bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

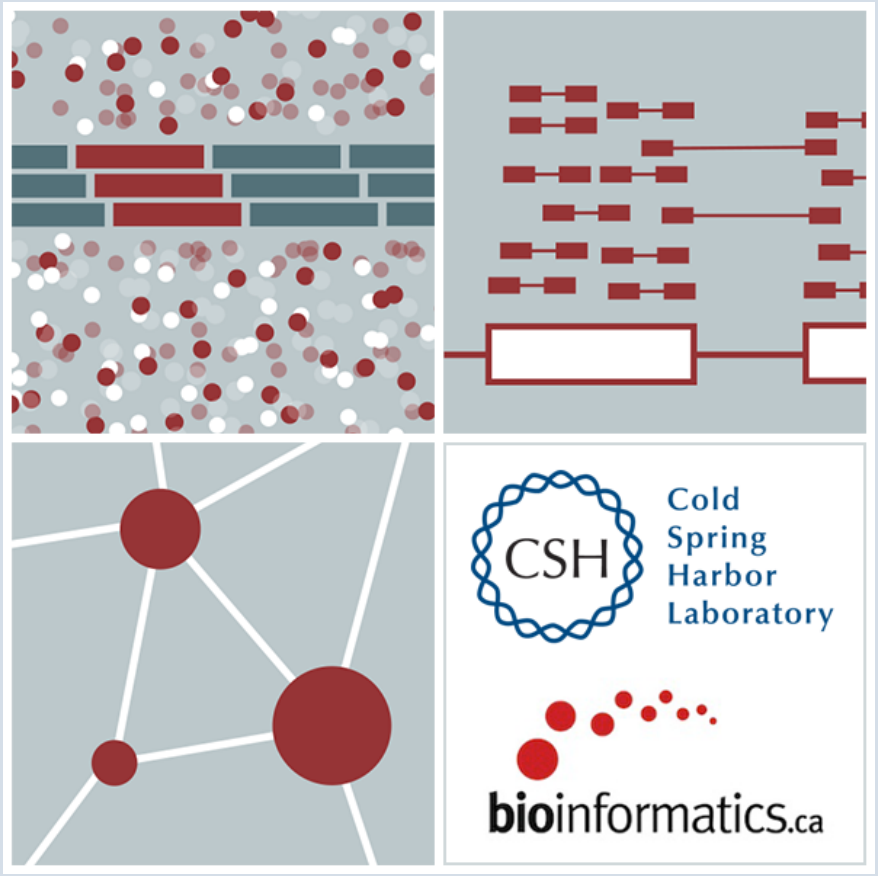
Introduction to cloud computing

Malachi Griffith, Obi Griffith, Fouad Yousif
Slides courtesy of Francis Ouellette

Informatics for RNA-seq Analysis

July 10-12, 2017

bioinformatics.ca

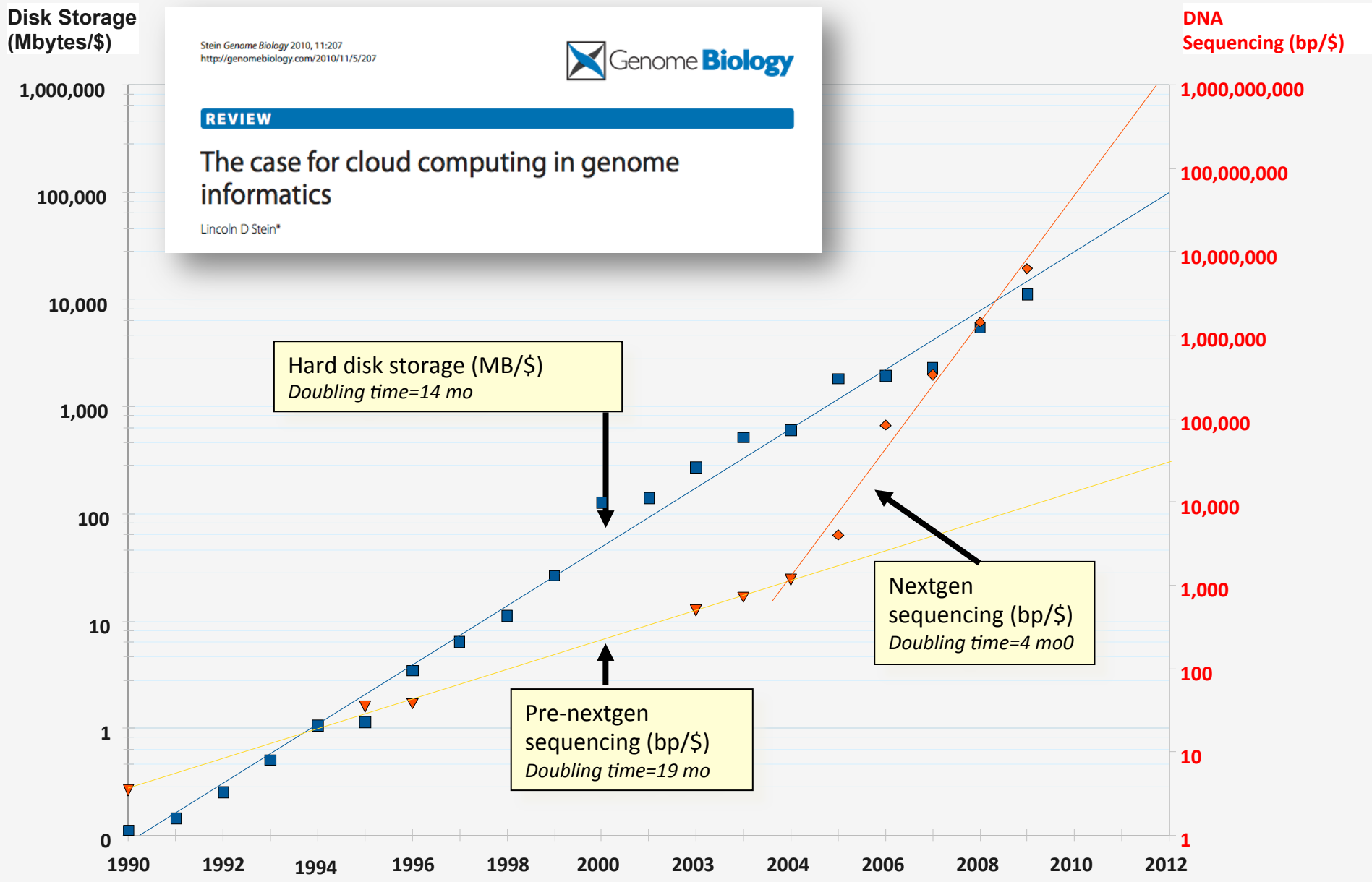


THE GENOME INSTITUTE
at Washington University

Learning Objectives

- Introduction to cloud computing concepts
- Introduction to cloud computing providers
- Use the Amazon EC2 console to create an instance for each student
  - Will be used for many hands-on tutorials throughout the course
- How to log into your cloud instance

Disk Capacity vs Sequencing Capacity, 1990-2012



Stein *Genome Biology* 2010, 11:207
http://genomebiology.com/2010/11/5/207

Genome Biology

REVIEW

The case for cloud computing in genome informatics

Lincoln D Stein*

About DNA and computers

- We'll hit the $1000 genome during 2015-?, then need to think about the $100 genome.
- The doubling time of sequencing has been ~5-6 months.
- The doubling time of storage and network bandwidth is ~12 months.
- The doubling time of CPU speed is ~18 months.
- The cost of sequencing a base pair will eventually equal the cost of storing a base pair

What is the general biomedical scientist to do?

- Lots of data
- Poor IT infrastructure in many labs
- Where do they go?
- Write more grants?
- Get bigger hardware?

Cloud computing providers

- Amazon AWS
  - https://aws.amazon.com/
- Google cloud
  - https://cloud.google.com/
- Digital ocean
  - https://www.digitalocean.com/
- Others I have not tried:
  - Microsoft Azure (https://azure.microsoft.com/en-us/)
  - Rackspace cloud (http://www.rackspace.com/cloud)

Amazon Web Services (AWS)

- Infinite storage (scalable): S3 (simple storage service)
- Compute per hour: EC2 (elastic cloud computing)
- Ready when you are High Performance Computing
- Multiple football fields of HPC throughout the world
- HPC are expanded at one container at a time:

Some of the challenges of cloud computing:

- Not cheap!
- Getting files to and from there
- Not the best solution for everybody
- Standardization
- PHI: personal health information & security concerns
- In the USA: HIPAA act, PSQIA act, HITECH act, Patriot act, CLIA and CAP programs, etc.
  - http://www.biostars.org/p/70204/

Some of the advantages of cloud computing:

- We received a grant from Amazon, so supported by 'AWS in Education grant award'.
- There are better ways of transferring large files, and now AWS makes it free to upload files.
- A number of datasets exist on AWS (e.g. 1000 genome data).
- Many useful bioinformatics AMI's (Amazon Machine Images) exist on AWS: e.g. cloudbiolinux & CloudMan (Galaxy) – now one for this course!
- Many flavors of cloud available, not just AWS

In this workshop:

- Some tools (data) are
  - on your computer
  - on the web
  - on the cloud.
- You will become efficient at traversing these various spaces, and finding resources you need, and using what is best for you.
- There are different ways of using the cloud:
  1. Command line (like your own very powerful Unix box)
  2. With a web-browser (e.g. Galaxy): not in this workshop

Things we have set up:

- Loaded data files to an ftp server
- We brought up an Ubuntu (Linux) instance, and loaded a whole bunch of software for NGS analysis.
  - Saved this as an Amazon Machine Instance (AMI)
- We will clone this and create separate instances for everybody in the class.
- We've simplified the security: you basically all have the same login and file access, and opened ports. In your own world you would be more secure.

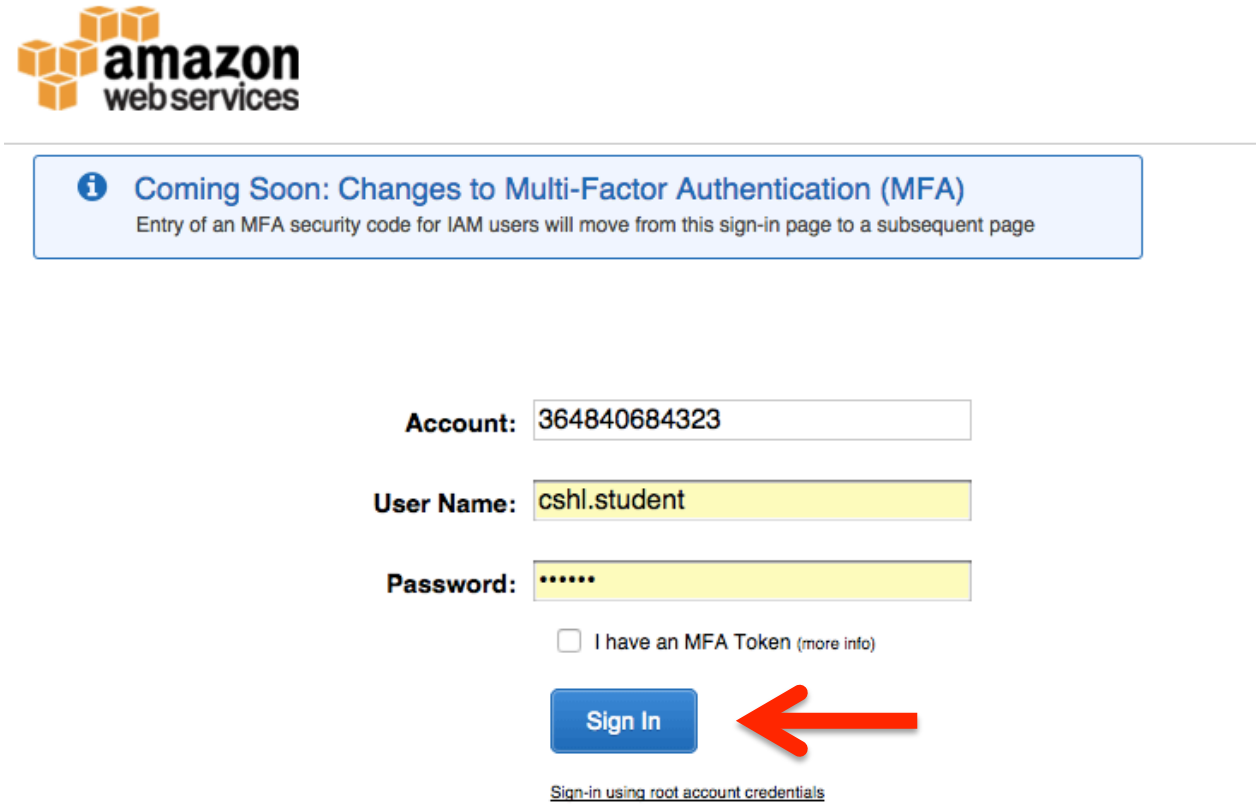
Amazon AWS documentation

https://github.com/griffithlab/rnaseq_tutorial/wiki/Intro-to-AWS-Cloud-Computing

http://aws.amazon.com/console/

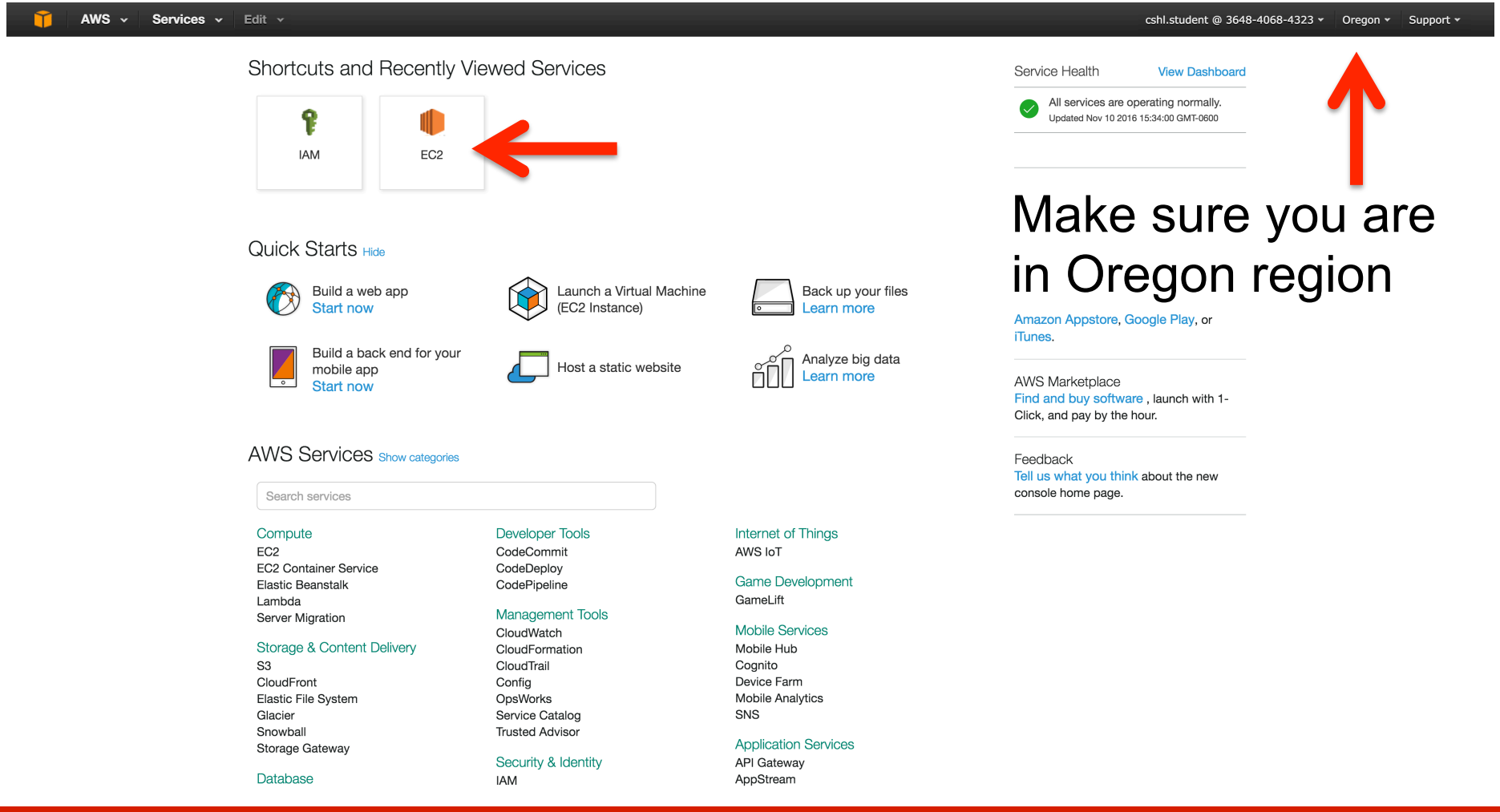
Logging into Amazon AWS

Login to AWS console

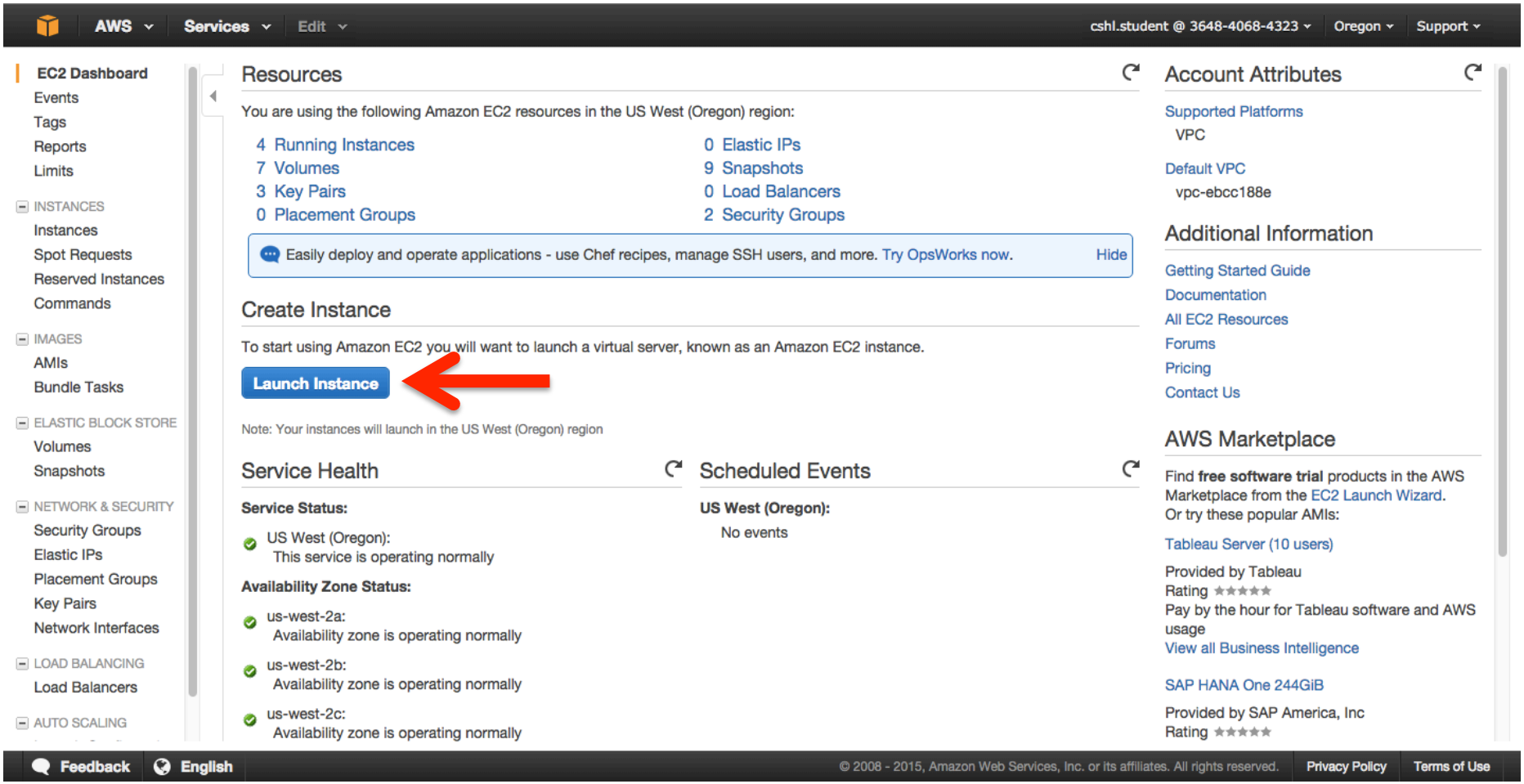


https://364840684323.signin.aws.amazon.com/console

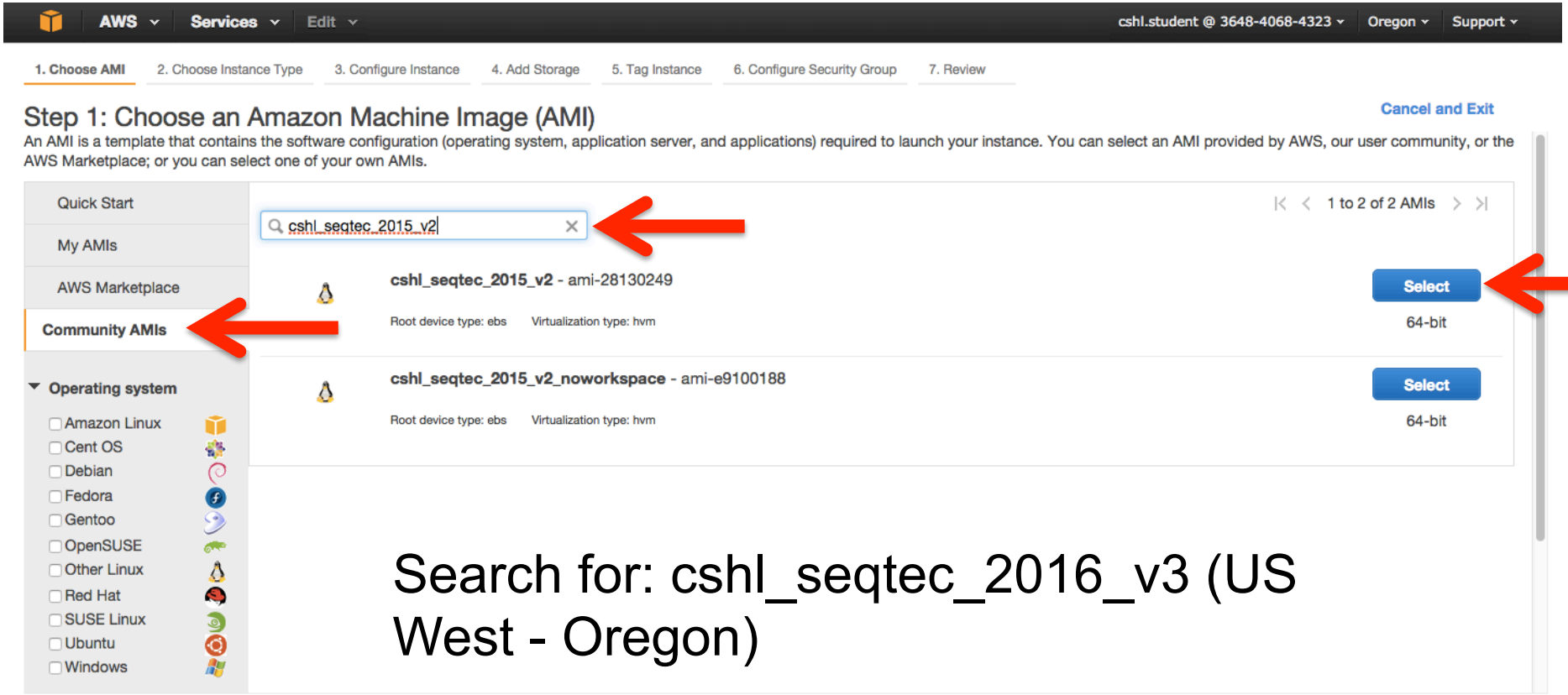
Select "EC2" service



Launch a new Instance

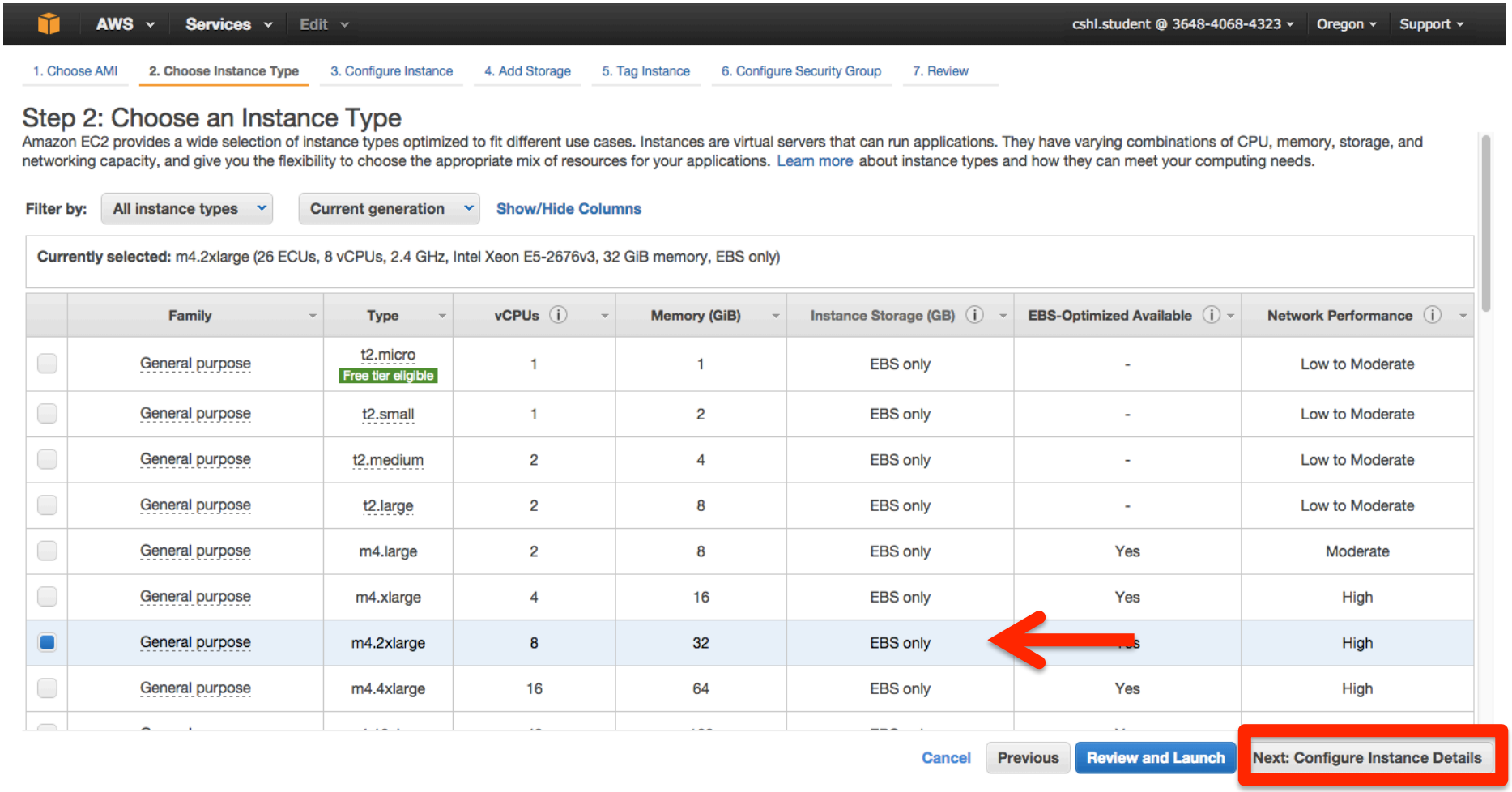


Choose an AMI – Find the CSHL SEQTEC 2016 AMI in the Community AMIs

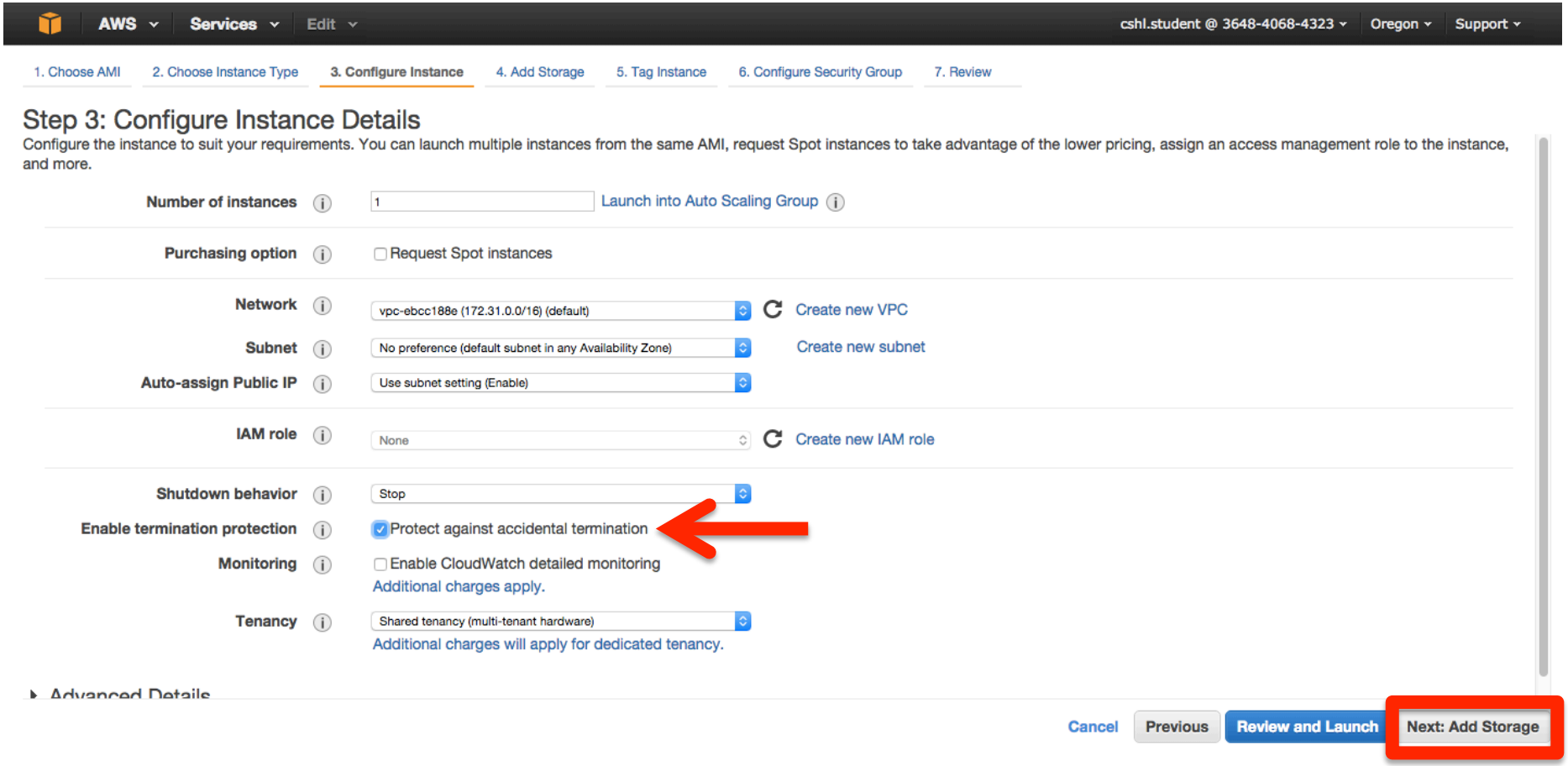


Search for: cshl_seqtec_2016_v3 (US West - Oregon)

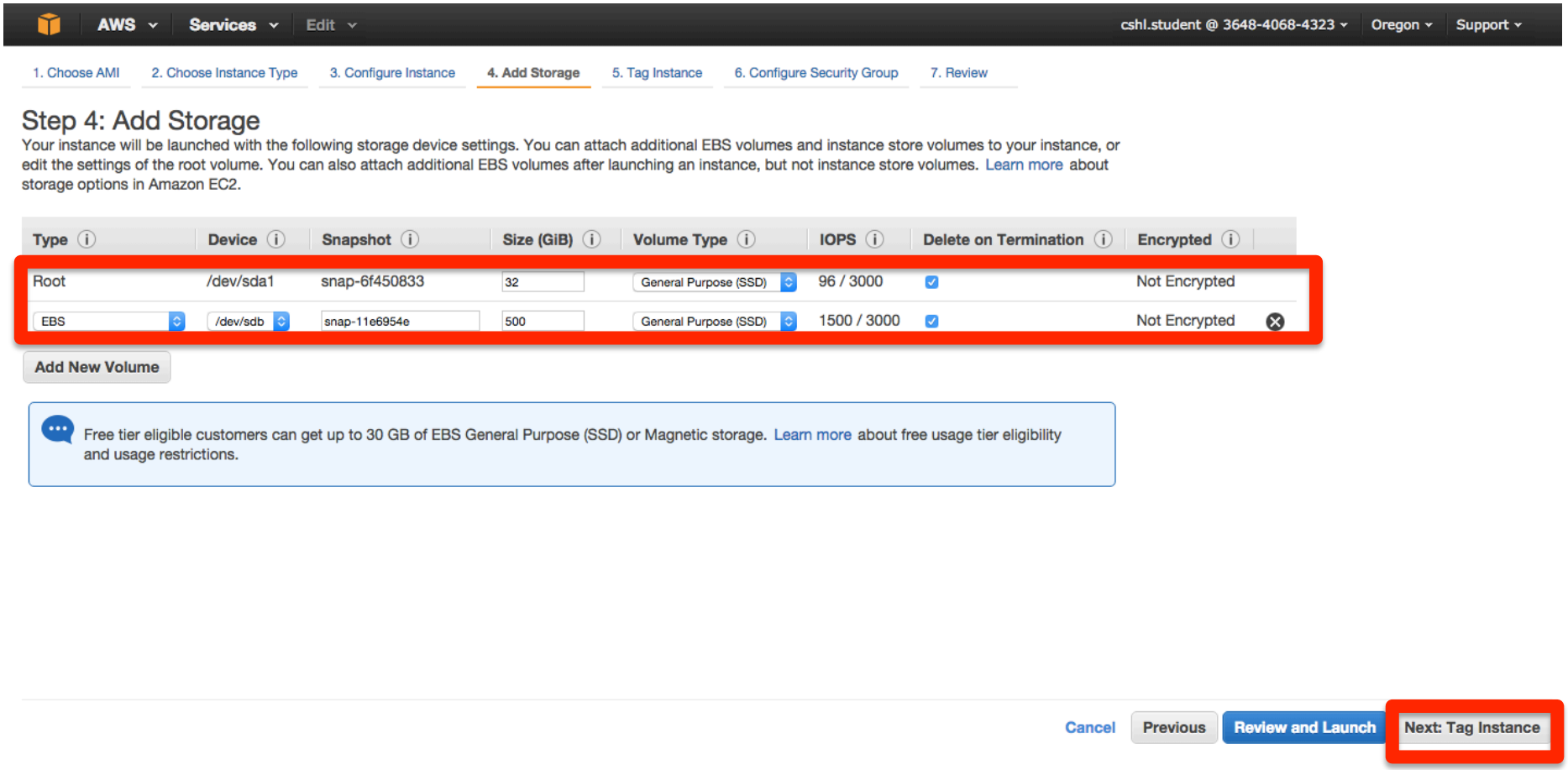
Choose “m4.2xlarge” instance type, then "Next: Configure Instance Details".



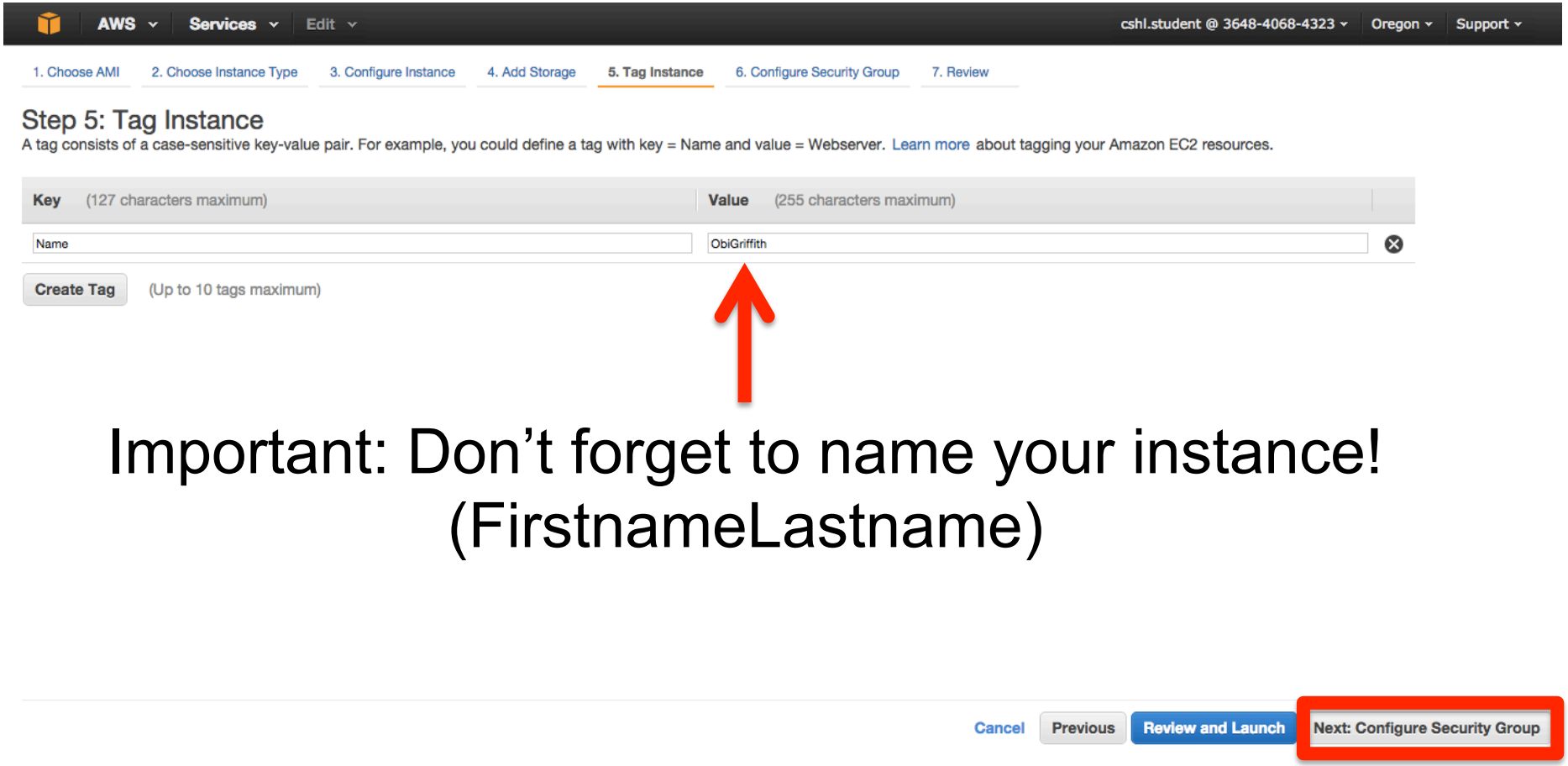
Select "Protect against accidental termination", then "Next: Add Storage".



You should see "snap-xxxxxxx" (32GB) and "snap-xxxxxxx" (80GB) as the two storage volumes selected. Then, "Next: Tag Instance"

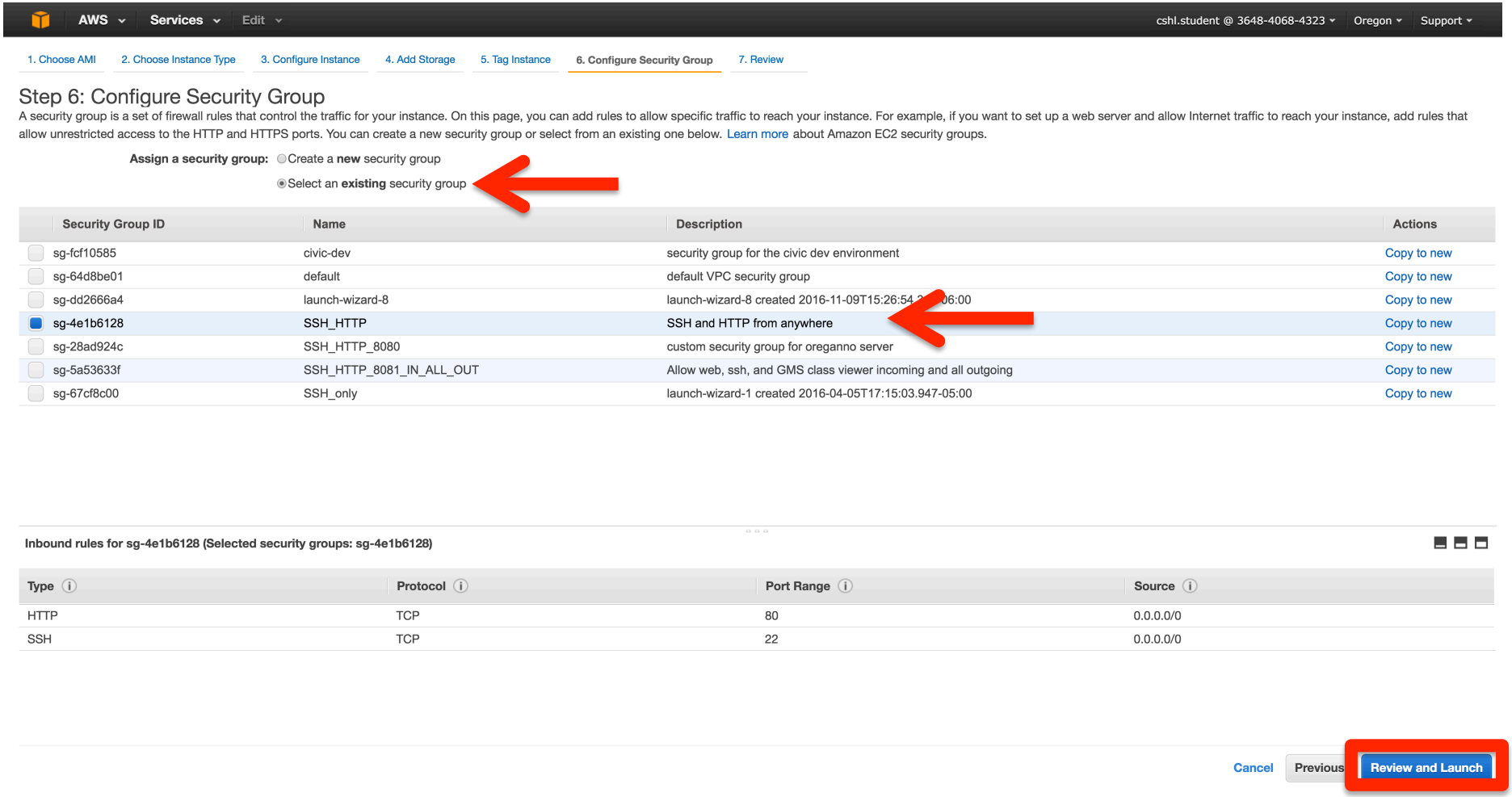


Create a tag like “Name=ObiGriffith” [use your own name]. Then hit "Next: Configure Security Group".

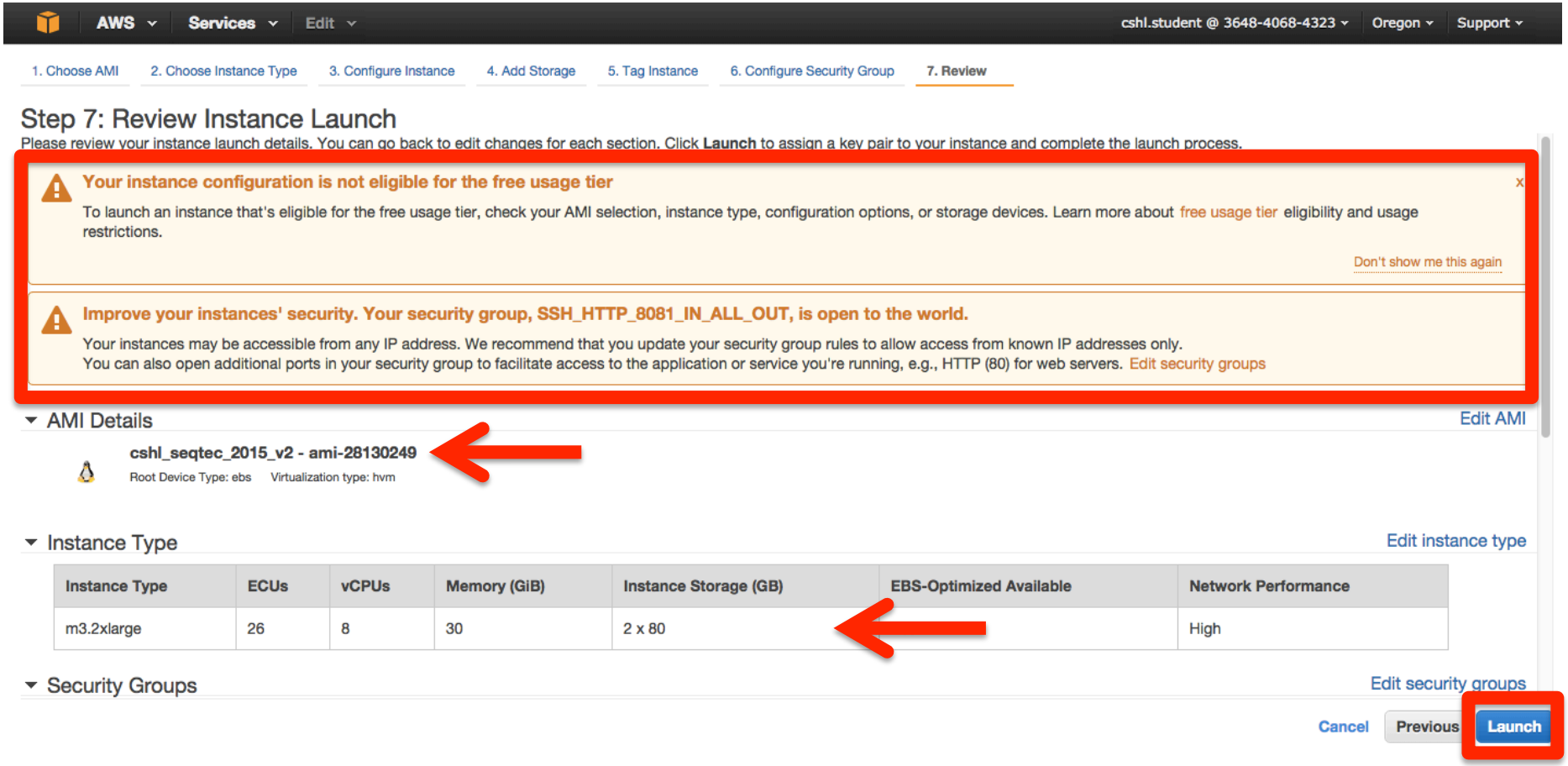


Important: Don’t forget to name your instance! (FirstnameLastname)

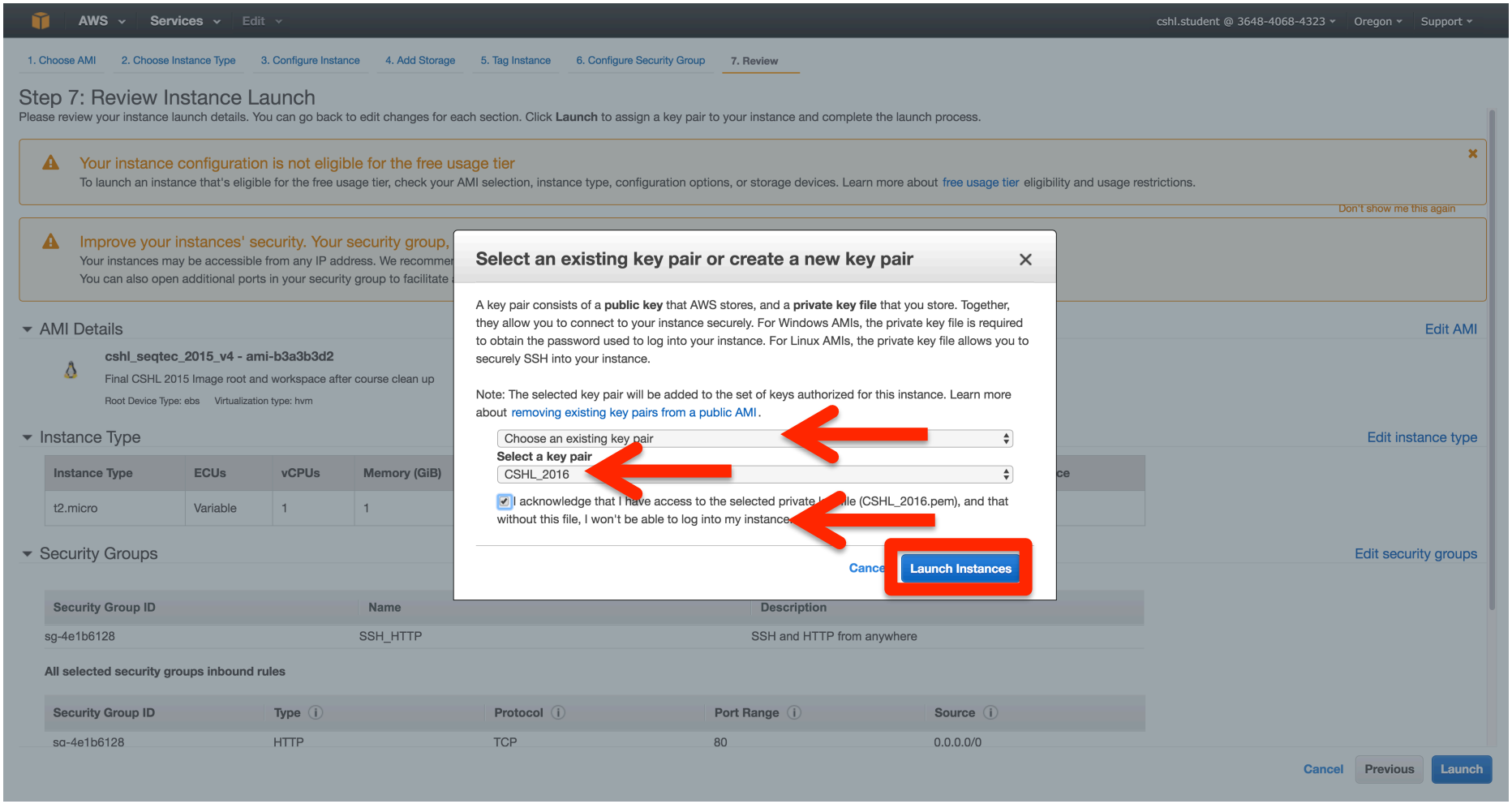
Select an Existing Security Group, choose "SSH_HTTP". Then hit "Review and Launch".



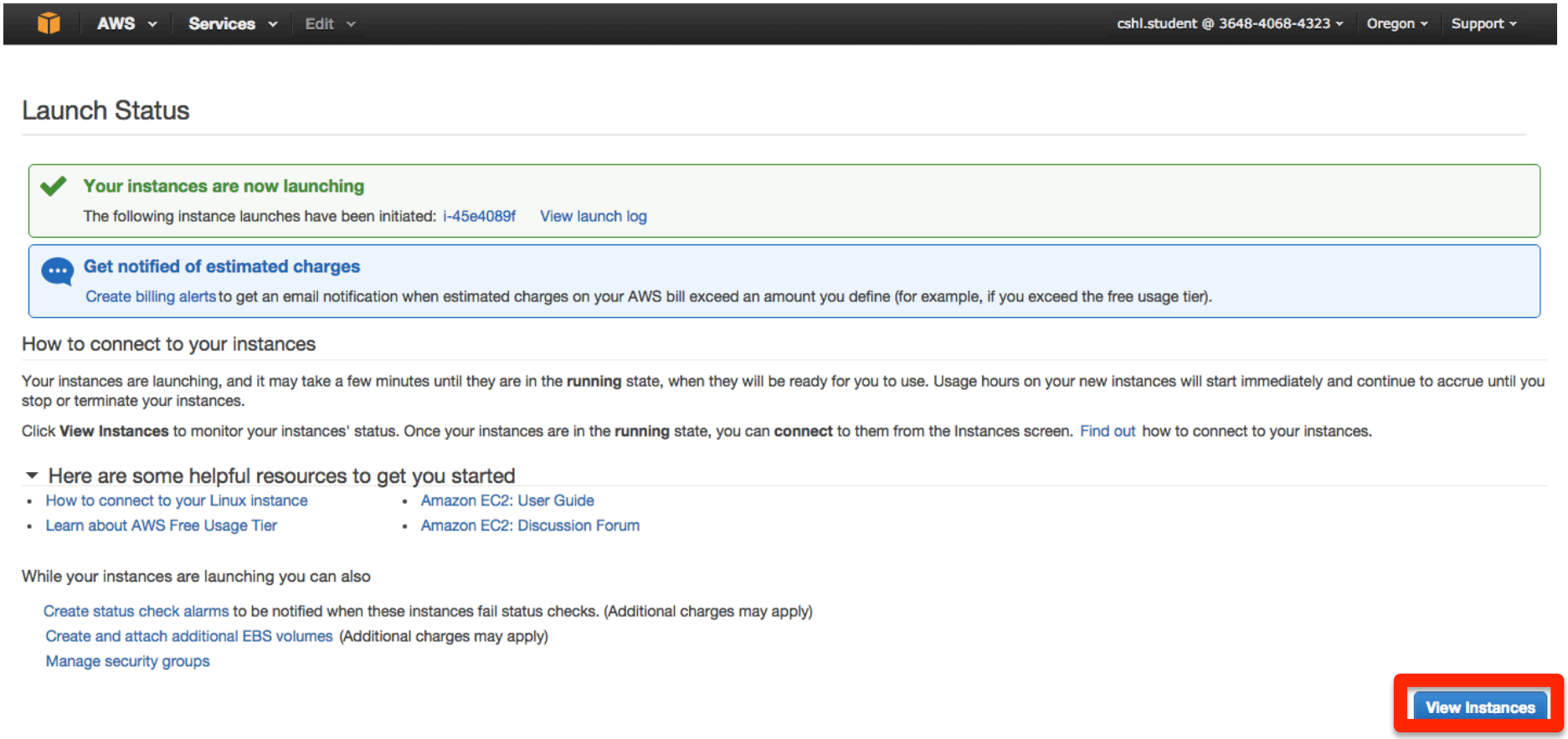
Review the details of your instance, note the warnings, then hit Launch



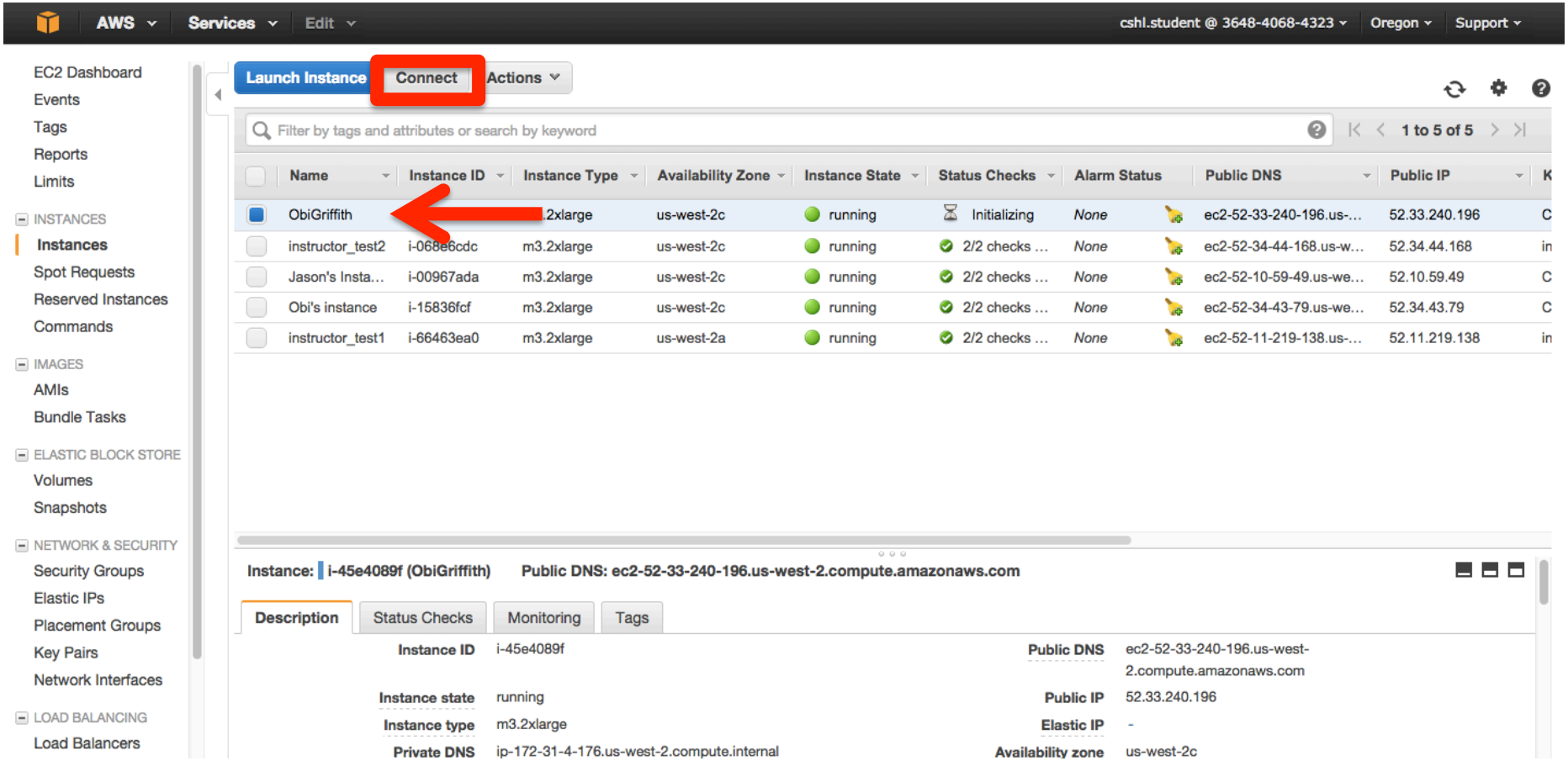
Choose an existing key pair: "CSHL_2016" and then Launch.



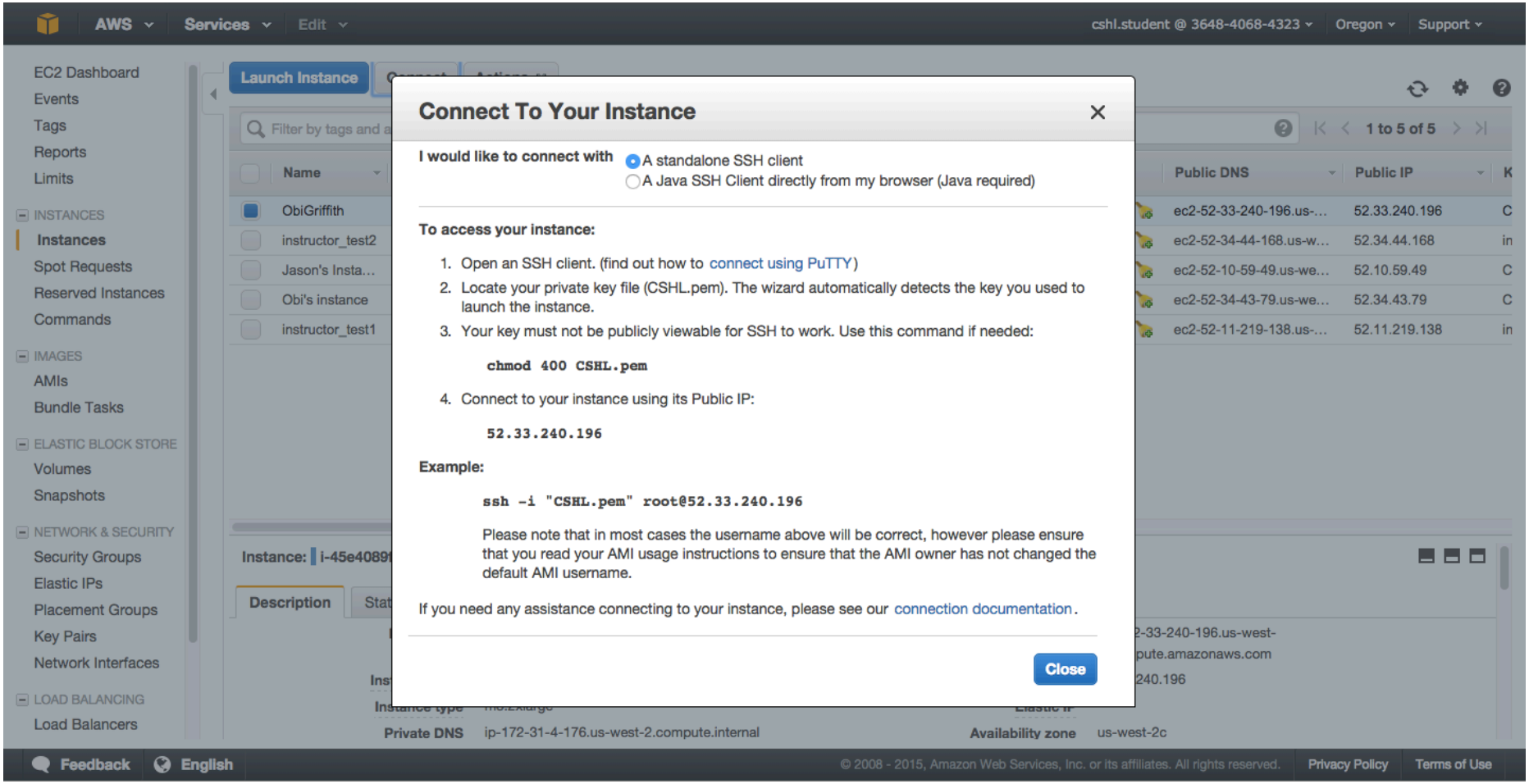
View Instances to see your new instance spinning up!



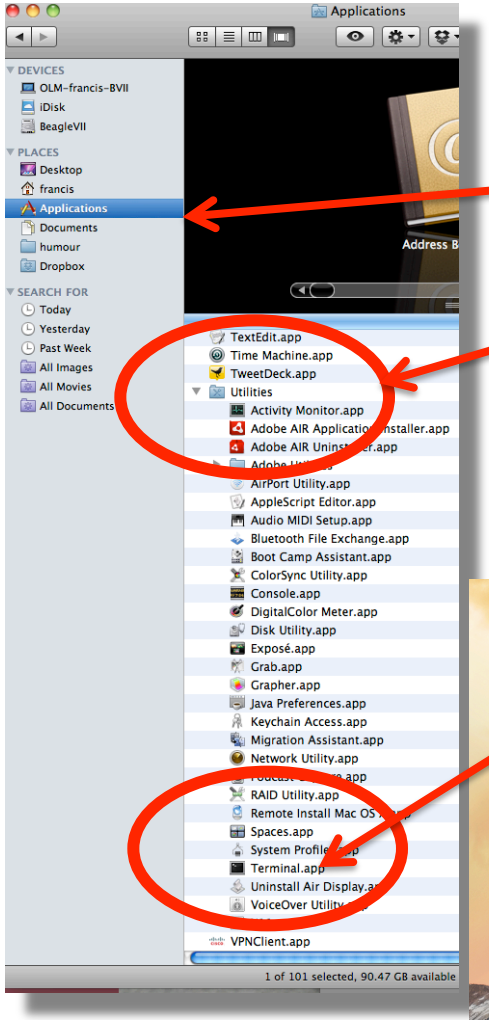
Find YOUR instance, select it, and then hit connect for instructions on how to connect



Take note of your Public DNS and the instructions on changing permissions for the key file (Note, we will login as ubuntu NOT root)



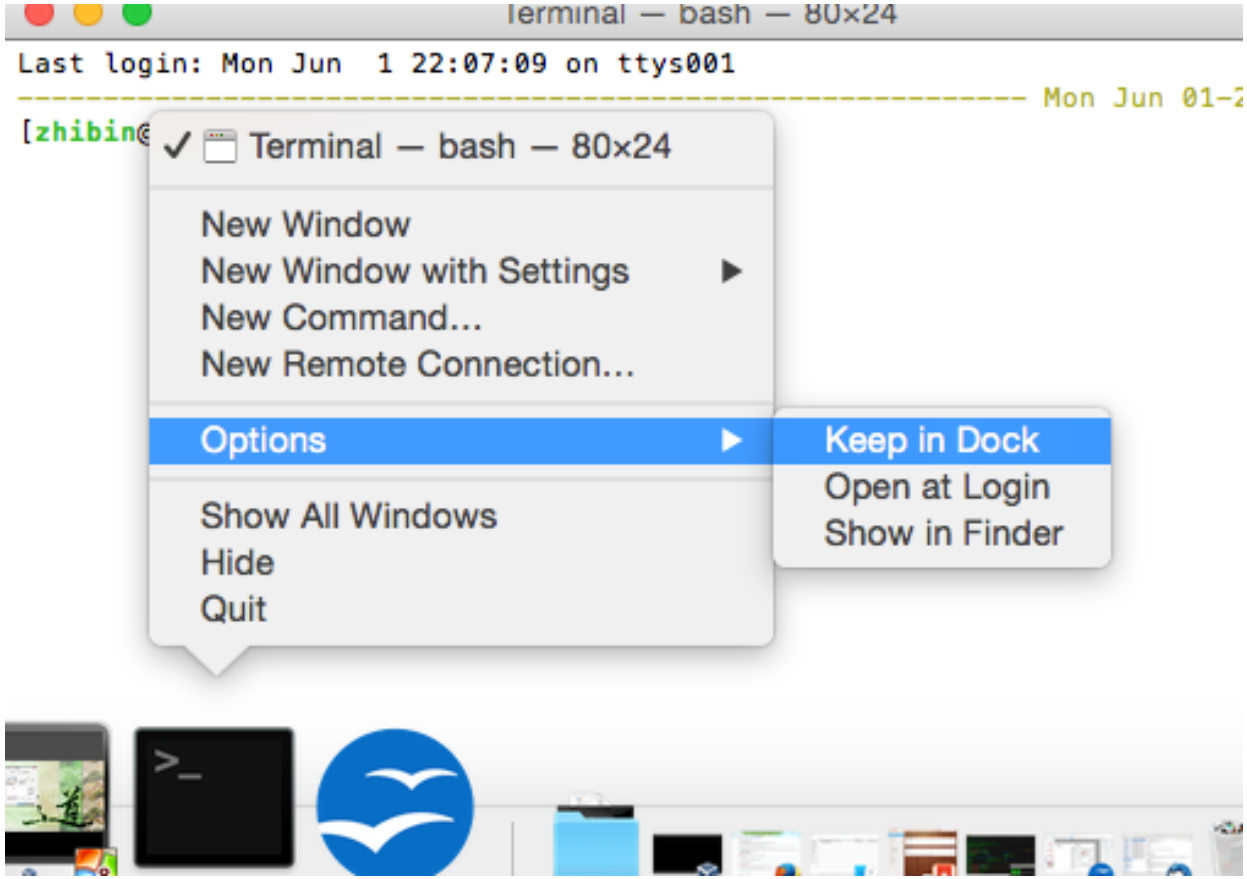
Opening a ‘terminal session’ on a Mac



In a Finder window
‘Applications’ -> ‘Utilities’ -> ‘Terminal’

Or on your dock

Add the terminal App to your dock

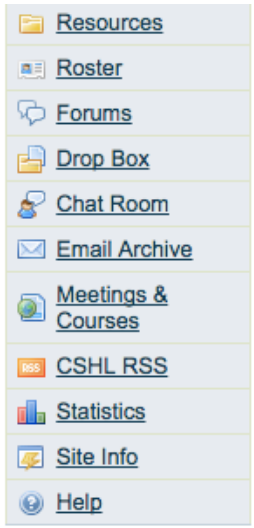


Creating a working directory on your Mac called ‘cshl’

```
obis-air:~ ogriffit$ pwd
/Users/ogriffit
obis-air:~ ogriffit$ ls
Applications    Desktop         Dropbox         Movies          Public          gittemp         temp
Attachments     Documents       Google Drive    Music           bin             igv
Box Sync        Downloads       Library         Pictures        git             ncbi
obis-air:~ ogriffit$ mkdir cshl
obis-air:~ ogriffit$ cd cshl
obis-air:cshl ogriffit$ ls -la
total 0
drwxr-xr-x   2 ogriffit  staff    68 Nov 13 22:18 .
drwxr-xr-x+ 58 ogriffit  staff  1972 Nov 13 22:18 ..
obis-air:cshl ogriffit$
```

mkdir cshl
cd cshl

Obtain your AWS 'key' file from course wiki



Presentations

20th November (Friday)

19th November (Thursday)

18th November (Wednesday)

17th November (Tuesday)

- Informatics for RNA-seq: A web resource for analysis on the cloud

16th November (Monday)

- **Obi, Malachi & Jason**: Informatics for RNA-seq: A web resource for analysis on the cloud

15th November (Sunday)

14th November (Saturday)

Malachi Griffith & Obi Griffith: In

AWS Sign In Console

- Username : cshl.student
- Password : seqtec

Connect to AWS via Terminal

- Download CSHL.pem
- chmod 400 CSHL.pem
- ssh -i CSHL.pem ubuntu@**YOUR_IP_ADDRESS**

Jason Walker: Unix Command-line Bootcamp

On Mac:
Control+
Save Link As

Go to course wiki, "Presentations" page

Save key file to your new 'cshl' directory

Viewing the 'key' file once downloaded

```
obis-air:cshl ogriffit$ cd ~/cshl/
obis-air:cshl ogriffit$ ls -la
total 8
drwxr-xr-x   3 ogriffit  staff   102 Nov 13 22:21 .
drwxr-xr-x+ 58 ogriffit  staff  1972 Nov 13 22:18 ..
-rw-r-----@  1 ogriffit  staff  1696 Nov 13 22:21 CSHL.pem
obis-air:cshl ogriffit$ cat CSHL.pem
-----BEGIN RSA PRIVATE KEY-----
MIIEpgIBAAKCAQEAvJ5gwmtby9QZ2Idz+ugiEQQHW6PsOZAZFvr+mWDnM4pKpccaVmDh7XjcEOLF
OkJzaP9+jj0kSFOyNinitoB32DgrmVhgNhyheEqH5XMn28szxUj1EuoNXAogNuY7mWMo6MoWssSW
Rqy+rj19vMGQn5rsnMLjCM1smebPoqYOL8EPa1ccRbdGXG1dMTlCC1ho/Hk9bZweamGiZLaAWVmf
zOK/L0zxgY3K4cwaL48HV6oGuMh5lTDpnobxXghQ4oC5Mej+DpCRF8C+EG2uNDuyuLzRJfQmFBV2
GKDWDwhdgGmKmX9IpMT9ubvNoQPy0vYLvM80eG3cMbz2IZpaNryihwIDAQABAoIBAQCZYT0TvFO4
a3DdCEEC/rN9HMaS+bjFkm0kp9RTi15XJhTPvBmptjzibA6gWJfDaXgKIQGbzxJrEkxwCR2IBO3v
0LV7jEcomZ2ggRMDPeJitFoUCuDnkZZtivppSk2azOzeaD+O/ZeqPx0L+Yr+7HSbpVLVoxEV/l5a
xDuCawBMSY2cnGWKfEBlSPnB6fGZj8luGzv0aP/CETx/K78TIS56m4yrTIQIeEPfFt/PQr/EUqoL
7co5oy9K3sD1noPLDhk3vJa1VNrMjHkMZLkbZuaoHPzgSQHninm80Ca25WWTGsSZ8vQsBIUTlGI1
W7lzXH3wD1jJNd+03QK4bnKaZ+DZAoGBAPVpisa49JY/6K2f9B8naqtX/ljzVWTl3Q7r6t6uh21Y
oexmC8eJ2wQwd0qNjZWVxSMVksIwdM6xcsBIJRMmltWTVdmD0fkDvOfjd8CM4nctH76tvSvZz02e
qI9wSshHY1fh+09CoLZeefFSURxqWbkJfREjoZ4UGUWMi3k1rxC9AoGBAMTBlBBOWQ+5ojzQYu0L
Q4YrsIPg1/ni0WmJ+05vcTCJ2aeI88VhK5c2PoXPWWiJ9CdD2VFZDiCm2XuJA5iwJmnhuwGGHHEn
BFBqEF/ueJrW+r43pRcYRuRIXjiH4mQQlK4Zemecym5fAHvxZxq4fs2kWfMPySFaVufcP0VC7X6T
AoGBAMhro0xbrFQwaUOyh9oRhMneGPhn8WtvVjNjc/LcMfmZEtRPGnuhF965/hJCvEhXgiH+8lXo
4NwUixSBVtXnA/P0WX5Ea2ykIth2Kkx0Qlb14SEGHqH7RZOsaRiLqmcZ9gXFpkm6rimByrDMezVr
nU7CcwNWSB0jaOgluZoJv6k5AoGBAJJuFsmD5ZhkaS+lTtpnlZtXDIk5XsMkYQGQpS0clzqufQPI
UtPEm3Jv9lwTktDQSpqmTifShUcbpaPgtoJ+JjiKvGhH7QbxKK7IIOOkULG76OSD+SOU972Rdj3Q
M1aRWHWxlH1kHOvDXFLhuAAU6poVBLR2PRPLbf4k1hmv05xtAoGBAJVQy1GF8uVNwkOCNzLIqmkY
uk9M24hfqn3N2GY3Zgqf43bD4kdYgL4rvsgpO8QzotPf+19kVlCv0ciolSjEHLyUdlyPGzj4CTTH
1f1RoGHmYzVn9VuFTu4hJ17J+uwgXgIr9Sx/UTjwkmCjPf7CEyIuGxaThG/ZoR9stufZB5db
-----END RSA PRIVATE KEY-----obis-air:cshl ogriffit$
```

Changing file permissions of your 'key' file (Mac/Linux)

ls -l (long listing)

drwx------+ 67 ogriffit staff 2278 22 May 21:25 ../

-rw-r--r--@ 1 ogriffit staff 1696 22 May 21:31 CSHL.pem

rwx : owner

rwx : group

rwx: world

r read (4)

w write (2)

x execute (1)

Which ever way you add these 3 numbers, you know which integers were used (6 is always 4+2, 5 is 4+1, 4 is by itself, 0 is none of them etc …)

So, when you have:

chmod 400 <file name>

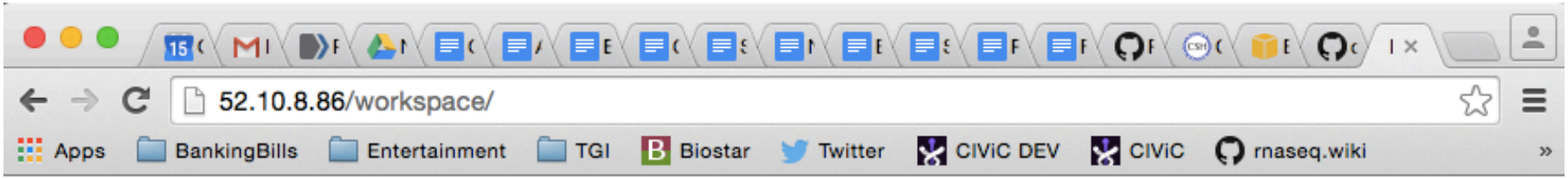
It is "r" for the the file owner **only**

Logging into your instance

Mac/Linux

```
cd cshl/
chmod 400 CSHL.pem
ssh -i CSHL.pem ubuntu@[YOUR PUBLIC DNS]
```

Copying files from AWS to your computer (using a web browser)



Index of /workspace

Name	Last modified	Size	Description
Parent Directory		-	
Homo_sapiens/	2015-11-13 06:45	-	
README.txt	2014-06-17 23:53	5.3K	
bam-demo/	2015-11-14 21:03	-	
data/	2015-11-13 01:39	-	
scratch/	2015-11-13 19:43	-	
tools/	2015-11-13 01:54	-	

Apache/2.4.7 (Ubuntu) Server at 52.10.8.86 Port 80

http://[YOUR PUBLIC DNS]/

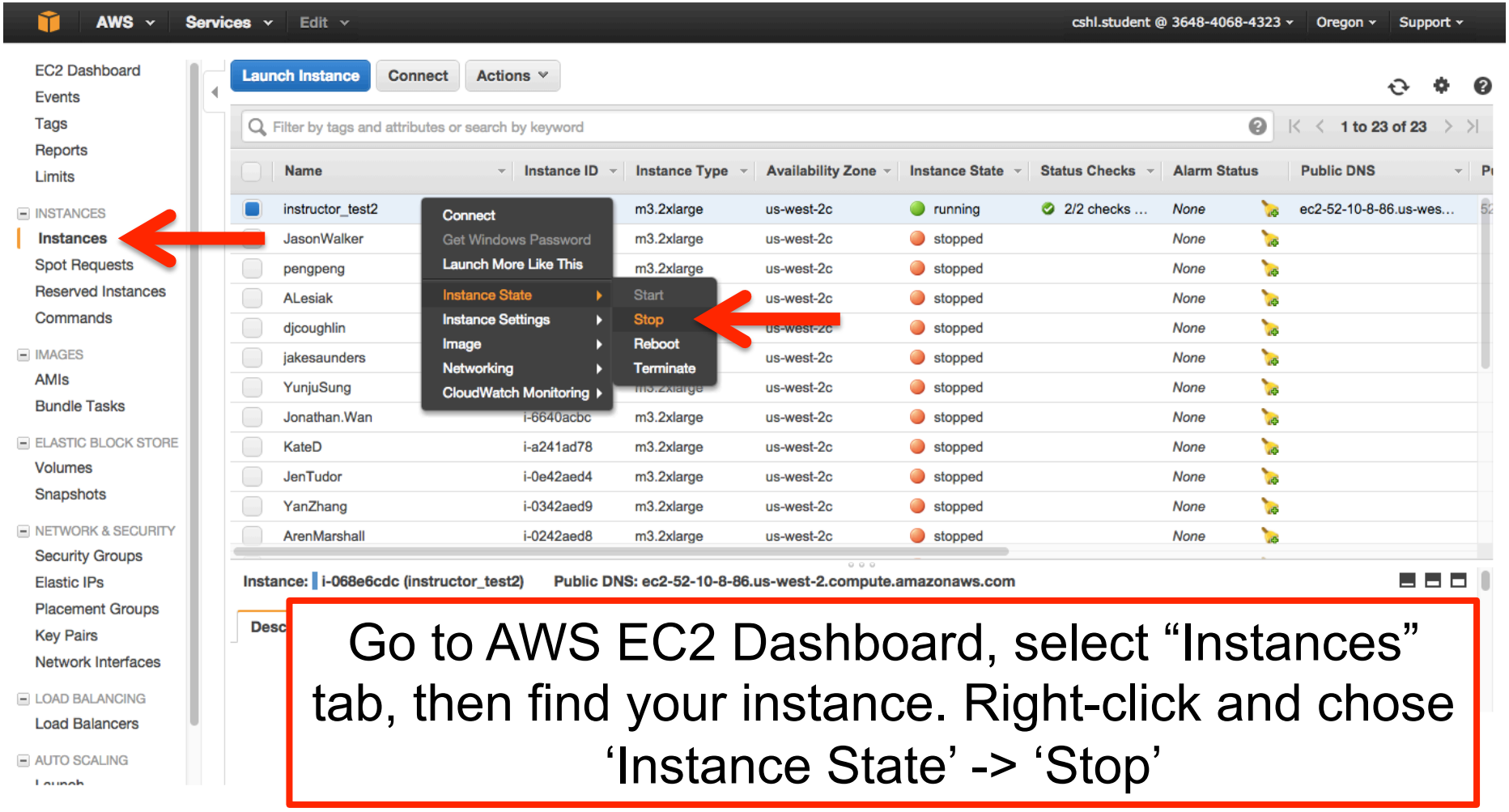
Logging out of your instance

Mac/Linux – simply type exit

```
exit
```

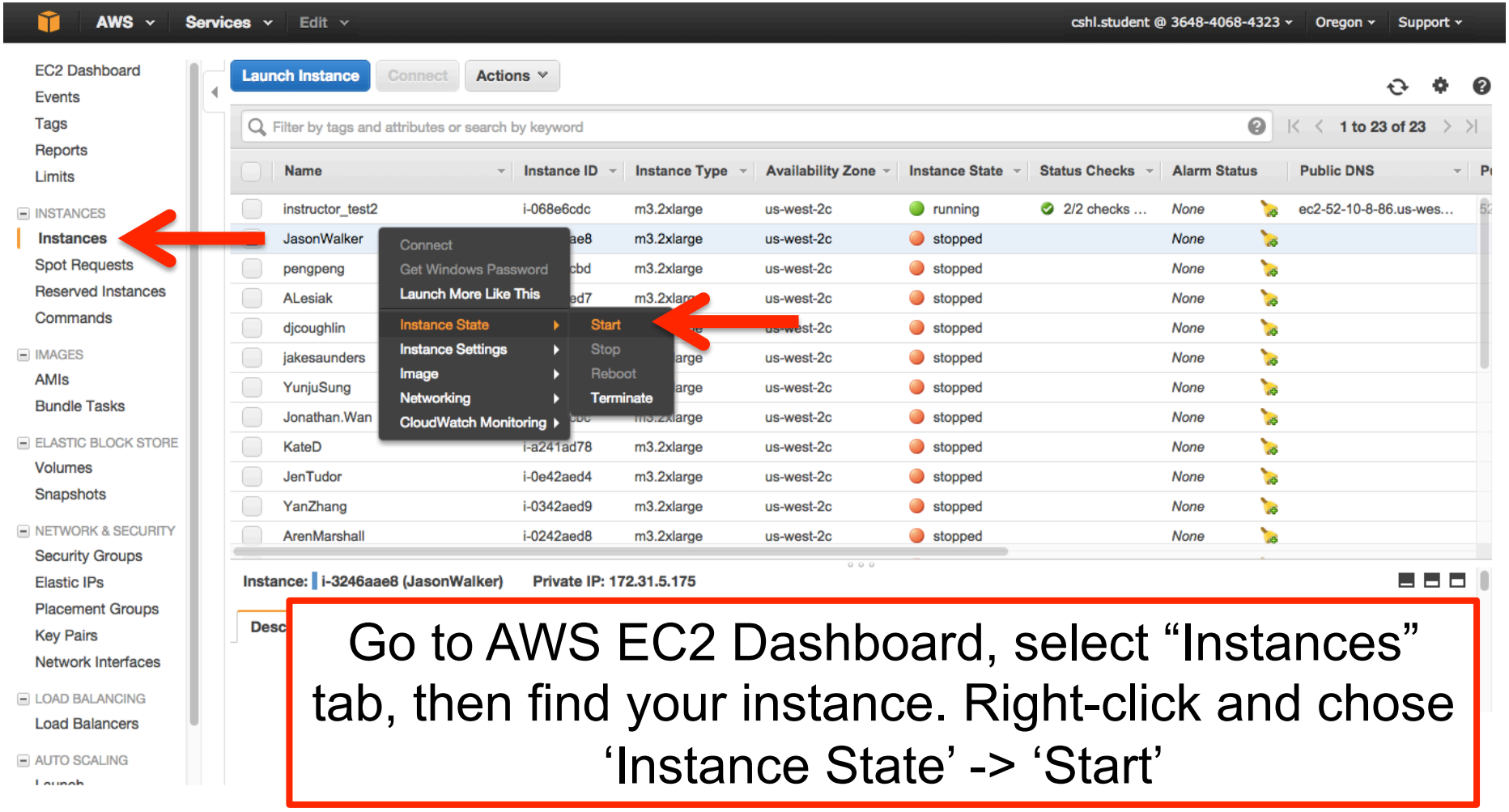
Note, this disconnects the terminal session (ssh connection) to your cloud instance. But, your cloud instance is still running! See next slide for how to stop your instance.

When you are done for the day you can "Stop" your instance – Don't Terminate!



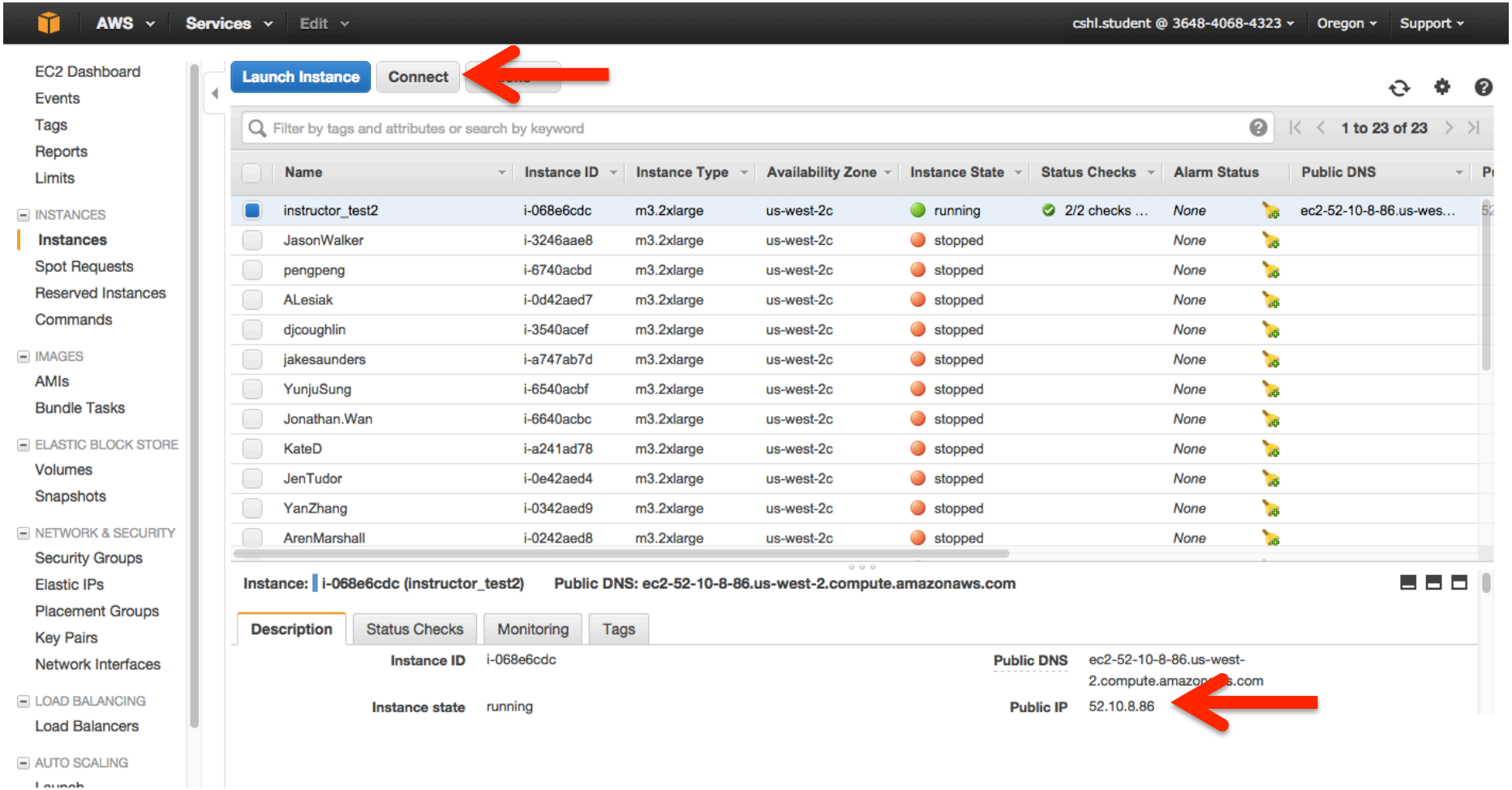
Go to AWS EC2 Dashboard, select "Instances" tab, then find your instance. Right-click and chose 'Instance State' -> 'Stop'

Next morning, you can “Start” your instance again



Go to AWS EC2 Dashboard, select “Instances” tab, then find your instance. Right-click and chose ‘Instance State’ -> ‘Start’

When you restart your instance you will need to find your new Public DNS or IP address. Select your instance and "Connect" or look in Description tab. Then go back to instructions for "Logging into your instance"



So, at this point:

- Your Mac desktop is ready for the workshop
- If it is not, you know where to get the information you need
- You know how to login to AWS
- The next step is to login to your linux machine on AWS and learn the basics of a linux command line

We are on a Coffee Break & Networking Session