Science Gateways and their Democratization and Acceleration of Science

Nancy Wilkins-Diehr

wilkinsn@sdsc.edu

HUB BUB 2012
the HUBzero conference

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Science Gateways

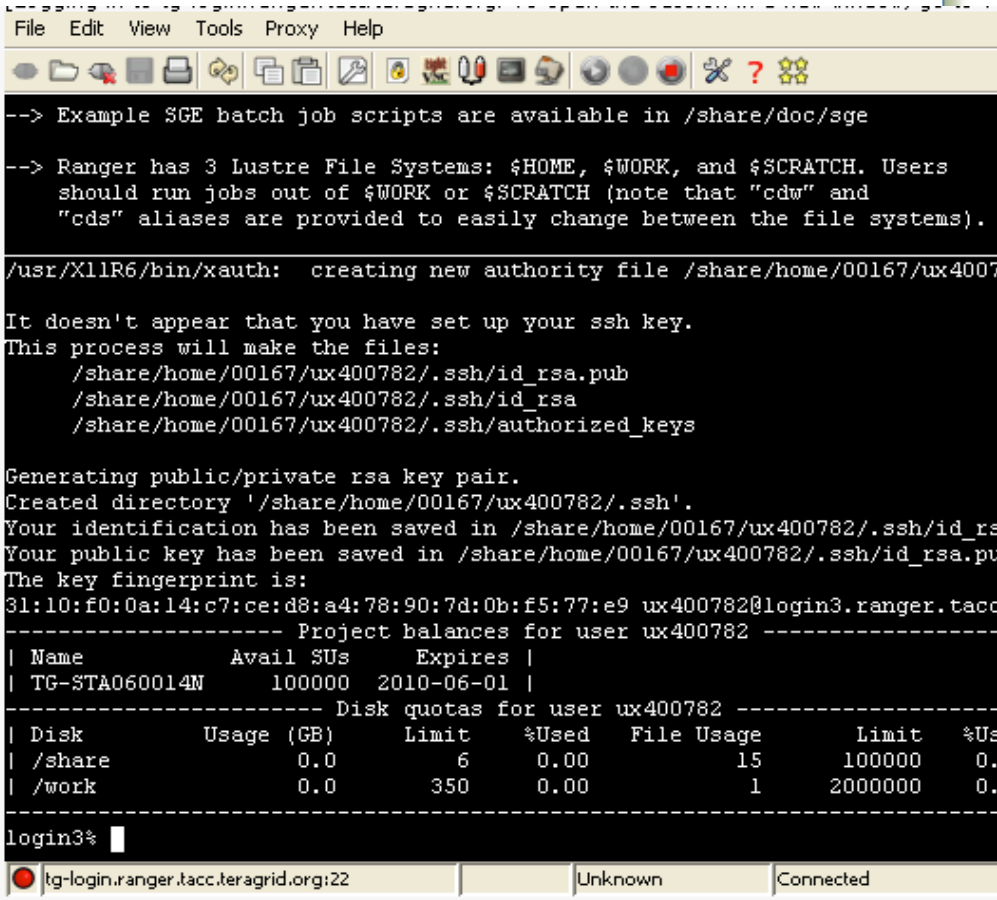
Gateways:
A natural result of the impact of the internet on worldwide communication and information retrieval

Only 20 years since the release of Mosaic!

- Implications on the conduct of science are still evolving
  - 1980's, Early gateways, National Center for Biotechnology Information BLAST server, search results sent by email, still a working portal today
  - 1989 World Wide Web developed at CERN
  - 1992 Mosaic web browser developed
  - 1995 "International Protein Data Bank Enhanced by Computer Browser"
  - 2004 TeraGrid project director Rick Stevens recognized growth in scientific portal development and proposed the Science Gateway Program
  - Today, Web 3.0 and programmatic exchange of data between web pages
- Simultaneous explosion of digital information
  - Growing analysis needs in many, many scientific areas
  - Sensors, telescopes, satellites, digital images, video, genome sequencers
  - #1 machine on Top500 today over 10,000x more powerful than *all combined* entries on the first list in 1993

My experiences
Science Gateways and high performance computing

- Interfaces to supercomputers haven't changed much in the last 20 years
- vt100 in the 1980s and a login window on the Ranger system at TACC today



HUB BUB 2012
the HUBzero conference

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Science Gateways

Why are gateways worth the effort?

- Increasing range of expertise needed to tackle the most challenging scientific problems
  - How many details do you want each individual scientist to need to know?
    - PBS, RSL, Condor
    - Coupling multi-scale codes
    - Assembling data from multiple sources
    - Collaboration frameworks

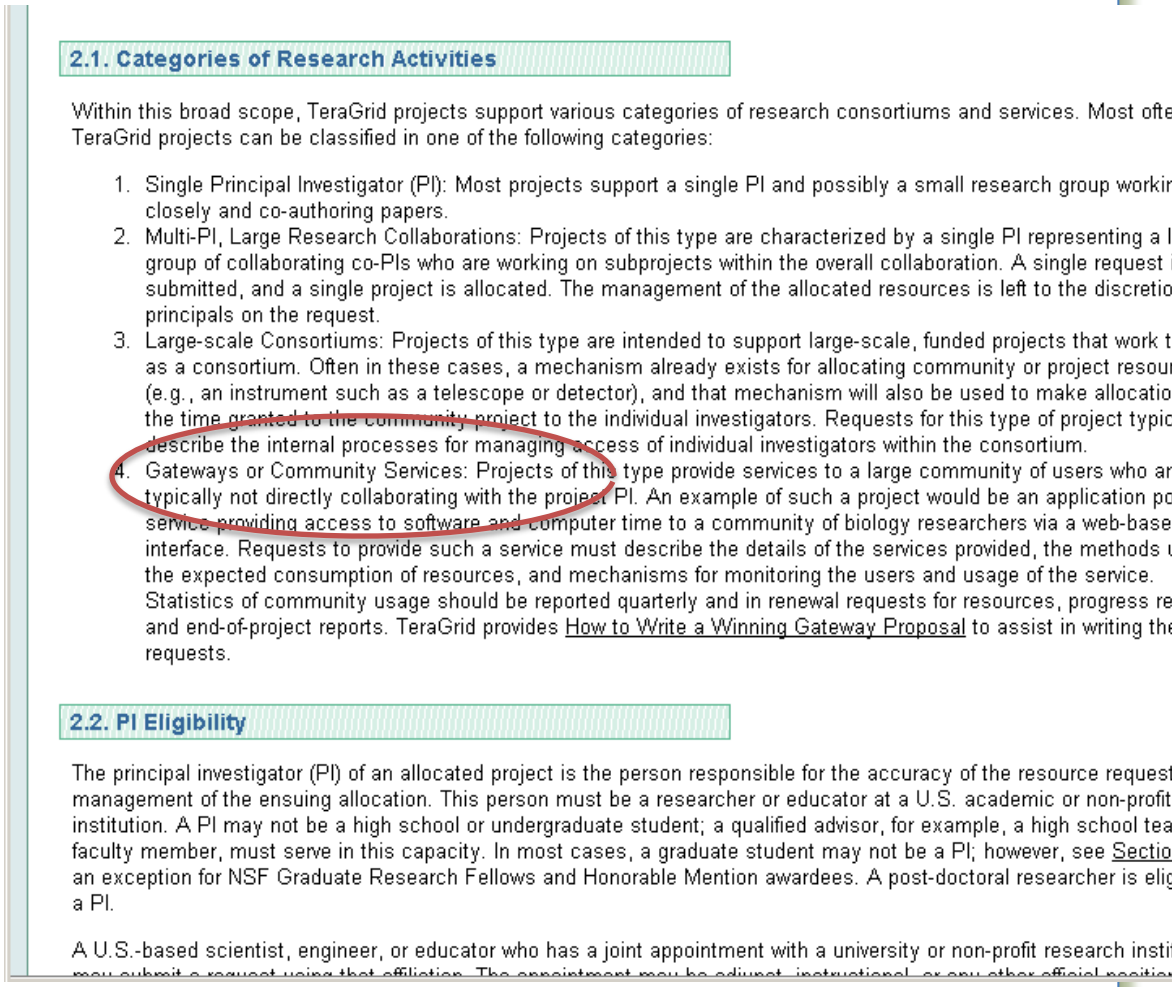
```
#! /bin/sh
#PBS -q dc
#PBS -l no
#PBS -l wa
#PBS -o pb
#PBS -e pb
#PBS -V
cd /users/w
../bin/mcell
```

```
+(
        &(re
login1.sd
pbs")

(executa
mcell")
        (arg
        (cou
  (hostCo
        (ma

(director
se_3")

(stdout="
```

```
=======
# Full path to executable
executable=/users/wilkinsn/tutor
ial/bin/mcell

# Working directory, where
Condor-G will write
# its output and error files on
the local machine.
initialdir=/users/wilkinsn/tutorial/
exercise_3

# To set the working directory of
the remote job, we
# specify it in this globus RSL,
which will be appended
# to the RSL that Condor-G
generates
globusrsl=(directory='/users/wilk
insn/tutorial/exercise_3')

# Arguments to pass to
```

Did you know?

- Supercomputers can be accessed through user-designed web interfaces as well as from the command line
- In 2011 40% of TeraGrid users came through the web



2.1. Categories of Research Activities

Within this broad scope, TeraGrid projects support various categories of research consortiums and services. Most ofte
TeraGrid projects can be classified in one of the following categories:

1. Single Principal Investigator (PI): Most projects support a single PI and possibly a small research group workin closely and co-authoring papers.
2. Multi-PI, Large Research Collaborations: Projects of this type are characterized by a single PI representing a l group of collaborating co-PIs who are working on subprojects within the overall collaboration. A single request submitted, and a single project is allocated. The management of the allocated resources is left to the discretio principals on the request.
3. Large-scale Consortiums: Projects of this type are intended to support large-scale, funded projects that work t as a consortium. Often in these cases, a mechanism already exists for allocating community or project resou (e.g., an instrument such as a telescope or detector), and that mechanism will also be used to make allocatio the time granted to the community project to the individual investigators. Requests for this type of project typic describe the internal processes for managing access of individual investigators within the consortium.
4. Gateways or Community Services: Projects of this type provide services to a large community of users who ar typically not directly collaborating with the project PI. An example of such a project would be an application po service providing access to software and computer time to a community of biology researchers via a web-base interface. Requests to provide such a service must describe the details of the services provided, the methods the expected consumption of resources, and mechanisms for monitoring the users and usage of the service. Statistics of community usage should be reported quarterly and in renewal requests for resources, progress re and end-of-project reports. TeraGrid provides How to Write a Winning Gateway Proposal to assist in writing the requests.

2.2. PI Eligibility

The principal investigator (PI) of an allocated project is the person responsible for the accuracy of the resource request management of the ensuing allocation. This person must be a researcher or educator at a U.S. academic or non-profit institution. A PI may not be a high school or undergraduate student; a qualified advisor, for example, a high school tea faculty member, must serve in this capacity. In most cases, a graduate student may not be a PI; however, see Sectio an exception for NSF Graduate Research Fellows and Honorable Mention awardees. A post-doctoral researcher is elig a PI.

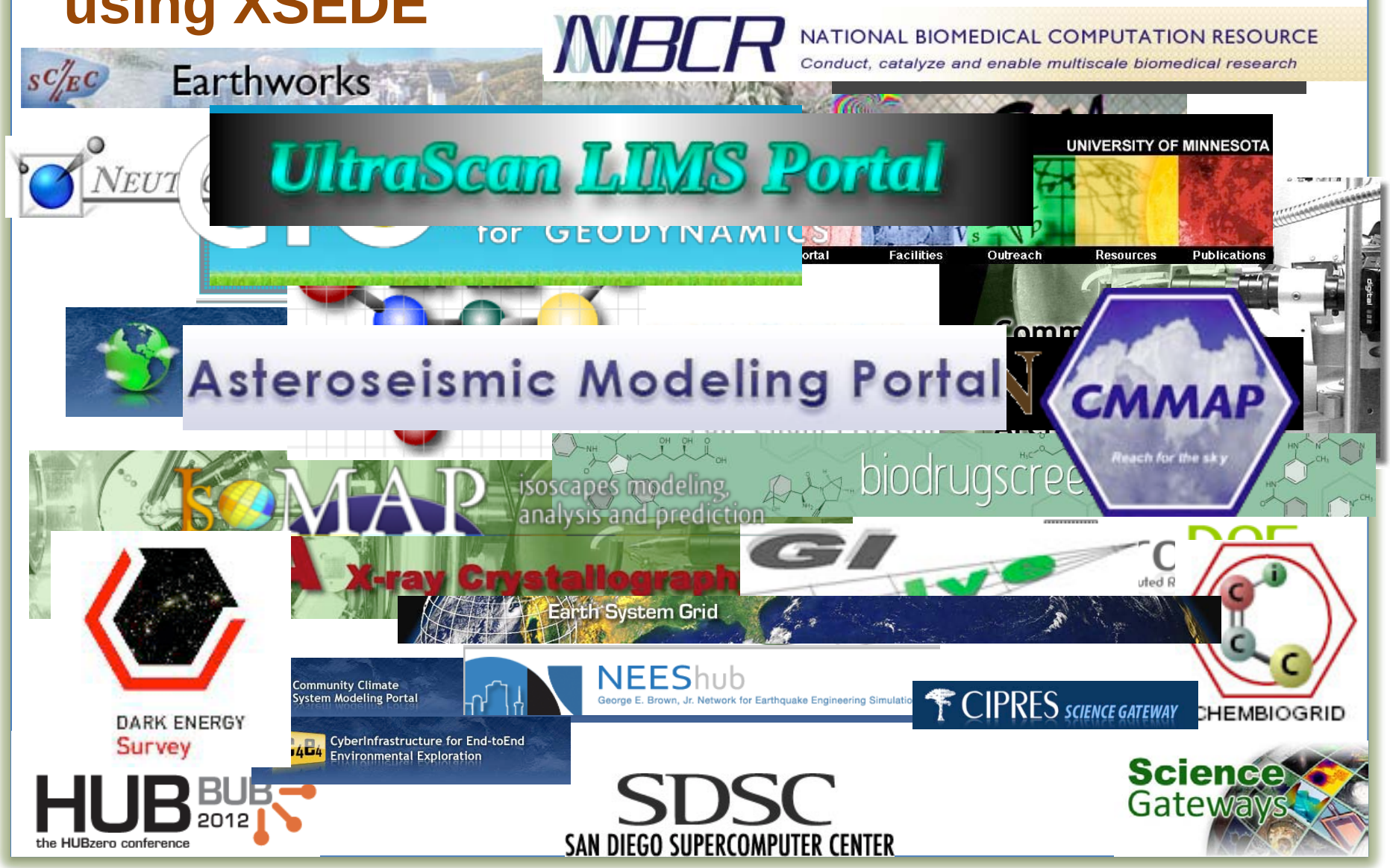
A U.S.-based scientist, engineer, or educator who has a joint appointment with a university or non-profit research insti

HUB BUB 2012
the HUBzero conference

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

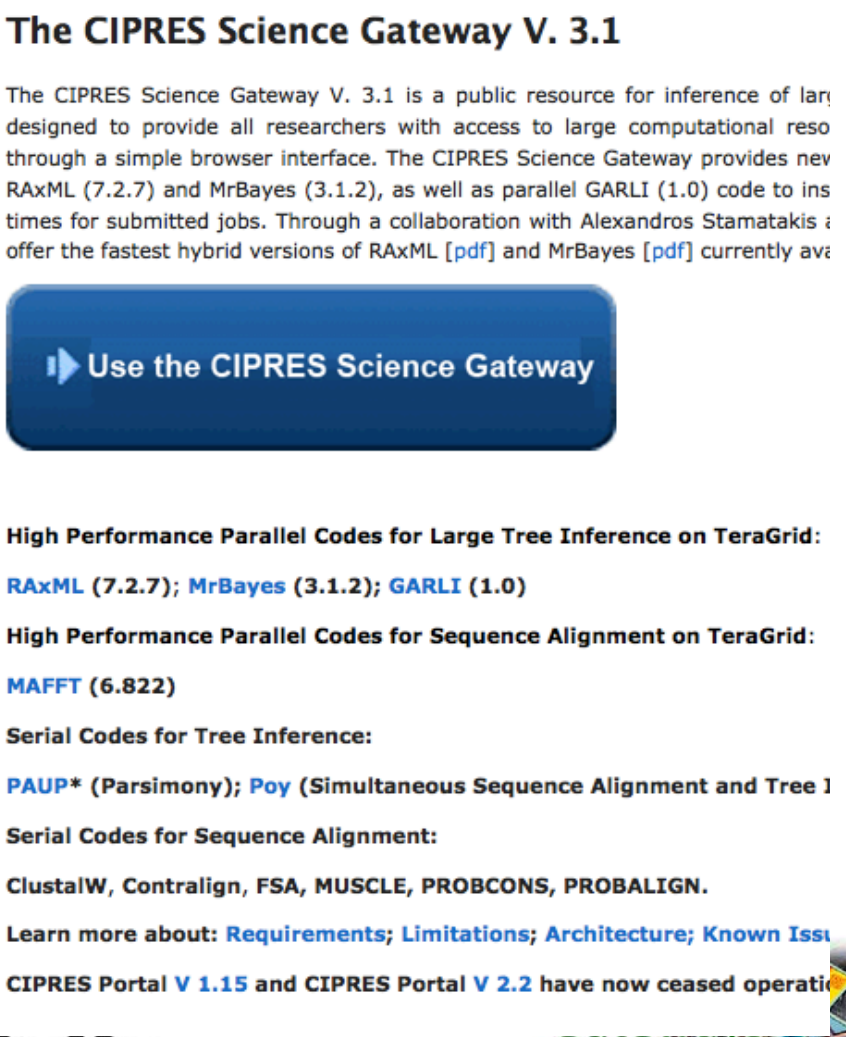
Science Gateways

Today, there are approximately 35 gateways using XSEDE



Cyberinfrastructure for Phylogenetic Research (CIPRES)

- Most popular science gateway
  - ~25% of all XSEDE users
- In use on 6 continents
- Cited in major journals (Cell, Nature, PNAS)
- Used at major research institutions (Stanford, Harvard, Yale)
- Used by 57 researchers for curriculum delivery
- Used in 80% of EPSCoR states
- Recently used by a 15-year-old high school student who won the Massachusetts state science fair with no support from gateway staff



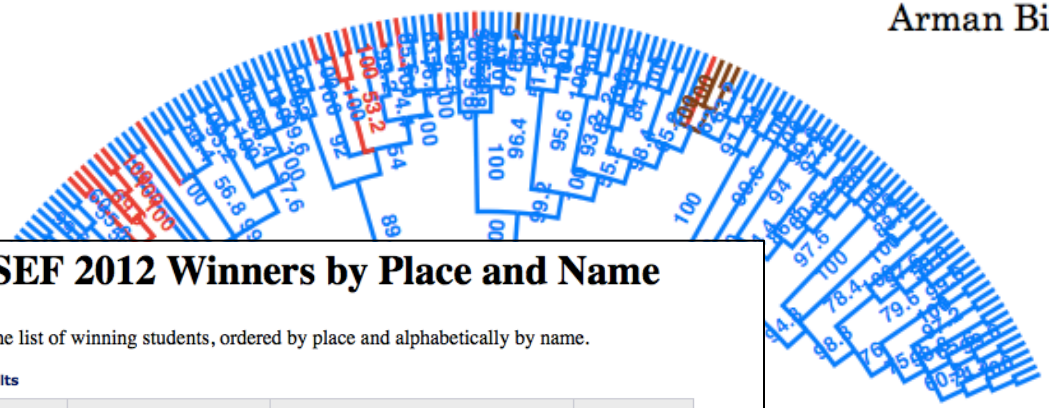
Locations of the last 1000 users of the CIPRES Science Gateway

http://www.phylo.org/tools/map.html

Arman Bilge, Lexington High School wins MA state science fair using CIPRES

The Origin & Spread of HIV-1 Subtype B in the Americas

Arman Bilge

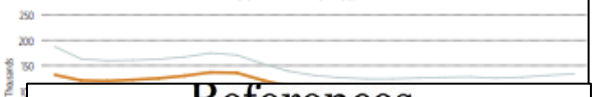


Number of New HIV Infections

North America & West Europe

Caribbean

South America



MSSEF 2012 Winners by Place and Name

This is the list of winning students, ordered by place and alphabetically by name.

160 Results

Exhibitor	School	Title	Place
Ackell, Ryan	Falmouth Academy	Effects of Nicotine and Vicodin on Memory Formation in Hermissenda	Team 1st Place
Athalye, Anish	Mass. Academy of Math & Science	Cooling without Electricity: Engineering a New Refrigerator	1st Place
Bhupatiraju, Surya	Lexington High School	Investigating the Spread of the Influenza A Virus: A Phylogenetic Anal	Team 1st Place
Bilge, Arman	Lexington High School	The Origin & Spread of HIV-1 Subtype B in the Americas	1st Place
Dodd, Oliver	Needham High School	Cancer Growth Regulators	1st Place
Fleron, Addie	Westfield High School	The Adverse Effects of Consumer and Pharmaceutical Goods on Plant Life	Team 1st Place

References

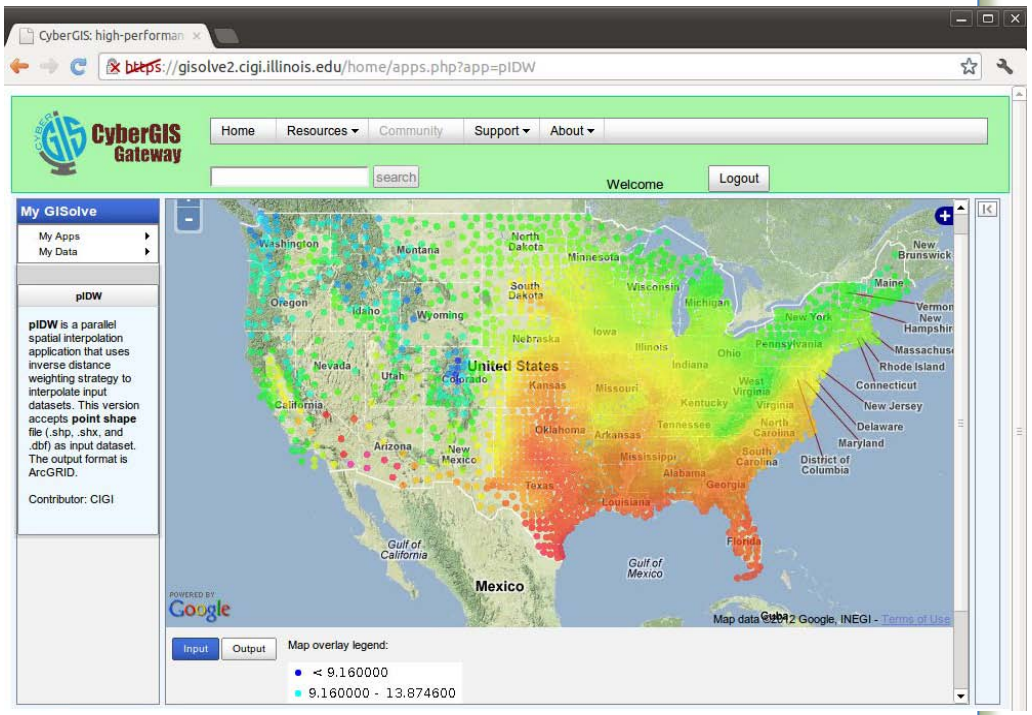
Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gateway Computing Environments Workshop*, 1–8. Retrieved from http://www.phylo.org/

CIPRES on NSF's front page today!

CyberGIS

Software Integration for Sustained Geospatial Innovation

- Application of high-end cyberinfrastructure to GIS
- Influence on multiple domains
- Improved decision support
- Spatial joins, layers of multiple datasets at different resolutions
- Goal is core set of composable, interoperable, manageable, and reusable software elements
- Collaborative geospatial problem solving environment



HUB BUB 2012
the HUBzero conference

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Science Gateways

Analytical Ultracentrifugation: Emerging computational tool for the study of proteins

- Samples from researchers all over the world
  - Some (Germany, Australia) have their own ultracentrifuges and use only the analysis capabilities, others send samples to UT to spin
- Spin the samples at high speeds, learn about macromolecule properties
- Monte Carlo simulations
- Observations are electronically digitized and stored for further mathematical analysis

UltraScan 9.0

The Center for Analytical Ultracentrifugation of Macromolecular Assemblies, UT Health Sciences, Borries Demeler, PI

Source: Suresh Marru, IU

Ultrascan provides a comprehensive data analysis environment

- Management of analytical ultracentrifugation data for single users or entire facilities
- Support for storage, editing, sharing and analysis of data
  - HPC facilities used for 2-D spectrum analysis and genetic algorithm analysis
    - XSEDE
    - Technische University of Munich
    - Juelich Supercomputing Center
- Portable graphical user interface
- MySQL database backend for data management
- Over 30 active institutions

Source: Suresh Marru, IU

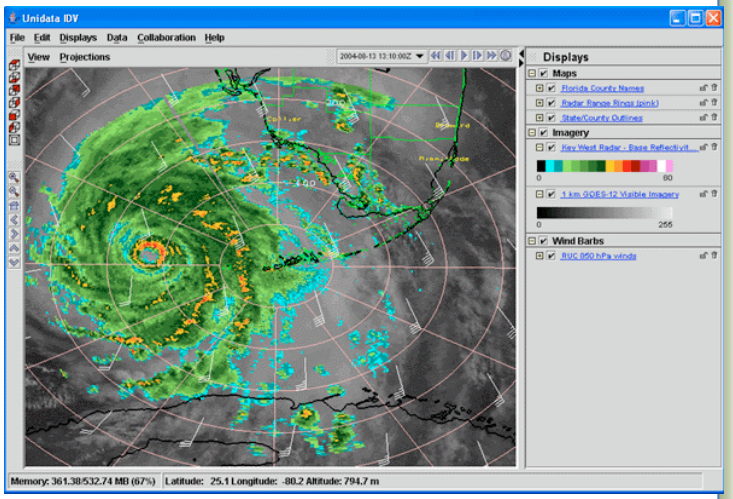
Gateways in the marketplace

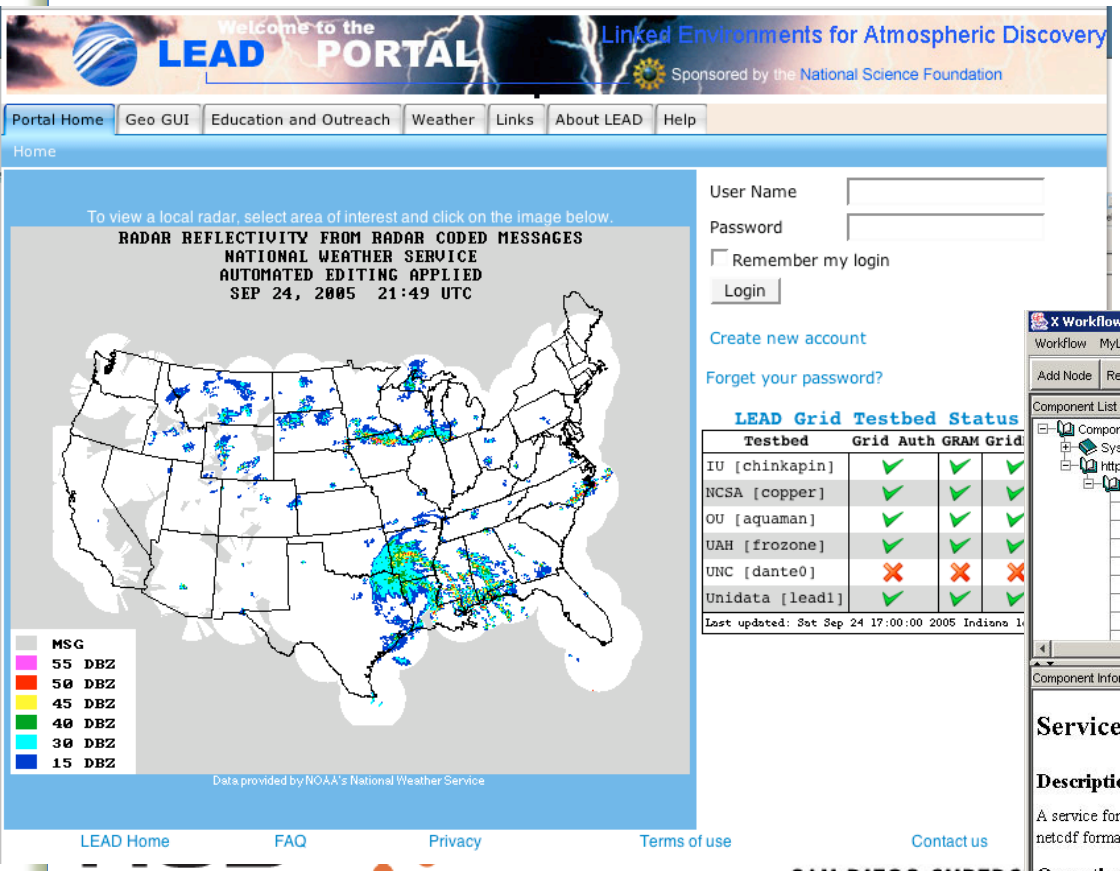
Kids control telescopes and share images

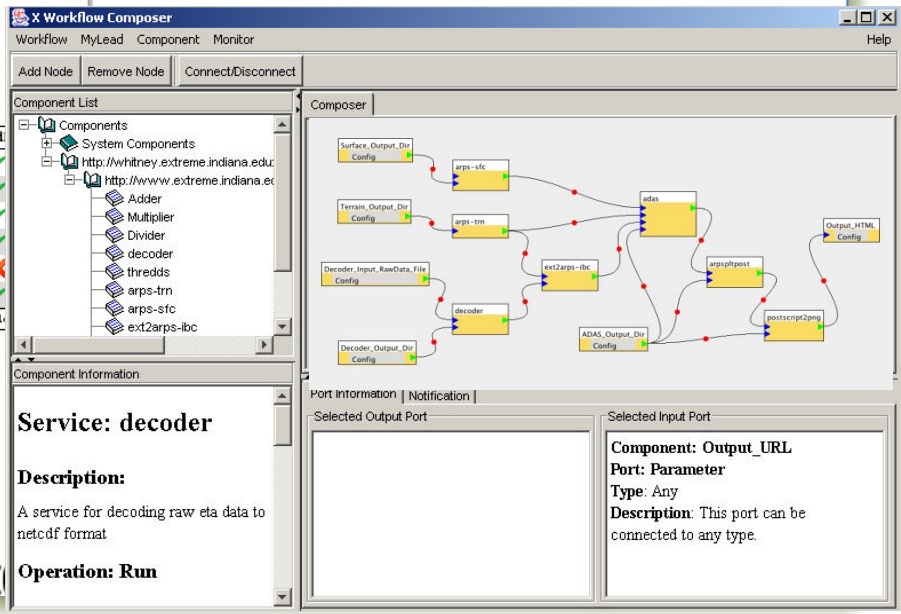
- "In seconds my computer screen was transformed into a live telescopic view"
  - "Slooh's users include newbies and professional astronomers in 70 countries"
- Observatories in the Canary Islands and Chile, Australia coming soon
- 5000 images/month since 2003
- Increases public support for investment in these facilities

Linked Environments for Atmospheric Discovery
LEAD

- Providing tools that are needed to make accurate predictions of tornados and hurricanes
- Data exploration and Grid workflow

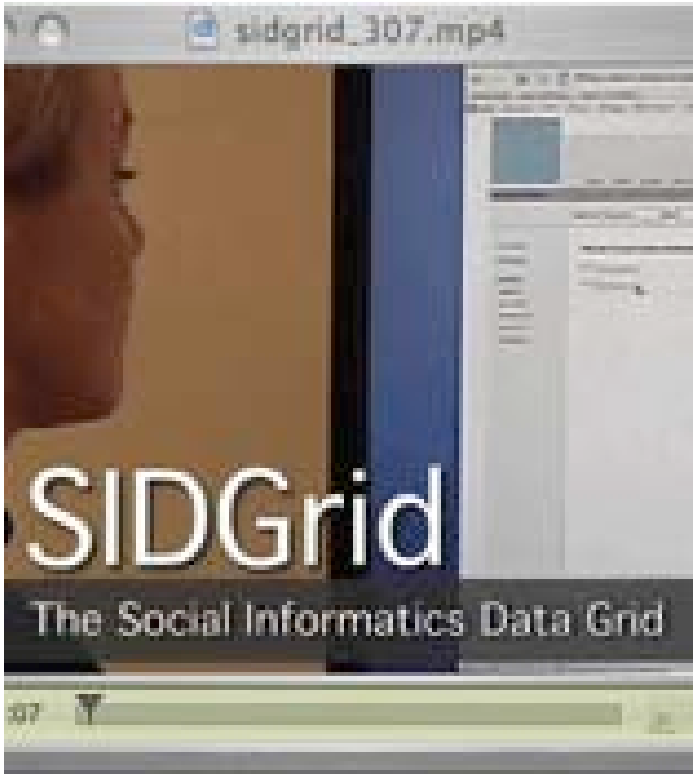






Social Informatics Data Grid
Collaborative access to large, complex datasets

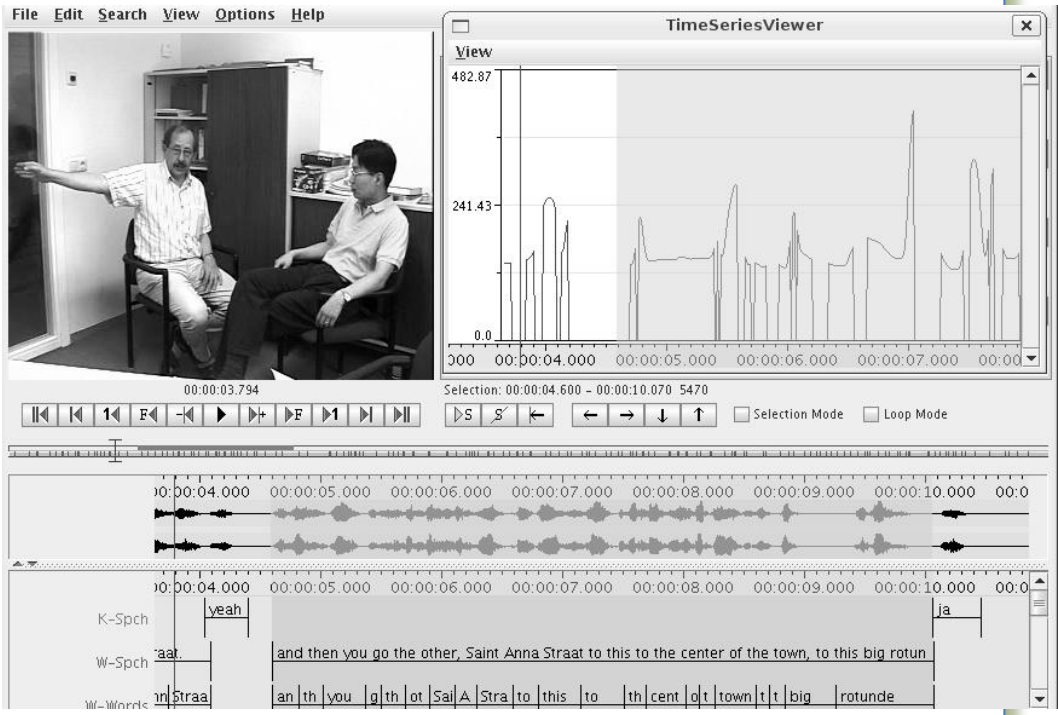
- SIDGrid is unique among social science data archive projects
  - Streaming data which change over time
    - Voice, video, images (e.g. fMRI), text, numerical (e.g. heart rate, eye movement)
  - Investigate multiple datasets, collected at different time scales, simultaneously
    - Large data requirements
    - Sophisticated analysis tools
- Especially valuable for multi-disciplinary course
  - Computer science, linguistics, psychology



http://www.ci.uchicago.edu/research/files/sidgrid.mov

Viewing multimodal data like a symphony conductor

- "Music-score" display and synchronized playback of video and audio files
  - Pitch tracks
  - Text
  - Head nods, pause, gesture references
- Central archive of multi-modal data, annotations, and analyses
  - Distributed annotation efforts by multiple researchers working on a common data set
    - History of updates
- Computational tools
  - Distributed acoustic analysis using Praat
  - Statistical analysis using R
  - Matrix computations using Matlab and Octave



Source: Studying Discourse and Dialog with SIDGrid, Levow, 2008

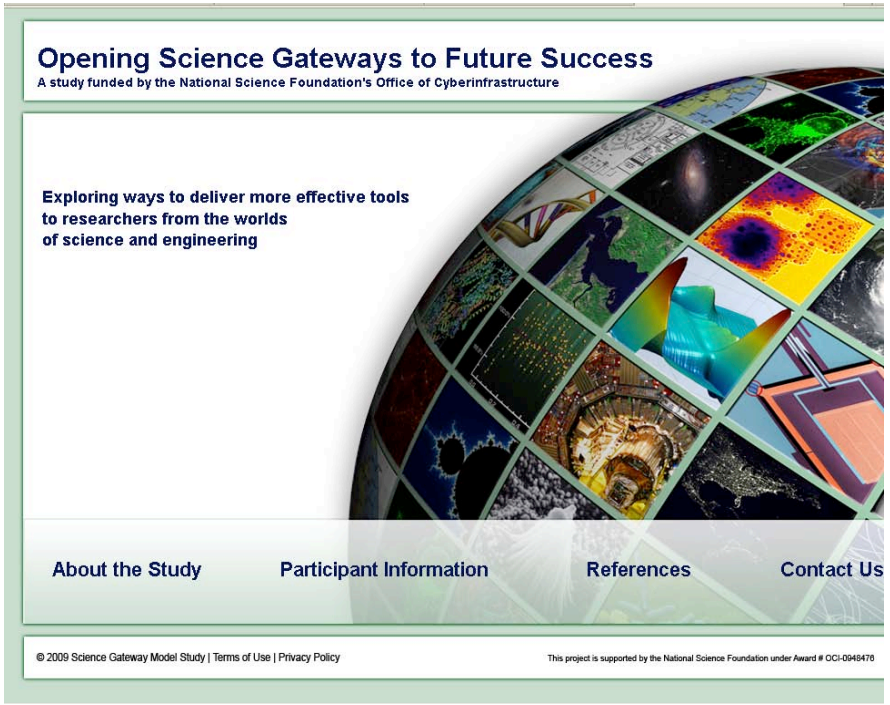
Gateways democratize access to high end resources

- Almost anyone can investigate scientific questions using high end resources
  - Not just those in high profile research groups
- Gateways allow anyone with a web browser to explore
- Foster new ideas, cross-disciplinary approaches
  - Encourage students to experiment
- But used in production, too
  - Significant number of papers resulting from gateways, including GridChem, nanoHUB
  - Scientists can focus on challenging science problems rather than challenging infrastructure problems

Gateway Sustainability Study

With Dr. Katherine Lawrence, U Michigan School of Information

- Characteristics of short funding cycles
  - Build exciting prototypes with input from scientists
  - Work with early adopters to extend capabilities
  - Tools are publicized, more scientists interested
  - Funding ends
  - Scientists who invested their time to use new tools are disillusioned
    - Less likely to try something new again
  - Start again on new short-term project
- Need to break this cycle
- Small grant to look at characteristics of successful gateways and domain areas where a gateway could have a big impact



4 focus group meetings over 2 years

First 2 held June, 2010

Final 3 held June, 2011

www.sciencegateways.org

After reading stacks of reports, we wanted to capture our findings in a memorable way

These findings lead nicely into NSF's Scientific Software Innovation Institutes (S2I2) program

- Cyberinfrastructure Framework for the 21st Century Science and Engineering (CIF21) vision
- Software as integral enabler of computation, experiment and theory
- NSF expects that software investment will result in robust, reliable, usable, sustainable software infrastructure
- **S2I2 institutes are viewed as long-term hubs of excellence that serve communities of substantial size and disciplinary breadth**

Science Gateway Institute

- One of several 1-year conceptualization grants awarded this past summer
- Collaboration involves
  - SDSC, Elizabeth City State, Indiana, Purdue, U Michigan, U Texas
- Gathering community feedback on Institute offerings this year
  - Incubator service
  - Extended support services
  - Gateway forum
  - Gateway framework
  - Workforce development
  - Community engagement

Incubator Service

Assist with the entire lifecycle of a gateway:

- Business plan development and review
- Development environment, consulting, documentation and software recommendations
- Software repositories
- Software engineering facilities
- Software assessment services
  - like Open Source Software Advisory Service, Apache assessment service, Software Sustainability Institute (UK)
- Build-and-test facilities
- Hosting service
- Offering gateways expertise in the following areas:
  - Usability assessment
  - Licensing
  - Sustainability
  - Project management
  - Security

Extended Support

- Peer-reviewed request process open to all
- Institute staff assigned to a project for months, up to a year
  - Assist with gateway development
  - Teach research teams what it takes to build, enhance, operate, and maintain gateways after support ends

Gateway Forum

- Gathering place for scientific web developers across NSF directorates, agencies, and international boundaries
- Social forums, white papers, blogs, testimonials and user stories
- Annual conference
- Broad and engaging symposium series
- Gateway training program
  - Synchronous and asynchronous, video tutorials
  - Best practices, case studies
- Showcase of successful projects
- Environment that enables continuous community feedback

Gateway Framework

- Modular, layered approach
  - Supports community contributions
  - Grocery store approach allows developers to pick and choose the components they need
- 4 levels of architecture
  - Value-added services such as a publication channel for delivering content to a wider audience, information repositories for good design practices, and information/code samples for best practices in user-interface and user-experience design
  - Core web framework which includes hosted site creation and content management
  - Platform API to provide a cohesive set of RESTful web services upon which the previous two layers rely
  - Systems layer where the hardware and low-level middleware reside
    - Clouds and cloud services, HPC systems, grid middleware, data warehouses, databases, instrumentation, and distributed data stores

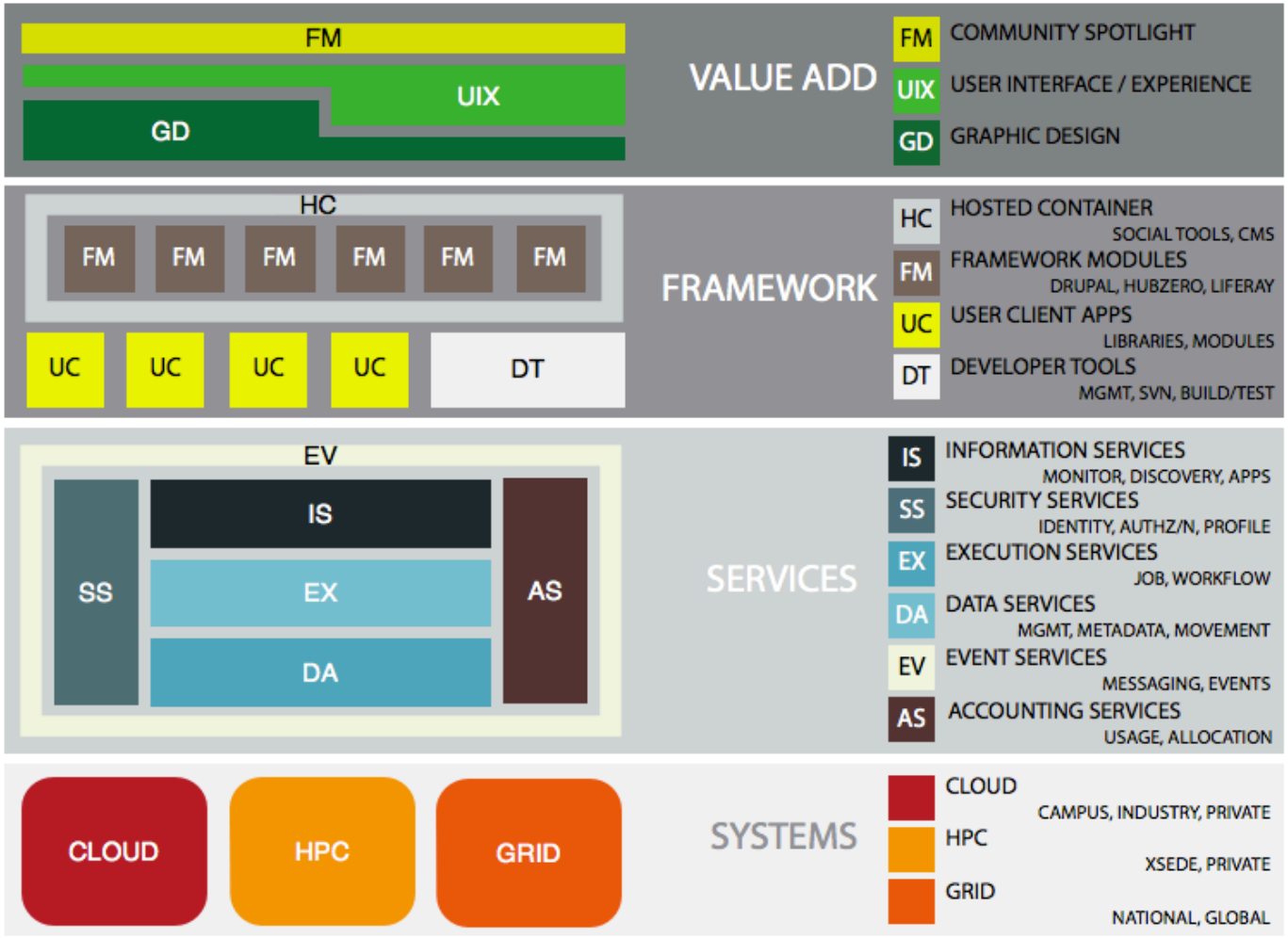


Figure 1. High-level architecture of software offerings and value-added services provided by the institute.

First community input session this afternoon!

- Lightning talks highlighting gateways from 1:30-4, 5-5:30
- Live polling and interactive session on what would be of value in an institute from 4-5
- Hope to see some of you this afternoon

- **Thank you for your attention**