

The iPlant Collaborative

Presented by
Sudha Ram

McClelland Professor of MIS and Computer Science

Eller College of Management

University of Arizona, Tucson AZ 85721

Email: ram@eller.arizona.edu

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



What is iPlant?

- The mission of the iPlant Collaborative is to build a cyberinfrastructure (CI) to support the solution of grand challenges in plant biology.
- A “unique” aspect is the grand challenges were not defined in advance, but are identified through an ongoing engagement with the plant biology community.

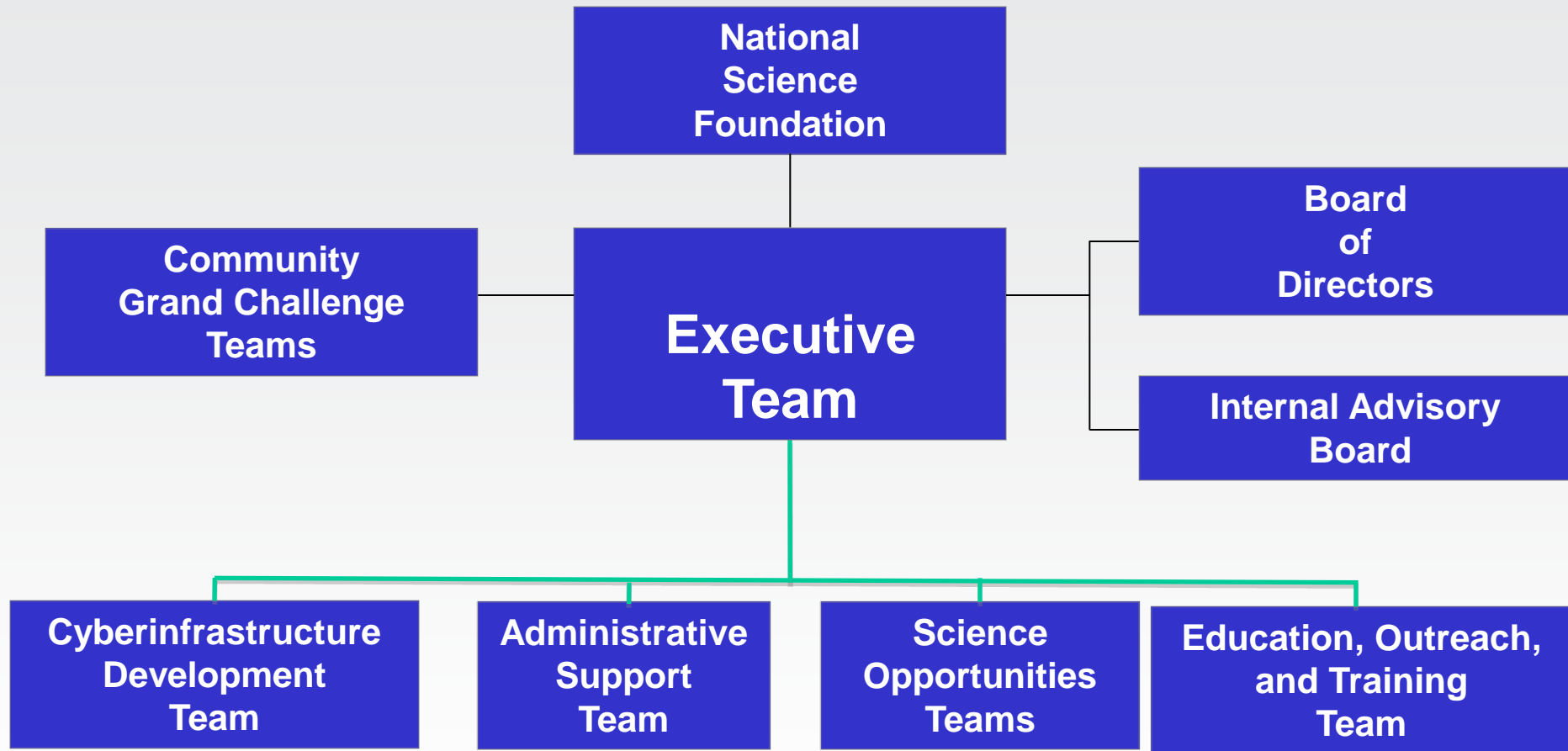


What is iPlant

- A virtual organization forming grand challenge teams from the scientific community.
- Long term focus on sustainable food supply, climate change, biofuels, pharma....
- **Now hundreds of participants from around the world... Working group members at more than 50 US academic institutions, USDA, DOE, etc.**



The iPlant Collaborative



What is the iPlant CI?

- Two grand challenges defined:
 - iPlant Tree of Life (IPTOL):

Build a single tree showing the evolutionary relationships of all green plant species on Earth
 - iPlant Genotype-to-Phenotype (IPG2P)

Construct a methodology whereby an investigator, given the genomic and environmental information about a given individual plant, can predict its characteristics.
 - Taken together, these challenges are the key to unlocking many “holy grails” of plant biology, such as the creation of drought resistant or pest resistant crops, or breaking reliance on fossil fuel based fertilizer



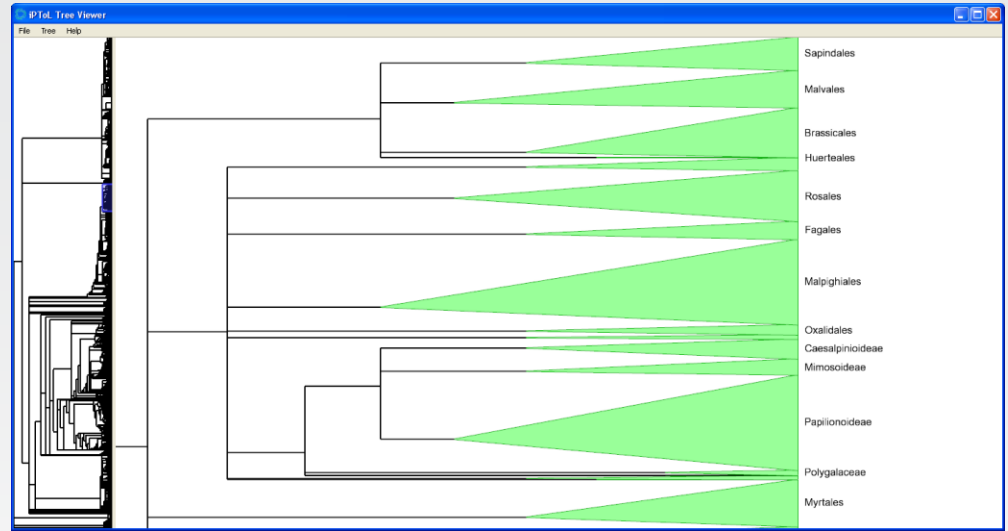
What is the iPlant CI?

- IPTOL CI:
 - Five areas: Data assembly and integration, visualization, scalable algorithms for large trees, trait evolution, tree reconciliation
- IPG2P CI:
 - Five areas: Data Integration, Visualization, Modeling, Statistical Inference, Next Gen Sequencing Tools
- In both, a combination of applying compute resources, developing or enhancing new tools, and creating web-based “discovery environments” to integrate tools and facilitate collaboration.



What is the iPlant CI?

- Two grand challenges:
- iPlant Tree of Life (IPTOL):
 - Build a single tree showing the evolutionary relationships of all green plant species on Earth
- iPlant Genotype-to-Phenotype (IPG2P)
 - Construct a methodology whereby an investigator, given the genomic and environmental information about a given plant, can predict it's characteristics.



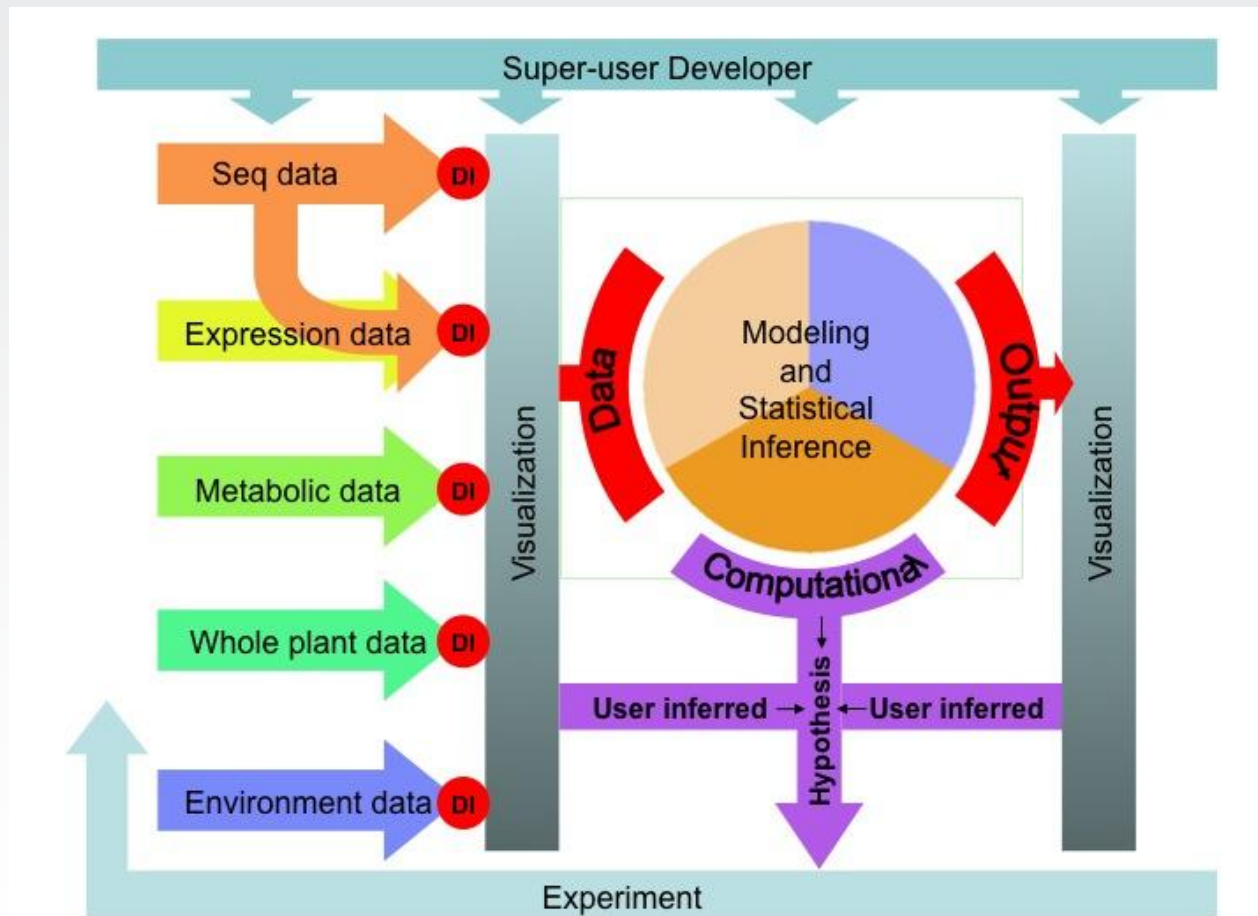
Prototype visualization tool, showing 220,000 taxa phylogenetic tree

Strong focus on *data integration*, not simulation:
Plant science is truly data driven.

Still many computational challenges (e.g. inferring phylogenies from genome data)



Diversity of Datasets in iPlant



Plant Science CI Needs

Through the Grand Challenge Workshop and proposal process, a great deal has been learned about the nature of the computational problems faced by plant biology.

- **Models are Immature**

- With a few exceptions (phylogenetic trees, molecular dynamics), modeling is still in it's infancy compared to other disciplines.

Data Integration is Key

- Vast numbers of databases with no single ontology; issues regarding data integration, **provenance**, and quality are key enabling technologies.
- *A successful plant science CI will look very different from the type of CI that has had so much impact in other disciplines (e.g. computational chemistry, particle physics).*



Building a Bridge to the CI Community

- A personal observation about Plant Science CI from Dan Stanzione:
- “...This is the first discipline I’ve worked in where my knowledge of differential equations and numerical methods wasn’t enough to understand the problems”.
 - Unlike electromagnetics, semiconductor design, weather forecasting, astrophysics, fluid mechanics, structural analysis, ocean modeling, mantle convection, earthquake modeling, hydrodynamics, signal processing, nanomaterials modeling, molecular dynamics, etc, etc.
- The language barrier between the bio community and the computing/CI community is *far* higher than any other discipline



Technology Evaluated and used IPTOL

- SoftParseMap
- Phylocom
- Phylomatic
- CIPRES WebPortal
- Dendroscope
- Phylum
- ATV
- SimMap
- Parafit/AXParafit
- NOTUNG
- TreeTapper
- TreeJuxtaposer
- AFTOL
- NexML data standard
- APWeb/APWeb2
- PhyloWidget
- PRIMETV
- Copycat
- APE
- PDAP/Mesquite
- Walrus
- PhyloWS
- CONTRASTS in Phylip



Recently Evaluated - Infrastructure

- Resiliency
 - BTMP
 - BLCR
 - Libratto
- Workflow
 - Pegasus
 - Neptune
 - Taverna
 - Kepler
 - Condor/OSG
- Web Portal
 - Nanohub
- Component Technologies
 - CCA
 - Hadoop
- User Interface Technologies
 - Adobe Flex
- Semantic Web
 - OWL
 - SSWAP



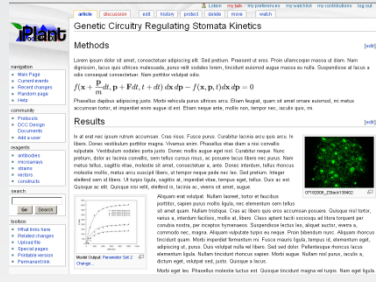
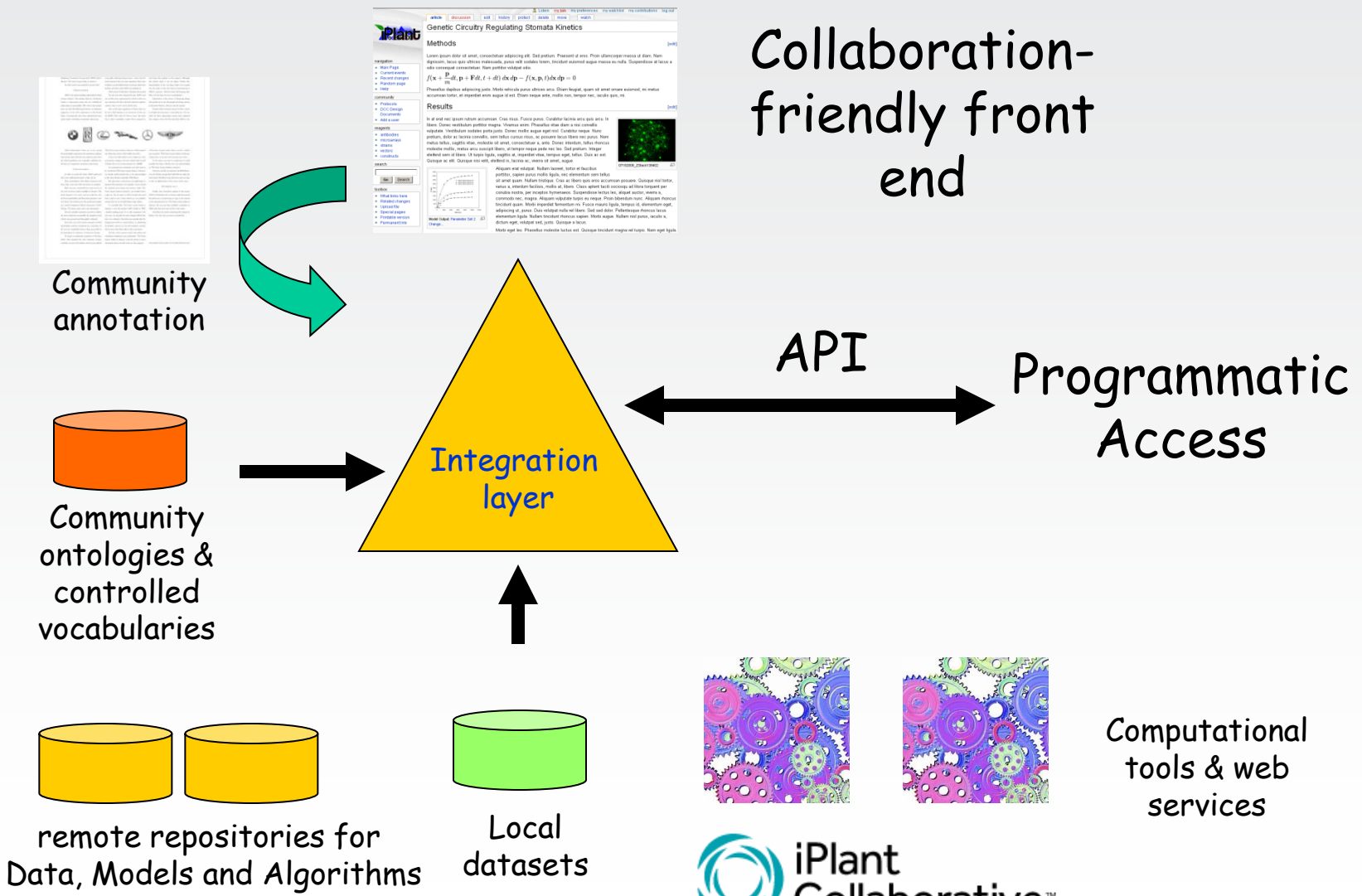
A Foundation of Computational and Storage Capability

- iPlant is positioned to take advantage of *tremendous* amounts of NSF and institutional compute and storage resources:
 - Compute: Ranger, Lonestar, Stampede (UT/TeraGrid) Saguaro, Sonora (ASU) Marin, Ice (UA)
 - **~700 Teraflops, more computing power than existed in all the Top 500 computers in the world 4 years ago**
 - Storage: Corral, Ranch (UT), Ocotillo (ASU)
 - **Well over 10 Petabytes of storage can be made available for the project, on scalable systems capable of growing much more.**
 - Visualization: Spur, Stallion (UT), Matinee (ASU), UA-Cave
 - **Among the world's largest visualization systems**
 - Virtualized/Cloud Services: iPlant (UA) and ASU virtual environments, vendor clouds
 - **iPlant is positioned to use cloud technologies to deliver persistent gateways and services to users.**

In short, the physical aspects of cyberinfrastructure employed via iPlant, utilizing large scale NSF investments, has capabilities second to none anywhere on the planet.



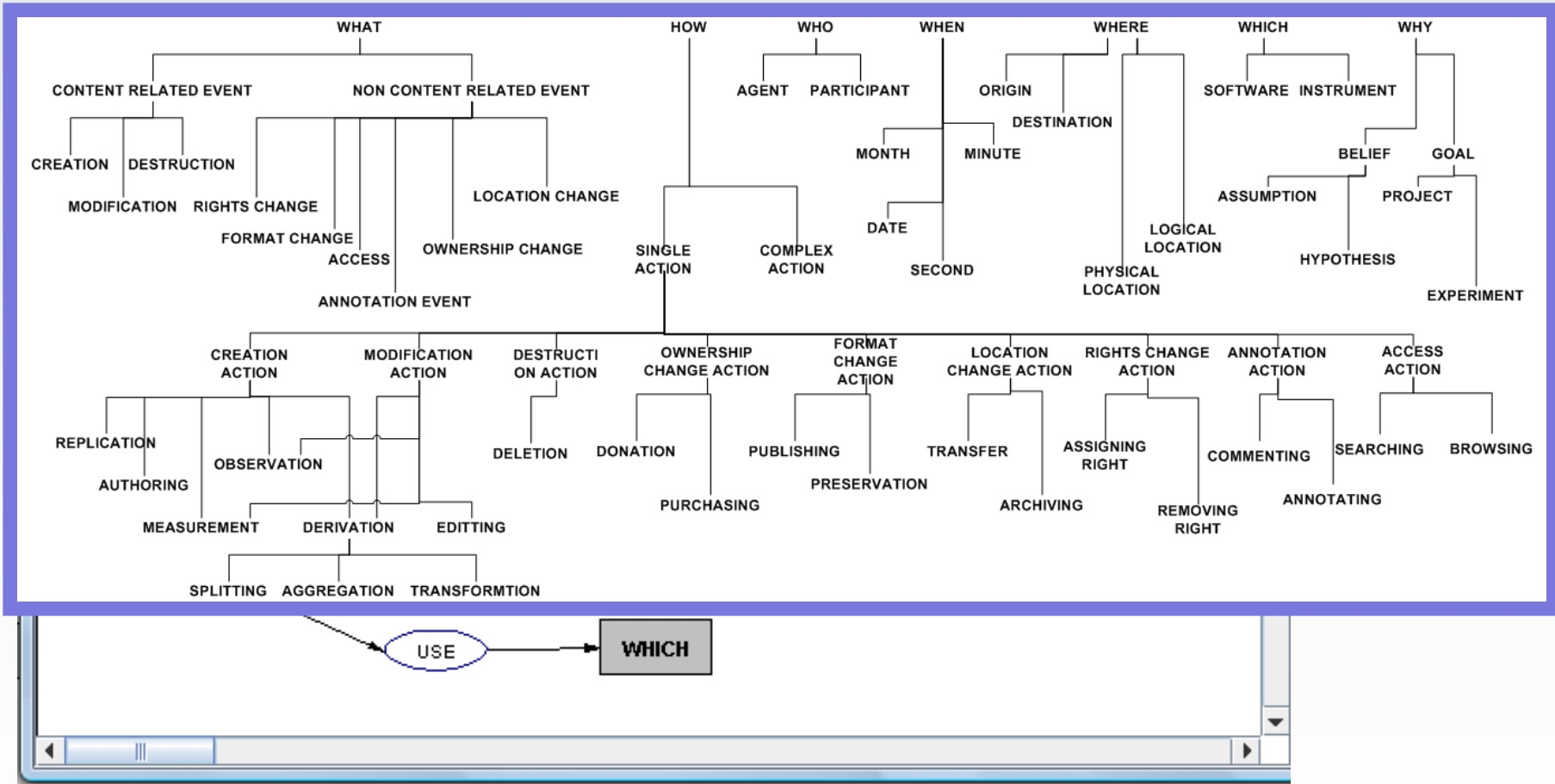
A Discovery Environment



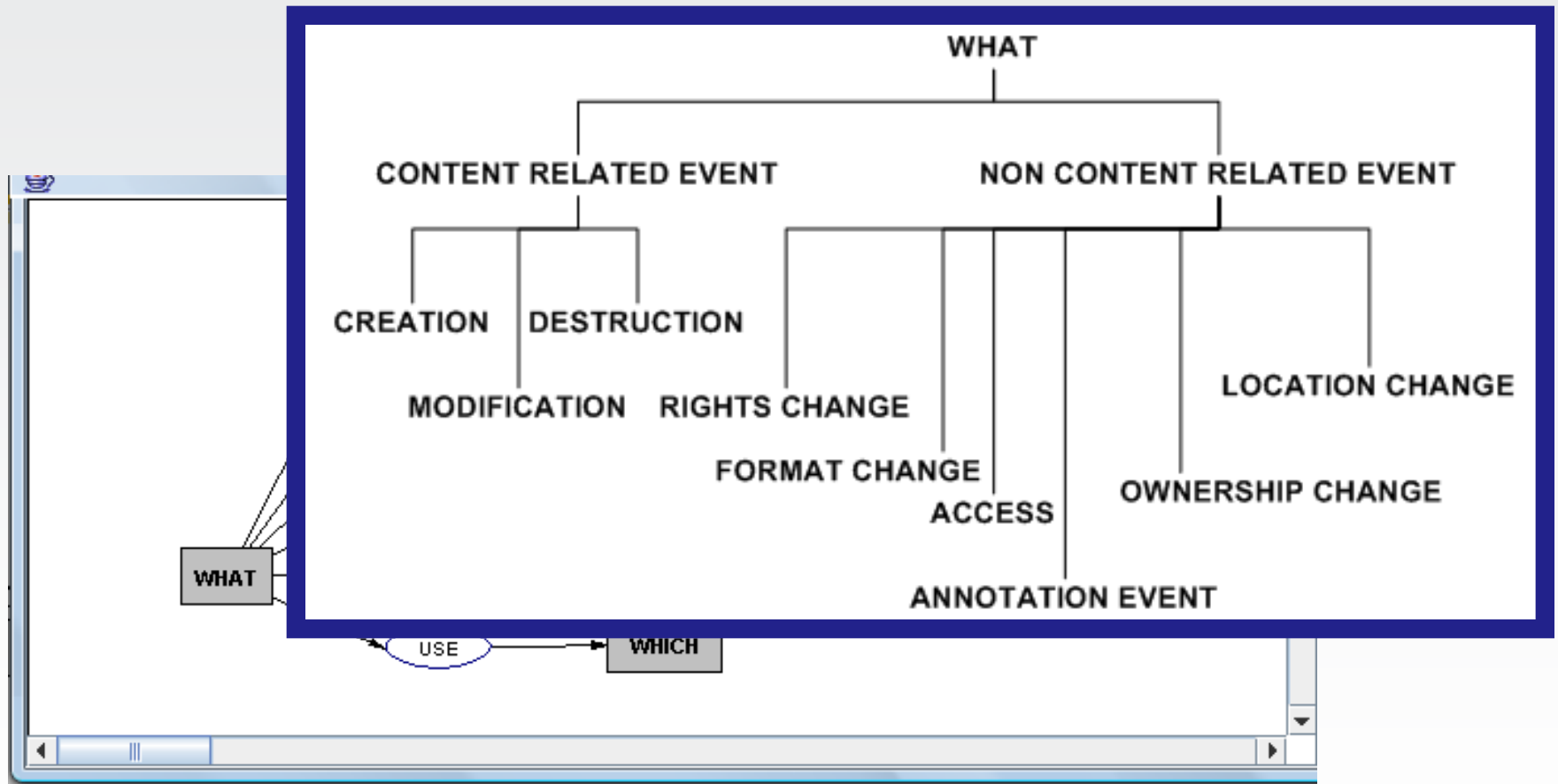
Provenance Management is Key!



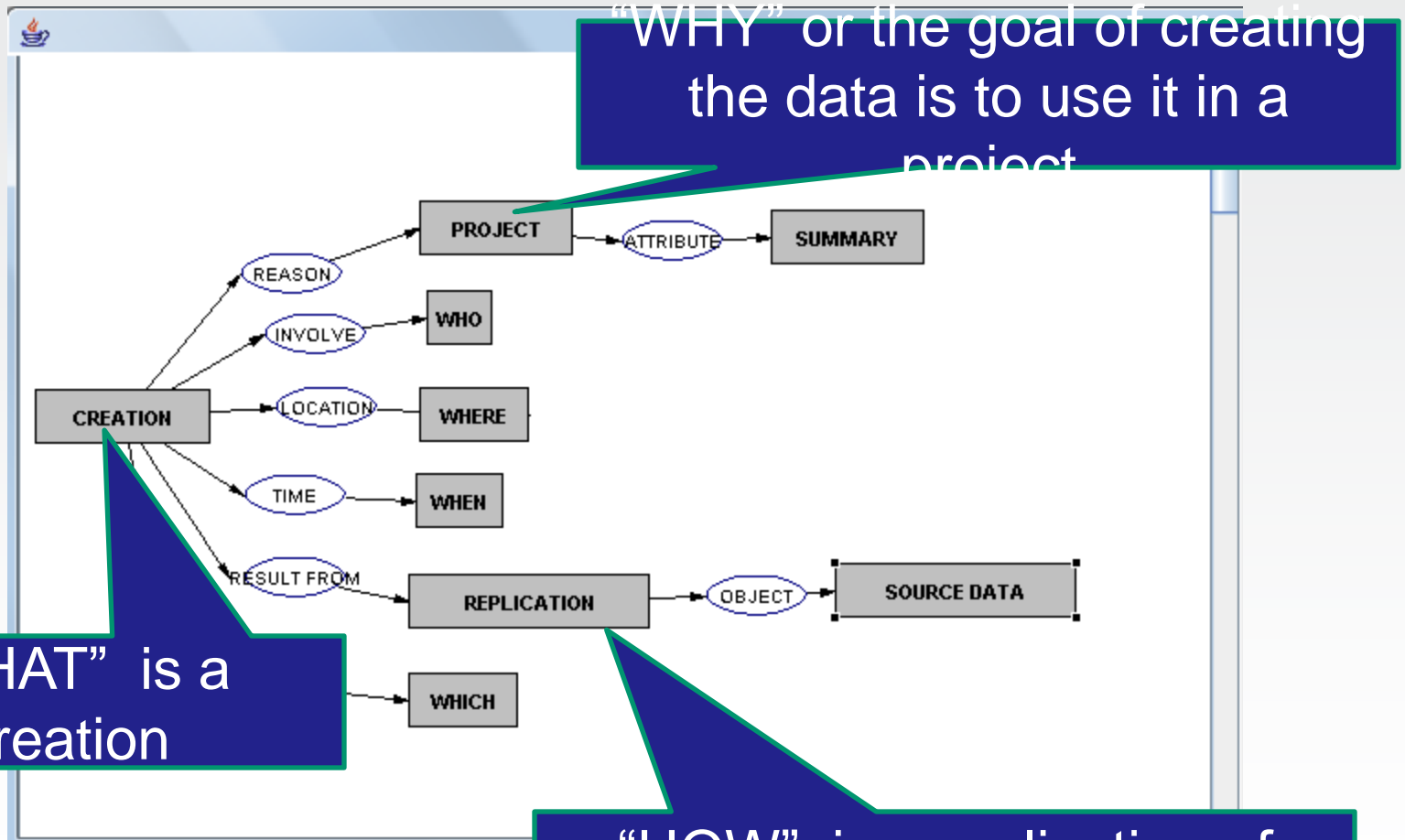
W7 model



W7 model of Provenance

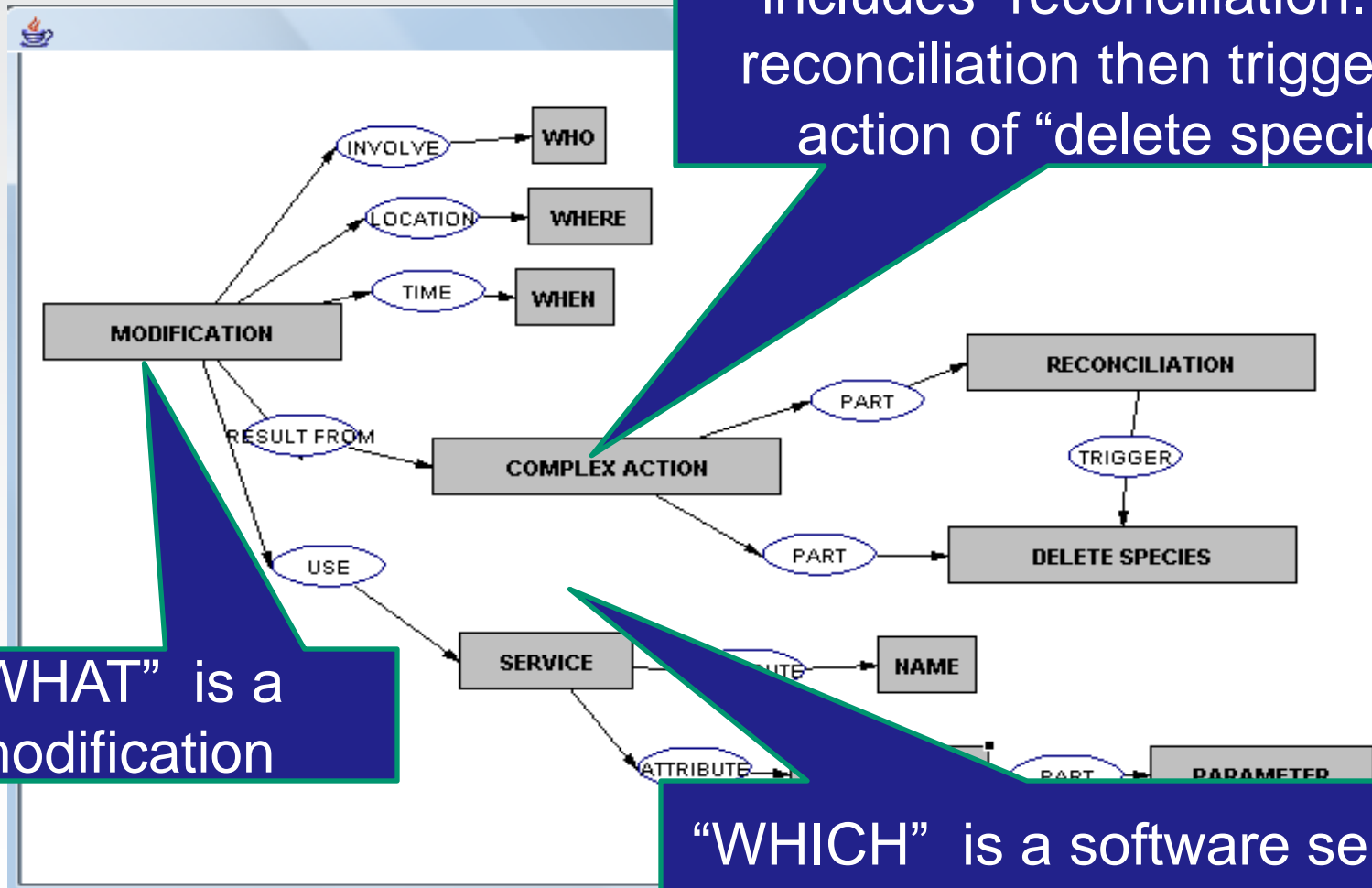


Provenance in iPlant



Provenance for iPTOL

“HOW” is a complex action that includes reconciliation. The reconciliation then triggers an action of “delete species”



“WHAT” is a modification

“WHICH” is a software service. We want to record the name and settings of the service



Uses of Provenance

- Experimental reproducibility
- Recipe for replication of scientific workflow
- Tracking ownership and resolving IP issues
- Audit Trail
- Evaluating Data Quality
- Informational Purposes
- Data Life Cycle Management



DNA Subway Demo

tour-all.swf (application/x-shockwave-flash Object)

file:///Users/dan/Desktop/subwaytour/tour-all.swf

Most Visited Latest Headlines Yahoo! Google Maps YouTube Wikipedia News Popular My Yahoo! ESPN Yahoo! Sports Fanta... Yahoo! Sports Fanta... Microsoft Exchange ...

PETTT... NSF Fell... Longho... iPlant C... Home - ... My-Pla... Google ... VA_sco... Universi... Google ... PETTT... tour-...

FAST TRACK TO GENE ANNOTATION AND GENOME ANALYSIS

ID:
Password:
Log In Enter As Guest
Forgot Password? Register

D N A
SUBWAY

Optimized FOR PLANTS
animals coming soon

Annotate a Genomic Sequence Prospect Genomes Using TARGET

Find Fragments Predict Genes Search Databases Build Models

Alignment & Tree Viewer

Browsers & Transfer

Annotate a Genomic Sequence
Input a DNA sequence of up to 100,000 nucleotides to identify the genes and transposons in it, and annotate (associate) the sequence with this information.

This site ties together key bioinformatics tools and databases used to annotate genes and analyze genome data. Roll over any of the "stations" on the subway map to find out more about the analysis steps. Start a project by selecting one of the "subway lines" (red, yellow). Register and login to be able to save or share your results.

About Credits Resources Feedback

All content is © 2009 by the iPlant Collaborative

Using the Red Line, you can predict and annotate genes in up to 100,000 basepairs of DNA.



Questions??

