Galaxy for NGS Data Analysis

Matt Shirley

Johns Hopkins Medicine
Pevsner Lab

Slides available at http://mattshirley.com/presentations

matt.shirley@jhmi.edu

Contents

- What is Galaxy?
- Interface elements
- Retrieving data
- Creating and running workflows
- A FASTQ quality statistics workflow
- Galaxy on Amazon Web Services (AWS)
  - Automatic configuration through cloudlaunch
  - Monitoring your AWS charges
  - (optional) Manual configuration through AWS console

matt.shirley@jhmi.edu

Who wants to do this? :(

matt — shirley@pevsner24tb:~/project_SWS/Strelka/inspect_sites — ssh — ⌘1

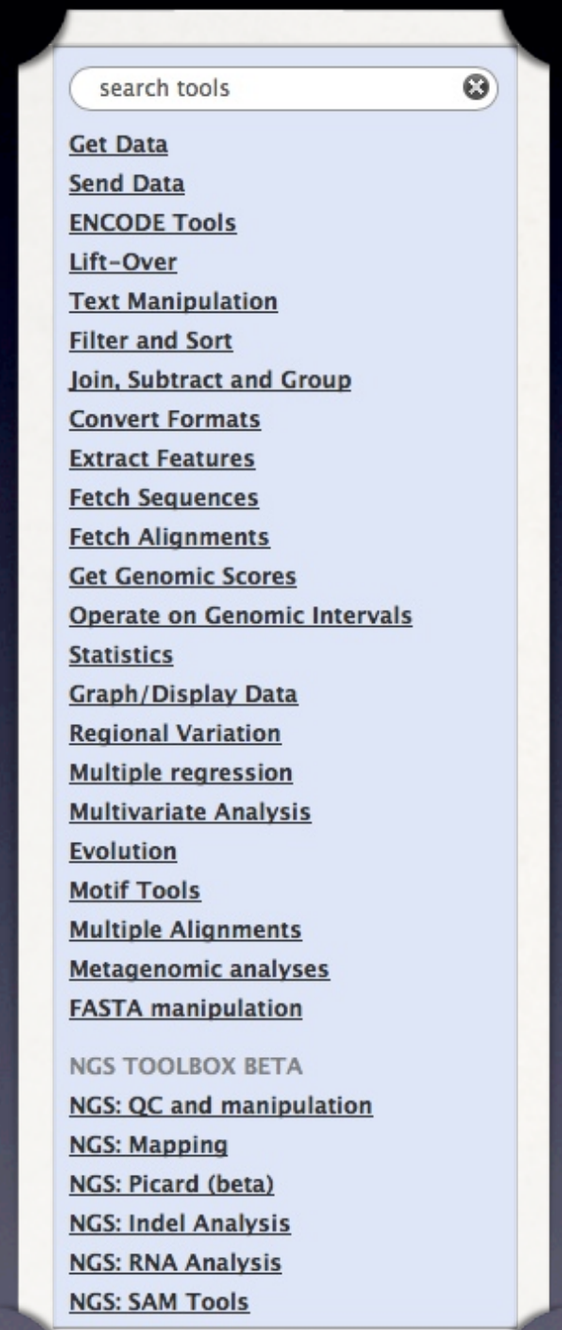
```
File Edit Options Buffers Tools Python Help
#!/usr/bin/env python
# This script will open each sam file specified as an argument,
# fetch the pileup at a site determined somatic by Strelka in at least one sample,
# and print out a read count and allele frequency for the somatic alternate allele.
# Usage: python inspect_variants.py variants [samfile ...]
# Variants is a file with the columns "chrom\tpos\tRef\tAlt"

from __future__ import division
import pysam
import sys
import csv

def printAlleles( samfile, chr, pos, ref, alt ): # Fetch aligned reads for a region from a (b)sam file, return a % alte\
rnate allele
    qfilter = False
    sam = pysam.Samfile( samfile, "rb" )
    flag = 0
    for pileupcolumn in sam.pileup( chr, int( pos ) - 1, int( pos ) ): # Convert from 1-based to 0-based
        flag = 1
        if pileupcolumn.pos == int( pos ) - 1:
            nt = {'A':0, 'C':0, 'G':0, 'T':0, 'N':0}
            for pileupread in pileupcolumn.pileups:
                base = pileupread.alignment.seq[pileupread.qpos]
                if qfilter is True:
                    phred = pileupread.alignment.qual[pileupread.qpos]
                    qualn = ord(phred) - 33 # Convert ASCII to integer
                    if qualn >= 30:
                        nt[base] += 1
                else:
                    nt[base] += 1
            if nt[ref] > 0:
                af = nt[alt] / ( nt[alt] + nt[ref] ) * 100
            else:
                af = 'Inf'
            nalt = nt[alt]
            nref = nt[ref]
            break
    if flag is not 1:
        af = str( 'NA' )
        nalt = str( 'NA' )
-UU-:----F1  inspect_variants.py    Top L1     (Python)----------------------------------------------------------------
For information about GNU Emacs and the GNU system, type C-h C-a.
```

Wouldn’t you rather do this?

What is Galaxy?

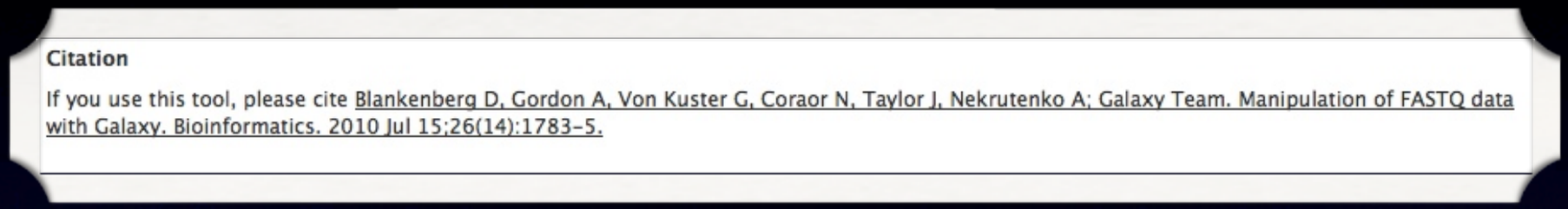
Galaxy is a collection of bioinformatics tools for:

- data conversion and manipulation
- statistical analysis
- next generation sequencing analysis
- much, much more!...



matt.shirley@jhmi.edu

What is Galaxy?

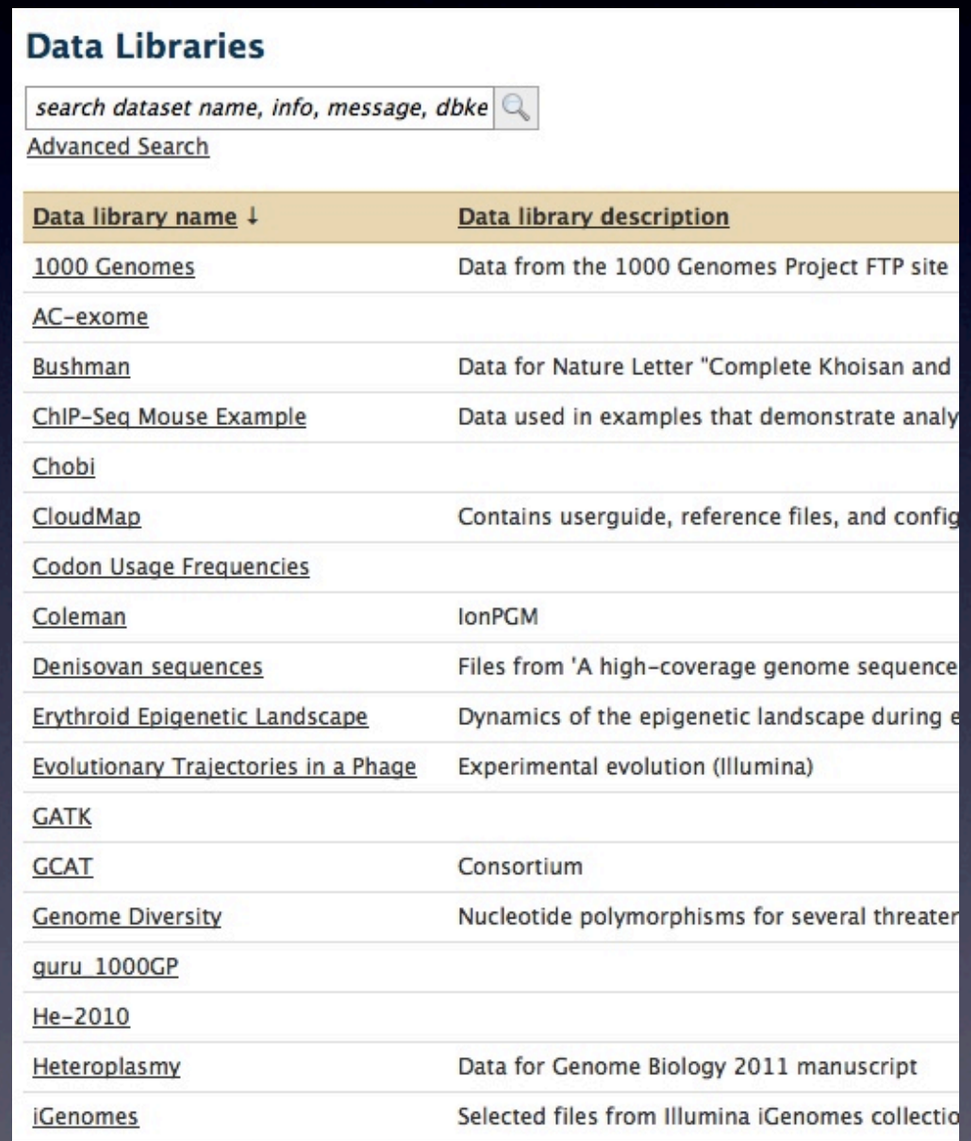


- Based on peer-reviewed and open-source implementations of each tool
- Galaxy provides integration with useful tools, targeted toward "bench" scientists
- Unified and consistent interface for easy exploration

matt.shirley@jhmi.edu

What is Galaxy?

- Data library: management and sharing for collaborative analysis
- Sequencer interface for direct intake of raw data



Data Libraries

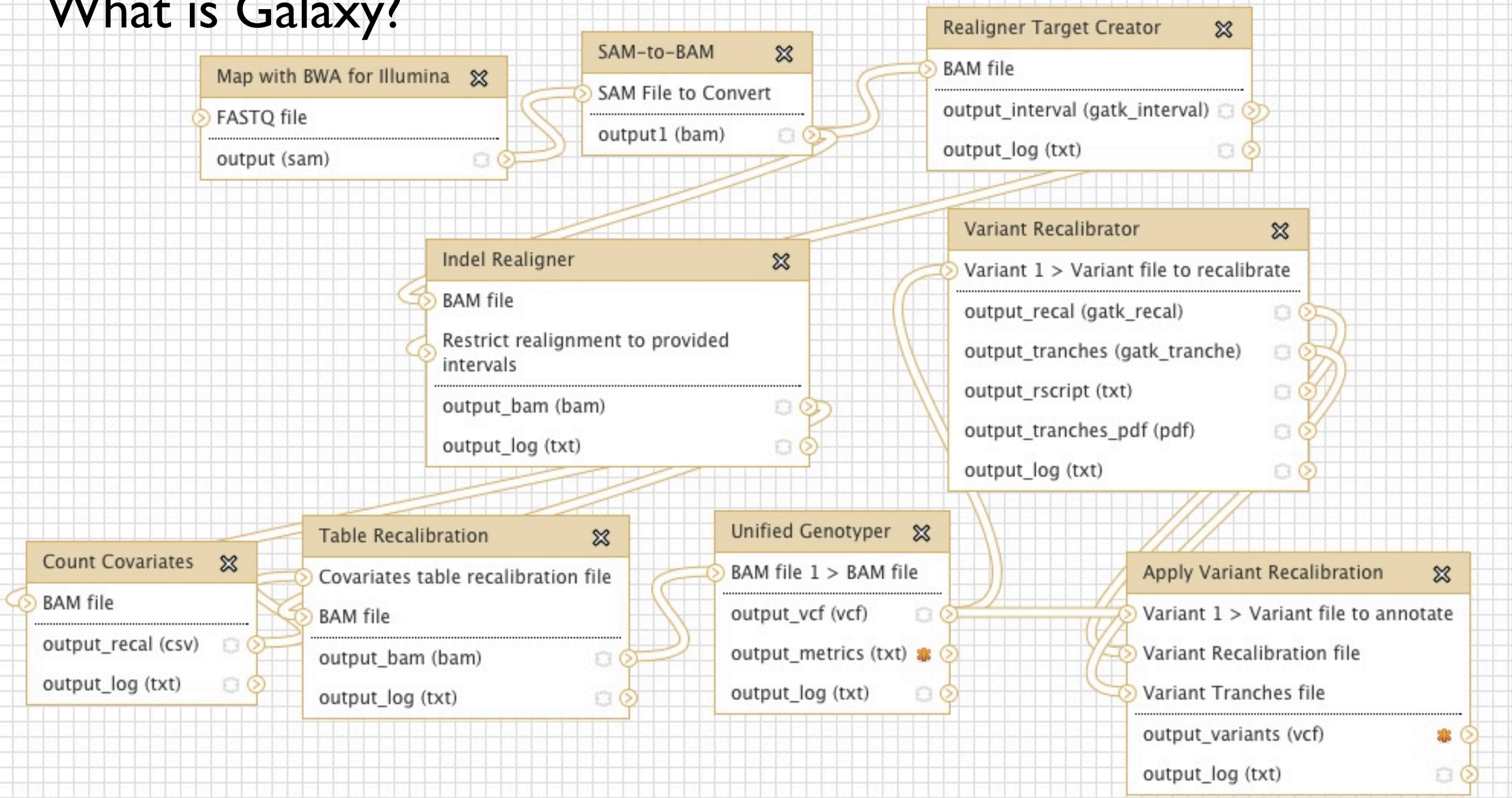
search dataset name, info, message, dbke

Advanced Search

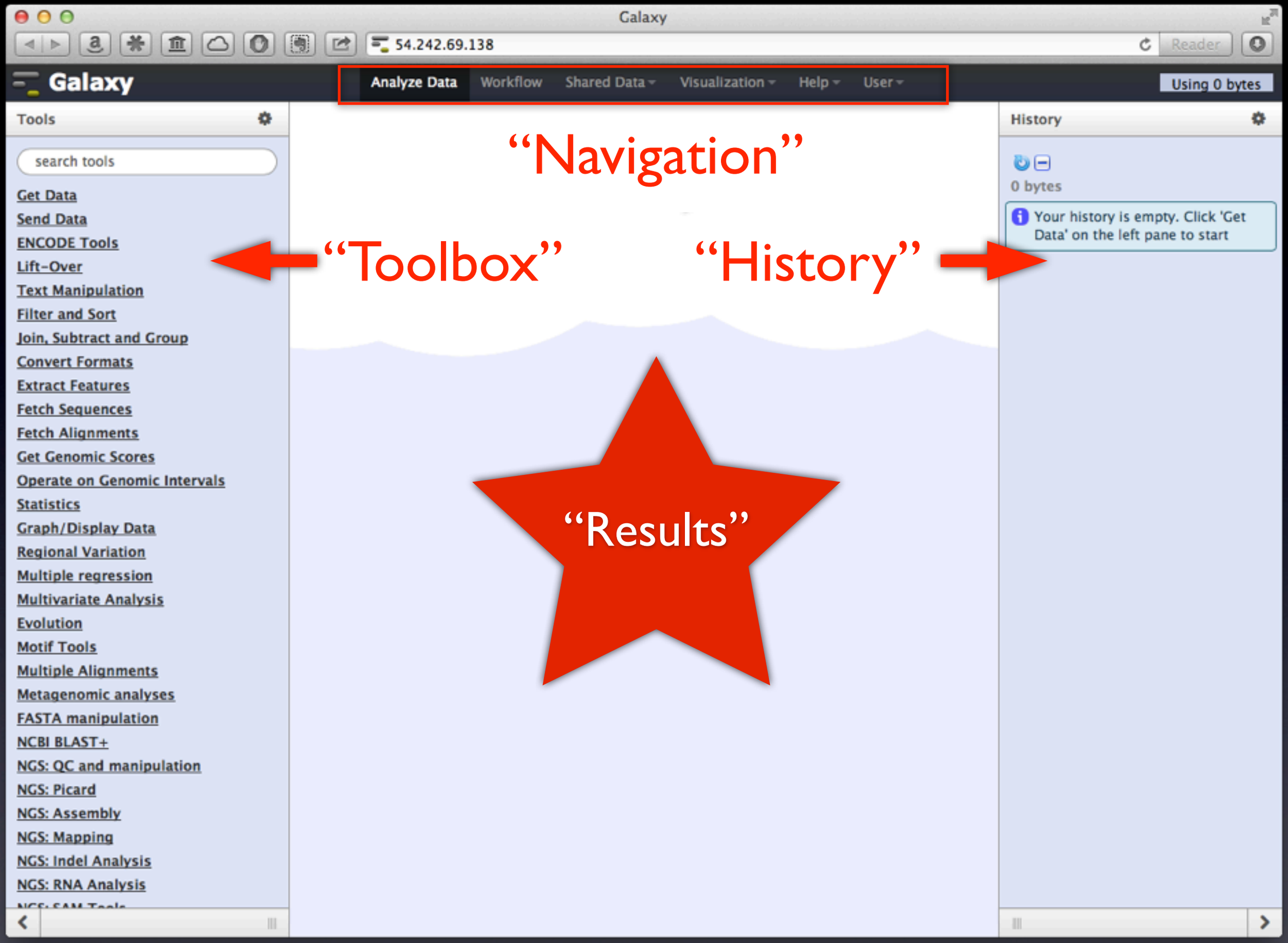
Data library name ↓	Data library description
1000 Genomes	Data from the 1000 Genomes Project FTP site
AC-exome	
Bushman	Data for Nature Letter "Complete Khoisan and
ChIP-Seq Mouse Example	Data used in examples that demonstrate analy
Chobi	
CloudMap	Contains userguide, reference files, and config
Codon Usage Frequencies	
Coleman	IonPGM
Denisovan sequences	Files from 'A high-coverage genome sequence
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during e
Evolutionary Trajectories in a Phage	Experimental evolution (Illumina)
GATK	
GCAT	Consortium
Genome Diversity	Nucleotide polymorphisms for several threater
guru_1000GP	
He-2010	
Heteroplasmy	Data for Genome Biology 2011 manuscript
iGenomes	Selected files from Illumina iGenomes collectio

matt.shirley@jhmi.edu

What is Galaxy?



Workflows that enable
reproducible research



matt.shirley@jhmi.edu

The "toolbox"

Contains links for :

- retrieving ("get") data
- manipulating data (lift-over, filter, sort, set operations, format conversions)
- data analysis (statistics, sequence alignment, variant calling and annotation)

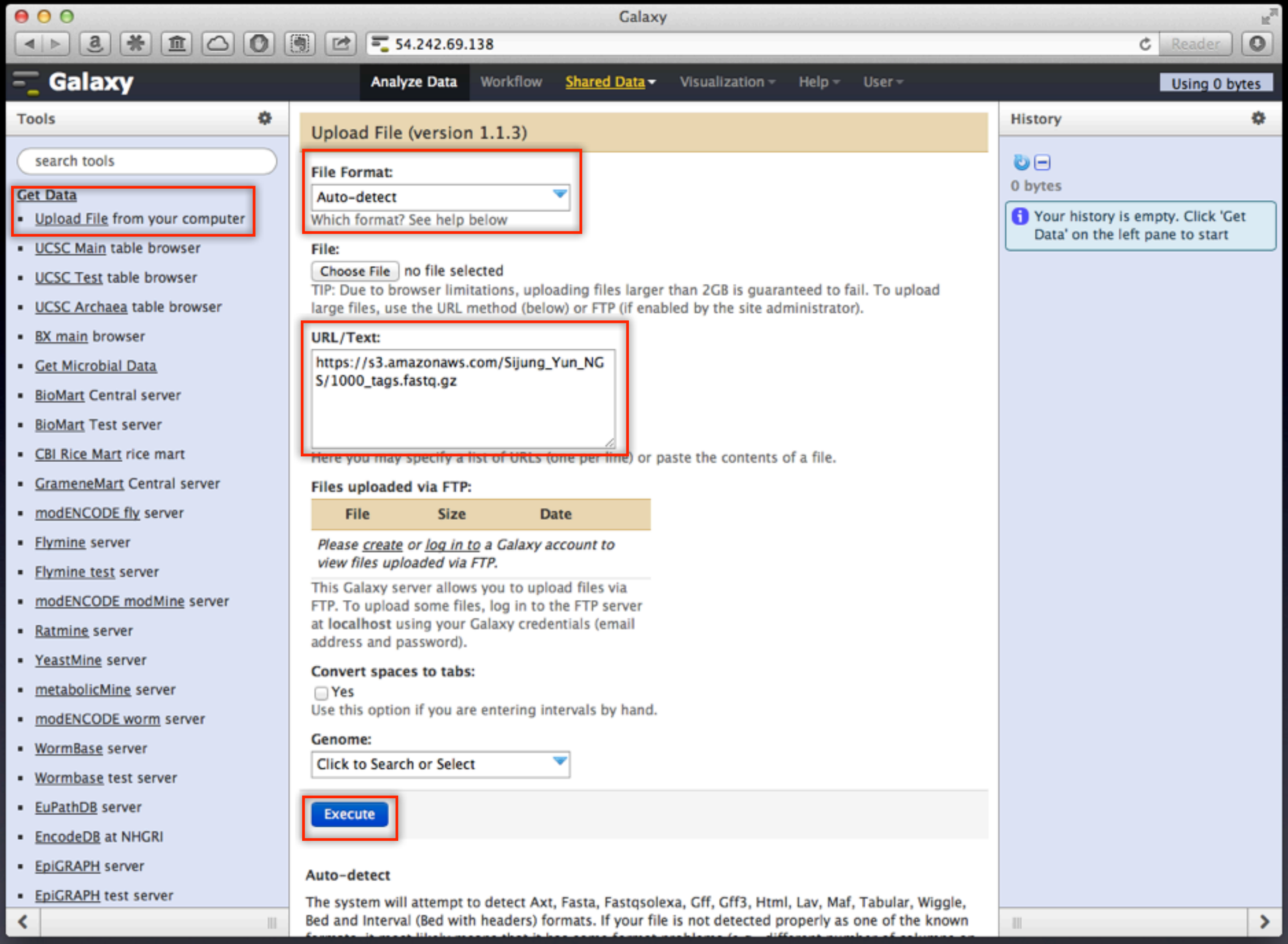
matt.shirley@jhmi.edu

“Get” data

In addition to uploading files from your computer, you may:

- Choose a file in the “shared data” library
- Import from UCSC, EBI SRA, BioMart, CBI Rice Map, modENCODE, Ratmine, Flymine, YeastMine, WormBase, EuPath, Microbial Genome Project, EncodeDB, EpiGRAPH, HbVar, GenomeSpace

matt.shirley@jhmi.edu



matt.shirley@jhmi.edu

Galaxy

54.242.69.138

Reader

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 2.1 Mb

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- Wormbase test server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- EpiGRAPH test server

This dataset is large and only the first megabyte is shown below.
Show all | Save

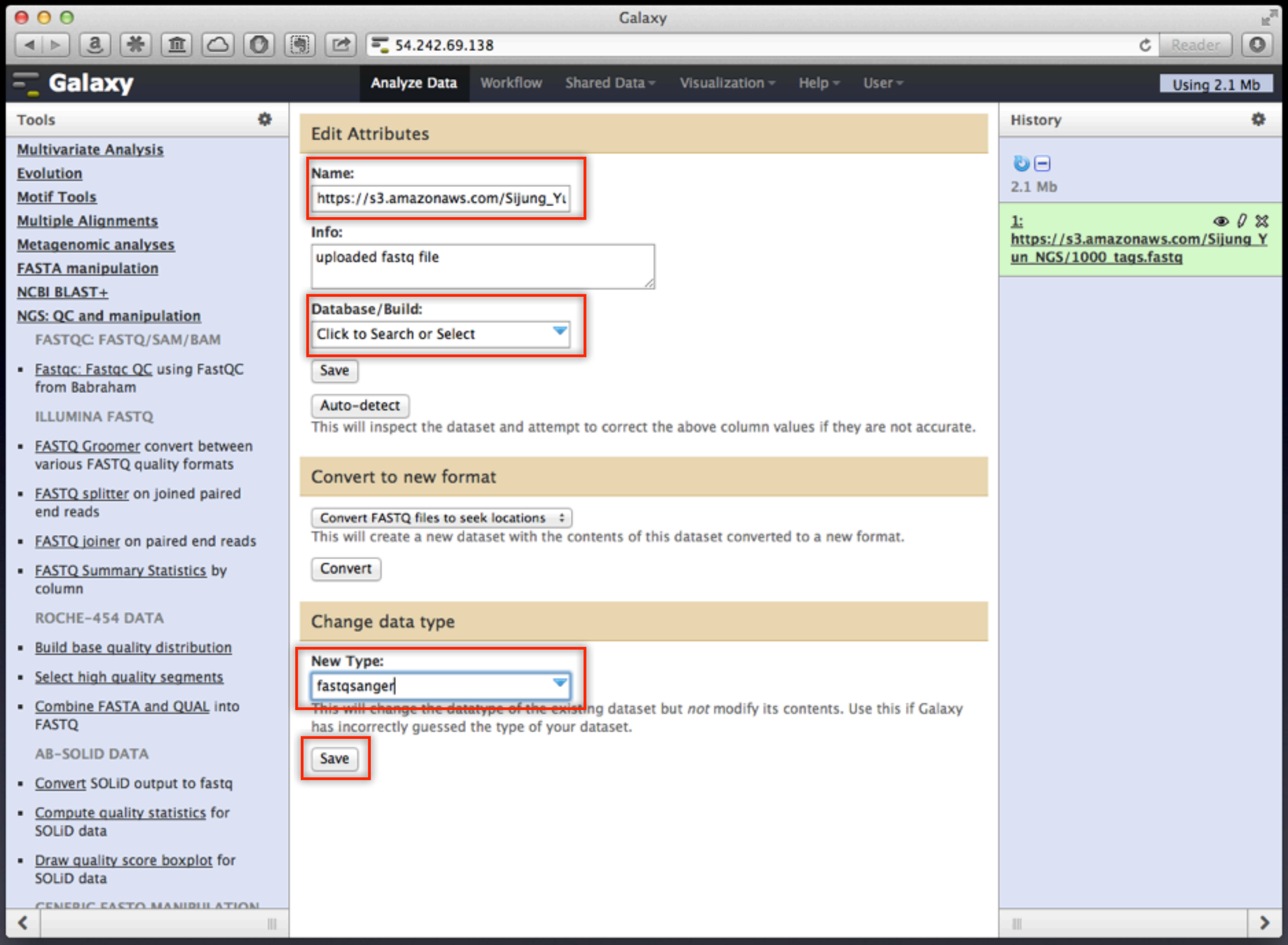
```
@HWI-EAS179_0001:5:1:7:119#0/1
CAGGGCGCGAATGNTTTGAGAGGGANATTGGAAANNNNNGATAGANNGGNCTATNNTGNNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:119#0/1
HIHHHGHHHFDHH#EHHH?HHHDH>#DGGG@7@?##########################################
@HWI-EAS179_0001:5:1:7:191#0/1
GAATCGATGAAGTNCGGCTTTTTGTNGATGATGTNNNNNATGCGGNNGCNCTGTNNGTNNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:191#0/1
G=EFHHDHF=9D<#D9:A@;HH=H0#B6?B##################################################
@HWI-EAS179_0001:5:1:7:135#0/1
TGGACCTCGCATTNACGTTGACAGTNCACTAAAGNNNNNGCAACTNNATNTCCANNGAANNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:135#0/1
HHHFHHHHHHHHH#HDFDDDEGEGG#HFHH###################################################
@HWI-EAS179_0001:5:1:7:1421#0/1
ATTGCTCAATGTAAAAGTTCAGCGGNGCCCCTCTCNNNNTCTCTTNNTCNTCACNNGAANNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1421#0/1
F@G@B;8EG92DA:?G5DG0=D9>D#5DD@CC)8C####CCC>CC##DD##############################
@HWI-EAS179_0001:5:1:7:1443#0/1
GTGTAGTCGAATTTTTTTAGTGCTTTGTTAGACTCNNNCTTGTCTGAAANCACCNAATTTNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1443#0/1
DBD<DHFHH@DHHHHHHG4H+BGGCHHHH=DBFDF###+A:?0A7@7::#=?##########################
@HWI-EAS179_0001:5:1:7:1350#0/1
TGAGAATTTTGTATGTAAGTTTTTTCTTTTGAAGTNNTTTGTTCAGTGAAACAANAATTGNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1350#0/1
HHDH@HHHHGHHHHHFF>FFEFFFFFFFFFEE@E>##A?AFA:GC15C############################
@HWI-EAS179_0001:5:1:7:72#0/1
CAAAGACCTCAAAGTCAAACTCGACTAGTTCGGAANNGGTCTTATTCTGTCAATNTCATANNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:72#0/1
HHHEEDGHHHCAGE4HGHHHGHAGE8:HEDFBF3<##A7.58=,1BECD/F0.B#######################
@HWI-EAS179_0001:5:1:7:746#0/1
TGCTAGCTTATATCATGGAGTTGGGCAGTGCATACNNTTCTTCGTGTATTTATTNATTTCNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:746#0/1
HHGHHHGHHCDAD;5DGGADGDGGGFAFFFC>CC9##=<2CCCC<D>7D<<5@@#####################
@HWI-EAS179_0001:5:1:7:1657#0/1
TGAAAGACGCAGACAATTTAATGCTAGAAATTTTCNNTAATTGGTCCTTGATCANTGTCGNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1657#0/1
HHDEHHFHHHHHHHH5GGG<DFFFF=D;>ACACCC##6:>@@C:FFCFF9@@?#@>;A@################
@HWI-EAS179_0001:5:1:7:1002#0/1
CCATCAGATCAAACAATTCAGCATGACTCTTTGGANNTCCATATCCAGTCCCTTNCGAATNNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1002#0/1
HGCHHGHHBHFA>G;ADDFDGGGEE.G8GD=9:C5##<=><9?<CFFF?FC7?#####################
@HWI-EAS179_0001:5:1:7:14#0/1
```

History

2.1 Mb

1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

matt.shirley@jhmi.edu



matt.shirley@jhmi.edu

Galaxy

54.242.69.138

Reader

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 2.1 Mb

Tools

- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

FASTQ Summary Statistics (version 1.0.0)

FASTQ File:

1: https://s3.amazon..._tags.fastq

Execute

This tool creates summary statistics on a FASTQ file.

TIP: This statistics report can be used as input for the **Boxplot** and **Nucleotides Distribution** tools.

The output file will contain the following fields:

column = column number (1 to 36 for a 36-cycles read Solexa file)
count = number of bases found in this column.
min = Lowest quality score value found in this column.
max = Highest quality score value found in this column.
sum = Sum of quality score values for this column.
mean = Mean quality score value for this column.
Q1 = 1st quartile quality score.
med = Median quality score.
Q3 = 3rd quartile quality score.
IQR = Inter-Quartile range (Q3-Q1).
lW = 'Left-Whisker' value (for boxplotting).
rW = 'Right-Whisker' value (for boxplotting).
outliers = Scores falling beyond the left and right whiskers (comma separated list).
A_Count = Count of 'A' nucleotides found in this column.
C_Count = Count of 'C' nucleotides found in this column.
G_Count = Count of 'G' nucleotides found in this column.
T_Count = Count of 'T' nucleotides found in this column.
N_Count = Count of 'N' nucleotides found in this column.
Other_Nucs = Comma separated list of other nucleotides found in this column.
Other_Count = Comma separated count of other nucleotides found in this column.
For example:

#column	count	min	max	sum	mean	Q1	med	Q3	IQR	lW	rW	outliers	A_Count	C_Cou
1	14336356	2	33	450600675	31.4306281875	32.0	33.0	33.0	1.0	31	33			
2	14336356	2	34	441135033	30.7703737965	30.0	33.0	33.0	3.0	26	34			
3	14336356	2	34	433659182	30.2489127642	29.0	32.0	33.0	4.0	23	34			
4	14336356	2	34	433635331	30.2472490917	29.0	32.0	33.0	4.0	23	34			
5	14336356	2	34	432498583	30.167957813	29.0	32.0	33.0	4.0	23	34			

History

Unnamed history — 2.1 Mb

1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

matt.shirley@jhmi.edu

Galaxy

54.242.69.138

Reader

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 2.1 Mb

Tools

- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

The following job has been successfully added to the queue:

2: FASTQ Summary Statistics on data 1

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

Unnamed history 2.1 Mb

2: FASTQ Summary Statistics on data 1

1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

matt.shirley@jhmi.edu

Galaxy

54.242.69.138

Reader

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 2.1 Mb

Tools

- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

#column	count	min	max	sum	mean	Q1	med	Q3	IQR	lW	rW
1	10000	2	40	358994	35.8994	35.0	38.0	39.0	4.0	29	40
2	10000	2	40	356075	35.6075	35.0	38.0	39.0	4.0	29	40
3	10000	2	40	355450	35.545	35.0	38.0	39.0	4.0	29	40
4	10000	2	40	354268	35.4268	35.0	38.0	39.0	4.0	29	40
5	10000	2	40	355691	35.5691	35.0	38.0	39.0	4.0	29	40
6	10000	2	40	357066	35.7066	35.0	38.0	39.0	4.0	29	40
7	10000	2	40	357335	35.7335	35.0	38.0	39.0	4.0	29	40
8	10000	2	40	356799	35.6799	35.0	38.0	39.0	4.0	29	40
9	10000	2	40	355891	35.5891	35.0	38.0	39.0	4.0	29	40
10	10000	2	40	356335	35.6335	35.0	38.0	39.0	4.0	29	40
11	10000	2	40	357109	35.7109	35.0	38.0	39.0	4.0	29	40
12	10000	2	40	357572	35.7572	35.0	38.0	39.0	4.0	29	40
13	10000	2	40	358956	35.8956	35.0	38.0	39.0	4.0	29	40
14	10000	2	40	360000	36.0	35.0	38.0	39.0	4.0	29	40
15	10000	2	40	356901	35.6901	35.0	38.0	39.0	4.0	29	40
16	10000	2	40	350223	35.0223	34.0	38.0	39.0	5.0	27	40
17	10000	2	40	351943	35.1943	34.0	38.0	39.0	5.0	27	40
18	10000	2	40	351757	35.1757	34.0	38.0	39.0	5.0	27	40
19	10000	2	40	351216	35.1216	34.0	38.0	39.0	5.0	27	40
20	10000	2	40	351567	35.1567	34.0	38.0	39.0	5.0	27	40
21	10000	2	40	349414	34.9414	34.0	38.0	39.0	5.0	27	40
22	10000	2	40	350206	35.0206	34.0	38.0	39.0	5.0	27	40
23	10000	2	40	350242	35.0242	34.0	38.0	39.0	5.0	27	40
24	10000	2	40	351490	35.149	34.0	38.0	39.0	5.0	27	40
25	10000	2	40	349360	34.936	34.0	37.0	39.0	5.0	27	40
26	10000	2	40	353087	35.3087	35.0	38.0	39.0	4.0	29	40
27	10000	2	40	351412	35.1412	35.0	38.0	39.0	4.0	29	40
28	10000	2	40	352640	35.264	35.0	38.0	39.0	4.0	29	40
29	10000	2	40	352542	35.2542	35.0	38.0	39.0	4.0	29	40
30	10000	2	40	349193	34.9193	34.0	38.0	39.0	5.0	27	40
31	10000	2	40	354332	35.4332	35.0	38.0	39.0	4.0	29	40
32	10000	2	40	347471	34.7471	34.0	38.0	39.0	5.0	27	40
33	10000	2	40	354691	35.4691	35.0	38.0	39.0	4.0	29	40
34	10000	2	40	352285	35.2285	35.0	38.0	39.0	4.0	29	40
35	10000	2	40	348150	34.815	34.0	38.0	39.0	5.0	27	40

History

Unnamed history — 2.1 Mb

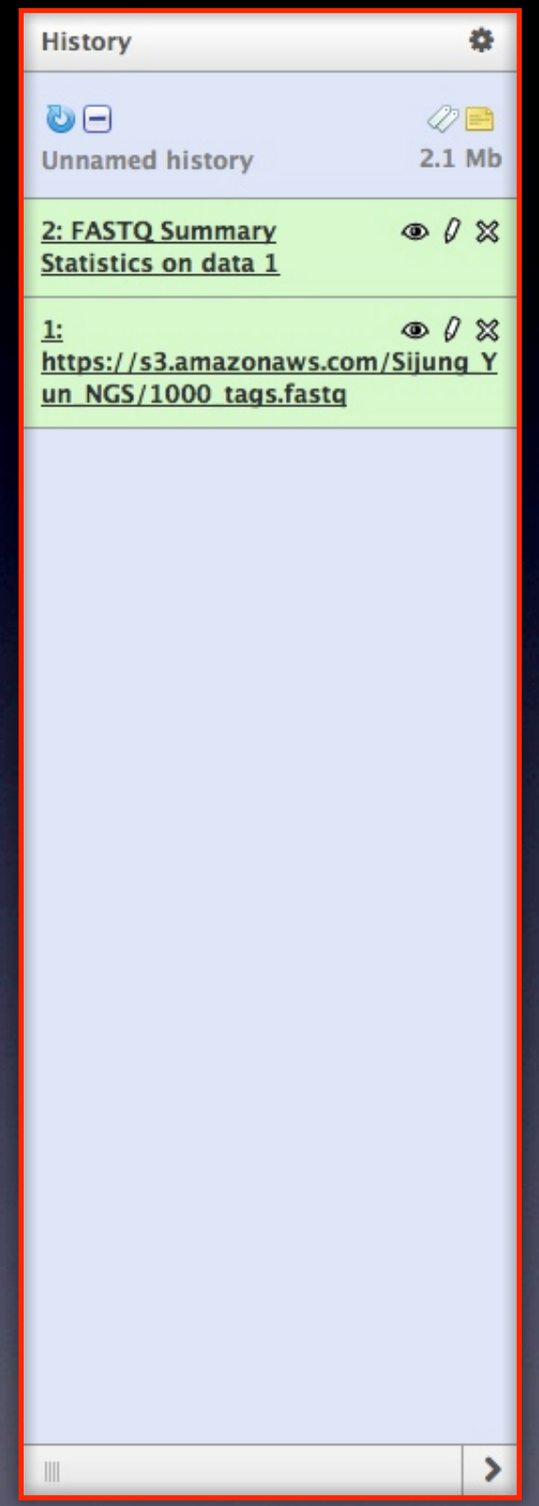
2: FASTQ Summary Statistics on data 1

1: https://s3.amazonaws.com/Sijung_Yun_NGS/1000_tags.fastq

matt.shirley@jhmi.edu

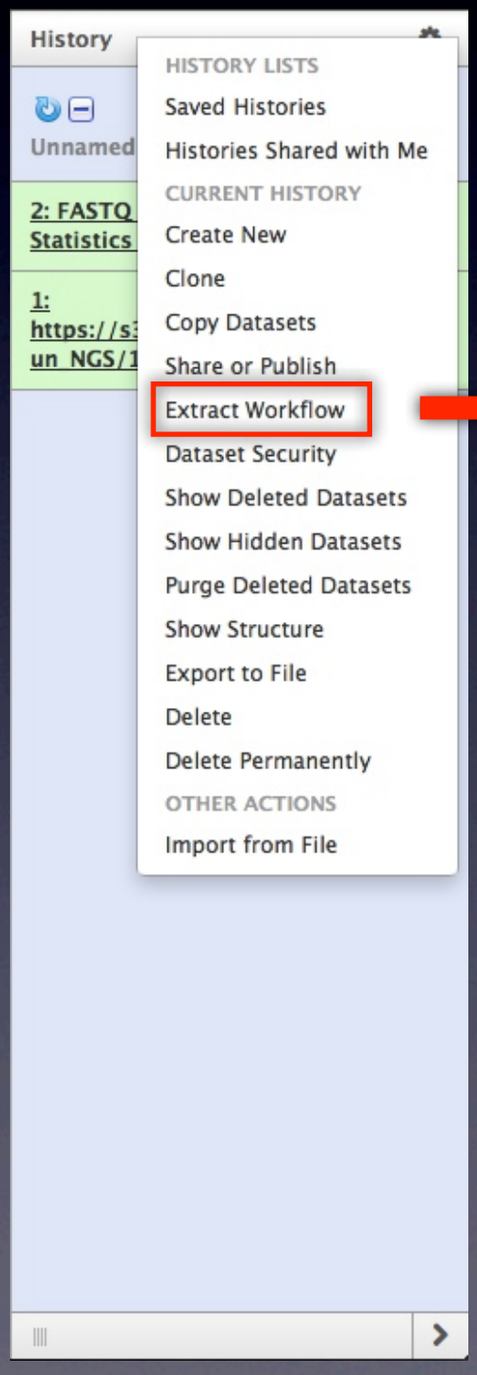
The “history”

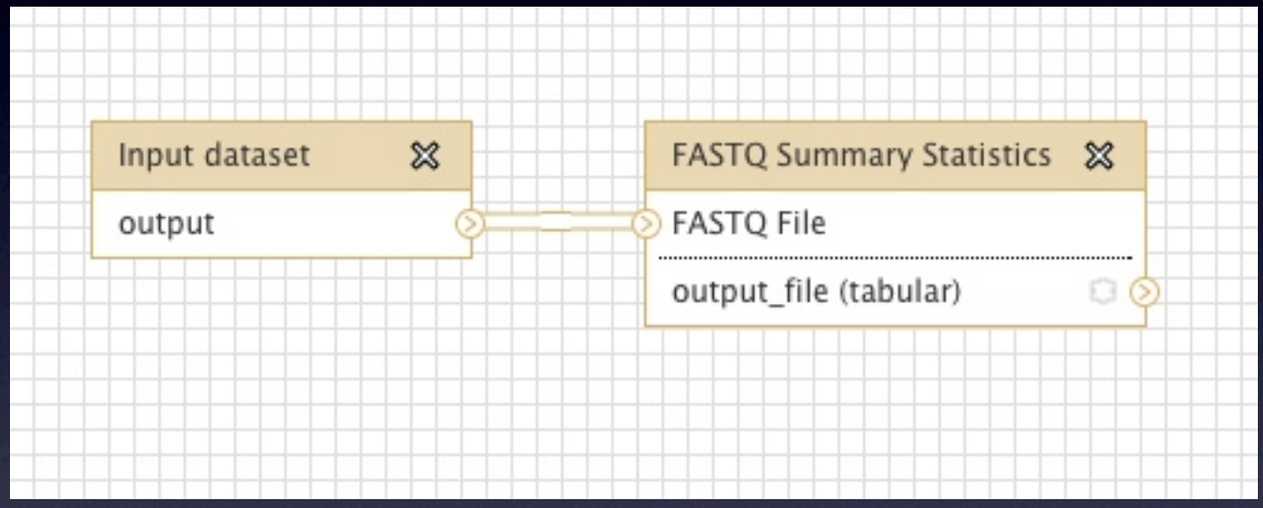
- Displays a list of your analysis steps
- Allows interaction with analysis results
- Each item in the history is a “data-set”
- Multiple concurrent histories allowed
- Maintains the order of analysis steps, allowing extraction of workflows on-the-fly



matt.shirley@jhmi.edu

Extracting workflows from histories





Histories and workflows result in reproducible research

matt.shirley@jhmi.edu

NGS analysis in Galaxy

- QC and manipulation: filter, trim, mask, and convert fastq files
- Picard: a Java implementation of many samtools functions
- Mapping: align to reference genome with BWA, Bowtie, Bowtie2, BFAST, PerM, Mosaik, Lastz
- RNA: Tophat, Cufflinks (gapped alignment and transcript assembly)
- GATK: advanced analysis tools from BROAD
- Peak Calling: ChIP-Seq analysis tools

matt.shirley@jhmi.edu

Visualizations

Trackster linear genome browser supports most interval, continuous, and discreet data formats

Circster "circos" style connectivity browser with interactive zooming

Visual parametric optimization allows the user to pick the most optimum local parameters, then optionally apply these globally

Strengths and Weaknesses

Strengths:

- Each tool has similar user interface elements, leading to a much lower learning curve
- Histories and workflows allow reproducibility
- Cluster and cloud compute-compatible
- Extensible tool set via Python scripting

Weaknesses:

- Inefficient resource management (for now)
- Limited set of parameters for some tools

matt.shirley@jhmi.edu

Local vs. Public

- Public Galaxy server is accessible at http://usegalaxy.org
- Learn about installing local instances at http://getgalaxy.org
- NGS analysis involves *larg*e data, and long compute times.
- For NGS analysis, a local (or cloud) installation of Galaxy is recommended.

matt.shirley@jhmi.edu

Questions?

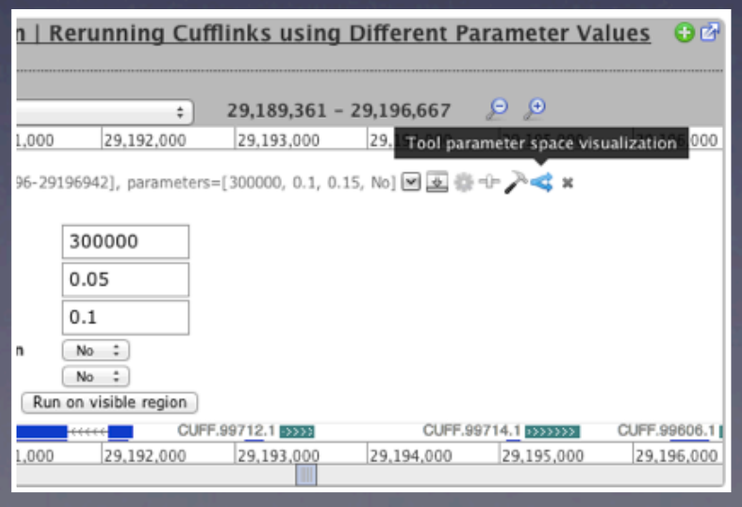
Slides available at http://mattshirley.com/presentations

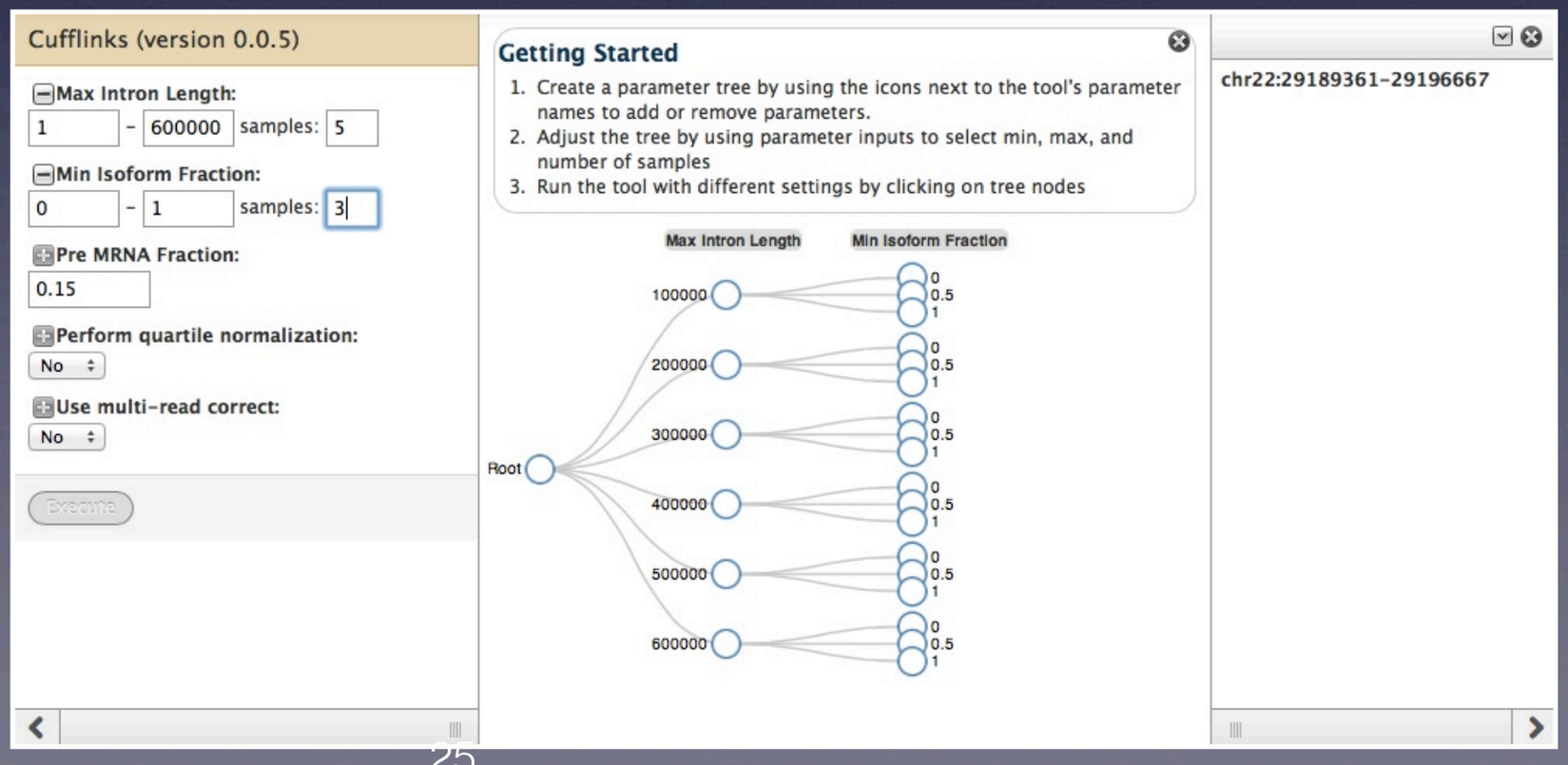
matt.shirley@jhmi.edu

Examples

- Basic protocols for Galaxy: Using Galaxy to Perform Large-Scale Interactive Data Analyses
- Parameter-space visualization: TopHat/CuffLinks

RNA-seq optimization





Galaxy on AWS ("the cloud")

http://xkcd.com/1117/

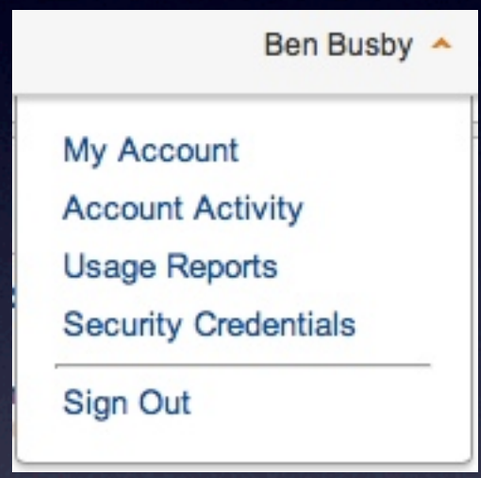
matt.shirley@jhmi.edu

New! Two options for cluster initialization

1. Use the new cloud launch tool from the main public instance.
2. Manually configure a cluster through Amazon Web Services management console.

Using the "cloud launch" tool at Galaxy Main

1. Log in to AWS EC2 management console http:/console.aws.amazon.com/ec2
   - Access you *Security Credentials page*
   - Save your *Access Key ID and Secret Access Key*



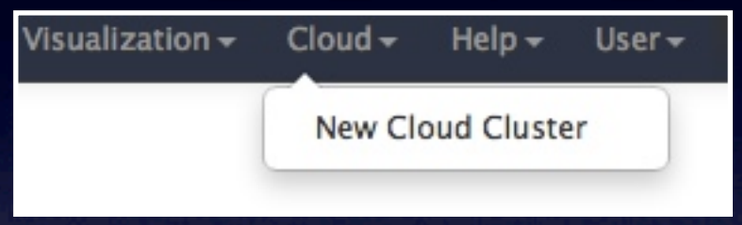
Your Access Keys

Created	Access Key ID	Secret Access Key	Status
October 3, 2012	AKIAJRBU4D36GXYROQWQ	Show	Active (Make Inactive)

matt.shirley@jhmi.edu

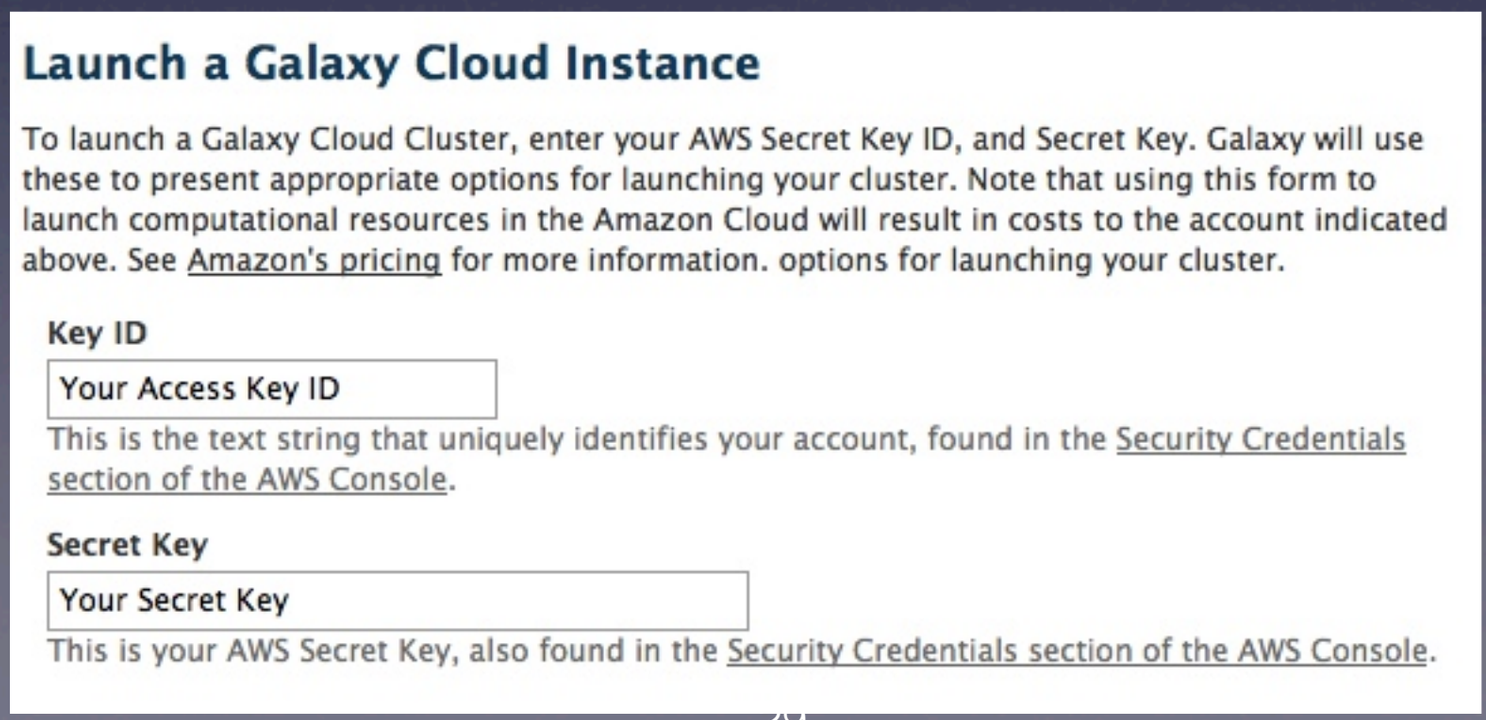
Automatic Galaxy cloud initialization

1. Click "New Cloud Cluster" from "Cloud" toolbar of the main public instance.



Alternative mirror (please use sparingly)

2. Enter your AWS access key ID and secret key



Final steps before initialization

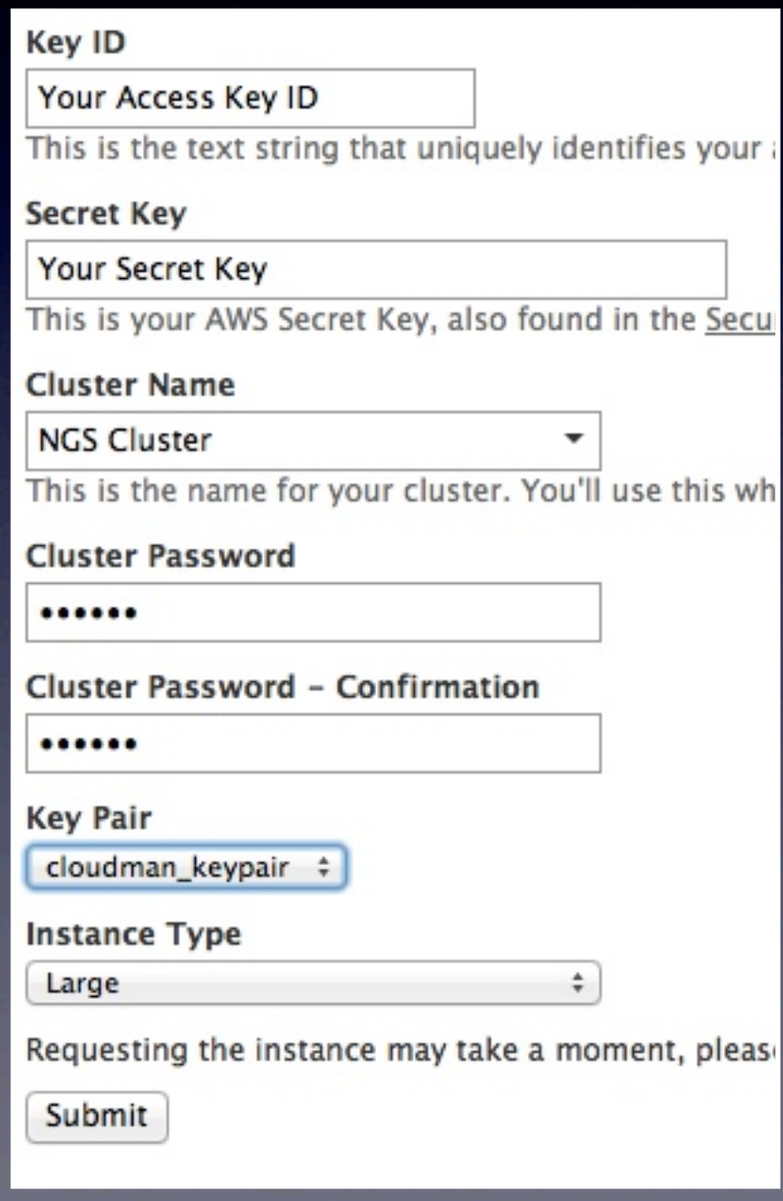
3. Enter a name for your cluster

4. Enter a password you can remember

5. Either choose an existing keypair or let the tool generate one for you

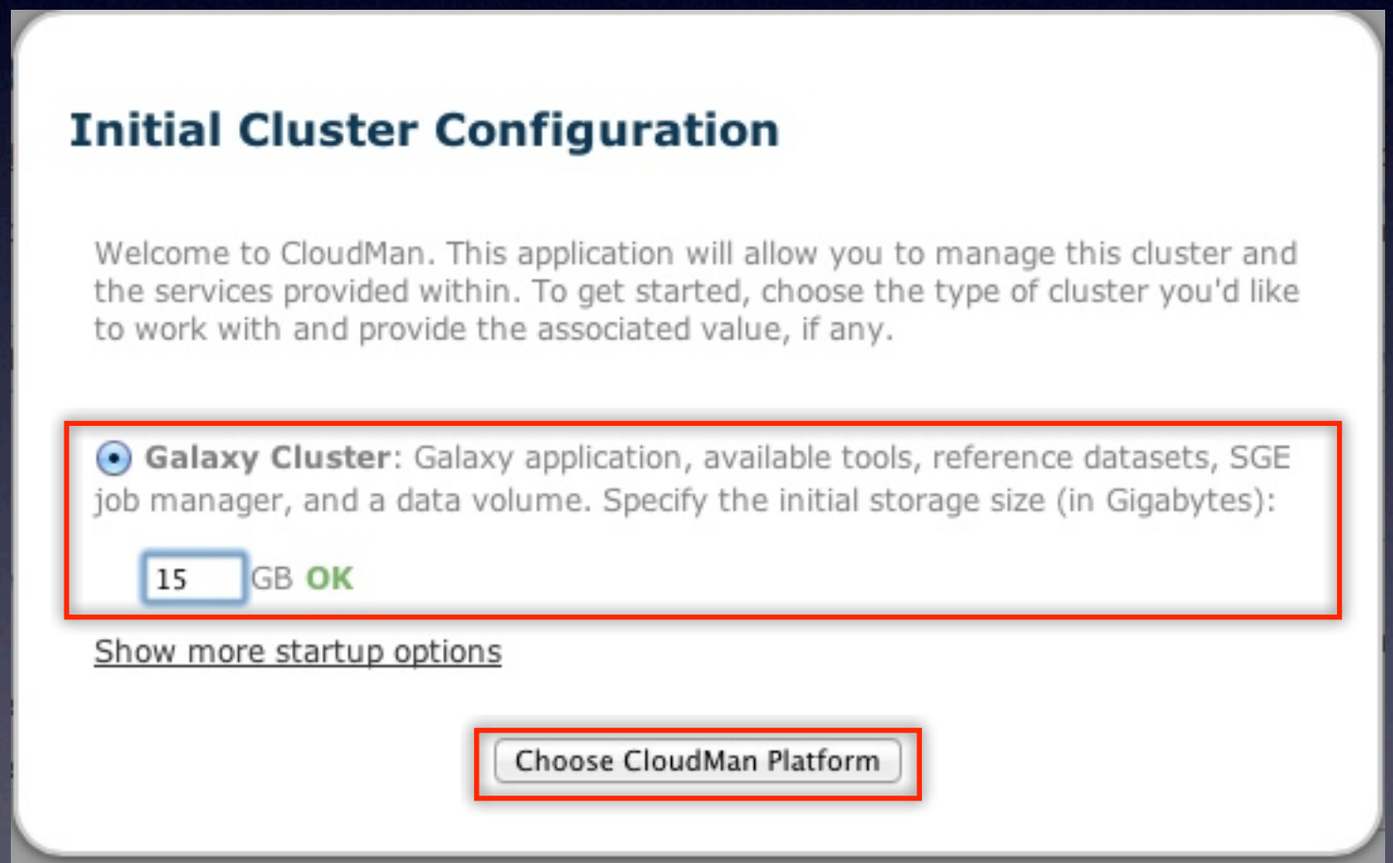
6. Select at least a “Large” instance type

7. Submit



Galaxy on AWS ("the cloud")

8. After logging in using the previously specified "cluster name" and "password", specify the initial storage for the Galaxy cluster

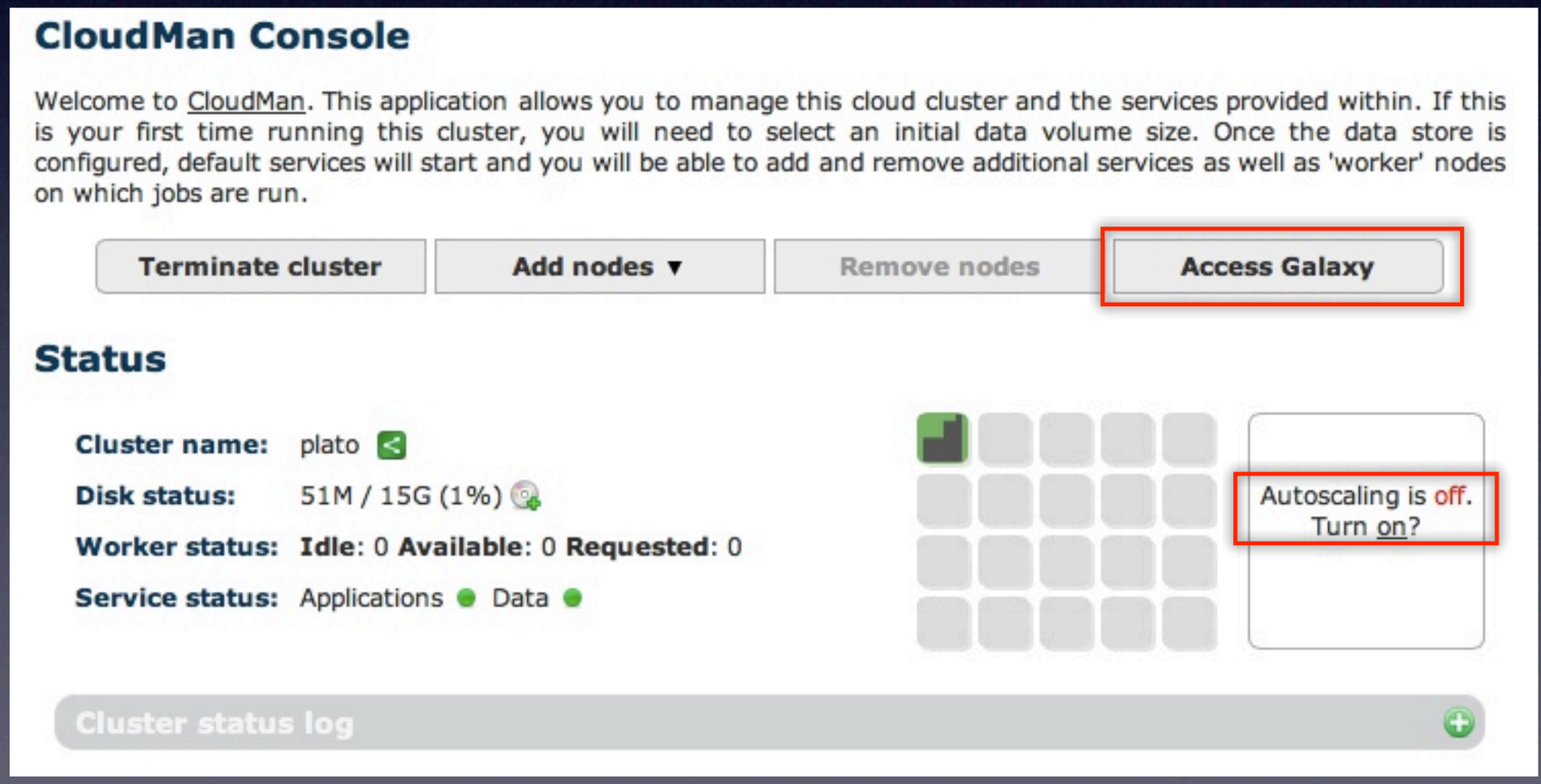


matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")

9. After a few minutes, the *Access Galaxy* button will become accessible, signaling success

- Note that performance will be improved if autoscaling is turned on



matt.shirley@jhmi.edu

You're ready to analyze some data!

Next:

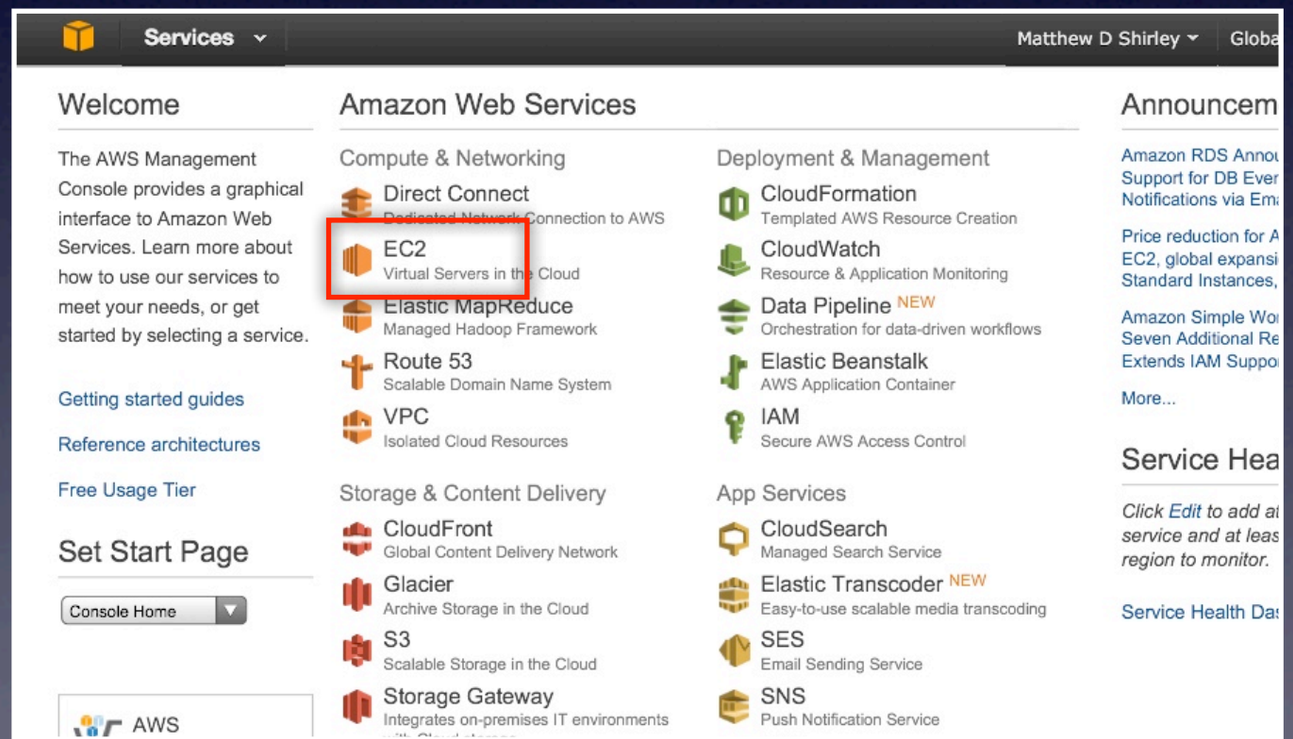
1. Learn how to shut down your cluster when you have finished.

2. Learn how to monitor your AWS usage.

3. Something didn't work? Try the hard way.

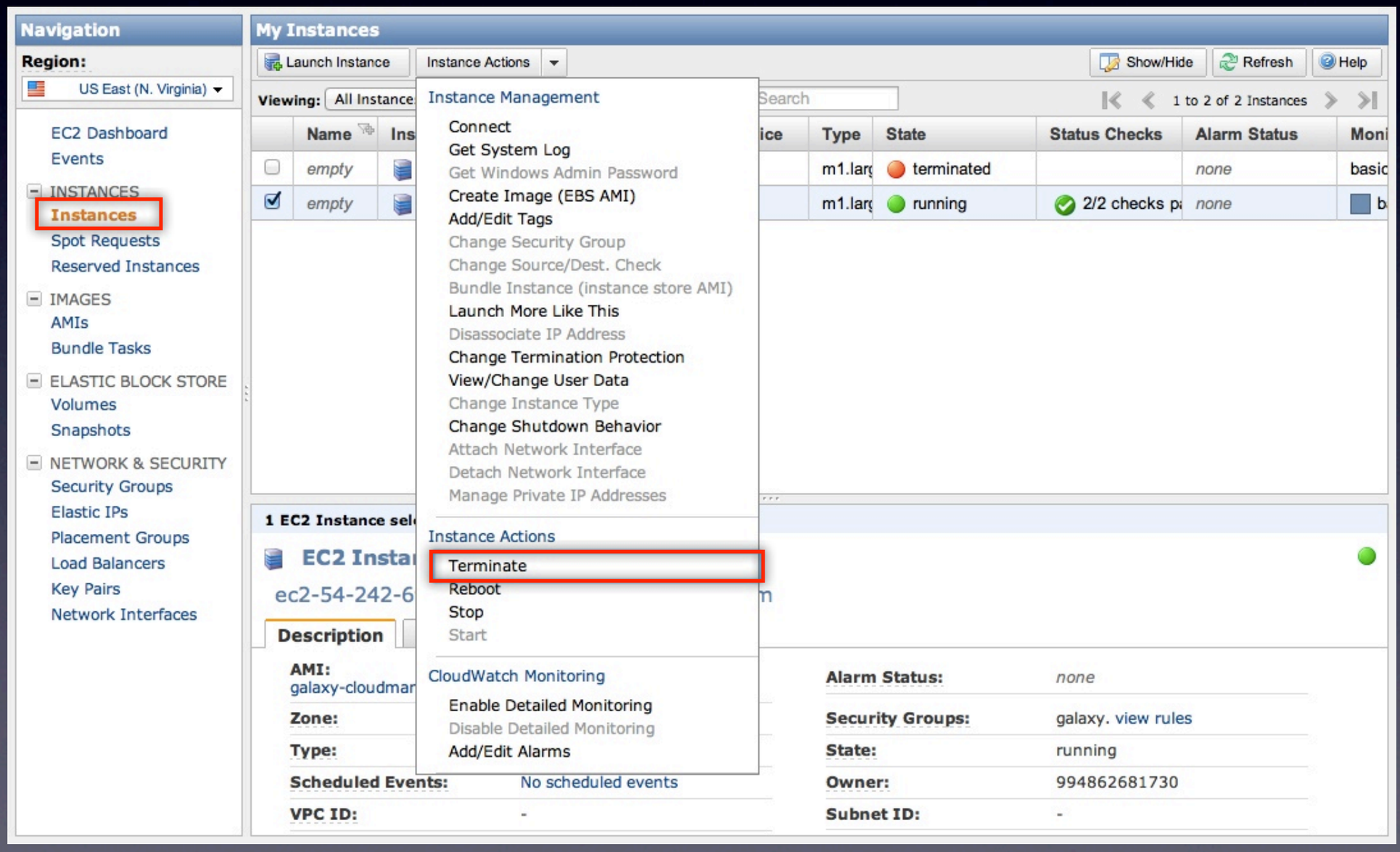
Shutting down your cluster

1. Log in to your AWS console
2. Select EC2



Shutting down your cluster

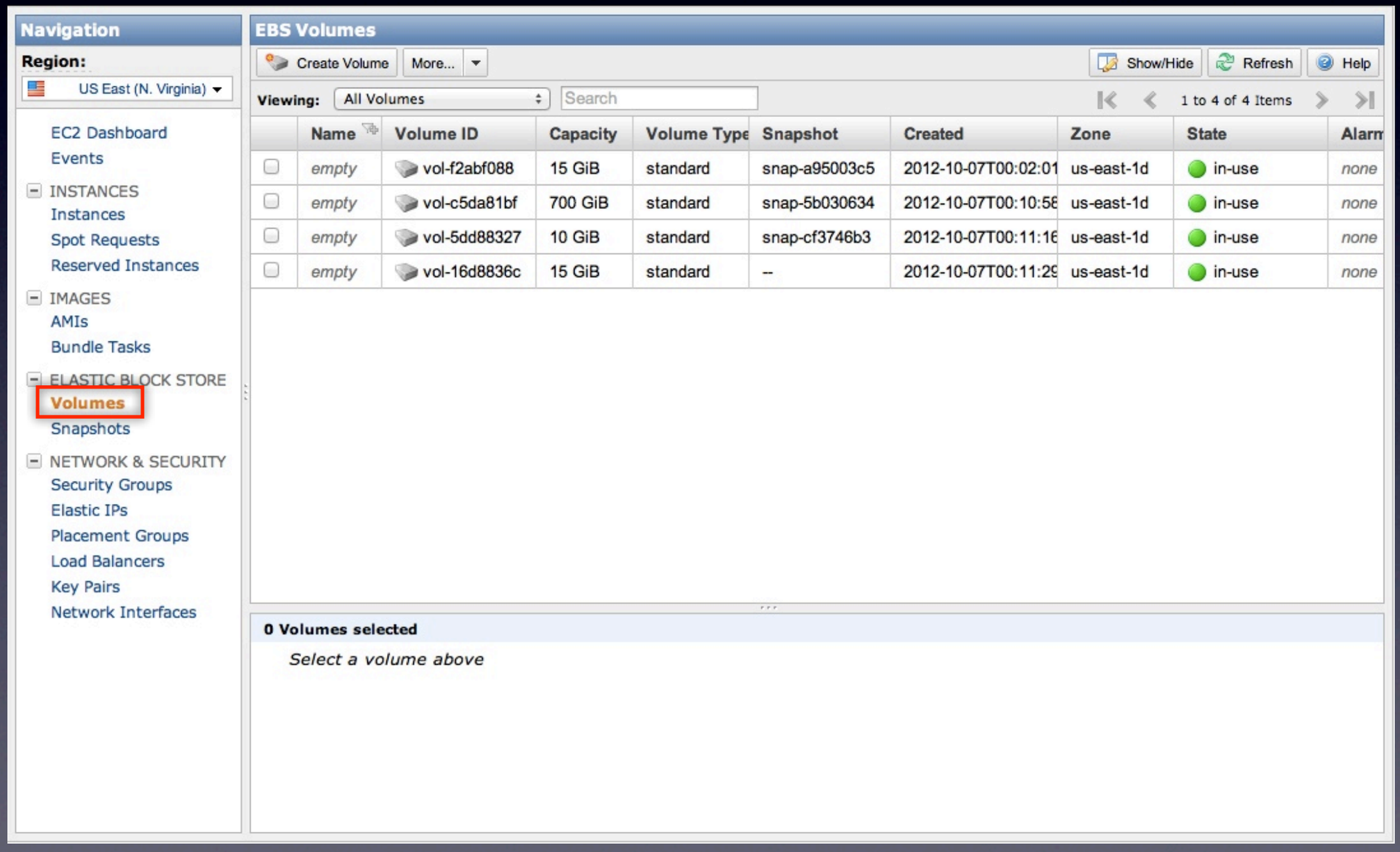
3. Select "instances" on the left and terminate any running EC2 instances



matt.shirley@jhmi.edu

Shutting down your cluster

4. Also remember to delete any EBS volumes that persist



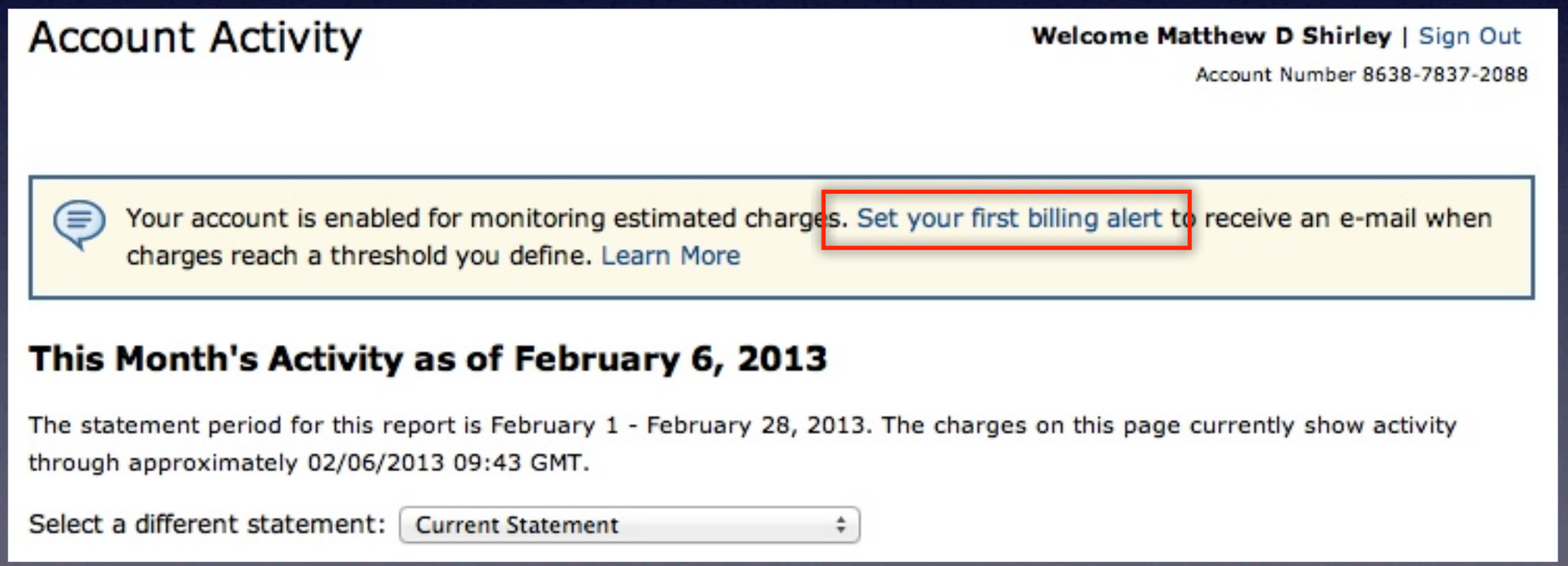
matt.shirley@jhmi.edu

Monitoring your usage!

1. Go to aws.amazon.com and select "Account Activity"

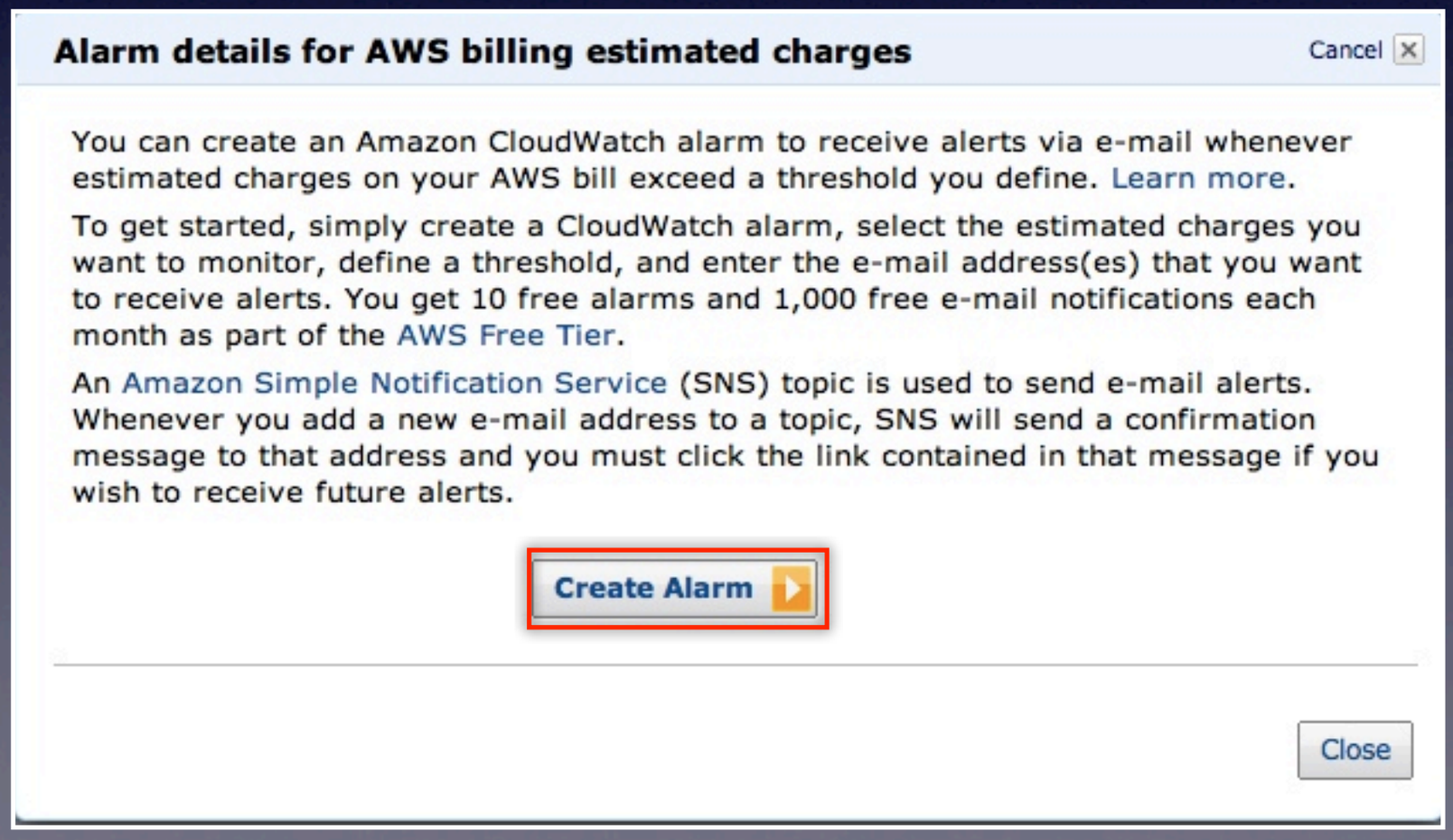
Monitoring your usage!

2. On your account activity page, select "Set your first billing alert"



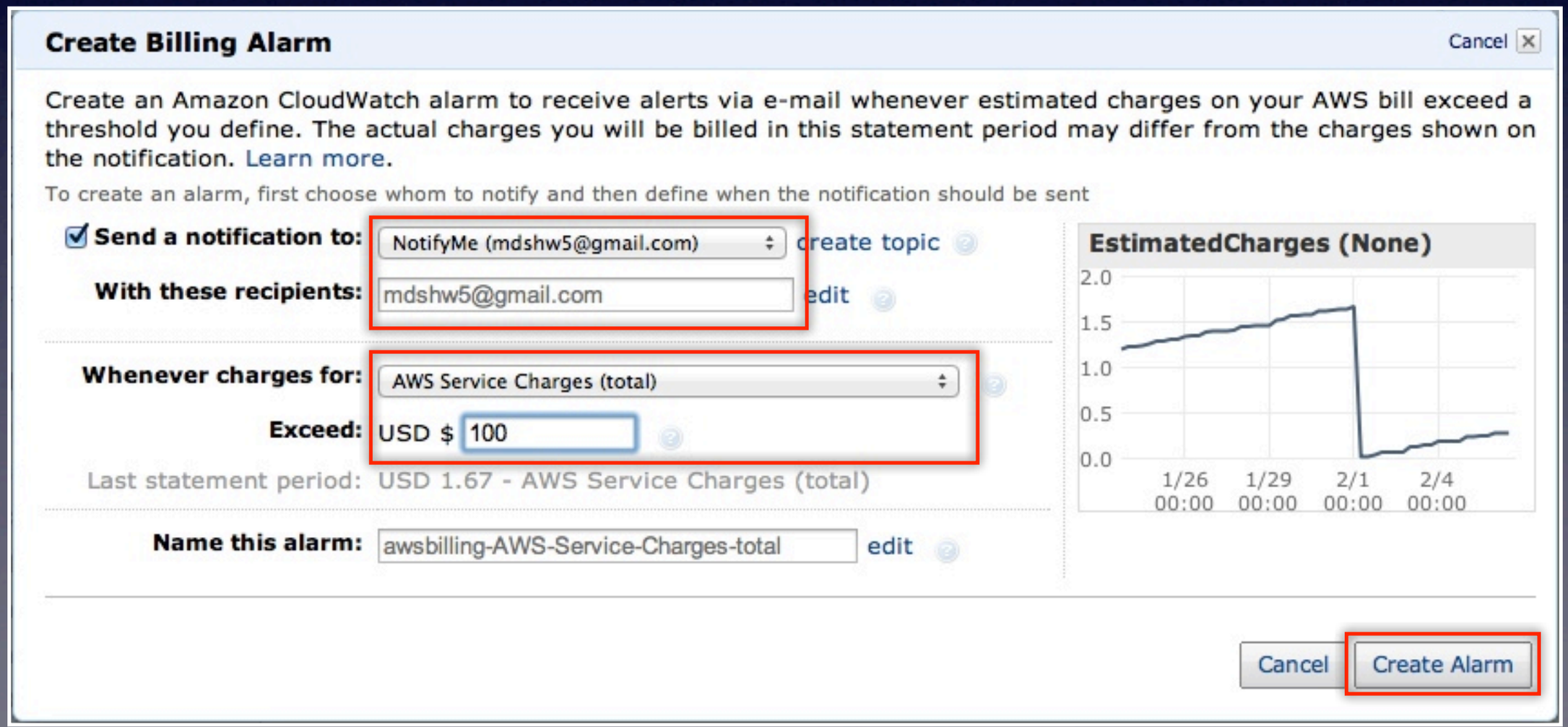
Monitoring your usage!

3. Select “Create Alarm”



Monitoring your usage!

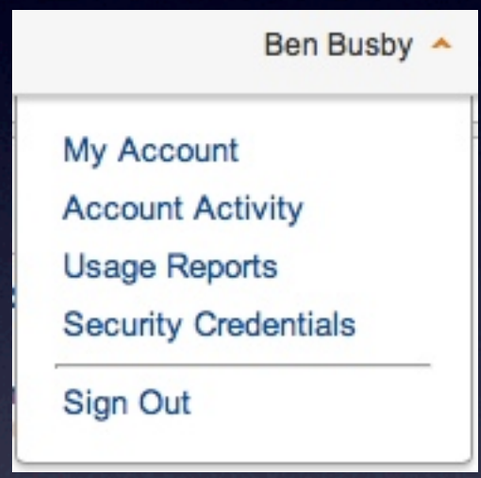
4. Select an email address to send notifications to, and enter a threshold of total AWS service charges above which you wish to be notified.



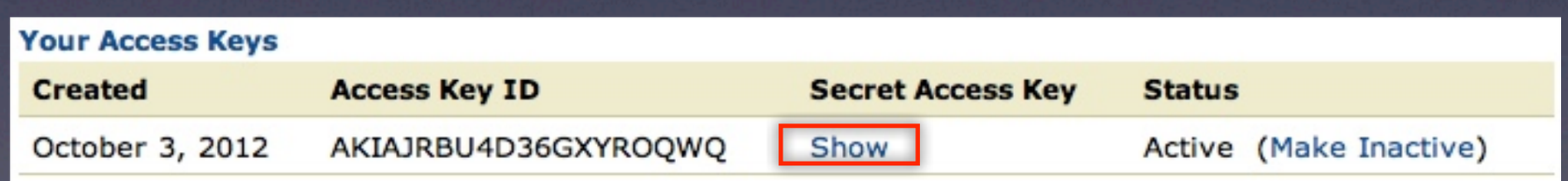
Manually configure a cluster through AWS management console

Steps adapted from http://wiki.g2.bx.psu.edu/CloudMan

1. Log in to AWS EC2 management console http:/console.aws.amazon.com/ec2



   - Access you *Security Credentials page*
   - Save your *Access Key ID and Secret Access Key*



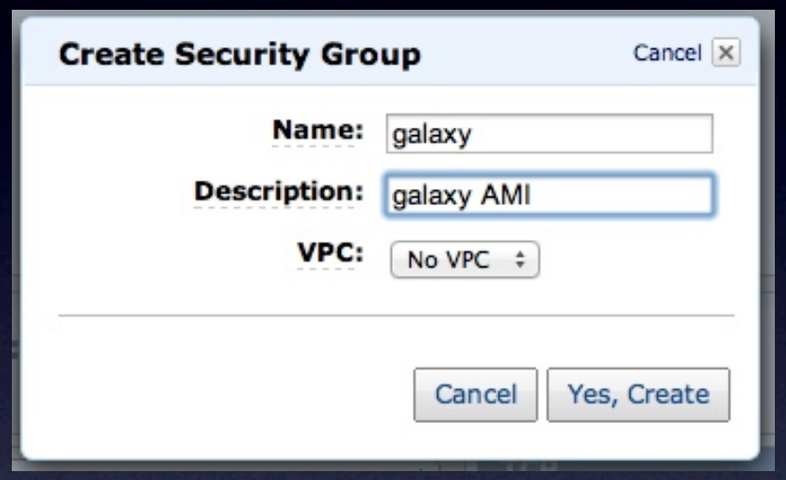
Your Access Keys

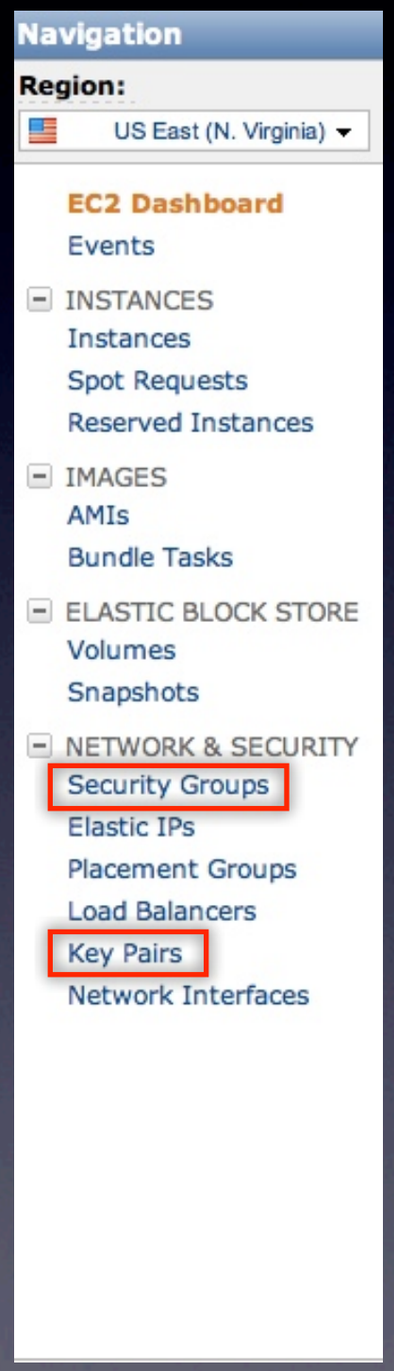
Created	Access Key ID	Secret Access Key	Status
October 3, 2012	AKIAJRBU4D36GXYROQWQ	Show	Active (Make Inactive)

matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")

2. Create a *Security Group* called "galaxy", description "galaxy AMI"





- Choose *Key Pairs*
- Create a key pair named "galaxy" and download it to your computer

matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")

3. Add *Inbound Rules* for the services you want to access on your AMI
    - HTTP, SSH, "Custom TCP Rule" (42284) (20-21) (30000-30100), "All TCP" source: galaxy

Security Group: galaxy

Details | **Inbound***

Create a new rule: Custom TCP rule

Port range:

(e.g., 80 or 49152-65535)

Source: 0.0.0.0/0

(e.g., 192.168.2.0/24, sg-47ad482e, or 1234567890/default)

Add Rule

Your changes have not been applied yet.

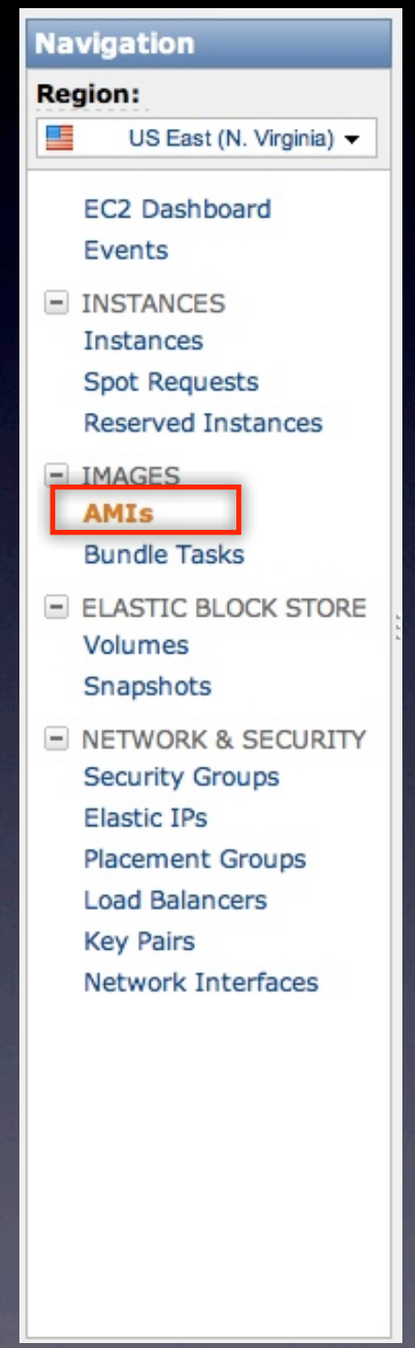
TCP Port (Service)	Source	Action
80 (HTTP)	0.0.0.0/0	Delete
22 (SSH)	0.0.0.0/0	Delete
42284	0.0.0.0/0	Delete
0 - 65535	sg-da5918b2	Delete
20 - 21	0.0.0.0/0	Delete
30000 - 30100	0.0.0.0/0	Delete

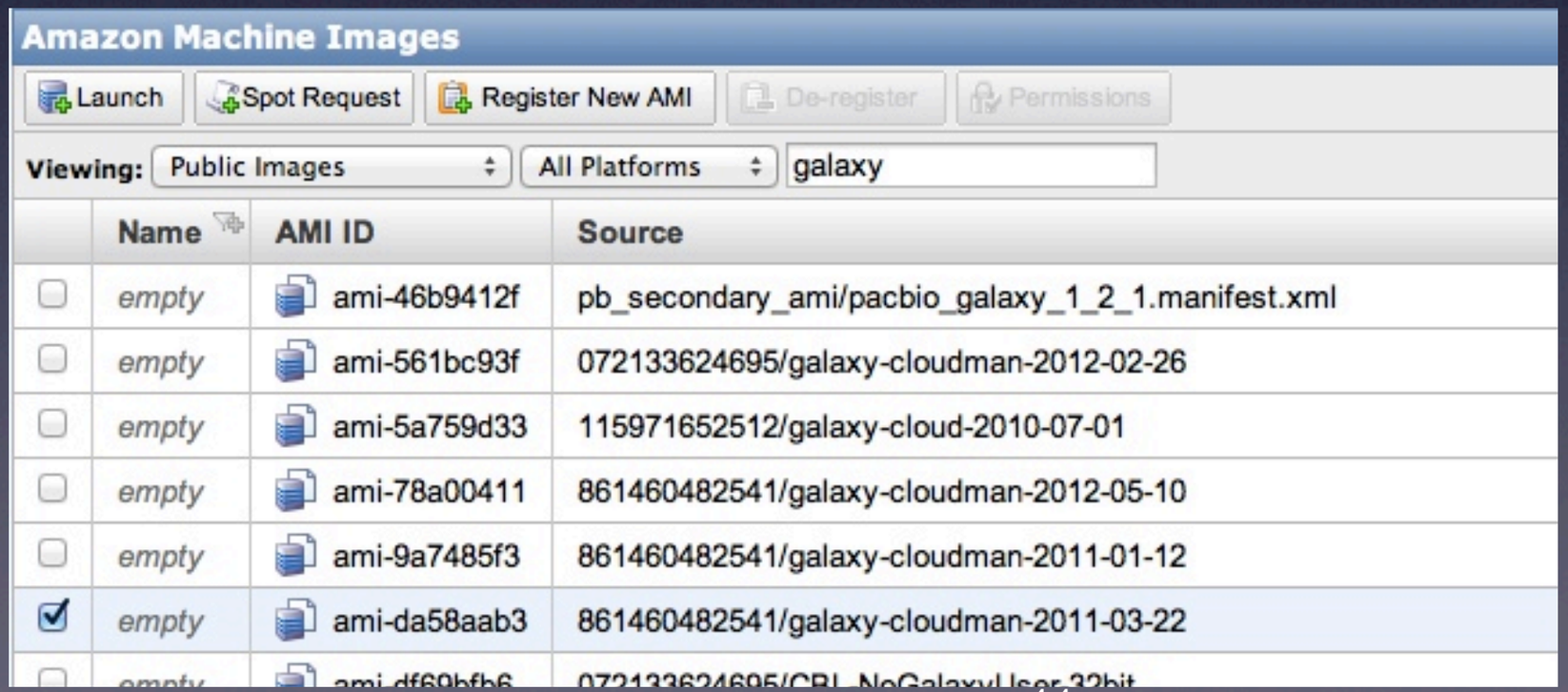
matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")

4. From the EC2 dashboard, select *AMIs*, and search for "galaxy" under *Public Images*

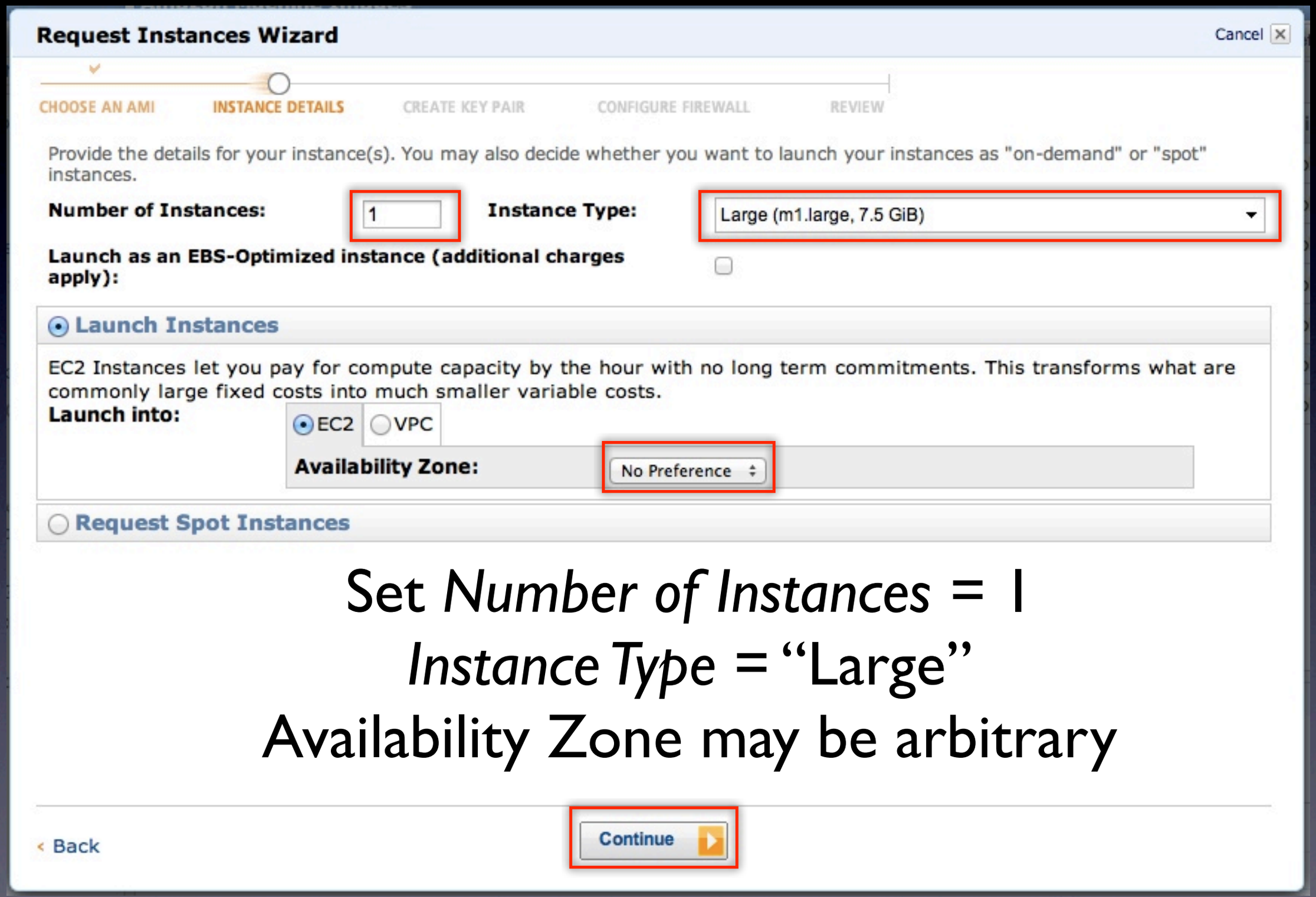
   - Choose "galaxy-cloudman-2011-03-22" and click *Launch*





matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")



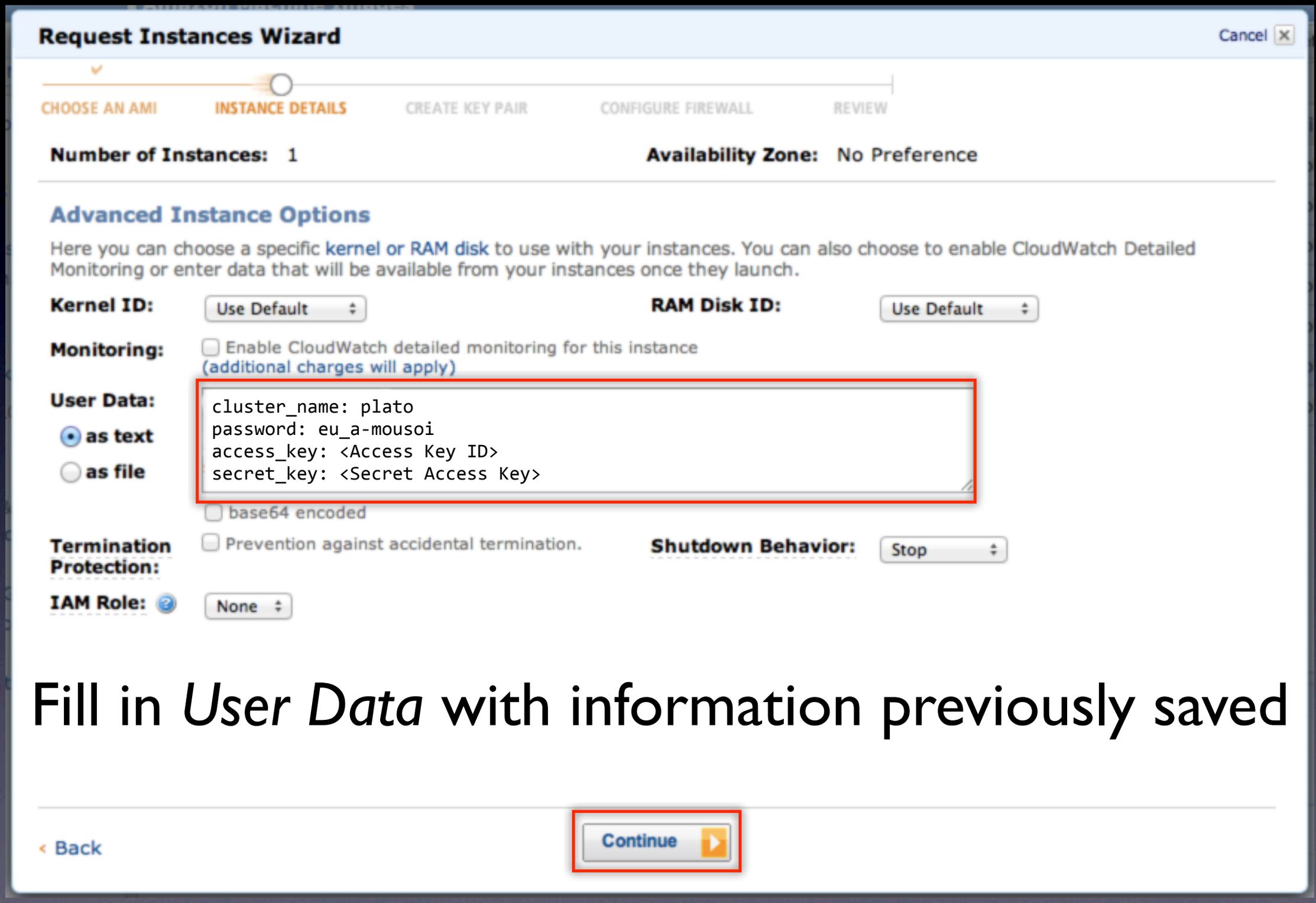
Set *Number of Instances* = 1

Instance Type = "Large"

Availability Zone may be arbitrary

matt.shirley@jhmi.edu

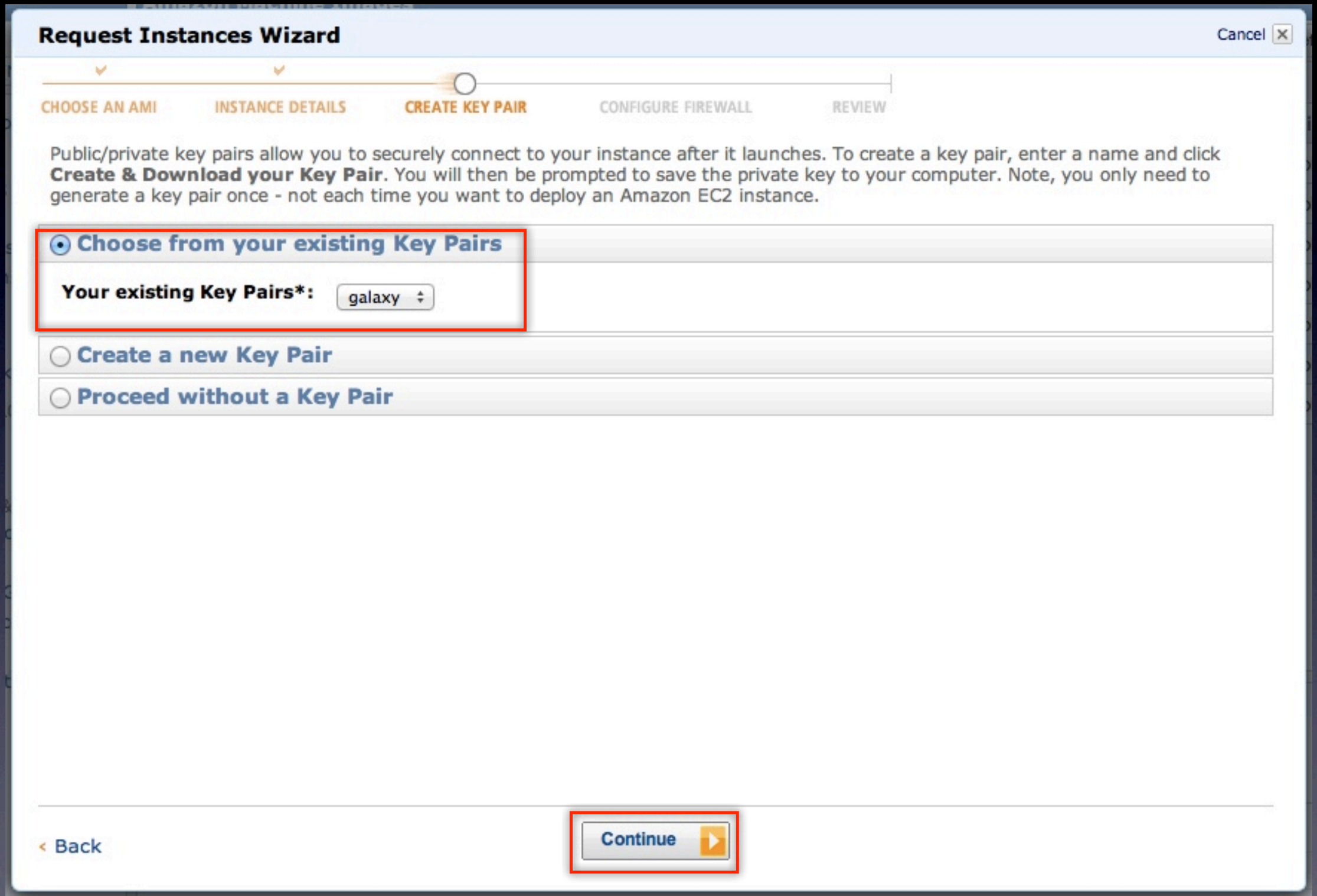
Galaxy on AWS ("the cloud")



Fill in *User Data* with information previously saved

matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")



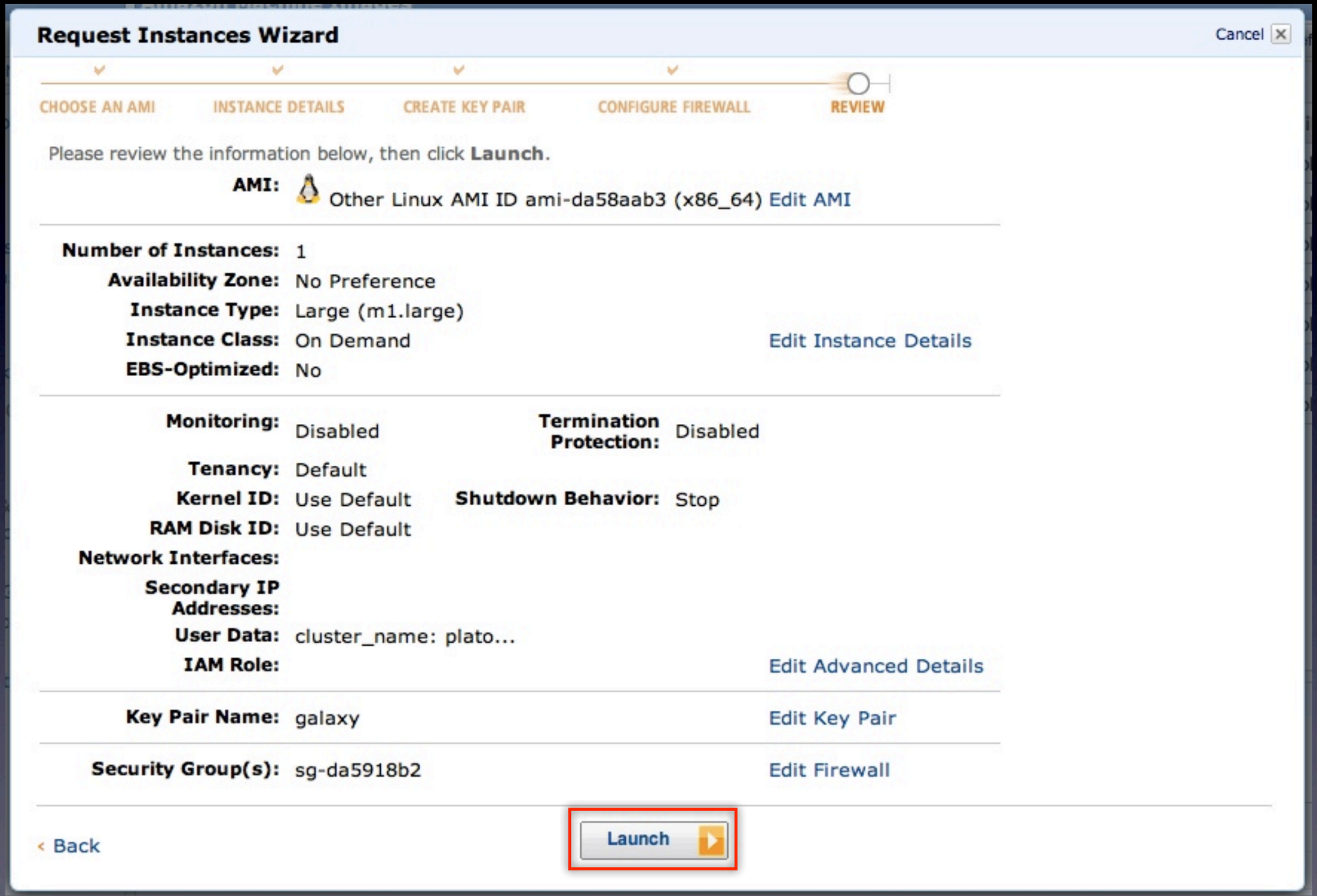
matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")

Choose your "galaxy" security group

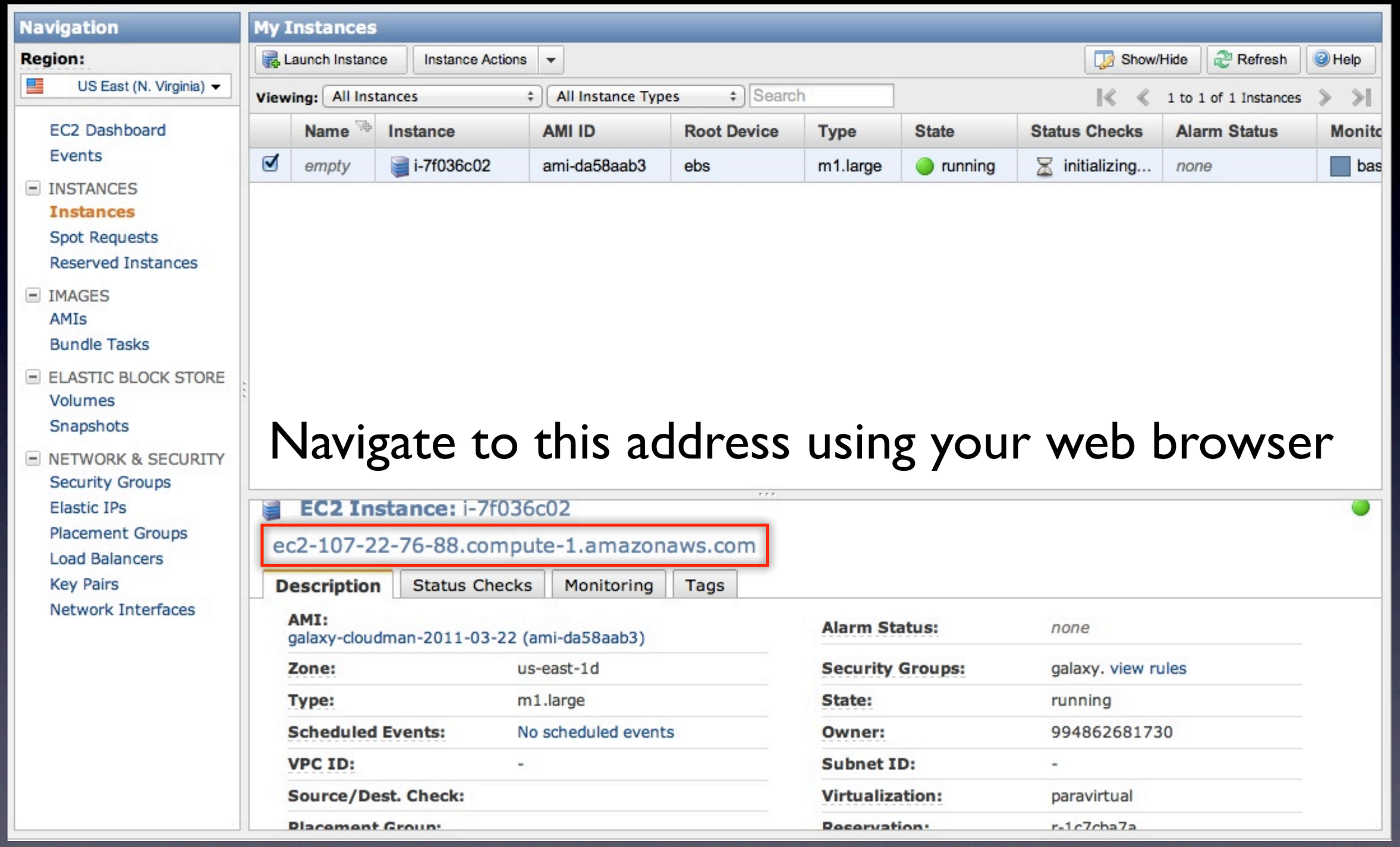
matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")



matt.shirley@jhmi.edu

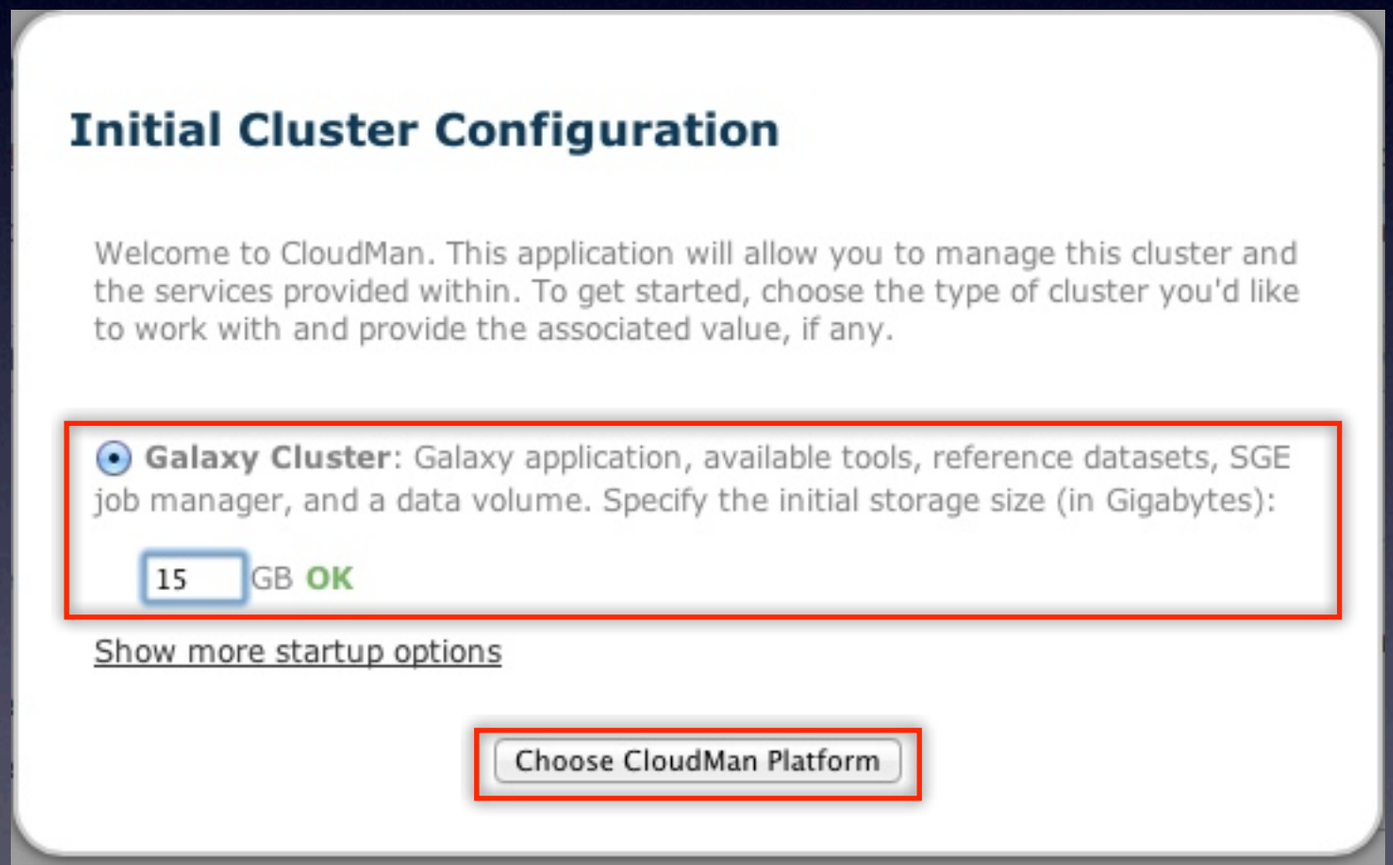
Galaxy on AWS ("the cloud")



matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")

5. After logging in using the previously specified "cluster name" and "password", specify the initial storage for the Galaxy cluster

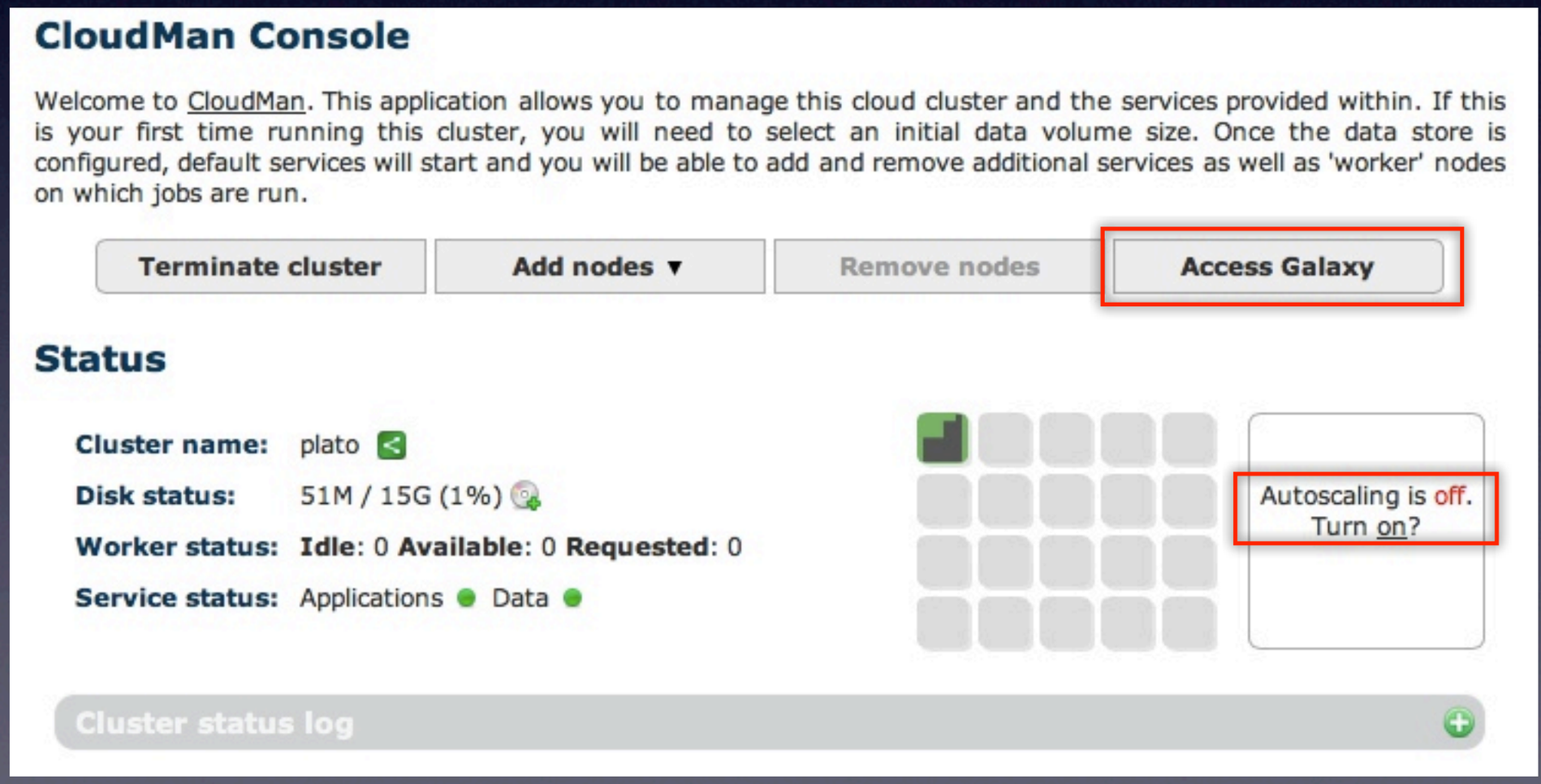


matt.shirley@jhmi.edu

Galaxy on AWS ("the cloud")

6. After a few minutes, the *Access Galaxy* button will become accessible, signaling success

- Note that performance will be improved if autoscaling is turned on



matt.shirley@jhmi.edu