Galaxy 101: Data Integration, Analysis and Sharing

Jennifer Hillman-Jackson
Penn State University

Thursday, November 8
Workshop 7-8:30 pm

Galaxy

galaxyproject.org

Agenda

7:00 **Welcome, Introduction to Galaxy**

7:15 Galaxy 101 Tutorial: Data, Tools, Workflows

8:30 Galaxy Usage, Sharing, Support, and Resources

8:45 Open discussion

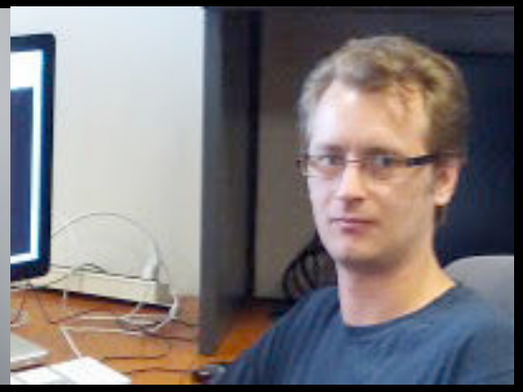
9:00 Done

Enis Afgan

Guru Ananda

Dannon Baker

Dan Blankenberg



Dave Bouvier

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Nuwan Goonasekera

Jen Jackson

Greg von Kuster

Ross Lazarus

Rémi Marenco

Scott McManus

Anton Nekrutenko

The Galaxy Team

http://galaxyproject.org/wiki/GalaxyTeam

James Taylor

Goals for this workshop

Introduction

- What is Galaxy? Project mission and goals.
- Core **Terminology**

Hands-on experience

- Perform **bioinformatics analysis with Galaxy**
- Use functions to **Visualize, Share, and Publish**

Going forward

- Learn **how to stay connected to the Galaxy community for support** once you begin your **own projects**
- Where to find more **tutorials** and advanced **'live supplimentals'**.
- **Galaxy Options**: Local, Cloud, Other Galaxies ....

Today is just the start

This workshop is about "Using Galaxy" and it will not cover in depth details about how the tools are implemented or new algorithm designs or which assembler or mapper or X ...

... is best for **all projects** or your **specific needs**.

BUT, we ***really*** like to help. **Tool forms are packed with useful information**. **All material is available online** and you can **review our support wiki** anytime, including the 101 protocol and by next week, these exact slides. If you get stuck or need more info, simply **send an email to one of our mailing lists** and we can get you on track.

http://wiki.galaxyproject.org/Support

So, ...What is Galaxy?

- A **data analysis and integration** tool
- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- These options result in several **ways to *use Galaxy***

http://galaxyproject.org

Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

Galaxy 101 Terminology

Dataset:

Any input, output or intermediate set of data + metadata. A record of a specific data or analysis step.

History:

A series of inputs, analysis steps, intermediate datasets, and outputs. A record of a group of data and analysis steps.

Tool:

An operation within Galaxy that acts upon dataset(s) as an analysis step. May be developed by Galaxy team or a 3rd party program that has been "wrapped" for Galaxy.

Workflow:

A series of analysis steps executed in a sequential stream

Agenda

7:00 Welcome, Introduction to Galaxy

7:15 Galaxy 101 Tutorial: Data, Tools, Workflows

8:30 Galaxy Usage, Sharing, Support, and Resources

8:45 Open discussion

9:00 Done

Galaxy 101 Tutorial

Located on the free Public Main Galaxy server at:

http://usegalaxy.org

Shared Data -> Published Pages ->

"Galaxy 101: The first thing you need to try"

http://usegalaxy.org/galaxy101

Galaxy 101 Tutorial

We're going to use a Cloud server for the workshop. To get started *now* ...

1. Open a **web browser**. One of the following is *strongly* recommended:

(a) Firefox, (b) Google Chrome, or (c) Safari

2. Go to ***http://cloud1.galaxyproject.org/***

3. Register for a new account:

User -> Register -> fill out form and submit

Agenda

7:00 Welcome, Introduction to Galaxy

7:15 Galaxy 101 Tutorial: Data, Tools, Workflows

8:30 Galaxy Usage, Sharing, Support, and Resources

8:45 Open discussion

9:00 Done

Galaxy's Mission & Project Goals

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

Accessible: Users without programming experience can easily specify parameters and run tools and workflows.

Reproducible: Galaxy captures information so that any user can repeat and understand a complete computational analysis.

Transparent: Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

Sharing and Publishing Your Work

CSH PRESS GENOME RESEARCH

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7],
Fran
Jame

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

Current Issue

October 2010, 20 (10)

Footnotes

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]

Histories, workflows, visualizations and ***pages*** can be shared with others or published to the world.

http://usegalaxy.org/u/aun1/p/windshield-splatter

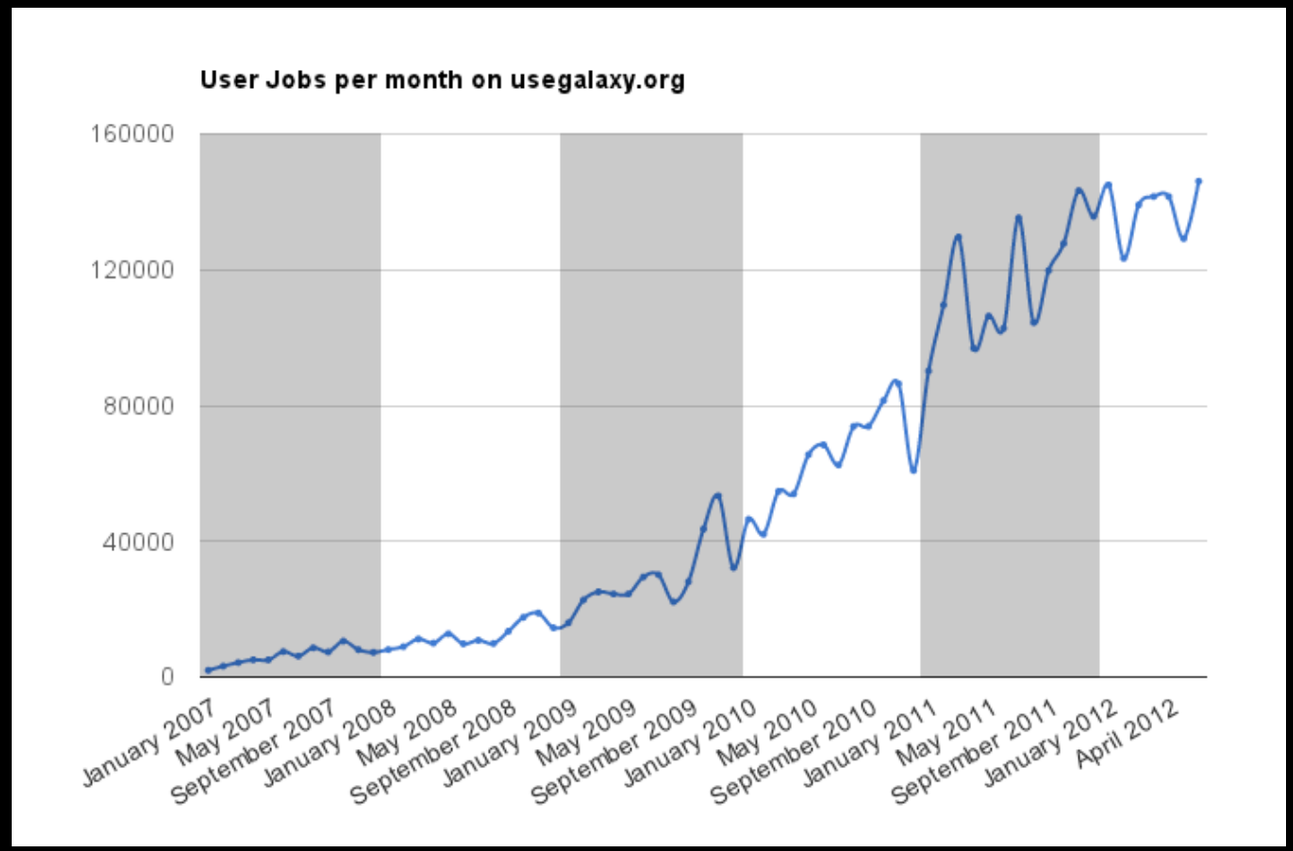
So, ...What is Galaxy? Redux

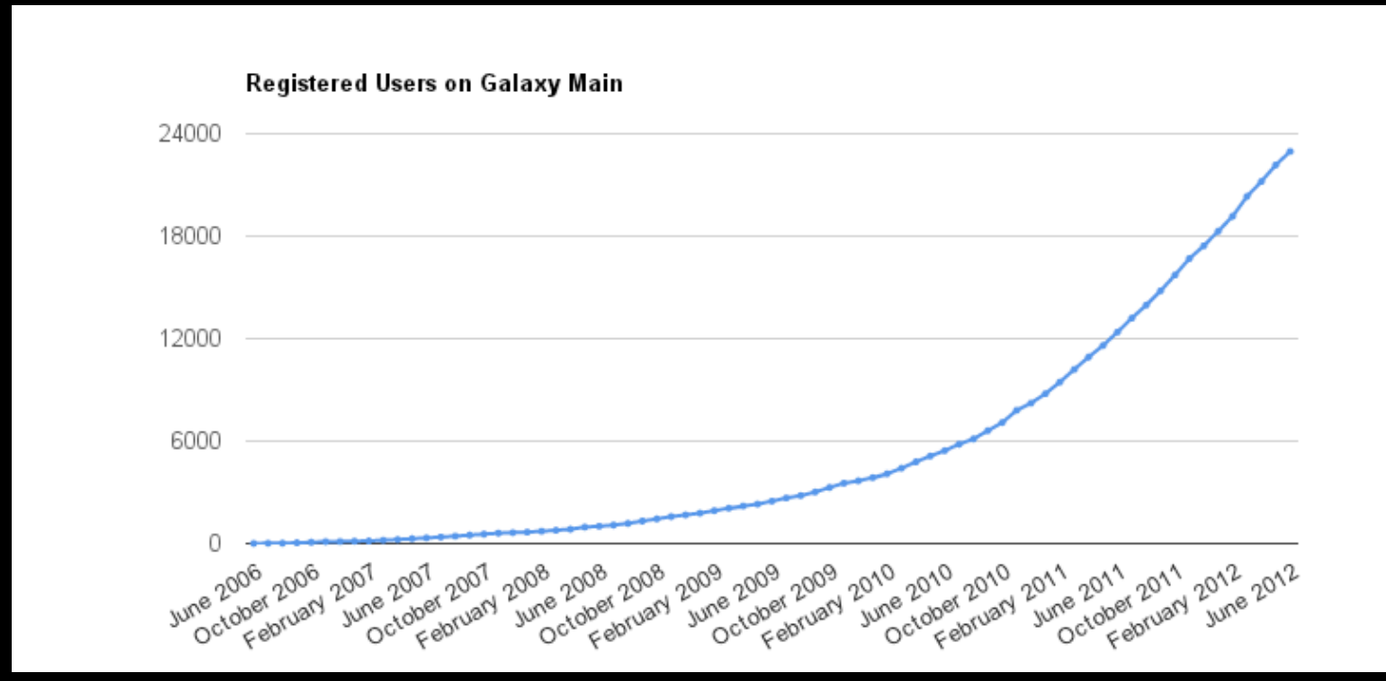
- A **data analysis and integration** tool
- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- These options result in several **ways to *use Galaxy***

http://galaxyproject.org

http://usegalaxy.org
(a.k.a Main)

- **Public web site**
- **Anybody can use it**
- Hundreds of tools
- **Persistent**
- + 500 users / month
- ~100 TB of user data
- ~140,000 jobs / month





http://bit.ly/gxystats

But, it's a big world

Main has lots of tools, storage, processor, users, ...

- But **not all tools** - there are thousands and adding new tools is not taken lightly
- But **not infinite storage and processors** - Main now has job limits and storage quotas

A centralized solution cannot scale to meet data analysis demands of the whole world

Scaling Galaxy & Options

- **Encourage local Galaxy instances and Galaxy on the cloud**
- Support **increasingly decentralized model** and *improve access to existing resources*
- Focus on building **infrastructure to enable the community to integrate and share** tools, workflows, and best practices
- Support Community & Alternate Public Galaxy

Local Galaxy Instances
http://getgalaxy.org

Galaxy is designed for local installation and customization

- Easily integrate new tools
- Easy to deploy and manage on nearly any (Unix) system

Got your own cluster?

- Move tool execution to other systems
- Galaxy works with any DRMAA compliant cluster job scheduler (which is most of them).
- Galaxy is just another client to your scheduler.

CLUSTER RESOURCES
TORQUE

GRID ENGINE

Platform Computing

DRMAA
Distributed Resource Management Application API — www.drmaa.org

Galaxy Tool Shed

- Allow users to share "suites" containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Integration with Galaxy instances to automate tool installation and updates

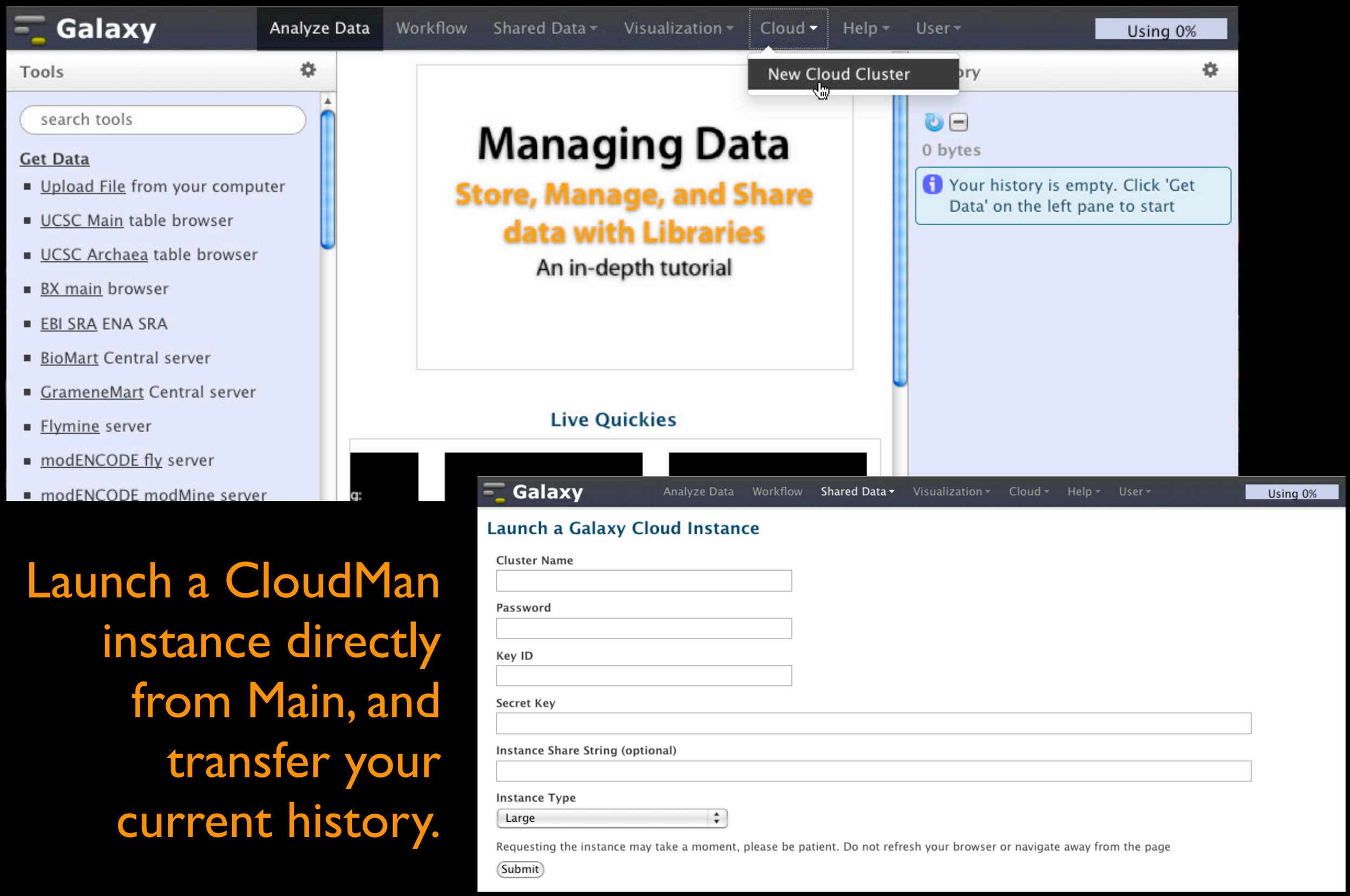
toolshed.g2.bx.psu.edu

Galaxy CloudMan
http://usegalaxy.org/cloud

- Start with a **fully configured and populated** (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center.
- **We used a Galaxy cloud *tonight* ...**

http://aws.amazon.com/education

Instant CloudMan



Launch a CloudMan instance directly from Main, and transfer your current history.

Public Galaxy Servers

Interested in:

ChIP-chip and ChIP-seq?

✓ Cistrome

Statistical Analysis?

✓ Genomic Hyperbrowser

Sequence and tiling arrays?

✓ Oqtans

Text Mining?

✓ DBCLS Galaxy

Reasoning with ontologies?

✓ GO Galaxy

Internally symmetric protein structures?

✓ SymD

http://galaxyproject.org/wiki/PublicGalaxyServers

Galaxy Community

Tool Shed
Mailing Lists (Sci User, Development, Private Data help)
Screencasts (Community and Galaxy Team)
Events Calendar, News Feed, Twitter, Monthly Updates
Distributions & Release Notes/Feature Descriptions
Community Wiki
Local Public Installs
CiteULike group, Mendeley mirror
Annual **Galaxy Community** Meeting

http://galaxyproject.org/wiki

Galaxy Community Conference

http://galaxyproject.org/GCC2013

Agenda

7:00 Welcome, Introduction to Galaxy

7:15 Galaxy 101 Tutorial: Data, Tools, Workflows

8:30 Galaxy Usage, Sharing, Support, and Resources

8:45 Open discussion

9:00 Done

Agenda

9:00 *Done - Thanks for using Galaxy !!*

Acknowledgments

Pauline Minhinnett & ASHG

UC Santa Cruz Genome Browser Group
http://genome.ucsc.edu/

Jeremy Goecks, Dave Clements, Dannon Baker, Anton Nekrutenko, James Taylor, & Galaxy Team

Galaxy Community & You!!

Hope to see you in Oslo!

Galaxy Community Conference

30 June - 2 July 2013

OSLO

UiO : University of Oslo

http://galaxyproject.org/GCC2013