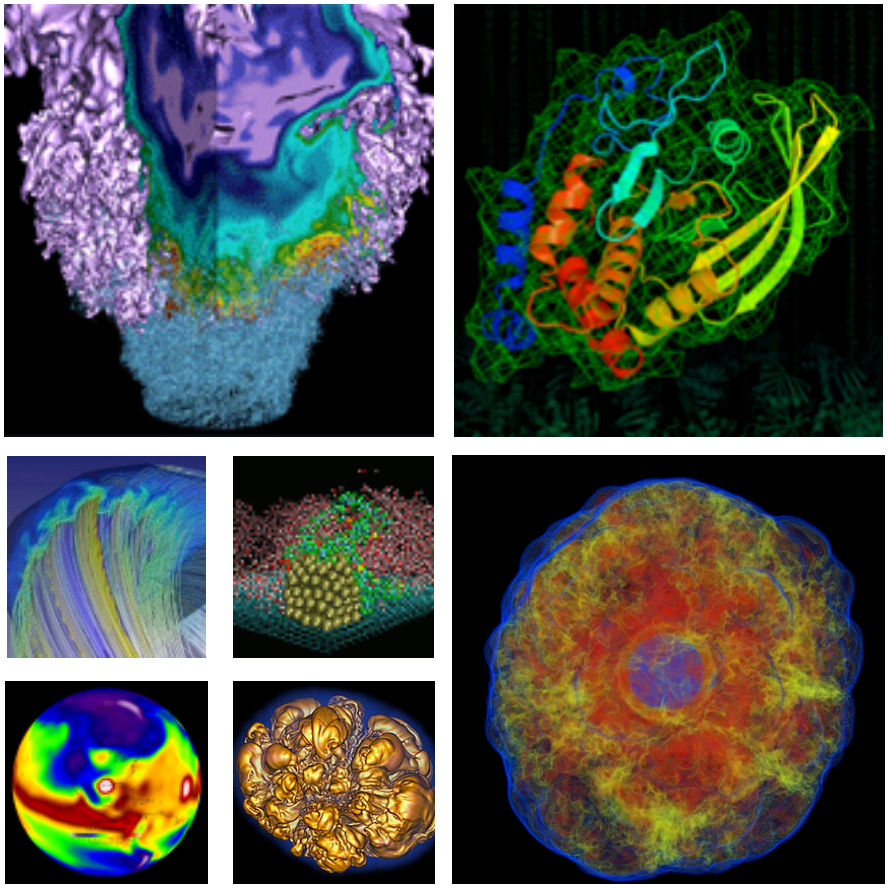
Design Patterns in Web Gateways for Scientific Data



Shreyas Cholia

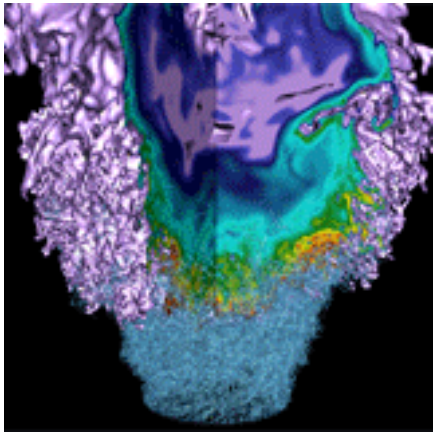
Data and Analytics Services, NERSC

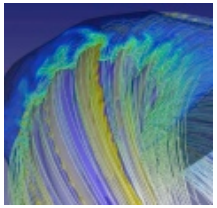
Joint Facilities User Forum on Data-Intensive Computing, Oakland, June 17, 2014

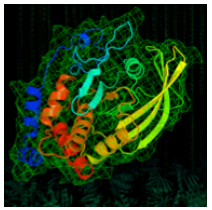
U.S. DEPARTMENT OF ENERGY | Office of Science

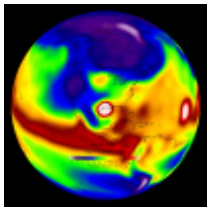
BERKELEY LAB

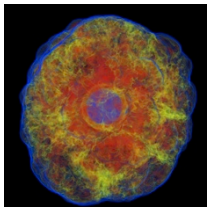
Introduction

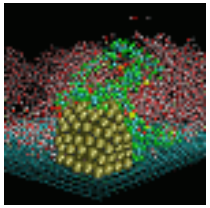












Overview

- **Motivation**
- **Patterns**
  - Web Data Sharing
  - End-to-End Frameworks
  - API-driven access
  - Web IDEs
  - Federated Identity
  - Social Data
- **Conclusions**

Motivation for Science Gateways

- **People expect web interfaces and applications for usability**

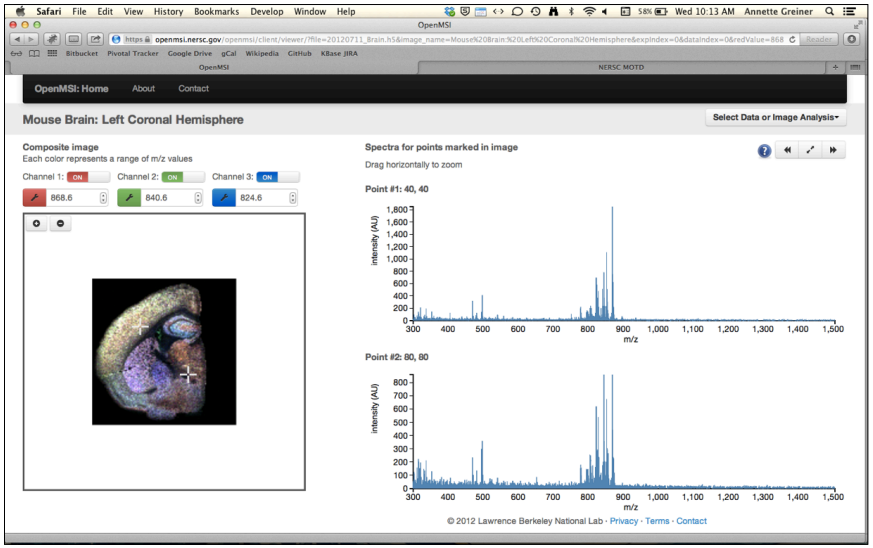
```
ssh carver.nersc.gov
qsub -I -q matgen_reg -l nodes=2:ppn=16 -l walltime=00:30:00
cd /project/projectdirs/matgen/vasp_test
mpirun -n 32 vasp
```

X don't want generic options/tools not applicable to your science

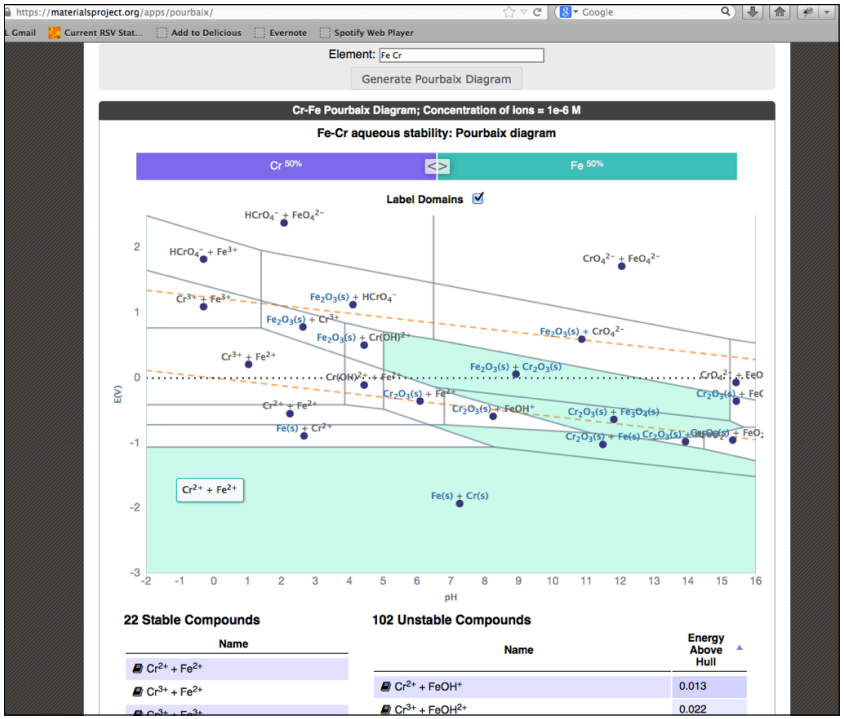
X don't want to deal with backend, middleware, UNIX CLI etc.

Harnessing The Power Of The Web

- **Web interfaces enable science-centric views to data**
- **Enable new discoveries and insights through collaborative tools and rich interfaces**



https://openmsi.nersc.gov

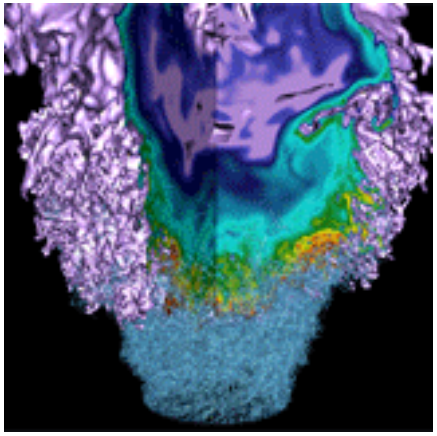


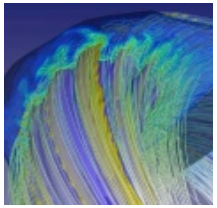
https://materialsproject.org

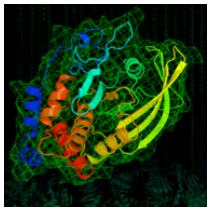
Science is Collaborative

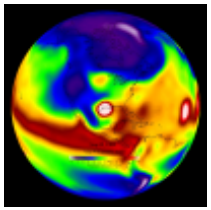
- Science is now a collaborative effort
- Large teams of people
- Requires shared access to compute and data resources

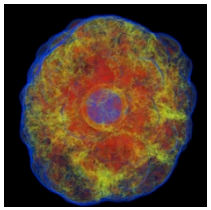
Patterns in Science Gateways

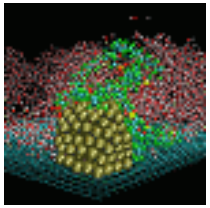






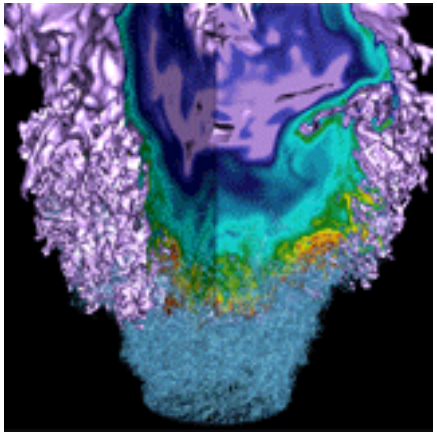


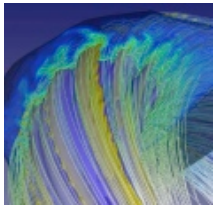


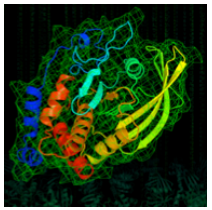


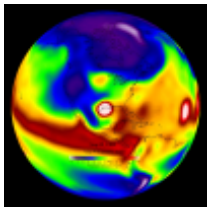
Attempt to catalog the interesting features of science gateways that span science domains in order to gain some insight on how to best expose scientific data over the web.

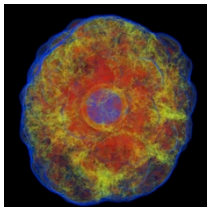
Data Sharing Over The Web

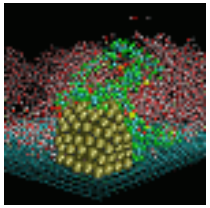








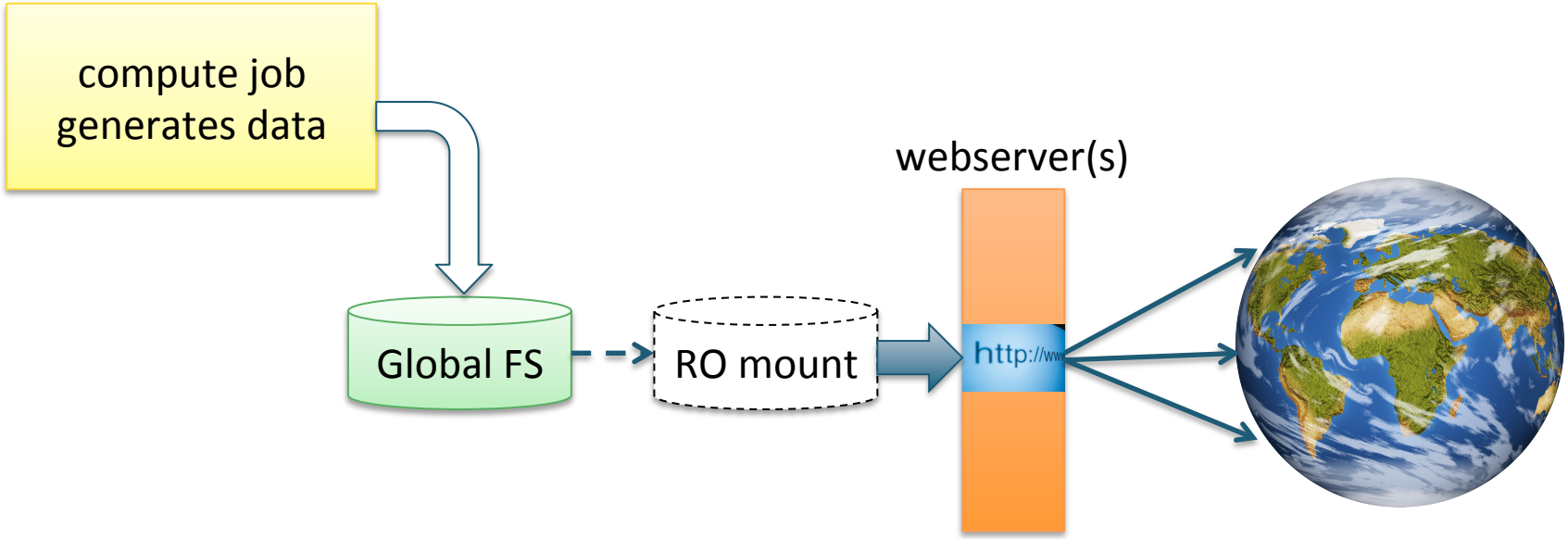




Data Sharing Over The Web

- OSTP requirements around data management
- What is the basic infrastructure required to make this happen?

Data Sharing Over The Web

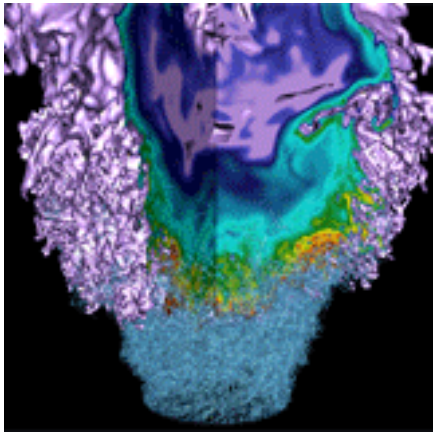


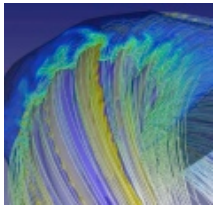
- **Store data on persistent global filesystem**
- **Designate an area for export**
- **Mount export area RO on a web node using a shared FS**
- **Run a webserver on web node to share with world**
- **Add AuthNZ to web server configuration as needed**

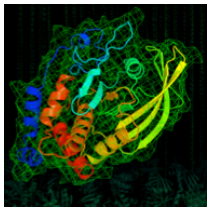
Scaling and Growth

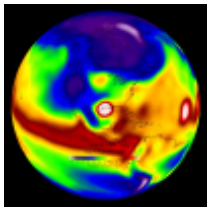
- **Scale out – add more nodes + loadbalancer**
- **Use sophisticated interfaces to search/browse data**
  - HTML+JS, PHP, Web Frameworks
- **High performance data transfer tools**
  - Globus
  - Xrootd + webdav
  - Custom downloaders

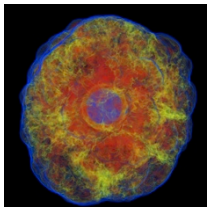
End-to-End Frameworks

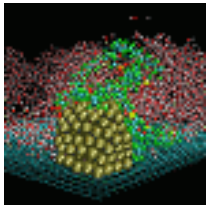








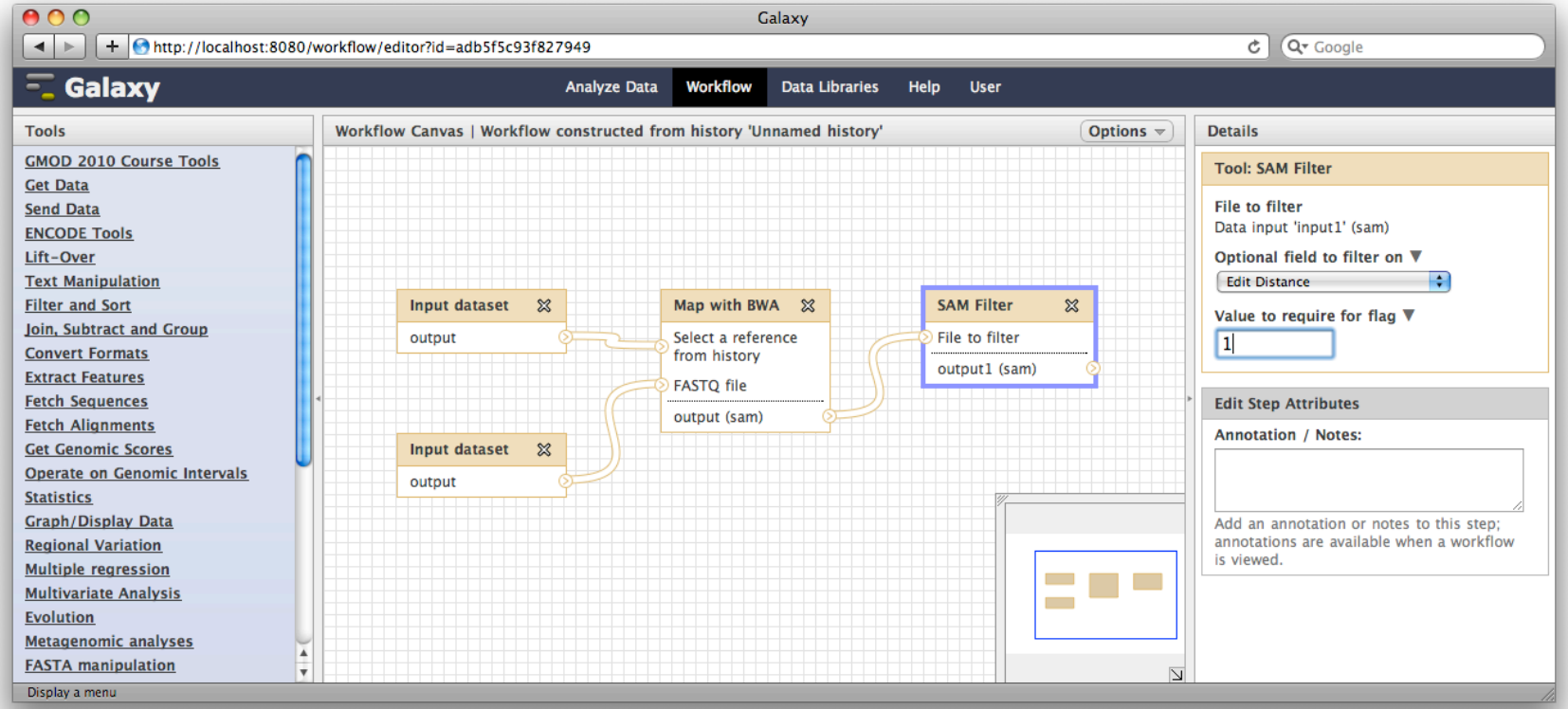




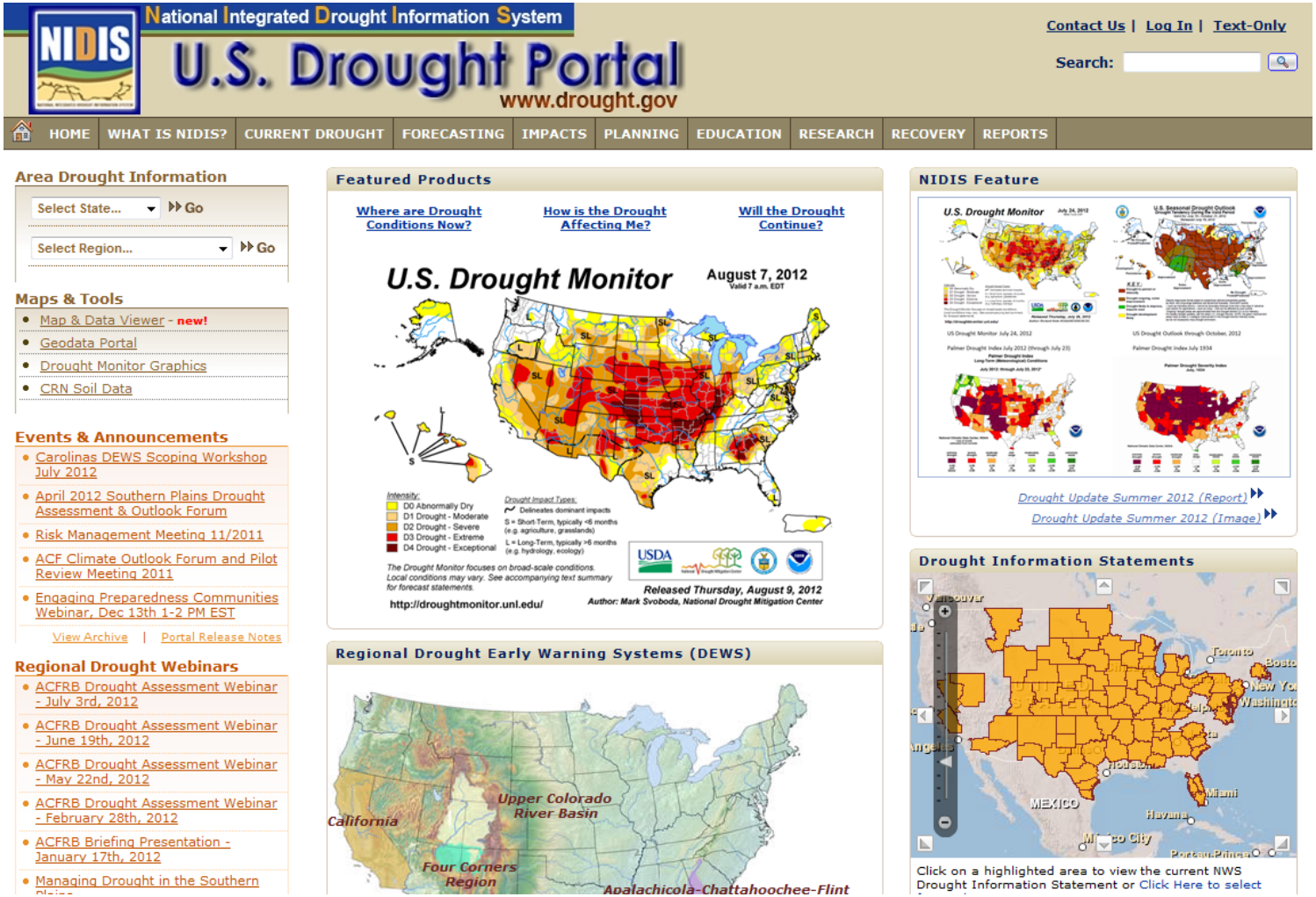
End-to-End Frameworks

- **Open source frameworks that can be customized for specific science use cases:**
  - HubZero
  - Galaxy
  - OGCE
- **Takes a plugin + configuration based approach to domain specific customizations.**
- **Very useful for managing canned workflows.**
- **Community and management tools often baked-in**

Galaxy


Galaxy
http://localhost:8080/workflow/editor?id=adb5f5c93f827949
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
GMOD 2010 Course Tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
FASTA manipulation
Workflow Canvas | Workflow constructed from history 'Unnamed history'
Options
Input dataset
output
Input dataset
output
Map with BWA
Select a reference from history
FASTQ file
output (sam)
SAM Filter
File to filter
output1 (sam)
Details
Tool: SAM Filter
File to filter
Data input 'input1' (sam)
Optional field to filter on
Edit Distance
Value to require for flag
1
Edit Step Attributes
Annotation / Notes:
Add an annotation or notes to this step; annotations are available when a workflow is viewed.
Display a menu

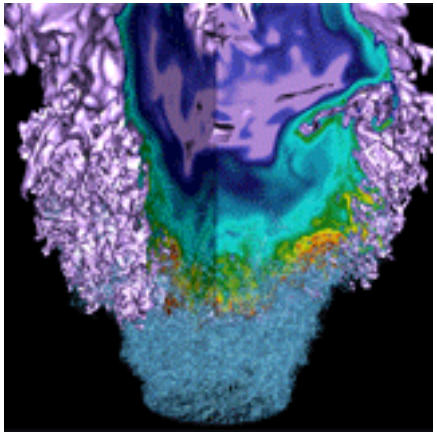
HubZero

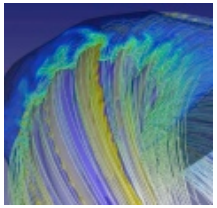

National Integrated Drought Information System
NIDIS
U.S. Drought Portal
www.drought.gov
Contact Us | Log In | Text-Only
Search:
HOME
WHAT IS NIDIS?
CURRENT DROUGHT
FORECASTING
IMPACTS
PLANNING
EDUCATION
RESEARCH
RECOVERY
REPORTS
Area Drought Information
Select State...
Go
Select Region...
Go
Maps & Tools
Map & Data Viewer - new!
Geodata Portal
Drought Monitor Graphics
CRN Soil Data
Events & Announcements
Carolinas DEWS Scoping Workshop July 2012
April 2012 Southern Plains Drought Assessment & Outlook Forum
Risk Management Meeting 11/2011
ACF Climate Outlook Forum and Pilot Review Meeting 2011
Engaging Preparedness Communities Webinar, Dec 13th 1-2 PM EST
View Archive | Portal Release Notes
Regional Drought Webinars
ACFRB Drought Assessment Webinar - July 3rd, 2012
ACFRB Drought Assessment Webinar - June 19th, 2012
ACFRB Drought Assessment Webinar - May 22nd, 2012
ACFRB Drought Assessment Webinar - February 28th, 2012
ACFRB Briefing Presentation - January 17th, 2012
Managing Drought in the Southern
Featured Products
Where are Drought Conditions Now?
How is the Drought Affecting Me?
Will the Drought Continue?
U.S. Drought Monitor
August 7, 2012
Valid 7 a.m. EDT
Intensity:
D0 Abnormally Dry
D1 Drought - Moderate
D2 Drought - Severe
D3 Drought - Extreme
D4 Drought - Exceptional
Drought Impact Types:
Delineates dominant impacts
S = Short-Term, typically <6 months (e.g. agriculture, grasslands)
L = Long-Term, typically >6 months (e.g. hydrology, ecology)
The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. See accompanying text summary for forecast statements.
http://droughtmonitor.unl.edu/
Released Thursday, August 9, 2012
Author: Mark Svoboda, National Drought Mitigation Center
USDA
NIDIS Feature
U.S. Drought Monitor
Drought Update Summer 2012 (Report)
Drought Update Summer 2012 (Image)
Drought Information Statements
Toronto
MEXICO
Miami
Havana
Click on a highlighted area to view the current NWS Drought Information Statement or Click Here to select
Regional Drought Early Warning Systems (DEWS)
California
Upper Colorado River Basin
Four Corners Region
Apalachicola-Chattahoochee-Flint

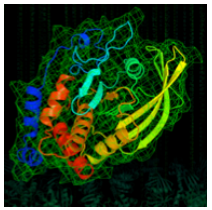
End-to-End Frameworks

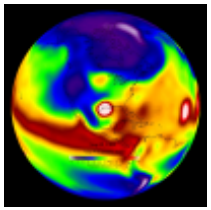
- Plugin/Portlet approach to development requires developers to learn the vernacular of the framework.
- Applications tend to be built around framework constraints rather than science use-cases. Rich set of front-end technologies in HTML5 and JavaScript much harder to leverage in this context.
- If you want a blank-slate that starts with science use cases and builds up interface from here, you might try a different approach.

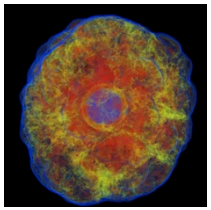
API Driven Access

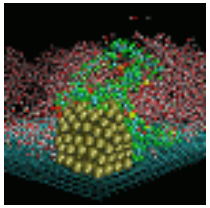












API Driven Access

- Web APIs provide programmatic access to data and resources to developers over the web
- Access to data as well-defined objects allows users to develop their own custom applications and code
- Build web application around this API using the software stack of your choice

The 1-minute intro to REST APIs

- **Use HTTP verbs in conjunction with URLs**
  - GET, POST, PUT, DELETE
- **Verb + URL + parameters = function call**
- **Return structured data**
- **For instance (from NEWT API):**

VERB	RESOURCE	DESCRIPTION
POST	/queue/R	Submits POST data to queue on R; returns job id
GET	/file/R/path/	Returns directory listing for /path/ on R
GET	/account/user/U	Returns user account info for U
DELETE	/queue/R/id	Deletes jobs from Queue

- **Structured Output (JSON):**

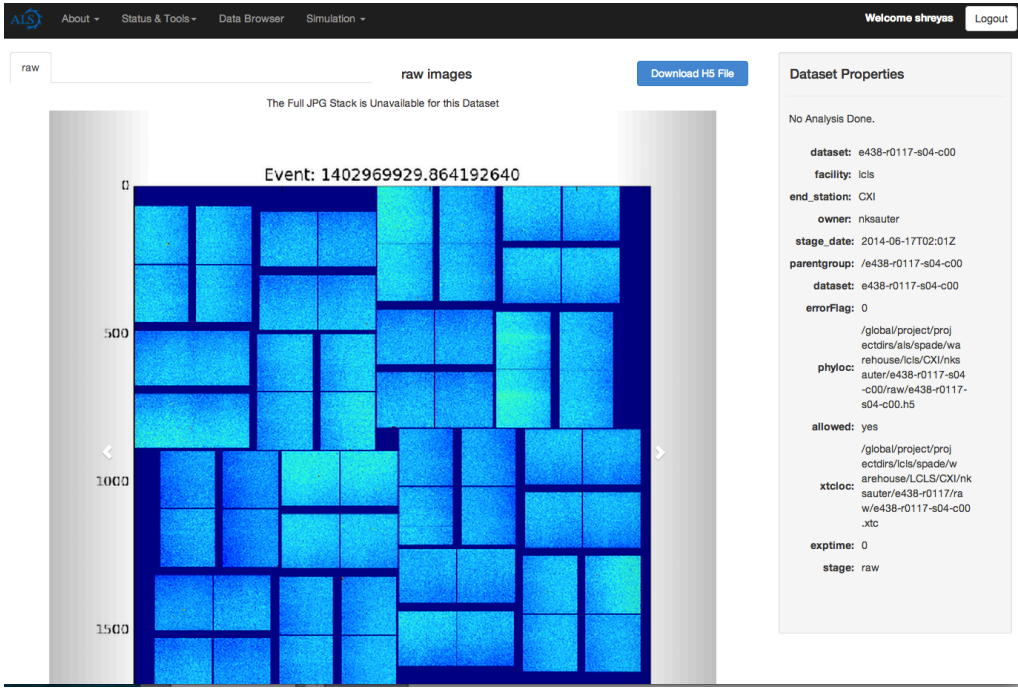
```
{"status": "OK",
"output":["status": "up", "system": "hopper"}, {"status": "up", "system": "carver"}]
"error": ""}
```

APIs Drive Programmatic Access to Data

- **Enable heavy duty server side operations and long running tasks**
- **Perform sub-selection of data – pull down only what is needed in client and render it in the browser**
  - eg. OpenDAP connects HDF5 datasets with web clients

eg. ALS Spot Suite

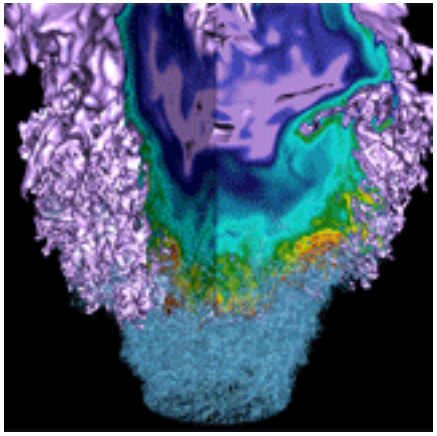
- All the client side logic resides in HTML5/JavaScript
- Make callouts to REST API for all server-side information

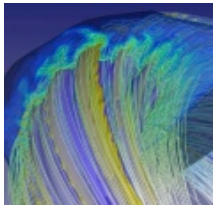


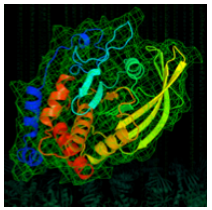
Aside - MVC Frameworks

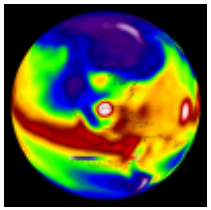
- **Increasingly, web frameworks are following Model-View-Controller pattern**
  - eg. Django, Ruby-On-Rails, Backbone.js
- **Enables separation of user interfaces (V) from backend data models (M) and scientific libraries for interacting with data (C)**
- **Provide you with just enough scaffolding to build out application without imposing front-end application constraints.**
- **Web APIs lend themselves nicely to this model**

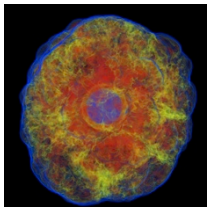
Web IDEs

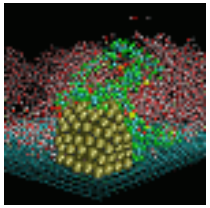






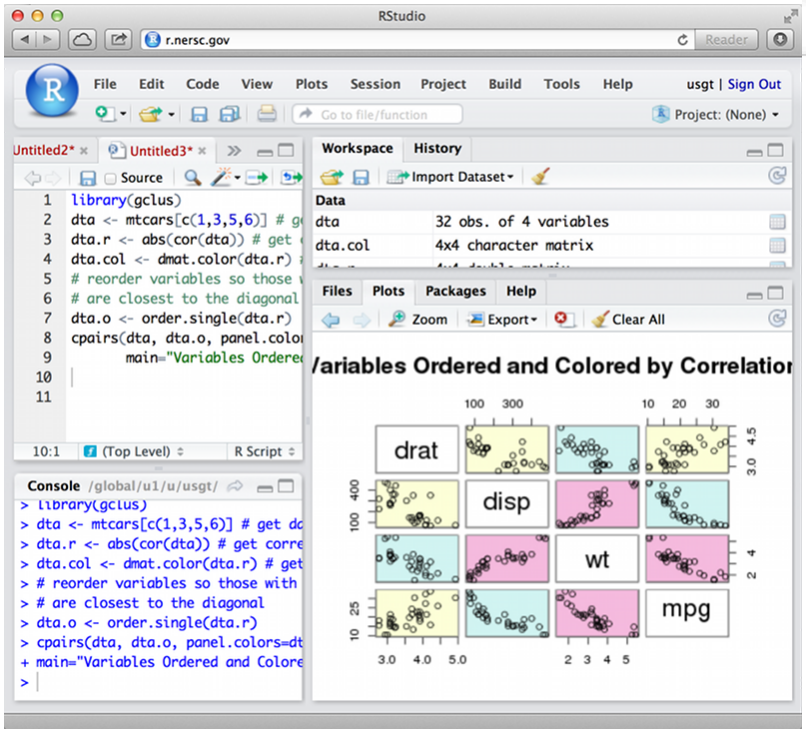


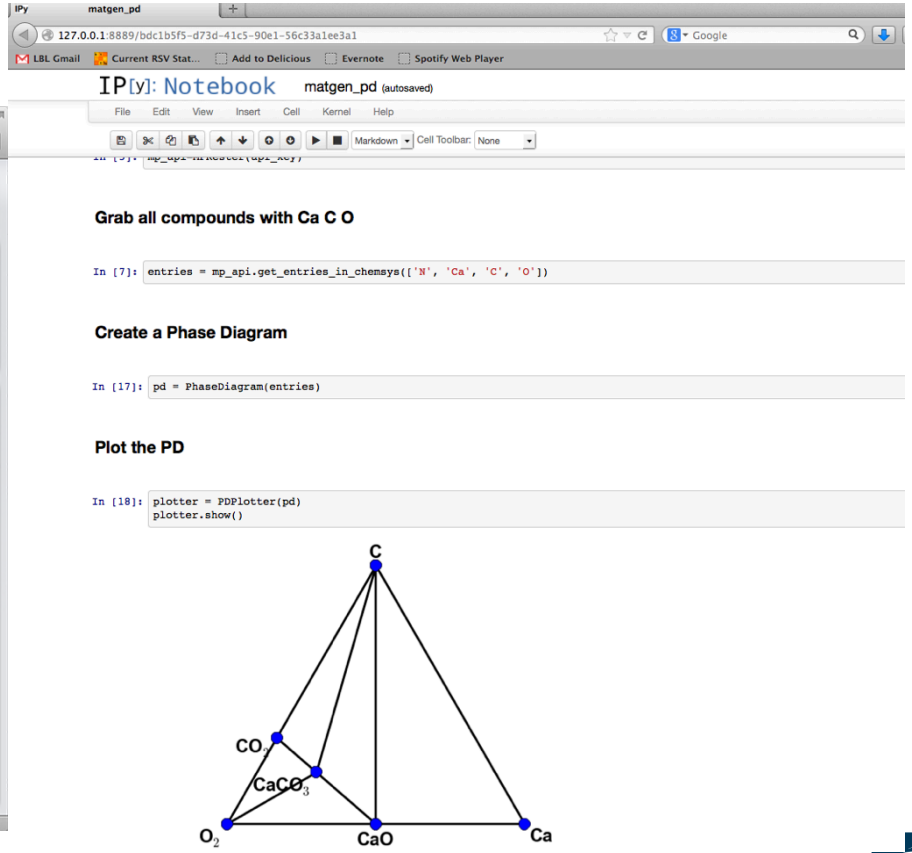




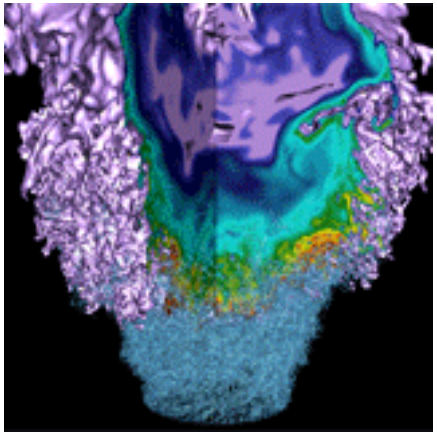
Web IDEs

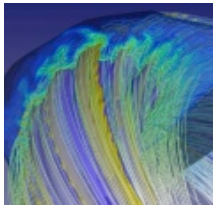
- **Combine the expressiveness of command line tools with web GUIs**
- **RStudio, Mathematica, iPython Notebook**

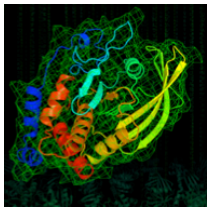


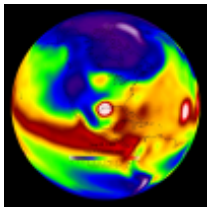


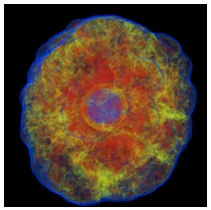
Federated Identity

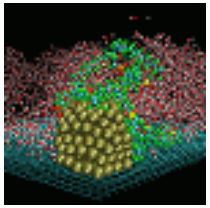












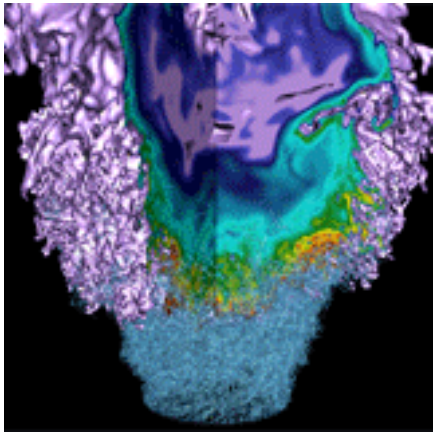
Authentication

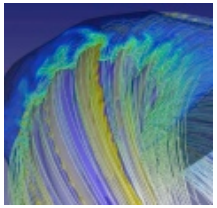
- **Authentication: verifying the identity of the user**
  - Are you who you say you are?
- **Local authentication**
  - Identity tied to local user credentials from institution that hosts the gateway
- **Federated authentication**
  - Identity verification delegated to external party (such as user's home institution, or verified 3rd party)
  - External party returns an assertion to the gateway with ID information.

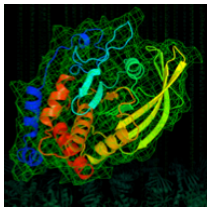
Federated Identity

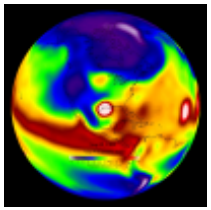
- **One identity to rule them all**
- **Allow users to log-in using common credentials from their home institution / 3rd party provider**
- **Assumes trust relationship between science gateway and identity provider**
- **Map federated ID to local account**
- **Examples:**
  - Globus Nexus (OAuth), CILogon (Shibboleth)
- **Enables access across users from multiple institutions**

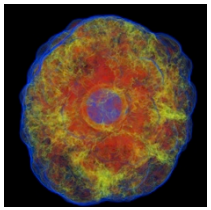
Social Data

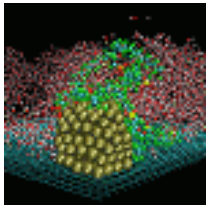










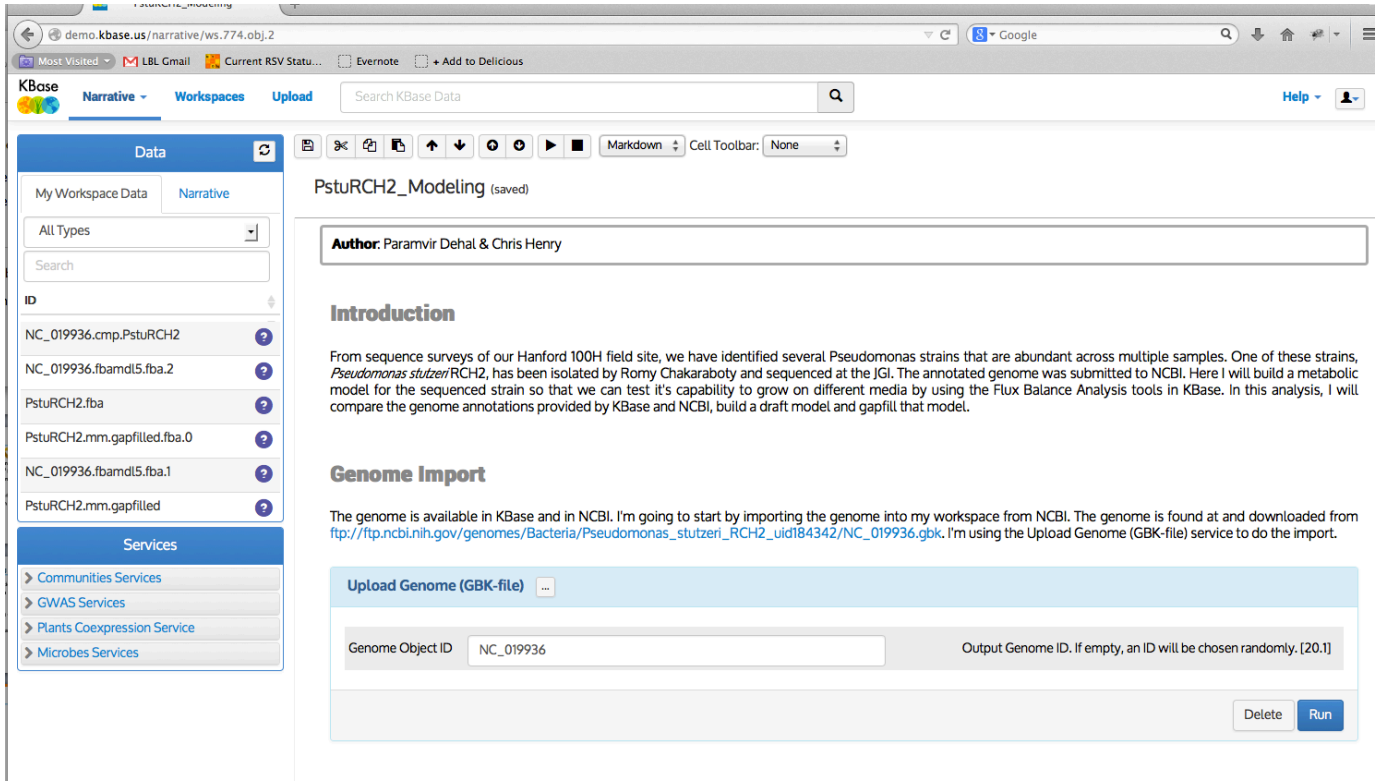


Communities and Social Data

- **Generate a discussion around science results**
- **Combine Science Gateways with social interaction tools**
  - Disqus
  - Forums integrated and cross-linked with datasets
  - Social Media integration
- **Referencing and sharing results and user generated datasets**
- **Users can upload datasets**

KBase Narrative

- **Allows you to reference data and run models from other users in system**



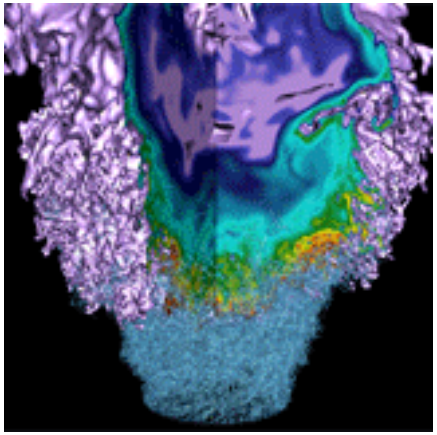
Data Upload and Provenance

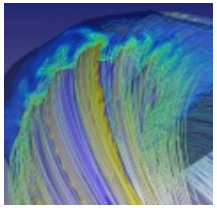
- **Users upload data to portals**
- **But this introduces challenges wrt data management**
  - Vetting/Validating the data
  - Managing Provenance
  - Data ownership and licensing
  - Public/Private/Protected data

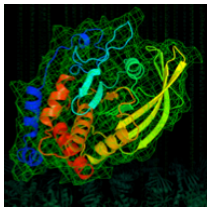
The End

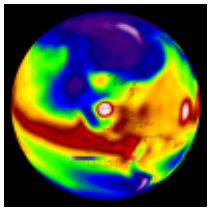
- **Probably not a complete list**
  - Intended to spur further discussion on other interesting patterns
- **NERSC Science Gateways**
  - http://portal.nersc.gov
- **Contact:**
  - Shreyas Cholia: scholia@lbl.gov
- **Questions? Comments?**

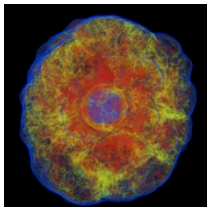
Fin

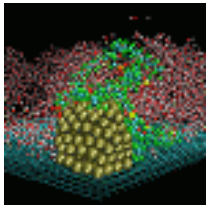




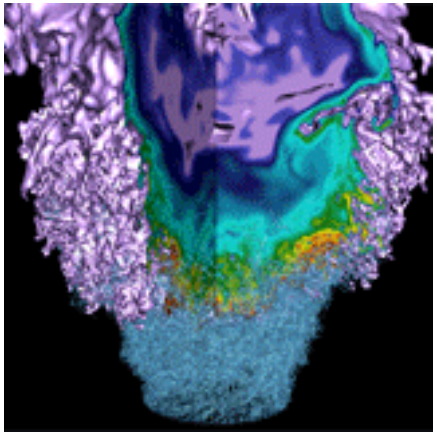


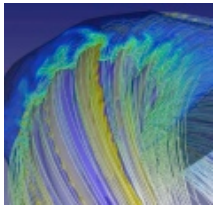


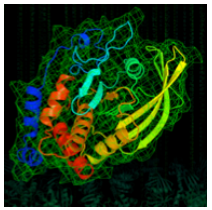


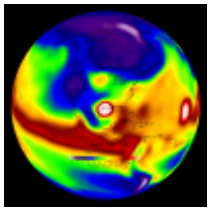


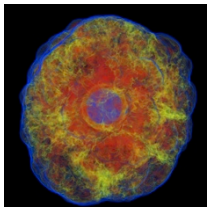
Extra

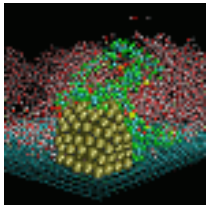












Authorization

- **Authorization: Access privileges for a given identity**
  - What is the user allowed to do?
- **Data portals see some combination of the following sharing modes:**
  - Public – data are publicly viewable to all users
  - Protected – data are only visible to members within a designated group or collaboration
  - Private – data are private to the individual user