

Galaxy Project Update

2014 GMOD Meeting
San Diego,



Dave Clements
Johns Hopkins University

Agenda

- Project Introduction
- Project Update

What is Galaxy?

- A web based data integration and analysis framework.
- **Open source software**

<http://galaxyproject.org>

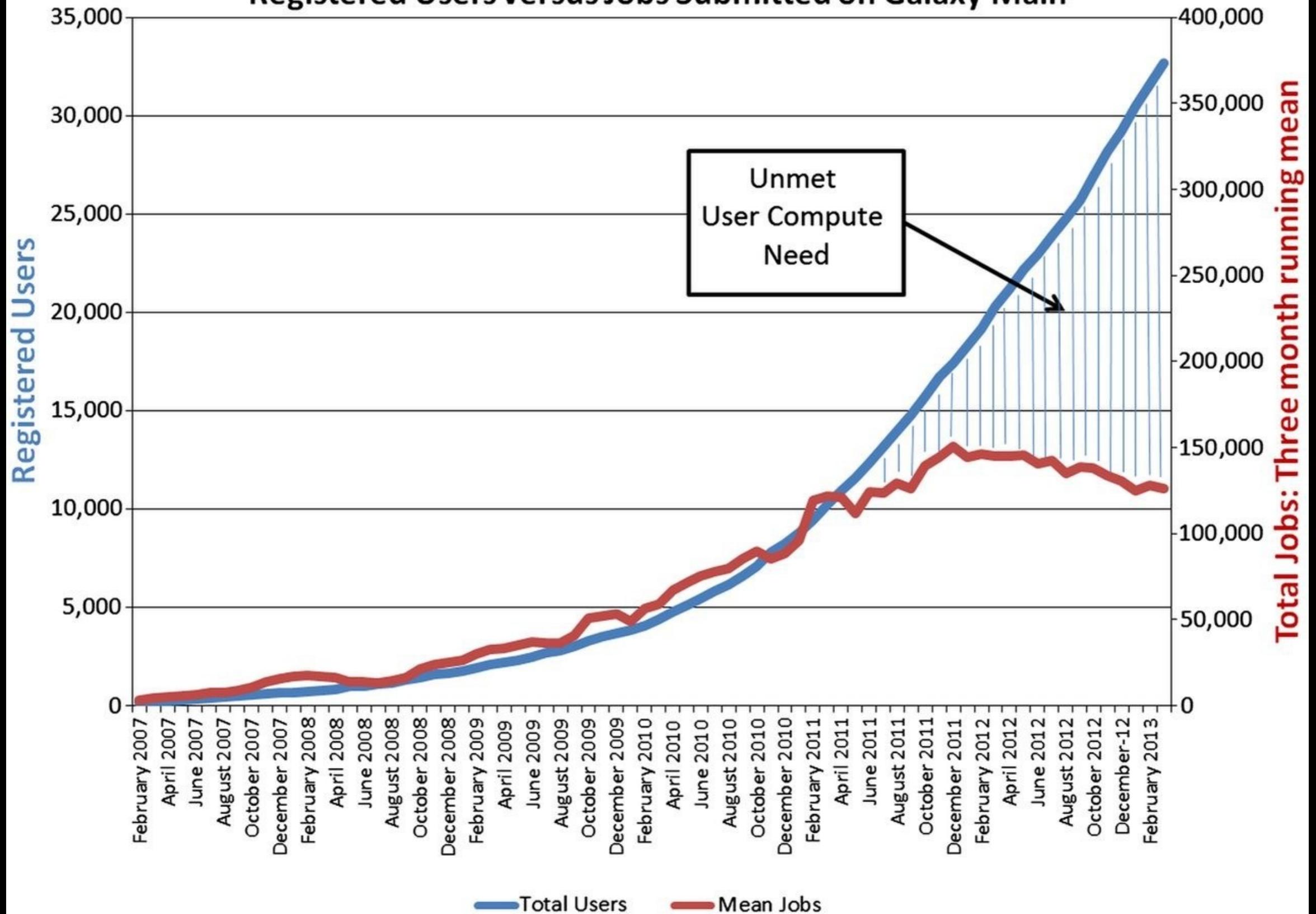
Galaxy is available as

As a free for everyone web service
(<http://usegalaxy.org>) integrating a wealth of tools,
compute resources, terabytes of reference data and
permanent storage.

However, *a centralized solution cannot support
the different analysis needs of the entire world.*

<http://galaxyproject.org>

Registered Users versus Jobs Submitted on Galaxy Main



Leveraging the national cyberinfrastructure for biomedical research
LeDuc, et al. *J Am Med Inform Assoc* doi:10.1136/amiajnl-2013-002059

Galaxy is available as

- As a free (for everyone) web service

<http://usegalaxy.org>

- **As open source software** that you can install locally

<http://getgalaxy.org>

Galaxy is available ...

- As a free (for everyone) web service
- As open source software
- ***On the Cloud***



<http://wiki.galaxyproject.org/Cloud>

AWS in Education Grants Program



<http://aws.amazon.com/education>

Galaxy is available ...

- As a free (for everyone) web service
- As open source software
- On the Cloud
- ***With Commercial Support***



A ready-to-use appliance (BioTeam)

Cloud-based solutions (ABgenomica, AIS, Appistry, GenomeCloud)

Consulting & Customization (Arctix, BioTeam, Deena Bioinformatics)

Agenda

- Project Introduction
- Project Update

New Development

- **Tool Shed Work**
 - Lots of enhancements to handle dependencies.
 - Continue moving tools out of source distribution and into Tool Shed

New Development

- **User Management**

- Disk quotas added a few years ago
- Limited execution concurrency before that
 - Led to widespread abuse w/ multiple accounts
- Put effort into ending that on Main
 - Email verification and duplicate detection
 - Added supporting code into distribution

Release Cycle: Galaxy

Experimented with 2-3 week release cycle:

2013/01/11 Release

2013/02/08 Release

Aiming for every 2 months or so

2013/04/01 Release

2013/06/03 Release

2013/08/12 Release

2013/11/04 Release

Releases

Project now makes extensive use of Trello cards for tracking work. This is reflected in release notes:

Core

1. Explicitly set TEMP dir in Local Runner, when a temp dir value is not already set.
<https://trello.com/c/HbFeoWRI>
2. Tool element return_code (under stdio) now functions from_work_dir or when setting metadata externally. <https://trello.com/c/JfB2w1Br>
3. Using Auto-detect and a cluster job runner now sets metadata only once.
<https://trello.com/c/Kc3NDGyN>
4. Upgrades to HierarchicalObjectStore, more planned. <https://trello.com/c/k4tovIFd>
5. New Plugin Framework lib/galaxy/web/base/pluginframework.py. <https://trello.com/c/lrfWbtw3>
6. Plugins define hook functions called by a Galaxy app when certain events/situations happen.
<https://trello.com/c/c2AzV3Xf>

Releases

Doing a much better job of incorporating pull requests from community into dist:

Pull Requests Merged

1. Björn Grüning contributed a method to implement the ability to change the tool-panel as user preference (Dynamic Toolbox Filtering). #179. This was a frequently requested feature by the community and full documentation on this can be found here [UserDefinedToolboxFilters](https://trello.com/c/XI7CZFMd). <https://trello.com/c/XI7CZFMd>
2. Björn Grüning also contributed several extensions allowing developers to utilize new actions simplifying various tool shed dependency definition idioms:
 - `make_install` action. #217
 - `autoconf` action. #218
 - `setup_r_environment` action. #219 Further extensions enhancing this last tag and a corresponding `setup_ruby_environment` tag from Björn will be forthcoming in the next release.
3. Additionally, Björn Grüning contributed other tool shed and tool related enhancements: #205, #216, and #239
4. Andrew Warren contributed an API method allowing coping datasets between histories as well as support for more secure e-mail settings. #199 and #198.
5. Nicola Soranzo contributed small fixes for various tools as well as enhancements for customizing and localizing data and time display in various parts of Galaxy. #222 and #211.
6. Kyle Ellrott contributed many enhancements for the API and the Galaxy search engine. #187, #241, and #234.
7. Lance Peterson contributed two enhancements to management scripts. #196 and #158. <https://trello.com/c/qzjBuljp>
8. Google Summer of Code Intern Saket Choudhary contributed enhancements for VCF 4.1 compatibility. #184.
9. Matthew Shirley contributed grammar fixes to the tool shed interface. #210.
10. Stephen McMahon contributed fixes to the PBS job runner's staging functionality. #194
11. Rémy Darnat contributed enhancements to the administrative interface allowing for management of user API keys. #134
12. Adam Brenner contributed an enhancement making it easier to deploy the histogram2 tool. #215.
13. A. Rretaud contributed extensions enabling data source tool developers to utilize the tool runners login e-mail address when implementing such tools. #206
14. John Chilton fixed job splitting to rewrite references in config files in addition to command-line. #169. <https://trello.com/c/FMPyde8L>
15. John Chilton and Simon Guest implemented configurable plugins for tool dependency resolution. #228. <https://trello.com/c/cP3tGSJv>
16. John Chilton implement GALAXY_SLOTS allowing tools to uniformly obtain allocated thread count. #236. <https://trello.com/c/cfOISfdP>
17. Kyle Ellrott contributed enhancements that allow API tool's POST to define history for tool state. #193. <https://trello.com/c/hpFanyx0>

Release Cycle: CloudMan

After a lull, now doing semi-annually

2013/07 Release

2014/01 Release

Where are we going?

Big Data: Supporting Analysis on a Massive Scale

Common request: run tools / workflows on many samples

Run each of a few dozen (paired) samples through a workflow of several dozen steps, and aggregate the results in some way

A simple analysis quickly results in dozens of workflow invocations and hundred of individual tool runs

Big Data: Plans

Rewrite default workflow engine

Histories will be able to contain pending workflows, dataset groups, other entities - not just datasets

Rather than scheduling all at once, monitor workflow progress, allow pausing in response to failure or user intervention, decision nodes, streaming data and intermediate datasets, ...

Make workflow scheduling engine pluggable

Once it is a background process, can afford the time to delegate

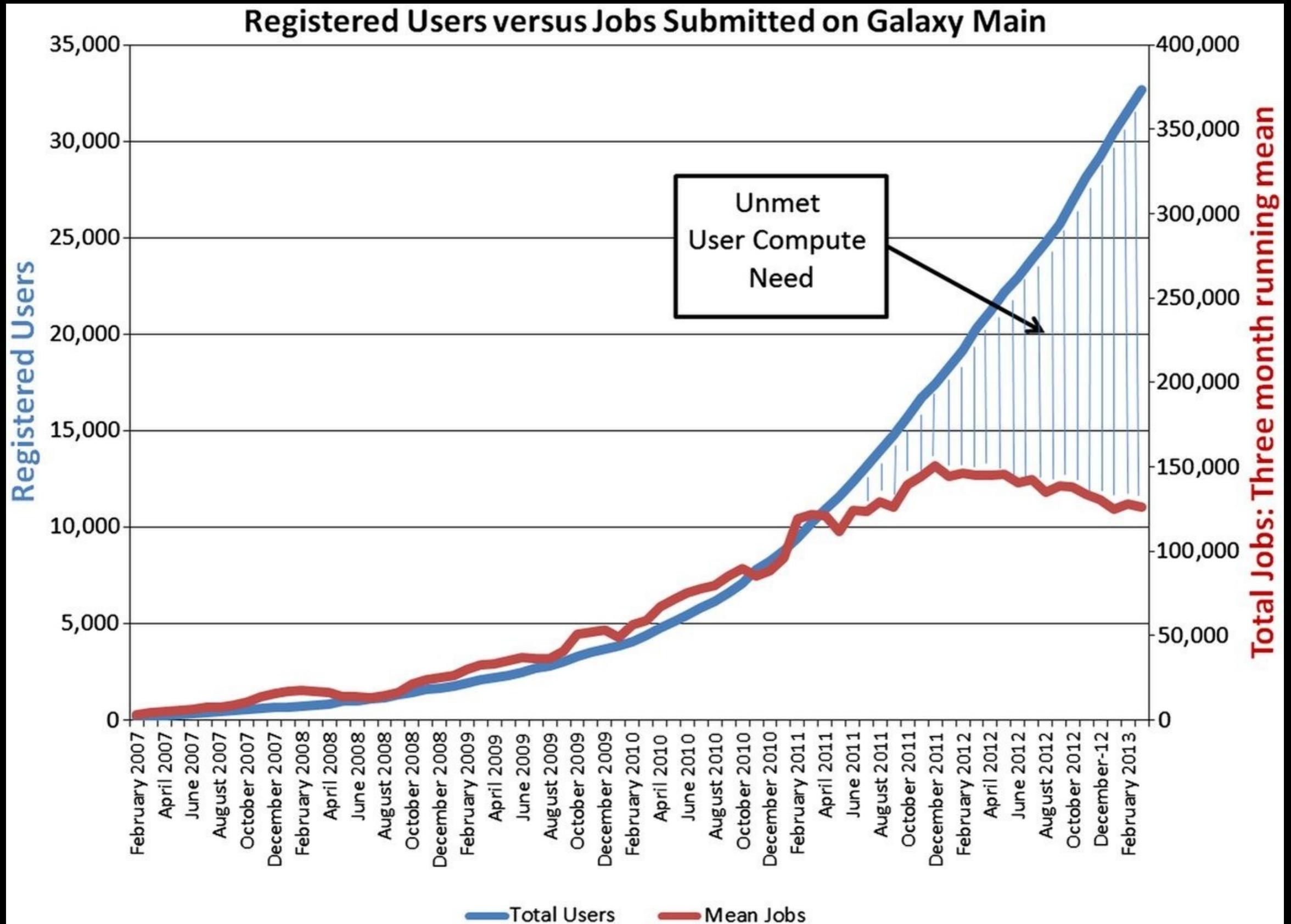
Pluggability / Extensibility / APIs

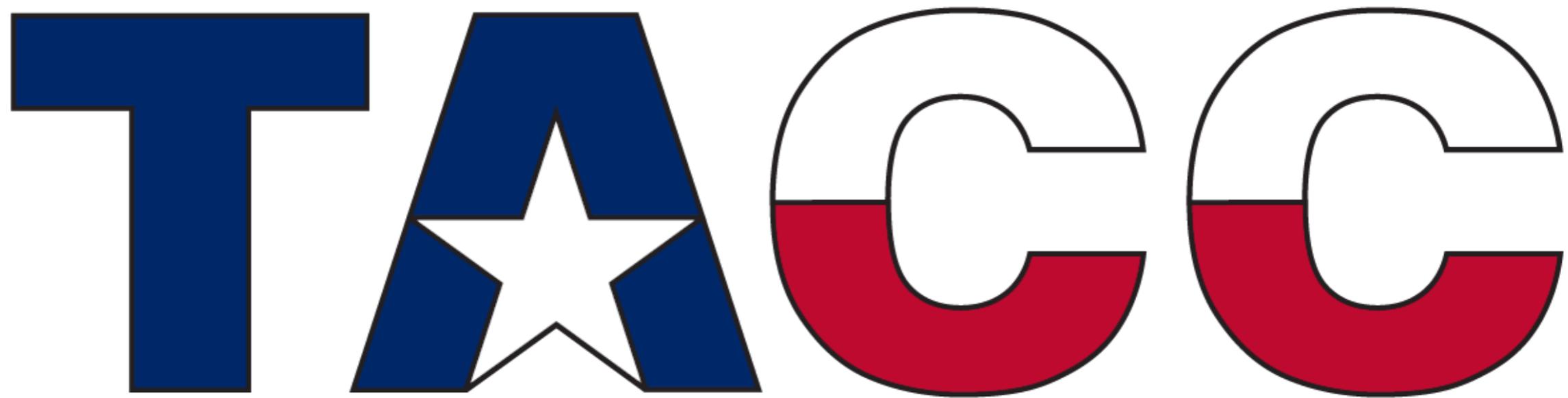
- Workflow rewrite
- Visualization framework
- ObjectStore storage api
- Galaxy API
- ...
- Make everything pluggable; start using those interfaces internally.

Galaxy toolshed vision

- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Version controlled
- Community annotation, rating, comments, review
- Dependency resolution
- Integration with Galaxy instances to automate tool installation and updates
- A key to intergalactic federation
- Intergalactic Utilities Commission

Meeting the Need





usegalaxy.org moved to the Texas Advanced
Computing Center on October 8

But it is still not enough to
meet the analysis needs of the world.

The Galaxy Project continues to emphasize
Cloud Installs
Local Installs
Public Servers

Community

Community Hubs

Galaxy Deployment Catalog



Welcome to the *Galaxy Deployment Catalog*. This catalog describes the details of how Galaxy is installed at different institutions. Details include [infrastructure information](#) as well as [user community and domain information](#) for each deployment.

You are strongly encouraged to [add your local Galaxy deployment](#) to this catalog so that the [Galaxy Community](#) can benefit from your experience.

Contents

1. [Deployments: User and Domain Information](#)
2. [Deployments: Implementation](#)
3. [Add Your Galaxy Deployment](#)
 1. [Using the Wiki](#)
 2. [Using the Google Form](#)

Deployments: User and Domain Information

Deployment	Domain	Owners	User Base	Audience
CSIRO Galaxy Service	NGS analysis: RNAseq, Genomics, metagenomics, custom tools.	CSIRO	209 registered users and 30 active users	
Galaxy server at the FMI	NGS analysis, MA analysis, custom tools.	Friedrich Miescher Institute for Biomedical Research		wet lab scientists
GalaxyAtUIowa	Local instance at the U of Iowa	IIHG		Human geneticists, biologists
Idaho State University MRCF	Bioinformatics, biology, next generation sequencing. Custom tools.	Idaho State University Molecular Research Core Facility (MRCF)		Idaho State University and Molecular Research Core Facility (MRCF) customers.
Sigenae Bioinfo-Genotoul	Genomics, esp. read alignment, SNP calling and annotation, RNA-Seq, and sRNAseq	Sigenae Team, GenoToul Bioinfo		bioinfo gentoul and INRA researchers and collaborators
URGI	Genomics	INRA	~ 30	researchers at the institution, lab members
UAB Galaxy	NGS analysis: Genomics (reference and de novo), RNAseq, metagenomics, custom tools.	Collaboration between Center for Clinical and Translational Science and Research Computing	currently ~200	Informatics core, sequencing core, researchers and students at the institution
ZBIT	Bioinformatics: SBML tools,	Center for Bioinformatics Tuebingen		



Training Day voting closes Friday

Use Galaxy

[Servers](#) • [Learn Main](#) • [Share](#) • [Search](#)

Communicate

[Support](#) • [News](#) [Events](#) • [Twitter](#) [Mailing Lists \(search\)](#)

Deploy Galaxy

[Get Galaxy](#) • [Cloud Admin](#) • [Tool Config](#) [Tool Shed](#) • [Search](#)



Contribute

[Tool Shed](#) • [Share Issues & Requests](#) [Teach](#) • [Support](#)

Galaxy Project

[Home](#) • [About Community](#)

Community Hubs



GalaxyAtUIowa

This is a local installation of Galaxy at the University of Iowa.

Domain

Human genetics, biology. Custom tools and reports.

Community

The University of Iowa community and members of the [Iowa Institute of Human Genetics](#).

Compute Infrastructure

This instances of Galaxy runs on a local [HPC cluster](#).

Compute

3600 cores, SGE DRM.

Storage

Deployment GalaxyAtUIowa

Domain

Local instance at the U of Iowa

Owners

IIHG

Audience

Human geneticists, biologists

User Base

Server Topology

SGE cluster

Compute

Memory

Storage

NFS

Disk Space

User Management

University or IIHG affiliation required



BALTIMORE, MD | JUNE 30 - JULY 2, 2011

Training Day voting closes Friday

Use Galaxy

[Servers](#) • [Learn](#)
[Main](#) • [Share](#) • [Search](#)

Communicate

[Support](#) • [News](#) 
[Events](#) • [Twitter](#)
[Mailing Lists \(search\)](#)

Deploy Galaxy

[Get Galaxy](#) • [Cloud](#)
[Admin](#) • [Tool Config](#)
[Tool Shed](#) • [Search](#)



Contribute

[Tool Shed](#) • [Share](#)
[Issues & Requests](#)

Community Hubs



The *Galaxy Community Log Board* is a place to share how you addressed a particular task in your Galaxy deployment. Log entries describe specific solutions to particular tasks, such as the details of what steps were taken to deploy Galaxy on particular platforms, or with particular software. It's an easy way to help others (and learn from others) by sharing what you've already done.

So, if you have figured out how to do something, and it took you a while, then this is the ideal forum for sharing that information with the community.

Logs

2014

Please use a more selective search term instead of search_term="regex:title:^Community/Log/2014/"

2013

Date	Topic	Resolution	Who
2013/12/05	Tool Integration Short Tutorial	Documents best tool integration practices from Institut Français de Bioinformatique Galaxy working group	Contributors
2013/11/13	VelvetG error on CloudMan instance: cannot find 'cov_cutoff'	Fixed XML wrapper	Dave Clements
2013/10/27	IGV Integration	How to set up IGV-Galaxy integration in Apache and Galaxy, including adding custom genomes.	Sarah Maman, Nabihoudine Ibou
2013/10/23	Sharing workflows between servers	Documents process for sharing workflows, including wrapping necessary tools	Contributors

Contents

- 1. [Logs](#)
 - 1. [2014](#)
 - 2. [2013](#)
- 2. [Add a Log Entry](#)
 - 1. [Use the Wiki](#)
 - 2. [Use the Online Form](#)



[Training Day voting g](#)
[closes Friday](#)

Use Galaxy

[Servers](#) • [Learn](#)
[Main](#) • [Share](#) • [Search ch](#)

Communicate

[Support](#) • [News](#)
[Events](#) • [Twitter](#)
[Mailing Lists \(search\)](#))

Deploy Galaxy

[Get Galaxy](#) • [Cloud](#)
[Admin](#) • [Tool Config](#)
[Tool Shed](#) • [Search](#)



Contribute

[Tool Shed](#) • [Share](#)
[Issues & Requests](#)
[Teach](#) • [Support](#)

Galaxy Project

[Home](#) • [About](#)
[Community](#)
[Big Picture](#)

Wiki

[Help](#) • [All Pages](#)
[Recent Changes](#)

Community: Public Galaxy Instances

<http://bit.ly/gxyServers>



Passed 50 This year

OSDD

CoSSci

Galaxy-P

GO Galaxy

Orione

GenOquest Tool Shed

Events: 2013

200+ Talks, workshops, tutorials, ...

90+ Events

80% With archived slides, video, exercises, ...

75% Presented by the **Galaxy Community**

Community Organized Events

Galaxy Day (France)

A swarming presence at

ASMS, eResearch Australasia

19 days of workshops from UC Davis Core

14 days of workshops from INRA (France)

11 workshops from QFAB (Australia)

6 days of workshops from PRABI (France)

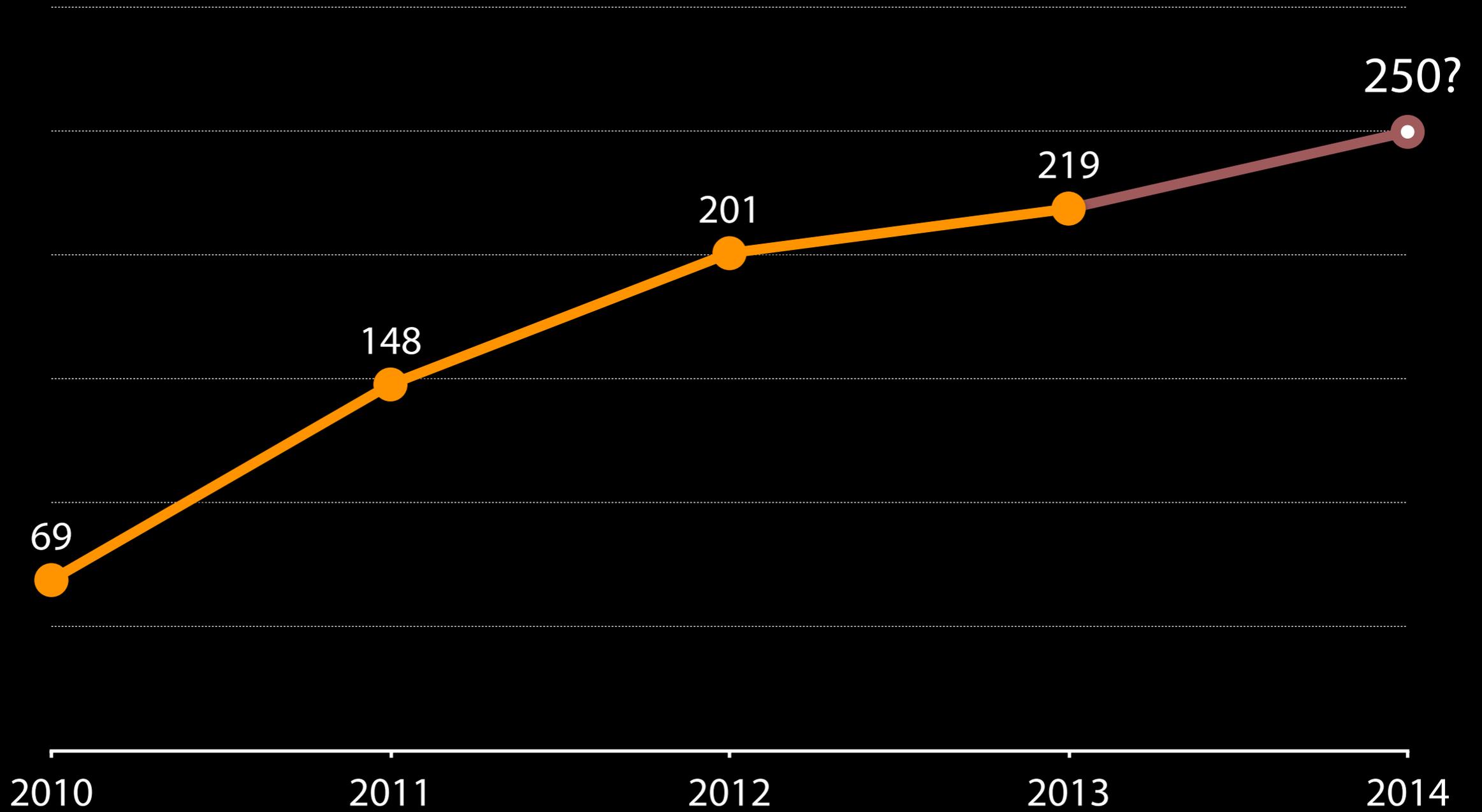
GCC2013



Added Poster sessions
Longer training sessions (2 hrs)
Breakout reformed into BoFs
2nd year with sponsors
Had a pub onsite!



GCC attendance over time





GALAXY

**COMMUNITY
CONFERENCE**

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Poster sessions, BoFs, Sponsors
Longer training sessions (2.5 hrs)

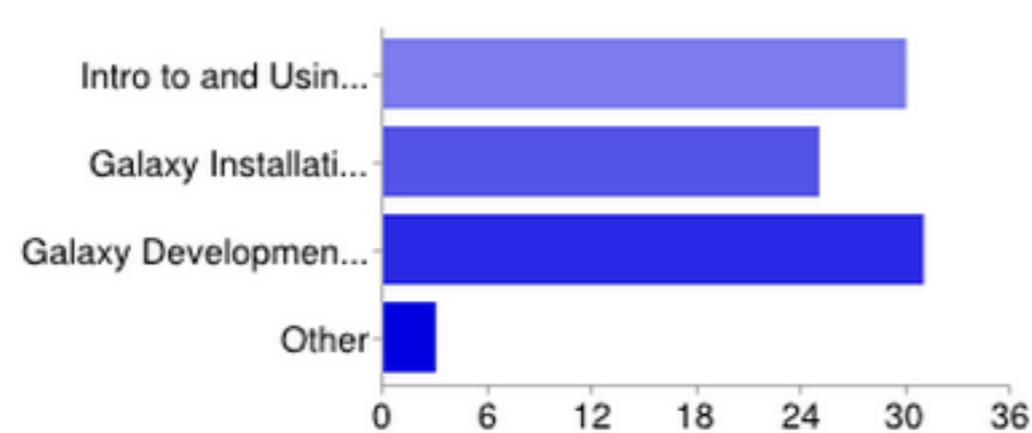
Adding Hackathon

Training Day Topic Voting ends Friday

Registration opens February 10

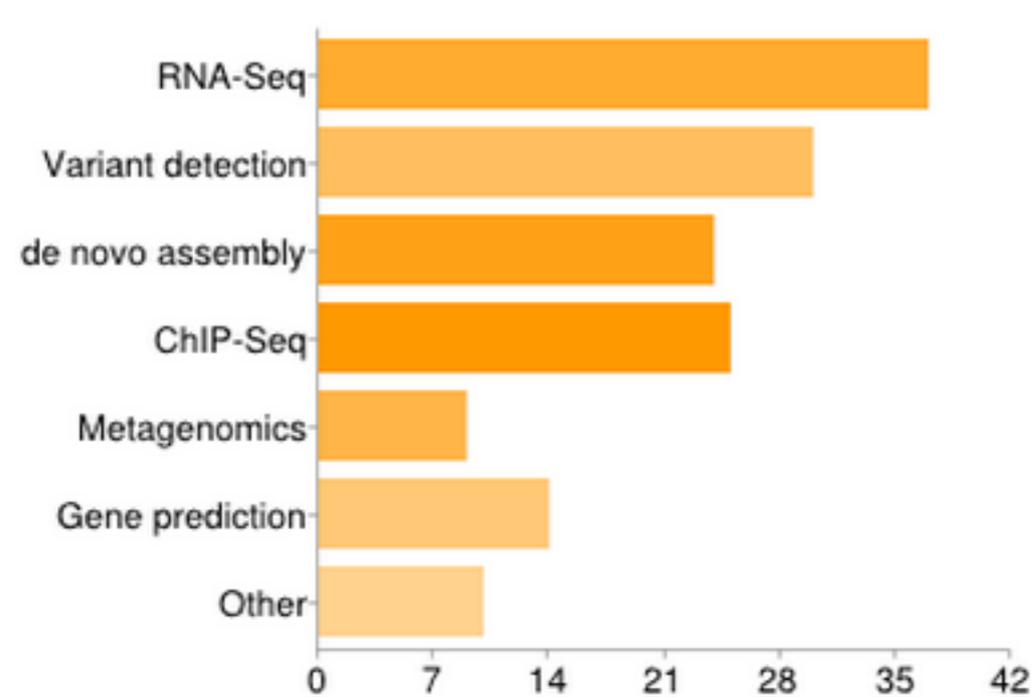
Training: Survey

What type of training are you most interested in?



Training Type	Count	Percentage
Intro to and Using Galaxy	30	34%
Galaxy Installation and Administration	25	28%
Galaxy Development and Tool Integration	31	35%
Other	3	3%

What type of data analysis are you interested in?



Data Analysis Type	Count	Percentage
RNA-Seq	37	25%
Variant detection	30	20%
de novo assembly	24	16%
ChIP-Seq	25	17%
Metagenomics	9	6%
Gene prediction	14	9%
Other	10	7%

Training in 2014: Scalability

Create a Training Network?

Make it easy to find and use workshops created by anyone.

Training AMIs, VMs

Look serious at MOOCs and Screencasts

Project

Core Team is at 18 people.

1 in Croatia

1 in Australia

10 with Penn State University

5 with Johns Hopkins University

1 with George Washington University

The Galaxy Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Marten Cech



John Chilton



Dave Clements



Nate Coraor



Carl Eberhard



Dorine Francheteau



Jeremy Goecks



Sam Guerler



Jen Jackson



Ross Lazarus



Anton Nekrutenko



Nick Stoler



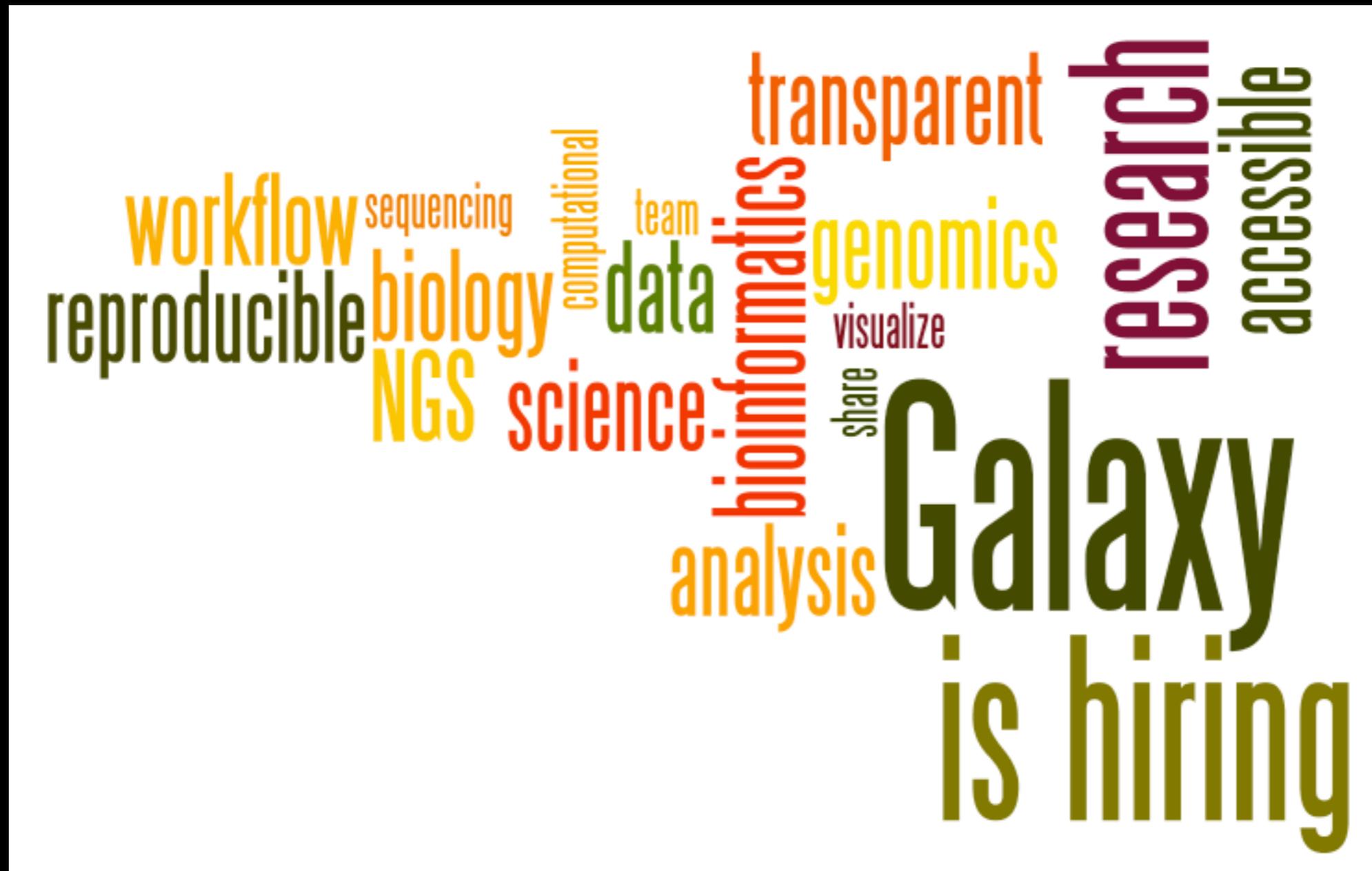
James Taylor



Greg von Kuster

<http://wiki.galaxyproject.org/GalaxyTeam>

Galaxy is hiring post-docs and software engineers
at both Emory and Penn State.



Please help.

<http://wiki.galaxyproject.org/GalaxyIsHiring>

Acknowledgements

GMOD:

Scott Cain

Amelia Ireland



Thank You!