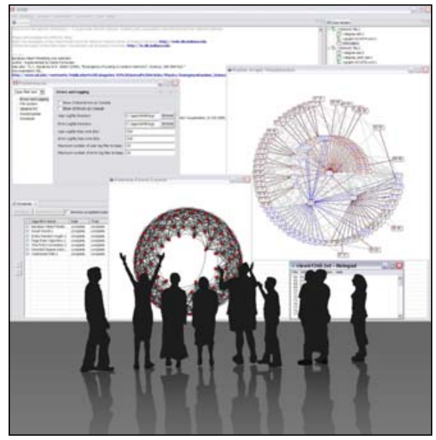
Creating Marketplaces for Science

The Emerging Characteristics of Modern Science Cyberinfrastructures

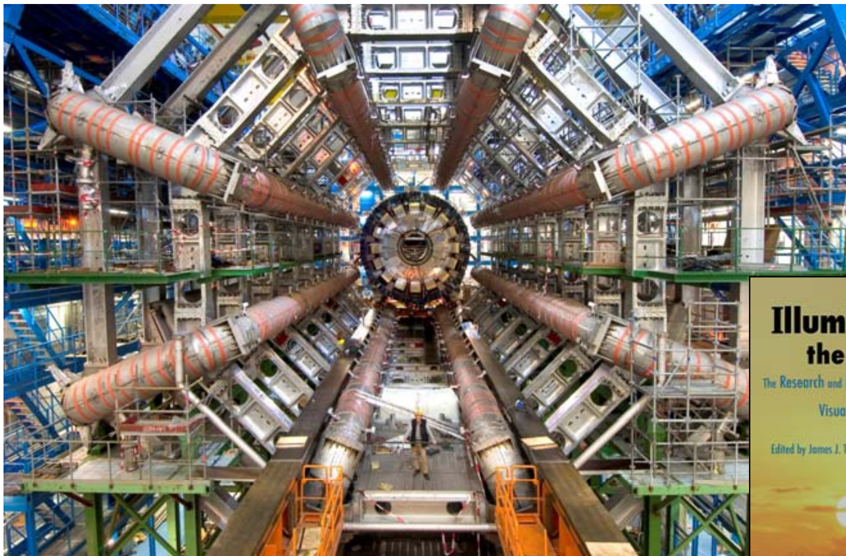
12th SANKEN International Symposium
Osaka University
22 January 2009



Russell J. Duhon
rduhon@indiana.edu
Cyberinfrastructure for Network Science Center
School of Library and Information Science
Indiana University
Bloomington, IN, USA

Table of Contents

The Problem of Increase



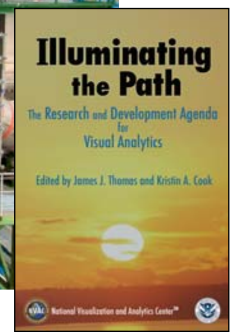


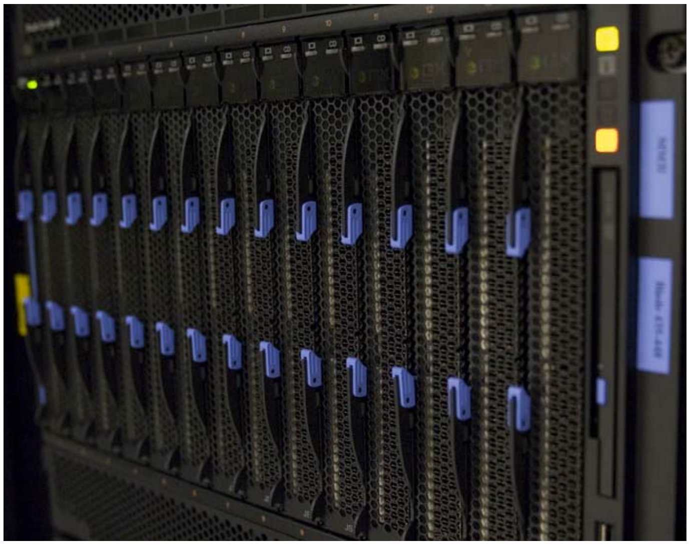
image courtesy www.gridpp.ac.uk

"For centuries we have been improving our ability to collect information, and this will continue. However, our ability to analyze this information is sorely lacking. This information is massive, complex, incomplete, and uncertain, and it encompasses all data forms, languages, and cultures."

- Illuminating the Path: The Research and Development Agenda for Visual Analytics

What is Cyberinfrastructure?

What is Cyberinfrastructure?



Cyberinfrastructure is Supercomputers!

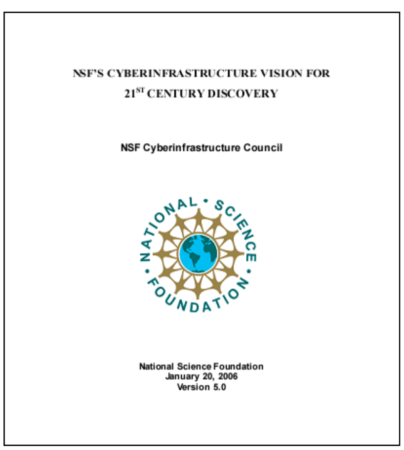
Cyberinfrastructure is the Grid?

. . .maybe Cyberinfrastructure includes Portals?

One way to think about Science Cyberinfrastructure

Computational tools deployed to support and enable scientific investigation.

Applied as a definition, this makes almost everything a Cyberinfrastructure.



NSF'S CYBERINFRASTRUCTURE VISION FOR
21ST CENTURY DISCOVERY

NSF Cyberinfrastructure Council

NATIONAL SCIENCE FOUNDATION

National Science Foundation
January 20, 2006
Version 5.0

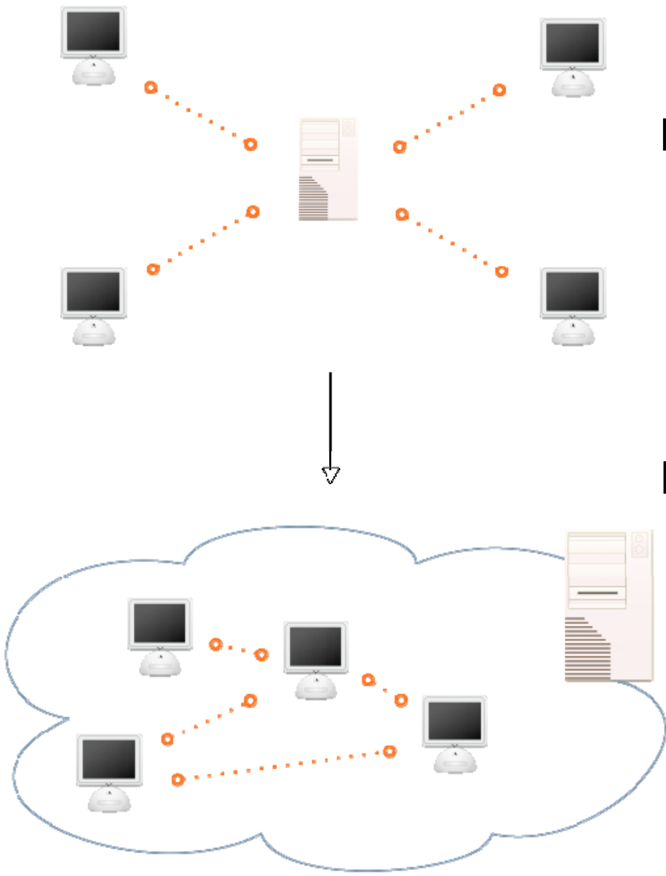
But as a guiding statement, it can focus motivation and approach to building Cyberinfrastructures.

"Fewer and fewer researchers working at the frontiers of knowledge can carry out their work without cyberinfrastructure of one form or another."
- Cyberinfrastructure Vision for 21st Century Discovery, a report by the US National Science Foundation.

The World's Largest Cyberinfrastructure

. . . is the World Wide Web.

Early on, the web was about people making and visiting websites.



However, the growth of the web saw sites about more than being visited flourish.

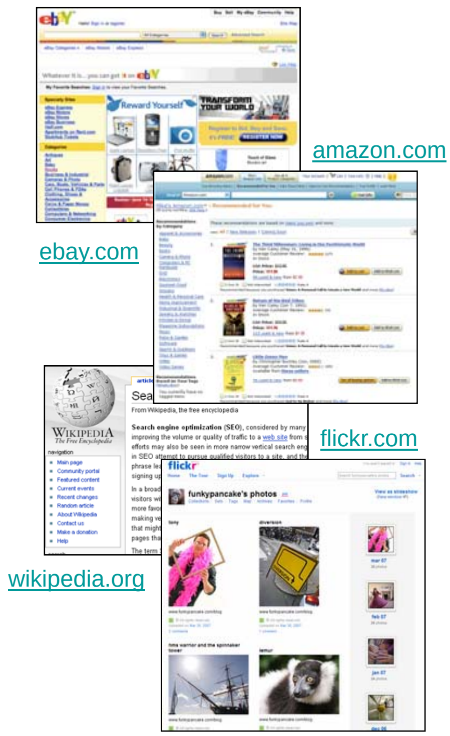
Instead of being about large numbers of visitors, they are about the interactions afforded those visitors.

The World's Largest Cyberinfrastructure, 2.0

One name given this model of the web is Web 2.0

Some use this to mean flashy interaction, but many of the most popular Web 2.0 sites are simple.

"[B]logging, tagging and social networking, dubbed Web 2.0, have . . . expanded people's ability not just to consume online information but to publish it, edit it and collaborate about it."
- M. Mitchell Waldrop. Science 2.0 – Is Open Access Science the Future?



eBay, bringing buyers and sellers together to synthesize reputation.

Amazon, where reviews and recommendations are the lifeblood.

Wikipedia, where everyone can contribute.

Flickr, where tags on user-uploaded images create folksonomies.

And many more.

Types of Human Interaction

Web 2.0 sites help humans:

- Contribute
- Rate
- Remix
- Collaborate

These capabilities also characterize modern science cyberinfrastructures.

Science. 7 March 2008: Vol. 319. no. 5868, pp. 1349 - 1350
DOI: 10.1126/science.1153539

Contributing means giving back to the community with content, effort, or thought.

Rating is implicitly or explicitly contributing to group estimations of the quality of contributions.

Users **Remix** to synthesize new contributions from existing ones.

Collaborations are joint creations that may become new contributions.

Cyberinfrastructures support these interactions by creating community marketplaces that use incentive structures to encourage participation.

Modern Science Cyberinfrastructures

- Connotea, Bibsonomy, CiteULike,
- Swivel, ManyEyes,
- nanoHUB, SciVee, ToBIG, LEAD,
- Network Workbench, CIShell,
- myExperiment,
- et cetera.

Connotea, Bibsonomy, CiteULike,

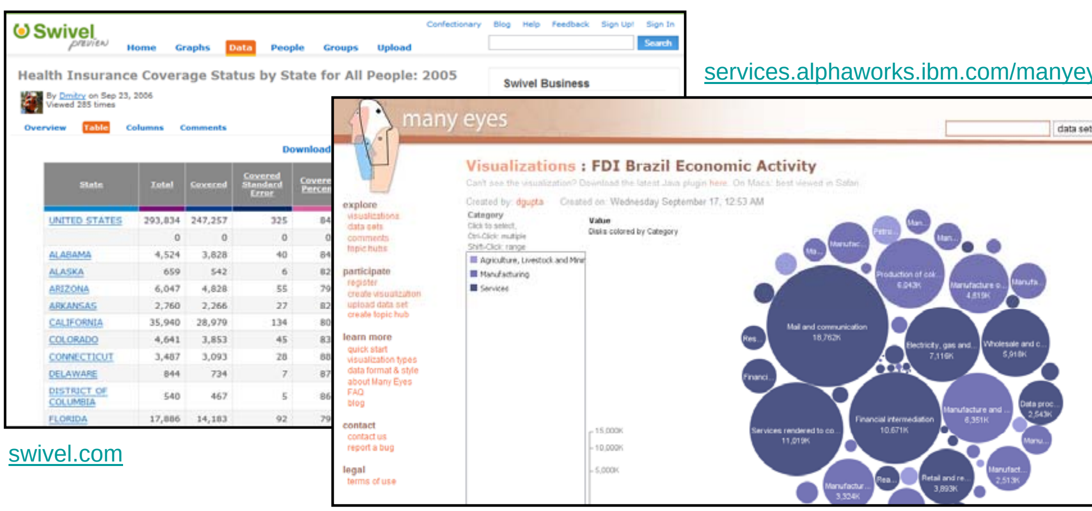
connotea.org

citeulike.org

not pictured: bibsonomy.org

These let scientists organize references, web sites, and other sources, **contributing** external memory for others to find and **remix**.

Swivel, ManyEyes,



services.alphaworks.ibm.com/manyeyes/

swivel.com

Uploaded data sets are **collaboratively remixed** into **contributed** visualizations that are **rated** by the community.

nanoHUB, SciVee, ToBIG, LEAD,

nanohub.org

scivee.tv

leadproject.org

sonicserver.northwestern.edu/tobig/

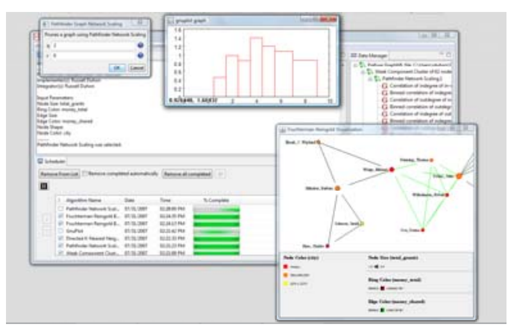
These are rich repositories of **contributed** educational content that the community **rates** and learns from.

nanoHUB is also a next-generation environment for **contributing**, running, and **remixing** nanotechnology algorithms.

On ToBIG, **contributed** content about tobacco research **collaboration** drives a recommendation system doing implicit **rating**.

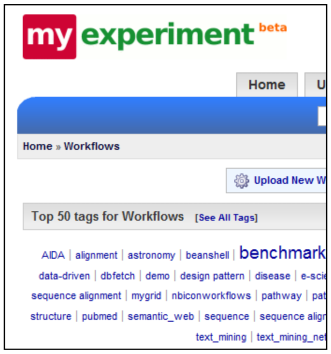
LEAD lets anyone **remix**, consume, and model weather-related data.

Network Workbench, CIShell

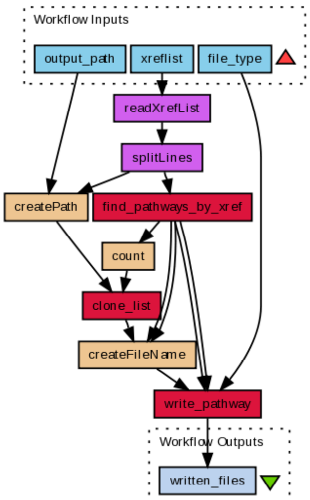


nwb.slis.indiana.edu cishell.org osgi.org

Network Workbench is built on top of CIShell, a container for **contributed** algorithms and data where better implementations rise to the top as users **remix** them into workflows and implicitly **rate** them by usage. CIShell is built on OSGi, an industry-standard services architecture.



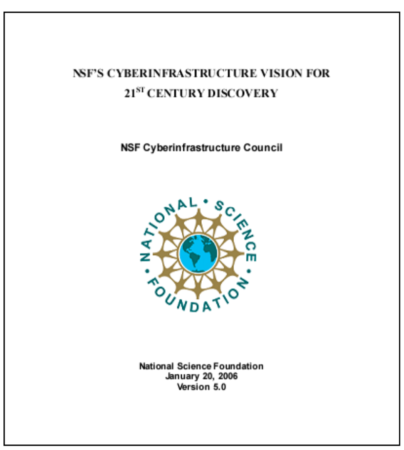
myExperiment



myexperiment.org

myExperiment takes things further, providing a social networking platform where users can **contribute**, **rate**, and **share** executable workflows that frequently **remix** cyberinfrastructure services around the web.

Thinking About Science Cyberinfrastructure


NSF'S CYBERINFRASTRUCTURE VISION FOR 21ST CENTURY DISCOVERY

NSF Cyberinfrastructure Council

National Science Foundation
January 20, 2006
Version 5.0

Scientists, policy makers, and the public need to think about science differently in order to do, enable, and understand it better.

"Fewer and fewer researchers working at the frontiers of knowledge can carry out their work without cyberinfrastructure of one form or another."
- Cyberinfrastructure Vision for 21st Century Discovery, a report by the US National Science Foundation.

Scientists: New Possibilities

Science can be done “just in time” more of the time, as data arrives and with the latest algorithms.

The selection of data and algorithms is increasingly important as algorithms and data become larger and more complex.

Collaboration and feedback involve more people in more places.

Scientists: New Challenges

Repeatability and complete records are harder and harder to define as services evolve rapidly.

Keeping track of increasing opportunities, algorithms, and data is increasingly difficult.

In particular, volumes of data increase faster than many approaches can scale to.

While there are more collaborators, connections with them are often weaker.

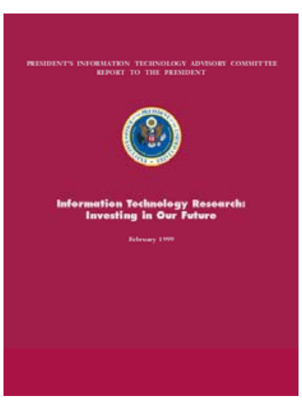
Policy Makers: New Possibilities

Science is increasingly available.

Modern tools drive increased discovery.

Creating cyberinfrastructures that function as strong marketplaces can drive the development of expertise and new knowledge.

Policy Makers: New Challenges



Policy must move faster to deal with rapid changes in science.

Risks are more concentrated in heavily-used cyberinfrastructure services.

Reducing barriers to interaction will be necessary to avoid delayed development.

"We have become dangerously dependent on large software systems whose behavior is not well understood and which often fail in unpredicted ways."
– the President's Information Technology Advisory Committee's report to the President in February 1999

The Public: New Possibilities

Educational materials help understand new discoveries.

New discoveries improve standards of living and create new possibilities for human experience.

It is much easier to contribute directly to science.

The Public: New Challenges

Engaging with science is more complex.

Navigating educational priorities is difficult.

Privacy is harder to maintain.

The Future in the Cloud

There is one more major development that is changing how the web works and will change how science works: the Cloud.

The Cloud consists of abstract, homogeneous platforms that provide easier, more scalable, cheaper, greener access to processing power, computing services, and utility applications.

image courtesy ascendtraining.com

The Cloud contains applications and centralized services as well as distributed identity and authorization services that increase the portability of personas.

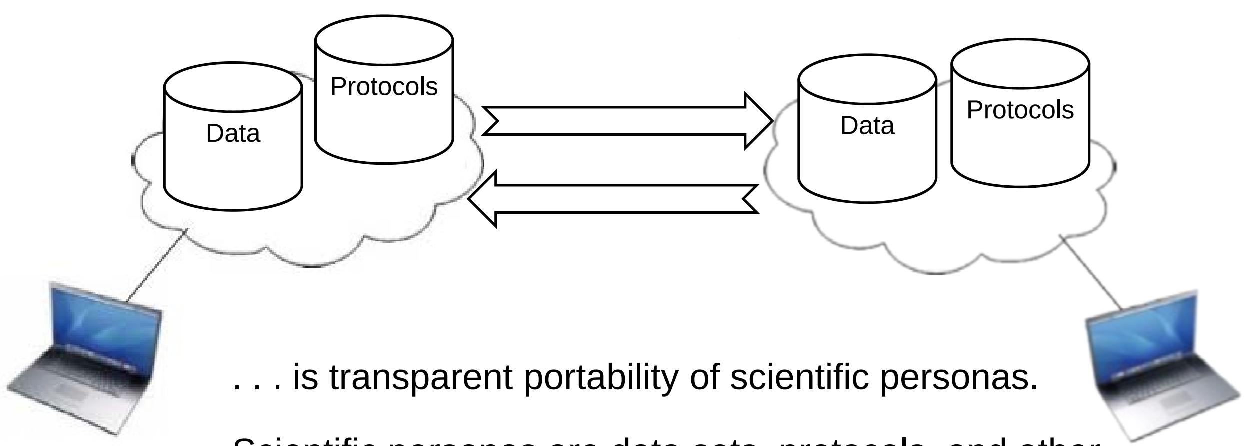
What Does It Mean to do Science in the Cloud?

Some parts are easy to integrate. Google Docs is like other word processors, for instance.

Centralized services like EC2 can do standard computational tasks in standard ways.

However, they can also be used to do science in new ways with more collaboration that scale on demand.

The Most Challenging Frontier



. . . is transparent portability of scientific personas.

Scientific personas are data sets, protocols, and other abstract but very coherent entities that consume, produce, and grow in the ecosystems of science.

Having such portability will make cyberinfrastructure marketplaces more successful.

In the Future with the Cloud

Continuously available and changing streams of data will be analyzed immediately by scientists using algorithms that have bubbled to the top in interdisciplinary use.

Novel analyses will be quickly spotted and integrated into the work of other scientists directly and dynamically, as the community discovers them.

Conclusions

Modern science cyberinfrastructures are marketplaces for humans to interact with contribution, rating, remixing and collaboration.

While this new ecology is very promising, it holds many challenges related to increased flexibility, complexity, and interdependency.

The possibilities of the Cloud may help deal with these issues by providing resources and resource portability for science on scales previously impossible.

Thank You

Russell Duhon
rduhon@indiana.edu

Special thanks to the symposium organizers, and the following grants:

TSEEN: NSF IIS-0534909

Network Workbench: NSF IIS-0513650 award