Powered by iPlant

The iPlant Collaborative as underlying infrastructure for bioinformatics projects

Plant and Animal Genome Meeting

1/16/2012

Dan Stanzione

iPlant Collaborative

Texas Advanced Computing Center

dan@tacc.utexas.edu

dan@iplantcollaborative.org

iPlant Collaborative™

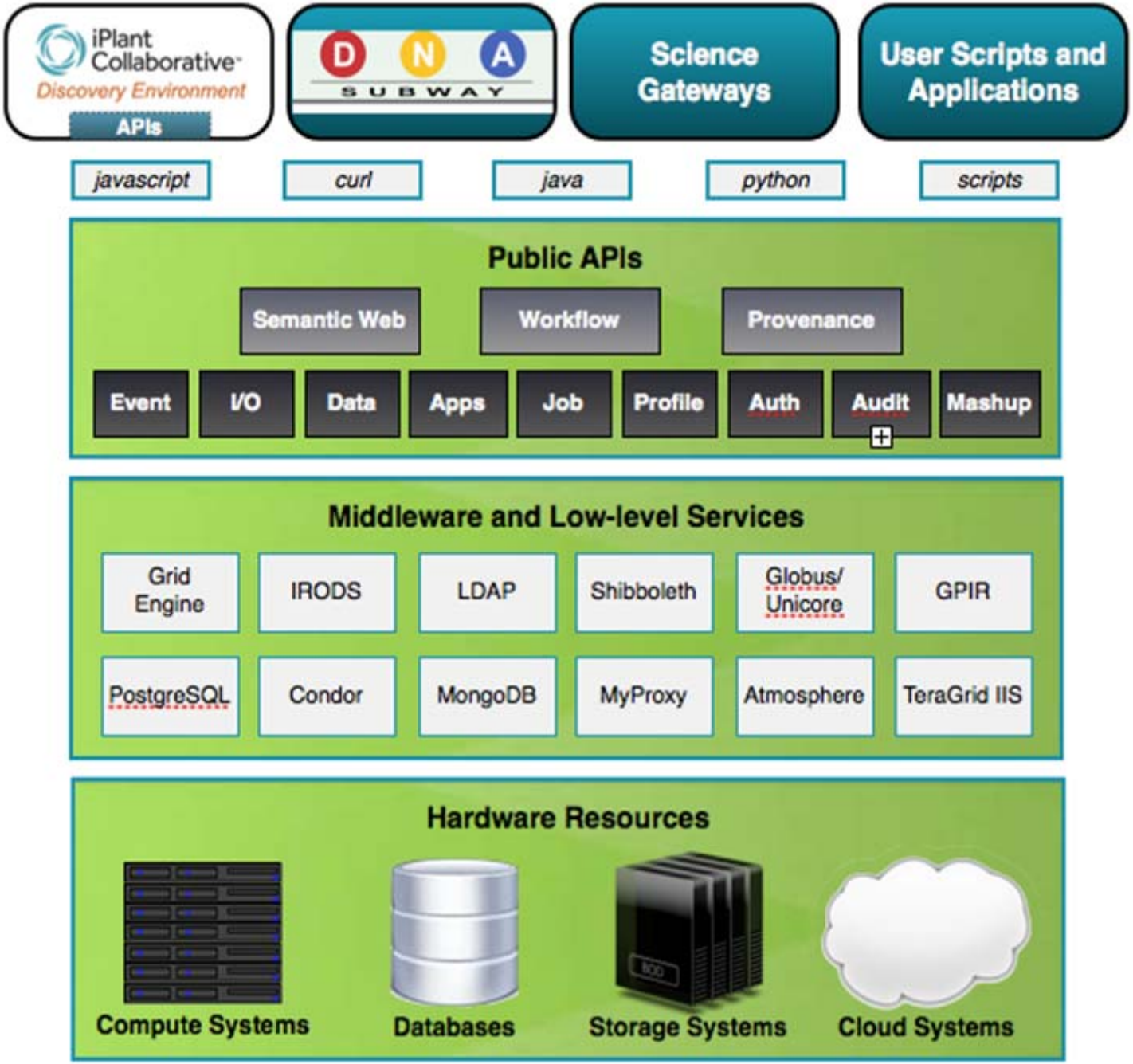
iPlant's Central Challenge

- To define what it means to build a lasting, community driven **Cyberinfrastructure** for the ***Grand Challenges*** of **Plant Science**, to get **community buy-in** of this vision, and to execute this vision.

iPlant Collaborative™

iPlant Cyberinfrastructure

- Universal, accessible, capacious storage
- Experimental reproducibility
- Sharing and collaboration
- Information Visualization
- Rapid, iterative workflow construction
- Ability to integrate community algorithms and practices
- Vast computing power (support to use it)
- **Support for community driven science**



Language/OS-independent application development

Complexity is abstracted behind API layer

Reuse and integration of Open Source Middleware

National-scale physical resources

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

iPlant Collaborative™

Powered by iPlant

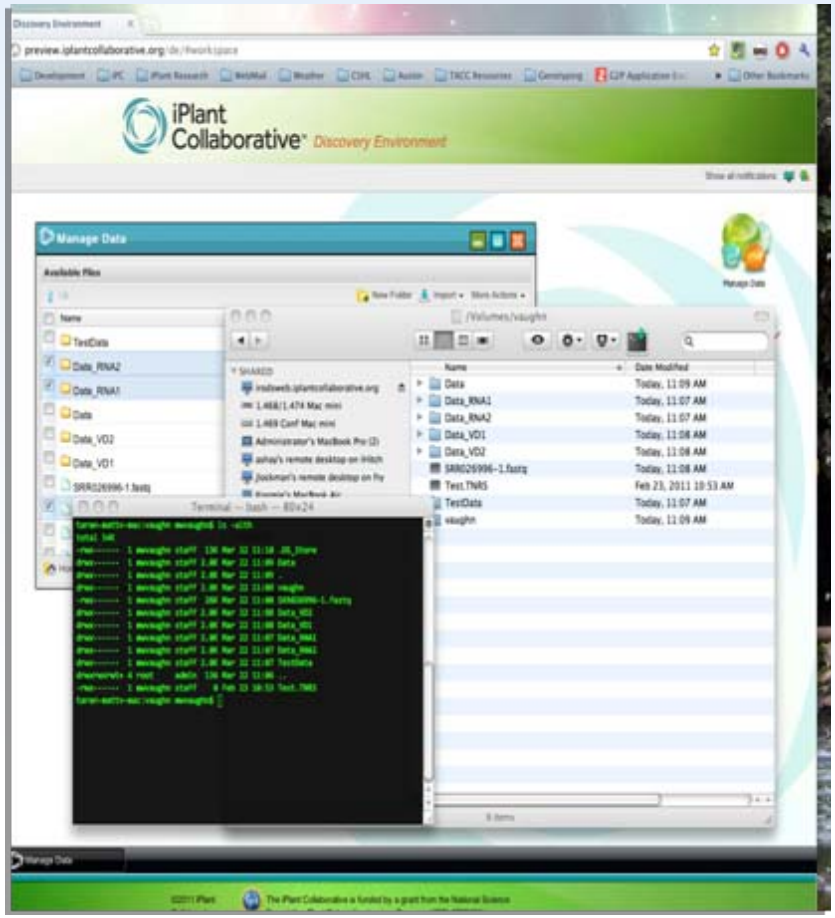
- Other major projects are beginning to adopt the iPlant CI as their underlying infrastructure (some completely, some in limited ways):
  - BioExtract *(web service platform)*
  - CiPRES *(computation)*
  - Gates Integrated Breeding Platform *(hosting, development)*
  - Galaxy (*storage, for now)*
 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

Major Ways to Access iPlant

- Atmosphere: For virtual hosting of web apps, sites, databases.
- iPlant Data Storage: All data large and small
- The Discovery Environment: Integrated Web apps.
- MyPlant: Social Networking.
- DNASubway: Annotation and more
- Standalone Apps: TNRS, TreeViewer, PhytoBisque, etc
- The API: For programmers embedding iPlant CI capabilities
- Command line for experts (thru TeraGrid/XSEDE)

iPlant Data Store

This is "Cloud Storage"… but it's not Amazon



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**

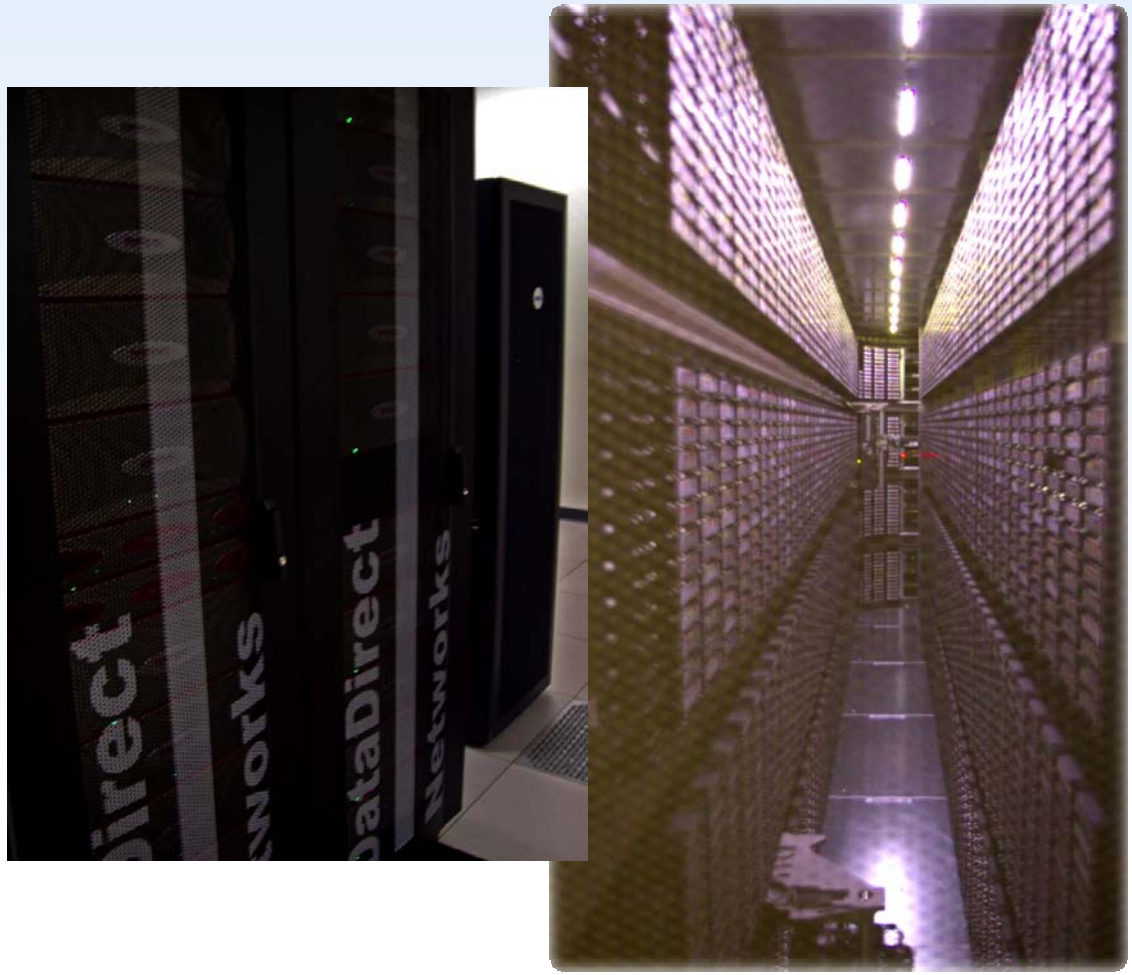
Fine-grained ACL permissions

- Sharing made simple

Access and a storage allocation is automatic with your iPlant account

iPlant Data Store

- When we say “capacious”, we mean >100,000 Terabytes of disk and tape



Project Atmosphere™

- 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
- Run servers, CloudBurst desktop use cases. Big data and the desktop are co-local again!

Eucalyptus

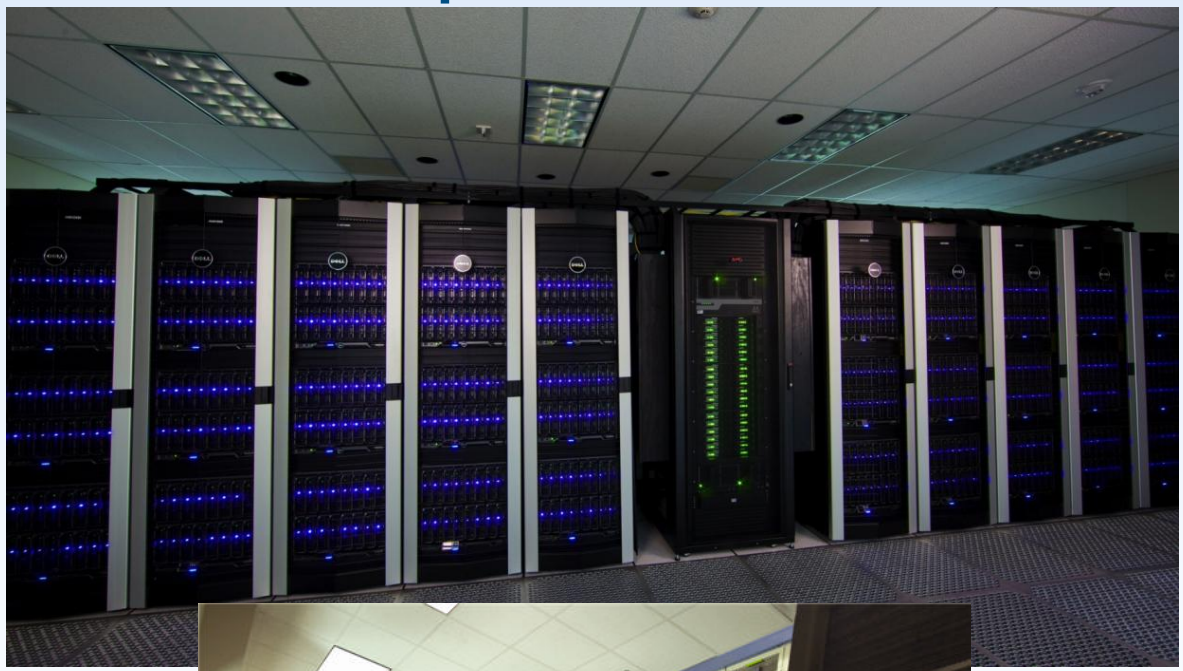
>60 hosted applications in Atmosphere today, including users from USDA, Forest Service, database providers, etc.

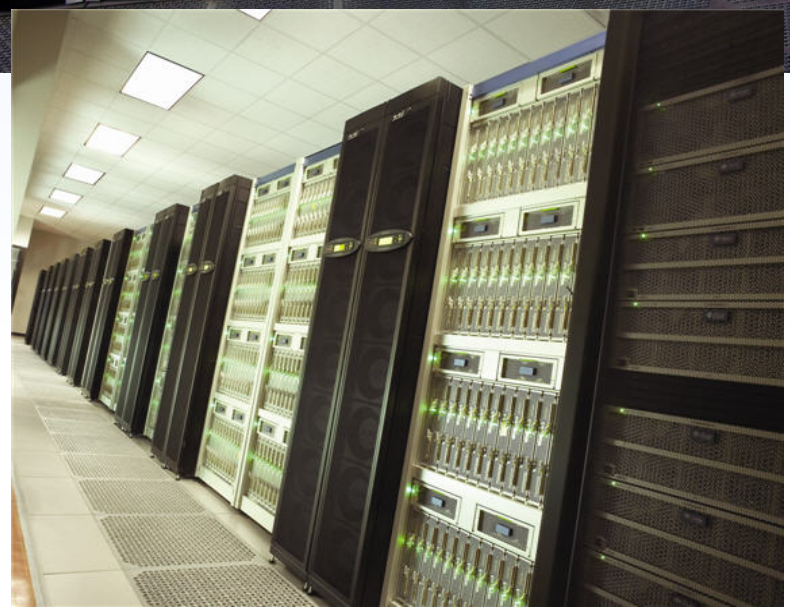
(30 more for postdocs and grad students for training classes)

iPlant Collaborative

Scalable Computation

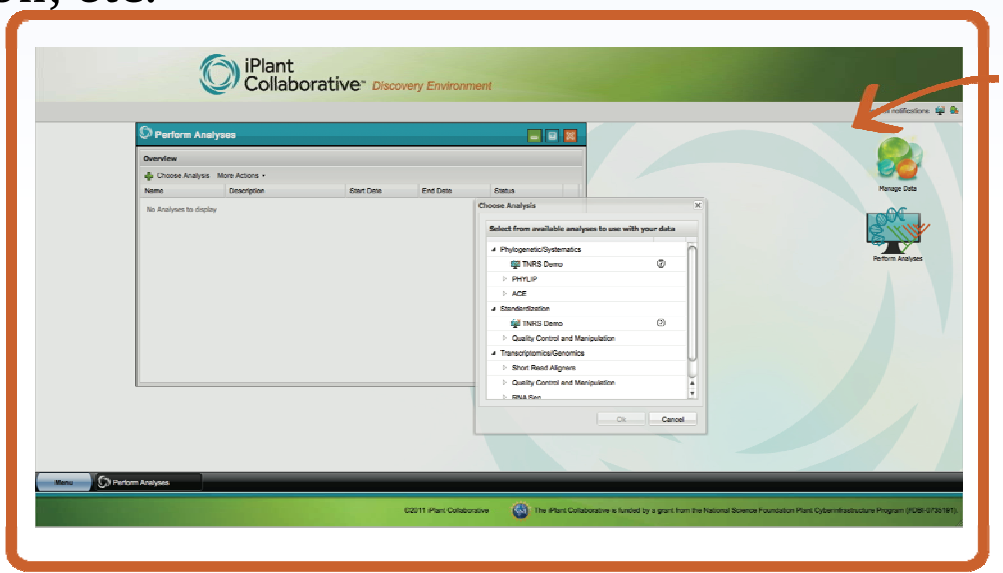
- 90,000 Compute Cores
- Up to 1TB shared memory
- Growing to ~500,000 cores by end of 2012





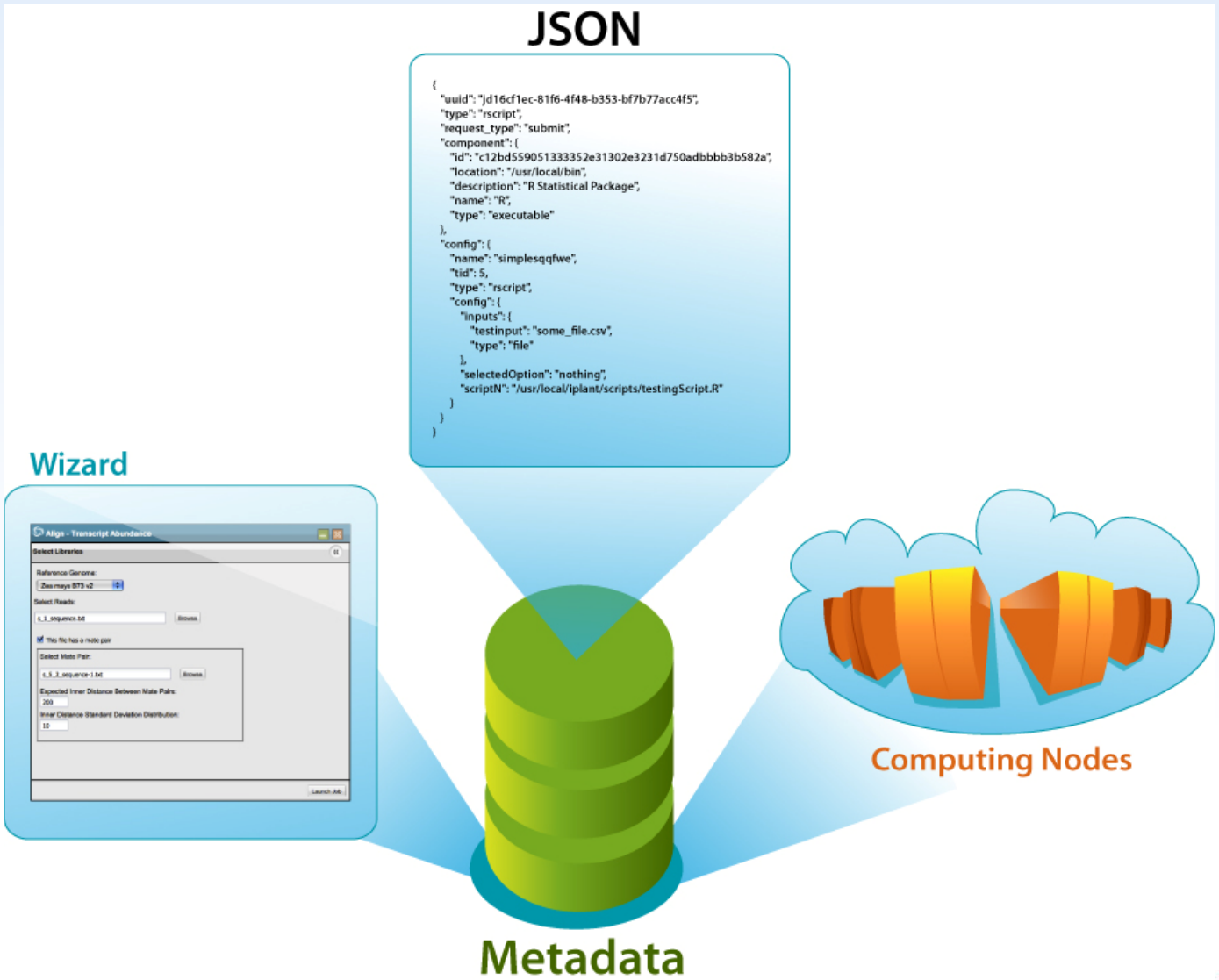
Discovery Environment

- A rich web client
  - Provides a consistent interface to a range of bioinformatics tools
  - Provides a portal to users not wishing to interact with lower level infrastructure
- An integrated, extensible system of applications and services
  - Provides additional intelligence above low level APIs – Provenance, Collaboration, etc.

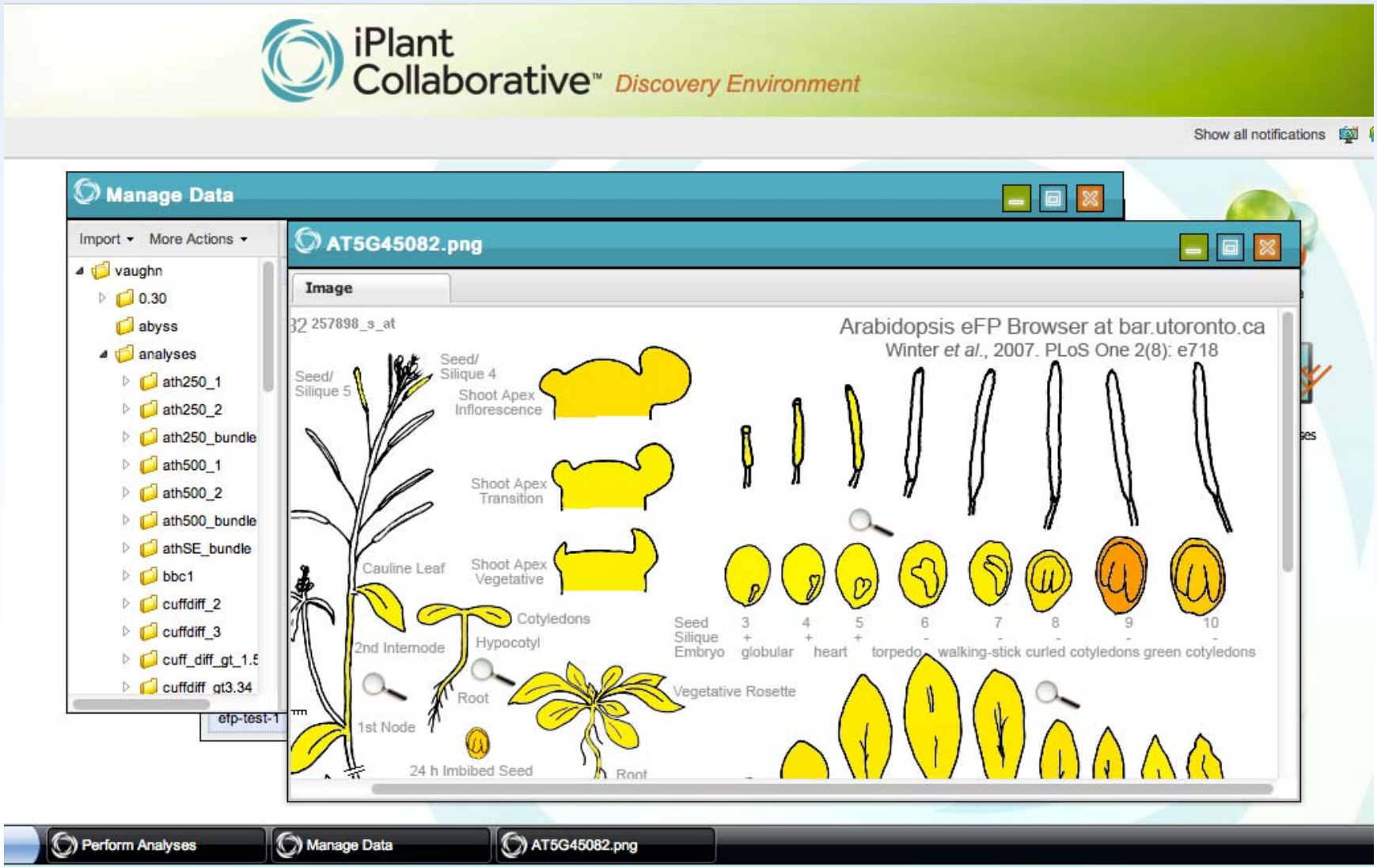


iPlant Collaborative

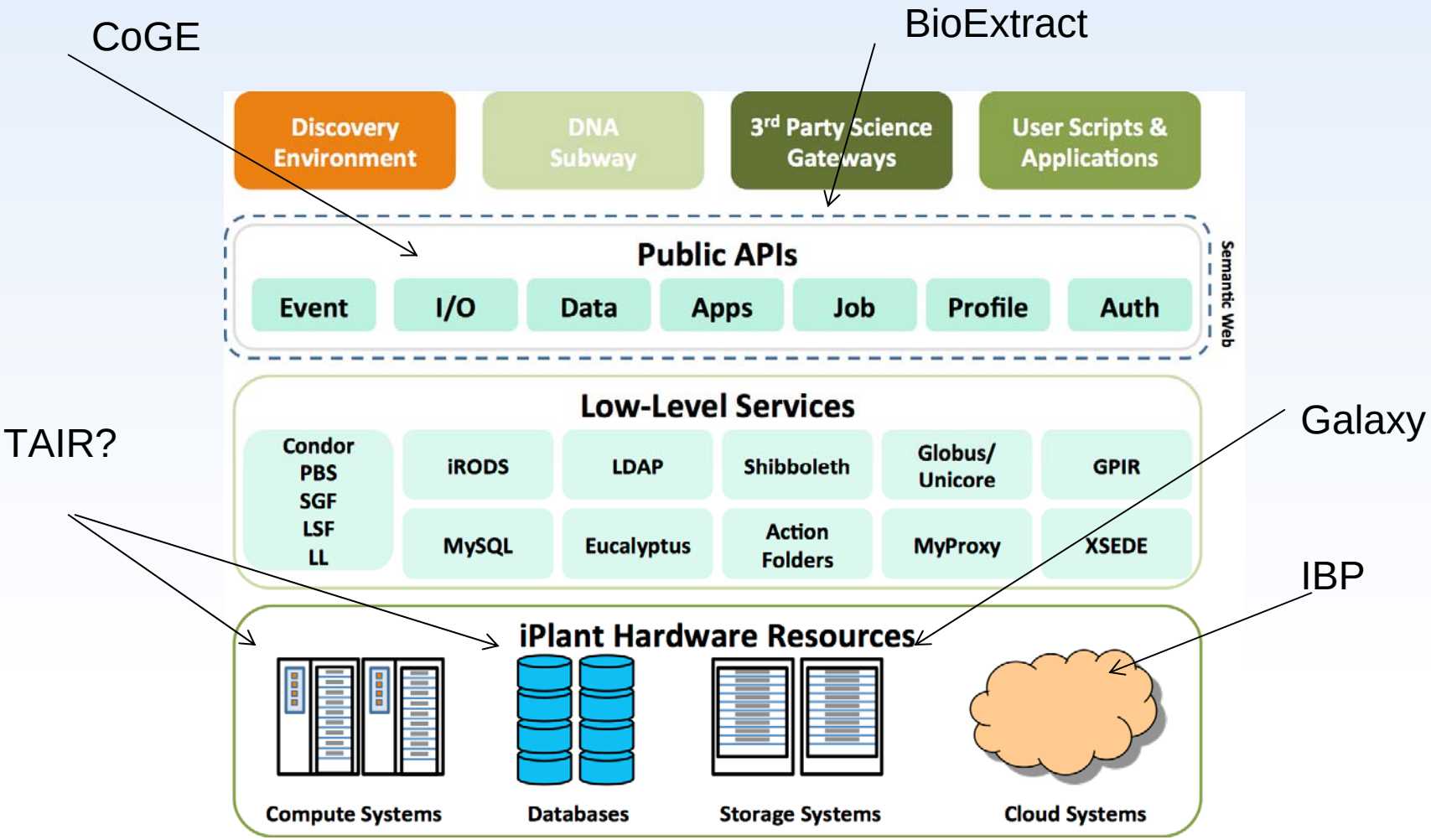
Integration of New Tools to UI without Programming



eFP in iPlant: in 38 minutes in September



Using the iPlant CI as a Foundation



Starting a New Collaboration

- If you just want to use the existing platforms, or add your own tools and workflows, just go to the web page, sign up, and read the documentation
- If you want to develop your own applications on top of the iPlant infrastructure, contact us directly:
  - dan@iplantcollaborative.org
  - support@iplantcollaborative.org
  - Best discussions are before proposals are submitted.
- Work with us to develop proposals for IAIC!

iPlant Collaborative™

TakeAways

- iPlant is building a robust CI, and many services and tools are available now.
  - >100 tools in discovery environment, and growing fast
  - >50 Applications in Atmosphere
  - Vast Computing and Storage resources available
- But iPlant is not about the specific tools available on a given day. It's about changing the way we build the future of bioinformatics.
  - Make use of the emerging international CI
  - Standard ways to interface tools**; things that work together!**
  - Standard ways to find and use them.

iPlant Collaborative™

Stop doing this…

```
tophat -r 160 -o top_SRR027863-65 ../../../reference/hg19
SRR027863_1.fastq,SRR027864_1.fastq,SRR027865_1.fastq
SRR027863_2.fastq,SRR027864_2.fastq,SRR027865_2.fastq

tophat -r 160 -o top_SRR027866-67 ../../../reference/hg19
SRR027866_1.fastq,SRR027867_1.fastq SRR027866_2.fastq,SRR027867_2.fastq


cufflinks -o cuff_SRR027863-65 top_SRR027863-65/accepted_hits.bam

cufflinks -o cuff_SRR027866-67 top_SRR027866-67/accepted_hits.bam


cuffmerge -s ../../../reference/hg19.fa assemblies.txt


cuffdiff merged_asm/merged.gtf top_SRR027863-65/accepted_hits.bam
top_SRR027866-67/accepted_hits.bam
```

Start doing this…