The Social Grid

Leveraging the Power of the Web
Focus on Development Simplicity

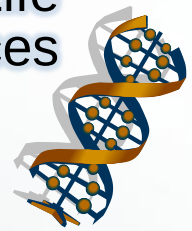
Roger S. Barga, PhD
Architect, Technical Computing at Microsoft

Technical Computing at Microsoft

http://www.microsoft.com/science

Earth Sciences

Life Sciences

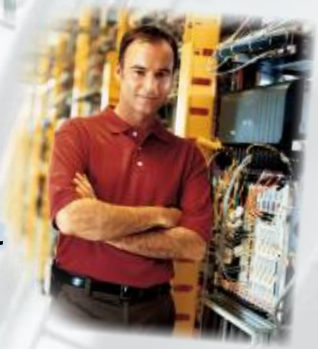


Social Sciences

Accelerating Discovery

Collaborative Research

New Materials, Technologies & Processes



Computer & Information Sciences

$E=MC^2$

Math and Physical Science

Emergence of a New Science Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

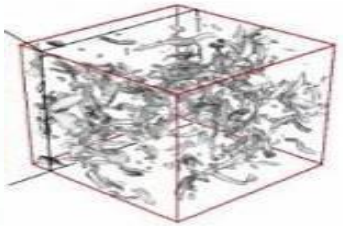
Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

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

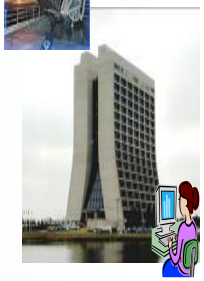
Last few decades – **Computational Science**

- Simulation of complex phenomena



Today – **Data-Centric Science or eScience**

- Unify theory, experiment, and simulation
- Using data exploration and data mining
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks



Slide thanks to Jim Gray

A New Set of Challenges

- Grid middleware and cyberinfrasturcture is fast becoming too complex and difficult to use
- Support the transition to data-centric eScience
  - *Computation is no longer the bottleneck*
- New dynamic in science investigations
  - *Collaborative, distributed, cross disciplinary*
  - *Science is becoming highly social (Research 2.0)*
- Need for simplicity and ease of development
- Can we learn lessons from the Web

The Web as a Platform for Research?

- Has tremendous momentum
- It is **the** channel for result dissemination
- The browser is the **universal canvas** for the delivery of information and functionality
- Web protocols, technologies, and middleware are well supported by the IT industry
- Today it is the contemporary **platform** for distributed, internet-scale applications
- Emergence of "*software-as-a-service*"
- Collection of 'Web 2.0' technologies is maturing

Scientific Data Servers for Hydrology

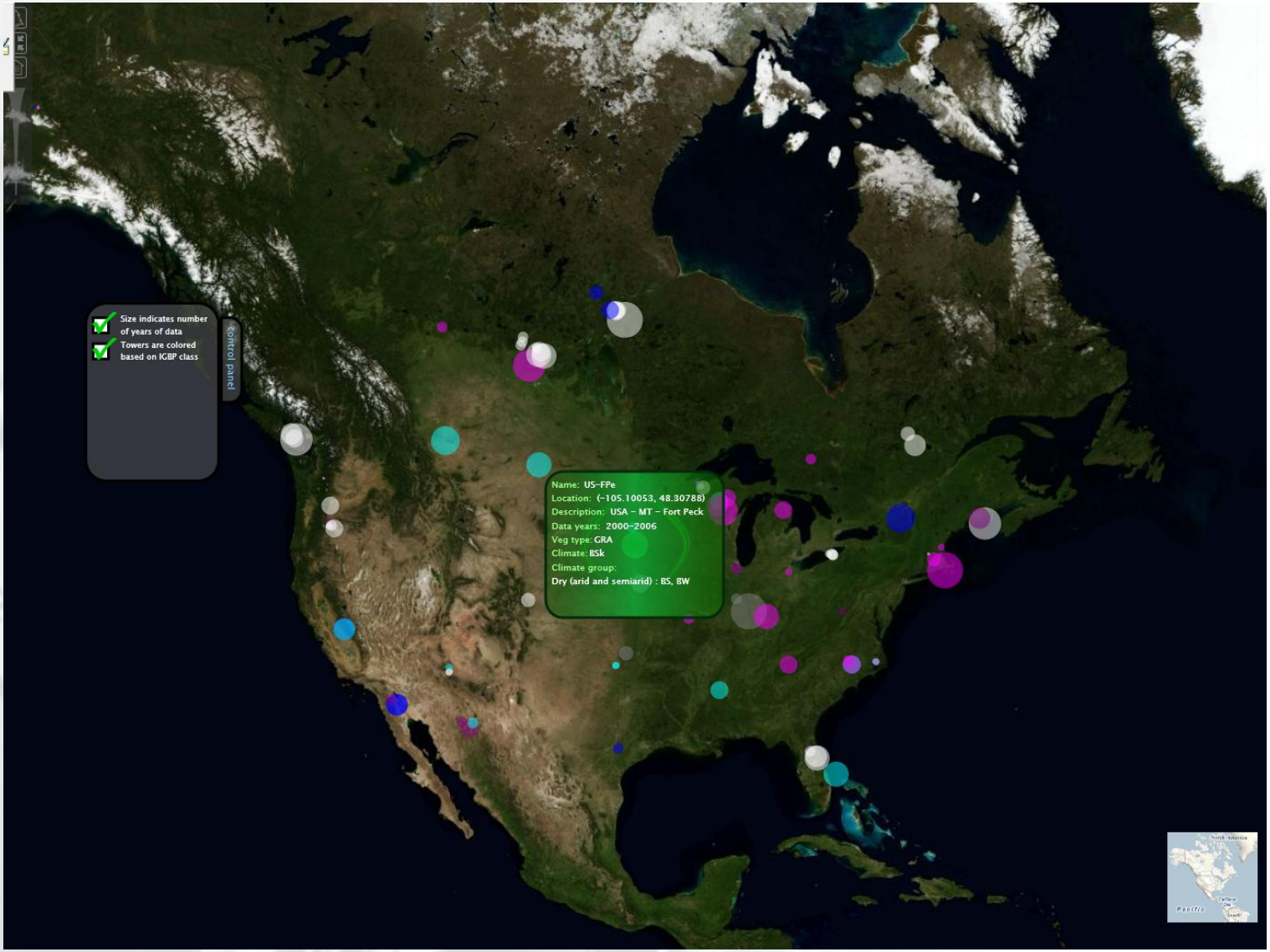
Work with Berkeley Water Center to use modern (relational) database technology

- 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/

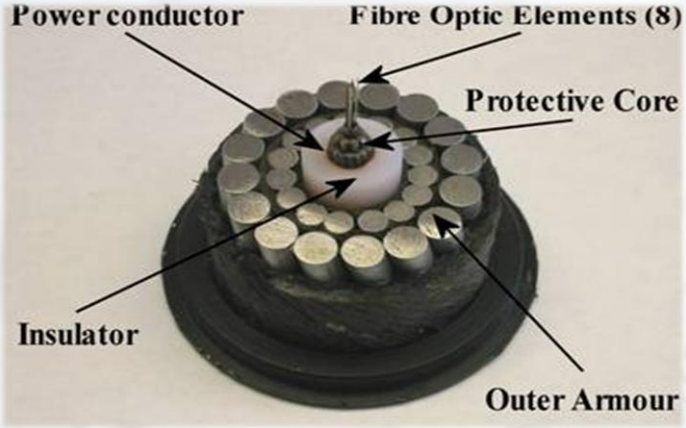
Microsoft Project Lead: Catharine van Ingen

Mashup of Ameriflux Sites



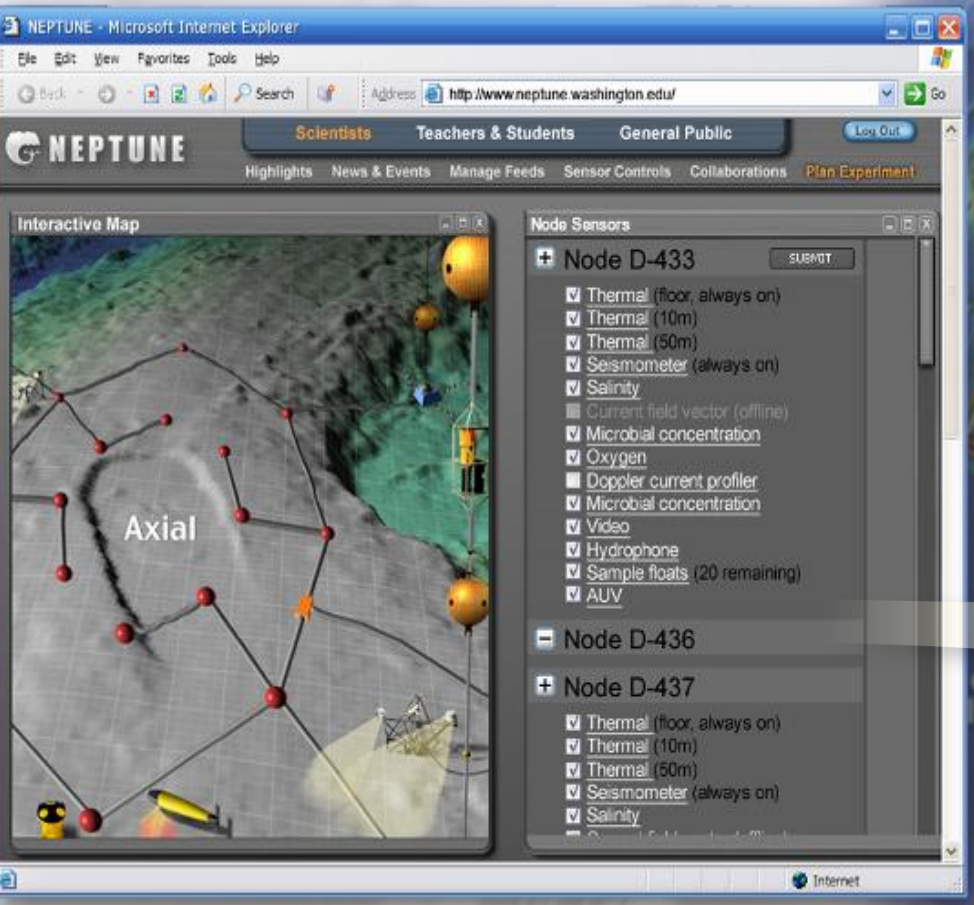
Work of Savas Parastatidis

Project Neptune



http://www.neptune.washington.edu/

Programmable Sensors & Remote Instruments



Undersea Sensor Network

Connected & Controllable Over the Internet

Explorer Plate

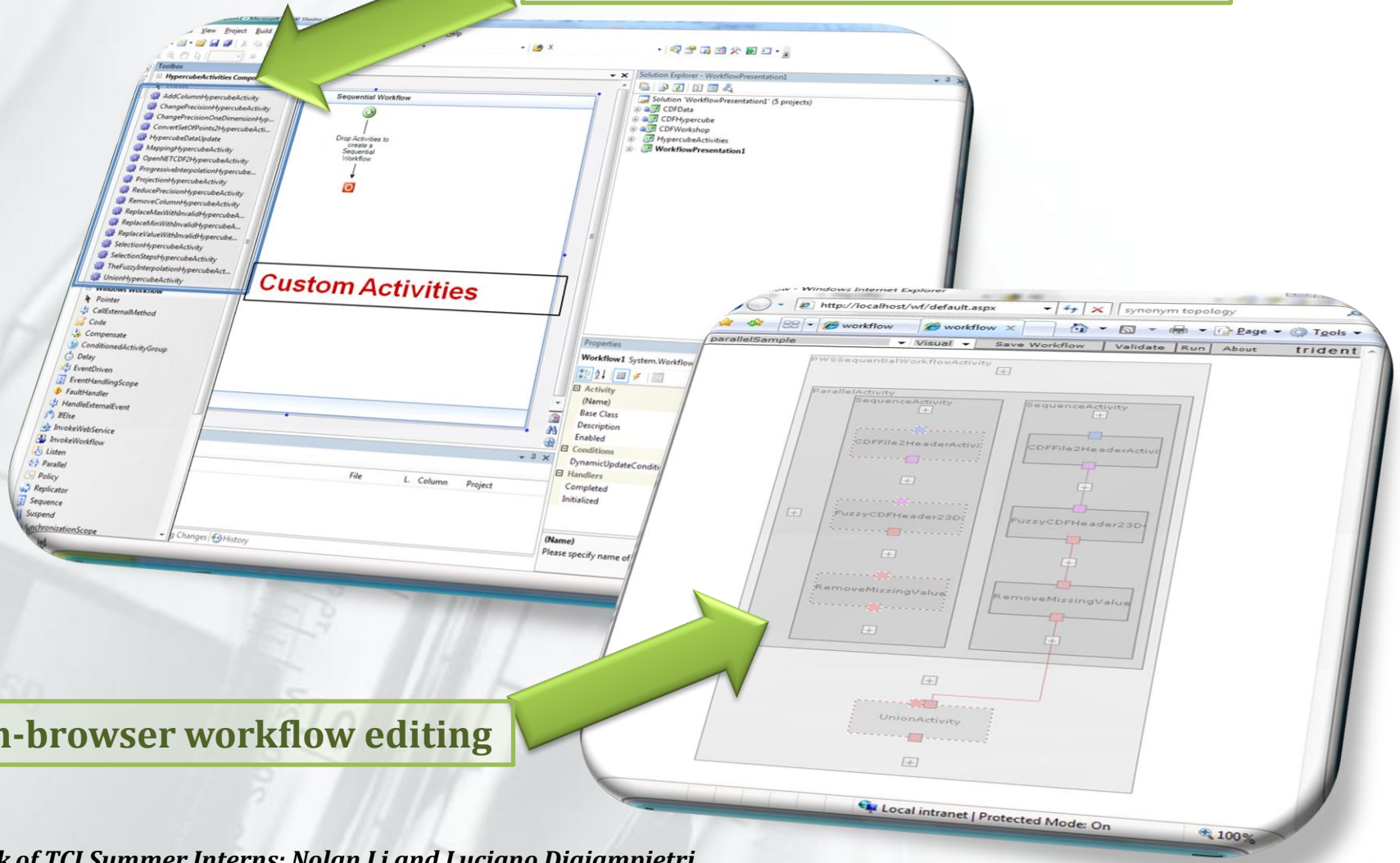
Juan de Fuca Plate

North American Plate

Gorda Plate

Trident – Scientific Workflow for Neptune

Custom activities for oceanographers



Work of TCI Summer Interns: Nolan Li and Luciano Digiampietri

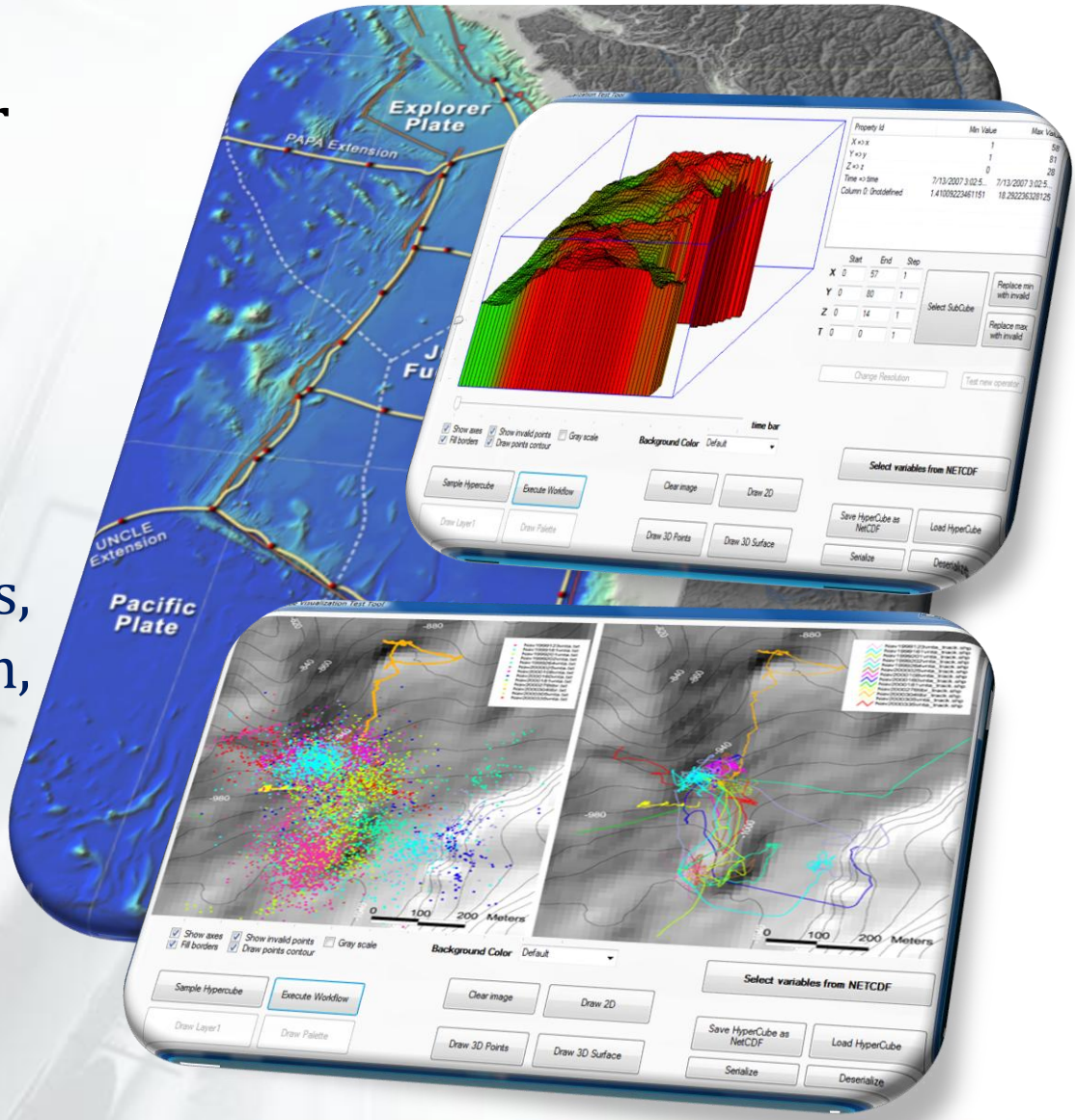
Trident for Neptune

Workflow workbench for oceanography

From raw sensor data to useable data products

- Automatic data cleaning,
- Integrates multiple models,
- Regridding & interpolation,
- Analysis

Real time, on-demand visualization from the Neptune sensor array



myGrid and the Taverna Workflow System

- Independent third party world-wide service providers of applications, tools and data sets – *in the cloud*.
  - 850 databases, 166 web servers: Nucleic Acids Research Jan 2006.
- My local applications, tools and datasets. In the Enterprise. In the laboratory.
- Easily incorporate new service without coding. So even more services from the cloud and enterprise.

- 3500+ service operations
- All major providers
- Integration application for service providers like BioMOBY and BioMART

Slide thanks to Carole Goble and David DeRoure

eScientists in the Cloud

Individual life scientists, in under-resourced labs, who use other people's applications, with little systems support.

- Exploratory workflows
- Developers (often) the users.
- Consumers are providers.

- A distributed, disconnected community of scientists.
- Decoupled suppliers and consumers of services and workflows.
- Scientists in an enterprise and in large projects
- Scientists out of the enterprise, in small projects or sole traders.

Slide thanks to Carole Goble and David DeRoure

200+ projects and sites, ~1000 individual users.
Users throughout UK, USA, Europe, and SE Asia

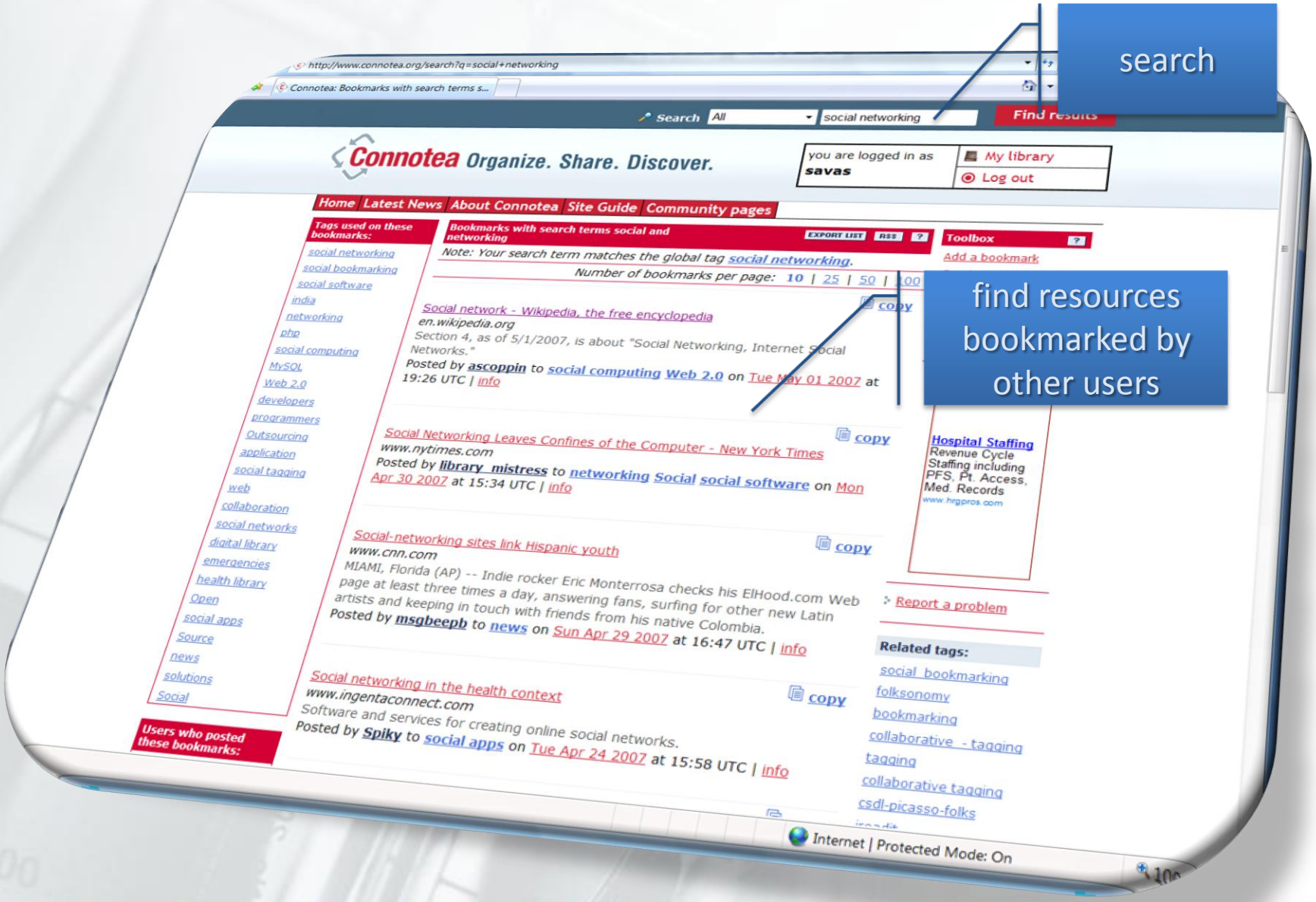
On the Web

- Users generate content on the Web
  - Blogs, wikis, photographs, videos, etc.
  - They do not have to know HTML
- They form communities
  - Social networks, virtual worlds
- They interact, collaborate, share
  - Instant messaging, web forums, content sites
- They consume information and services
  - Search, annotate, syndicate

And Scientists Today…

- Annotate, share, discover data
- Collaborate, exchange ideas over the Web
- Create communities, social networks
- Use workflow tools to compose services

Example – Connotea (Nature Publishing)



Mashups: Composing Data and Functionality

SensorMap

Functionality: Map navigation

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

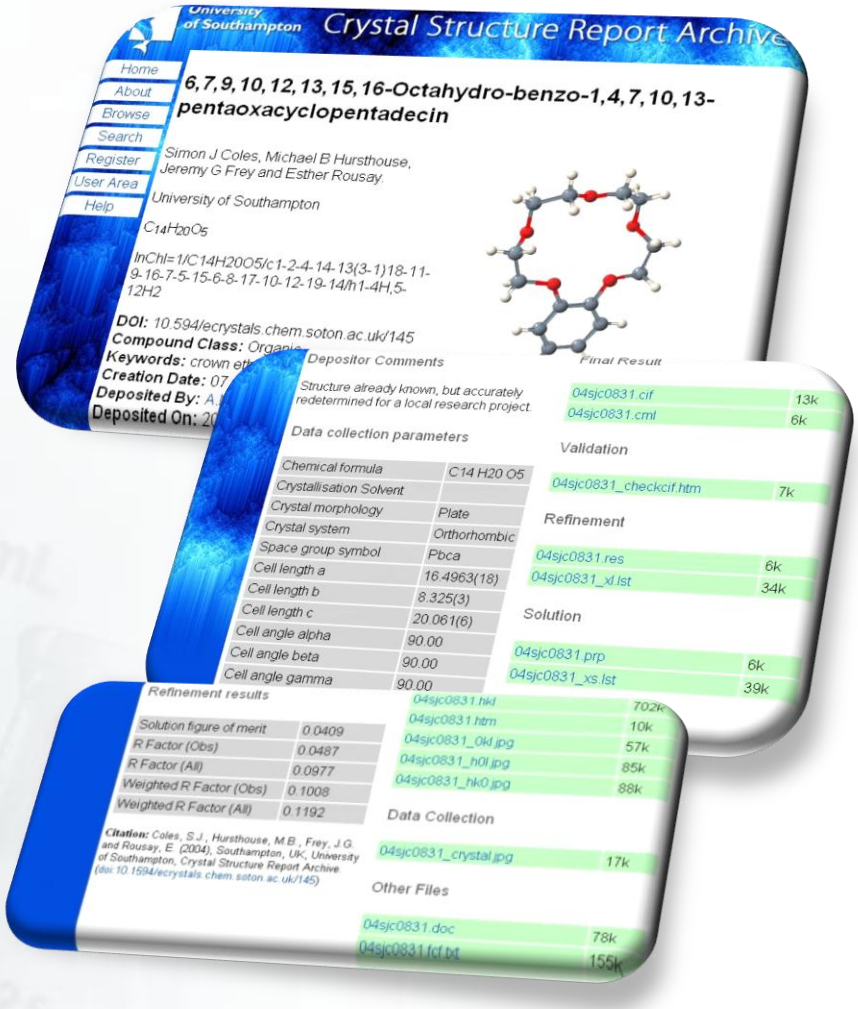
The Web as a Platform for eResearch

Services not middleware

- *No need to install many thousands of lines of middleware*

Core Services in the Cloud

- Identity
- Blogging, Messaging
- Search, Discovery
- Data processing/visualization
- Content upload, sharing, discovery
- Computation and Storage

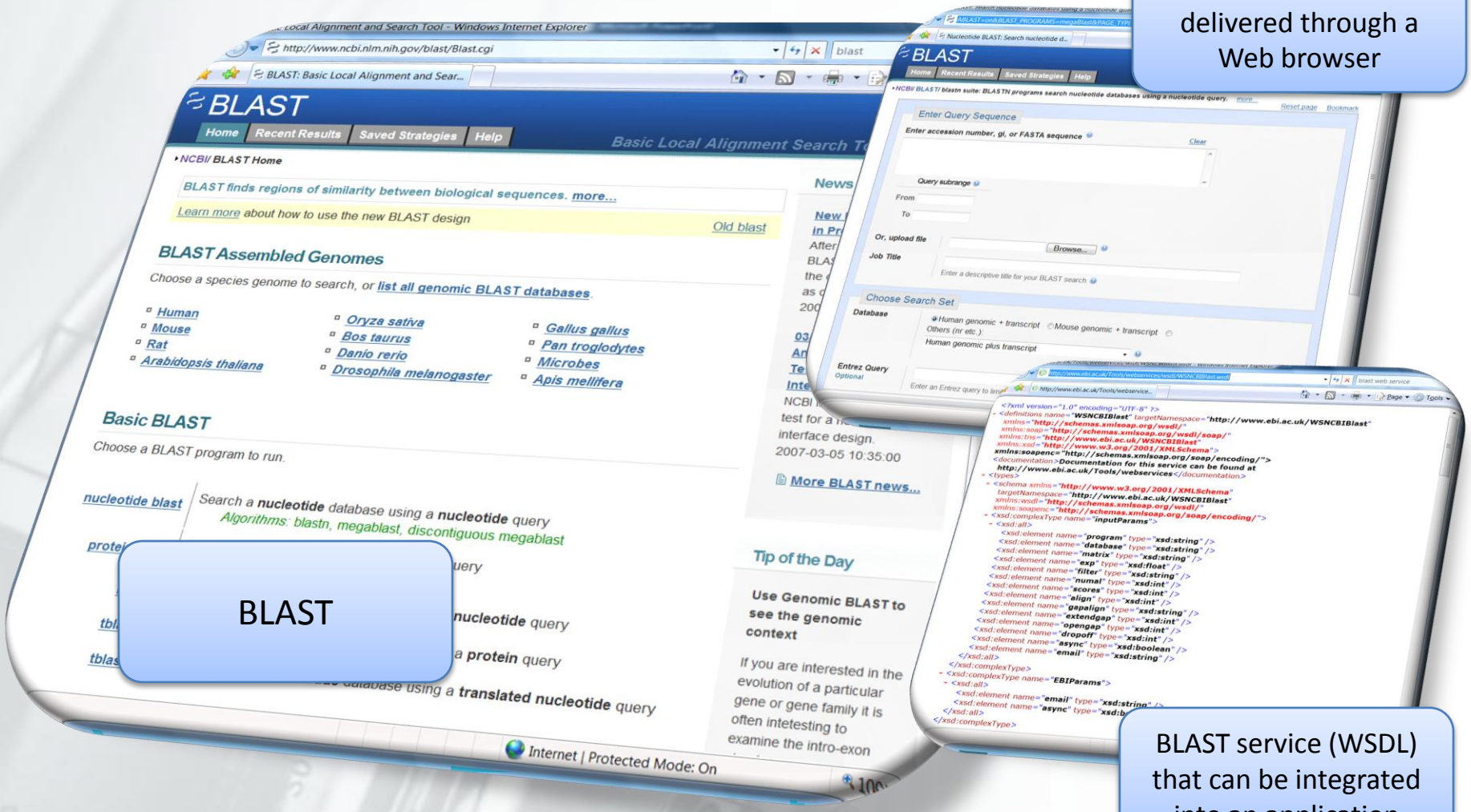


http://ecrystals.chem.soton.ac.uk

Thanks to Jeremy Frey

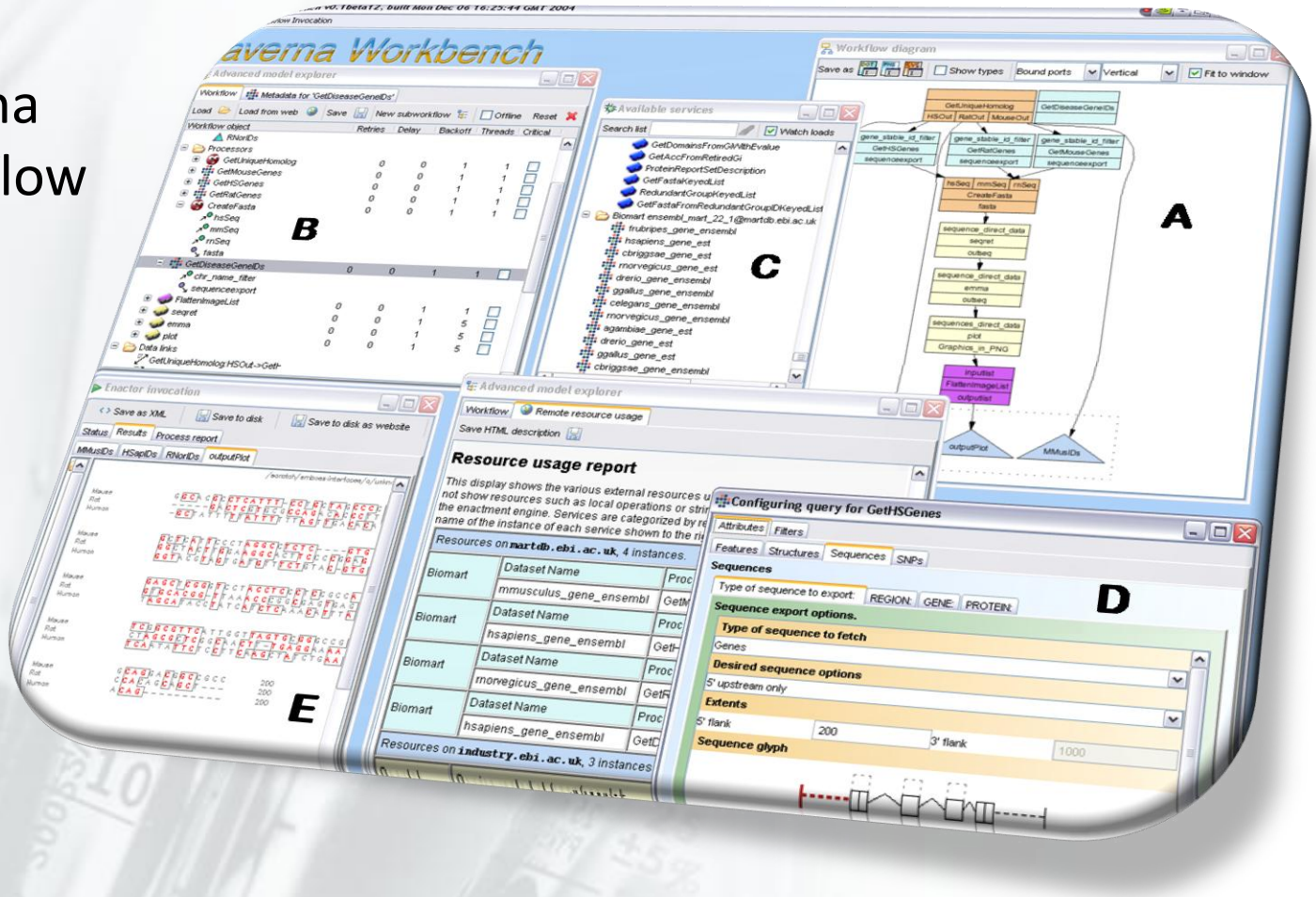
Data and Services can be Accessed Securely

Services Expose Functionality



Services can be Composed

Taverna Workflow



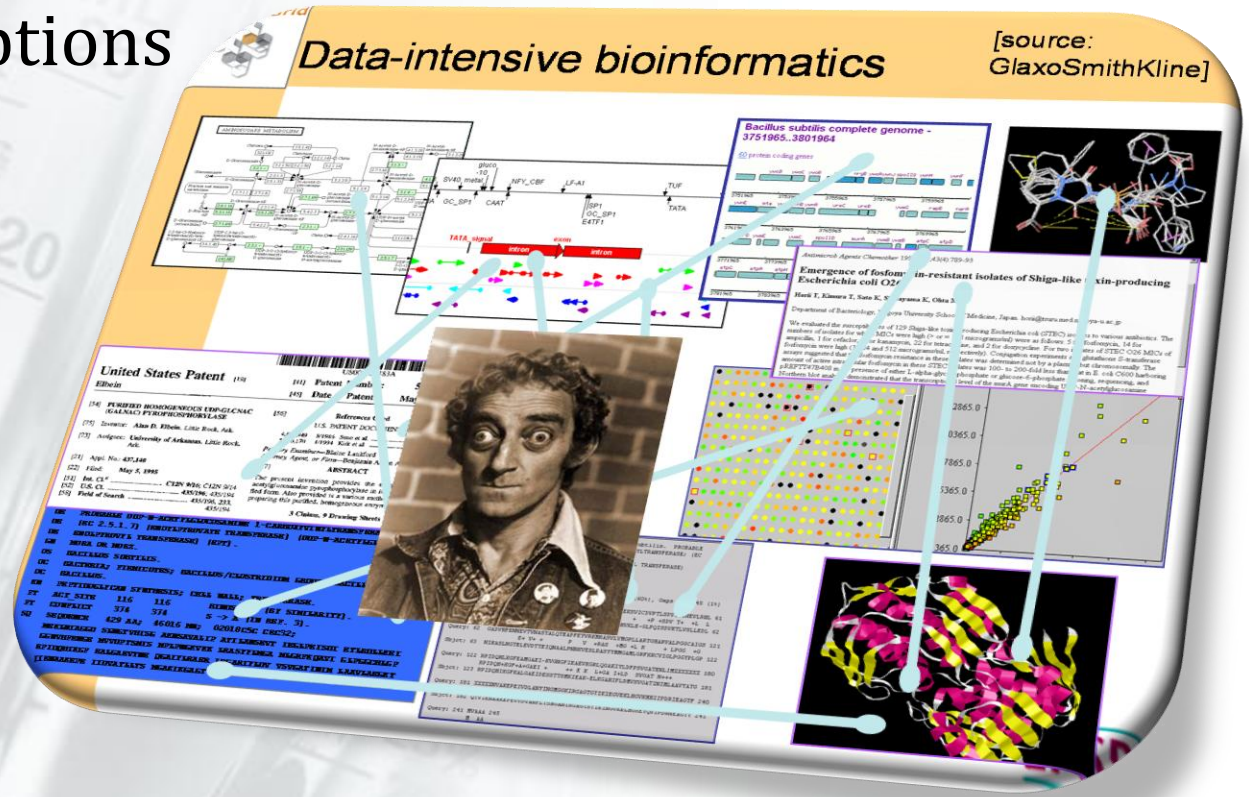
Data is Easily Shareable

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

Knowledge can be created/published/archived/discovered

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

myGrid



Grids in Industry

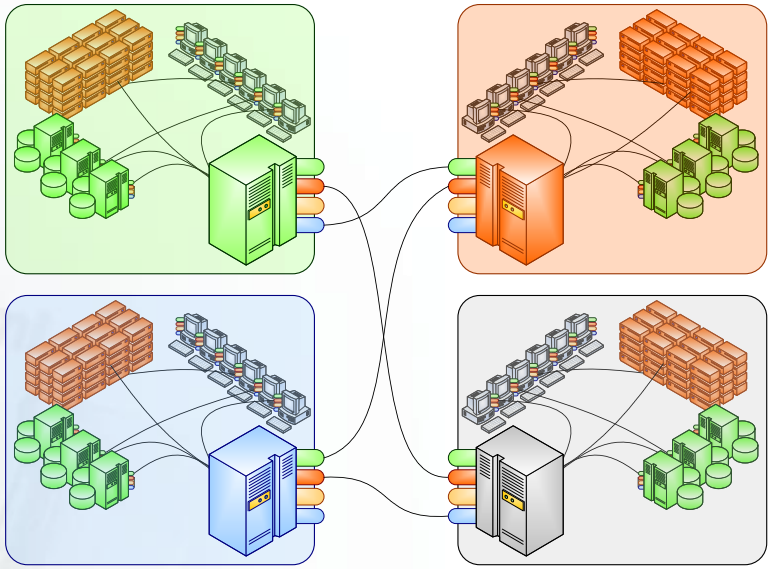
- Google, Amazon, Yahoo, eBay and Microsoft are the major 'Cloud Platform' providers
  - All have infrastructures of hundreds of thousands of servers
  - Many large data centers, distributed across multiple continents
  - Have developed proprietary technologies for job scheduling, data sharing and management
  - Care about power consumption, fault tolerance, scalability, operational costs, performance, etc.

They are living the "Grid dream" on a daily basis

Google as an Example

- Estimated 450,000* servers distributed around the world

*source: Wikipedia



- Google File System - highly distributed, resilient to failures, parallel, etc.
- Schedulers and load balancers for the distribution of work
  - Use their 'Map-Reduce' middleware as parallel computational model

Amazon web services: simple storage service (s3)

- S3 is storage for the Internet
  - Designed to make web-scale computing easier for developers
- Provides a simple Web Services interface to store and retrieve any amount of data from anywhere on the Web
  - 'CRUD' philosophy – Create, Read, Update and Delete operations
- Uses simple standards-based REST and SOAP Web Service interfaces
  - Built to be flexible so that protocol or functional layers can easily be added

Amazon s3 Functionality

- Intentionally built with a minimal feature set
  - Write, read, and delete objects containing from 1 byte to 5 gigabytes of data each
- Can store unlimited number of objects
  - Each object is stored and retrieved via a unique, developer-assigned key
- Authentication mechanisms provided
  - Objects can be made private or public, and rights can be granted to specific users
- Default download protocol is HTTP
  - BitTorrent protocol interface is provided to lower costs for high-scale distribution

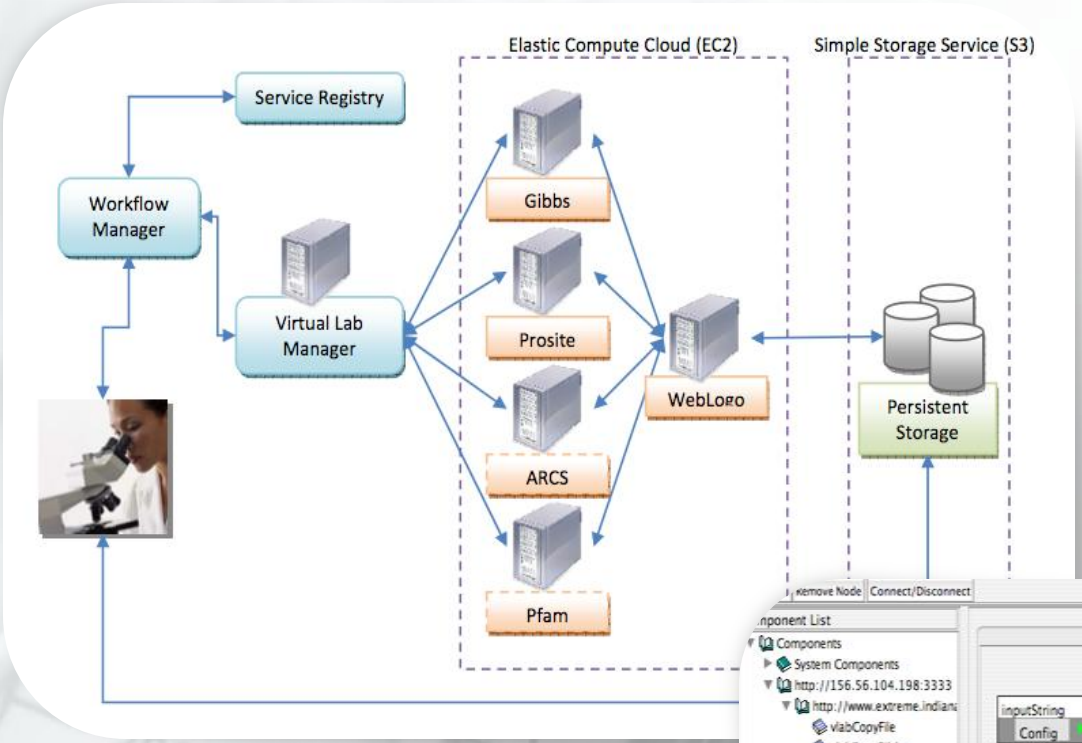
Amazon web services: elastic compute cloud (ec2)

- Compute on demand service that works seamlessly with their S3 storage service
- Create Amazon Machine Image (AMI) containing application, libraries and data
- Use EC2 Web Service to configure security and network access
- Use EC2 to start, terminate and monitor as many instances of your AMI as you want

  Each instance has:
  - 1.7 GHz x86 Processor
  - 1.75 GB RAM
  - 160 GB local disk
  - 250 MB/s network bandwidth

Used by Catlett and Beckman as capacity computing alternative to TeraGrid 'SPRUCE' capability computing for emergency urgent response

A Grad Student Project Using S3 and EC2

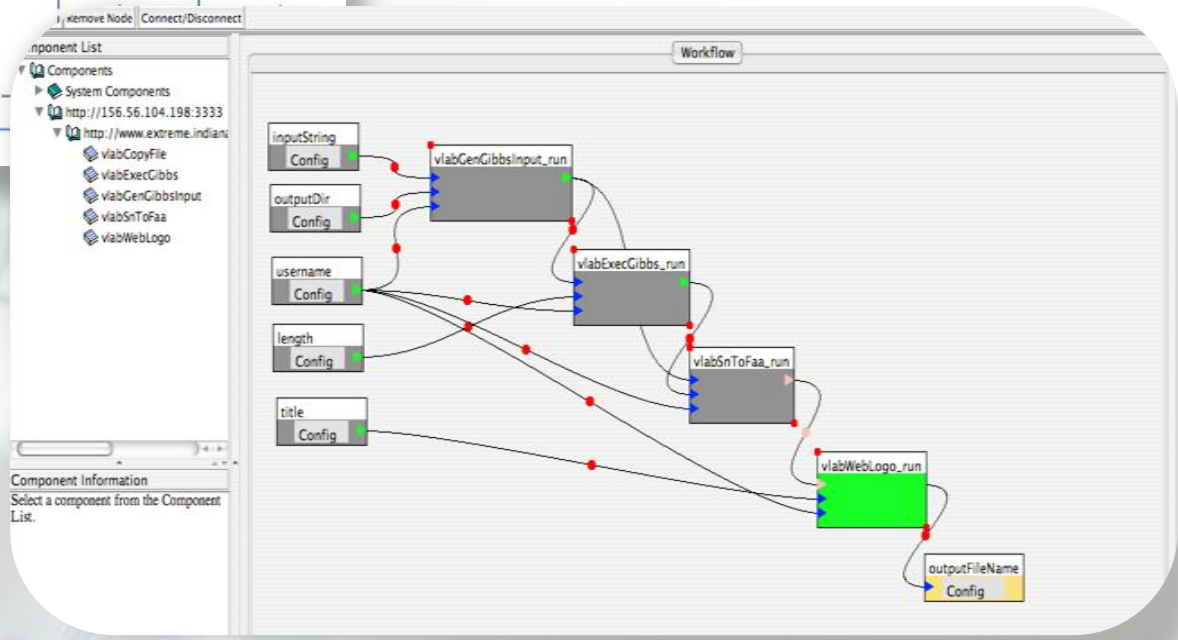


Gene Analysis Virtual Lab Experiment

by Jong Youl Choi

at Indiana

(For Beth Plale and Sun Kim)



Data-Intensive High-Performance Computing

- A new generation of facilities to support eResearch on the cloud
- Data-intensive
  - Large storage capacity
  - Functional-style programming for data filtering, searching (e.g. MapReduce)
  - Storage-as-a-service
- Compute-intensive
  - State-of-the-art clusters
  - No need to be the fastest in the world; few top100 ones
  - Scientific applications-as-services

Social Grids and the Web

A Call to Action

- Focus on solutions for scientific/technical computing and not just on infrastructure
- Focus on "data-centric eScience"
  - Help domain experts define formats for representing and annotating domain-specific data
- Keep it simple, build on known Web technologies
  - Solutions that "just work" without the need for complicated middleware platforms
  - Leverage only existing, Web infrastructure (HTTP, XML, simple Web Services, services in the cloud)

Microsoft®

Your potential. Our passion.™