Galaxy Australasia Workshop 2014

Galaxy

past, present and future

Nate Coraor
The Galaxy Team
Penn State University

Galaxy is...

- getgalaxy.org
  - Free, open source software
  - Bring your own compute, storage, tools
  - Maximize privacy and security
- usegalaxy.org
  - Was: Galaxy Main, main.g2.bx.psu.edu
  - 4,000+ jobs/day, 100,000+ jobs/month
- usegalaxy.org/cloud
  - Galaxy cluster in Amazon EC2
  - Buy as much compute, storage as you need
  - More on Galaxy cloud from Enis Afgan
- A thriving community of scientists, developers, bioinformaticians, admins, students

Past

- Development timeline
- usegalaxy.org history
- Community growth

Etymology

- *GALA* = Genome Alignment and Annotation Database
- *Galaxy* = *Gala* + *XL* (Bob Harris, author of Lastz)
- Brainchild of Ross Hardison
- Taking over the universe was not our *original* intent

Ross Hardison

Development Timeline

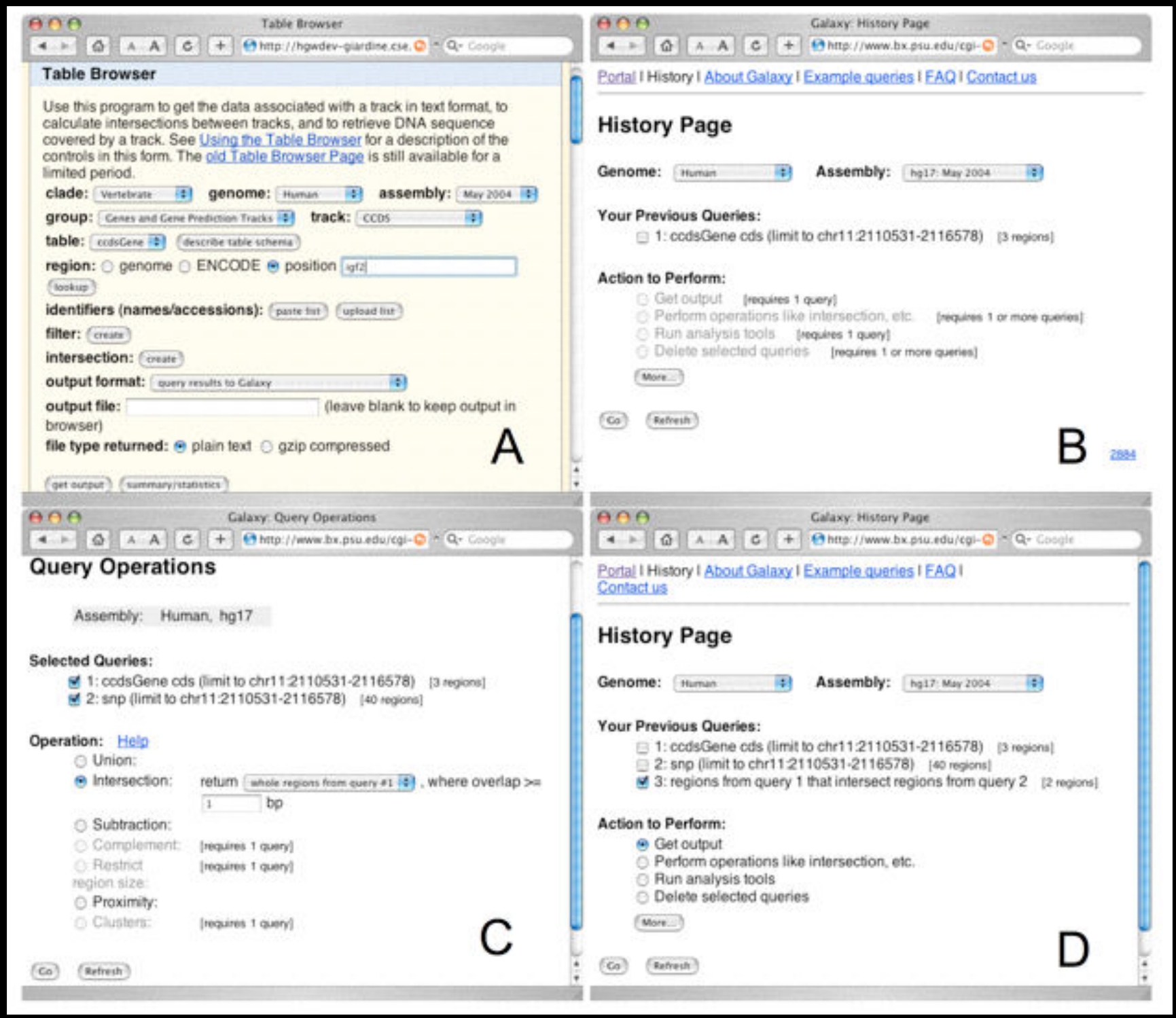
2003
GALA

2005
Galaxy 1

Galaxy: A platform for interactive large-scale genome analysis

Belinda Giardine,[1] Cathy Riemer,[1] Ross C. Hardison,[1] Richard Burhans,[1] Laura Elnitski,[2] Prachi Shah,[1,2] Yi Zhang,[1] Daniel Blankenberg,[1] Istvan Albert,[1] James Taylor,[1] Webb Miller,[1] W. James Kent,[3] and Anton Nekrutenko[1,4]

Galaxy as a single Perl script (!)



Development Timeline

2005
Galaxy 2 (Python)

2005
Multipanel layout
UCSC direct connection
Standardized tool form

Development Timeline

2006

EMBOSS/PHYLIP
ENCODE
Genomic interval operations
User accounts: persistence
PBS job execution
History sharing

Development Timeline

2007

Workflows

Tool repeat/conditional

Pluggable datatypes

Pluggable display apps

jQuery

Advanced metadata

Development Timeline

2008

Short read tools

Evolution, taxonomy,
liftover, metagenomics

Tool versions

Dataset cleanup

Implicit datatype conversion

Dataset abstraction

Grid Engine support

Development Timeline

2009

NGS: bwa, samtools, tuxedo suite

Data libraries

Data security (roles, permissions)

Trackster

Development Timeline

2010

Tool Shed

API

Tagging

DRMAA job runner

Pages and publishing

Visual analytics

Sample tracking

Development Timeline

2011

Picard, Freebayes, GATK

Tool Shed installs

Data accounting and quotas

Object Store

“Actual user” jobs

Development Timeline

2012

Tool dependencies
Tool lineage
Circster/Sweester
Pause/resume jobs
Cloud launch
Tophat2/Bowtie2

Development Timeline

2013

Data managers
Advanced job config
Job limits

2014

Past

- Development timeline
- usegalaxy.org history
- Community growth

usegalaxy.org is...

- Still the fastest, easiest way to start performing analysis right now
- A public service providing free compute and data storage (you're mad!)
- A collection of hardware located at Penn State University

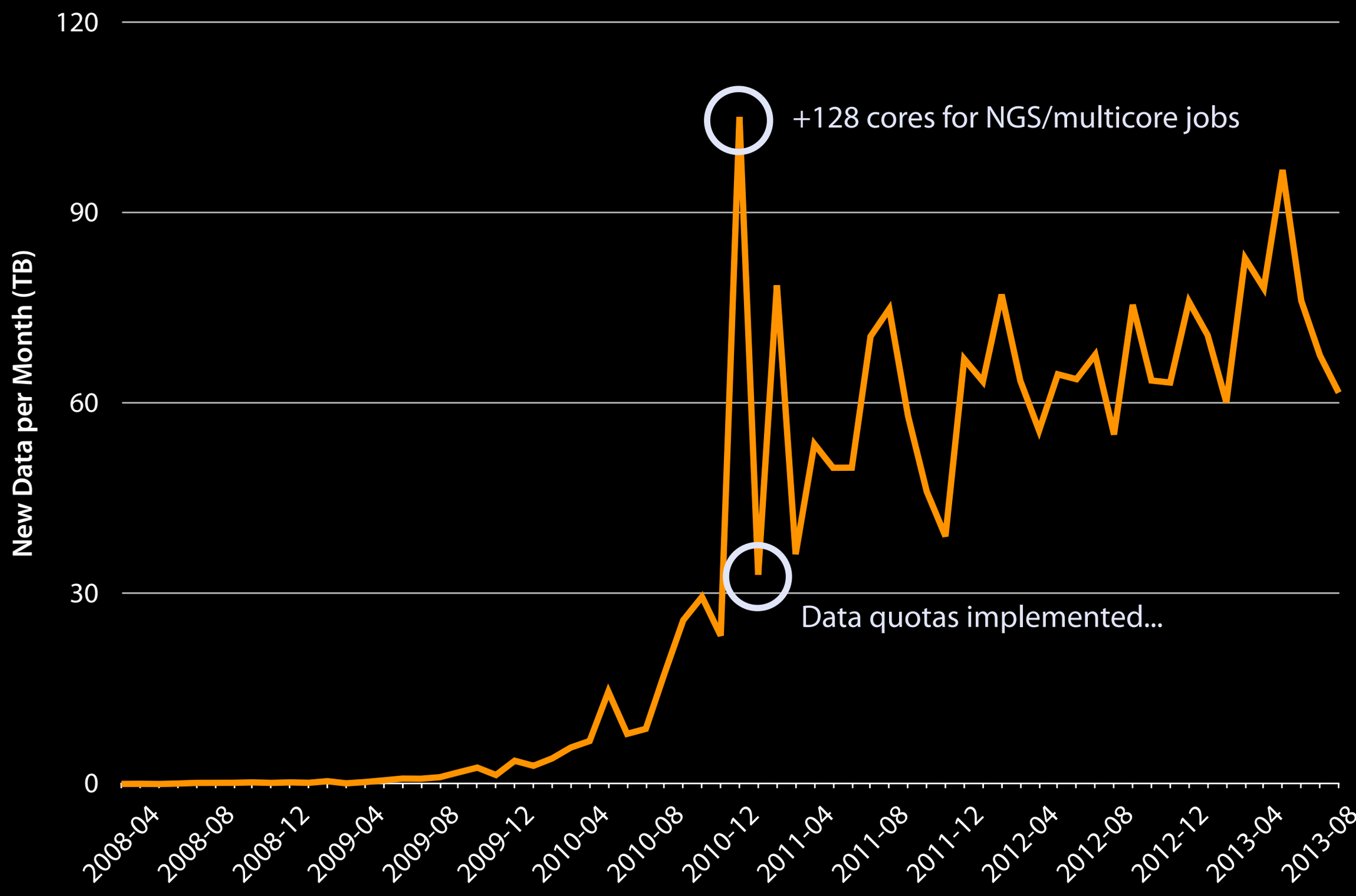
usegalaxy.org Infrastructure

- In 2006
  - 2 CPU, 32 bit, 2 GB RAM, 500 GB server
    - Used for development, Test and Main Galaxy sites, PostgreSQL databases
- In 2007
  - Acquired 8 nodes from decommissioned cluster (2 CPU, 32 bit, 4 GB)
  - Bought an Xserve RAID (7 TB)

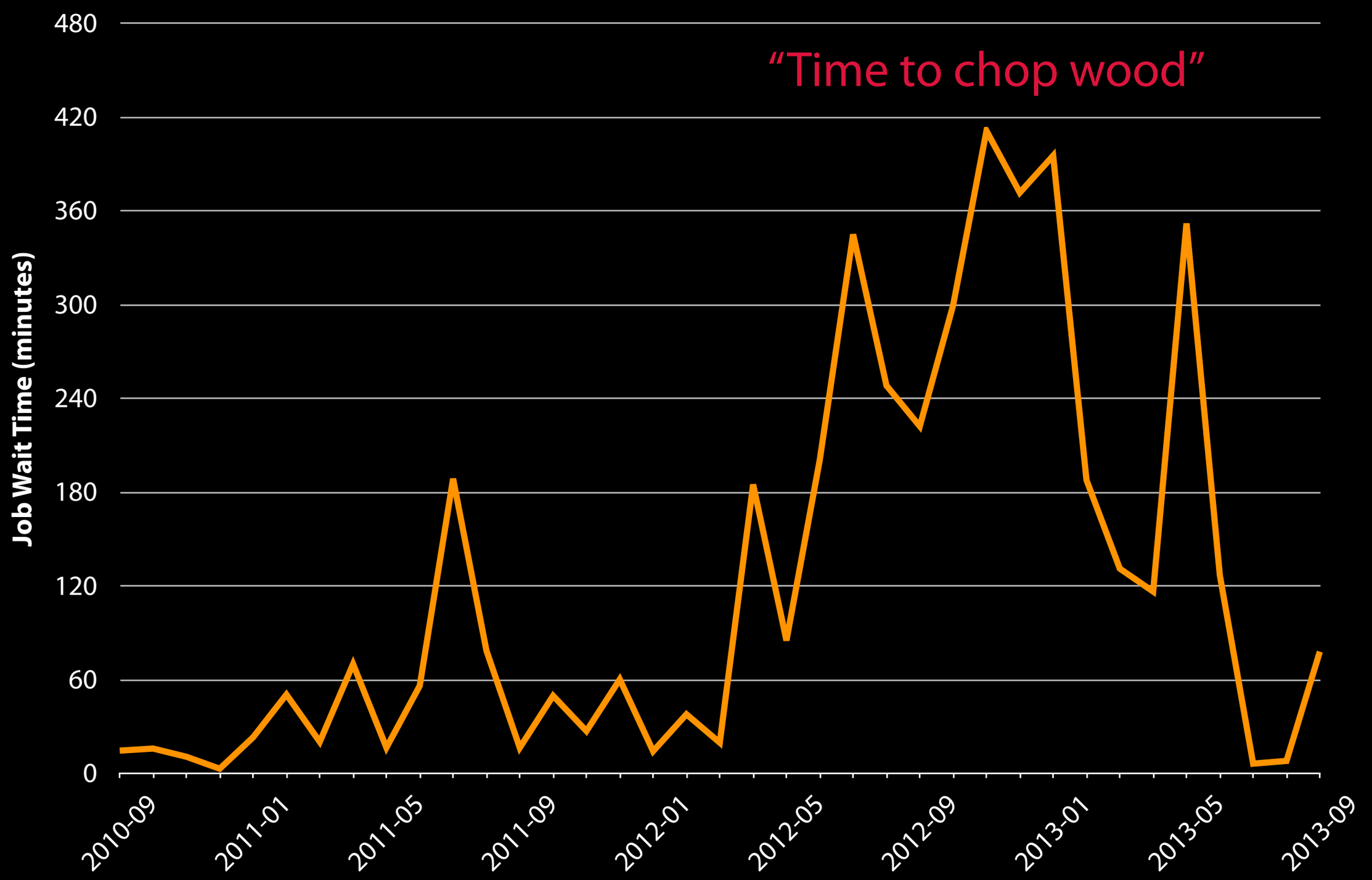
usegalaxy.org Infrastructure

- From 2008 to 2013
  - Server
    - 8 core/32 GB application/db server in 2008
  - Compute
    - 120 core cluster in 2009
    - +28 cores assimilated from SOLiD
    - +128 cores on loan
  - Storage
    - 48 TB in 2008
    - +48 TB in 2009…
    - +24 TB in 2010…
    - +240 TB in 2011…
    - +96 TB in 2012…
  - NIH support for people, but not for hardware

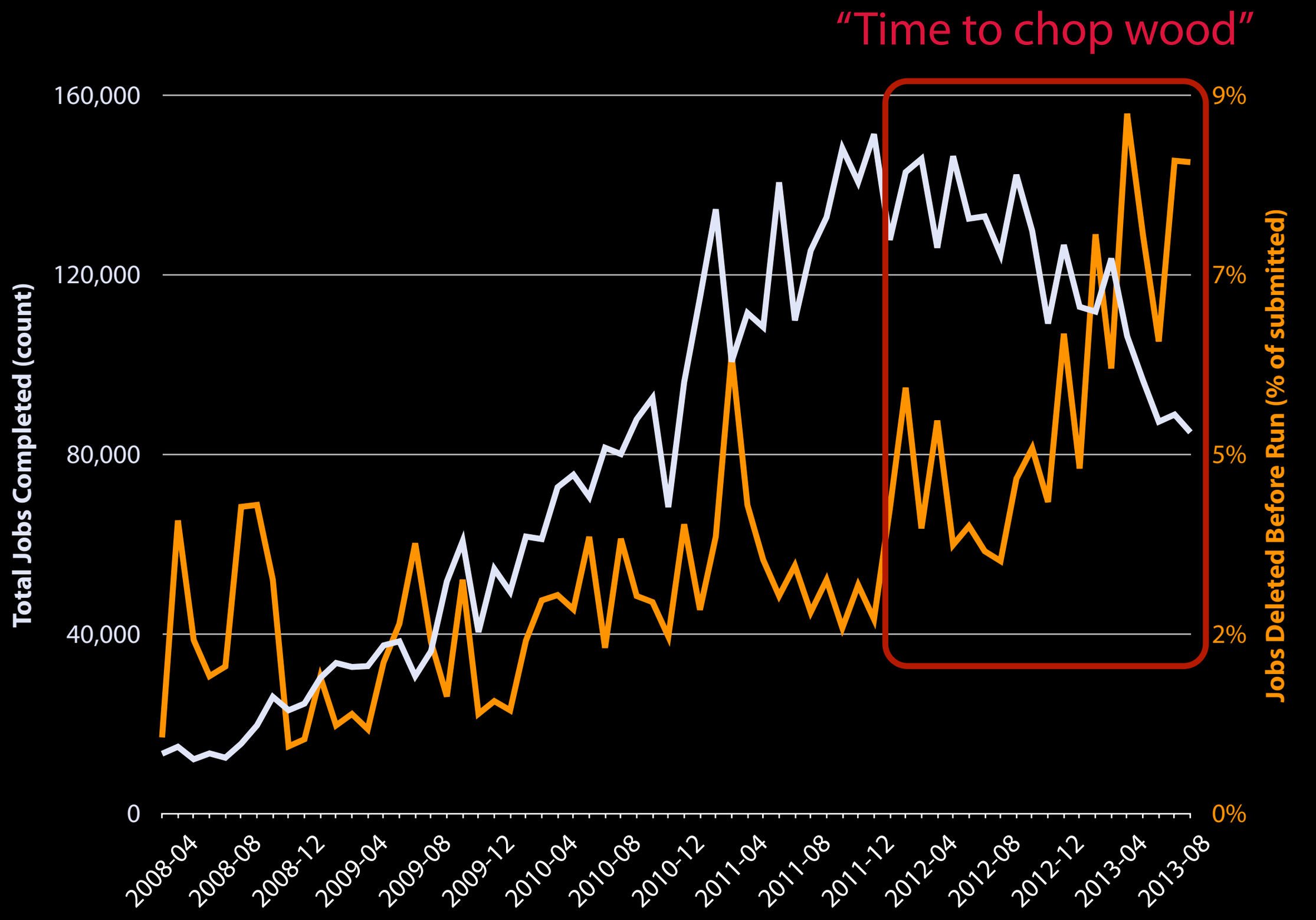
usegalaxy.org data growth



usegalaxy.org queue growth



usegalaxy.org frustration growth

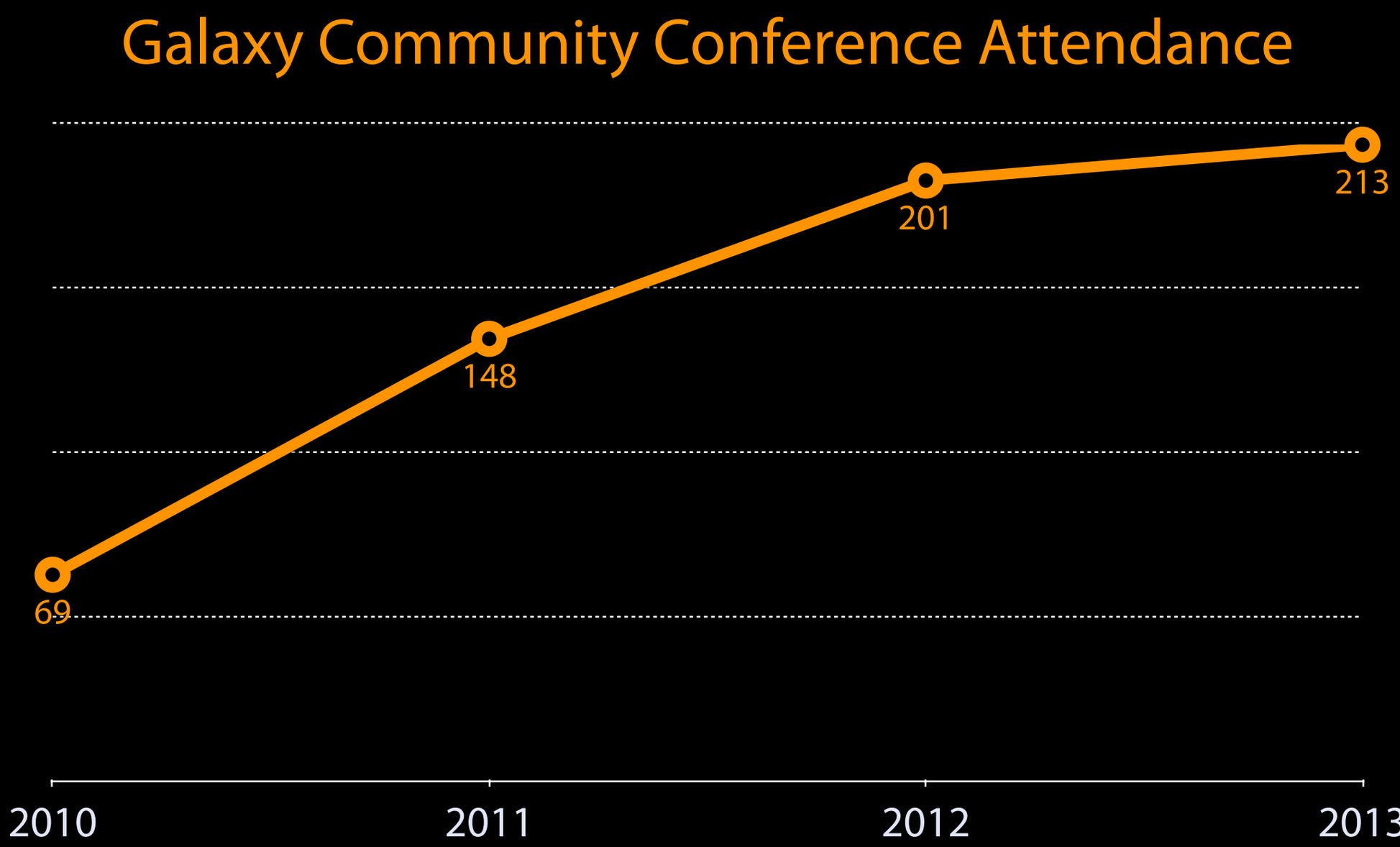


How can usegalaxy.org be sustained while still providing a useful service?

Past

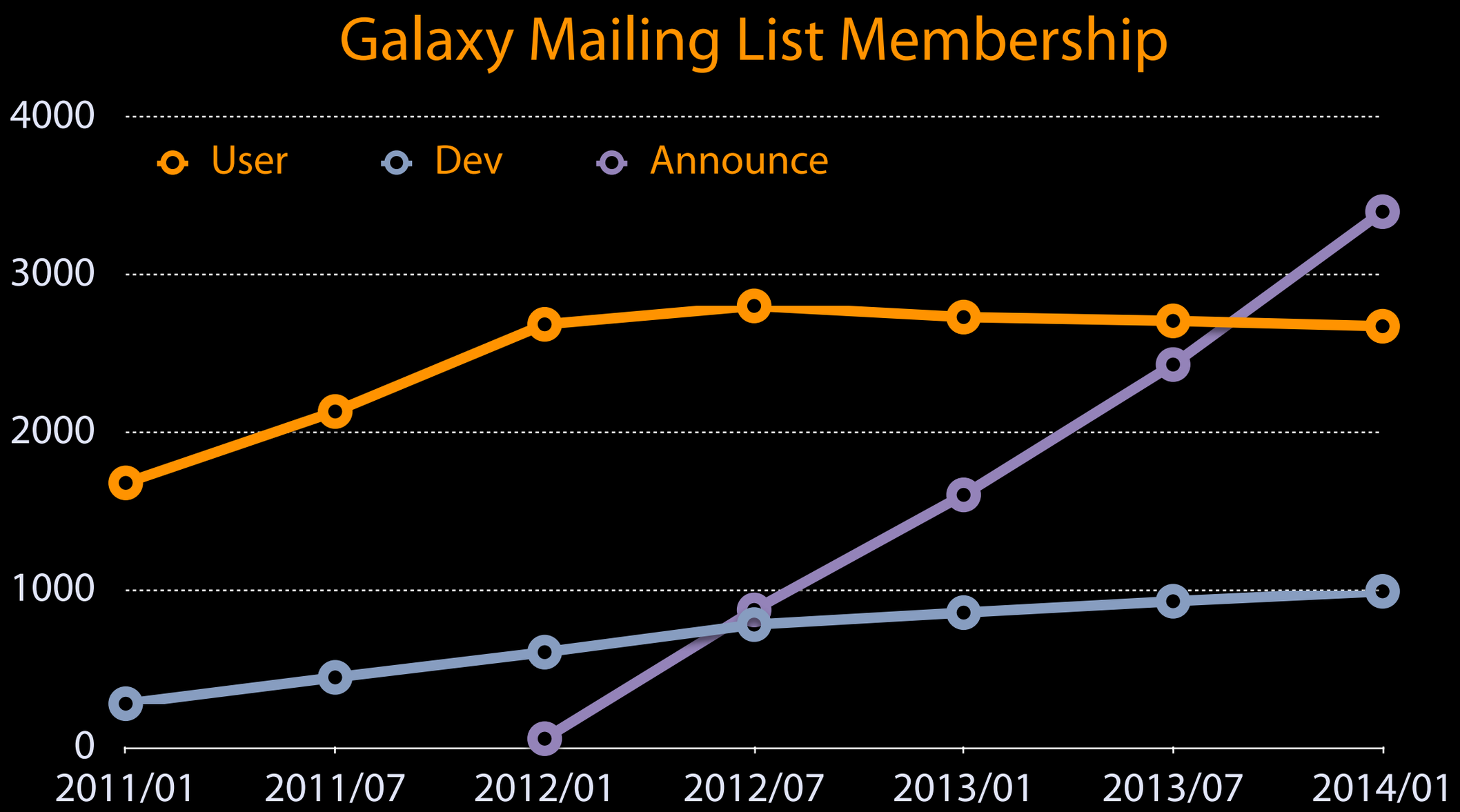
- Development timeline
- usegalaxy.org history
- Community growth

Community Growth



https://wiki.galaxyproject.org/GalaxyProject/Statistics

Community Growth



https://wiki.galaxyproject.org/GalaxyProject/Statistics

Great, now what have you done for me lately?

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Martin Čech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Sam Guerler

Jeremy Goecks

Jen Jackson

Greg Von Kuster

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

http://wiki.galaxyproject.org/GalaxyTeam

Development and Outreach

- Formalized release process
  - 2 month release process
  - 2 week feature freeze prior to release
  - Continual bugfixes to “stable” branch
    - Security releases as necessary
- Releases accompanied by extensive development news
- Galaxy Update
- Workshops and training

Development Transparency

- http://bit.ly/gxytrello
- All of our development goals, feature requests, reported bugs, etc.
- Vote!

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

http://bit.ly/gcc2014

Galaxy SlipStream Appliance

- The first turnkey Galaxy server
- Preconfigured tools, data, “production server” setup

BIOTEAM
Enabling Science

SLIPSTREAM APPLIANCE GALAXY EDITION
Galaxy made easy.

http://bioteam.net/slipstream/galaxy-edition/

Globus Genomics

- Scalable Galaxy services in Amazon Web Services
- Managed and requires no IT expertise
- Data transfer with Globus Online

https://www.globus.org/genomics/

BioStar Integration

- Find answers to usage-related questions
- Get direct help on how to use a tool from the tool interface
- On Galaxy Test now

http://biostar.galaxyproject.org/

Publications from others on Galaxy

~1,500 at CiteULike

349 Pull Requests

59 Public Servers

8 Data Managers

fetch fasta, bowtie, bowtie2, bwa, samtools, blastdb, fasta, 2bit, picard

Tool Shed

- 913 Repositories
- 2,335 Tools
- 226 Owners
- 55 Workflows
- 461 Custom datatypes
- 63,007 Clones/downloads

- Complete tool and dependency management, no sysadmin required

Visualizations and Scratchbook

usegalaxy.org Sustained (for now)

- Direct resources from TACC/iPlant
  - VMs for Galaxy servers
  - 256 cores dedicated to Galaxy jobs
  - Infinite (ha ha) space on Corral (with some drawbacks…)
- Direct resources from PSC
  - Data Supercell
- XSEDE/iPlant Allocations (~1 million SUs)
  - TACC Stampede, Lonestar
  - PSC Blacklight (16 TB shared memory!)
- Job throughput record!

XSEDE Extreme Science and Engineering Discovery Environment

iPlant Collaborative™

TACC

PITTSBURGH SUPERCOMPUTING CENTER

Now what?

Enabling Large-scale Analysis

- What Galaxy does well
  - Interactive, iterative tool process on a small-to-moderate number of datasets

- What Galaxy does less well
  - Repetitive processes on large numbers of datasets

- How it can do better
  - Dataset collections
  - Workflow improvements

Workflows

- Recovery from error
- More intelligent scheduling
- Mid-flow conditionals

Deployment

- Deployment recipe catalog (Ansible)
- Docker
  - Provide Galaxy server and production dependencies (nginx, PostgreSQL, etc.)
  - Tool dependencies(!)
    - Ensure your tools always run on
      - Exact same version of the OS
      - With the exact same system-level deps
      - With the exact same dependencies

usegalaxy.org Sustainability

- TACC Stampede (522,080 cores)
- Per-user XSEDE allocations
- Cloud
  - Cloud burst
  - S3 archiving
  - Data federation

Remaining Priorities

- Tools
  - Update existing tools
  - Add new tools
- Modernize data libraries
- Finish conversion from “legacy” UI to client-side (via backbone.js)
- Continue work on visualizations

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Martin Čech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Sam Guerler

Jeremy Goecks

Jen Jackson

Greg Von Kuster

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

http://wiki.galaxyproject.org/GalaxyTeam