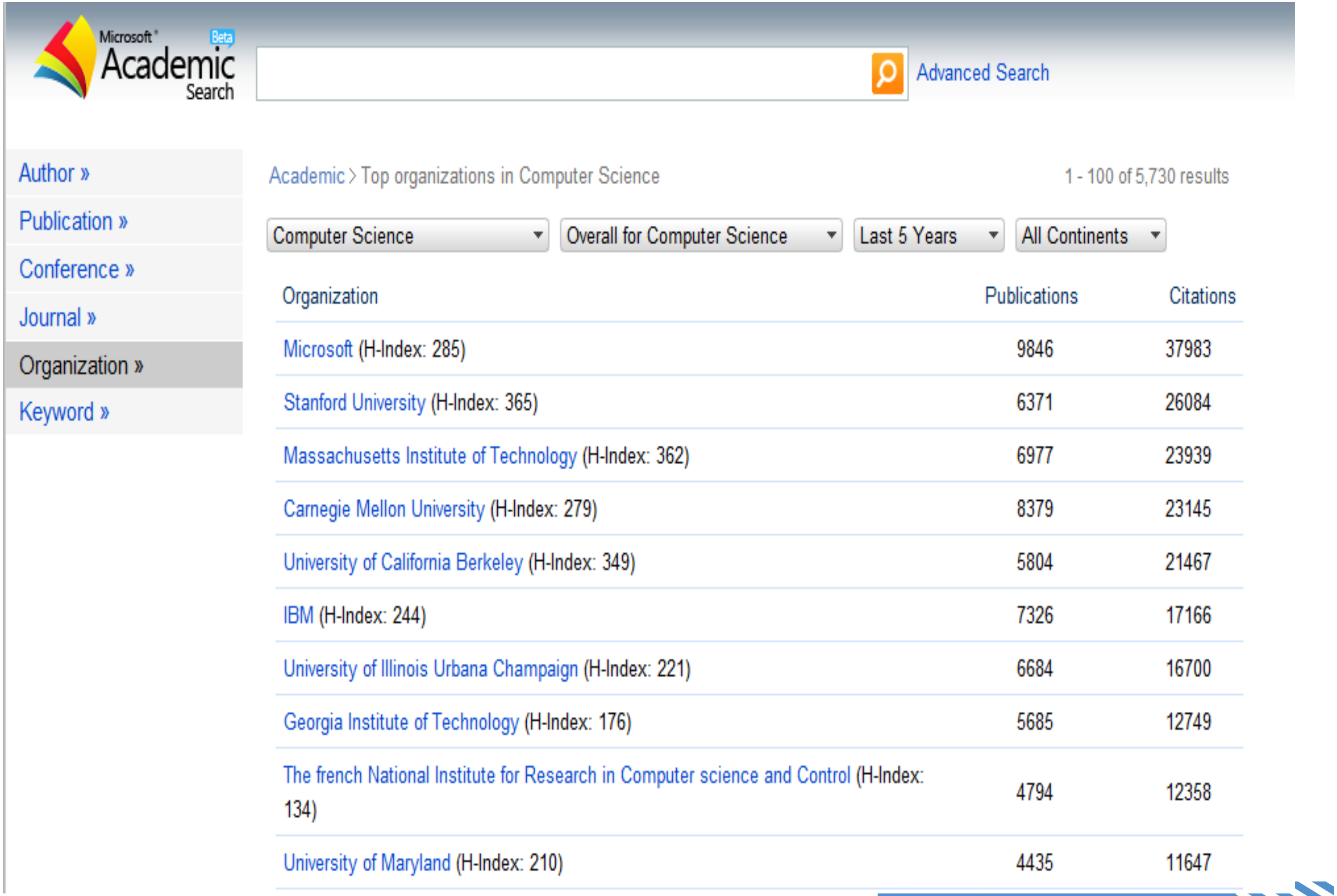
http://academic.research.microsoft.com

Microsoft Academic Search Beta



- **Powerful search tool for academic papers**
- **From our MSR Asia Lab (Beijing)**
  - For researchers, by researchers
- **Historically focused on Computer Science**
  - Now targeting all academic fields/domains
- **Key functionality includes**
  - Find top papers in a domain
  - Easily search the top papers, authors, conferences, and journals for a topic
  - See details about a specific paper, author, conference or journal
  - Quickly find relationships between authors (with Visual Explorer)
  - Get a related Bing Answer
- **Currently 37M papers across 20+ domains**
  - With 100M papers in the queue …

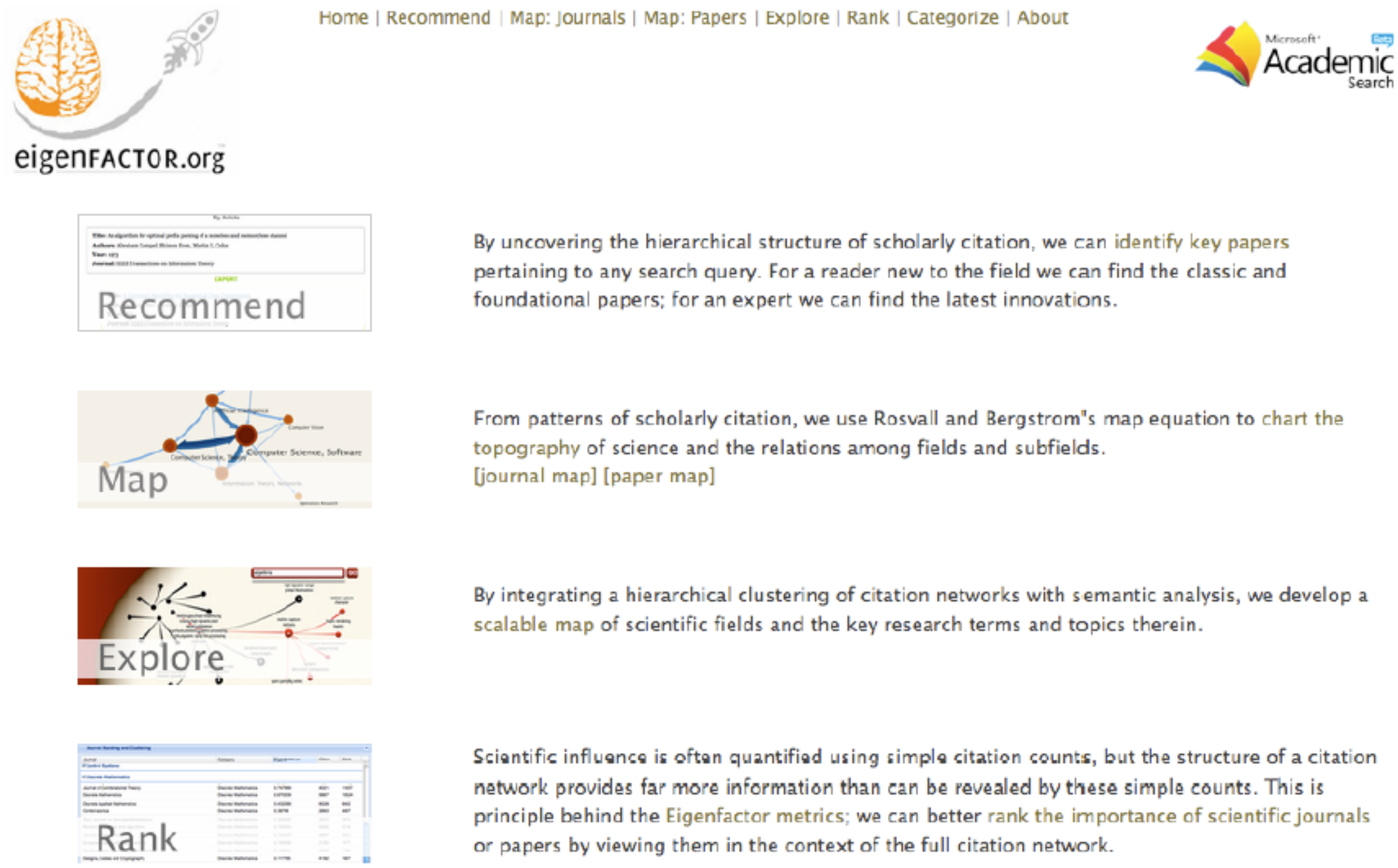
Top 10 Computer Science Organizations



Academic > Top organizations in Computer Science — 1 - 100 of 5,730 results

Computer Science ▾ | Overall for Computer Science ▾ | Last 5 Years ▾ | All Continents ▾

Organization	Publications	Citations
Microsoft (H-Index: 285)	9846	37983
Stanford University (H-Index: 365)	6371	26084
Massachusetts Institute of Technology (H-Index: 362)	6977	23939
Carnegie Mellon University (H-Index: 279)	8379	23145
University of California Berkeley (H-Index: 349)	5804	21467
IBM (H-Index: 244)	7326	17166
University of Illinois Urbana Champaign (H-Index: 221)	6684	16700
Georgia Institute of Technology (H-Index: 176)	5685	12749
The french National Institute for Research in Computer science and Control (H-Index: 134)	4794	12358
University of Maryland (H-Index: 210)	4435	11647


eigenFACTOR.org
Home | Recommend | Map: Journals | Map: Papers | Explore | Rank | Categorize | About
Microsoft Academic Search
Recommend
Map
Explore
Rank

By uncovering the hierarchical structure of scholarly citation, we can identify key papers pertaining to any search query. For a reader new to the field we can find the classic and foundational papers; for an expert we can find the latest innovations.

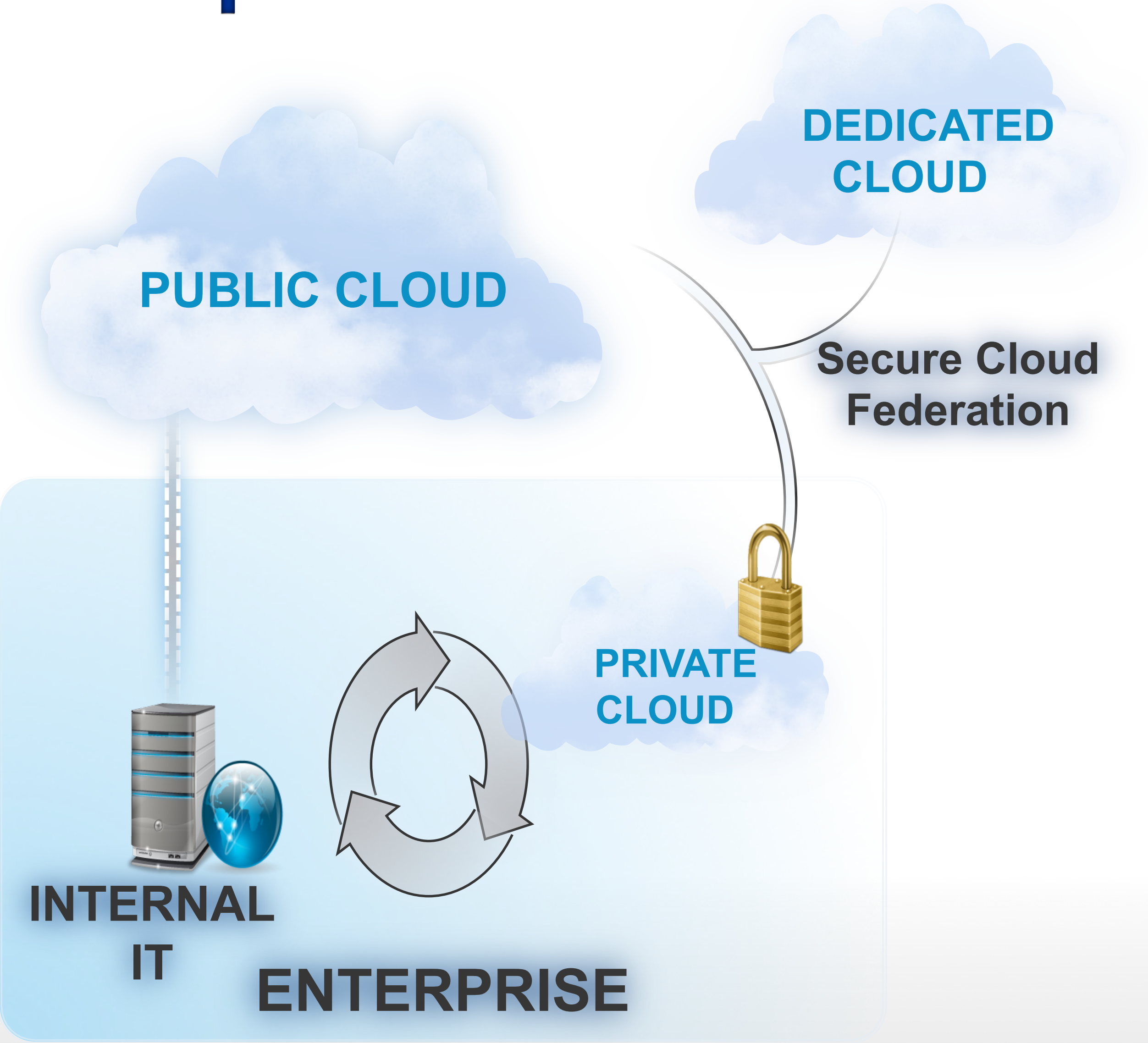
From patterns of scholarly citation, we use Rosvall and Bergstrom's map equation to chart the topography of science and the relations among fields and subfields.
[journal map] [paper map]

By integrating a hierarchical clustering of citation networks with semantic analysis, we develop a scalable map of scientific fields and the key research terms and topics therein.

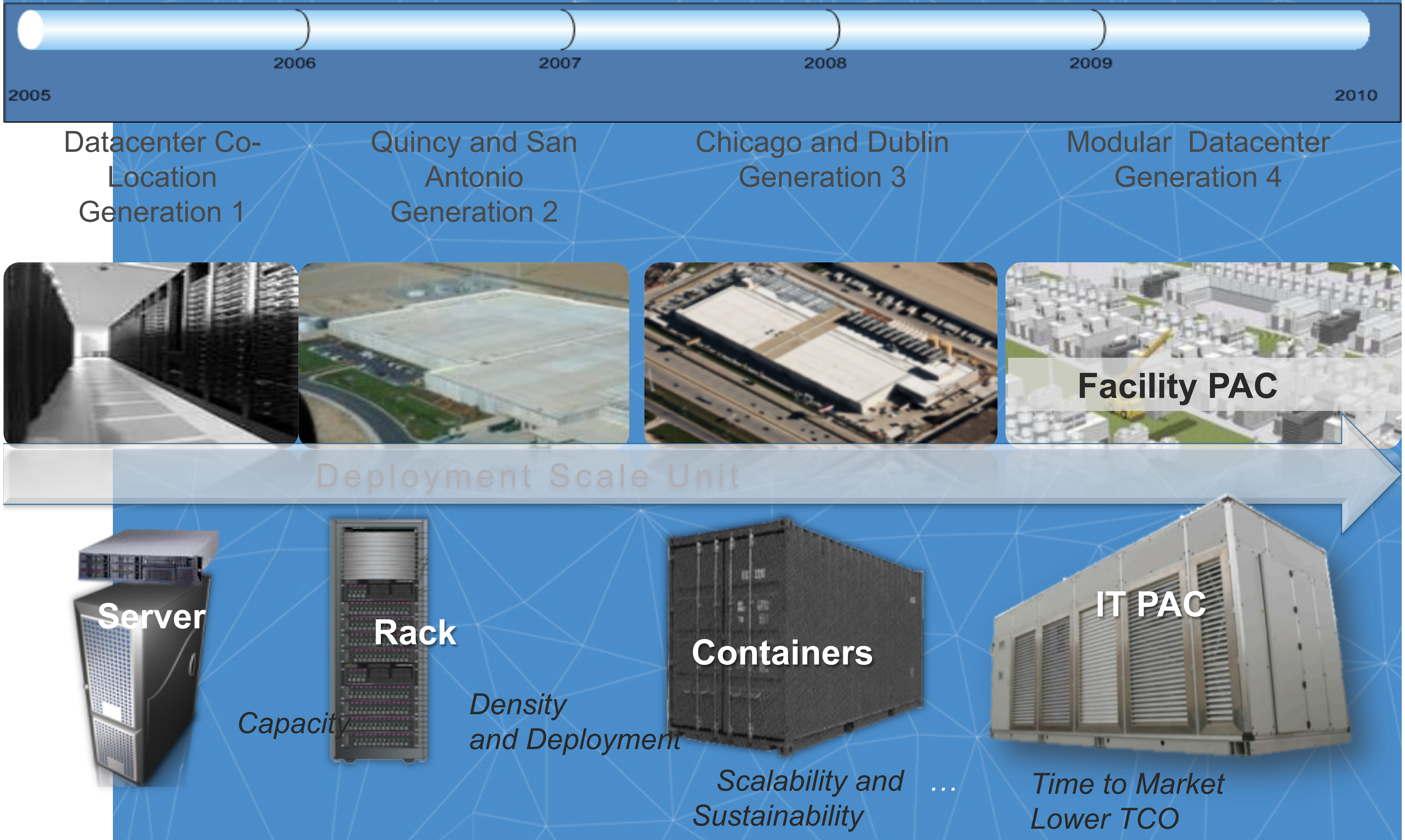
Scientific influence is often quantified using simple citation counts, but the structure of a citation network provides far more information than can be revealed by these simple counts. This is principle behind the Eigenfactor metrics; we can better rank the importance of scientific journals or papers by viewing them in the context of the full citation network.

http://www.eigenfactor.org/

The Cloud - Options



Microsoft's Datacenter Evolution



Datacenter Co-Location Generation 1

Quincy and San Antonio Generation 2

Chicago and Dublin Generation 3

Modular Datacenter Generation 4

Facility PAC

Deployment Scale Unit

Server

Capacity

Rack

Density and Deployment

Containers

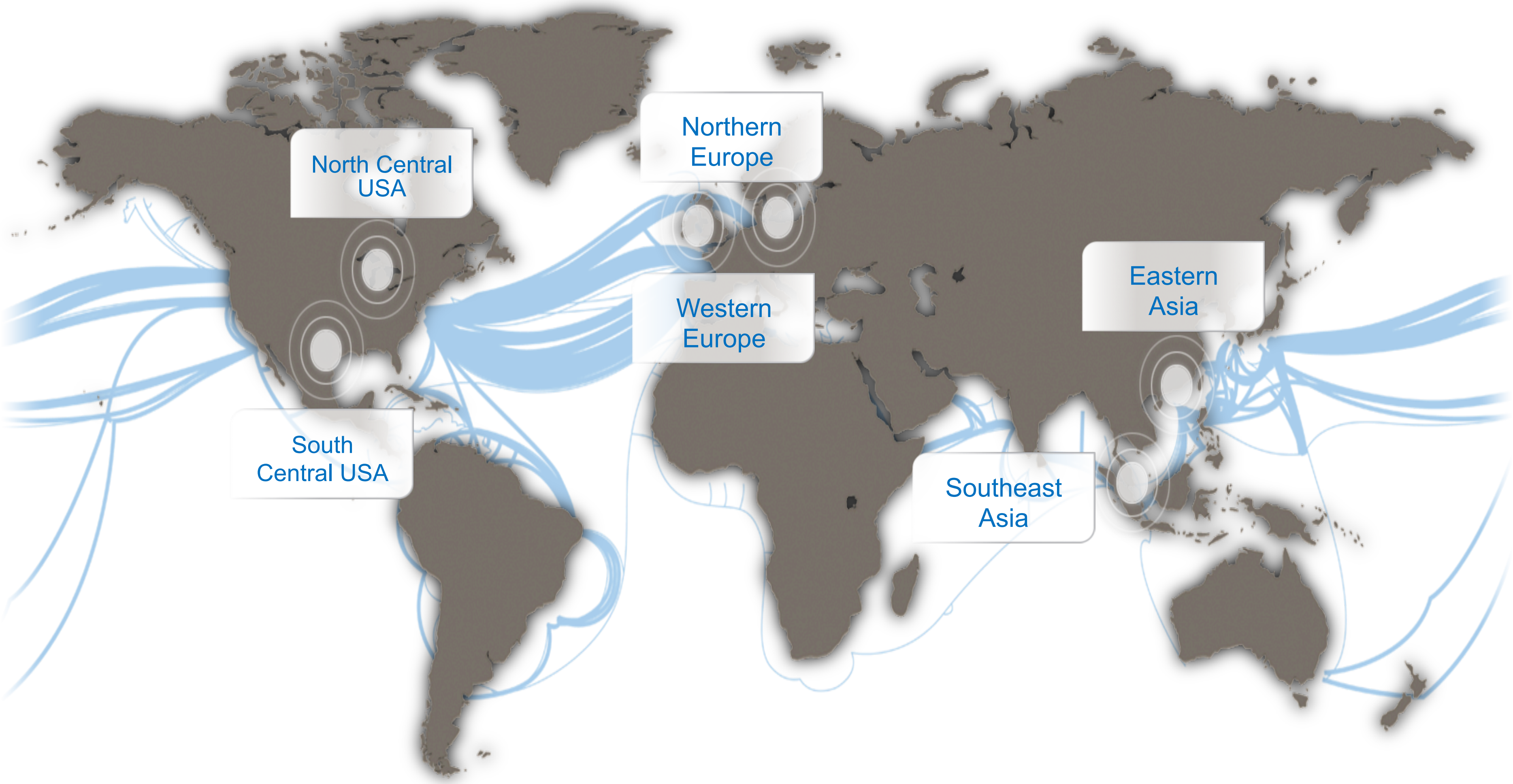
Scalability and Sustainability

…

IT PAC

Time to Market
Lower TCO

Windows Azure Platform Availability

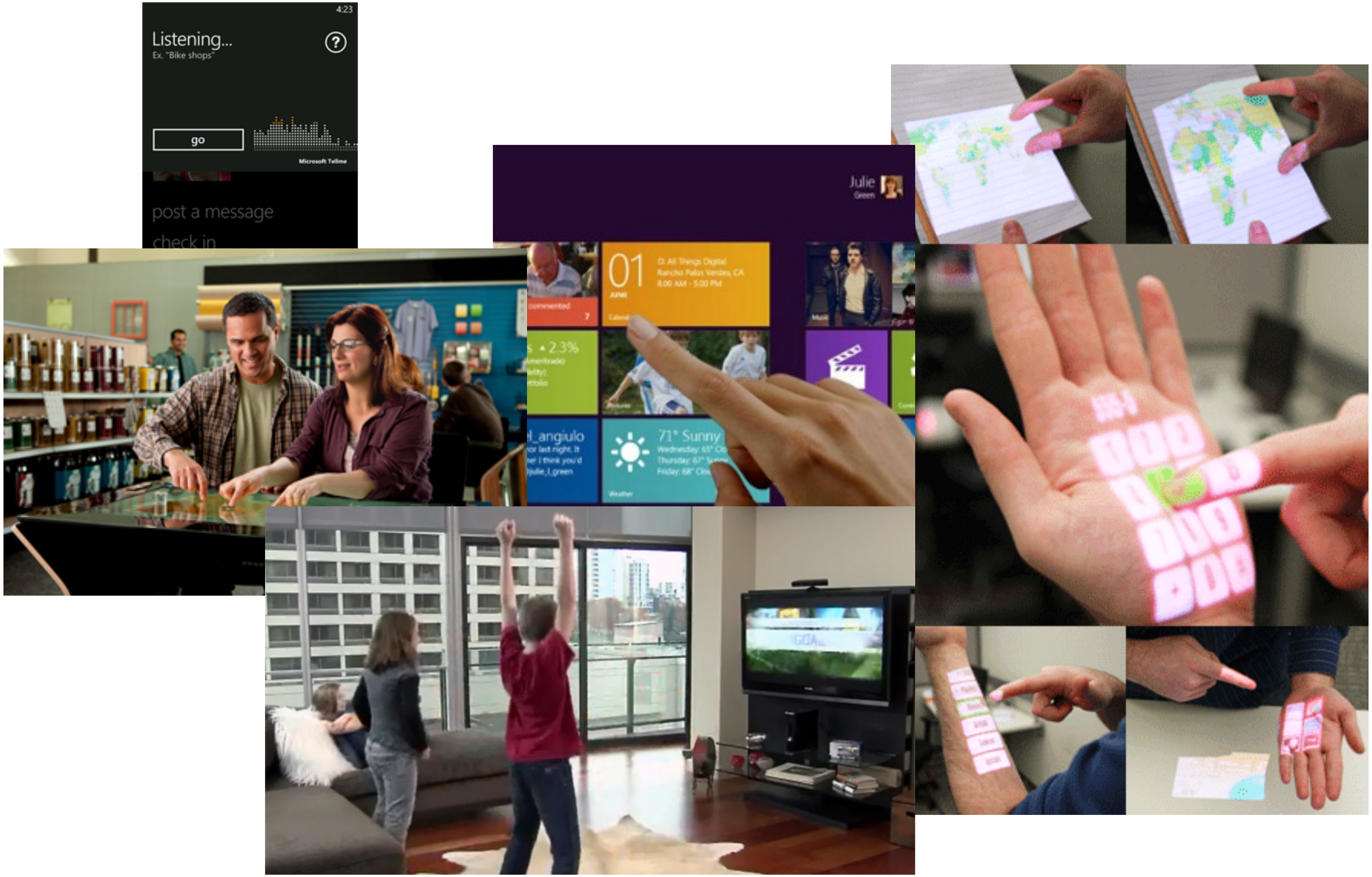


EU FP7 Project: VENUS-C

Venus-C

Virtual multidisciplinary EnviroNments USing Cloud infrastructures

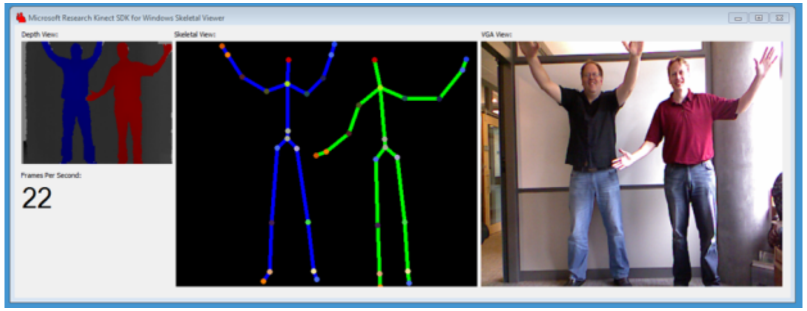
Natural User Interfaces or NUIs



Microsoft® Research Kinect™ for Windows® SDK beta

KINECT™ for XBOX 360

- Rethinking ways in which people will interact with computers/technologies of the future
- Re-evaluating everything from their (non-) physical design to the human needs and interaction models
- Revolutionize the way we think about technology and what it can do on our behalf



Frames Per Second:
22

http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/

Semantic Computing

computers are
great **tools** for

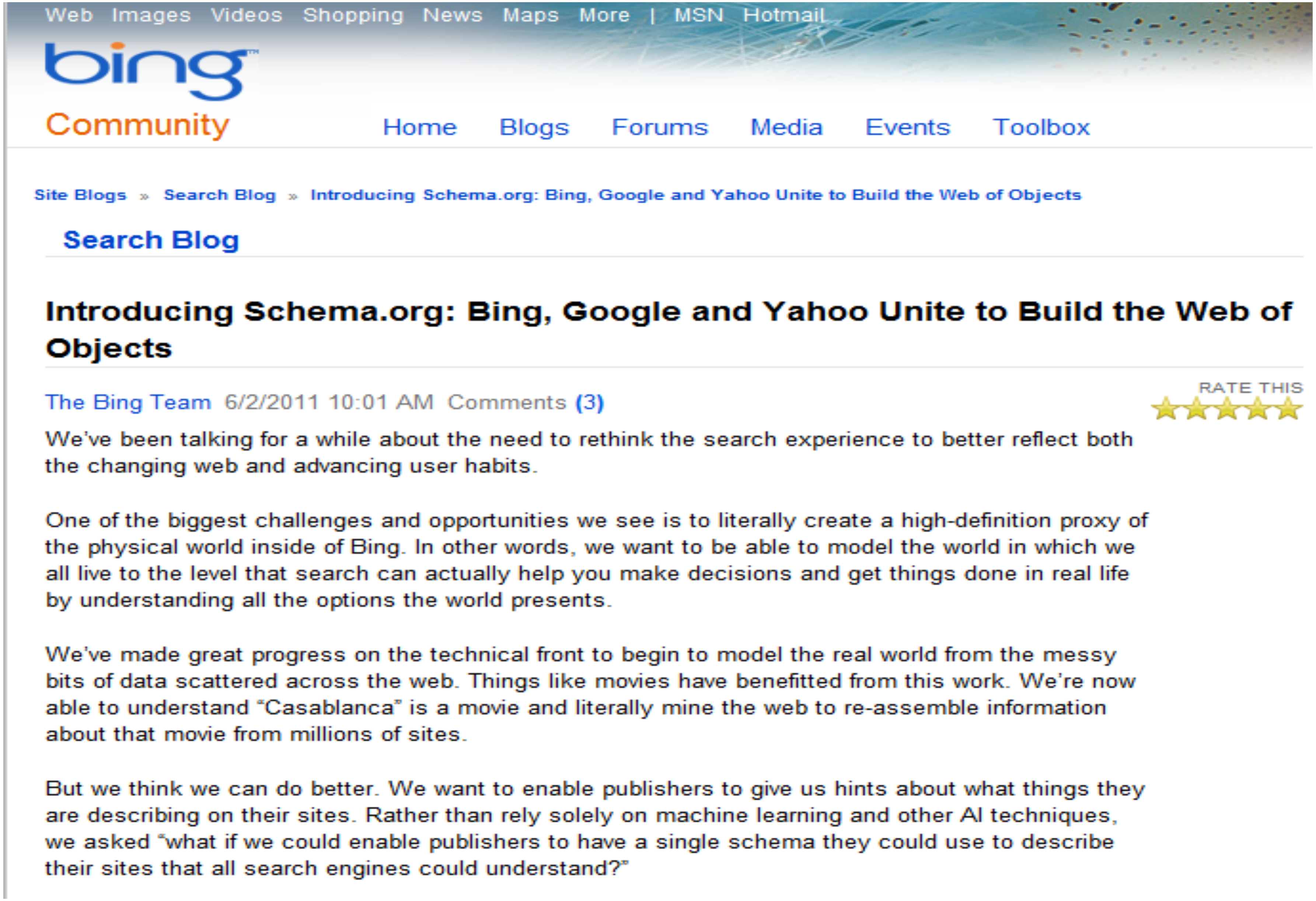
storing
computing
managing
indexing

huge amounts
of **data**

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

acquisition
discovery
aggregation
organization
correlation
analysis
interpretation
inference

of the world's **information**
and **knowledge**


Web Images Videos Shopping News Maps More | MSN Hotmail
bing
Community
Home Blogs Forums Media Events Toolbox
Site Blogs » Search Blog » Introducing Schema.org: Bing, Google and Yahoo Unite to Build the Web of Objects

Search Blog

Introducing Schema.org: Bing, Google and Yahoo Unite to Build the Web of Objects

The Bing Team 6/2/2011 10:01 AM Comments (3)

We've been talking for a while about the need to rethink the search experience to better reflect both the changing web and advancing user habits.

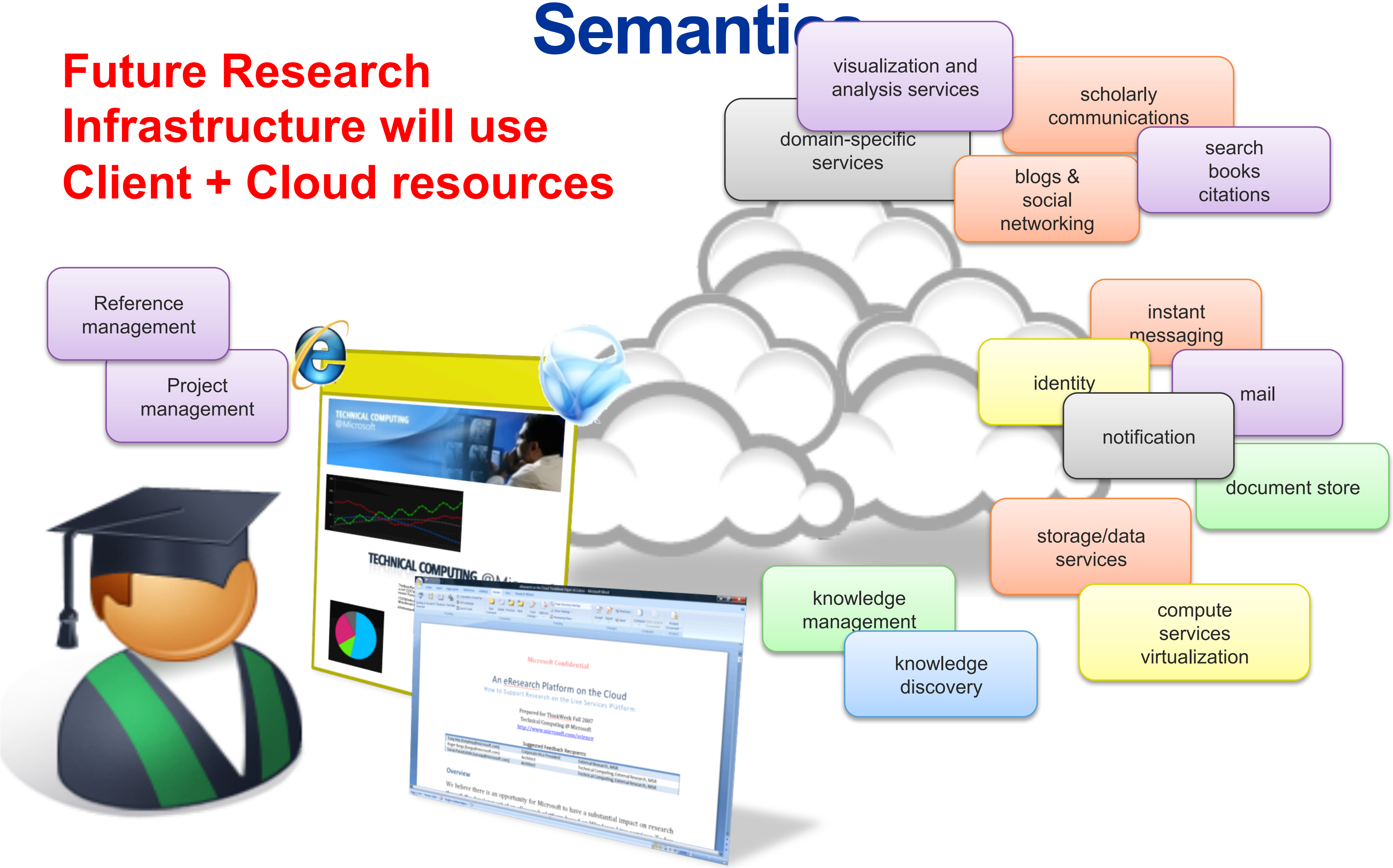
One of the biggest challenges and opportunities we see is to literally create a high-definition proxy of the physical world inside of Bing. In other words, we want to be able to model the world in which we all live to the level that search can actually help you make decisions and get things done in real life by understanding all the options the world presents.

We've made great progress on the technical front to begin to model the real world from the messy bits of data scattered across the web. Things like movies have benefitted from this work. We're now able to understand "Casablanca" is a movie and literally mine the web to re-assemble information about that movie from millions of sites.

But we think we can do better. We want to enable publishers to give us hints about what things they are describing on their sites. Rather than rely solely on machine learning and other AI techniques, we asked "what if we could enable publishers to have a single schema they could use to describe their sites that all search engines could understand?"

Forecast: Cloudy with a Chance of Semantics

Future Research Infrastructure will use Client + Cloud resources



visualization and analysis services
scholarly communications
domain-specific services
blogs & social networking
search books citations
Reference management
Project management
instant messaging
identity
mail
notification
document store
storage/data services
knowledge management
knowledge discovery
compute services virtualization
TECHNICAL COMPUTING @Microsoft
TECHNICAL COMPUTING @Microsoft
Microsoft Confidential
An eResearch Platform on the Cloud
How to Support Research on the Live Services Platform
Prepared for ThinkWeek Fall 2007
Technical Computing @ Microsoft
http://www.microsoft.com/science
Suggested Feedback Recipients
Overview
We believe there is an opportunity for Microsoft to have a substantial impact on research

Some Resources

- Microsoft Research
  - http://research.microsoft.com
  - Microsoft Research downloads: http://research.microsoft.com/research/downloads
- Microsoft Research Connections
  - http://research.microsoft.com/en-us/collaboration/
- Science at Microsoft
  - http://www.microsoft.com/science
- Scholarly Communications
  - http://www.microsoft.com/scholarlycomm
- CodePlex
  - http://www.codeplex.com
- Outercurve Foundation
  - http://www.outercurve.org/
- Tony Hey on eScience
  - http://tonyhey.net/

Microsoft®

Your potential. Our passion.™