

iPlant: Supporting the Lifecycle of Data

Eric Lyons



One Big Problem...

2008



2011



One Big Problem...

Published online 3 September 2008 | *Nature* 455, 16-21 (2008) | doi:10.1038/455016a

News Feature

Big data: Welcome to the new world of data

What does it take to store bytes by the trillions? Can you analyze them in a way which it's a

Big data: Distilling meaning from the mountains of information

Felice Frankfort¹, Clifford Lynch², Rosalind Wiseman³

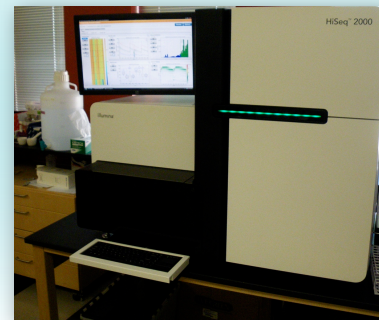
Lifecycle of Data

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

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



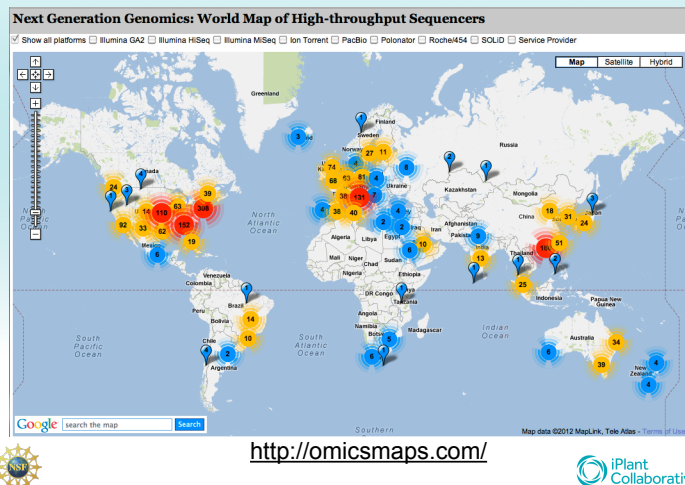
High-throughput Data Acquisition



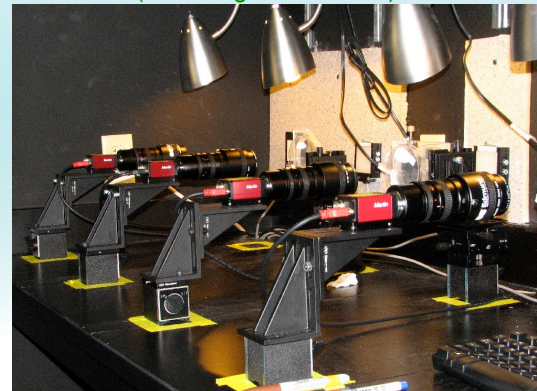
- In 11 Days
- Generates 4TB of raw data
- 600,000,000,000 bases of DNA sequence (200 human genomes)



and it is a global phenomenon



High-throughput Phenotyping (Watching Grass Grow)



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



Big Data in Ecology Global Multidimensional Data



Biologist

Lifecycle of Data

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

gnc wbbj wbbL

iPlant Collaborative

Biologist

iPlant Data Store

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

gnc wbbj wbbL

iPlant Collaborative

The journey of data to iDS and challenges along the way !

Hard Drive Network card Building network Campus network Internet

Internet UA/TACC network iDS Network card Hard Drive

http://en.wikipedia.org/wiki/List_of_device_bandwidths
Check: USB, HDD, Network capabilities

iPlant Collaborative

The Flow of Data

iPlant Collaborative

How fast is our network?

Wireless

Wired

AT&T (iPhone 3G)

iPlant Collaborative

iPlant Layered Services and Access

Community Facing Resources

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

End Users

iPlant Data Store

**Scalable
Reliable
Redundant
High-Performance**

Computational Users

Job, Apps, IRC/Federated, Grid

Cloud Systems, High Performance Computers, Databases, Storage

iPlant Data Store Free Your Data

**Different Users,
Different Access Needs:
One Data Store**

iPlant Data Store Free Your Data

WebDAV **DE**

iPlant Data Store

**Desktop Folder
Discovery Environment
Command Line (HPT)
i-Drop (HPT)
API**

iPlant Data Store Web-Integrated High Performance Big Data Transfers

iPlant Data Store

Discovery Environment (HPT)

Up to 4x faster than "regular" web transfers

iPlant Data Store: Metadata Data About Data

iPlant Collaborative Discovery Environment

Name	Value	Unit
Metadata	iPlant/home/elyons/data/data_big	
Date created	July 2012	
Experiment	Banana genome	

iPlant Collaborative

iPlant Data Store: Metadata Data About Data

OrganismView

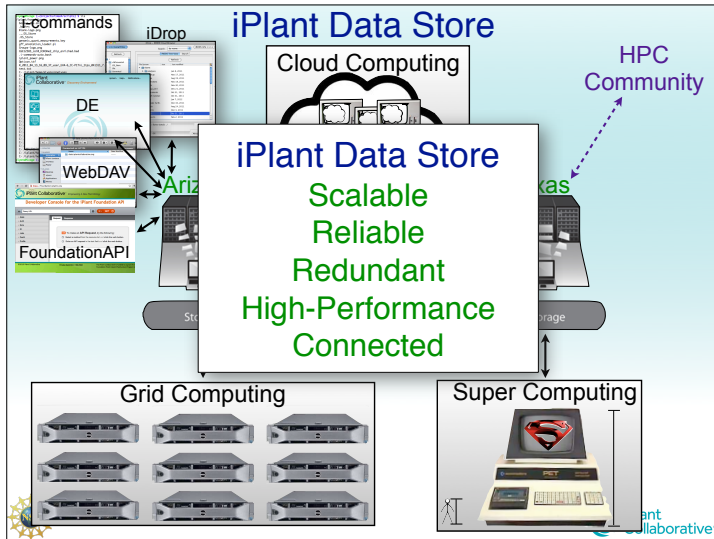
Search for organisms and genomes

Organism Name: maize
 Organism Description: Search
 Dataset Name: Zea mays (maize:com) 11265.faa

More Data; Smarter Data

source_link	http://ftp.maizegenome.org/maize_pseudo_v2.tar.gz
source_name	ZmB73_sb_FGS.gff.gz
version	2
organism	Zea mays (maize:com)

**GC content
GFP Names and Annotations
CodeBlast (Send To iPlant Data Store)**



iPlant Data Store Performance

The UC Berkeley Connection

39,000 Students
2900 Faculty/Staff

Dec 5th, 2011:
100GB: 29m15s

36,000 Students
2000 Faculty

iPlant Data Store Performance

The UC Berkeley Connection

100GB: 29m15s

1 GB / 17.5 seconds

Source	Destination	Copy Method	Time (seconds)
CD	My Computer	cp	320
Berkeley Server	My Computer	scp	150
External Drive	My Computer	cp	36
USB2.0 Flash	My Computer	cp	30
iDS	My Computer	iget	18
My Computer	My Computer	cp	15

My Computer (UA): 7.2K Internal Hard Drive
External Drive: USB 2.0: 5.4k Hard Drive
Flash Drive: USB2.0 Patriot XT

<https://pods.iplantcollaborative.org/wiki/display/start/How+fast+is+the+iPlant+Data+Store>

iPlant Data Store

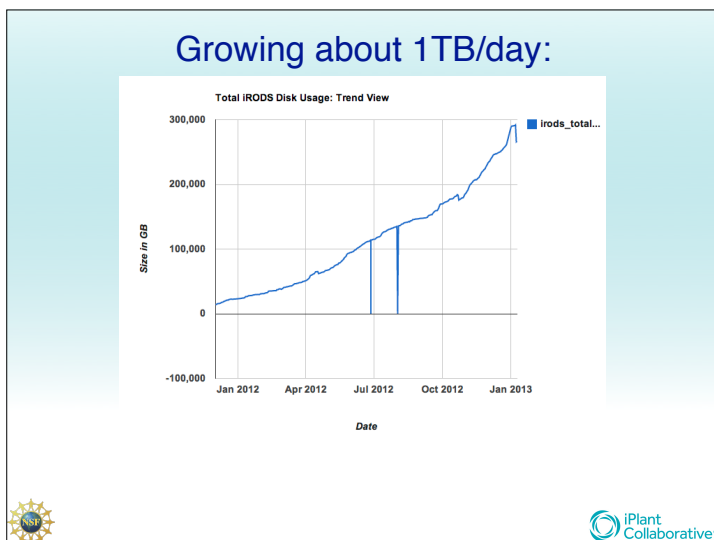
Connecting people with data and computation:

Lifecycle of Data

Powered By iPlant

Cyberinfrastructure for Life Sciences

Scalable
Capable
Extensible



Where to Get Information

Data Store Quick Start:
<http://www.iplantcollaborative.org/Zki>

Data Store Manual:
<http://www.iplantcollaborative.org/Zko>

iPlant Forums:
<http://forums.iplantcollaborative.org>

Exercise Your Data Store

For each (WebDAV, DE, iDROP, i-command):

{

1. Upload your data
2. Rename your data
3. Delete your data

}

