

# iPlant: Cyberinfrastructure for Plant Sciences (and Beyond)

Nirav Merchant

[nirav@email.arizona.edu](mailto:nirav@email.arizona.edu)



# One Big Problem...



# One Big Problem...

2008



2011



# One Big Problem...





# One Big Problem...

Published online 3 September 2008 | *Nature* **455**, 16-21 (2008) |

doi:10.1038/455016a

## News Feature

### Big data: Welcome to the

What does it take to store bytes by the te  
trillions? *Nature* **455**, 30 (4 September 2008) | doi:10  
which it's a

### Big data: Distilling mean

Felice Franke *Nature* **455**, 28-29

Buried in va Big data:  
need to cra  
Rosalind Re Clifford Lynch

1. Clifford L  
Circ *Natur*  
Univ  
Ema

Scientist  
long hau Res  
Lynch. resp  
scie

- Transfer
- Storage
- Analysis
- Visualization
- Metadata Mark-up
- Search and Discover
- Share/Collaborate
- Publish

Winston Hide<sup>6,7</sup>, David P. Hill<sup>8</sup>, Renate Kania<sup>9</sup>, Mary Schaeffer<sup>10,11</sup>, Susan St  
Pierre<sup>12</sup>, Simon Twigger<sup>13</sup>, Owen White<sup>14</sup> & Seung Yon Rhee<sup>15</sup>

**To thrive, the field that links biologists and their data urgently  
needs structure, recognition and support.**



# got data ?



# Flavors of Bio-Information

- Sequences (most popular)
- Structures
- Images
- Video
- Audio
- Pathways (graphs)
- Text (Publications)
- Traces
- Combination (eg Video & Traces)
- And much more ...



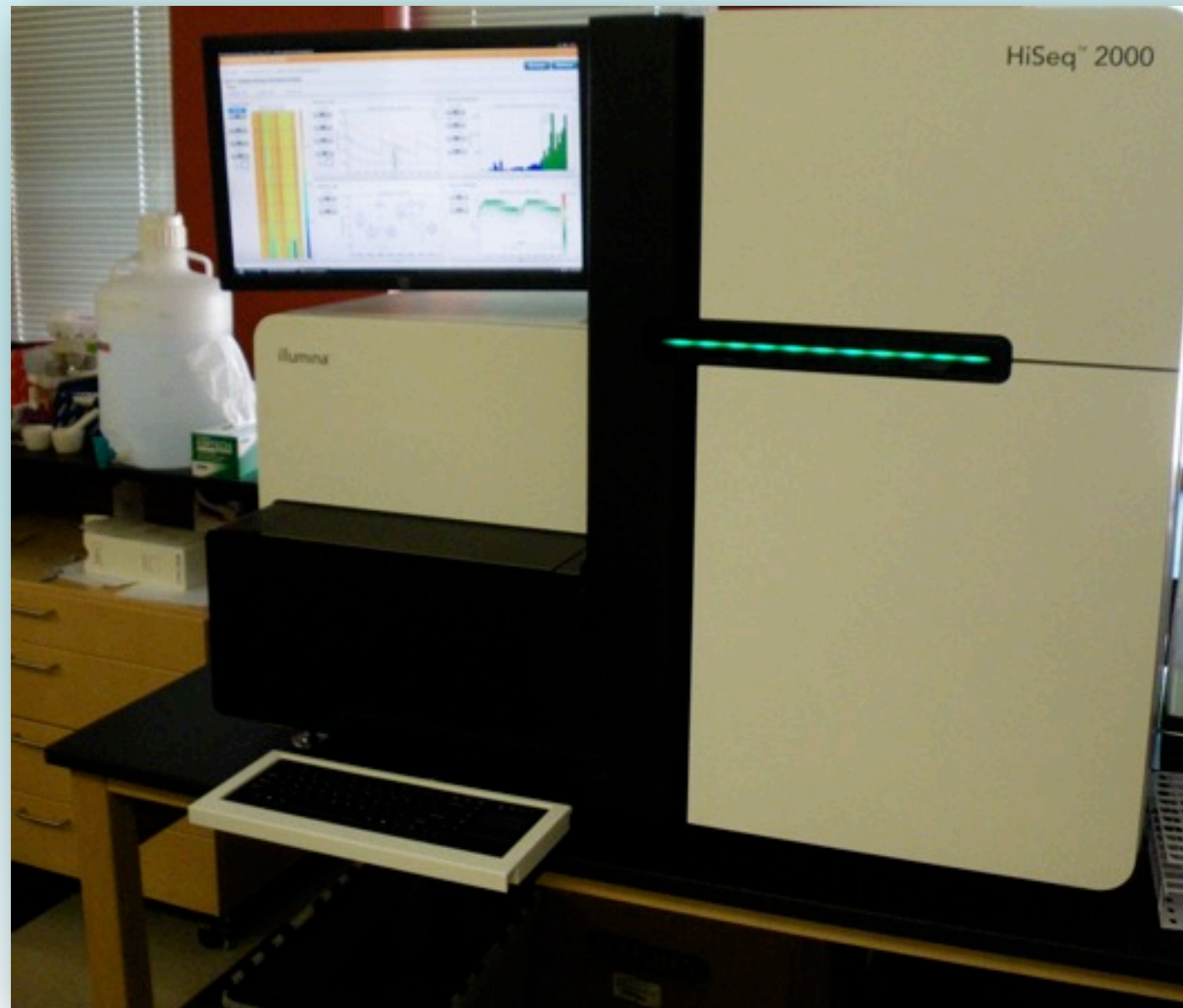


# Growth

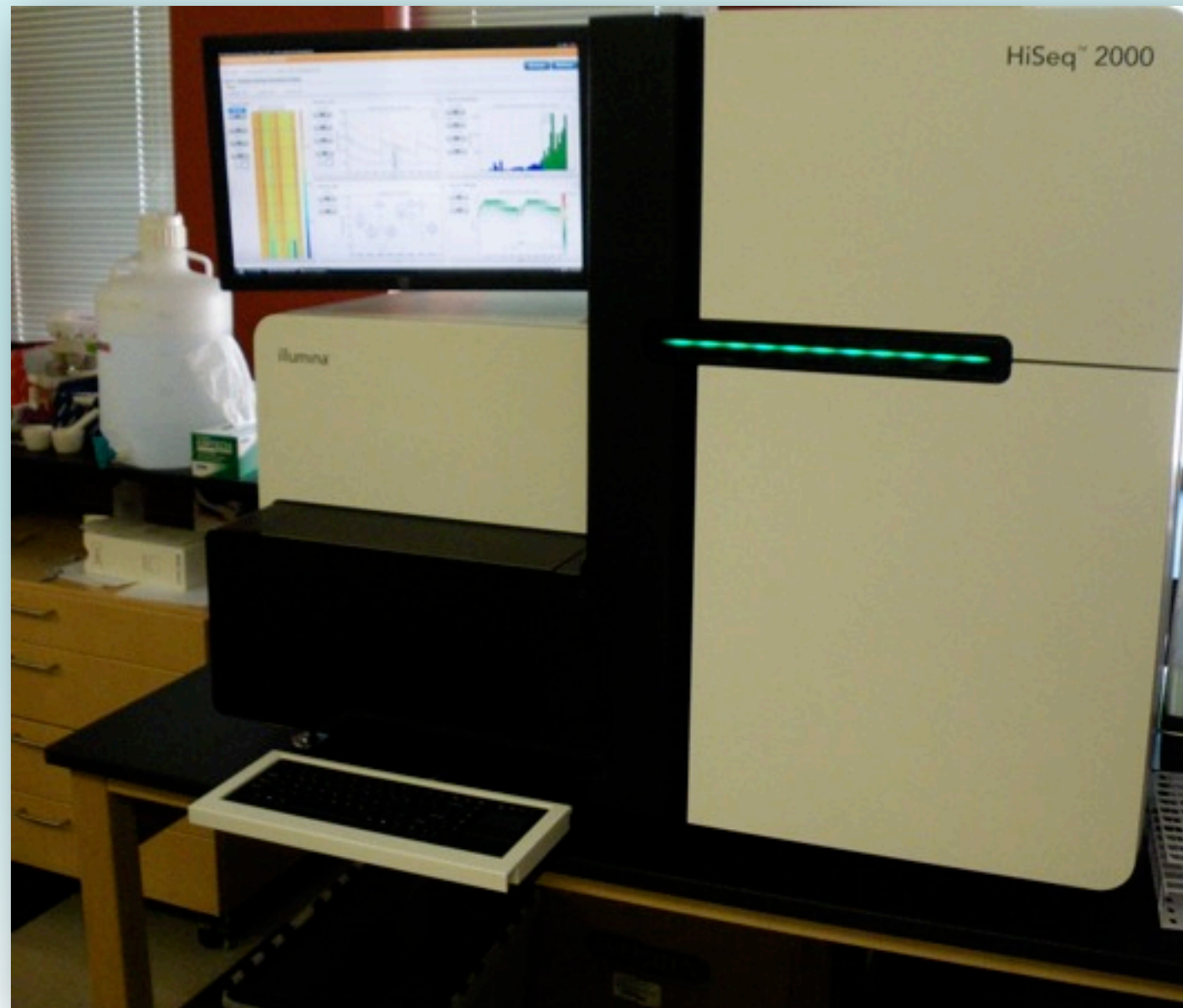
- The Human genome project (1990-2003)
  - 13 years • 23 labs • \$500 Million.
- A Human genome today (2010)
  - ~3 days • ~1 machine. • ~\$10,000.
- Many 1000 and 10,000 genomes projects (human, plant etc)
- Bench-top sequencers are
- “Next-next-next” generation sequencers will produce \$500 genome in 5 years



# High-throughput Data Acquisition

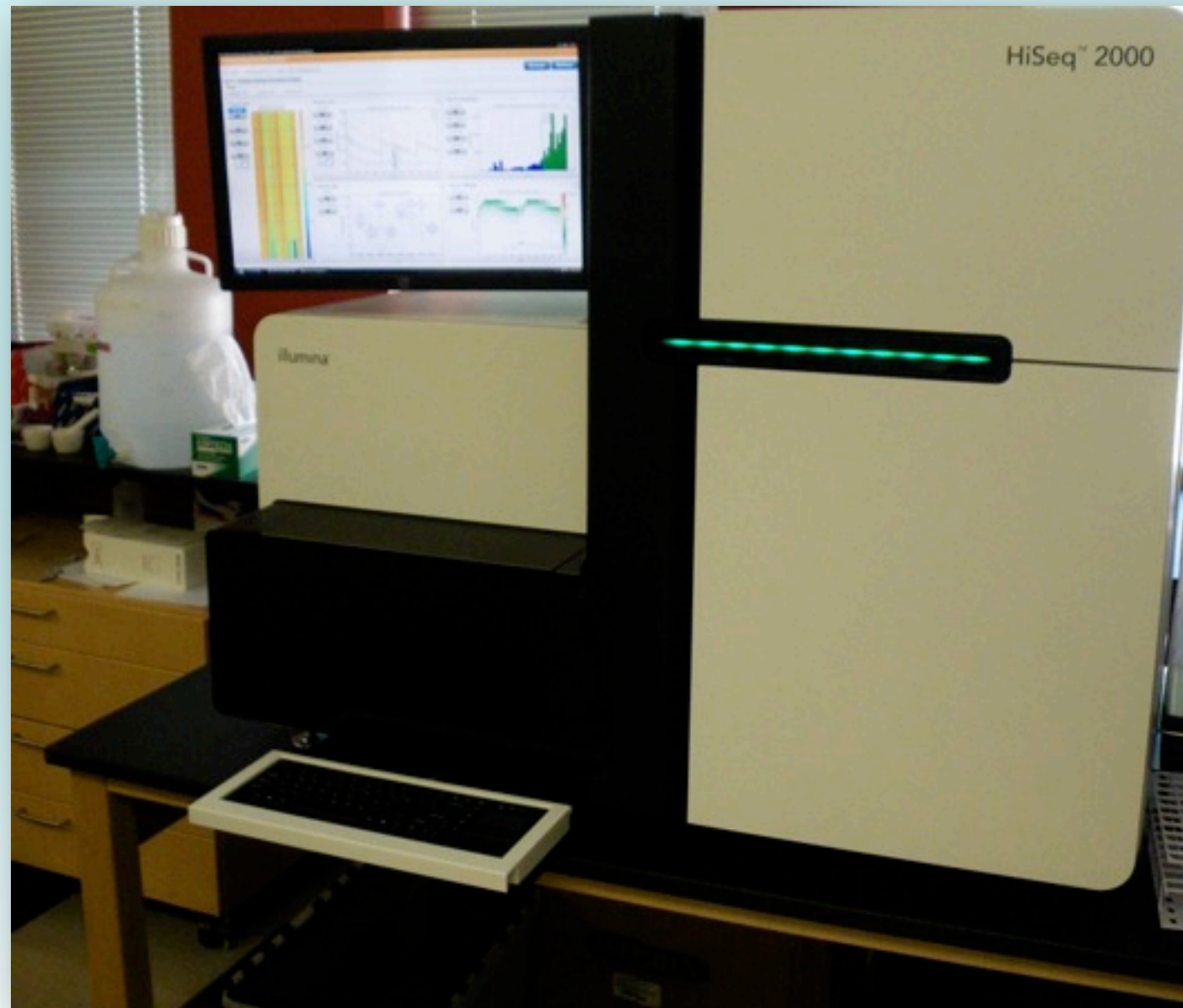


# High-throughput Data Acquisition



In 11 Days

# High-throughput Data Acquisition

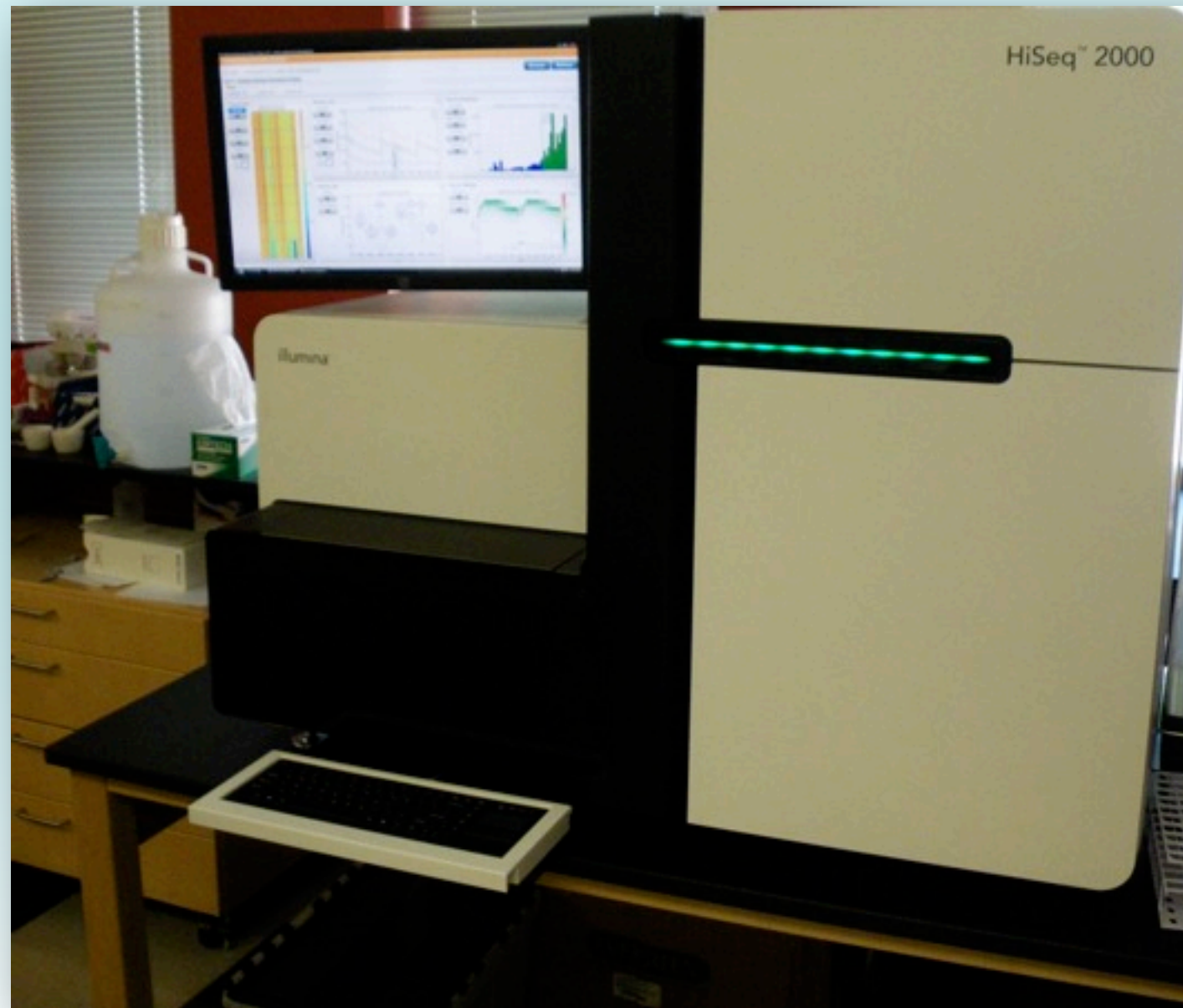


In 11 Days

- Generates 4TB of raw data



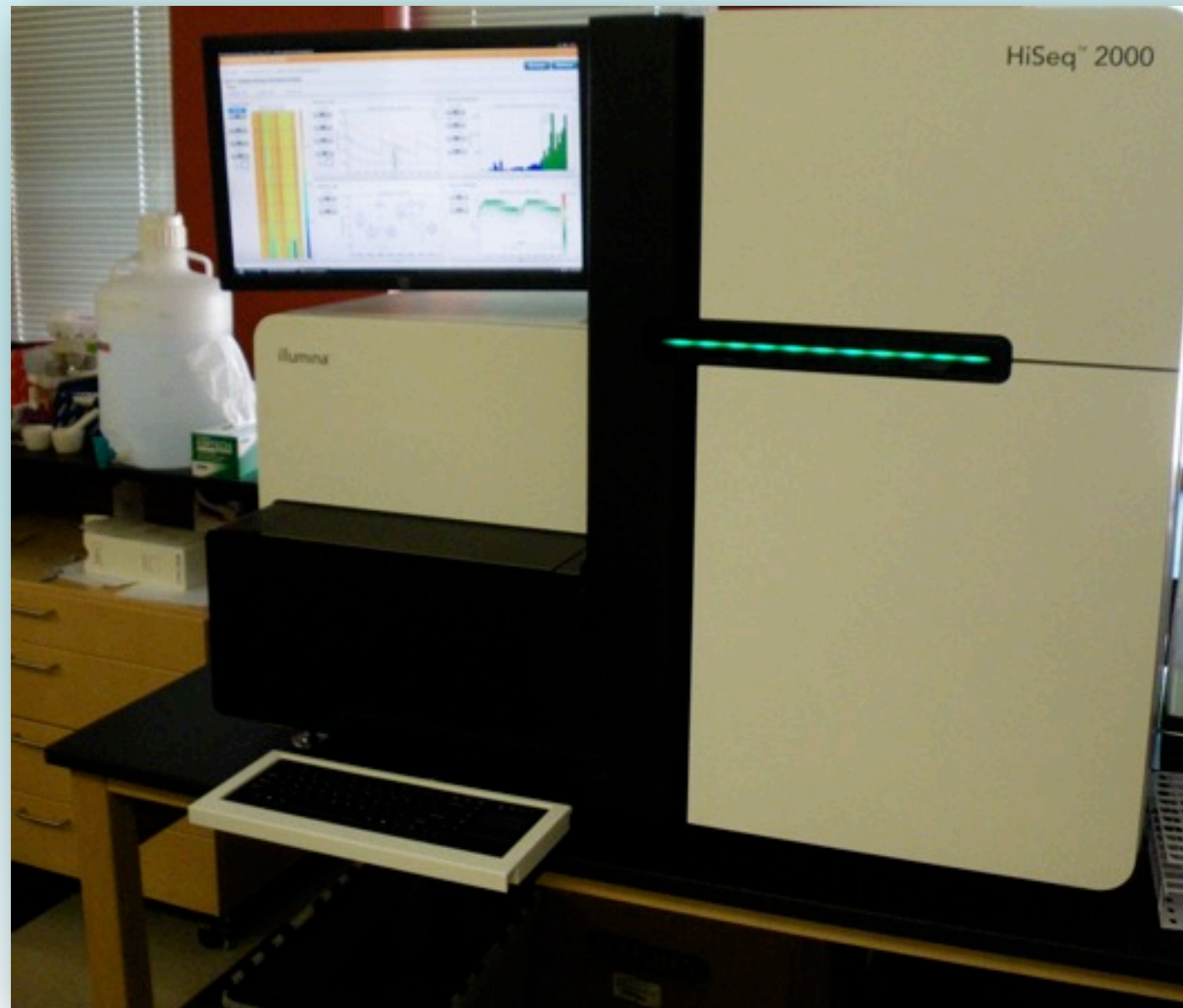
# High-throughput Data Acquisition



In 11 Days

- Generates 4TB of raw data
- 600,000,000,000 bases of DNA sequence

# High-throughput Data Acquisition



In 11 Days

- Generates 4TB of raw data
- 600,000,000,000 bases of DNA sequence
  - (200 human genomes)





# High-throughput Phenotyping (Watching Grass Grow)

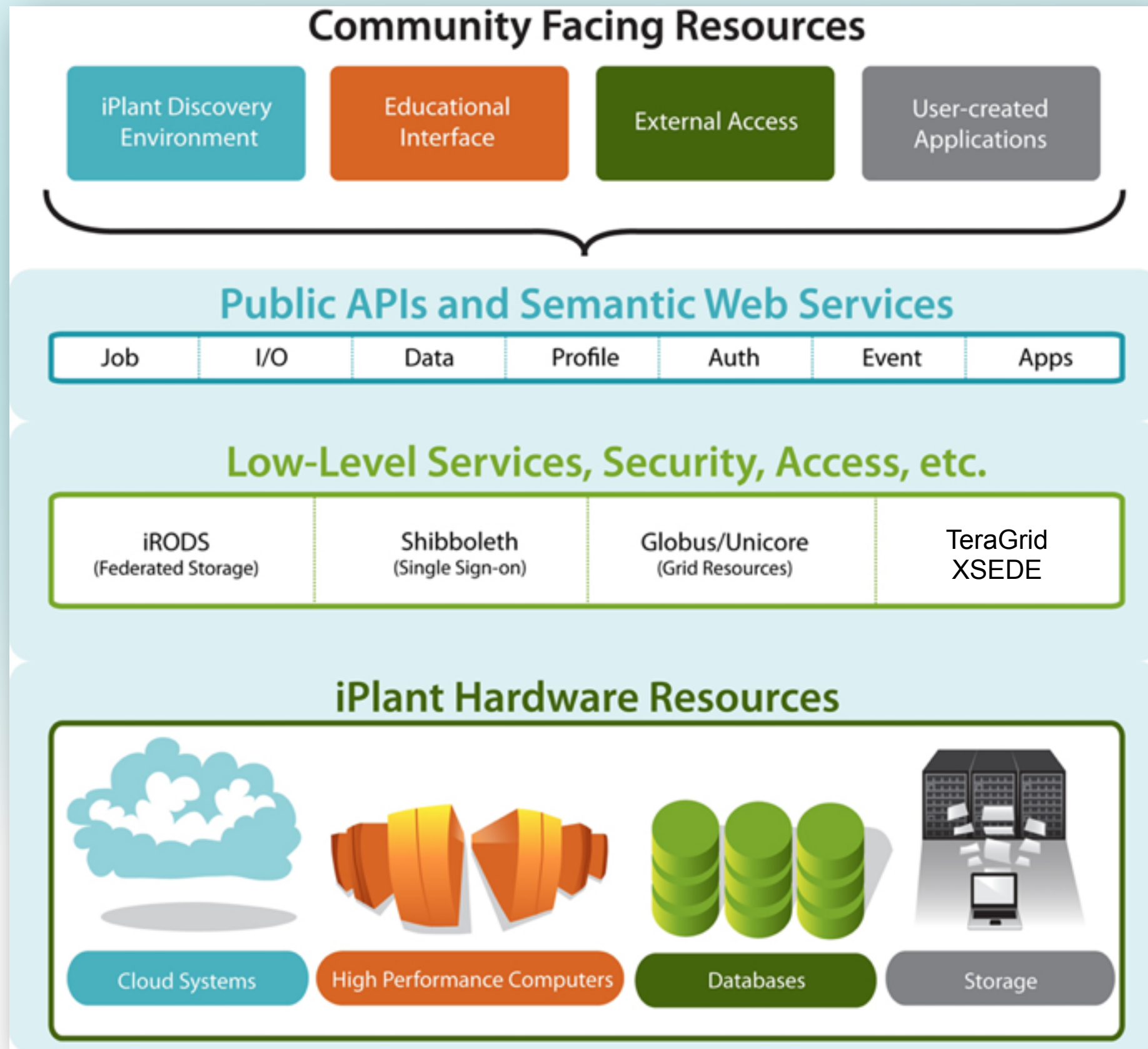


- \$70K for~30 camera sets
- ~200 movies of plants undergoing a dynamic growth process
- “Only” 4GB a day





# iPlant Layered Services and Access

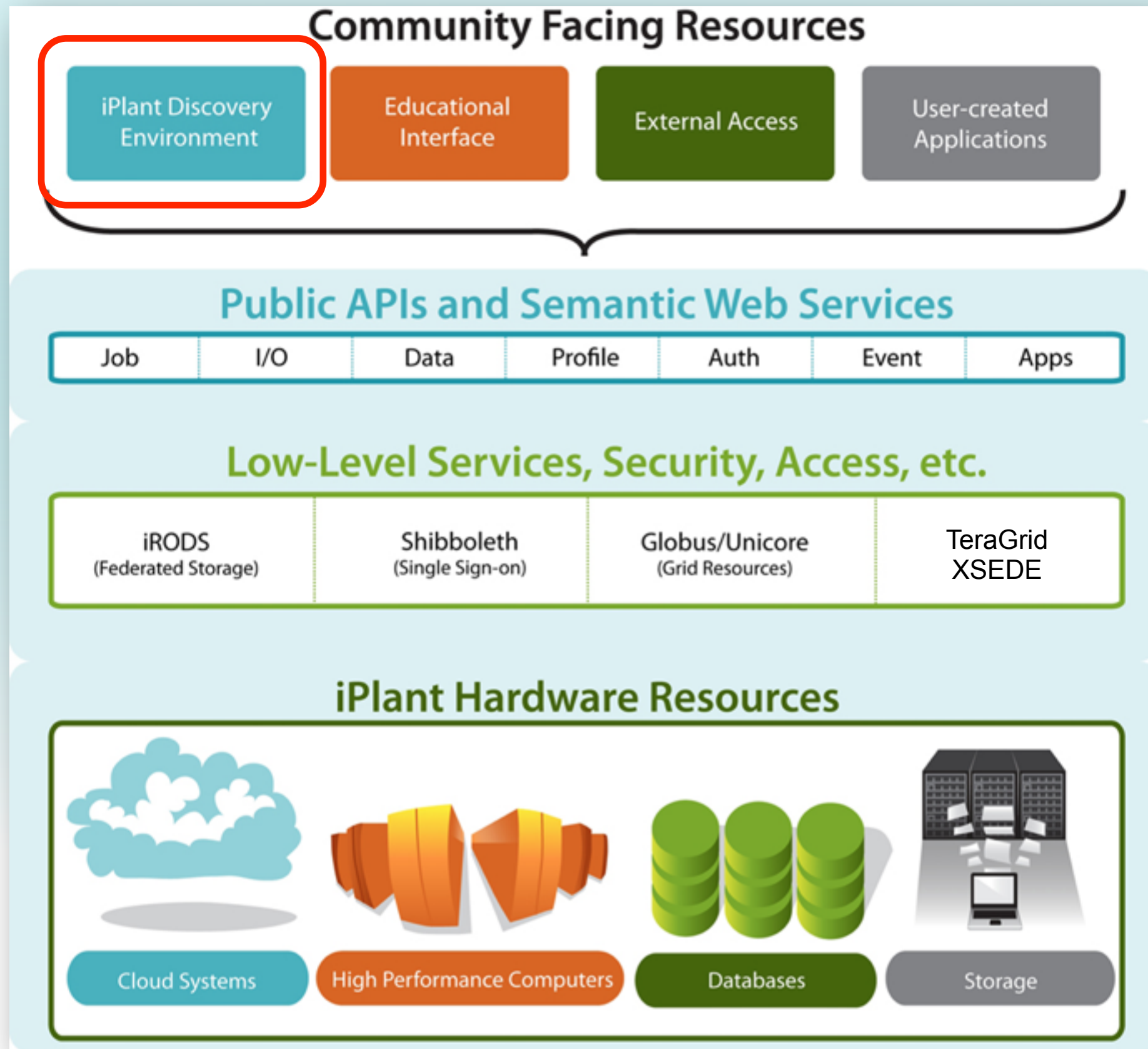


End  
Users

Computational  
Users



# iPlant Layered Services and Access



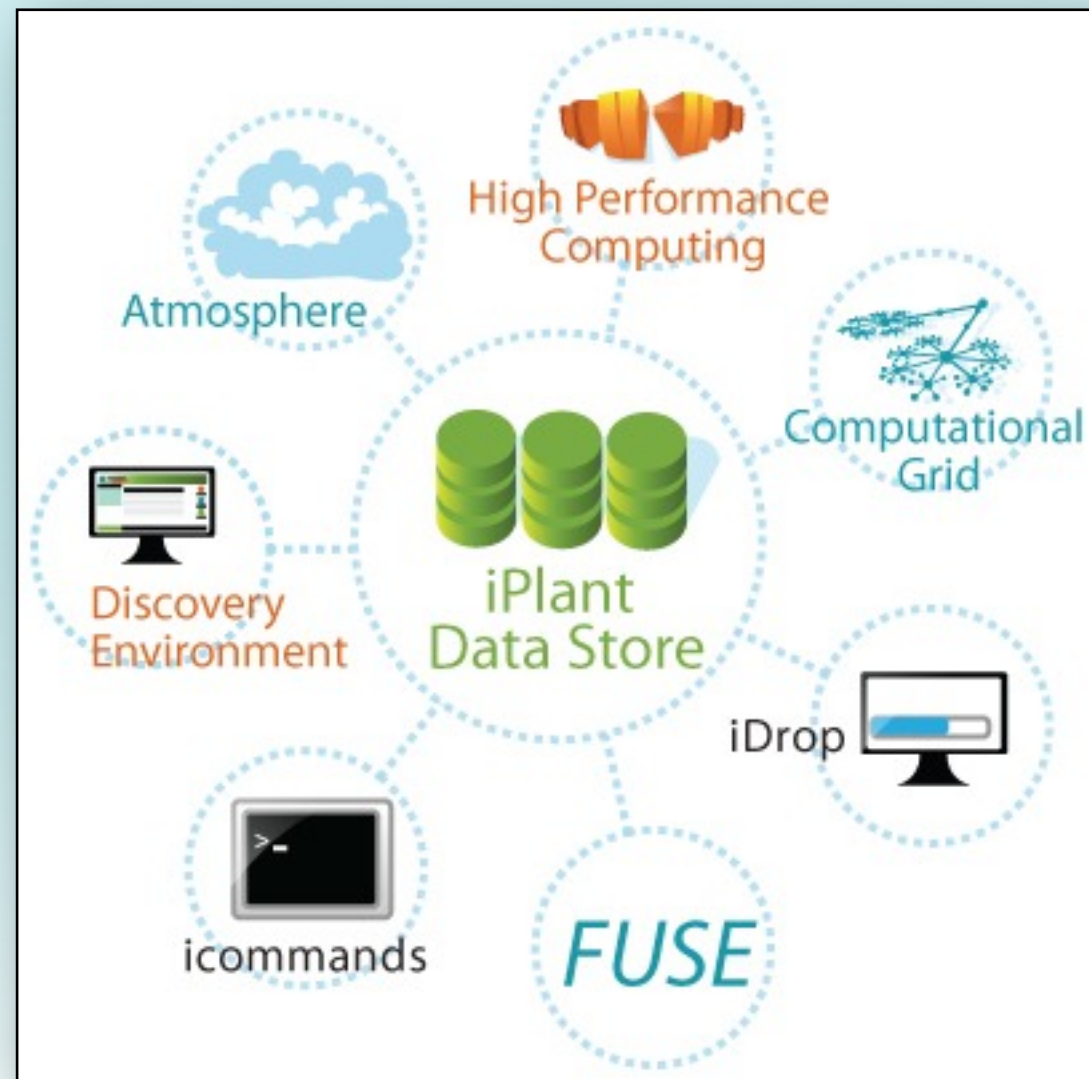
End  
Users

Computational  
Users



# iPlant Data Store

## Free Your Data



Different Users,  
Different Access Needs:  
One Data Store



# iPlant Data Store

## Free Your Data

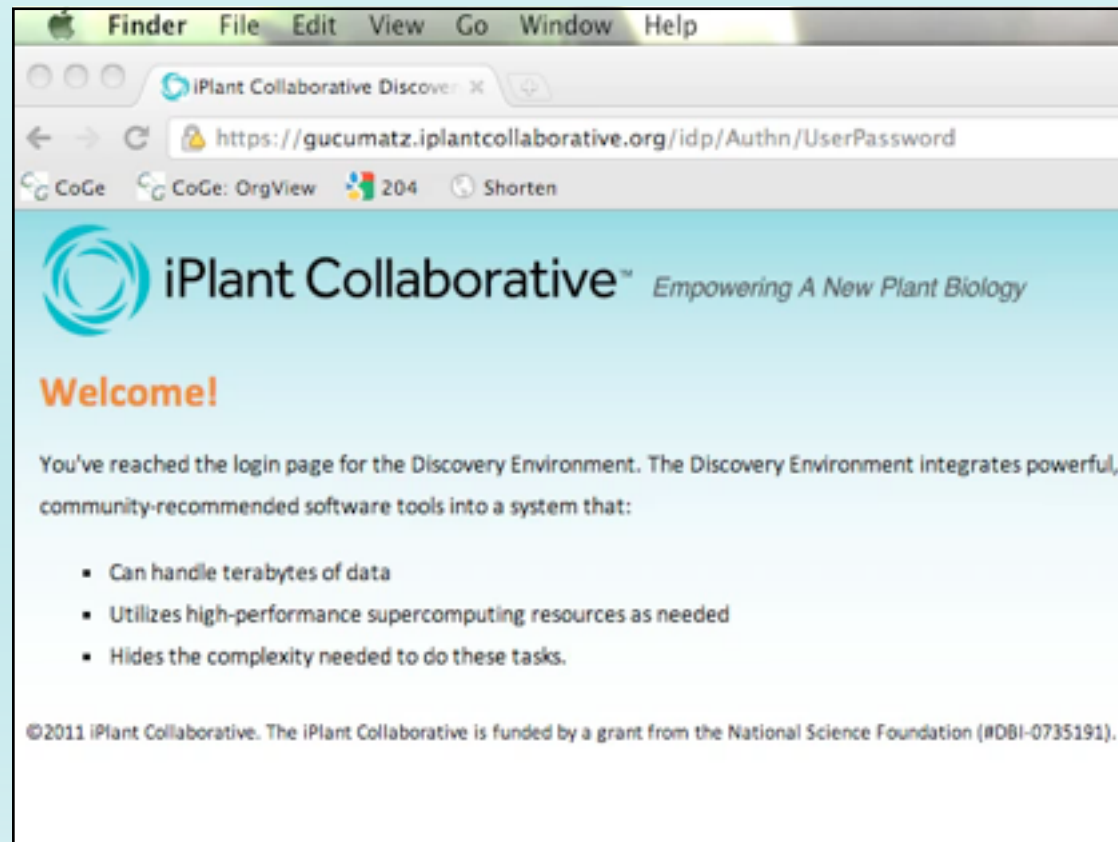




# WebDAV

# iPlant Data Store

## Free Your Data

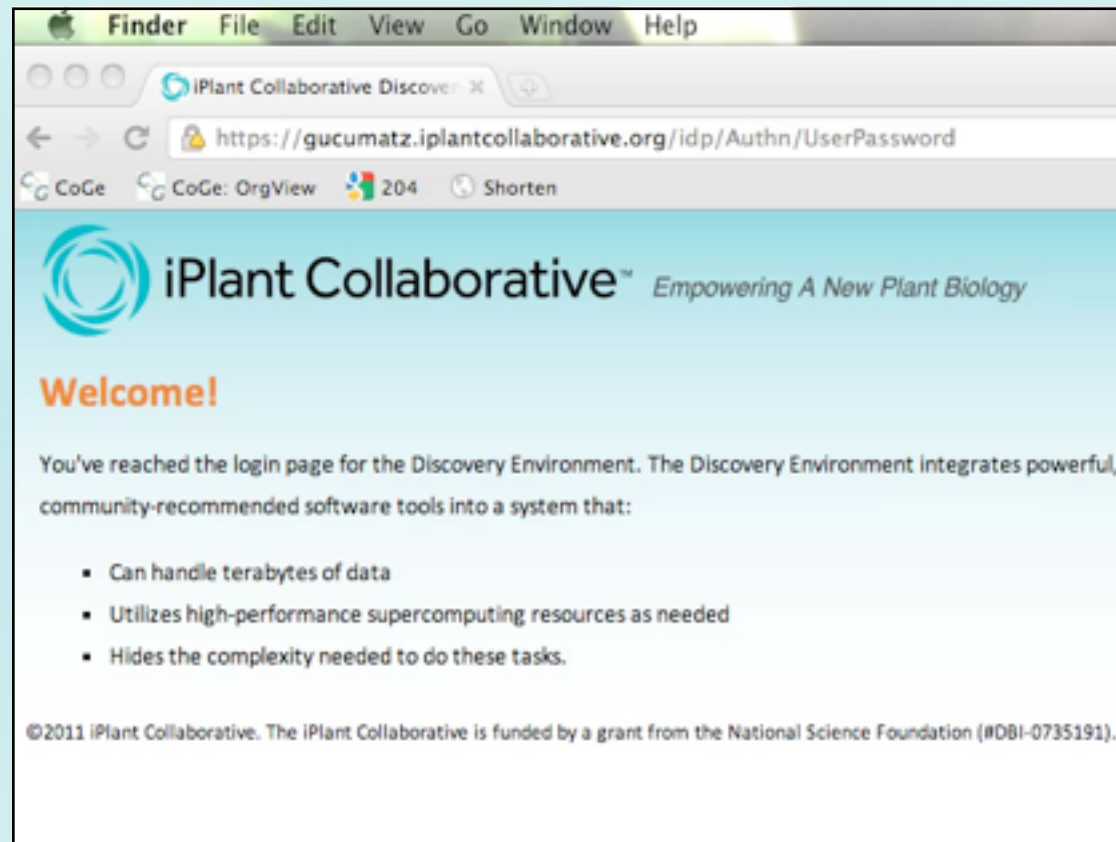


# iPlant Data Store

WebDAV

Free Your Data

DE

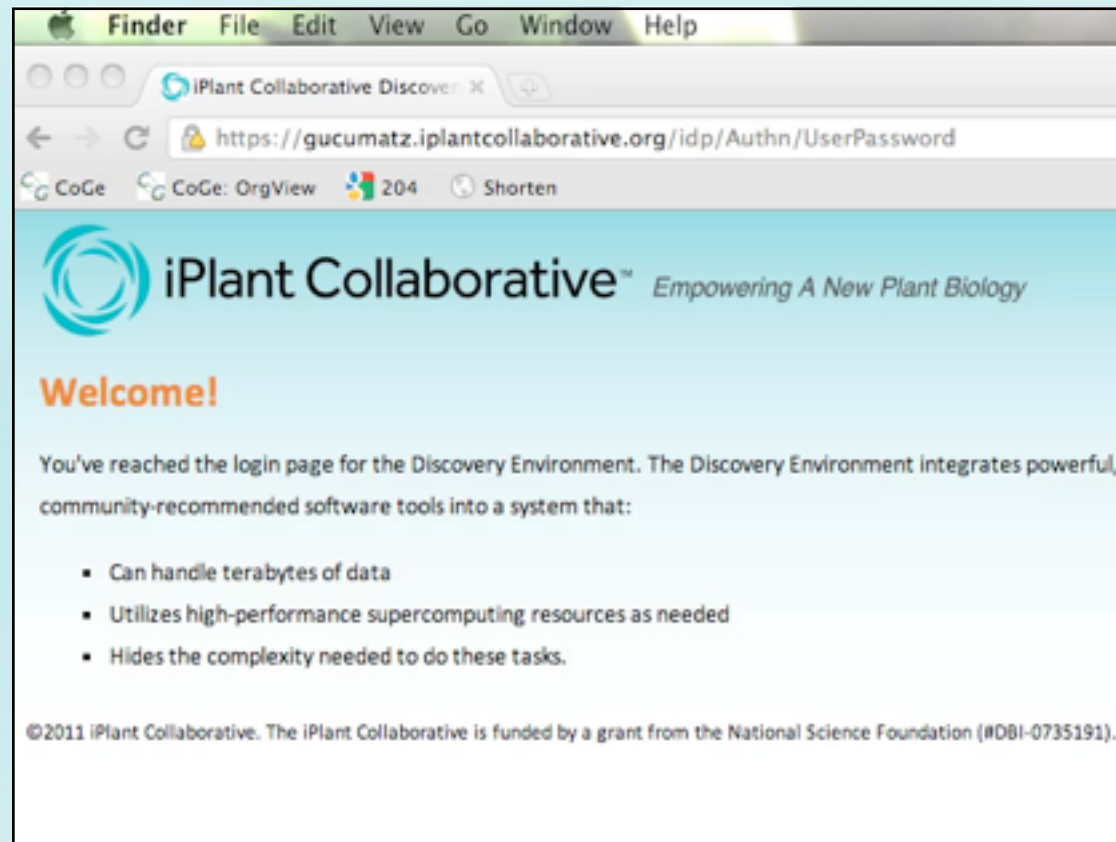


# iPlant Data Store

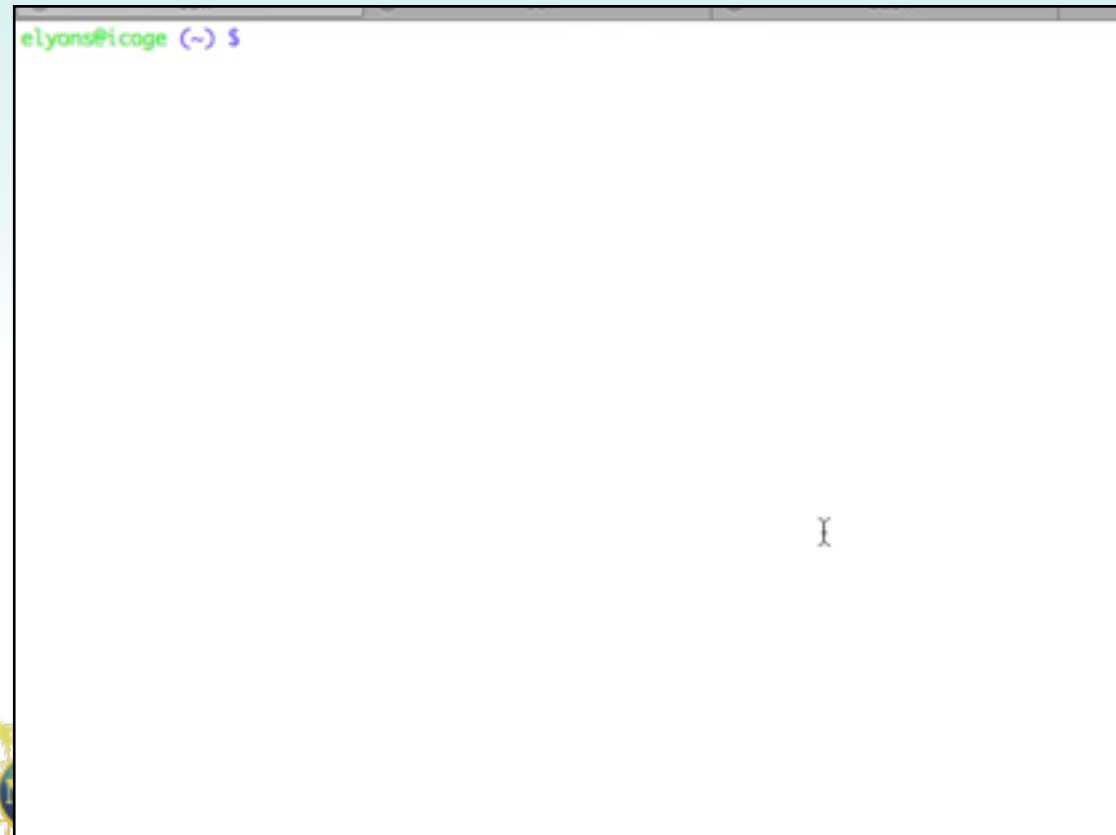
WebDAV

Free Your Data

DE



i-commands

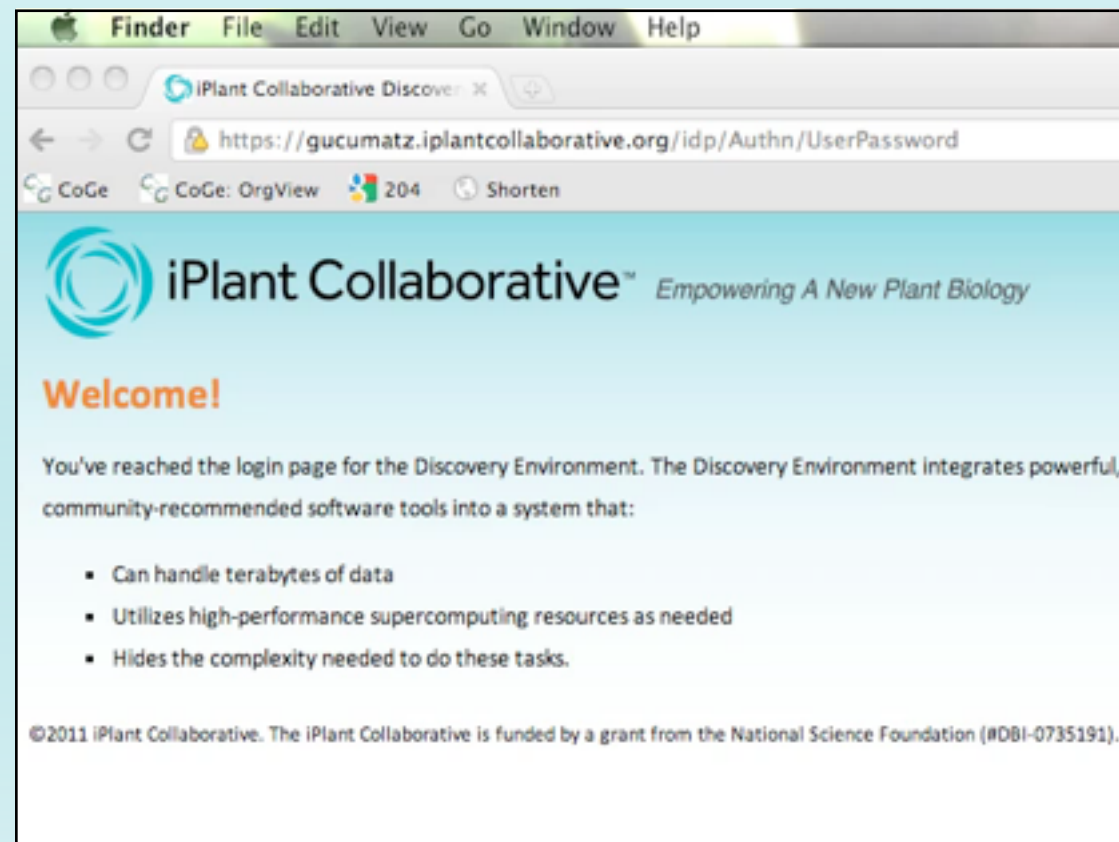


# iPlant Data Store

WebDAV

Free Your Data

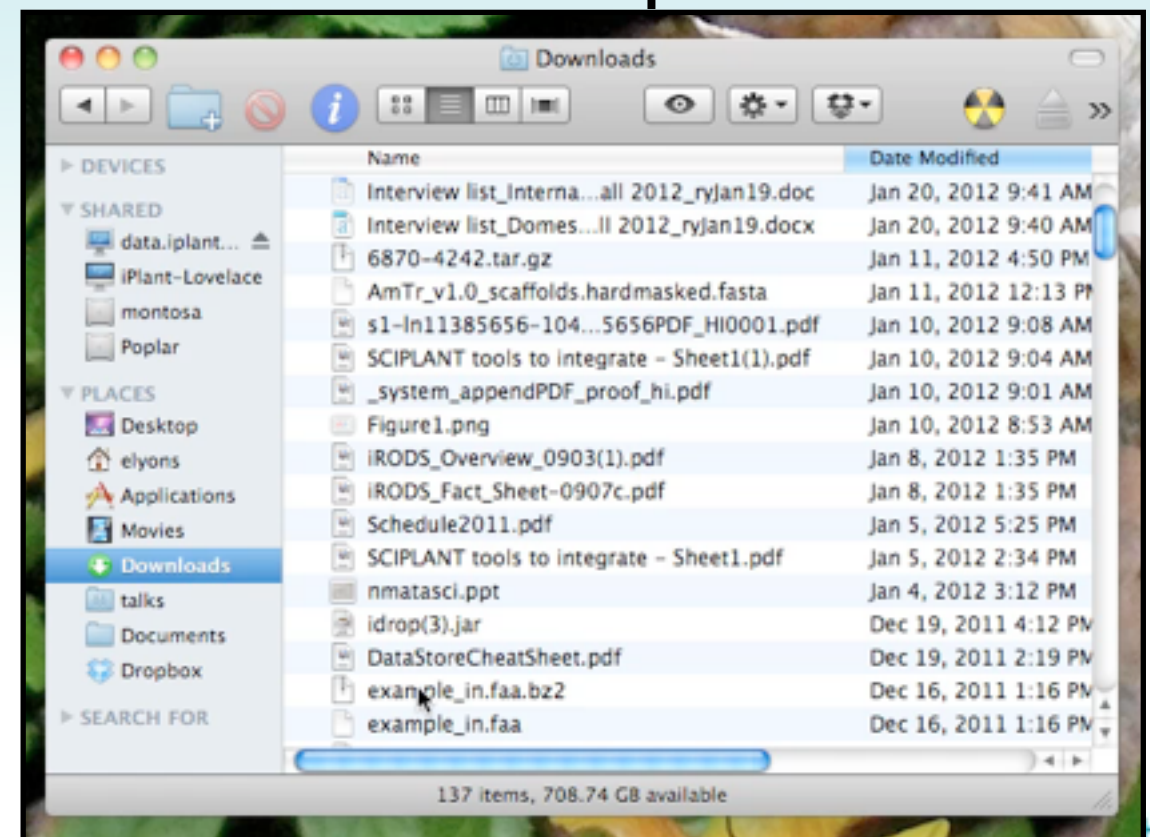
DE



i-commands



iDrop





# iPlant Data Store

WebDAV

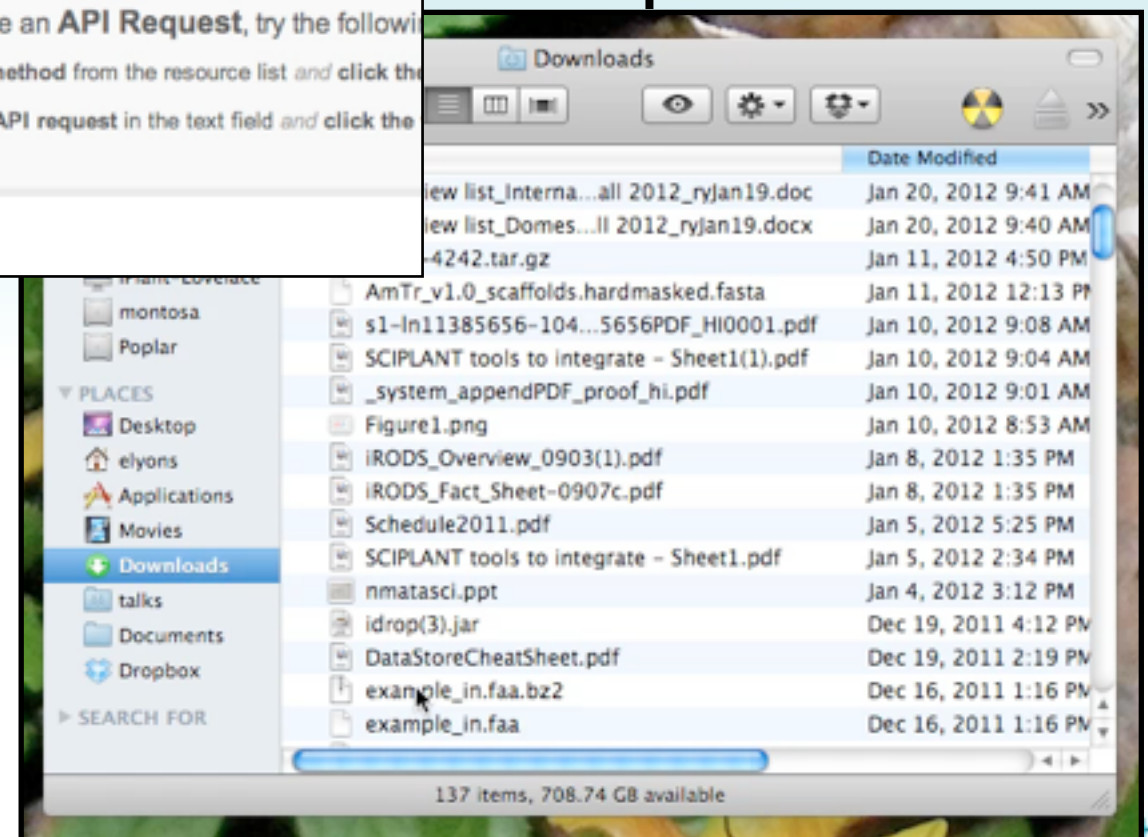
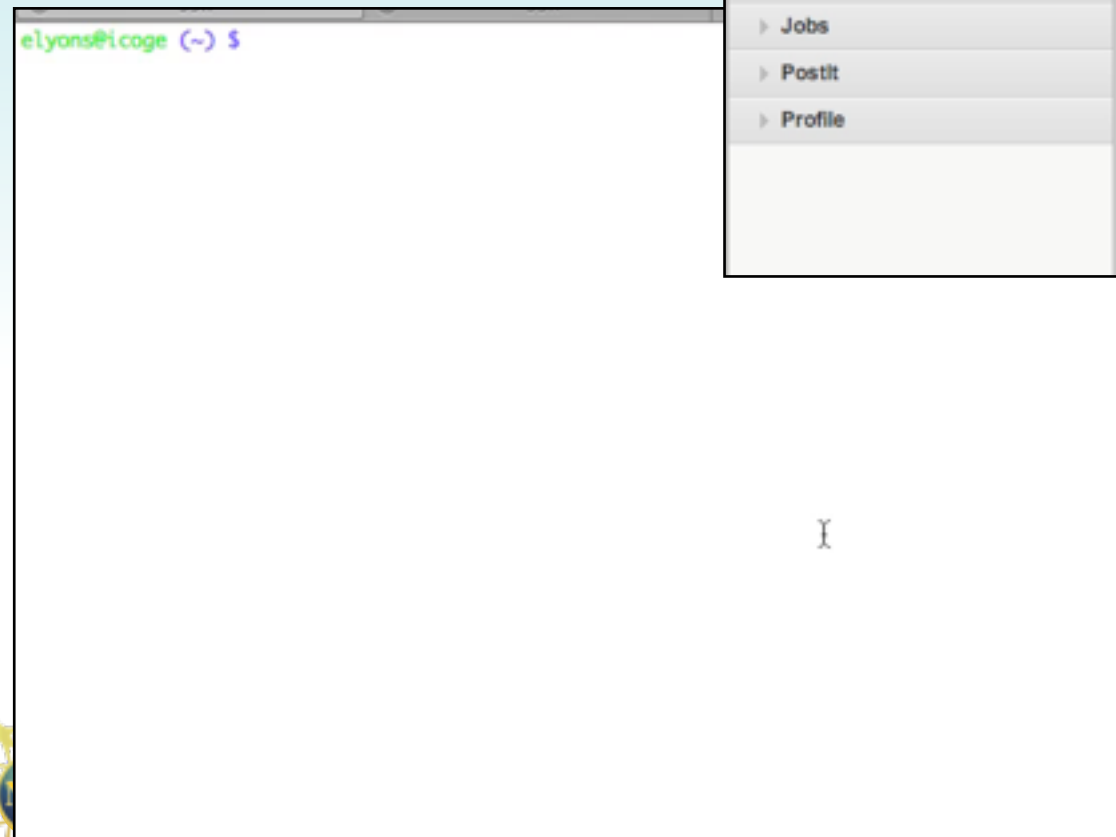
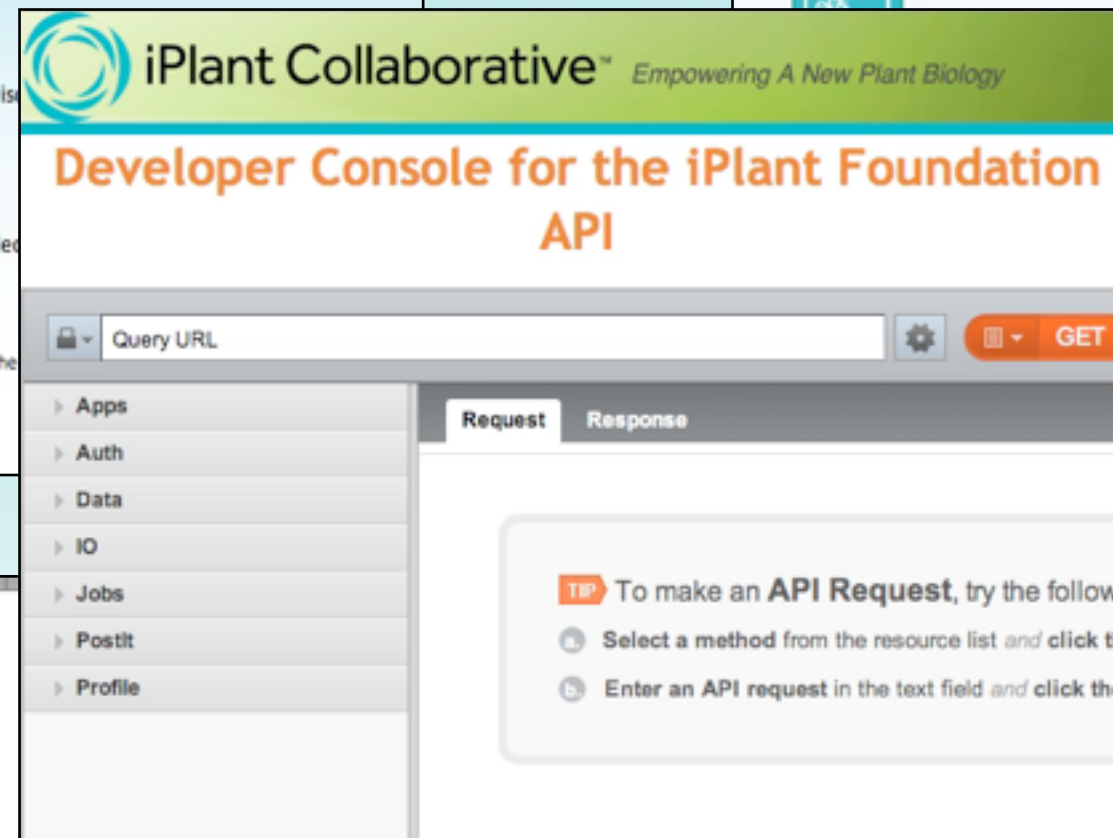
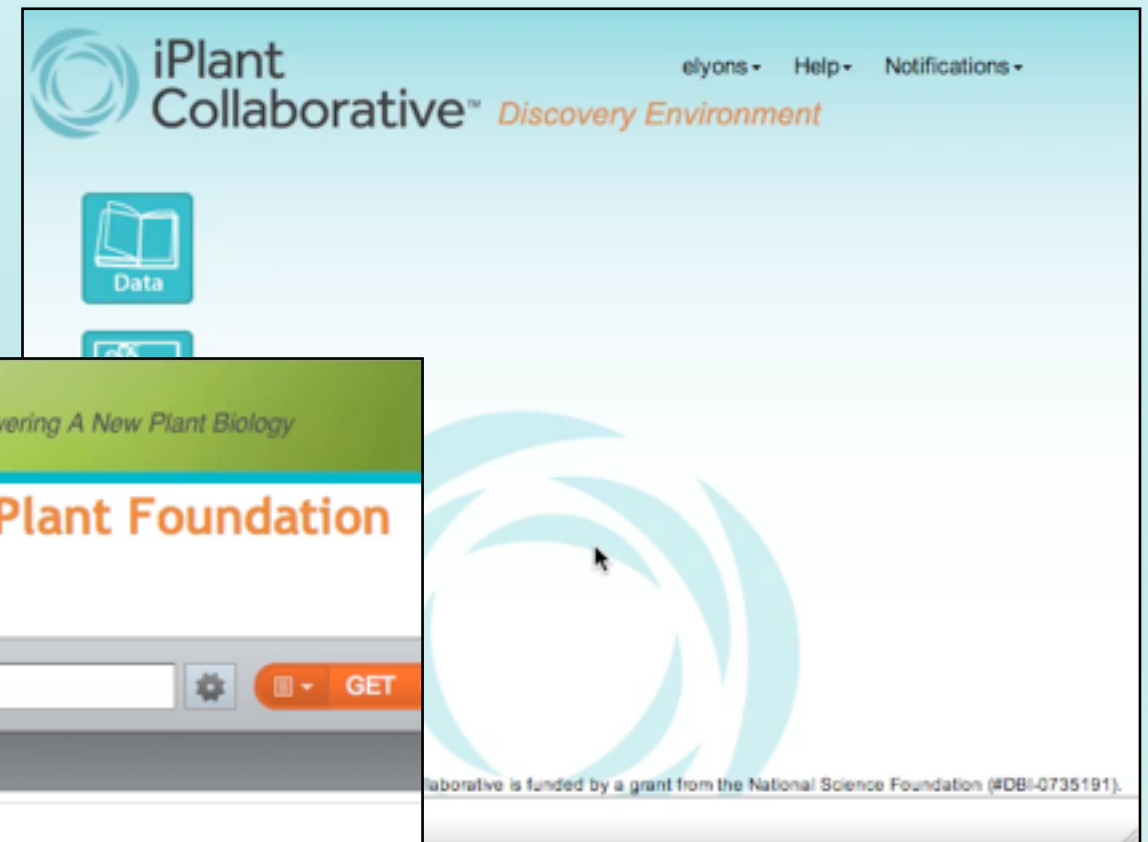
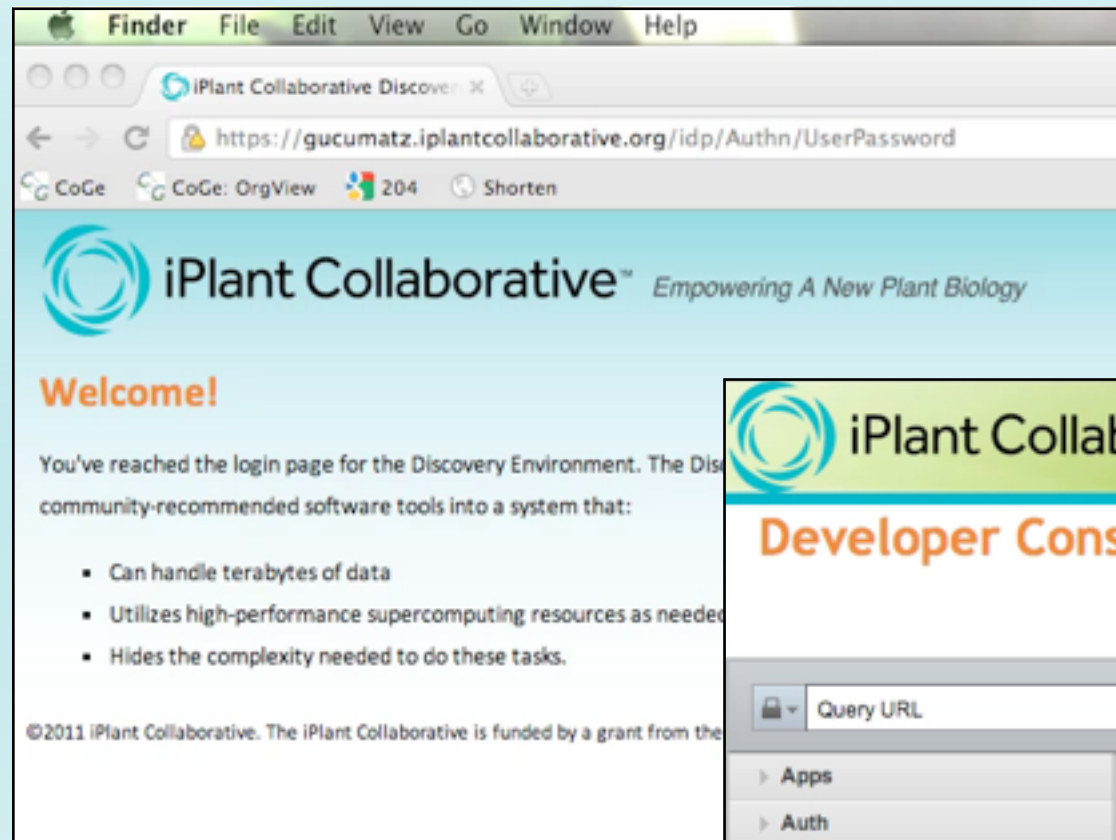
Free Your Data

DE

API

iDrop

i-commands



# iPlant Data Store

WebDAV

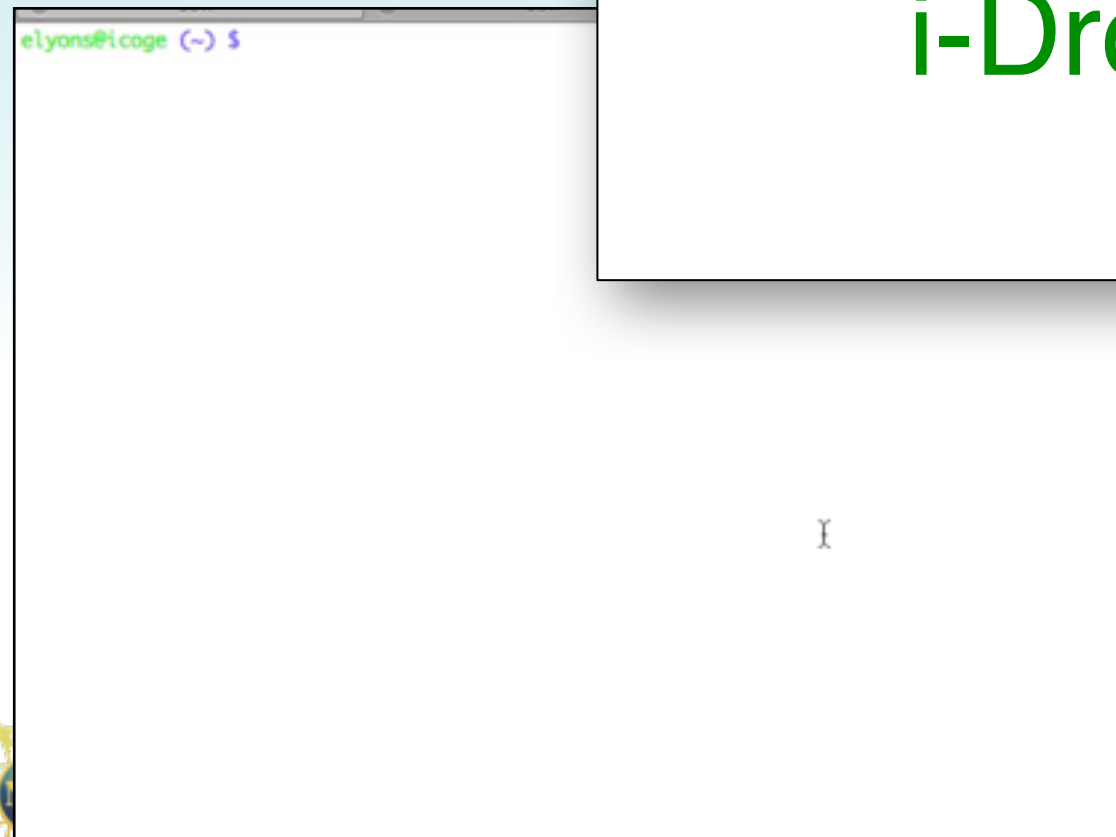
Free Your Data

DE

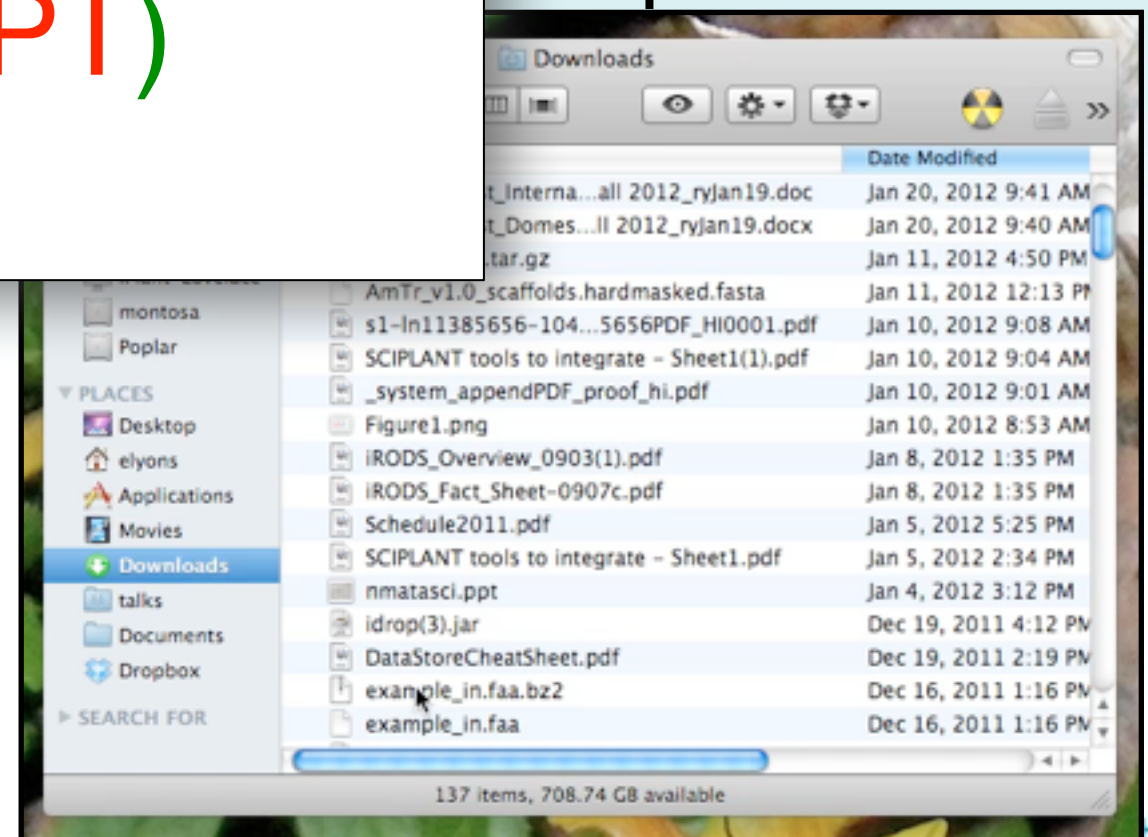


iPlant Data Store  
Desktop Folder  
Discovery Environment  
Command Line (HPT)  
i-Drop (HPT)  
API

i-command



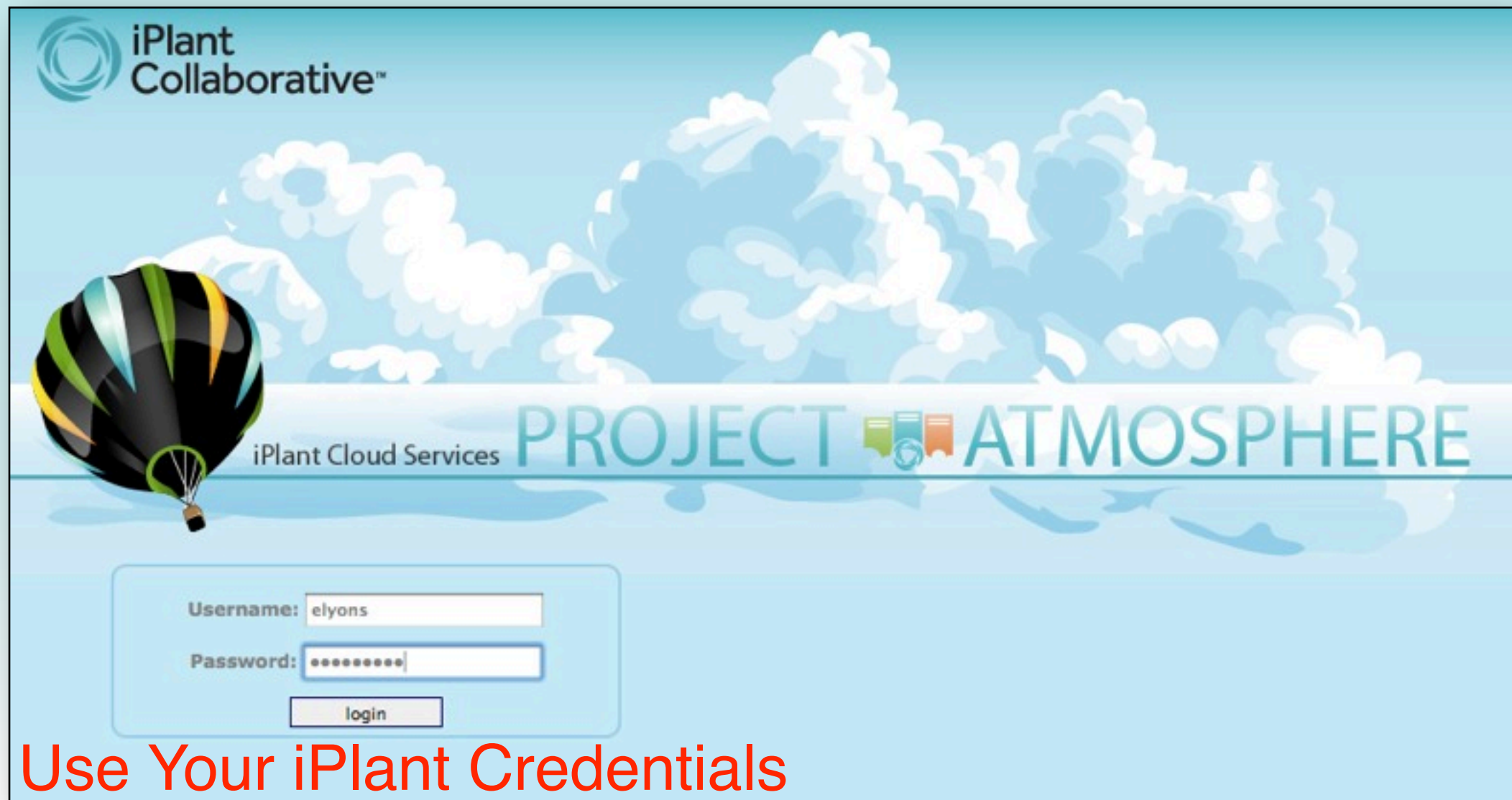
Drop





# Atmosphere: Servers and Software on Demand

<http://atmosphere.iplantcollaborative.org>



iPlant Collaborative™

iPlant Cloud Services

PROJECT ATMOSPHERE

Username: elyons

Password: [masked]

login

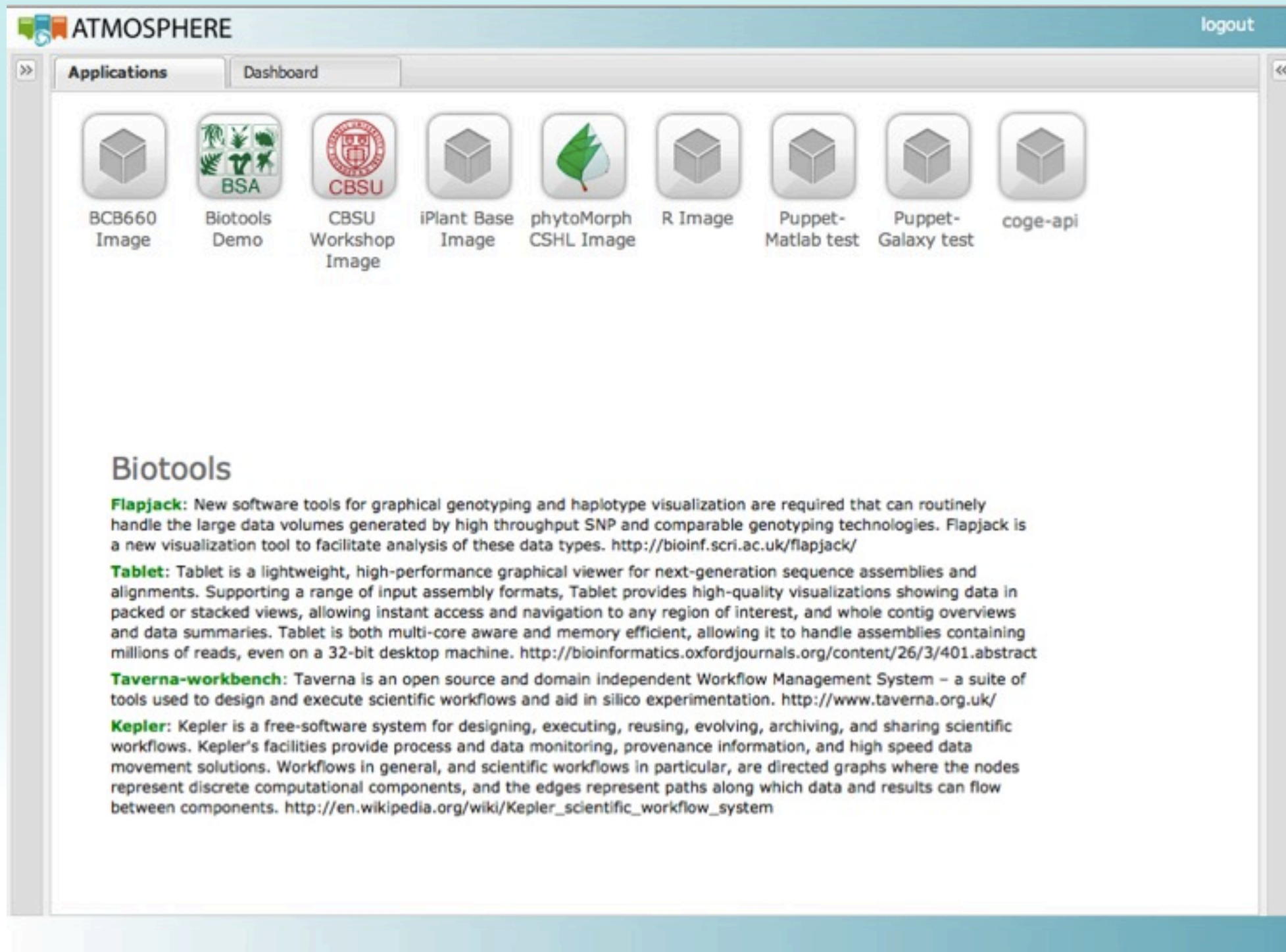
Use Your iPlant Credentials





# Atmosphere

<http://atmosphere.iplantcollaborative.org>



**ATMOSPHERE** logout

**Applications** Dashboard

BCB660 Image   Biotools Demo   CBSU Workshop Image   iPlant Base Image   phytoMorph CSHL Image   R Image   Puppet-Matlab test   Puppet-Galaxy test   coge-api

### Biotools

**Flapjack:** New software tools for graphical genotyping and haplotype visualization are required that can routinely handle the large data volumes generated by high throughput SNP and comparable genotyping technologies. Flapjack is a new visualization tool to facilitate analysis of these data types. <http://bioinf.scri.ac.uk/flapjack/>

**Tablet:** Tablet is a lightweight, high-performance graphical viewer for next-generation sequence assemblies and alignments. Supporting a range of input assembly formats, Tablet provides high-quality visualizations showing data in packed or stacked views, allowing instant access and navigation to any region of interest, and whole contig overviews and data summaries. Tablet is both multi-core aware and memory efficient, allowing it to handle assemblies containing millions of reads, even on a 32-bit desktop machine. <http://bioinformatics.oxfordjournals.org/content/26/3/401.abstract>

**Taverna-workbench:** Taverna is an open source and domain independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation. <http://www.taverna.org.uk/>

**Kepler:** Kepler is a free-software system for designing, executing, reusing, evolving, archiving, and sharing scientific workflows. Kepler's facilities provide process and data monitoring, provenance information, and high speed data movement solutions. Workflows in general, and scientific workflows in particular, are directed graphs where the nodes represent discrete computational components, and the edges represent paths along which data and results can flow between components. [http://en.wikipedia.org/wiki/Kepler\\_scientific\\_workflow\\_system](http://en.wikipedia.org/wiki/Kepler_scientific_workflow_system)



# Atmosphere

<http://atmosphere.iplantcollaborative.org>

The screenshot shows the Atmosphere web application interface. At the top, there is a header bar with the "ATMOSPHERE" logo on the left and a "logout" link on the right. Below the header, there are two tabs: "Applications" (selected) and "Dashboard". The main content area displays a grid of application icons, each with a label below it: BCB660 Image, Biotools Demo, CBSU Workshop Image, iPlant Base Image, phytoMorph CSHL Image, R Image, Puppet-Matlab test, Puppet-Galaxy test, and coge-api. A modal dialog box titled "Message" is overlaid on the interface, containing the text: "Successfully initiated your app. Please wait 10 to 15 min to finish the request. Your App id is i-3C8206D5". Below the message, there is an "OK" button. In the background, the "Biotools" section is partially visible, featuring descriptions for Flapjack, Tablet, Taverna-workbench, and Kepler.

**Message**

Successfully initiated your app.  
Please wait 10 to 15 min to finish the request.  
Your App id is **i-3C8206D5**

**Biotools**

**Flapjack:** New software tools for graph handling the large data volumes generated by a new visualization tool to facilitate analysis.

**Tablet:** Tablet is a lightweight, high-performance tool for visualizing genomic alignments. Supporting a range of input formats, Tablet allows users to view data in packed or stacked views, allowing instant access and navigation to any region of interest, and whole contig overviews and data summaries. Tablet is both multi-core aware and memory efficient, allowing it to handle assemblies containing millions of reads, even on a 32-bit desktop machine. <http://bioinformatics.oxfordjournals.org/content/26/3/401.abstract>

**Taverna-workbench:** Taverna is an open source and domain independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation. <http://www.taverna.org.uk/>

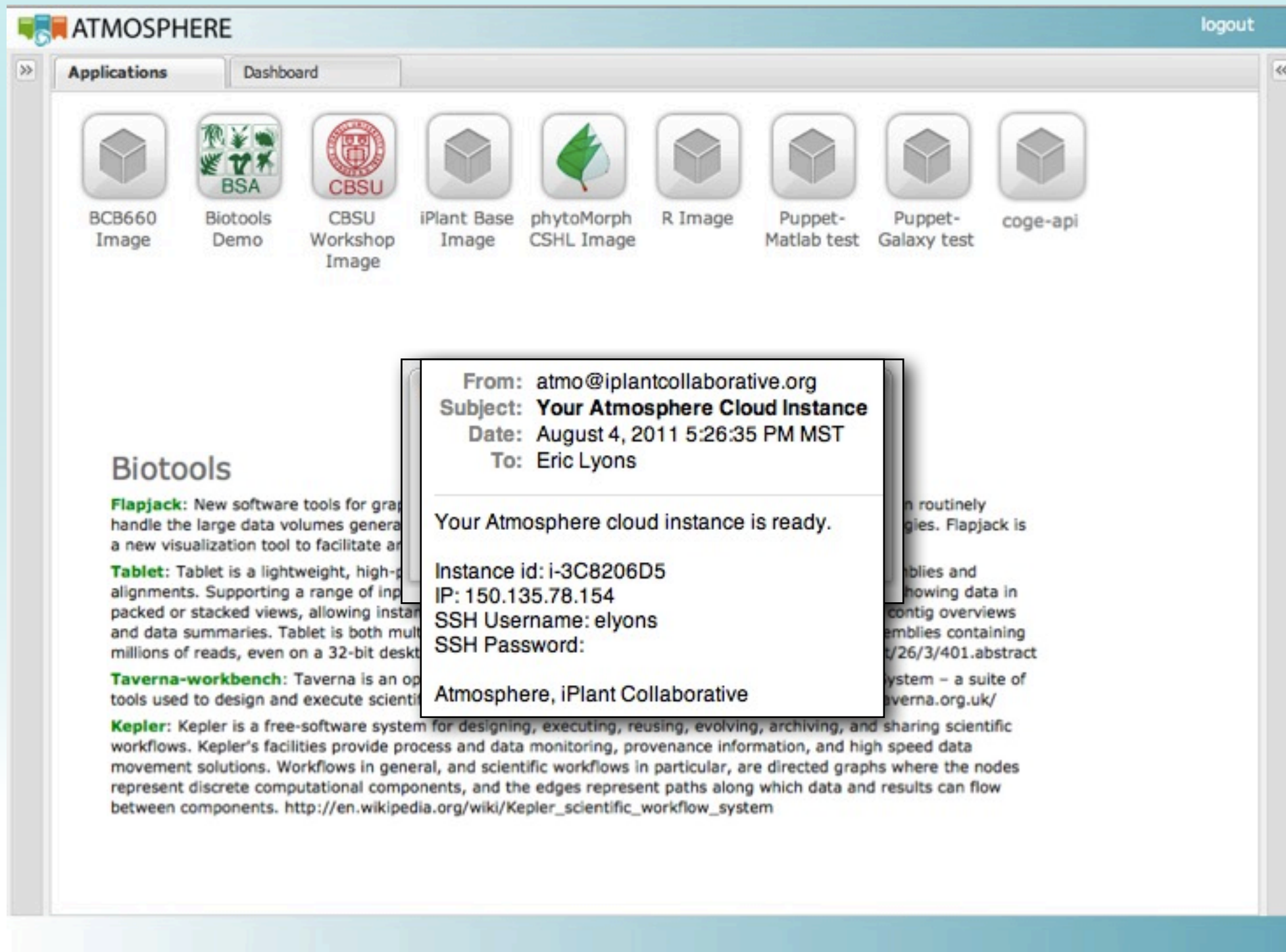
**Kepler:** Kepler is a free-software system for designing, executing, reusing, evolving, archiving, and sharing scientific workflows. Kepler's facilities provide process and data monitoring, provenance information, and high speed data movement solutions. Workflows in general, and scientific workflows in particular, are directed graphs where the nodes represent discrete computational components, and the edges represent paths along which data and results can flow between components. [http://en.wikipedia.org/wiki/Kepler\\_scientific\\_workflow\\_system](http://en.wikipedia.org/wiki/Kepler_scientific_workflow_system)





# Atmosphere

<http://atmosphere.iplantcollaborative.org>



The screenshot shows the ATMOSPHERE web interface. At the top, there's a header with the ATMOSPHERE logo and a 'logout' link. Below the header, there are two tabs: 'Applications' and 'Dashboard'. The 'Applications' tab is active, displaying a grid of application icons. The icons are labeled: BCB660 Image, Biotools Demo, CBSU Workshop Image, iPlant Base Image, phytoMorph CSHL Image, R Image, Puppet-Matlab test, Puppet-Galaxy test, and coge-api. Below the grid, there's a section titled 'Biotools' with descriptions for Flapjack, Tablet, Taverna-workbench, and Kepler. A pop-up window is overlaid on the Biotools section, displaying an email notification from atmo@iplantcollaborative.org with the subject 'Your Atmosphere Cloud Instance'. The email content states that the cloud instance is ready and provides the instance ID, IP address, SSH username, and SSH password. The email is signed 'Atmosphere, iPlant Collaborative'.

**From:** atmo@iplantcollaborative.org  
**Subject:** Your Atmosphere Cloud Instance  
**Date:** August 4, 2011 5:26:35 PM MST  
**To:** Eric Lyons

Your Atmosphere cloud instance is ready.

Instance id: i-3C8206D5  
IP: 150.135.78.154  
SSH Username: elyons  
SSH Password:

Atmosphere, iPlant Collaborative

**Biotools**

**Flapjack:** New software tools for graph handling the large data volumes generated by a new visualization tool to facilitate analysis.

**Tablet:** Tablet is a lightweight, high-performance tool for visualizing genomic alignments. Supporting a range of input data, Tablet can display genomic data in packed or stacked views, allowing instant zooming and data summaries. Tablet is both multi-platform and multi-user, capable of displaying millions of reads, even on a 32-bit desktop.

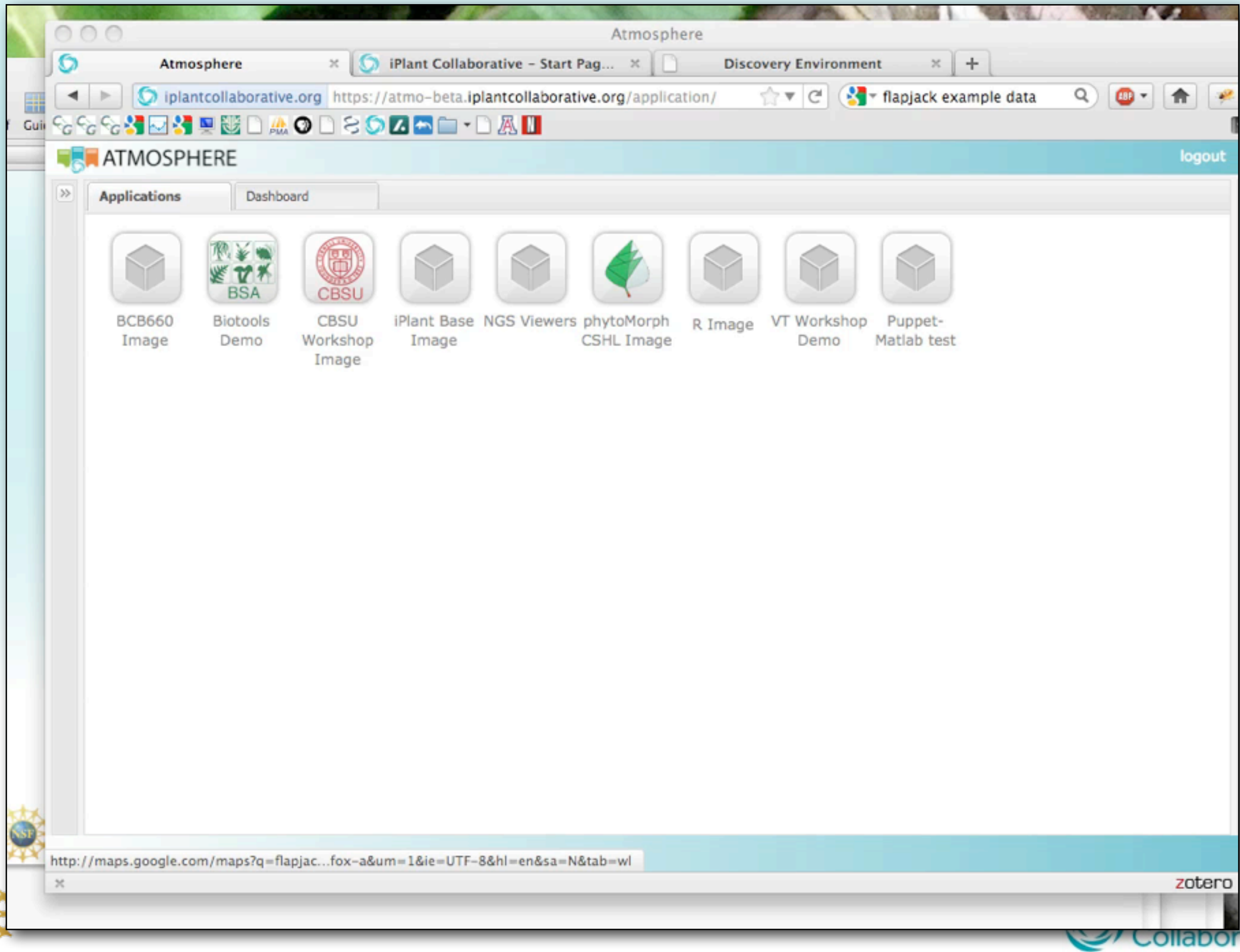
**Taverna-workbench:** Taverna is an open-source system for designing and executing scientific workflows.

**Kepler:** Kepler is a free-software system for designing, executing, reusing, evolving, archiving, and sharing scientific workflows. Kepler's facilities provide process and data monitoring, provenance information, and high speed data movement solutions. Workflows in general, and scientific workflows in particular, are directed graphs where the nodes represent discrete computational components, and the edges represent paths along which data and results can flow between components. [http://en.wikipedia.org/wiki/Kepler\\_scientific\\_workflow\\_system](http://en.wikipedia.org/wiki/Kepler_scientific_workflow_system)



# Atmosphere

Plus VNC



# “Powered by iPlant CI”

- Make it easy for community to develop applications based on iPlant CI
- Authentication
- Data store
- Compute
- Applications can be distributed or hosted by iPlant
- Provide developer with tool kit to get started





# Training end users !!

← → ↻ <https://pods.iplantcollaborative.org/wiki/display/start/Storing+Your+Data+with+iPlant+and+Accessing+that+Data>

Atmosphere

Dashboard > Getting Started with iPlant > Getting Started with iPlant > Storing Your Data with iPlant and Accessing that Data Browse ▾ Nirav Merchant ▾

This box searches only this space. The box at the upper right searches the entire iPlant wiki.

- ☐ Getting Help with iPlant Services
- ☐ Acquiring an iPlant Account
- ☐ Accessing iPlant Services (List Included)
- ☐ Resetting Your iPlant Account Password
- ☒ **Storing Your Data with iPlant and Accessing that Data**
- ☐ List of Available Materials, Documentation, and User Help Options
- ☐ Frequently Asked Questions (FAQ)
- ☐ Reporting Issues with JIRA
- ☒ Employee Information

## Uploading data

There are several ways to upload your data. These include:

- Directly to the [Discovery Environment](#) from your desktop (for files under 2 GB)
- Directly to the [Discovery Environment](#) from a URL (for all files)
- Using the [DAVIS Web Interface](#) (for files under 2 GB) ([video tutorial](#))
- Using the [iRODS web client](#) (file size limit unknown, may also be limited by web browser limit of 2 GB)
- Using [WebDAV](#) (file size limit unknown)
- Using [iDrop](#) (file size limit unknown) ([video tutorial](#))
- Using [icommands](#) (ideal for bulk transfers and best for large data and files over 2 GB)
- From [Atmosphere](#) using FUSE (useful if you have already been working in a virtual machine (VM) environment)

*Colors indicate relative difficulty: green=easier, yellow=intermediate, red=more difficult. For most users, we recommend green, the*

## Accessing data

There are several ways to access your data. These include:

- Directly from the Discovery Environment: [viewing, managing, deleting](#)
- Using the [Davis web interface](#)
- Using the [iRODS web client](#)
- Using [WebDAV](#)
- Using [iDrop](#)
- From [Atmosphere](#) virtual machines using FUSE
- Using [icommands](#)
- Installing the [iRODS FUSE](#) client for Linux to create a mountable volume

Powered by a free **Atlassian Confluence Open Source Project License** granted to The iPlant Collaborative, University of Arizona. Evaluate


Powered by Atlassian Confluence 3.4.5, the Enterprise Wiki | Report a bug | Atlassian News





# Training end users !!

w.iplantcollaborative.org/learn/events/2012/02/02/iplant-tools-and-services-workshop-uc-davis-march-12th-and-13th-2012



**iPlant Collaborative™** Empowering A New Plant Biology

Login or Register About Contact Us Feedback

CHALLENGE DISCOVER LEARN CONNECT



Learn

Home > Learn > Events > iPlant Tools and Services Workshop, UC Davis. March 12th and 13th 2012

## iPlant Tools and Services Workshop, UC Davis. March 12th and 13th 2012

### Date and Time

March 12 2012	8:00am - 2:30pm SciLab 2060
March 13 2012	9:30am - 4:00pm Genome Center 4202

### Location

University of California, Davis

### Details

**Registration for this event is CLOSED.**

Biological research depends increasingly on high-throughput data collection methods and complex computational analyses. This **free** two-day workshop provides a comprehensive look at the tools and services provided by the iPlant Collaborative, a major cyberinfrastructure project of the National Science Foundation ([www.iplantcollaborative.org](http://www.iplantcollaborative.org)). Workshop topics build progressively to cater to the needs of general and advanced audiences.

- Events
- Education, Outreach & Training
- Manuals & Tutorials
- News
- Publications
- Media





# Training end users !!

## Agenda

### Monday, March 12th, 2012

Time	Description
08:00 AM - 08:30 AM	Arrive / Sign-in / Verify iPlant Accounts
08:30 AM - 09:00 AM	Welcome - Presenter/Participant Self-Introductions
09:00 AM - 09:20 AM	An Overview of the iPlant Collaborative
09:20 AM - 09:45 AM	Overview of the iPlant Discovery Environment (DE)
09:45 AM - 10:00 AM	Break
10:00 AM - 10:30 AM	iPlant Data Store - Managing "Big Data"
10:30 AM - 11:30 AM	Using the DE to Examine Differential Expression Within an RNA-seq Dataset
11:30 AM - 12:00 PM	Atmosphere - Custom Cloud Computing
12:00 PM - 12:30 PM	Lunch
12:30 PM - 01:00 PM	Building and Using Workflows Within the DE; Phylogenetics
01:00 PM - 01:45 PM	Extending the DE for Your Research: Tool Integration and Customization
01:45 PM - 02:30 PM	Collaborating with iPlant: Future Projects and Workshop Summary

### Tuesday, March 13th, 2012

Time	Description
09:30 AM - 11:30 AM	Advanced ChIP-seq in the iPlant Cyberinfrastructure
11:30 AM - 12:00 PM	Comparative Phylogenetic Methods in the iPlant Cyberinfrastructure
	<b>Data Clinic Appointments</b>
01:30 PM - 02:00 PM	<b>Powered by iPlant</b> - Consuming iPlant Services in Your Portals
02:00 PM - 02:30 PM	Using XSEDE for Bioinformatics



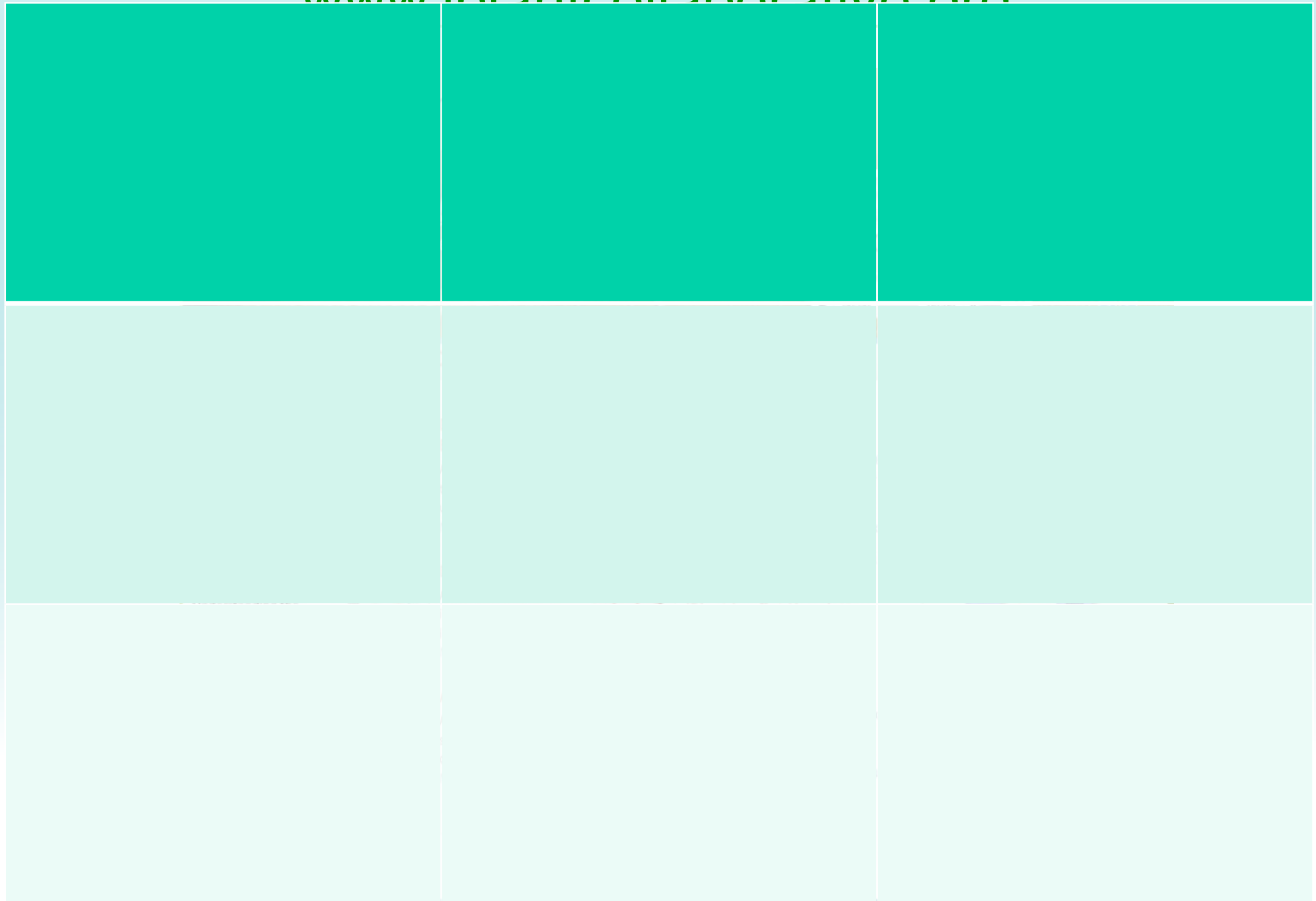
# Future directions

- More data driven collaboration interfaces (tickets/tokens\*)
- Deeper Metadata handling (extending AVU)
- Content/Repository searching (Solr)
- Data enrichment, decoration (enhanced markup, semantic integration)
- Social computing (Data de-duplication, suggestion engine)



# How to get access:

[www.iplantcollaborative.org](http://www.iplantcollaborative.org)



©2011 iPlant Collaborative

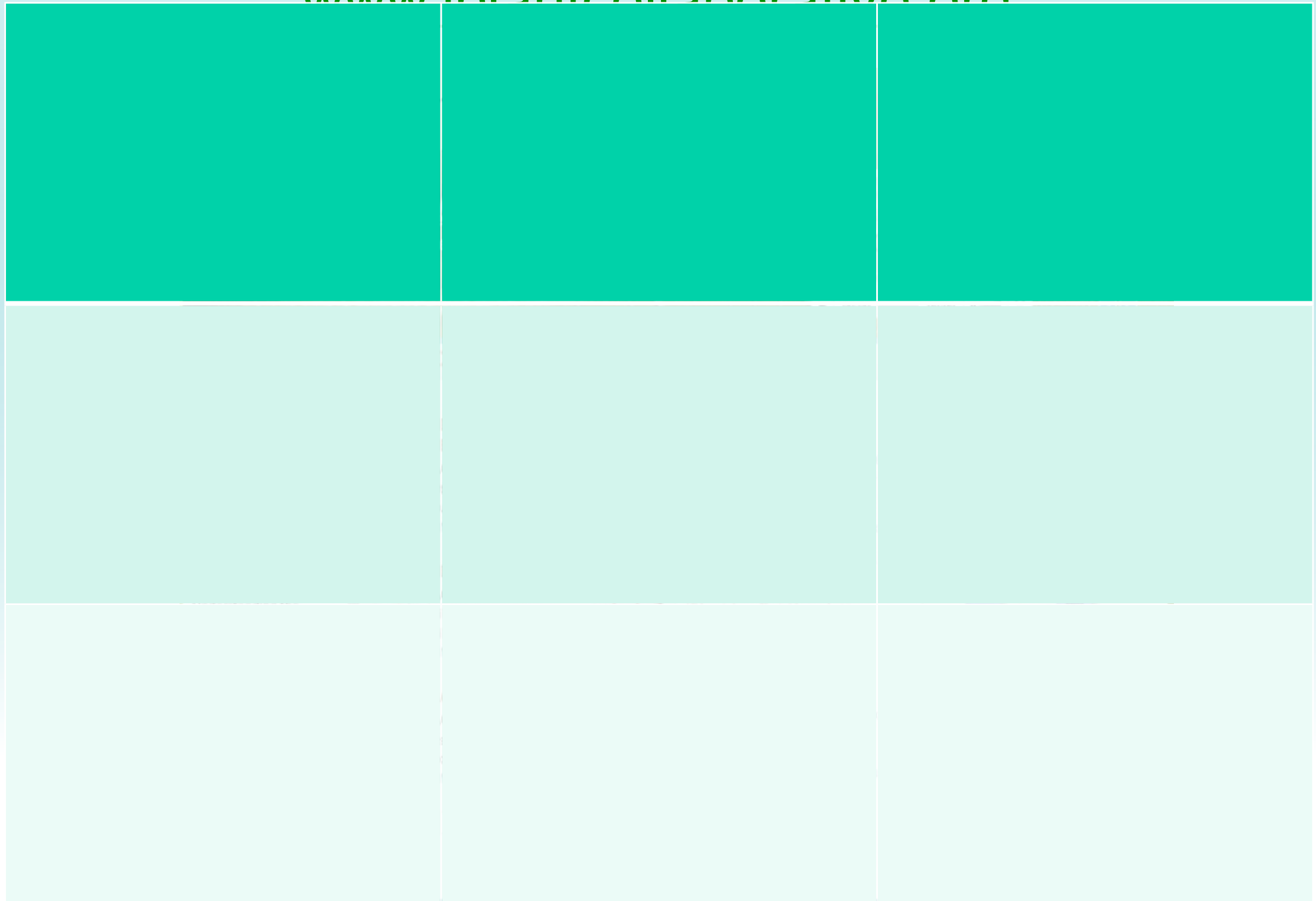
[Privacy](#) | [RSS Feed](#) | [Staff Login](#)

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



# How to get access:

[www.iplantcollaborative.org](http://www.iplantcollaborative.org)



©2011 iPlant Collaborative

[Privacy](#) | [RSS Feed](#) | [Staff Login](#)

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







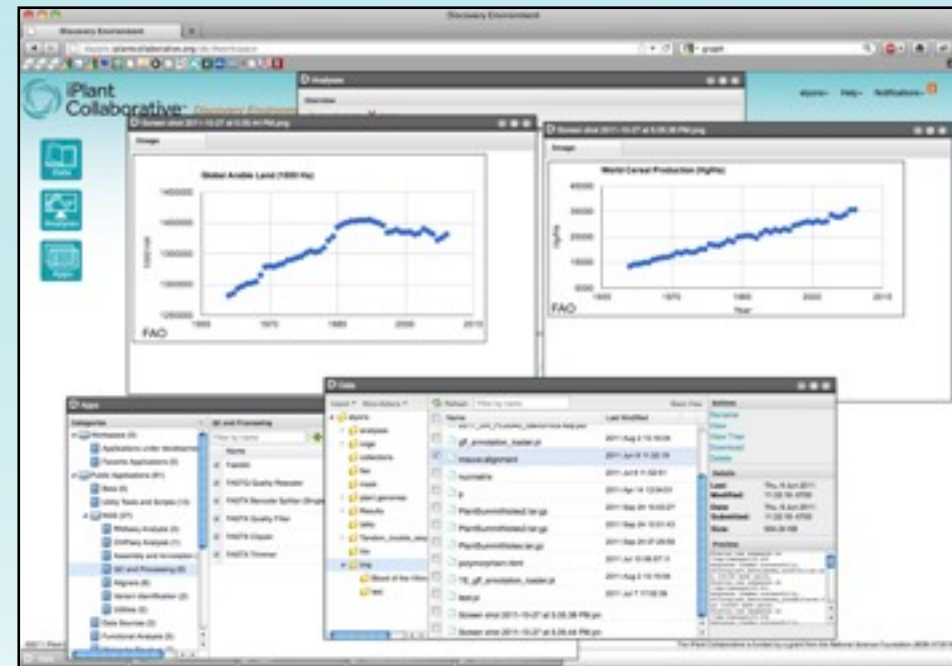
Powered By iPlant

iPlant APIs  
Resources





# Powered By iPlant

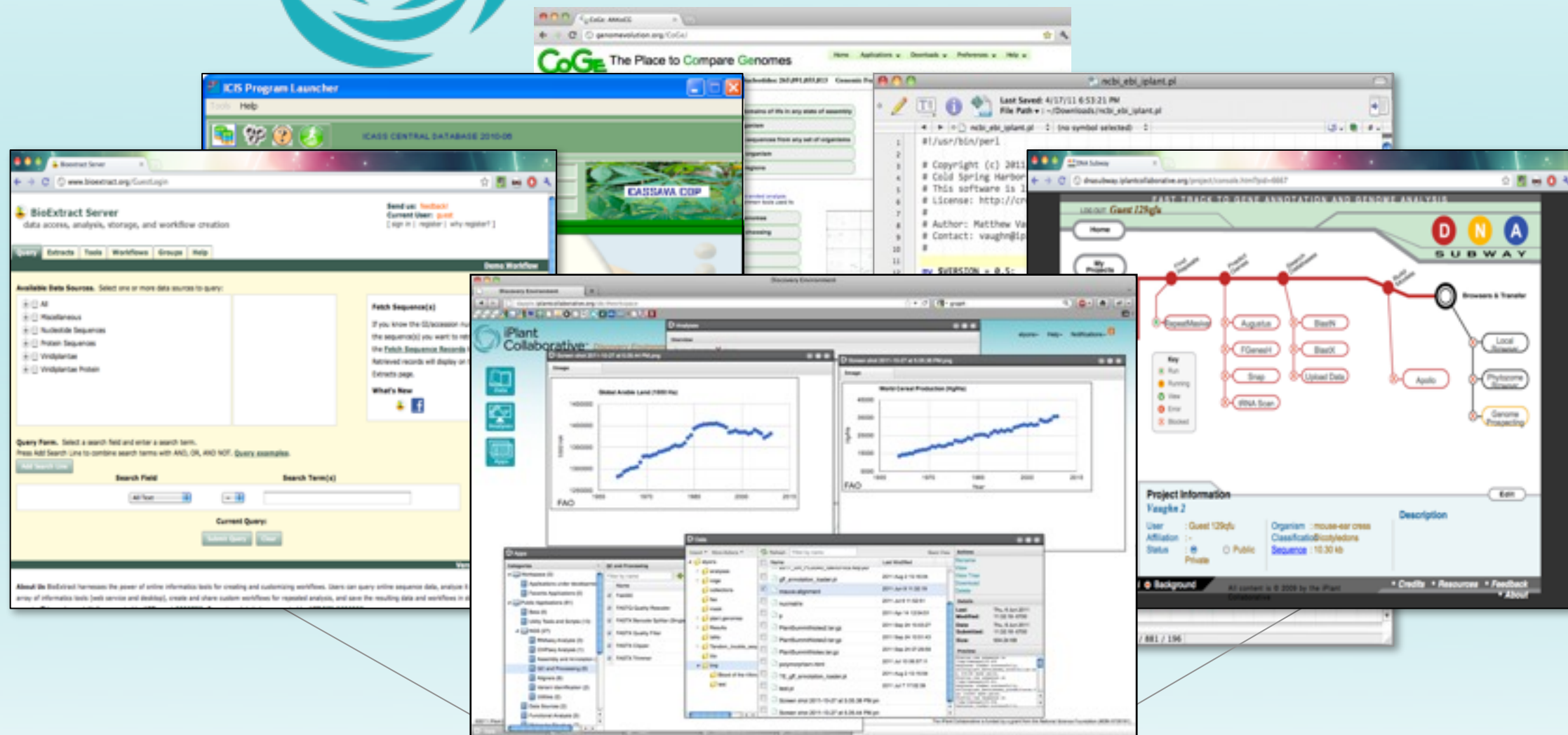


iPlant APIs  
Resources





# Powered By iPlant



iPlant APIs  
Resources







Powered By iPlant



# Cyberinfrastructure for Life Sciences

Scalable  
Capable  
Extensible



# Where to Get More Information

- Register:
- iPlant DE:
- iPlant Data Store:
- iPlant Atmo:
- iPlant Quick-Start Guides:
- iPlant Tutorials:
- iPlant Forums:
- iPlant Wiki:



# iPlant's Building Blocks

## Postdocs:

Barbara Banbury  
Jamie Estill  
Bindu Joseph  
Christos Noutsos  
Brad Ruhfel  
Stephen A. Smith  
Chunlao Tang  
Lin Wang  
Liya Wang  
Norman Wickett

## Executive Team: Steve Goff Dan Stanzione

## Students:

Peter Bailey  
Jeremy Beaulieu  
Devi Bhattacharya  
Storme Briscoe  
Ya-Di Chen  
John Donoghue  
Steven Gregory  
Yekatarina Khartianova  
Monica Lent  
Amgad Madkour  
Aniruddha Marathe  
Kurt Michaels  
Dhanesh Prasad  
Andrew Predoehl  
Jose Salcedo  
Shalini Sasidharan  
Gregory Striemer  
Jason Vandeventer  
Kuan Yang

Metadata

Data

Tools

Workflows

Viz

## Faculty Advisors & Collaborators:

Ali Akoglu  
Greg Andrews  
Kobus Barnard  
Sue Brown  
Thomas Brutnell  
Michael Donoghue  
Casey Dunn  
Brian Enquist  
Damian Gessler  
Ruth Grene  
John Hartman  
Matthew Hudson  
Dan Kliebenstein  
Jim Leebens-Mack  
David Lowenthal  
Robert Martienssen

B.S. Manjunath  
Nirav Merchant  
David Neale  
Brian O'Meara  
Sudha Ram  
David Salt  
Mark Schildhauer  
Doug Soltis  
Pam Soltis  
Edgar Spalding  
Alexis Stamatakis  
Ann Stapleton  
Lincoln Stein  
Val Tannen  
Todd Vision  
Doreen Ware  
Steve Welch  
Mark Westneat

## Staff:

Greg Abram  
Sonali Aditya  
Roger Barthelson  
Brad Boyle  
Todd Bryan  
Gordon Burleigh  
John Cazes  
Mike Conway  
Karen Cranston  
Rion Doodey  
Andy Edmonds  
Dmitry Fedorov  
Michael Gatto  
Utkarsh Gaur  
Cornel Ghiban  
Michael Gonzales  
Hariolf Häfele  
Matthew Hanlon

Anthony Heath  
Barbara Heath  
Matthew Helmke  
Natalie Henriques  
Uwe Hilgert  
Nicole Hopkins  
Eun-Sook Jeong  
Logan Johnson  
Chris Jordan  
B.D. Kim  
Kathleen Kennedy  
Mohammed Khalfan  
Seung-jin Kim  
Lars Koersterk  
Sangeeta Kuchimanchi  
Kristian Kvilekval  
Aruna Lakshmanan  
Sue Lauter  
Tina Lee

Andrew Lenards  
Zhenyuan Lu  
Eric Lyons  
Naim Matasci  
Sheldon McKay  
Robert McLay  
Angel Mercer  
Dave Micklos  
Nathan Miller  
Steve Mock  
Martha Narro  
Praveen Nuthulapati  
Shannon Oliver  
Shiran Pasternak  
William Peil  
Titus Purdin  
J.A. Raygoza Garay  
Dennis Roberts  
Jerry Schneider

Bruce Schumaker  
Sriramu Singaram  
Edwin Skidmore  
Brandon Smith  
Mary Margaret Sprinkle  
Sriram Srinivasan  
Josh Stein  
Lisa Stillwell  
Kris Urie  
Peter Van Buren  
Hans Vasquez-Gross  
Matthew Vaughn  
Fusheng Wei  
Jason Williams  
John Wregglesworth  
Weijia Xu  
Jill Yarmchuk





# iPlant's Building Blocks

## Postdocs:

Barbara Banbury  
Jamie Estill  
Bindu Joseph  
Christos Noutsos  
Brad Ruhfel  
Stephen A. Smith  
Chunlao Tang  
Lin Wang  
Liya Wang  
Norman Wickett

## Students:

Peter Bailey  
Jeremy Beaulieu  
Devi Bhattacharya  
Storme Briscoe  
Ya-Di Chen  
John Donoghue  
Stephen Gregory  
Khatianova  
Khatianova  
Khatianova

Aniruddha Marathe  
Kurt Michaels  
Dhanesh Prasad  
Andrew Predoehl  
Jose Salcedo  
Shalini Sasidharan  
Gregory Striemer  
Jason Vandeventer  
Kuan Yang

## Executive Team

Steve  
Dan

## Metadata

## Data

## Workflows

## Viz

## Faculty Advisors & Collaborators:

Ali Akoglu  
Greg Andrews  
Kobus Barnard  
Sue Brown  
Thomas Brutnell  
Michael Donoghue  
Casey Dunn  
Brian Enquist  
Damian Gessler  
Ruth Grene  
John Hartman  
Matthew Hudson  
Dan Kliebenstein  
Jim Leebens-Mack  
David Lowenthal  
Robert Martienssen

B.S. Manjunath  
Nirav Merchant  
David Neale  
Brian O'Meara  
Sudha Ram  
David Salt  
Mark Schildhauer  
Doug Soltis  
Pam Soltis  
Edgar Spalding  
Alexis Stamatakis  
Ann Stapleton  
Lincoln Stein  
Val Tannen  
Todd Vision  
Doreen Ware  
Steve Welch  
Mark Westneat

Gordon T. ...  
John ...  
Michael Gatto  
Utkarsh Gaur  
Cornel Ghiban  
Michael Gonzales  
Hariolf Häfele  
Matthew Hanlon  
Dmitry Fedorov  
Michael Gatto  
Utkarsh Gaur  
Cornel Ghiban  
Michael Gonzales  
Hariolf Häfele  
Matthew Hanlon

Andrew Lenards  
Zhanyuan Lu  
Eric Lyons  
Mam Matasci  
Gordon McKay  
McLay  
Angel Mercer  
Dave Micklos  
Nathan Miller  
Steve Mock  
Martha Narro  
Praveen Nuthulapati  
Shannon Oliver  
Shiran Pasternak  
William Peil  
Titus Purdin  
J.A. Raygoza Garay  
Dennis Roberts  
Jerry Schneider

Bruce Schumaker  
Sriramu Singaram  
Edwin Skidmore  
Brandon Smith  
Mary Margaret Sprinkle  
Sriram Srinivasan  
Josh Stein  
Lisa Stillwell  
Kris Urie  
Peter Van Buren  
Hans Vasquez-Gross  
Matthew Vaughn  
Fusheng Wei  
Jason Williams  
John Wregglesworth  
Weijia Xu  
Jill Yarmchuk



# iPlant Discovery Environment



Managing and Integrating: Data, Tools, Analysis



# iPlant Discovery Environment

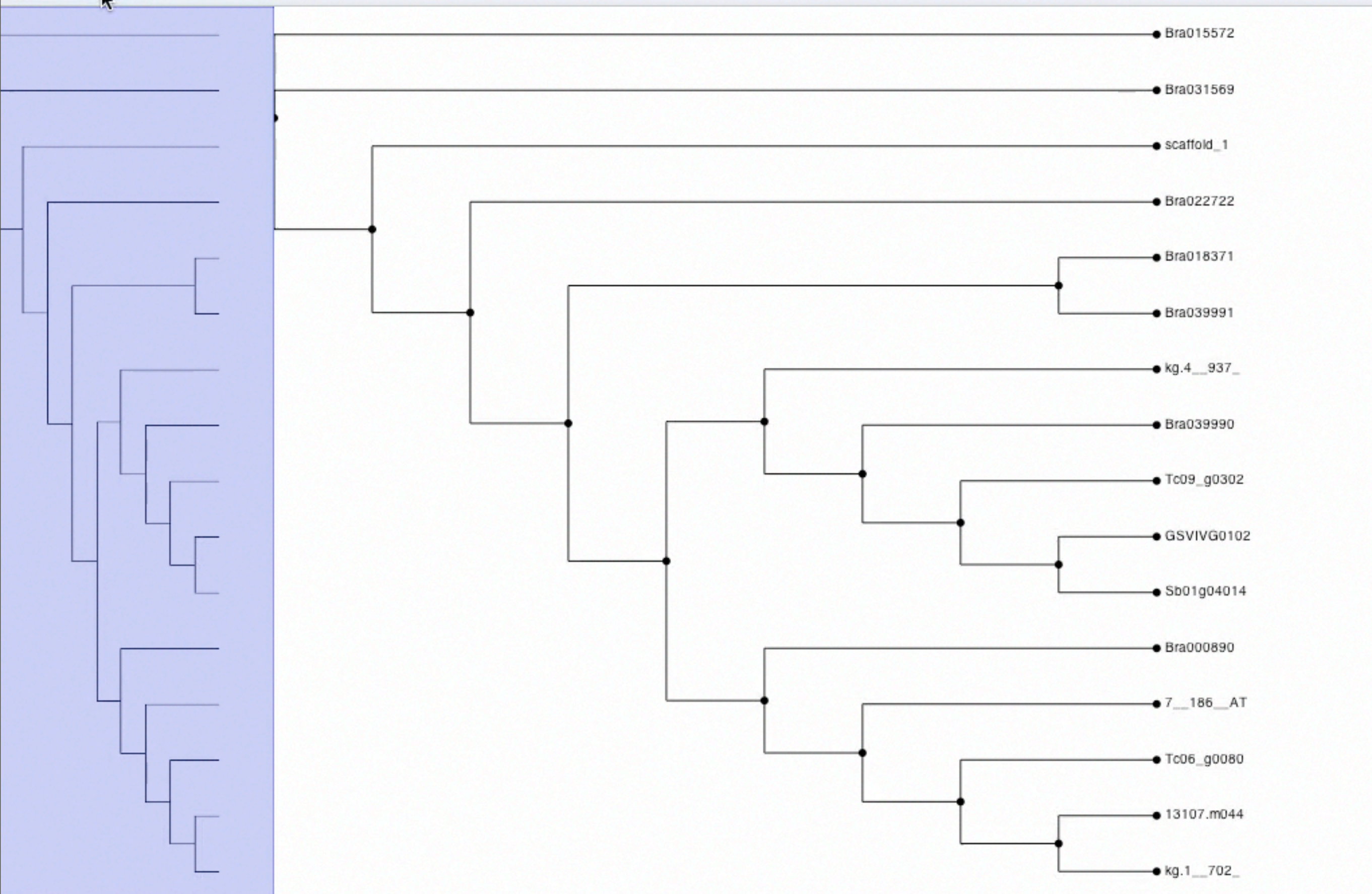


Managing and Integrating: Data, Tools, Analysis





File View

Search: