The iPlant Collaborative: A model for collaborative science cyberinfrastructure

(at 30 seconds a slide)

Dan Stanzione

Co-Director, The iPlant Collaborative

Deputy Director

Texas Advanced Computing Center, UT-Austin

American Geophysical Union 2012 Meeting

What is iPlant

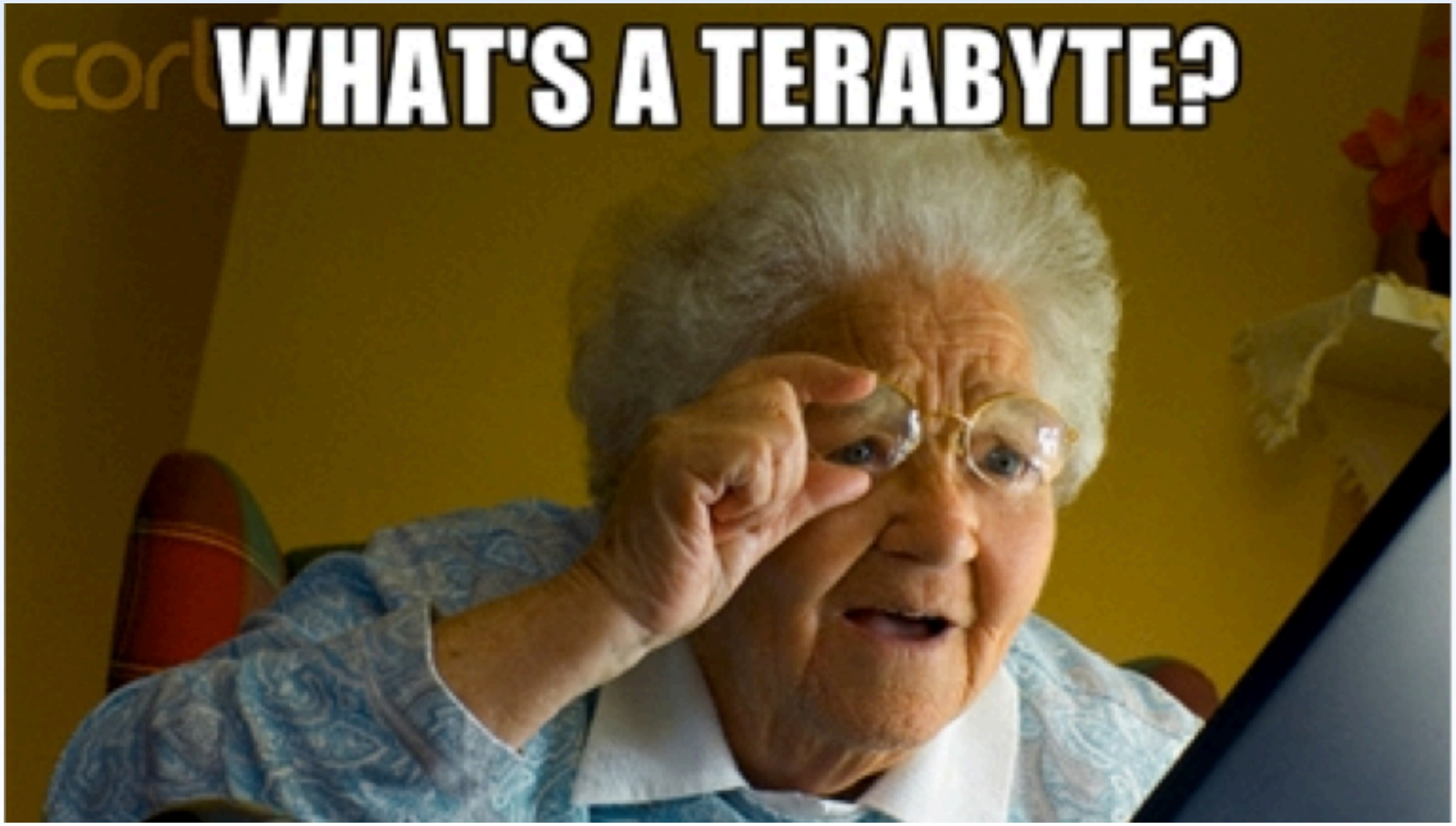
- The iPlant Cyberinfrastructure Collaborative is building a comprehensive informatics *infrastructure* for plant biology.
- (and lately, some animals as well).
- This rapidly evolving infrastructure is sometimes very visible in your work, and sometimes hides in the background.

iPlant – EarthCube

- iPlant began with many goals and a community building process, somewhat similar to EarthCube.
- The starting point in bio is wayyy behind geosciences.
  - Biology wasn't really computational before the late '90s; pre-sequencer not much digital data, very little modeling and simulation, etc.
- Presuming we have done anything right, there might be some lessons for Earthcube (or perhaps just what not to do!).

NSF

iPlant Collaborative™



Data-intensive biology will mean getting biologists comfortable with new technology…

One key goal in our infrastructure, training and outreach is to minimize the emphasis on technology and return the focus to the biology.

1958
Matt Meselson &
Ultracentrifuge, $500,000

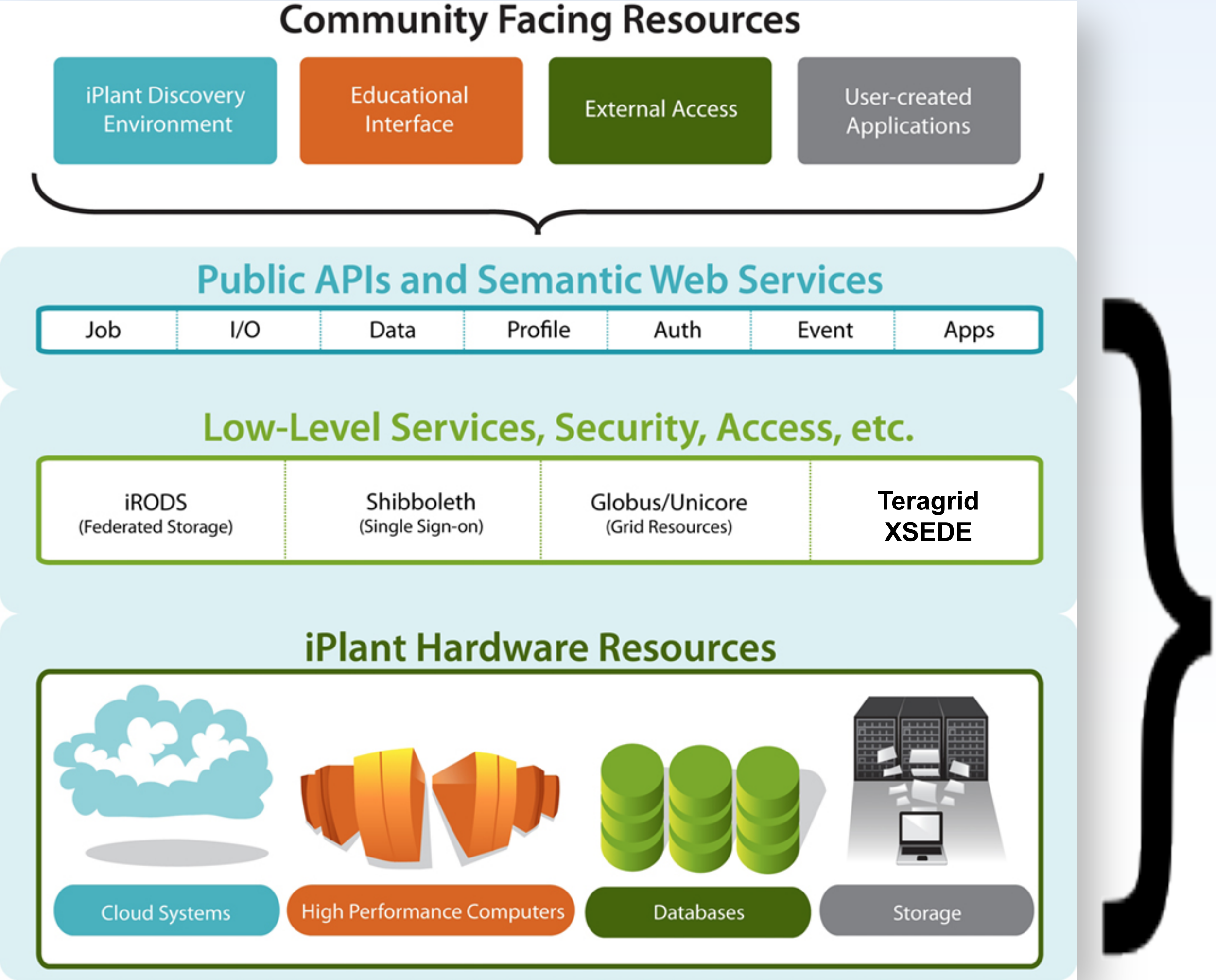
1973
Sharp, Sambrook, Sugden
Gel Electrophoresis Chamber,
$250

iPlant Collaborative™

What does iPlant Provide

- DATA
  - iPlant Data Storage: All data large and small
- COMPUTING:
  - *Large Scale:* Up to hundreds of thousands of processors
  - *Virtual:* "Cloud Style" server hosting
- A Programmer's Interface
  - Easily embed iPlant resources in your applications
- User Interfaces
  - The Discovery Environment: Integrated Web apps.
  - More than 200 bioinformatics applications
  - MyPlant, DNASubway, TNRS, TreeViewer, PhytoBisque, etc

The iPlant Cyberinfrastructure



Community Facing Resources

- iPlant Discovery Environment
- Educational Interface
- External Access
- User-created Applications

Public APIs and Semantic Web Services

Job	I/O	Data	Profile	Auth	Event	Apps

Low-Level Services, Security, Access, etc.

iRODS (Federated Storage)	Shibboleth (Single Sign-on)	Globus/Unicore (Grid Resources)	**Teragrid** **XSEDE**

iPlant Hardware Resources

- Cloud Systems
- High Performance Computers
- Databases
- Storage

End Users

Computational Users

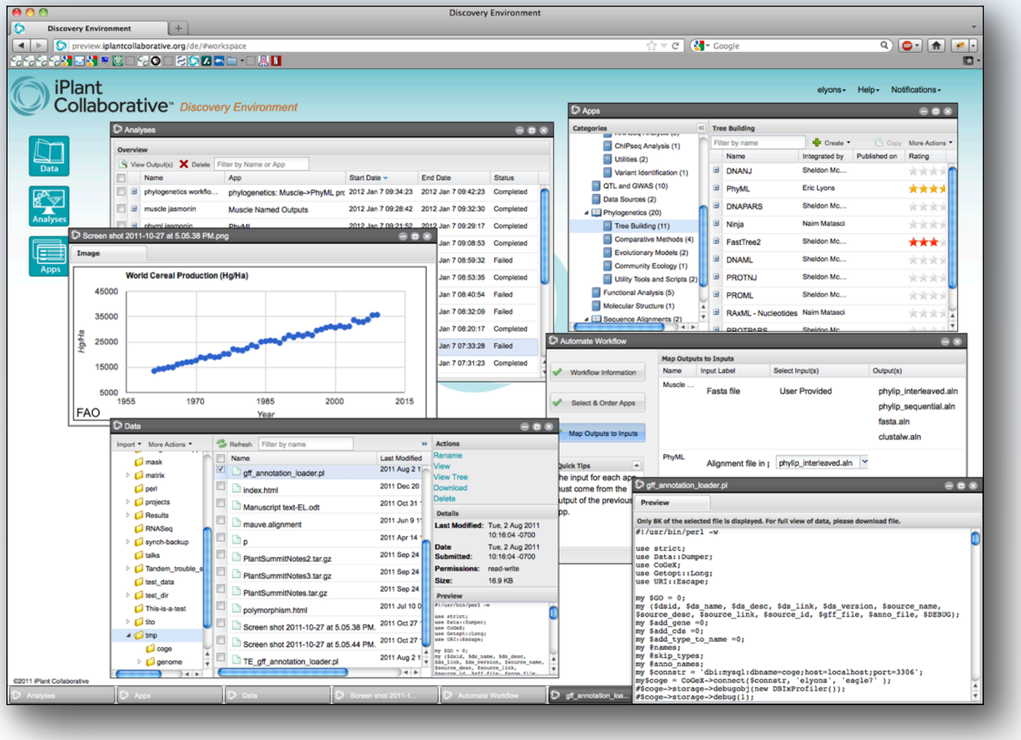
iPlant Collaborative™

Ways to Access iPlant

- Atmosphere: a free cloud computing platform
- Data Store: secure, cloud-based data storage
- Discovery Environment: a web portal to many integrated applications
- DNA Subway: genome annotation, DNA bar-coding (and more) for science educators
- The API: For programmers embedding iPlant infrastructure capabilities
- Command line: for expert access (thru TeraGrid/XSEDE)

The iPlant Discovery Environment

- A rich web client
  - Consistent interface to bioinformatics tools
  - Portal for users who won't want to interact with lower level infrastructure
- An integrated, extensible system of applications and services
  - Additional intelligence above low level APIs – Provenance, Collaboration, etc.



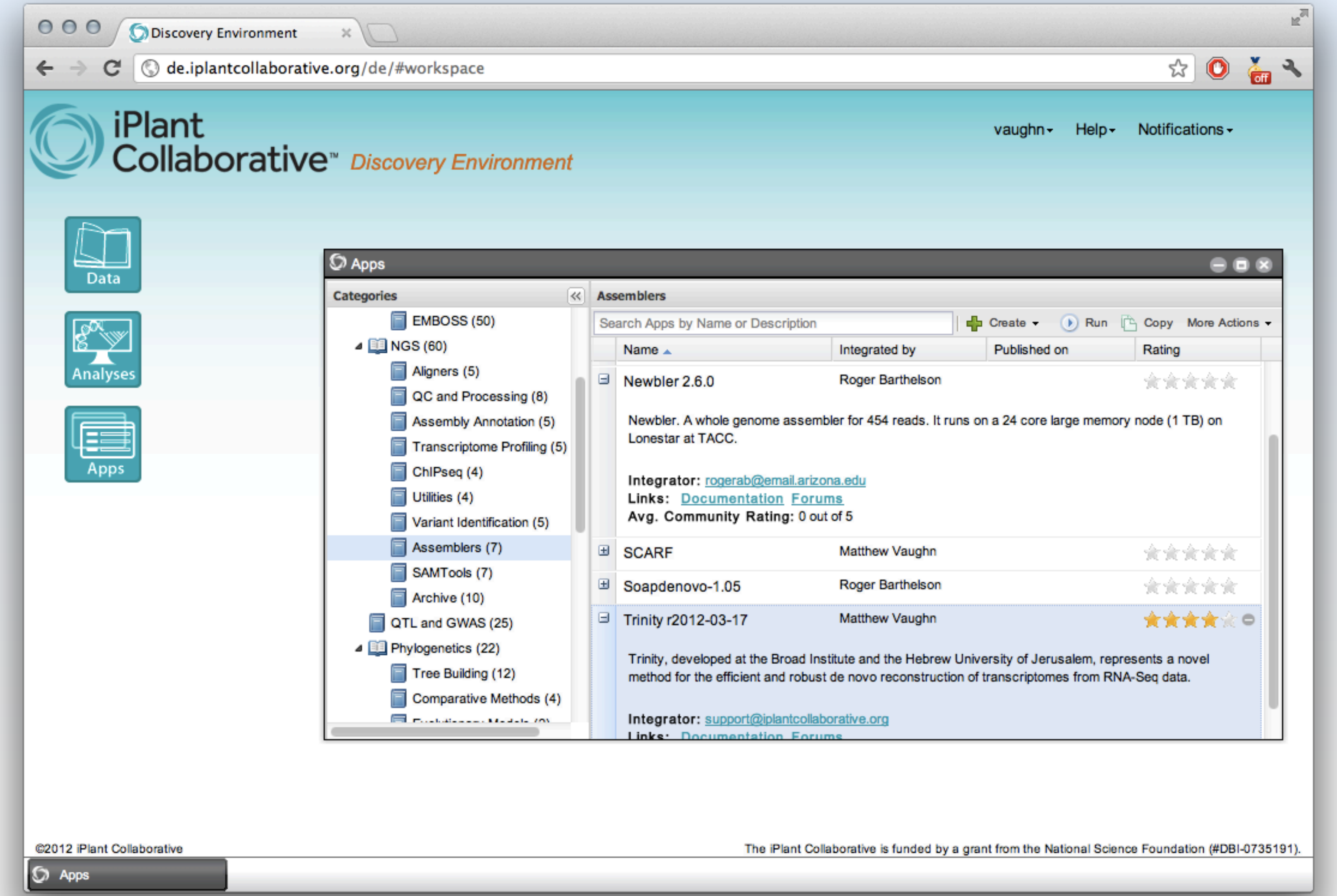
Name	App	Start Date	End Date	Status
phylogenetics workflo...	phylogenetics: Muscle->PhyML prc	2012 Jan 7 09:34:23	2012 Jan 7 09:42:23	Completed
muscle jasmonin	Muscle Named Outputs	2012 Jan 7 09:28:42	2012 Jan 7 09:32:30	Completed
phyml jasmonin	PhyML	2012 Jan 7 09:21:52	2012 Jan 7 09:29:17	Completed

Only 8K of the selected file is displayed. For full view of data, please download file.

```
#!/usr/bin/perl -w

use strict;
use Data::Dumper;
use CoGeX;
use Getopt::Long;
use URI::Escape;

my $GO = 0;
my ($dsid, $ds_name, $ds_desc, $ds_link, $ds_version, $source_name,
$source_desc, $source_link, $source_id, $gff_file, $anno_file, $DEBUG);
my $add_gene =0;
my $add_cds =0;
my $add_type_to_name =0;
my @names;
my @skip_types;
my @anno_names;
my $connstr = 'dbi:mysql:dbname=coge;host=localhost;port=3306';
my$coge = CoGeX->connect($connstr, 'elyons', 'eagle7' );
#$coge->storage->debugobj(new DBIxProfiler());
#$coge->storage->debug(1);
```



Name	Integrated by	Published on	Rating
Newbler 2.6.0	Roger Barthelson		☆☆☆☆☆

Newbler. A whole genome assembler for 454 reads. It runs on a 24 core large memory node (1 TB) on Lonestar at TACC.

Integrator: rogerab@email.arizona.edu
Links: Documentation Forums
Avg. Community Rating: 0 out of 5

Name	Integrated by	Published on	Rating
SCARF	Matthew Vaughn		☆☆☆☆☆
Soapdenovo-1.05	Roger Barthelson		☆☆☆☆☆
Trinity r2012-03-17	Matthew Vaughn		★★★★☆

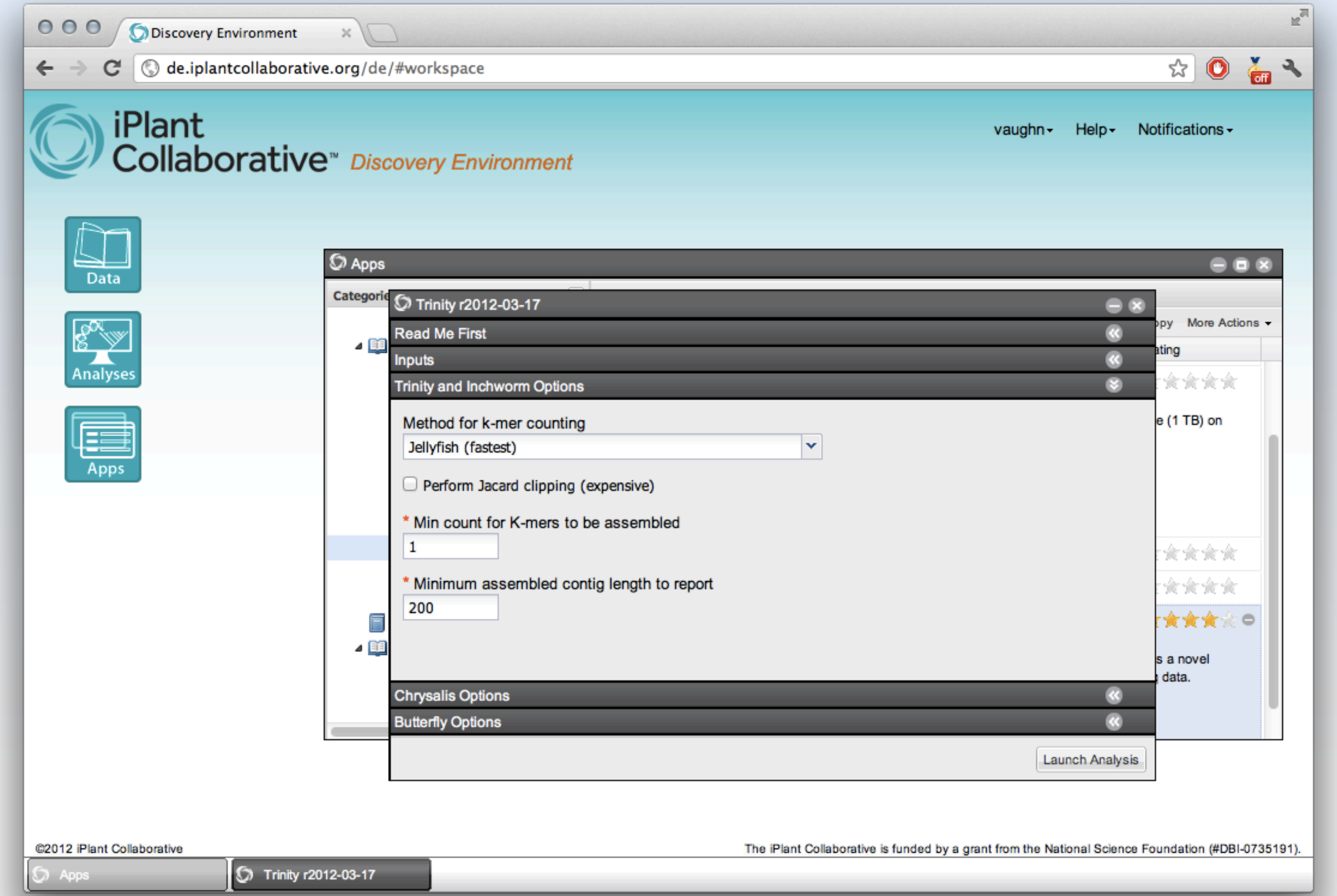
Trinity, developed at the Broad Institute and the Hebrew University of Jerusalem, represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-Seq data.

Integrator: support@iplantcollaborative.org
Links: Documentation Forums

©2012 iPlant Collaborative

The iPlant Collaborative is funded by a grant from the National Science Foundation (#DBI-0735191).

iPlant Collaborative™



Method for k-mer counting

Jellyfish (fastest)

□ Perform Jacard clipping (expensive)

* Min count for K-mers to be assembled

1

* Minimum assembled contig length to report

200

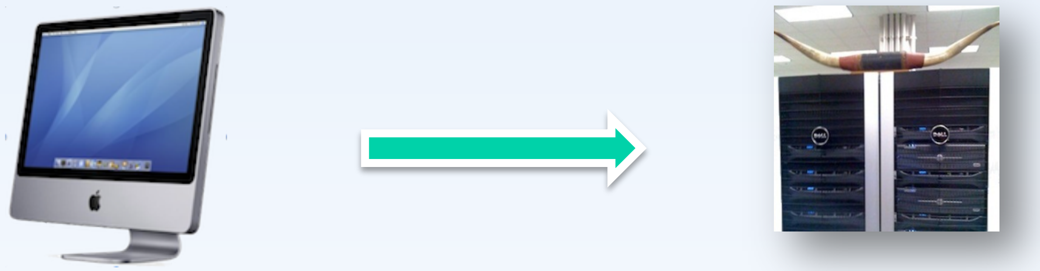
©2012 iPlant Collaborative

The iPlant Collaborative is funded by a grant from the National Science Foundation (#DBI-0735191).

iPlant Collaborative™

Workflows within the DE; Phylogenetics

Trees also present computational challenges



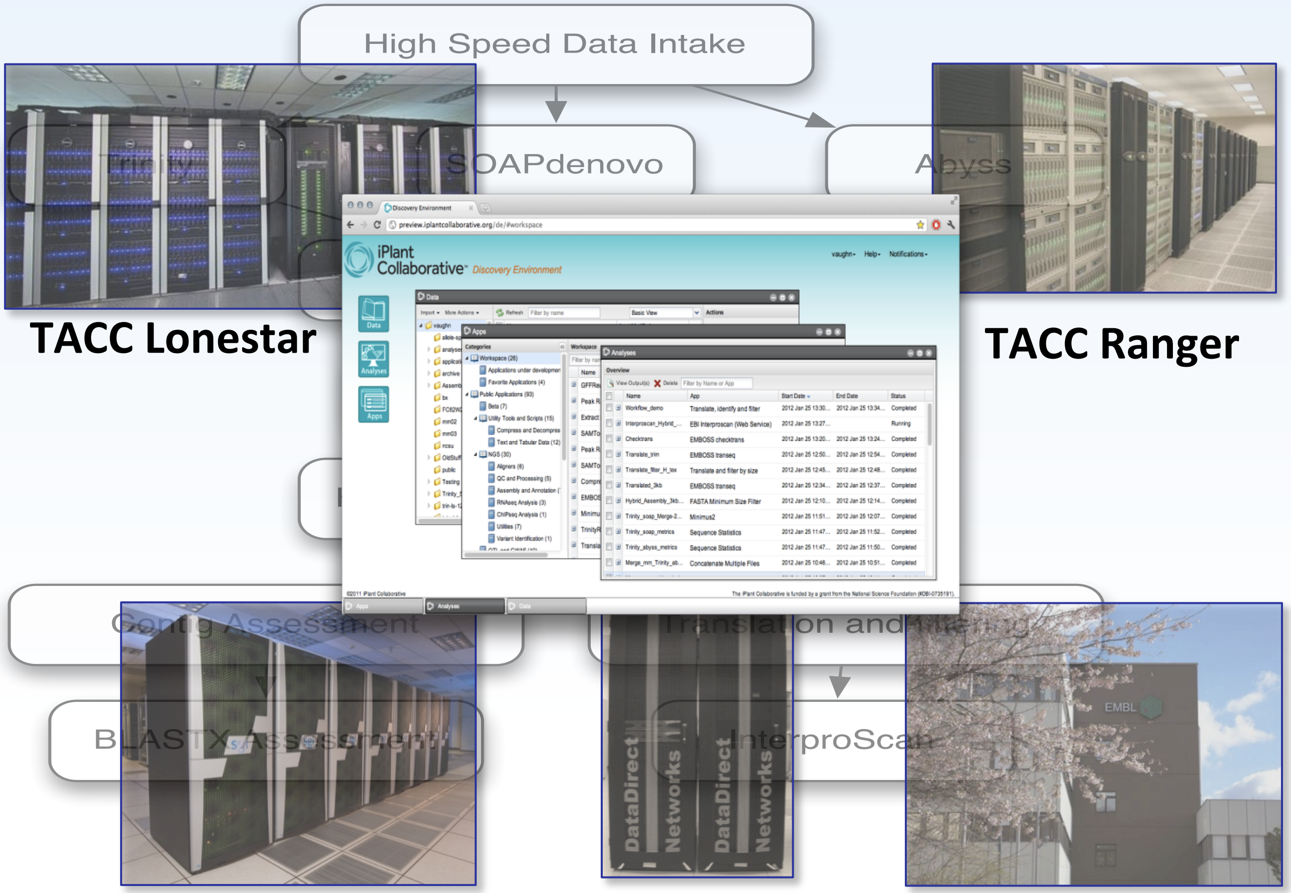
It can take weeks or months to analyze data sets with > 100, 000 species. Example of iPlant contribution:

NINJA/WINDJAMMER	(Neighbor-Joining)
-- NINJA	216K species, ~8 days
-- WINDJAMMER	216K species, ~4 hours

iPlant Collaborative™

Scalable Computation for High-Throughput Inquiry

- 90,000 Compute Cores
- Up to 1TB shared memory
- Growing to ~500,000 cores by end of 2012



TACC Lonestar

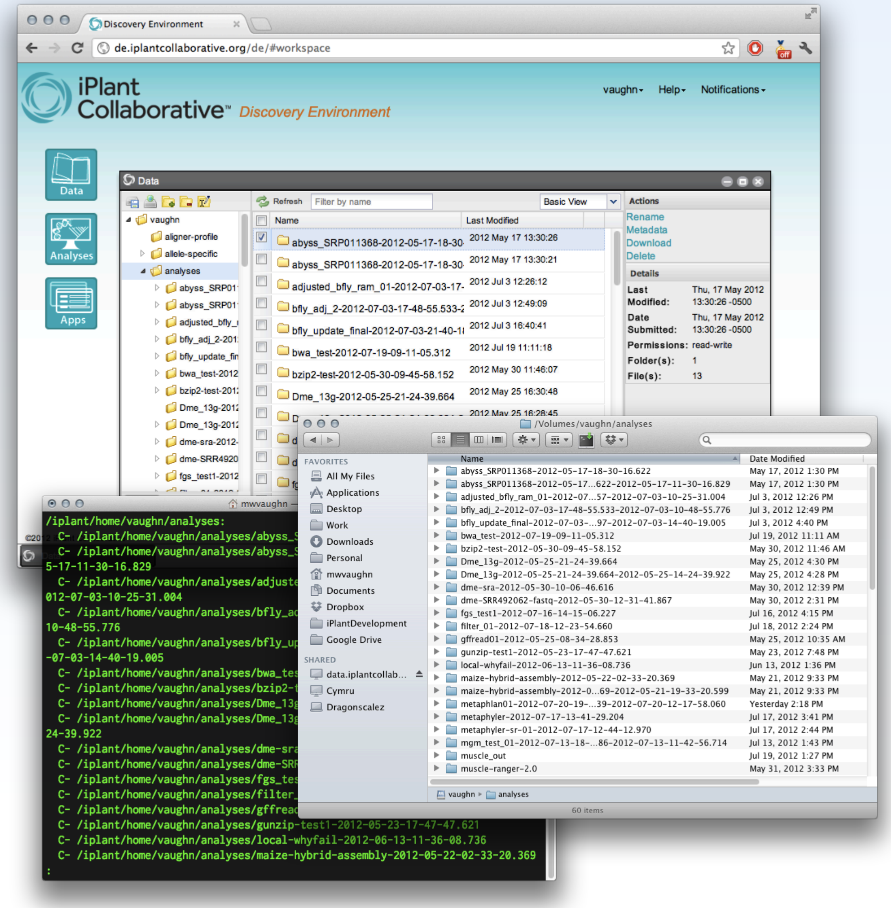
TACC Ranger

PSC Blacklight

TACC Corral

EBI Web Services

The iPlant Data Store



	Name	Last Modified
☑	abyss_SRP011368-2012-05-17-18-30	2012 May 17 13:30:26
□	abyss_SRP011368-2012-05-17-18-30	2012 May 17 13:30:21
□	adjusted_bfly_ram_01-2012-07-03-17-	2012 Jul 3 12:26:12
□	bfly_adj_2-2012-07-03-17-48-55.533-2	2012 Jul 3 12:49:09
□	bfly_update_final-2012-07-03-21-40-1	2012 Jul 3 16:40:41
□	bwa_test-2012-07-19-09-11-05.312	2012 Jul 19 11:11:18
□	bzip2-test-2012-05-30-09-45-58.152	2012 May 30 11:46:07
□	Dme_13g-2012-05-25-21-24-39.664	2012 May 25 16:30:48
□	D	2012 May 25 16:28:45

Last Modified: Thu, 17 May 2012 13:30:26 -0500
Date Submitted: Thu, 17 May 2012 13:30:26 -0500
Permissions: read-write
Folder(s): 1
File(s): 13

Name	Date Modified
abyss_SRP011368-2012-05-17-18-30-16.622	May 17, 2012 1:30 PM
abyss_SRP011368-2012-05-17...622-2012-05-17-11-30-16.829	May 17, 2012 1:30 PM
adjusted_bfly_ram_01-2012-07...57-2012-07-03-10-25-31.004	Jul 3, 2012 12:26 PM
bfly_adj_2-2012-07-03-17-48-55.533-2012-07-03-10-48-55.776	Jul 3, 2012 12:49 PM
bfly_update_final-2012-07-03-...97-2012-07-03-14-40-19.005	Jul 3, 2012 4:40 PM
bwa_test-2012-07-19-09-11-05.312	Jul 19, 2012 11:11 AM
bzip2-test-2012-05-30-09-45-58.152	May 30, 2012 11:46 AM
Dme_13g-2012-05-25-21-24-39.664	May 25, 2012 4:30 PM
Dme_13g-2012-05-25-21-24-39.664-2012-05-25-14-24-39.922	May 25, 2012 4:28 PM
dme-sra-2012-05-30-10-06-46.616	May 30, 2012 12:39 PM
dme-SRR492062-fastq-2012-05-30-12-31-41.867	May 30, 2012 2:31 PM
fgs_test1-2012-07-16-14-15-06.227	Jul 16, 2012 4:15 PM
filter_01-2012-07-18-12-23-54.660	Jul 18, 2012 2:24 PM
gffread01-2012-05-25-08-34-28.853	May 25, 2012 10:35 AM
gunzip-test1-2012-05-23-17-47-47.621	May 23, 2012 7:48 PM
local-whyfail-2012-06-13-11-36-08.736	Jun 13, 2012 1:36 PM
maize-hybrid-assembly-2012-05-22-02-33-20.369	May 21, 2012 9:33 PM
maize-hybrid-assembly-2012-0...69-2012-05-21-19-33-20.599	May 21, 2012 9:33 PM
metaphlan01-2012-07-20-19-...39-2012-07-20-12-17-58.060	Yesterday 2:18 PM
metaphyler-2012-07-17-13-41-29.204	Jul 17, 2012 3:41 PM
metaphyler-sr-01-2012-07-17-12-44-12.970	Jul 17, 2012 2:44 PM
mgm_test_01-2012-07-13-18-...86-2012-07-13-11-42-56.714	Jul 13, 2012 1:43 PM
muscle_out	Jul 19, 2012 1:27 PM
muscle-ranger-2.0	May 31, 2012 3:33 PM

```
/iplant/home/vaughn/analyses:
  C- /iplant/home/vaughn/analyses/abyss_S
  C- /iplant/home/vaughn/analyses/abyss_S
5-17-11-30-16.829
  C- /iplant/home/vaughn/analyses/adjuste
012-07-03-10-25-31.004
  C- /iplant/home/vaughn/analyses/bfly_ad
10-48-55.776
  C- /iplant/home/vaughn/analyses/bfly_up
-07-03-14-40-19.005
  C- /iplant/home/vaughn/analyses/bwa_tes
  C- /iplant/home/vaughn/analyses/bzip2-t
  C- /iplant/home/vaughn/analyses/Dme_13g
  C- /iplant/home/vaughn/analyses/Dme_13g
24-39.922
  C- /iplant/home/vaughn/analyses/dme-sra
  C- /iplant/home/vaughn/analyses/dme-SRR
  C- /iplant/home/vaughn/analyses/fgs_tes
  C- /iplant/home/vaughn/analyses/filter_
  C- /iplant/home/vaughn/analyses/gffread
  C- /iplant/home/vaughn/analyses/gunzip-test1-2012-05-23-17-47-47.621
  C- /iplant/home/vaughn/analyses/local-whyfail-2012-06-13-11-36-08.736
  C- /iplant/home/vaughn/analyses/maize-hybrid-assembly-2012-05-22-02-33-20.369
:
```

Fast data transfers via parallel, non-TCP file transfer

- Move large (>2 GB) files with ease

Multiple, consistent access modes

- iPlant API
- iPlant web apps
- Desktop mount (FUSE/DAV)
- Java applet (iDrop)
- Command line

Fine-grained ACL permissions

- Sharing made simple

Access and a storage allocation is automatic with your iPlant account

iPlant Collaborative™

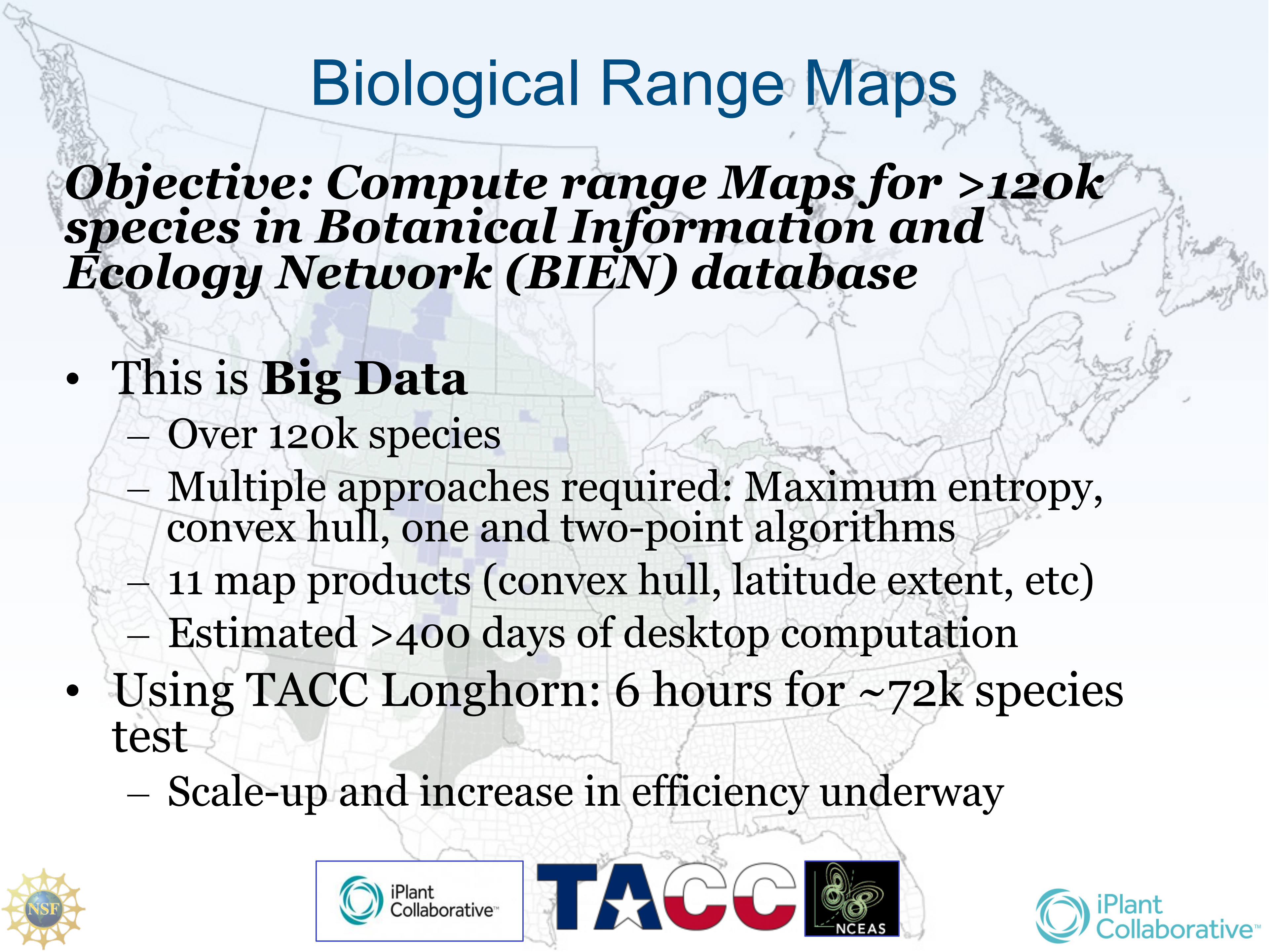
Powered by iPlant

- The iPlant CI is designed as infrastructure. This means it is a platform upon which other projects can build.
- Use of the iPlant infrastructure can take one of several forms:
  - Storage
  - Computation
  - Hosting
  - Web Services
  - Scalability

Powered by iPlant

- Other major projects are beginning to adopt the iPlant CI as their underlying infrastructure (some completely, some in limited ways):
  - BioExtract *(web service platform)*
  - CiPRES *(computation)*
  - Gates Integrated Breeding Platform *(hosting, development)*
  - Galaxy (*storage, for now)*
  - CoGE *(authentication, hostingo*
  - TAIR

Biological Range Maps



Objective: Compute range Maps for >120k species in Botanical Information and Ecology Network (BIEN) database

- This is **Big Data**
  - Over 120k species
  - Multiple approaches required: Maximum entropy, convex hull, one and two-point algorithms
  - 11 map products (convex hull, latitude extent, etc)
  - Estimated >400 days of desktop computation
- Using TACC Longhorn: 6 hours for ~72k species test
  - Scale-up and increase in efficiency underway

The iPlant Collaborative

Project Atmosphere™: Custom Cloud Computing

- API-compatible implementation of Amazon EC2/S3 interfaces
- Virtualize the execution environment for applications and services
- Up to 12 core / 48 GB instances
- Access to Cloud Storage + EBS
- Big data and the desktop are co-local again
  - Bring your data to Atmosphere VM for interactive access and analysis
  - Send it back to the DE for transactional analysis

Eucalyptus

>60 hosted applications in Atmosphere today, including users from USDA, Forest Service, database providers, etc.

(30 more for postdocs and grad students for training classes)

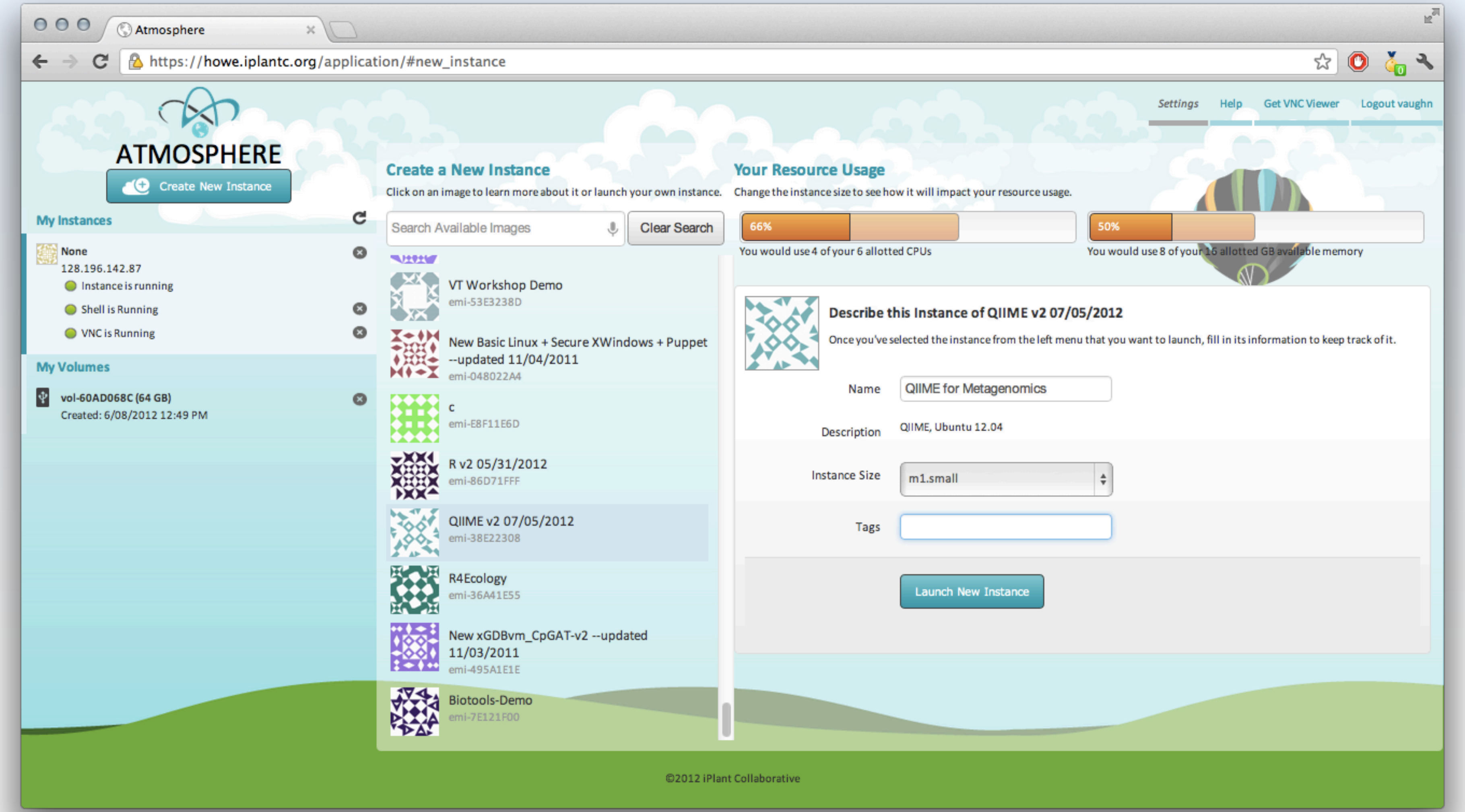
Atmosphere: Motivations

- Standalone GUI-based applications are frequently required for analysis
- GUI apps not easily to transform into web apps
- Need to handle complex software dependencies (e.g specific bioperl version and R modules)
- Users needing full control of their software stack (occasional sudo access)
- Need to share desktop/applications for collaborative analysis (remote collaborators)

Atmosphere: What is it?

- Self-service cloud infrastructure
- Designed to make underlying cloud infrastructure easy to use by novice user
- Built on open source Eucalyptus
- Fully integrated into iPlant authentication and storage and HPC capabilities
- Enables users to build custom images/ **appliances** and share with community
- Cross-platform desktop access to GUI applications in the cloud (using VNC)
- Provide easy web based access to resources

Atmosphere: Launch a new VM



https://howe.iplantc.org/application/#new_instance

My Instances

None
128.196.142.87
Instance is running
Shell is Running
VNC is Running

My Volumes

vol-60AD068C (64 GB)
Created: 6/08/2012 12:49 PM

Create a New Instance

Click on an image to learn more about it or launch your own instance.

VT Workshop Demo
emi-53E3238D

New Basic Linux + Secure XWindows + Puppet --updated 11/04/2011
emi-048022A4

c
emi-E8F11E6D

R v2 05/31/2012
emi-86D71FFF

QIIME v2 07/05/2012
emi-38E22308

R4Ecology
emi-36A41E55

New xGDBvm_CpGAT-v2 --updated 11/03/2011
emi-495A1E1E

Biotools-Demo
emi-7E121F00

Your Resource Usage

Change the instance size to see how it will impact your resource usage.

66%
You would use 4 of your 6 allotted CPUs

50%
You would use 8 of your 16 allotted GB available memory

Describe this Instance of QIIME v2 07/05/2012

Once you've selected the instance from the left menu that you want to launch, fill in its information to keep track of it.

Name: QIIME for Metagenomics

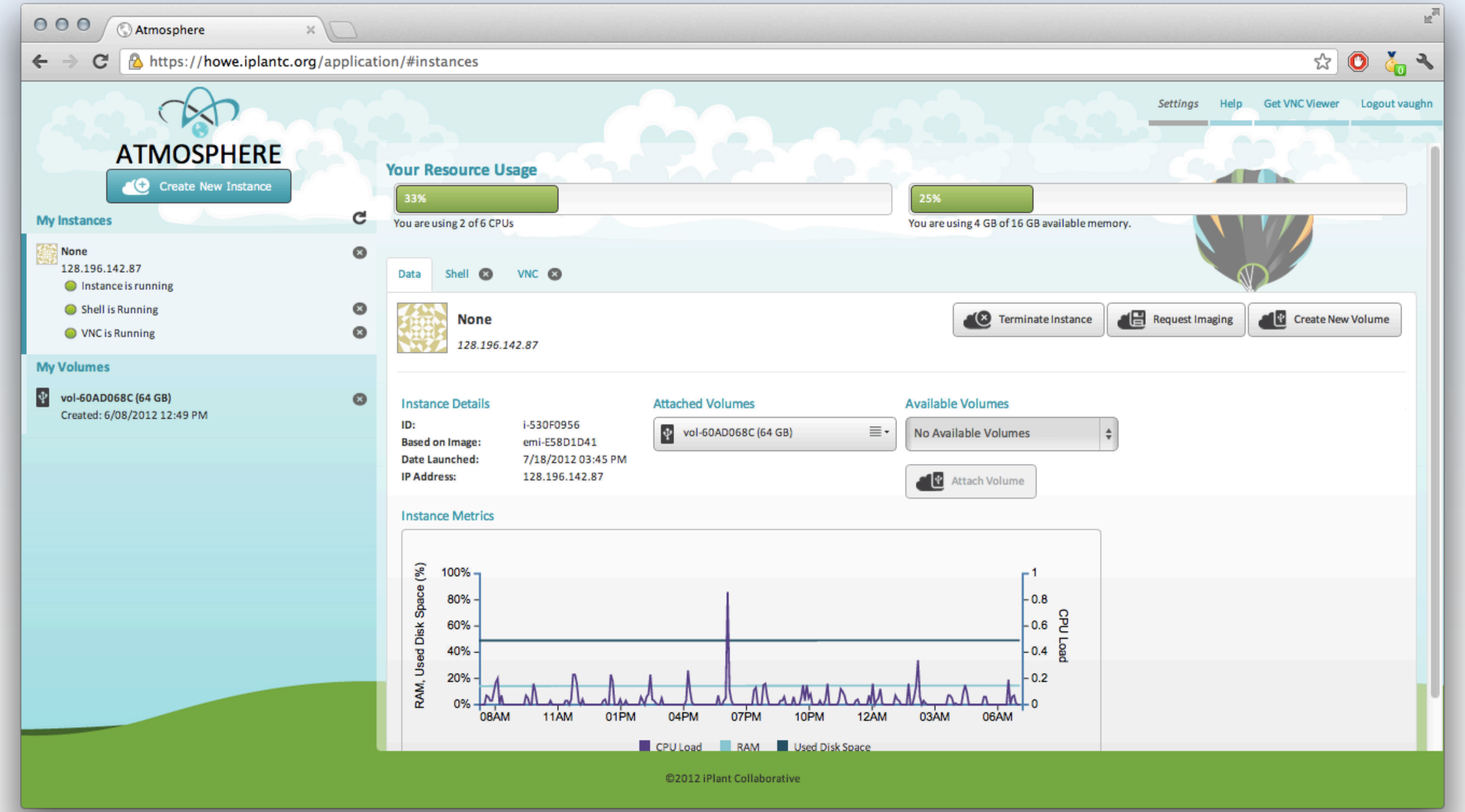
Description: QIIME, Ubuntu 12.04

Instance Size: m1.small

Tags:

©2012 iPlant Collaborative

Atmosphere: Access a running VM



My Instances

None
128.196.142.87
Instance is running
Shell is Running
VNC is Running

My Volumes

vol-60AD068C (64 GB)
Created: 6/08/2012 12:49 PM

Your Resource Usage

33% You are using 2 of 6 CPUs

25% You are using 4 GB of 16 GB available memory.

None
128.196.142.87

Instance Details

ID:	i-530F0956
Based on Image:	emi-E58D1D41
Date Launched:	7/18/2012 03:45 PM
IP Address:	128.196.142.87

Attached Volumes

vol-60AD068C (64 GB)

Available Volumes

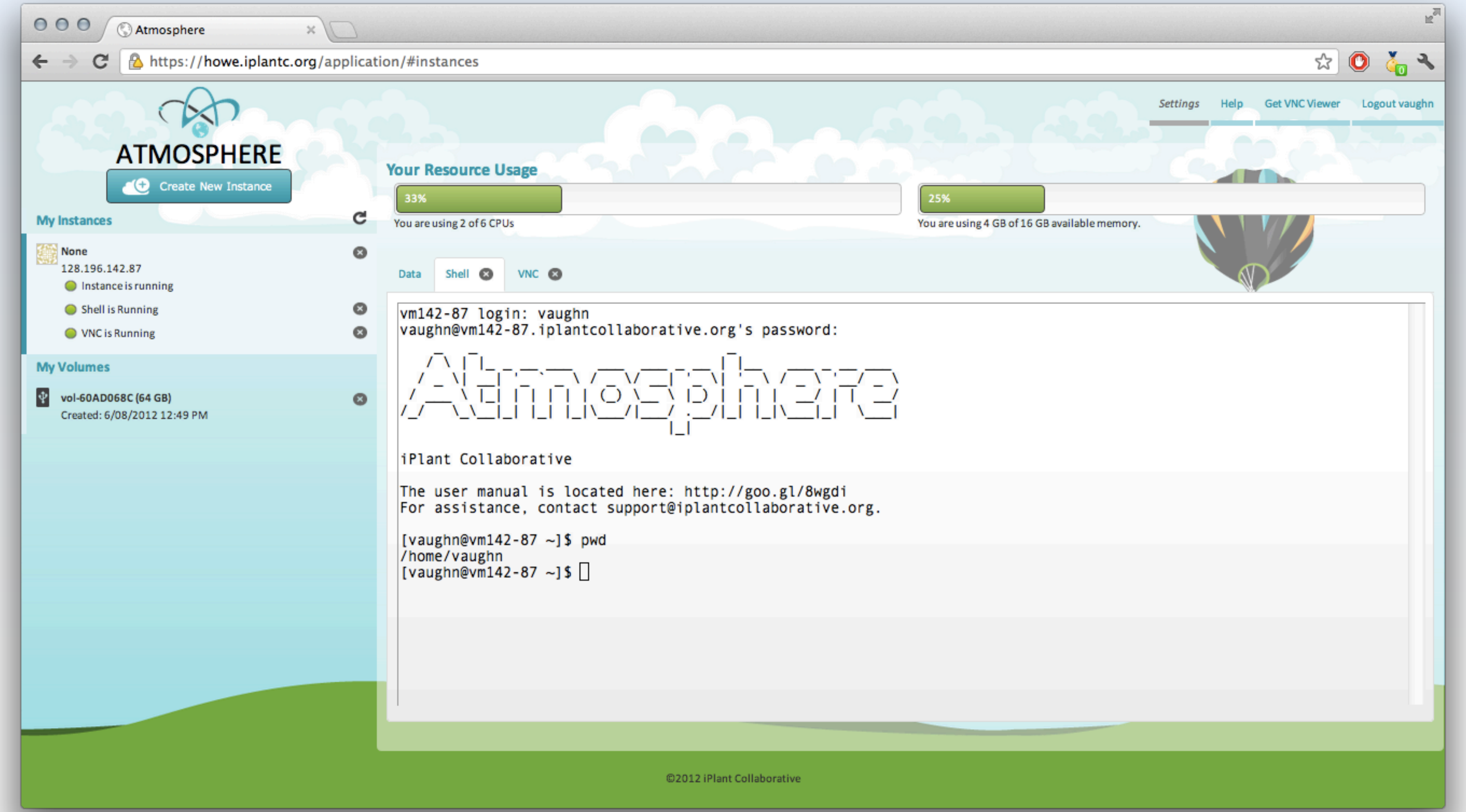
No Available Volumes

Instance Metrics

©2012 iPlant Collaborative

iPlant Collaborative™

Atmosphere: Log in via shell



```
vm142-87 login: vaughn
vaughn@vm142-87.iplantcollaborative.org's password:
    _   _                             _
   / \ | |_ _ __ ___   ___  ___ _ __ | |__   ___ _ __ ___
  / _ \| __| '_ ` _ \ / _ \/ __| '_ \| '_ \ / _ \ '__/ _ \
 / ___ \ |_| | | | | | (_) \__ \ |_) | | | |  __/ | |  __/
/_/   \_\__|_| |_| |_|\___/|___/ .__/|_| |_|\___|_|  \___|
                               |_|

iPlant Collaborative

The user manual is located here: http://goo.gl/8wgdi
For assistance, contact support@iplantcollaborative.org.

[vaughn@vm142-87 ~]$ pwd
/home/vaughn
[vaughn@vm142-87 ~]$ 
```

Collaborations within iPlant

- At the conception of the project, there were, ahem, “wide ranging” notions of what CI collaboration meant.
- Videoconferencing, chat, social networking of all shapes and sizes, wikis, interactive visualization, matchmaking services, virtual conferences, etc.
- Remember that “behind” starting point

Data sharing is the key collaboration mechanism

- Lots of these things were essentially solved problems.
- Re-implementing them was usually a bad idea.
- The ability to have a shared dataspace mattered more than the communication mechanisms.
  - A working dataspace, not just an archive.
  - Ability to share the workflows used to analyze the data, not just the data.

A 15 second case study

- A large plant transcriptome project chose to work with iPlant (who will remain nameless to protect the guilty).
- They wanted to use iPlant storage, and wanted help building ingest pipelines.
- They were not interested in changing their workflow, so they didn't use things like metadata support.
- Now with 5,000+ directories, no search capability beyond "grep", no unique identifier scheme, and no notion of how to pull data to release from the overall corpus, they are seeing the pain of not using proper data management, metadata, etc.

Additional Lessons

- Consensus is not going to merge on the collaboration issue (or most other ones).
- Listen carefully to the community, but don't be paralyzed.
  - Build some things. Convince people to use them. Give up on the ones that don't work, and focus on the ones that do.
  - Then try some more.
  - Find users in the community who want to *invest* their time in doing things differently, and make them successful.
- There are other projects in the world; build with federation in mind from day 1 (how does your data repository talk to other ones?).

Additional Lessons(2)

- There are other projects in the world; build with federation in mind from day 1
  - How does your data repository talk to other ones?.
  - Can you use a standard interface? Or build on top of one? (i.e. REST).
- (Strongly) encouraging best practices is really important… but requiring compliance can be self-defeating
  - Semantic web technologies is a good example)
  - Data formats, Ontologies, Metadata/tagging, versioning, unique IDS, retention policies, transfer mechanisms, etc.

Does it work?

- 7,000+ users have DE accounts.
- 200+ Atmosphere images exist.
- 300+ TB is stored in the Data store.
- iPlant users have run thousands of jobs using millions of hours on supercomputers through iPlant in the last year.
- 100+ NSF awards mention iPlant

- It's hard to authoritatively say if this means transformational science… but something is happening.

Questions?

dan@tacc.utexas.edu

dan@iplantcollaborative.org