eResearch in the Cloud: Data Intensive High Performance Computing

Tony Hey

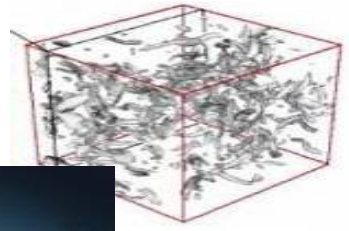
Corporate Vice President for External Research
Microsoft Research

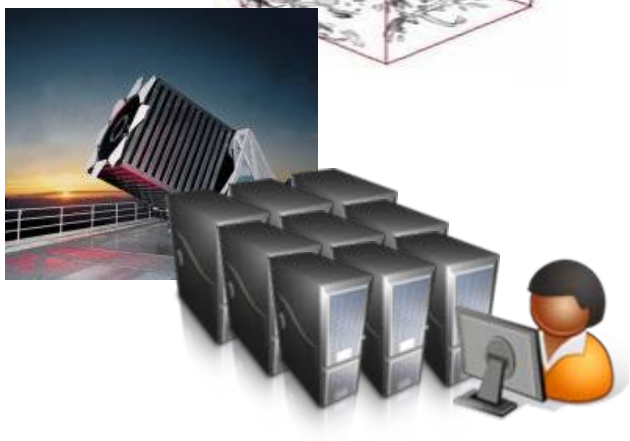
Emergence of a New Research Paradigm?

- Thousand years ago – **Experimental Science**
  - Description of natural phenomena
- Last few hundred years – **Theoretical Science**
  - Newton's Laws, Maxwell's Equations…
- Last few decades – **Computational Science**
  - Simulation of complex phenomena
- Today – **eScience or Data-centric Science**
  - Unify theory, experiment, and simulation
  - Using data exploration and data mining
    - Data captured by instruments
    - Data generated by simulations
    - Data generated by sensor networks
  - ➢ Scientists over-whelmed with data
  - ➢ Computer Science and IT companies have technologies that will help

(With thanks to Jim Gray)

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



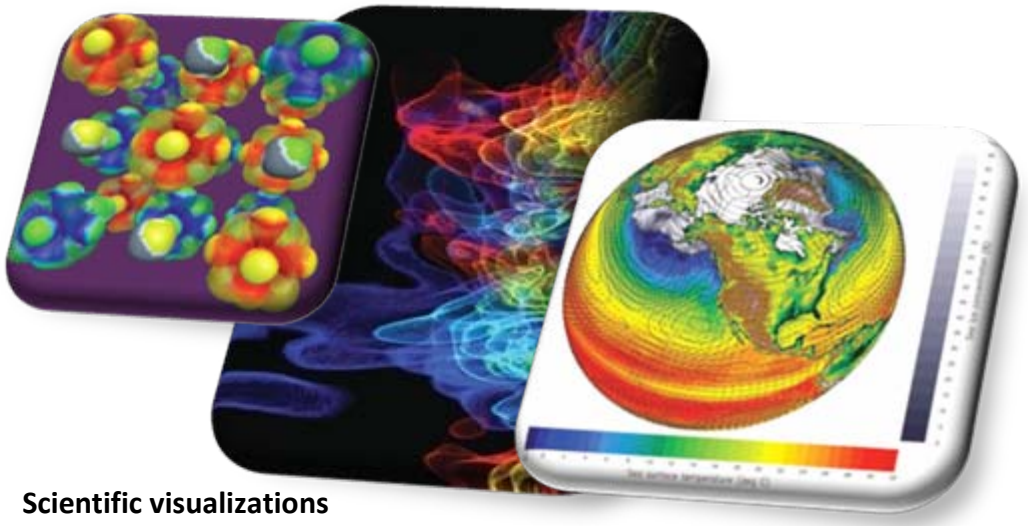


A Scientific Data Deluge

- Data collection
  - Sensor networks, global databases, local databases, desktop computer, laboratory instruments, observation devices, …
- Data processing, analysis, visualization
  - Legacy codes, workflows, data mining, indexing, searching, graphics, screens, …
- Archiving
  - Digital repositories, libraries, preservation, …

SensorMap

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



Scientific visualizations

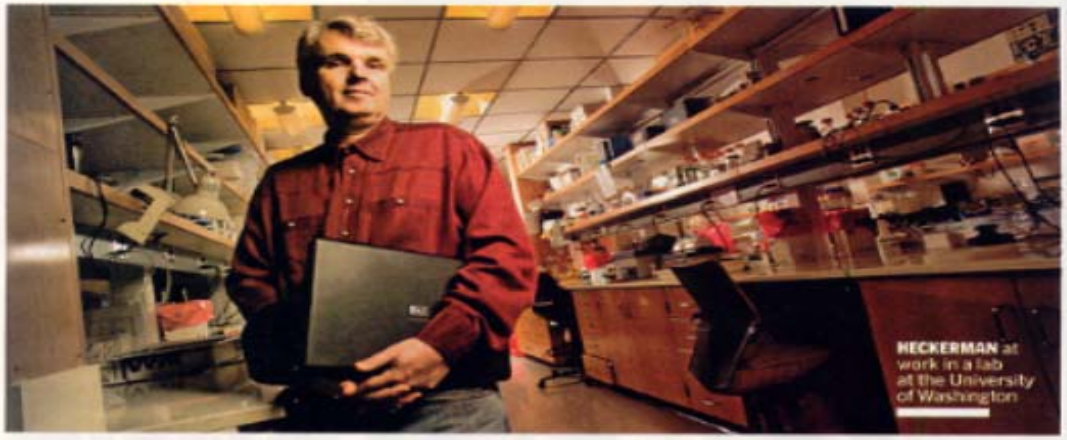
NSF Cyberinfrastructure report, March 2007

Technical Computing @ Microsoft

- Advanced Computing for Science and Engineering
  - Application of new research algorithms, tools and technologies to science and engineering
- High Productivity Computing
  - Application of today's software technologies – HPC, SQLServer, Office, Sharepoint … - to scientific applications
- Radical Computing
  - Research in potential breakthrough technologies

Fighting HIV with Computer Science

- A major problem: Over 40 million infected
  - Drug treatments are effective but are an expensive life commitment
- Vaccine needed for third world countries
  - Effective vaccine could eradicate disease
- Methods from computer science are helping with the design of vaccine
  - Machine learning: Finding biological patterns that may stimulate the immune system to fight the HIV virus
  - Optimization methods: Compressing these patterns into a small, effective vaccine



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

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

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

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

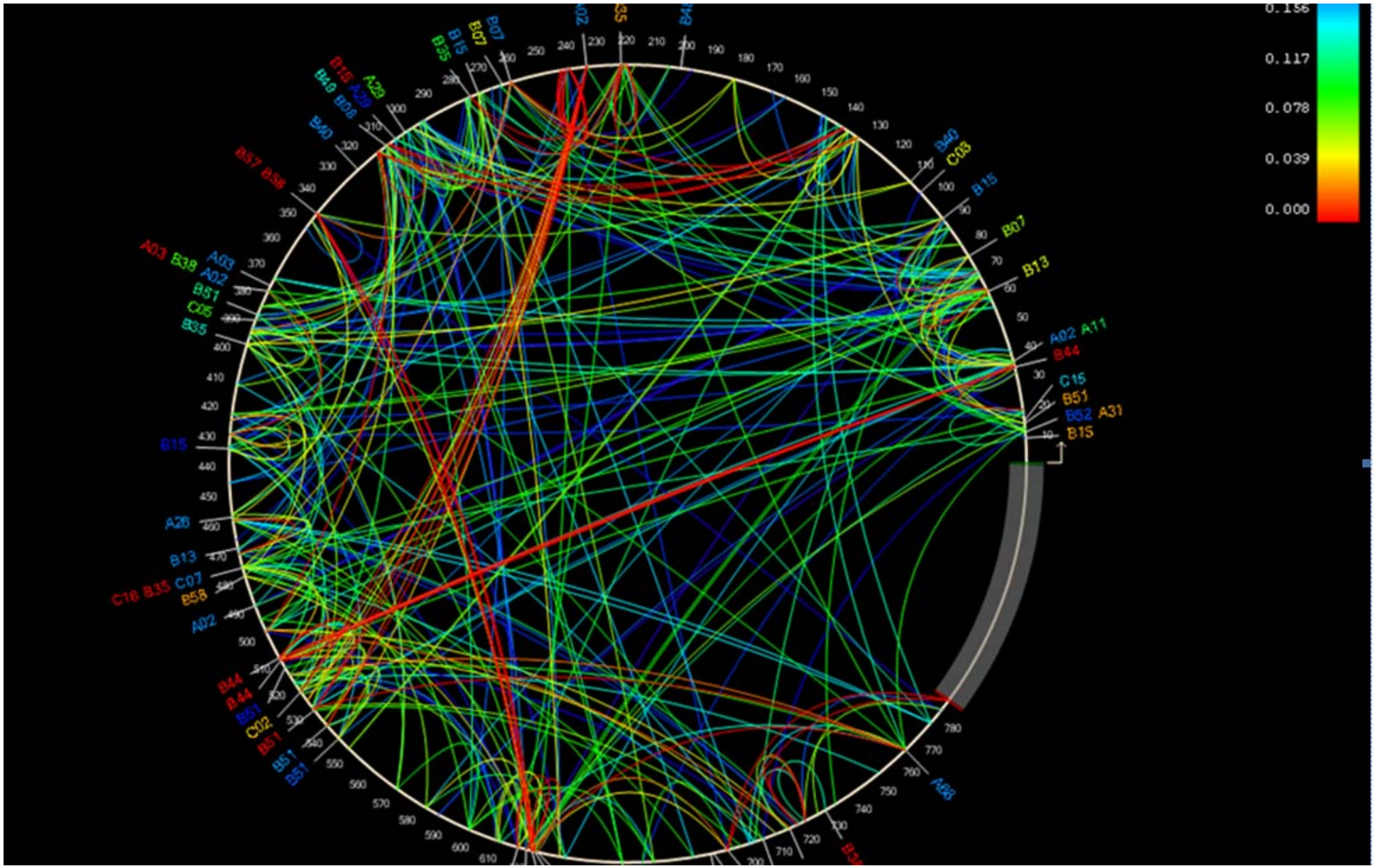
Heckerman was plunging into medicine—and carrying Microsoft with him. When he brought his plan to Bill Gates, the company chairman "got really excited," Heckerman says. Well versed on HIV from his philanthropy work, Gates lined up Heckerman with AIDS researchers at Massachusetts General Hospital, the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

Similar mutations may crop up in computer and medical viruses

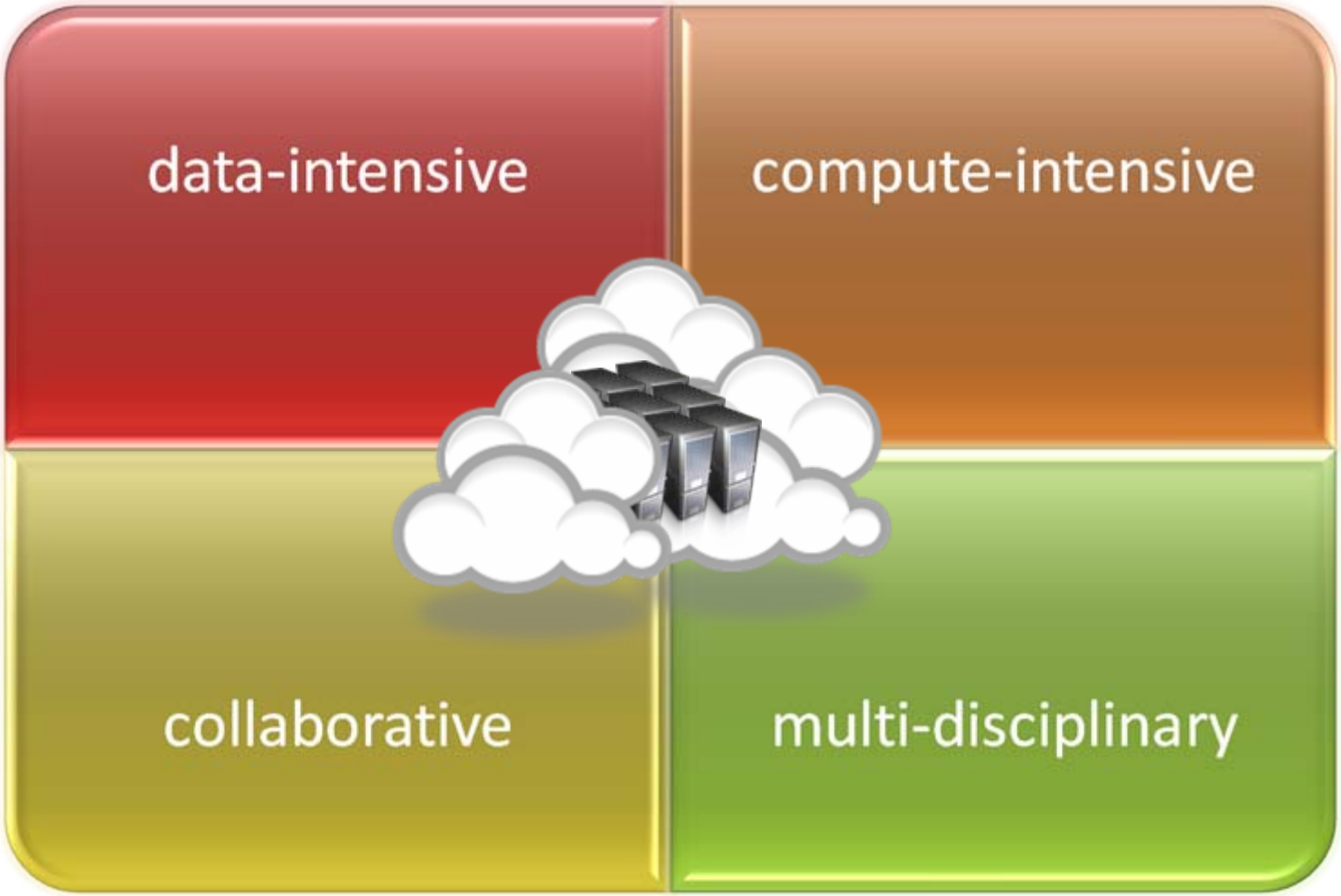
Since then, the 50-year-old Heckerman and two colleagues have created their own biology niche at Microsoft, where they build HIV-detecting software. These are research tools to spot infected cells and correlate the viral mutations with the individual's genetic profile. Heckerman's team runs mountains of data through enormous clusters of 320 computers, operating in parallel. Thanks to smarter algorithms and more powerful machines, they're sifting through the data 480 times faster than a year ago. In June, the team released its first batch of tools for free on the Internet.

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

HIV Protein PoI Evolving in Human Host



Research today is…



Today …

Web users…

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

Researchers…

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

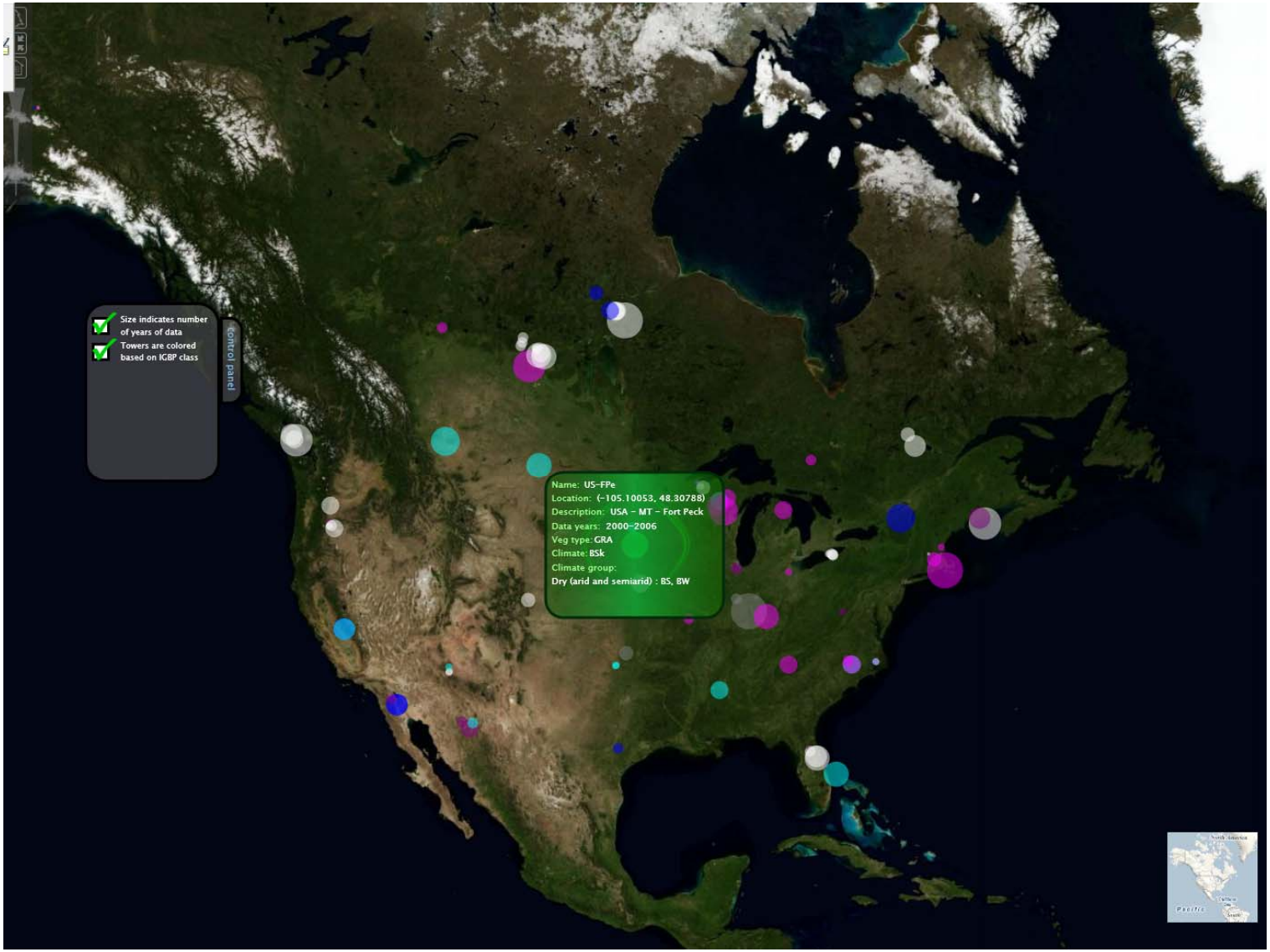
eResearch: Data is easily accessible

With thanks to
Catharine van Ingen

Scientific Data Servers for Hydrology

- Working with Berkeley Water Center using modern Database technologies
  - 149 Ameriflux sites across the Americas reporting minimum of 22 common measurements
  - Carbon-Climate Data published to and archived at Oak Ridge
  - Total data reported to date on the order of 192M half-hourly measurements since 1994
- http://public.ornl.gov/ameriflux/

Mashup of Ameriflux Sites



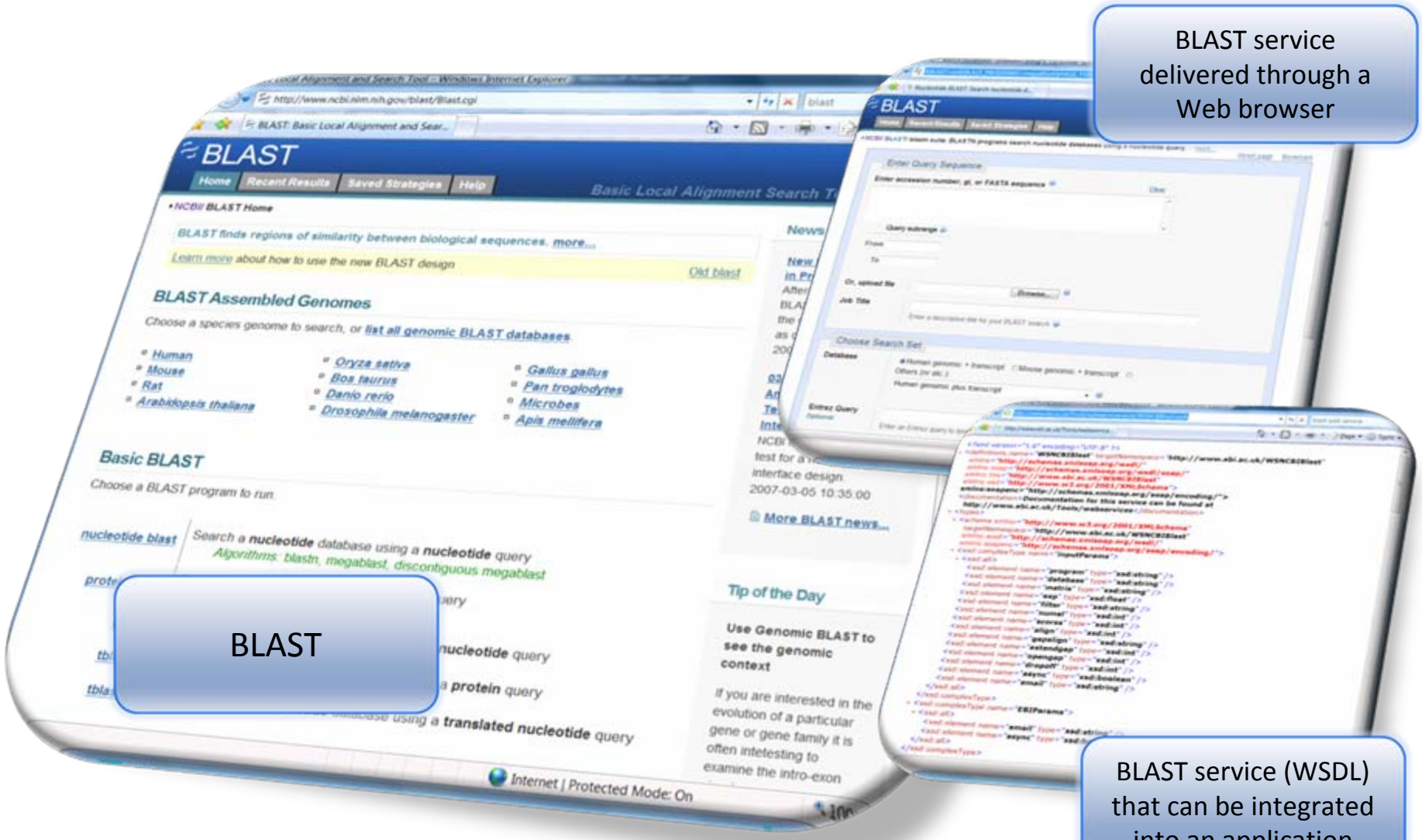
Ameriflux: An eResearch Mashup

Deb Agarwal and Savas Parastatidis
Easy mashups using PopFly and Silverlight
SilverLight Technologies

eResearch: Data is easily shareable

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

eResearch: Services expose functionality



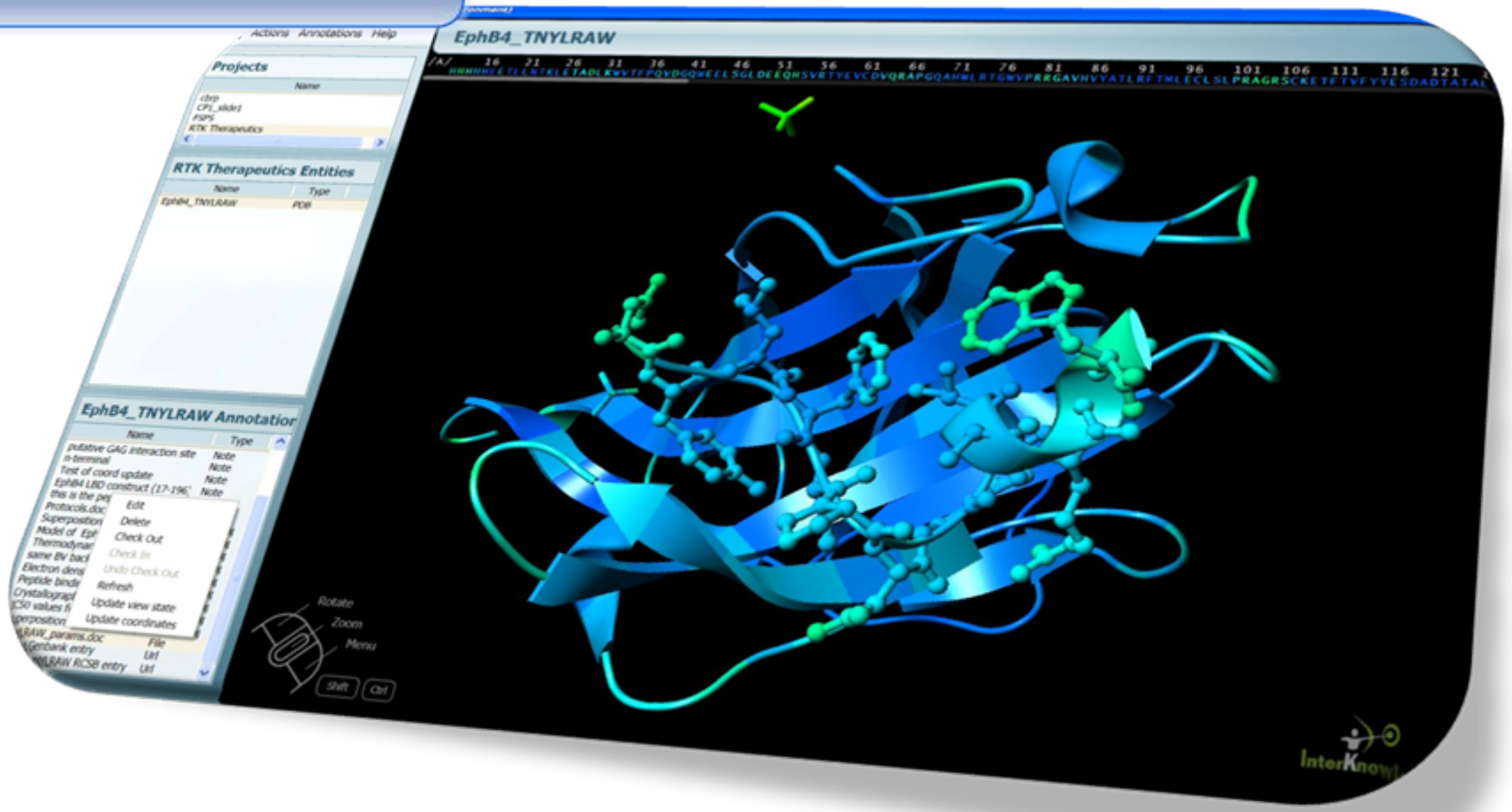
eResearch: Services can be composed

Taverna Workflow

eResearch: Data and Services can be accessed securely

eResearch: New software tools can make a difference

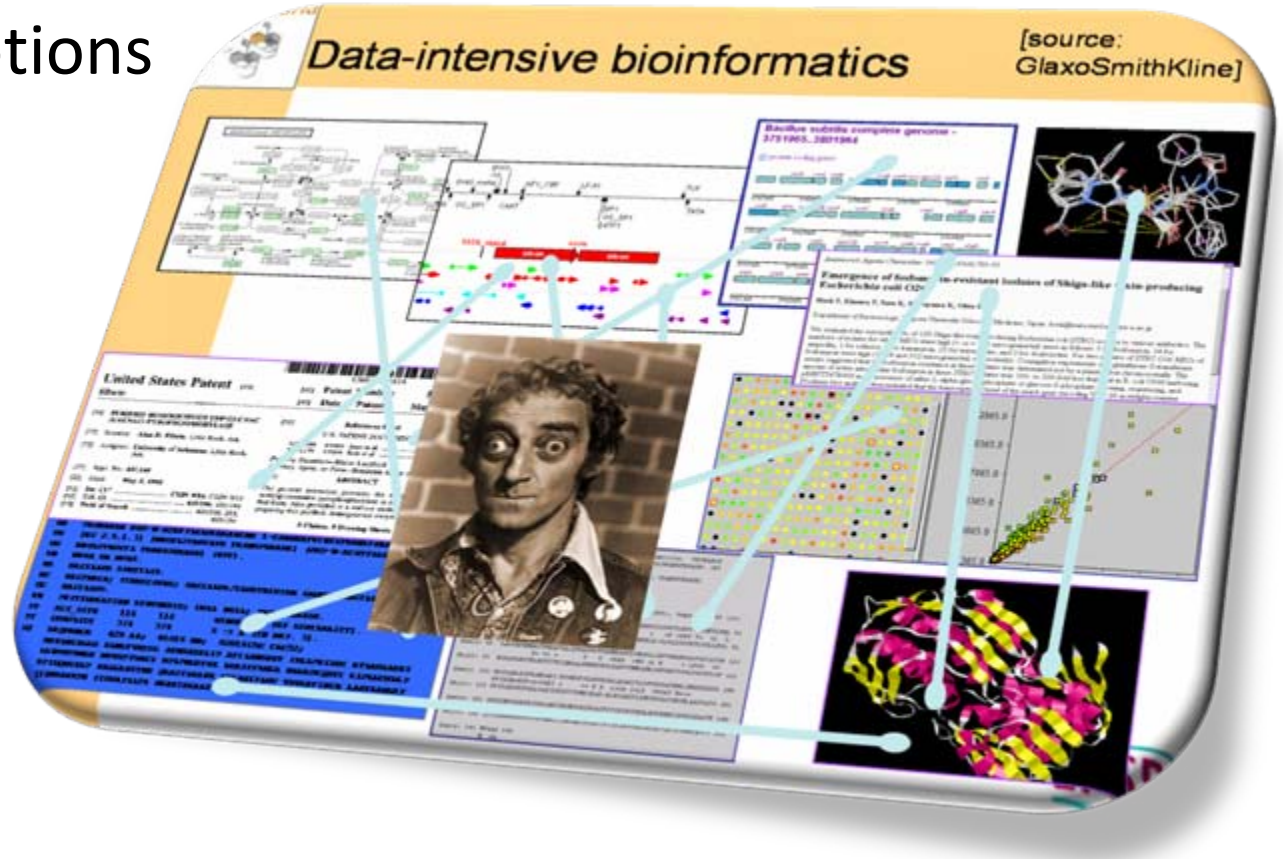
SCRIPPS
(see www.microsoft.com/science)



eResearch: Knowledge can be created/published/archived/discovered

- Semantic relationships between different data
- Semantic descriptions of services
- Annotations
- Provenance
- Repositories
- Ontologies
- Folksonomies

myGrid



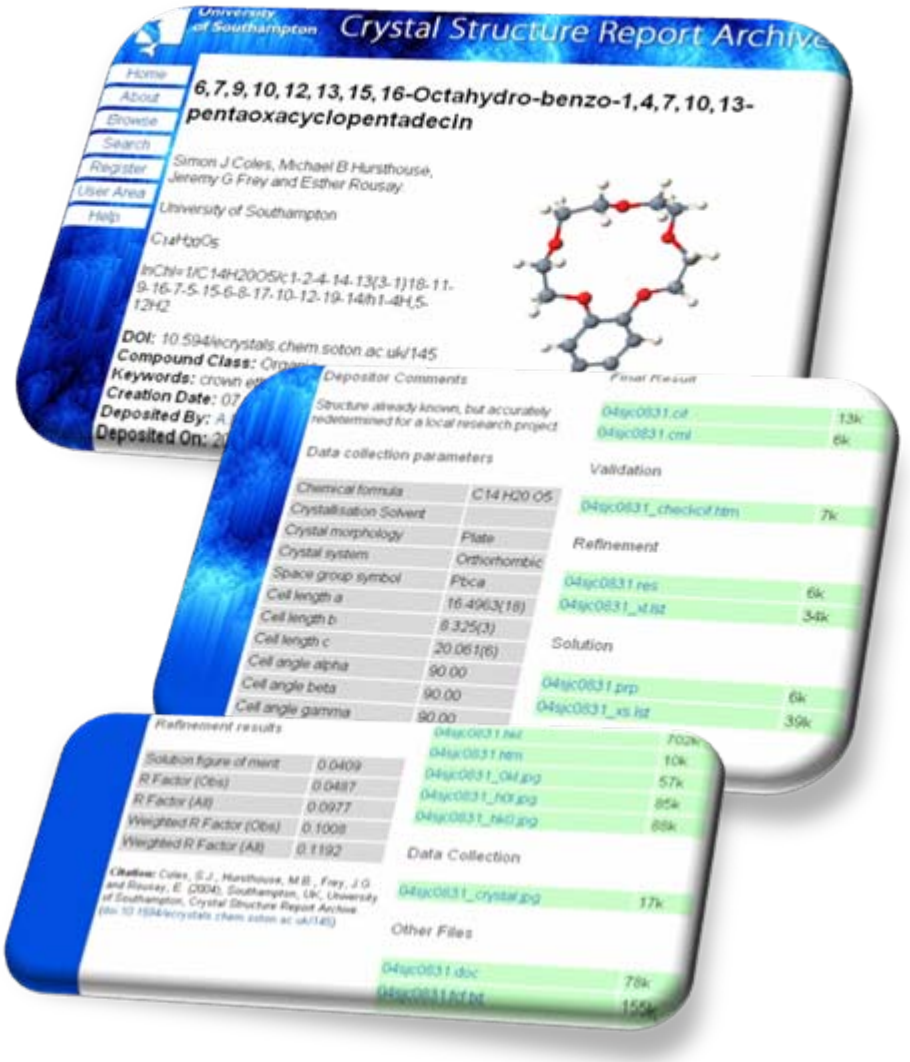
eResearch: Services in the Cloud?

Services not middleware

- No need to install many thousands of lines of middleware
- Take the Web2.0 example and empower researchers through their browser

Services in the cloud

- Blogs, Wikis
- RSS, Tagging
- Data processing, data mining
- Content upload, sharing, discovery
- Storage, computation, messaging



http://ecrystals.chem.soton.ac.uk

Thanks to Jeremy Frey

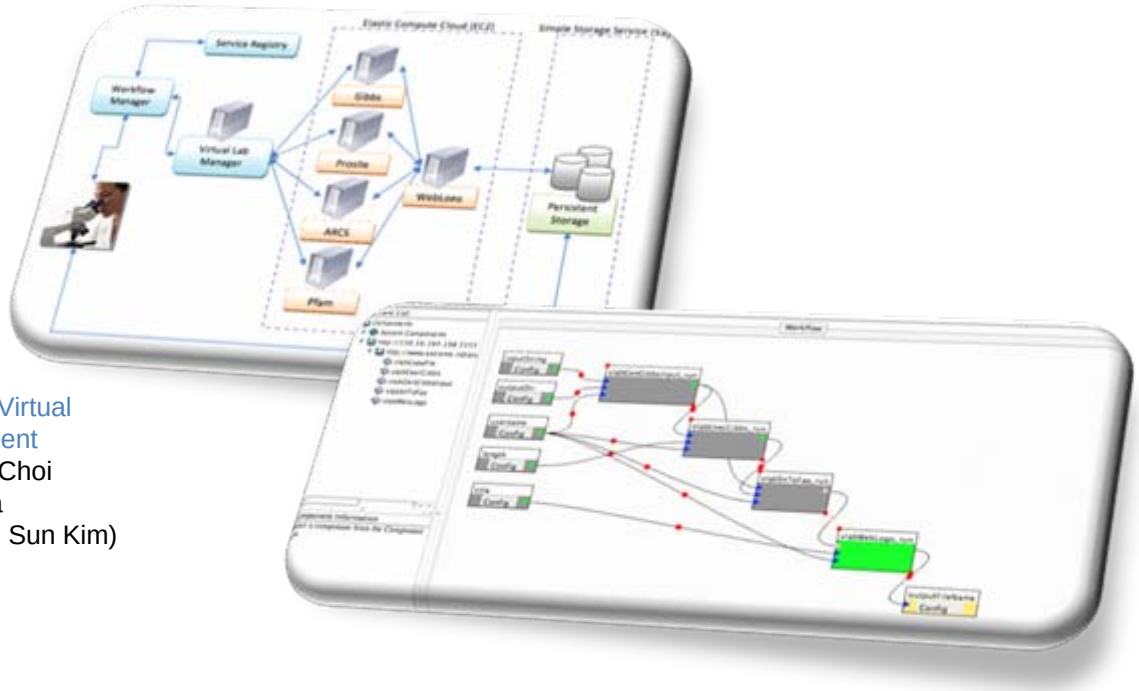
Example: Amazon Web Services

Simple Storage Service (S3)

- Storage for the Internet
- Simple Web Services interface to store and retrieve any amount of data from anywhere on the Web
- Standards-based REST and SOAP Web Service interfaces

Elastic Compute Cloud (EC2)

- Compute on demand
- Virtualization
- Integration with S3



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

eResearch: in the Cloud

General infrastructure and eResearch-oriented services

visualization and analysis services

scholarly communications

domain-specific services

search
books
citations

blogs &
social networking

Reference management

Project management

instant messaging

identity

mail

notification

document store

storage/data services

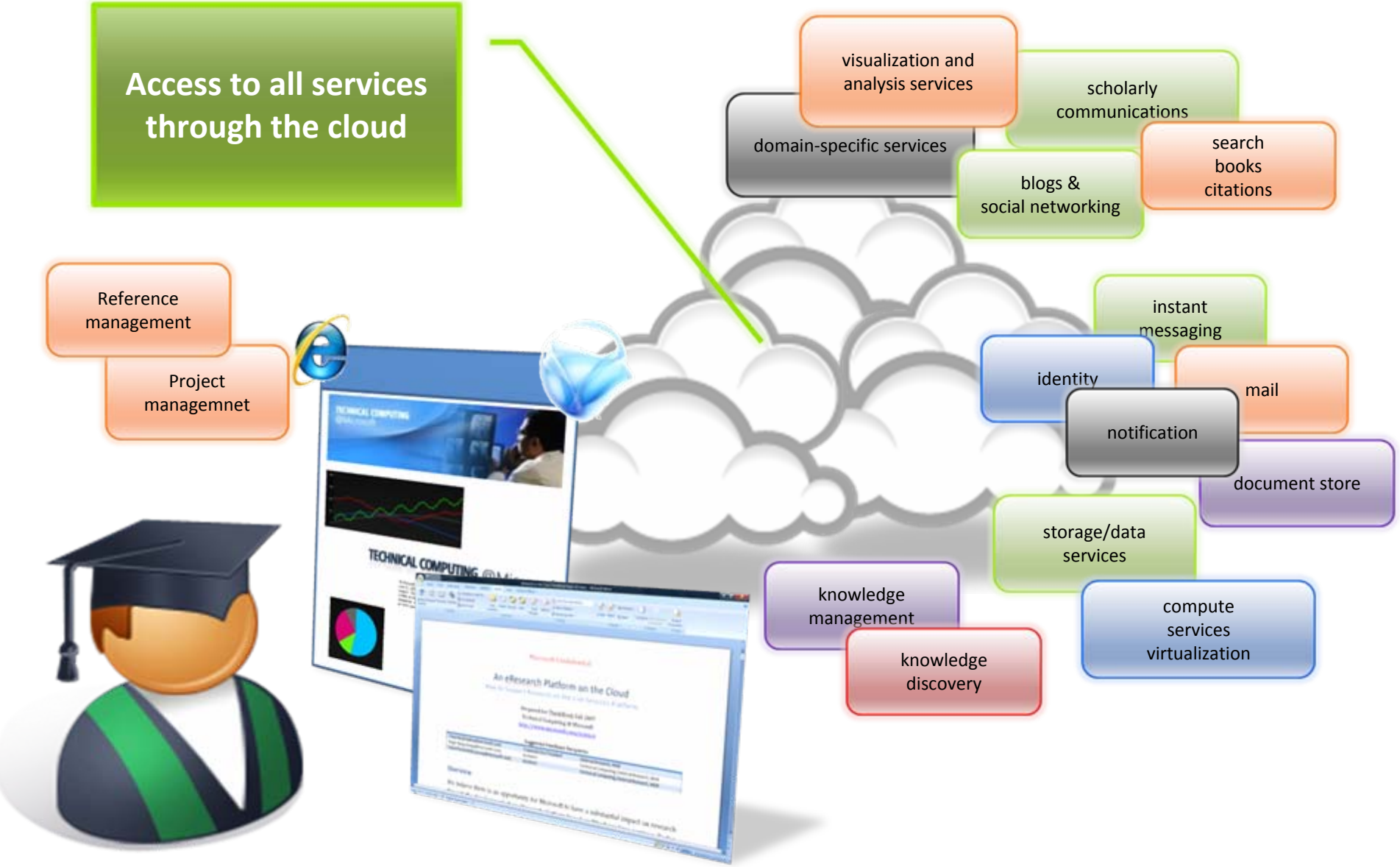
knowledge management

compute services virtualization

knowledge discovery

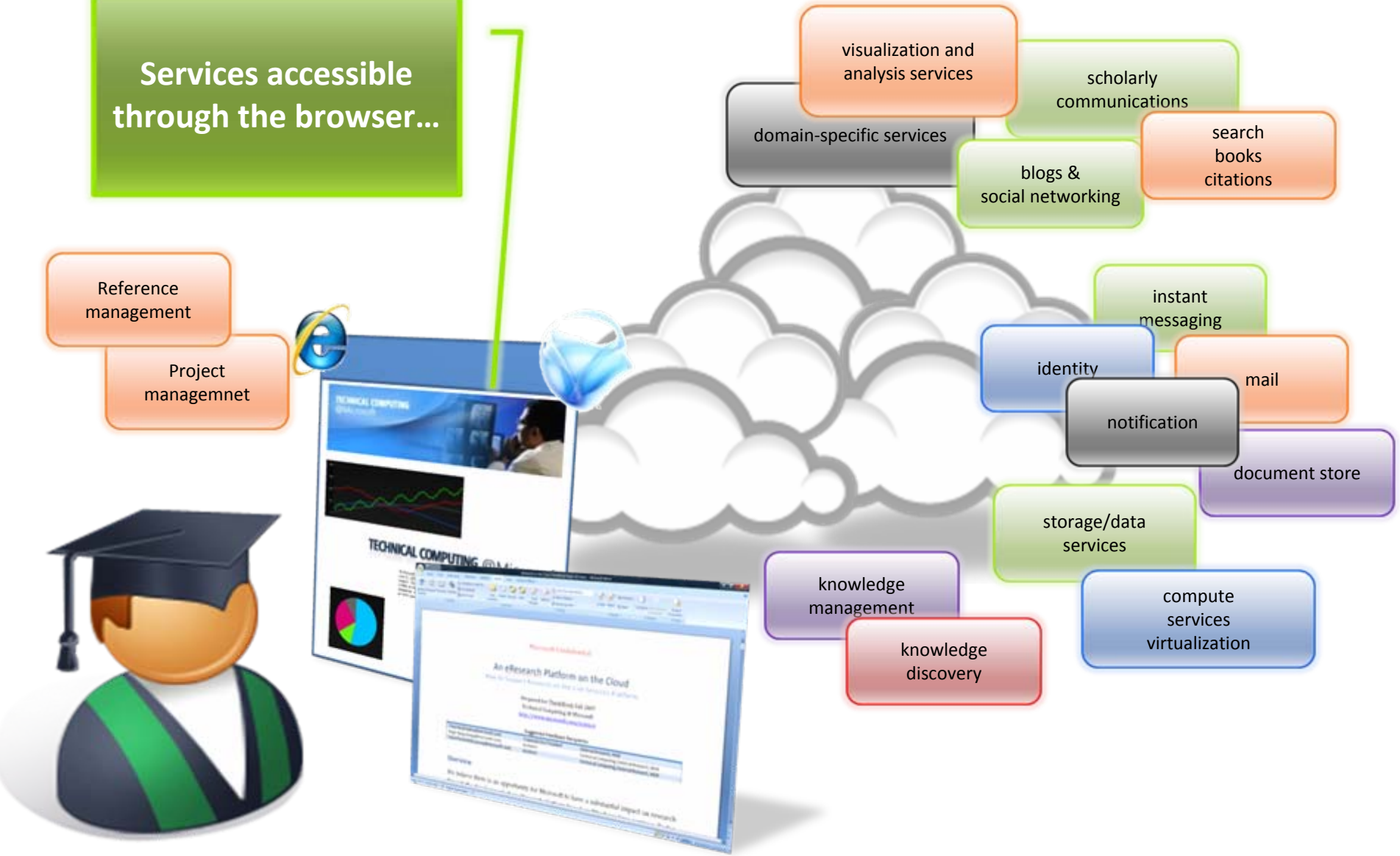
eResearch: in the Cloud

Access to all services through the cloud



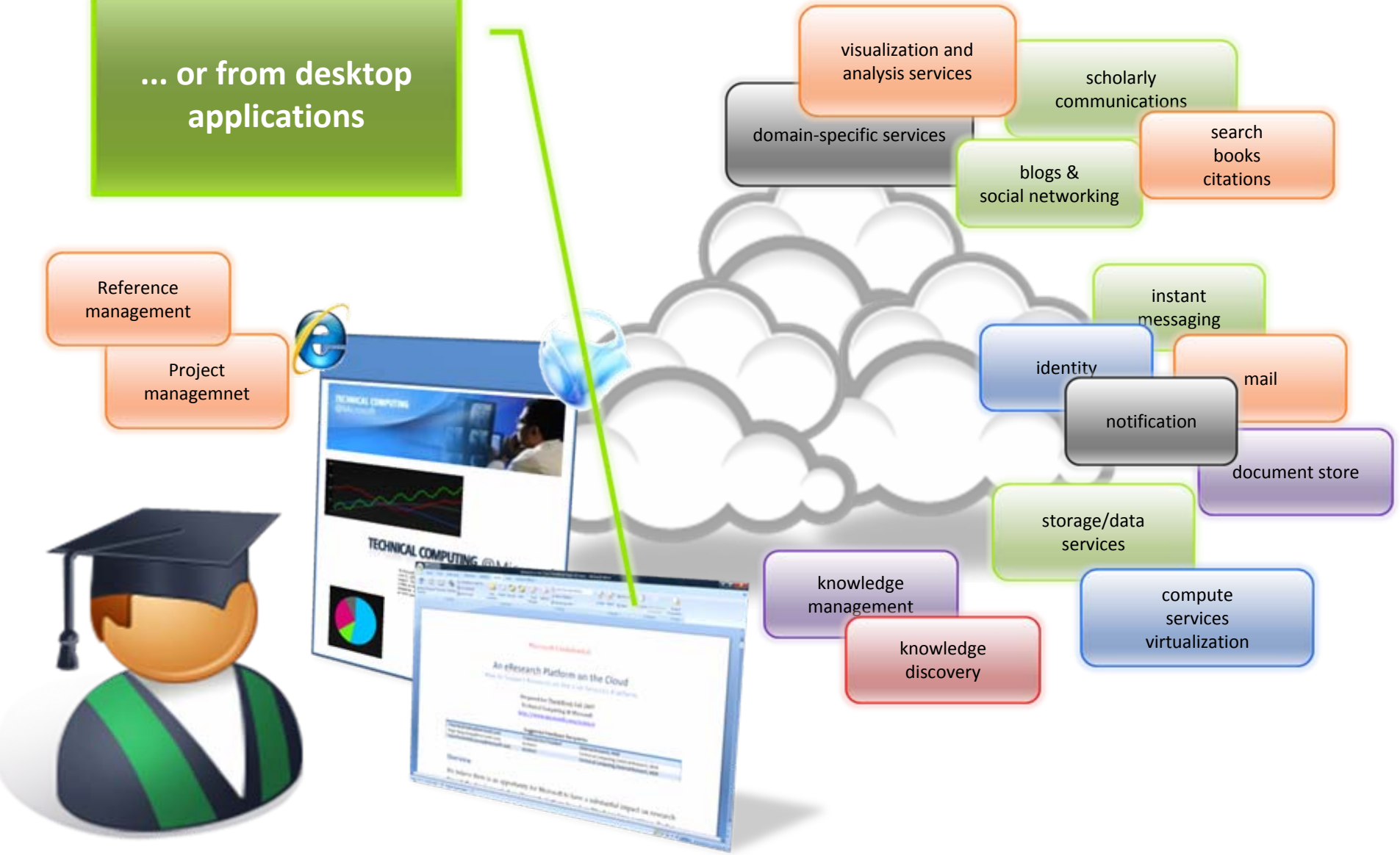
eResearch: in the Cloud

Services accessible through the browser…



eResearch: in the Cloud

... or from desktop applications



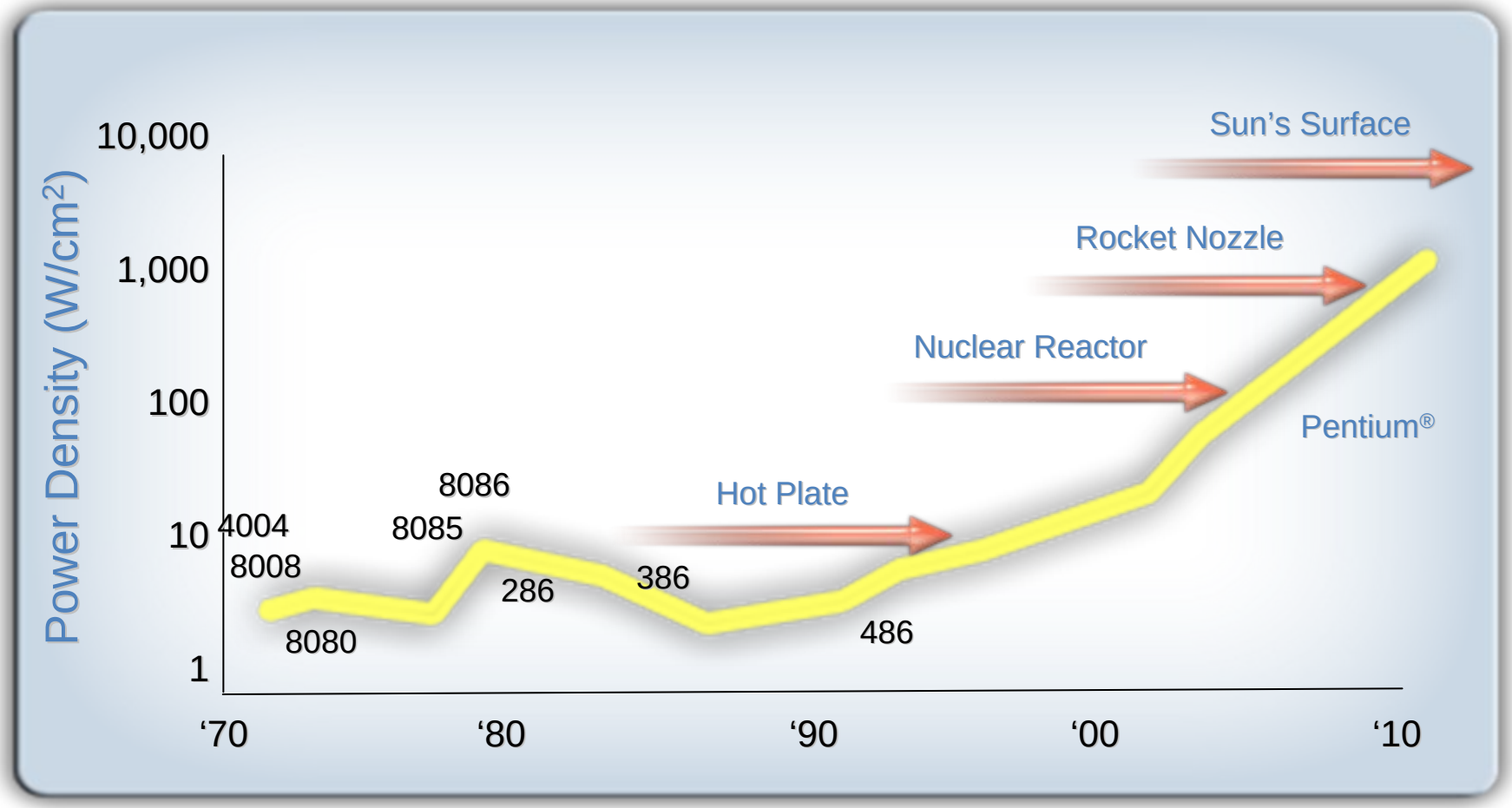
Data Intensive High Performance Computing

- Combination of Data Intensive HPC facilities and Software + Services
  - Will enable research community to address global scientific grand challenges
- More affordable/sustainable infrastructure for research organizations
  - Major IT companies bringing down the cost of the infrastructure
  - Will be cheaper to pay-per-use than own/manage private infrastructure

➢ "Research-as-a-service"

Today's CPU Architecture

Heat becoming an unmanageable problem



Intel Developer Forum, Spring 2004 - Pat Gelsinger

Center for Research on Multi-Core Computing

Project leaders:

Rice University – Keith Cooper & John Meller-Crummey
Univ. of Tennessee – Jack Dongarra
Indiana University – Geoffrey Fox, Dennis Gannon & Beth Plale

Project Focus:

- Rice: Compiler and tools support
  - Tools to understand and predict performance
  - Compiler transformations to improve memory hierarchy
- Tennessee: Libraries and fault tolerance support
  - Linear Algebra on Multicore architectures
- Indiana: Run-time systems and applications
  - Analyze three application classes - machine learning, computer chess, & event driven simulation for use w/ multicore languages and STM
  - Programming data stream driven applications
  - Evaluation of CCR and DSS

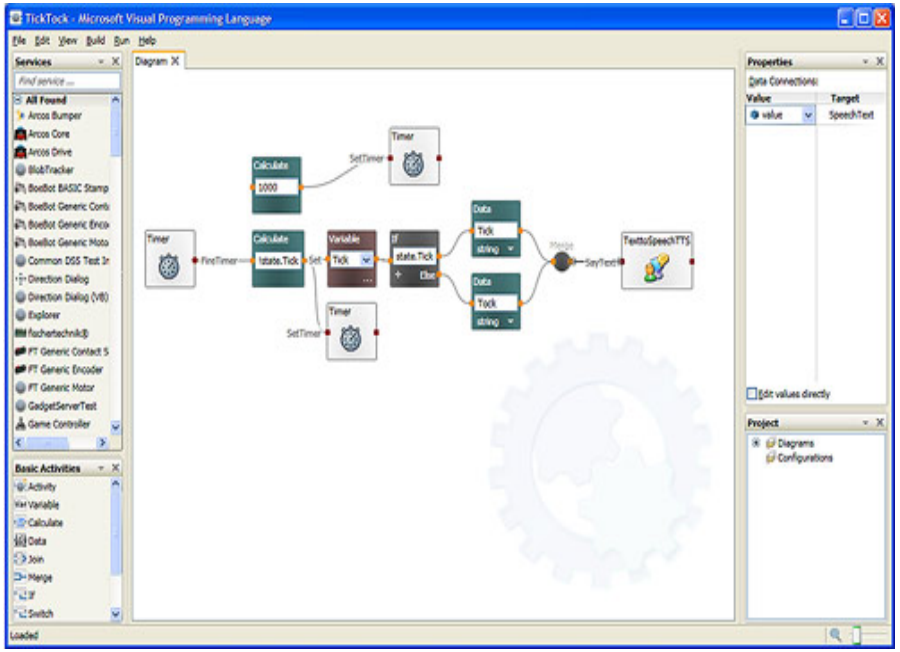
INDIANA UNIVERSITY

RICE

THE UNIVERSITY of TENNESSEE

Microsoft Robotics Studio

Authoring



Visual Programming Language

- Dataflow diagrams
- Drag and drop
- Services as blocks
  - Extensible
  - Connections = messages
  - Concurrency automatically determined and managed (using DSS semantics)
- Novice to expert

The Multi-Core Project at Barcelona Supercomputer Centre

- Project leader: Prof. Mateo Valero, Barcelona Supercomputer Centre of the Polytechnic University of Barcelona
- Project Focus: Research and development of architectural ideas for the next generation of multiprocessor chips
- Objectives: Use of transactional programming models and possible optimizations to
  - Reduce implementation and content costs
  - Support for single assignment variables
  - Nested parallelism and non-structured parallelism
  - Relevance of thread scheduling, load balance and locality issues
- Collaboration with Simon Peyton-Jones and Tim Harris

UNSW
THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY • AUSTRALIA

Multi-Core Project
Nested Data Parallel Haskell

- Project leader: Simon Peyton Jones – MSR Cambridge
  - Manuel Chakravarty & Gabriele Keller – Univ of New South Wales
  - Post-Doc: Roman Leshchinskiy
- Project Scope: Implement Nested Data Parallelism in Glasgow Haskell Compiler (GHC)
  - Design performs very aggressive flattening and loop fusion
  - Programmer can write modular programs that operate on successive intermediate, nested arrays, while the implementation generates a small number of complex fused loops that operate over flat arrays, with few intermediates.
- Project deliverables
  - Complete design and implementation, resulting in a complete implementation of nested data parallelism in Haskell
  - demonstrate its effectiveness on two or three realistic exemplars

Microsoft® Research

Quantum Computing

- Multidisciplinary Institute at UCSB directed by Michael Freedman of MSR
- Looking at novel material physics and the 2-dimensional Fractional Quantum Hall effect
- Exploring non-Abelian ‘Anyon’ excitations to protect coherence of qubits and quantum gates

➢ See paper by Sarma, Freedman and Nayak, in Physics Today July 2006 for a more detailed description of ‘Project Q’s search for quantum immunity’

Microsoft®

Your potential. Our passion.™