Cold Spring Harbor Laboratory

# Introduction to cloud computing

John Chamberlin, Kelsy Cotto, Felicia Gomez, Obi Griffith, Malachi Griffith, Simone Longo, Allegra Petti, Aaron Quinlan, Megan Richters, Huiming Xia

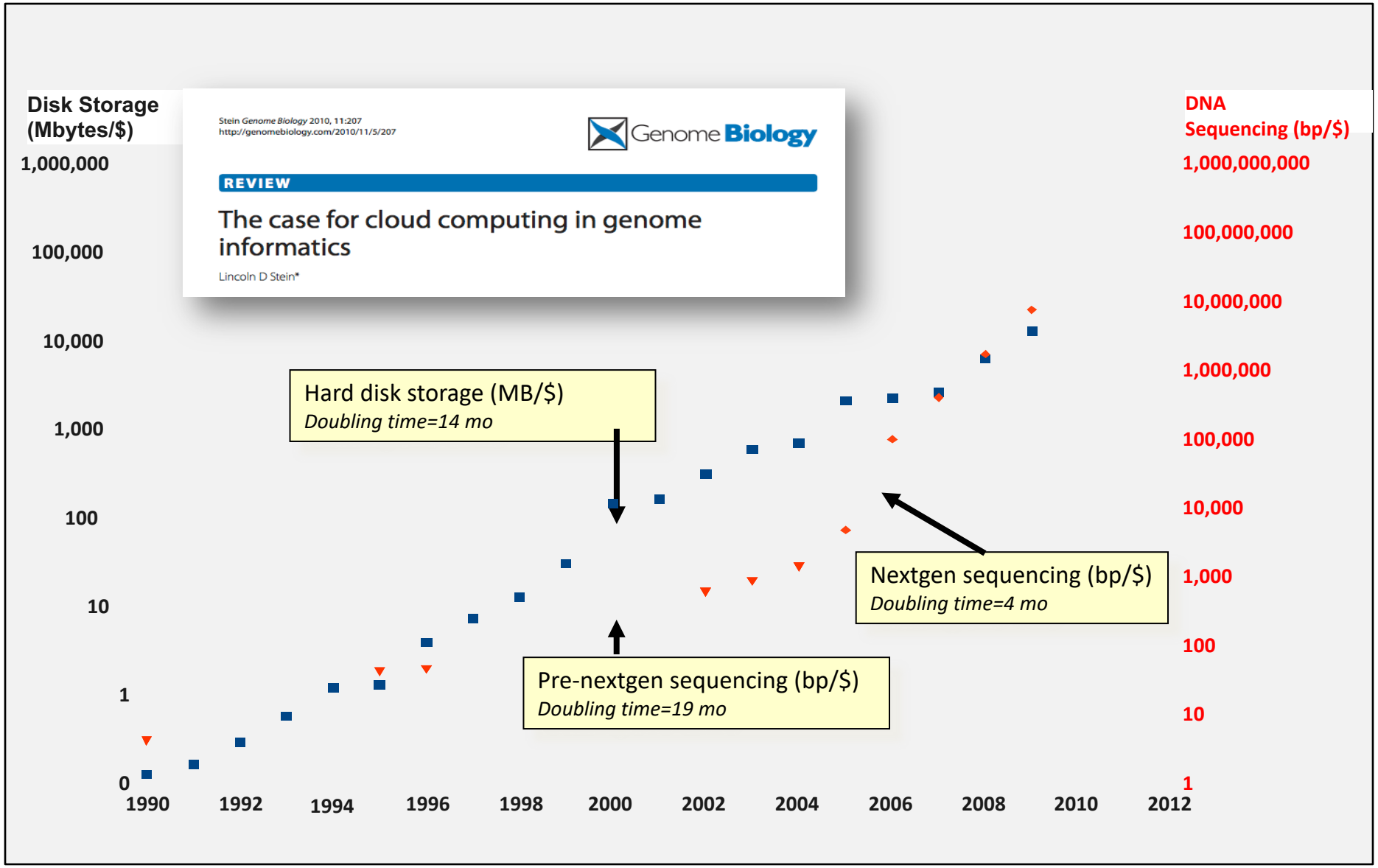
Advanced Sequencing Technologies & Bioinformatics Analysis November 16-20, 2020

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 hit the $1000 genome* in ~2016
  - 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?
- Get bigger hardware?
- Write more grants?

# Cloud computing providers

- Amazon AWS
  - https://aws.amazon.com/
- Google cloud
  - https://cloud.google.com/
- Microsoft Azure
  - https://azure.microsoft.com/en-us/
- More…

# 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

# Some of the challenges of cloud computing:

- Not cheap
- Getting files to and from there
- Standardization can be a challenge if you don't control hardware
- 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

# Key AWS concepts and terminology

- **AWS** - Amazon Web Services. A collection of cloud computing services provided by Amazon.
- **EC2** - Elastic Compute. An AWS service that allows you to configure and rent computers to meet your compute needs on an as needed basis.
- **EBS** - Elastic Block Storage. A data storage solution that allows you to rent disk storage and associate that storage with your compute resources. EBS volumes are generally backed by SSD devices.

# Key AWS concepts and terminology

- **S3** - Simple storage service. Cheaper than EBS and allows for storage of larger amounts of data with some drawbacks compared to EBS. S3 volumes store data as objects that are accessed by an API or command line interface or other application designed to work with S3. EBS volumes on the other hand can be mounted as if they were a local disk drive associated with the Instance.
- **SSD** - Solid state drive. A particular type of storage hardware that is generally faster and more expensive than traditional hard drives.

# What is difference between the 'Start', 'Stop', 'Reboot', and 'Terminate' (Instance States)?

- Start – turn on an EC2 instance that you have previously created
- Stop – turn off an EC2 instance that you have previously created
- Reboot – restart an EC2 instance
- Terminate – permanently stop and destroy an EC2 instance. Any associated EBS volumes may also be destroyed at this time depending on configuration

# What is an AMI/snapshot?

- AMI (Amazon Machine Image) – a template that specifies how to launch EC2 instances
  - Root volume with operating system (OS), pre-installed applications, etc
  - Launch permissions the determine who can use the AMI
  - Specification of (data) volumes to attach when launched
- You can create an AMI for any instance you have created/configured
- AMI can be made public for sharing (region-specific)
- Creating an AMI involves creating a snapshot of the root and any attached volumes. You will be charged to store this snapshot.

# I can not log into my EC2 instance, what might have gone wrong?

- Is your instance running?
- Are you providing the correct path to your key file?
- Is it the correct key file?
- Have you set the permissions for your key file correctly?
- Did you specify a valid user for your AMI (e.g., ubuntu)?
- Did you specify the correct IP address?
- Does the Security Group for the instance allow access for your connection protocol (e.g., SSH) and location?

# How much does it cost to use AWS EC2 resources?

Linux | RHEL | SLES | Windows | Windows with SQL Standard | Windows with SQL Web | Windows with SQL Enterprise

Region: US West (Oregon)

| | vCPU | ECU | Memory (GiB) | Instance Storage (GB) | Linux/UNIX Usage |
|---|---|---|---|---|---|
| **General Purpose - Current Generation** | | | | | |
| t2.nano | 1 | Variable | 0.5 | EBS Only | $0.0058 per Hour |
| t2.micro | 1 | Variable | 1 | EBS Only | $0.0116 per Hour |
| t2.small | 1 | Variable | 2 | EBS Only | $0.023 per Hour |
| t2.medium | 2 | Variable | 4 | EBS Only | $0.0464 per Hour |
| t2.large | 2 | Variable | 8 | EBS Only | $0.0928 per Hour |
| t2.xlarge | 4 | Variable | 16 | EBS Only | $0.1856 per Hour |
| t2.2xlarge | 8 | Variable | 32 | EBS Only | $0.3712 per Hour |
| m4.large | 2 | 6.5 | 8 | EBS Only | $0.1 per Hour |
| m4.xlarge | 4 | 13 | 16 | EBS Only | $0.2 per Hour |
| m4.2xlarge | 8 | 26 | 32 | EBS Only | $0.4 per Hour |

Data transfer (GB): In: free or $0.01; Out: free, $0.01 or $0.02
EBS storage (GB/Month): $0.10
S3 storage (GB/Month): $0.023 standard, $0.0125 infrequent access, or $0.004 glacier

# Why am I still getting a monthly bill?

- Generally you get an accounting of usage and cost on a 30 day cycle
  - Pricing is per instance-hour (now instance-second!) consumed for each instance type.
  - Also charges for storage, transfers, etc
- Be aware of regions!
- Even when an instance is stopped, storage for root or other EBS volumes persist
- Creating AMIs/snapshots requires storage
- Explore the billing and cost management tools of AWS to track your spending, set warnings, etc

# Amazon AWS documentation

https://rnabio.org/module-00-setup/0000/06/01/Intro_to_AWS/

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

# 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 a web server
- We brought up an Ubuntu (Linux) instance, and loaded a whole bunch of software for NGS analysis.
- 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.

# Logging into Amazon AWS

# Go to course wiki, “Log into AWS” page


## Log into AWS

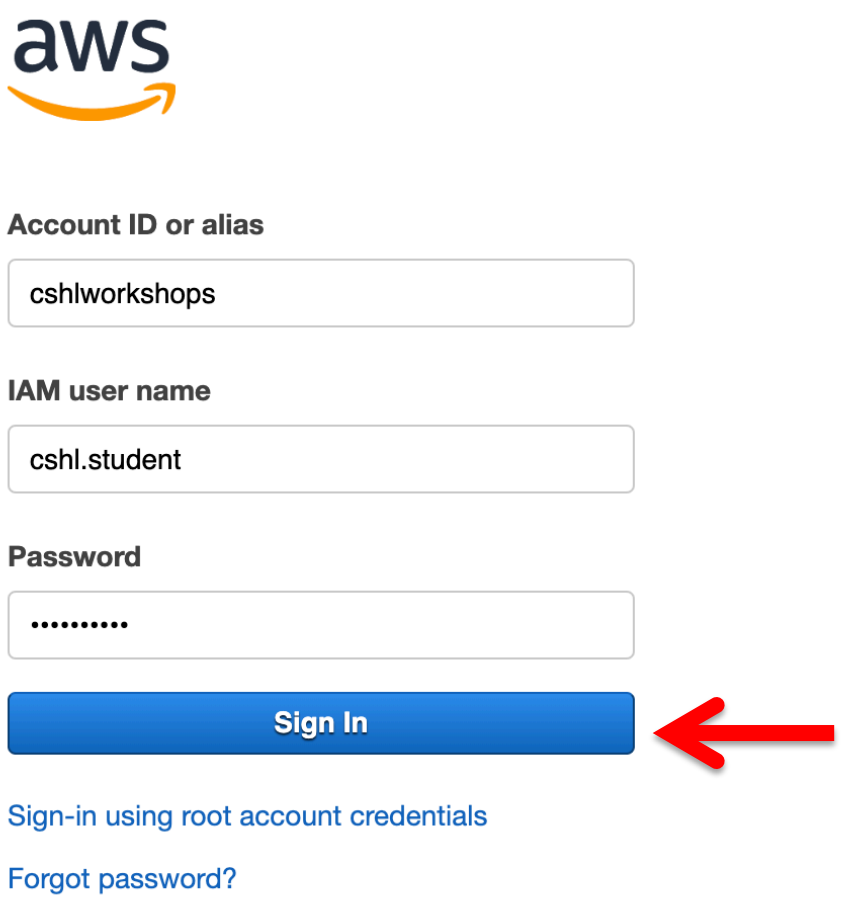
« Introduction to AWS  Course  Unix »

Using cloud computing to complete this course involves two major components: (1) Launching an instance on the cloud (essentially renting a virtual computer by the hour and turning it on) and (2) logging into that instance).

Covered in this section: logging into AWS EC2 console, starting an instance from the course AMI, configuring it in the console (select instance AMI, instance type, instance details, storage volumes, tags, security group, and key pairs).

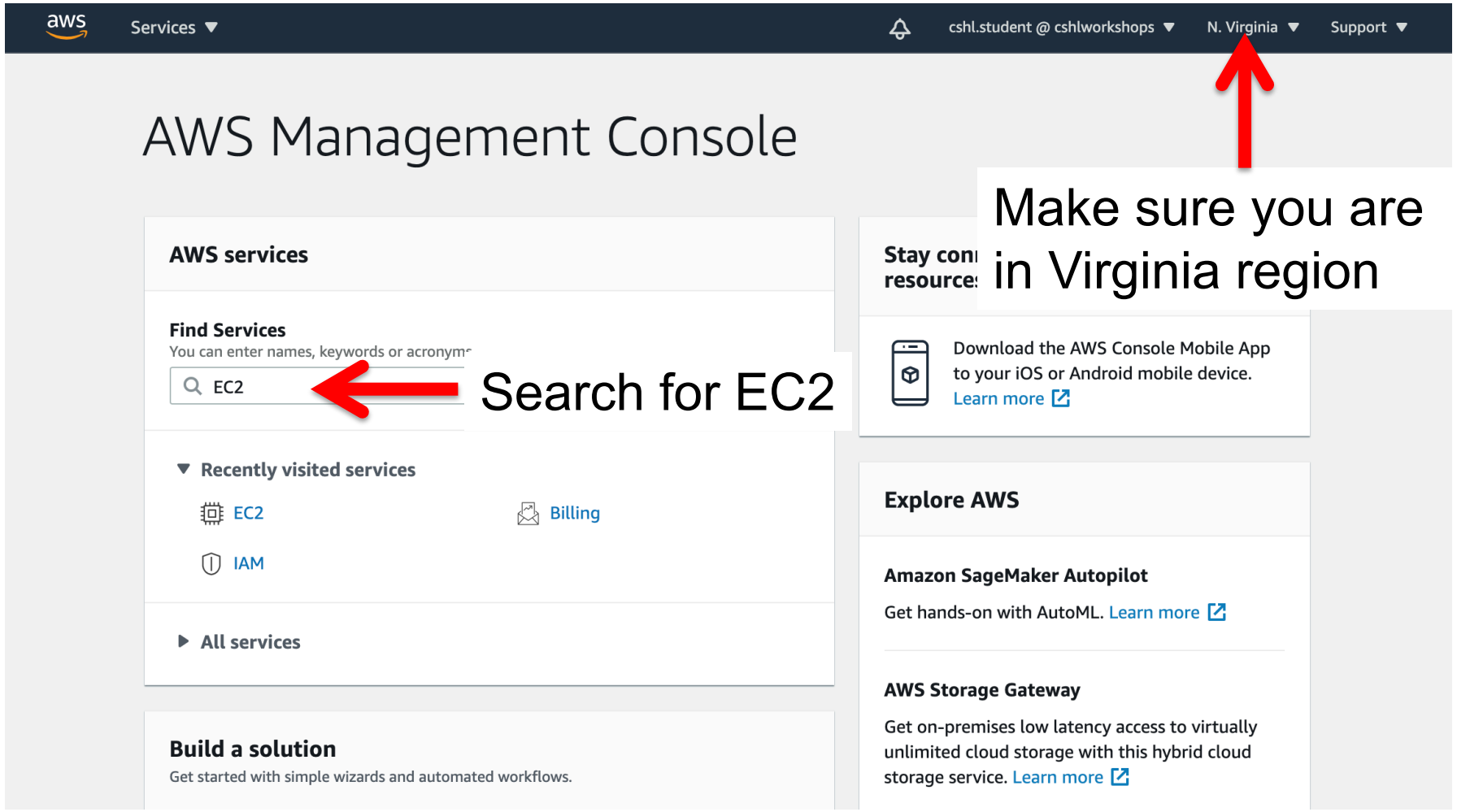
https://rnabio.org/module-00-setup/0000/07/01/Log_into_AWS/

# Login to AWS console

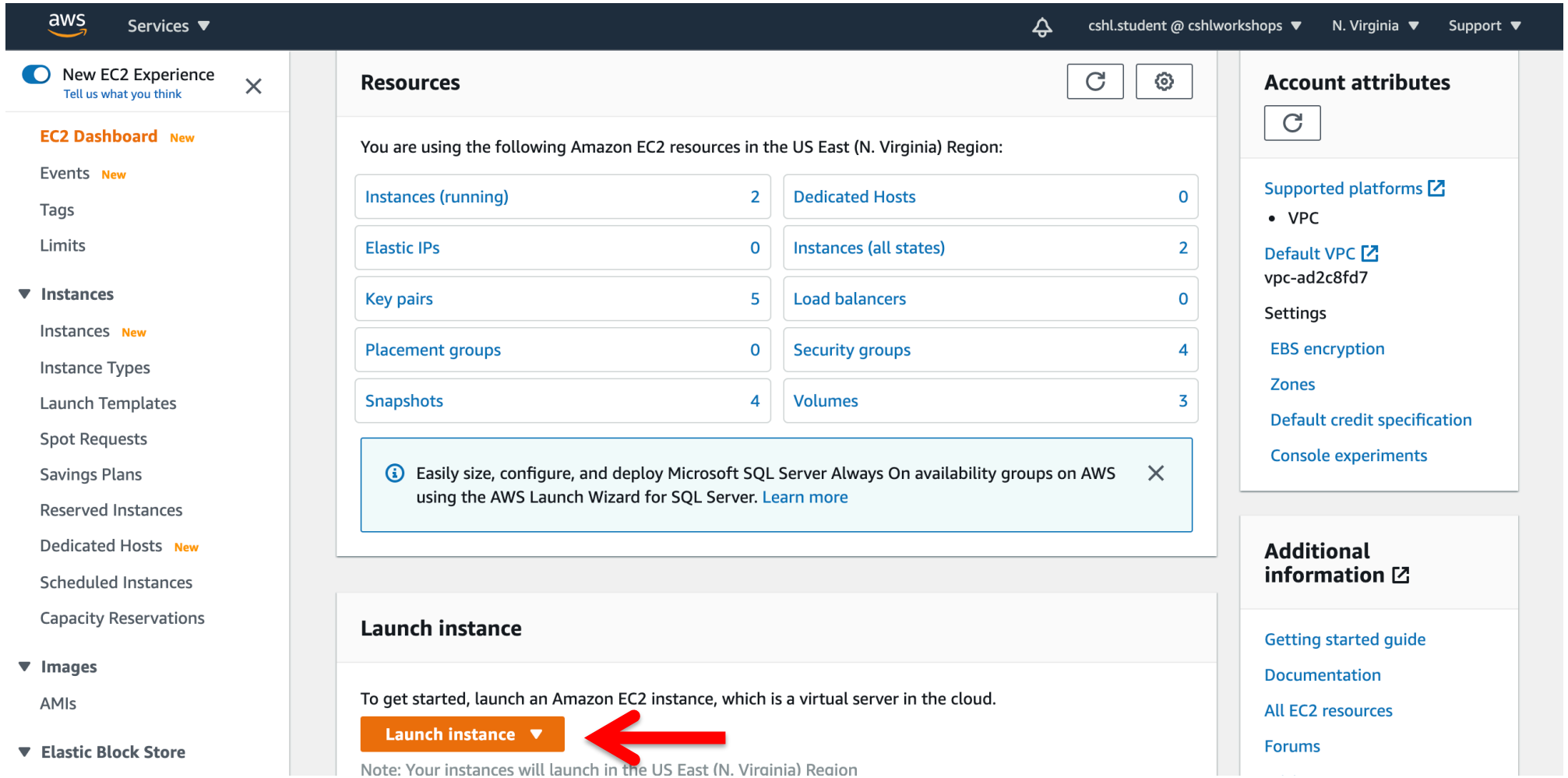



https://cshlworkshops.signin.aws.amazon.com/console

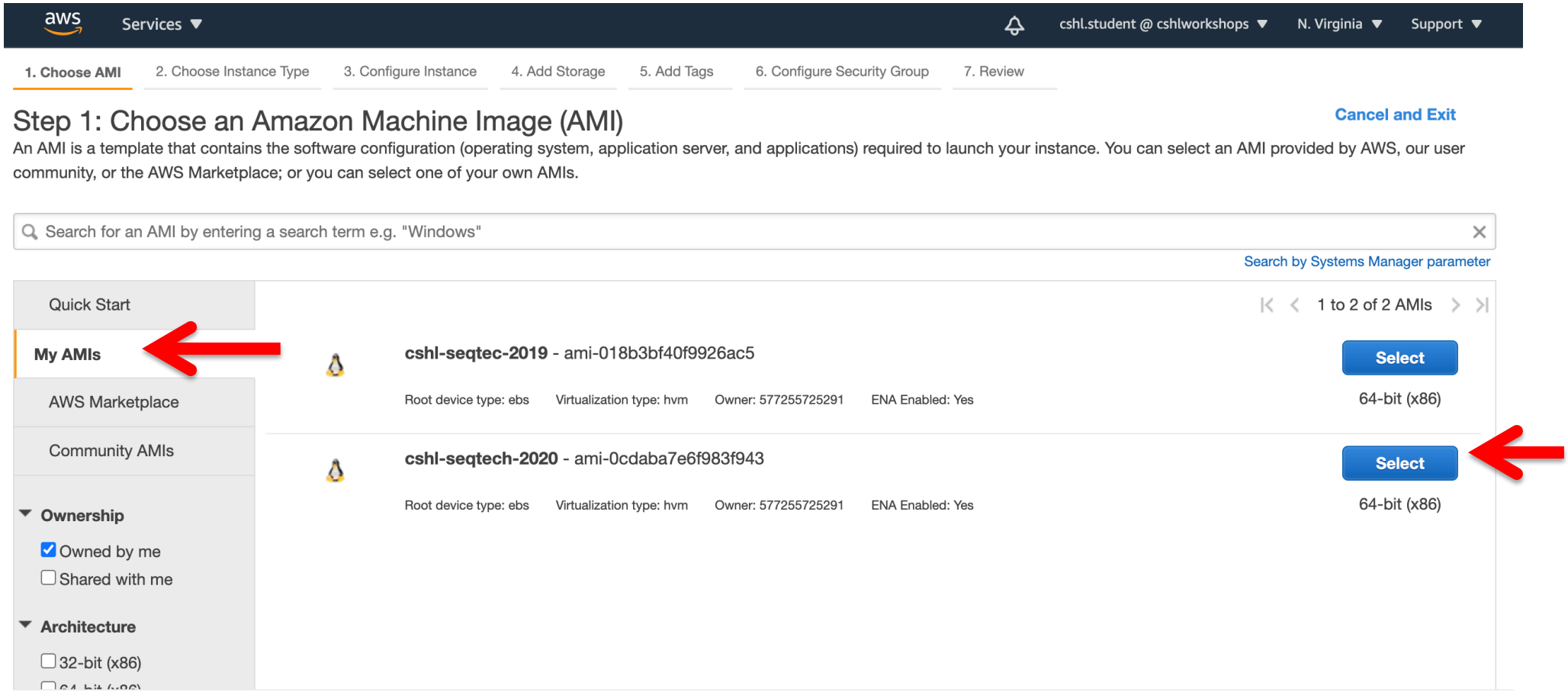
# Select "EC2" service



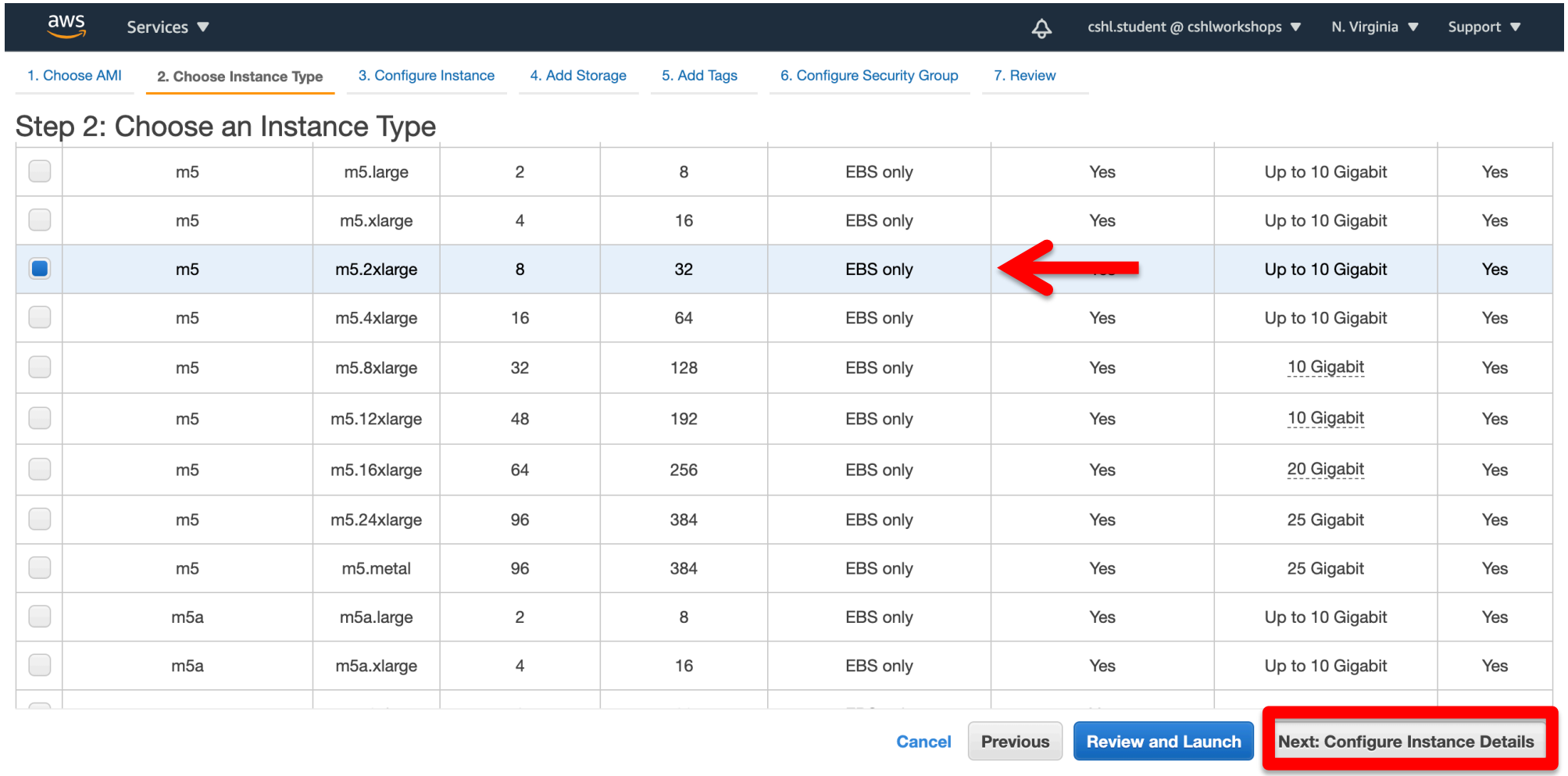
# Launch a new Instance



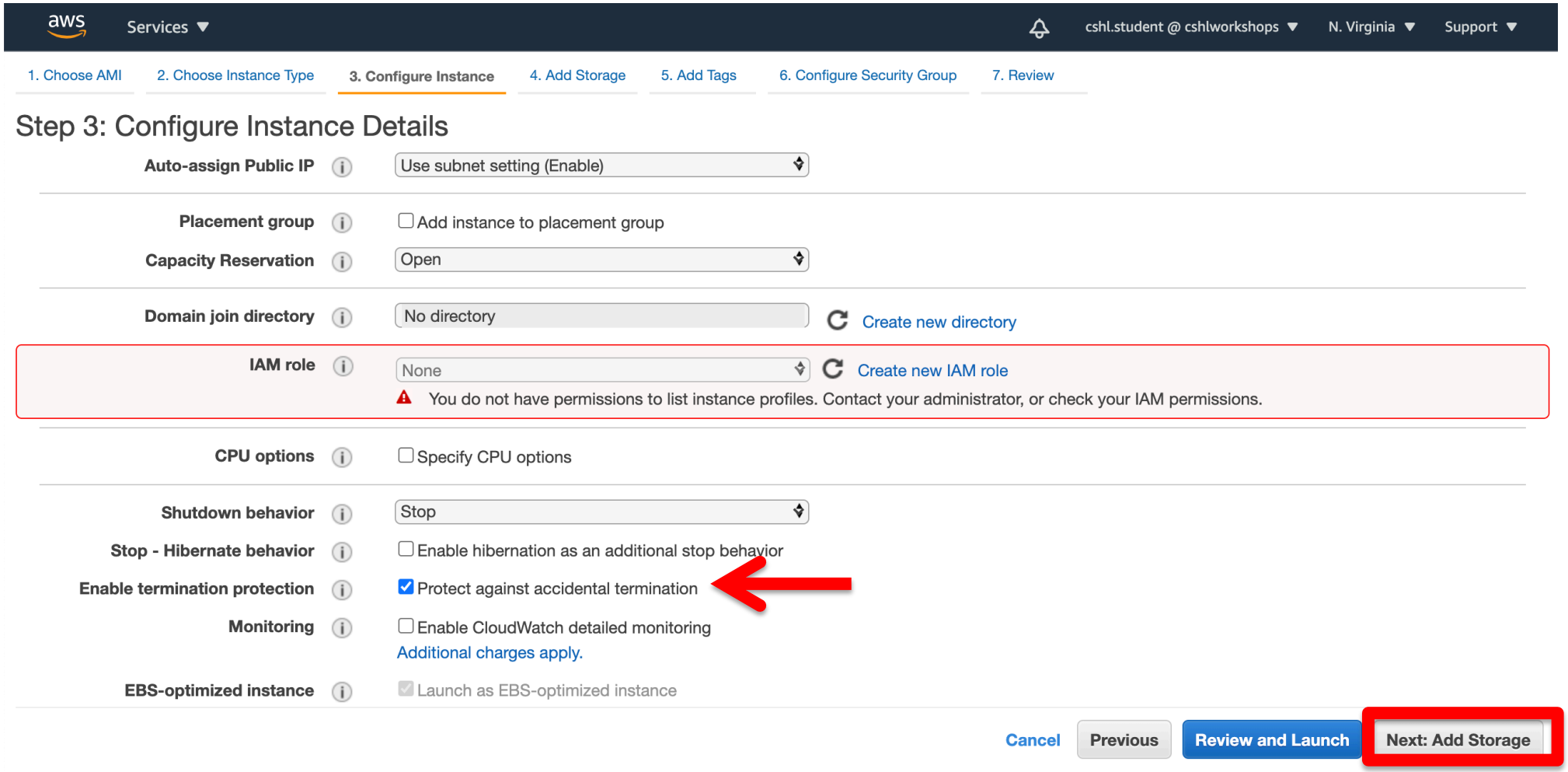
# Choose an AMI – Find the CSHL SEQTEC 2020 AMI in the My AMIs



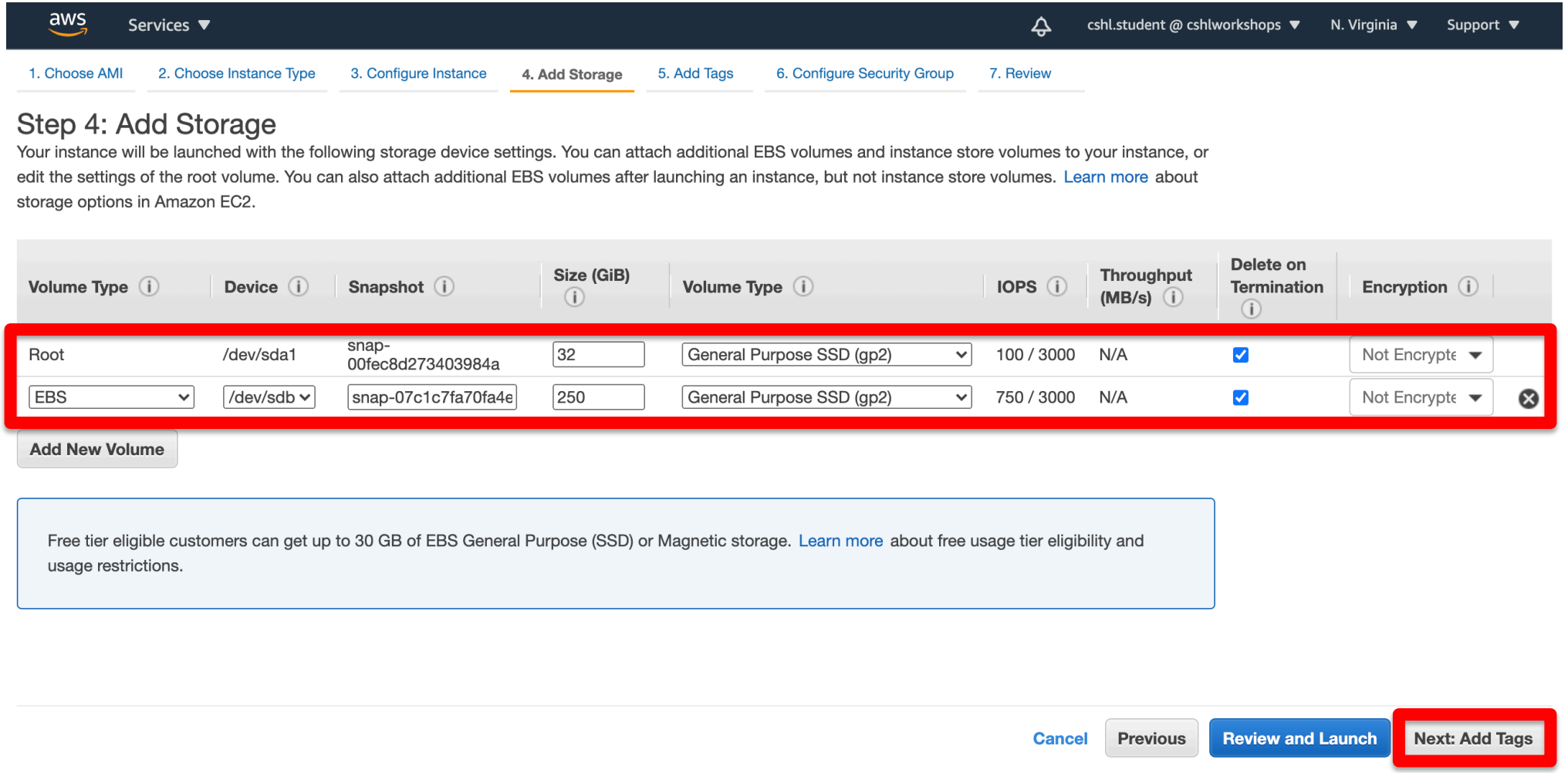
# Choose “m5.2xlarge” instance type, then "Next: Configure Instance Details".



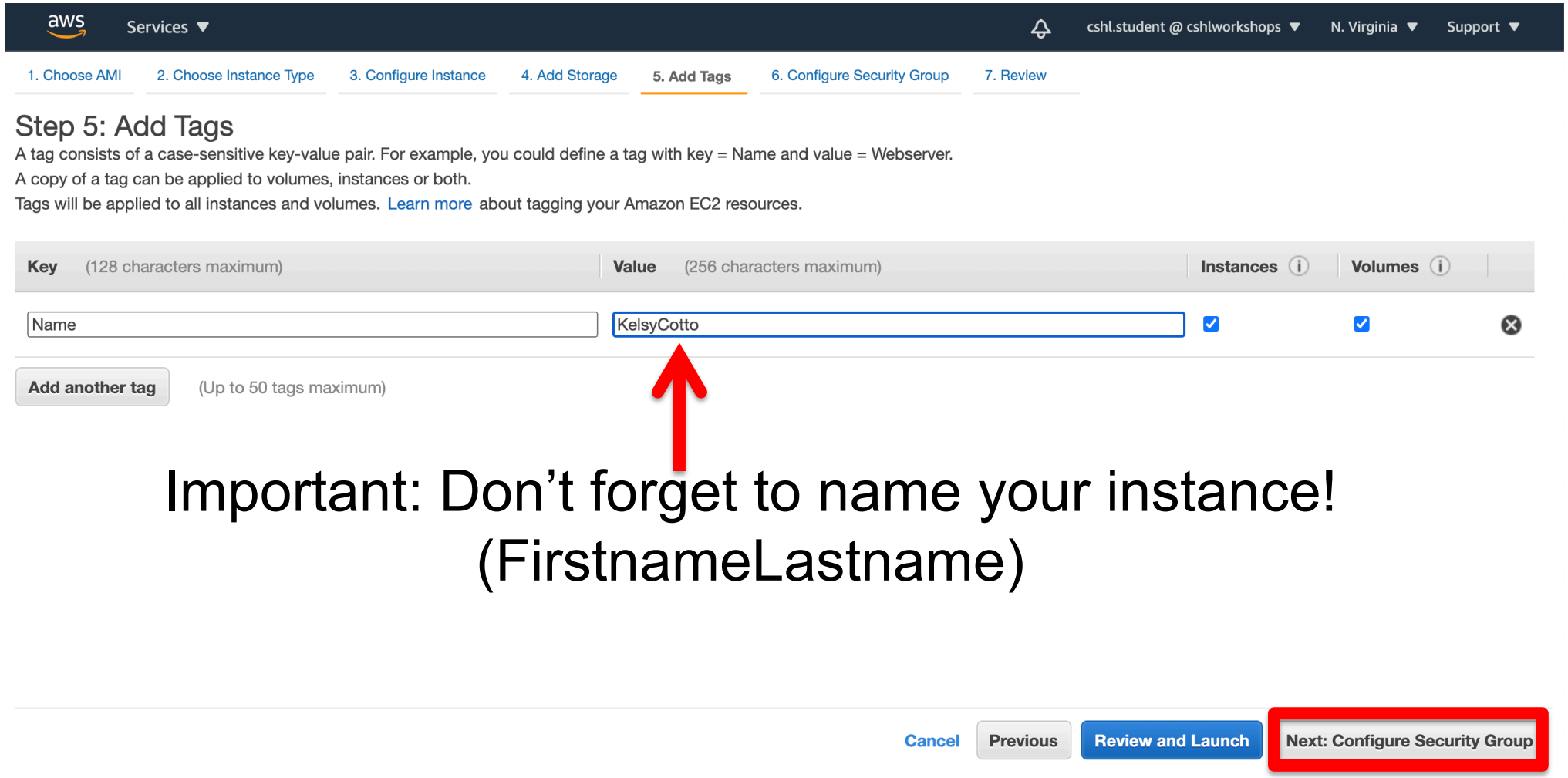
# Select "Protect against accidental termination", then "Next: Add Storage".



# You should see "snap-xxxxxxx" (32GB) and "snap-xxxxxxx" (250GB) as the two storage volumes selected. Then, "Next: Tag Instance"

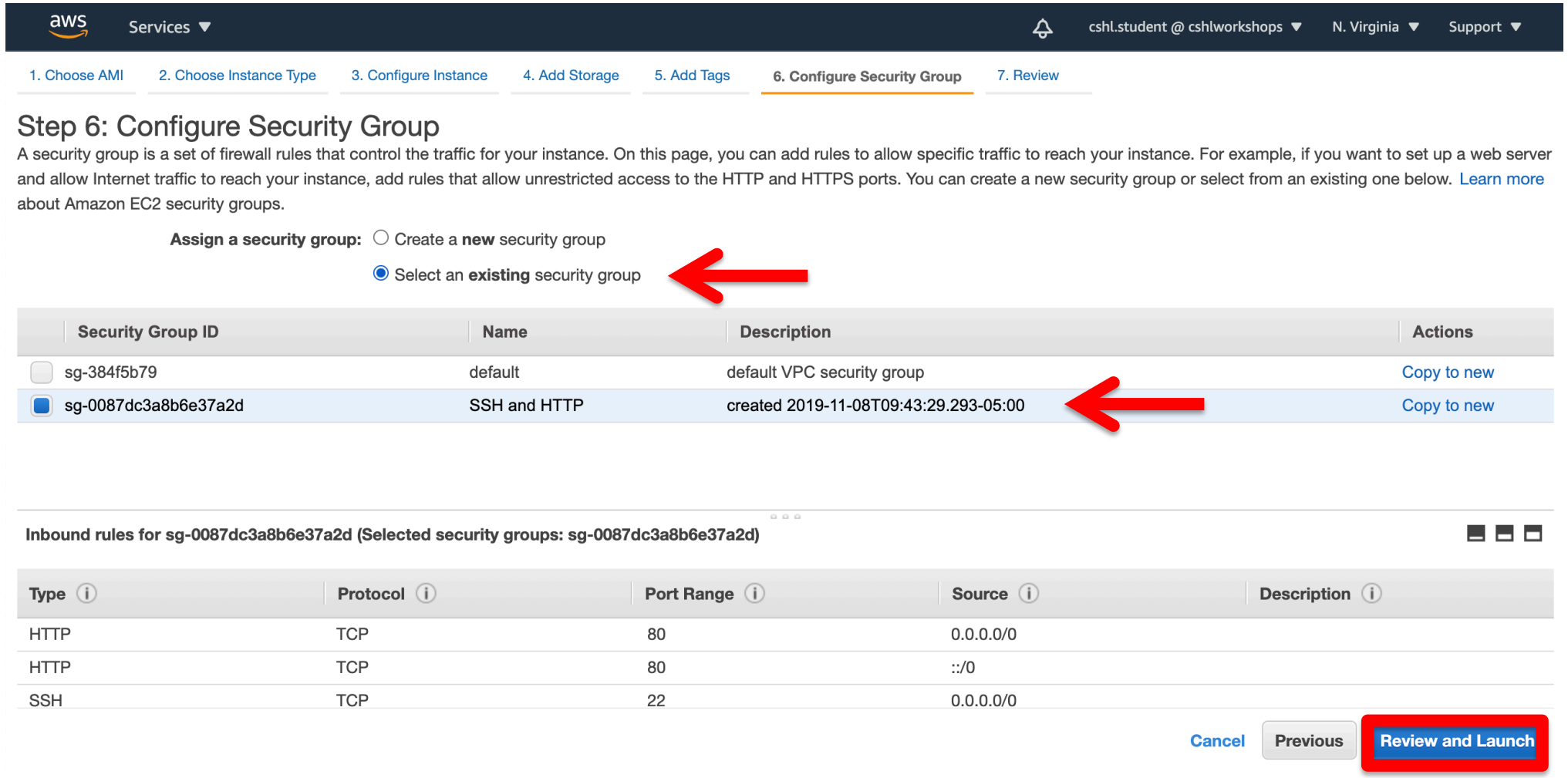


# Create a tag like “Name=KelsyCotto” [use your own name]. Then hit "Next: Configure Security Group".

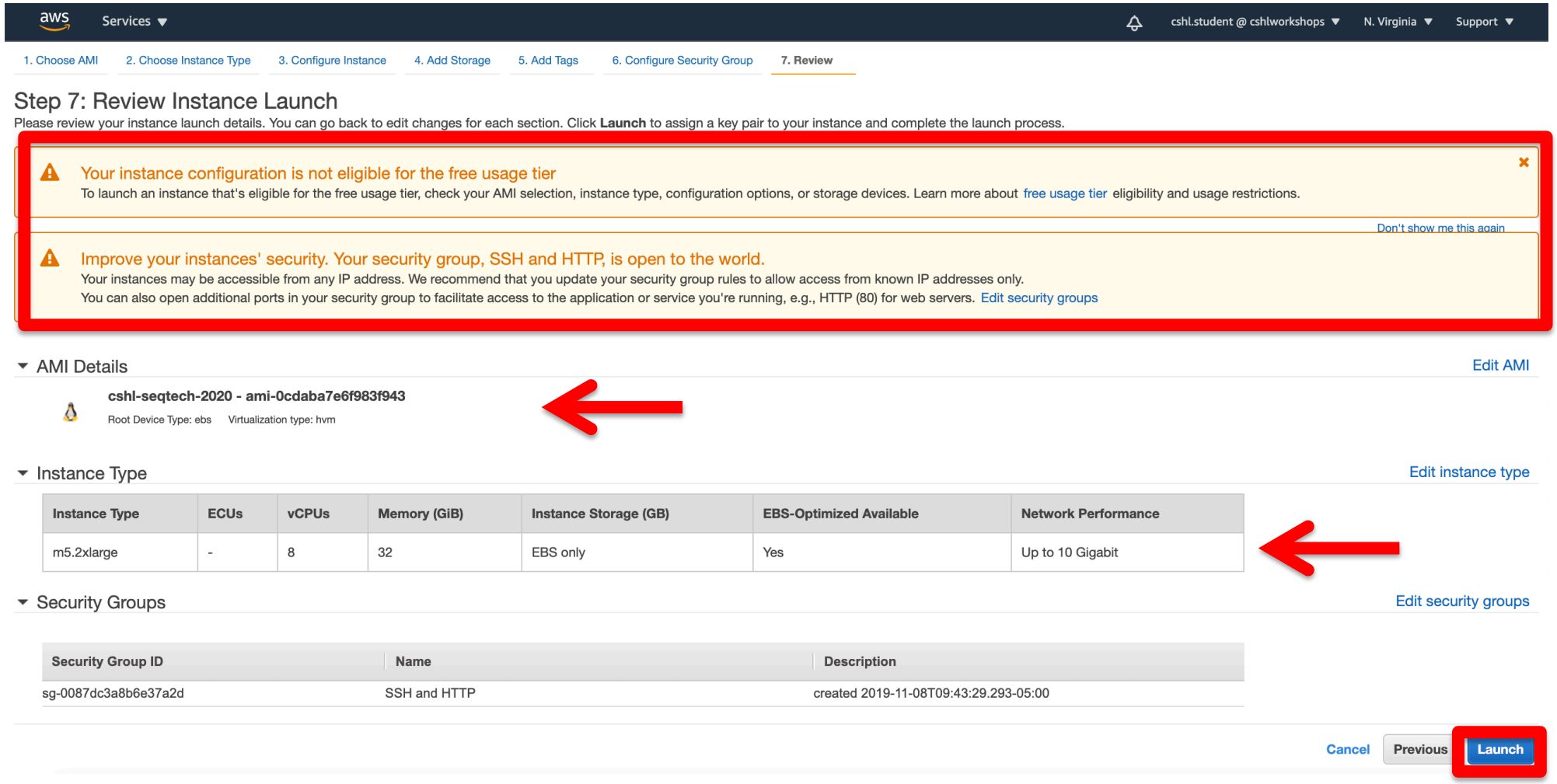



Important: Don’t forget to name your instance! (FirstnameLastname)

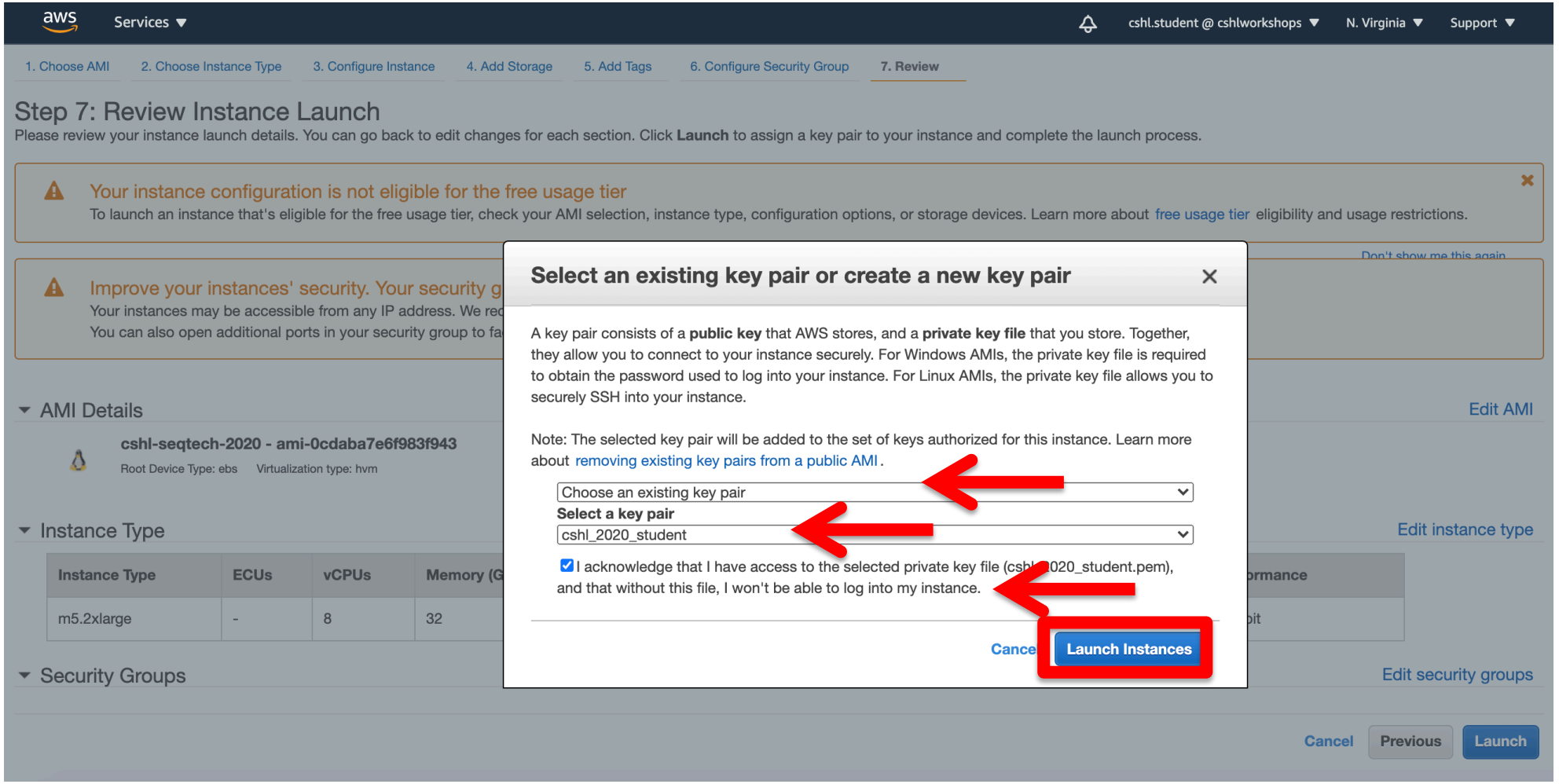
# Select an Existing Security Group, choose "SSH_HTTP". Then hit "Review and Launch".



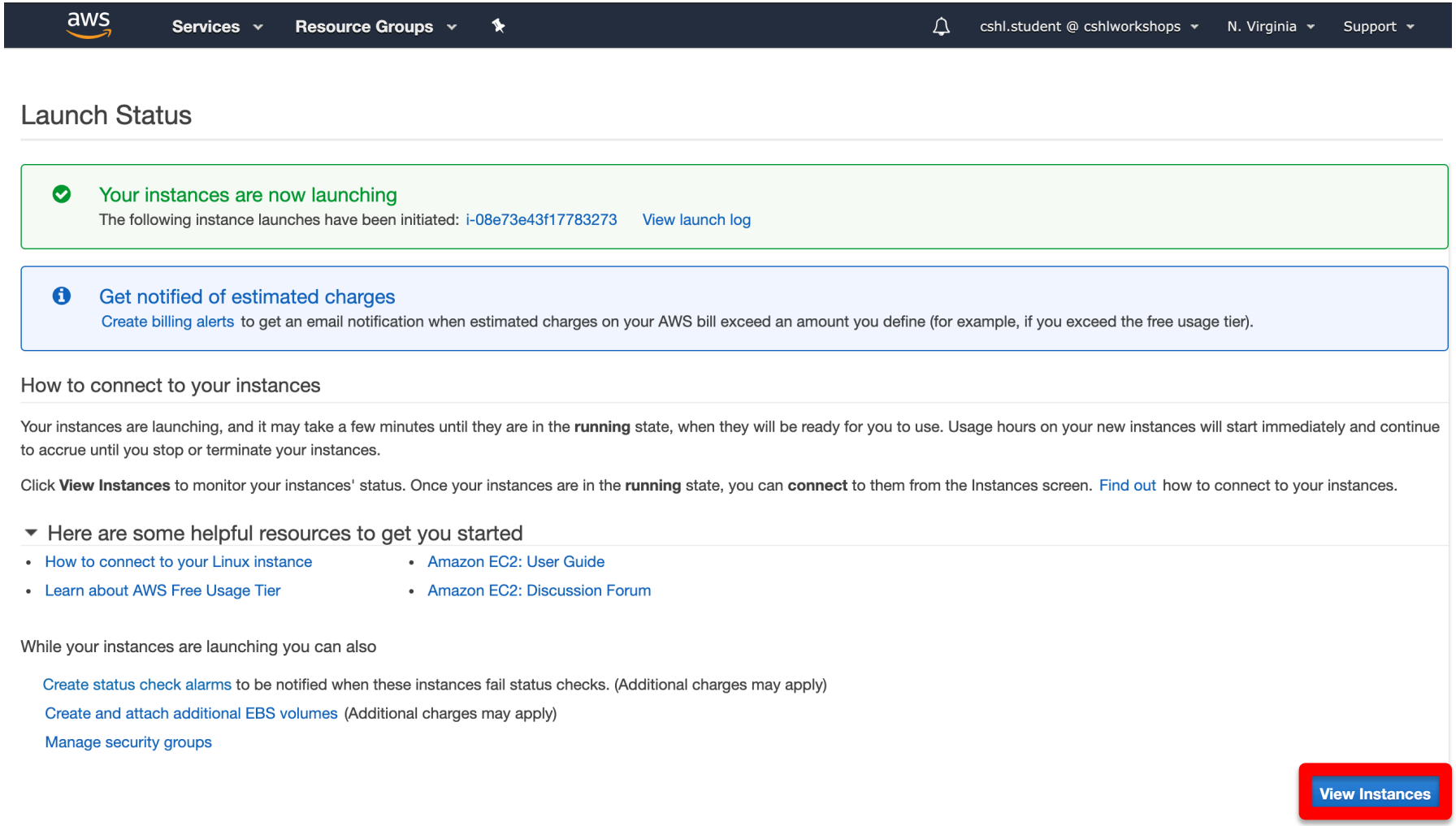
# Review the details of your instance, note the warnings, then hit Launch



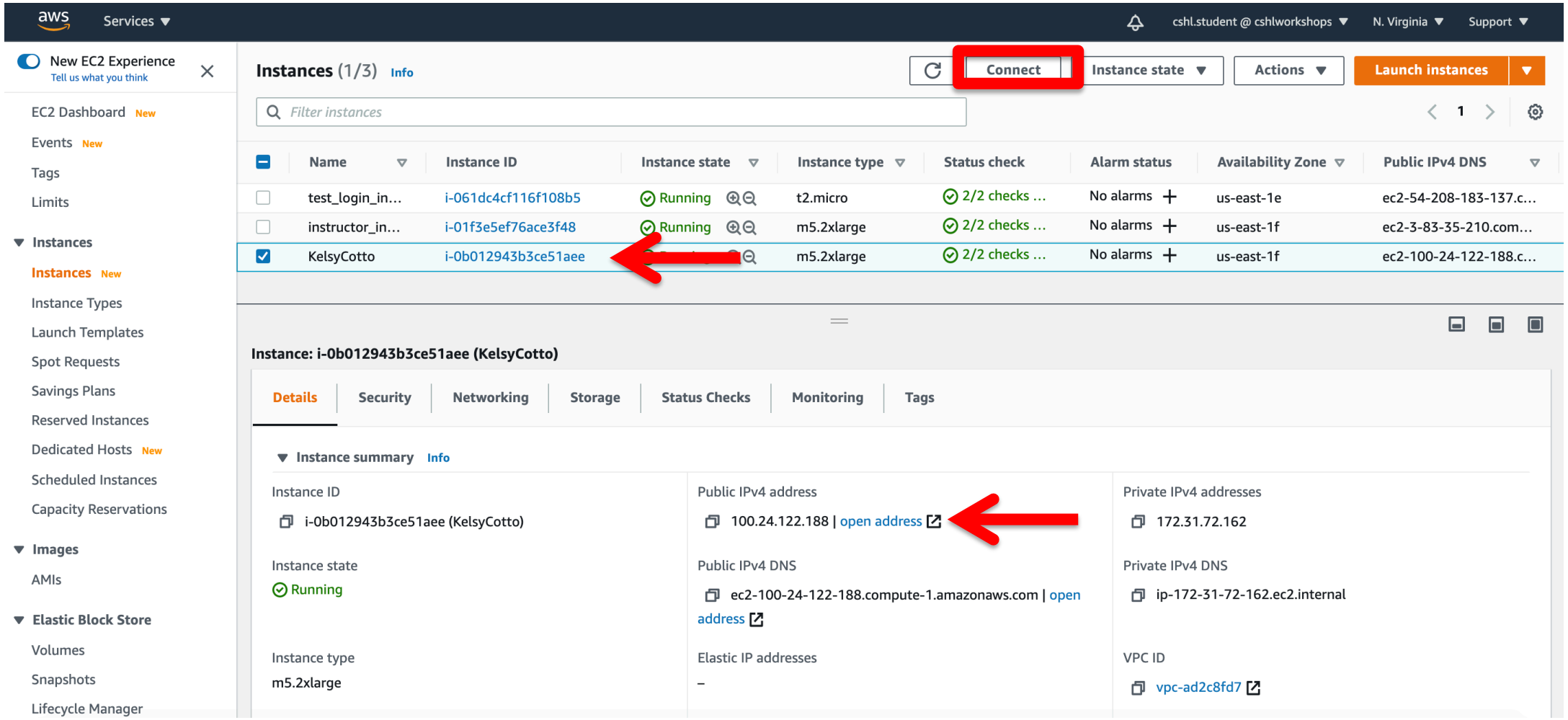
# Choose an existing key pair: ”cshl_2020_student" and then Launch.



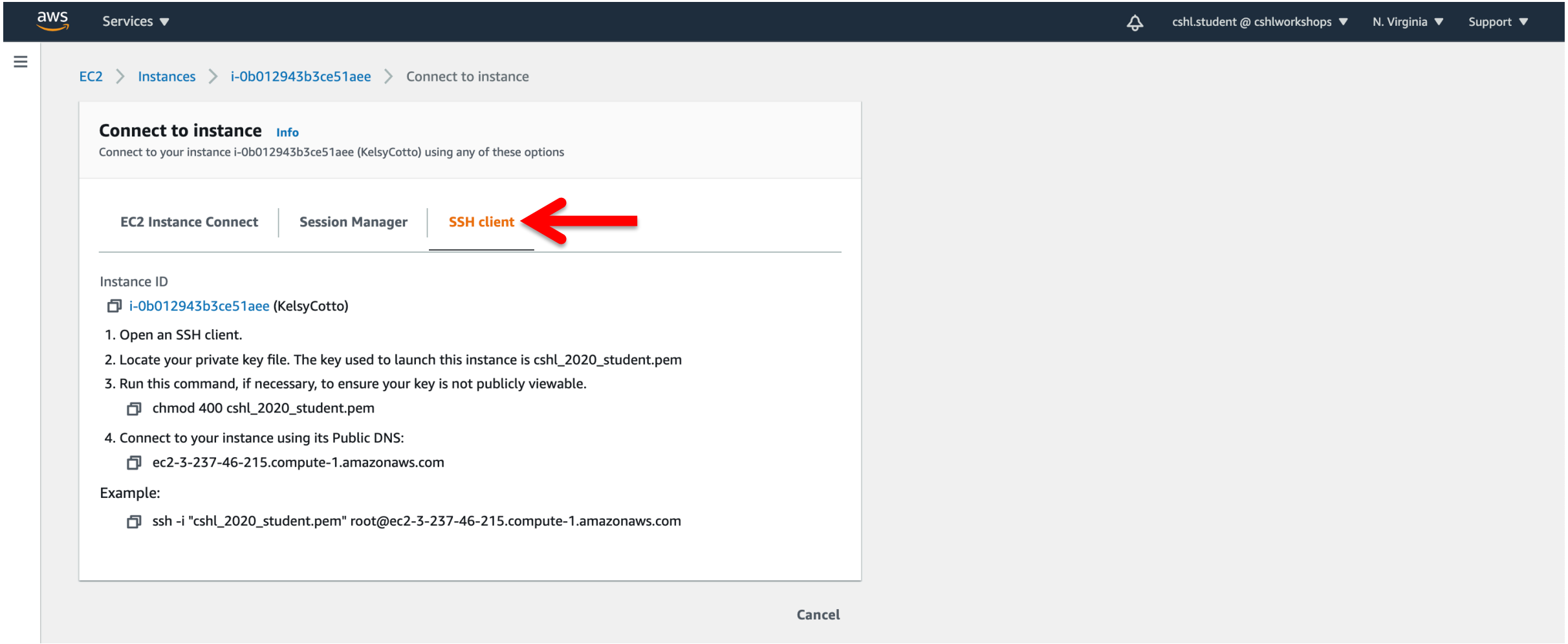
# View Instances to see your new instance spinning up!



# Find YOUR instance, select it, and then hit connect for instructions on how to connect (It may take some time for your instance to be ready)

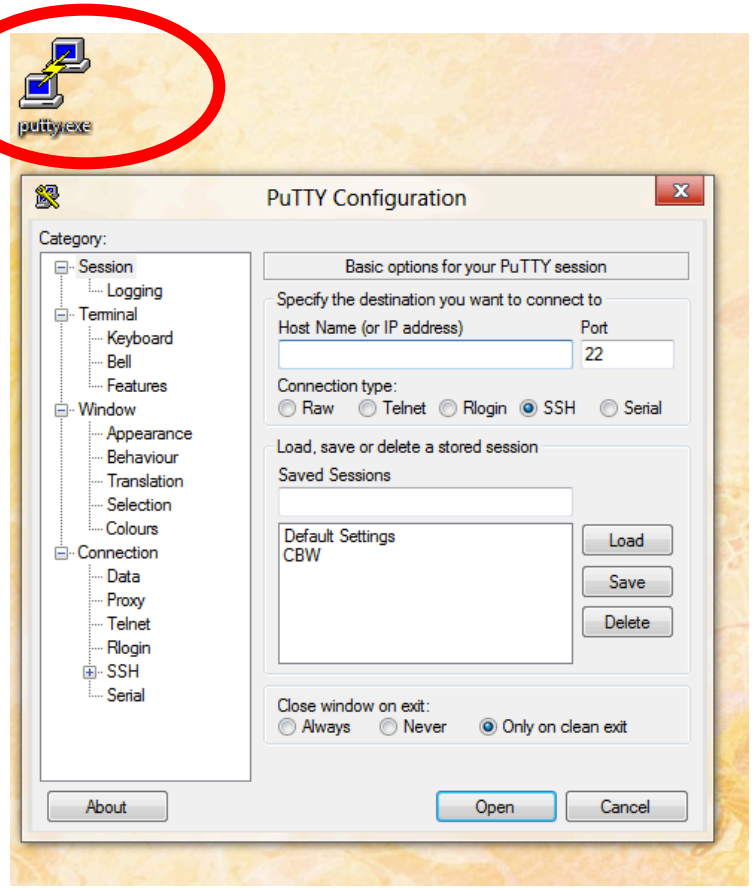


# Take note of your Public DNS/IP and the instructions on changing permissions for the key file (Note, we will login as ubuntu NOT root)

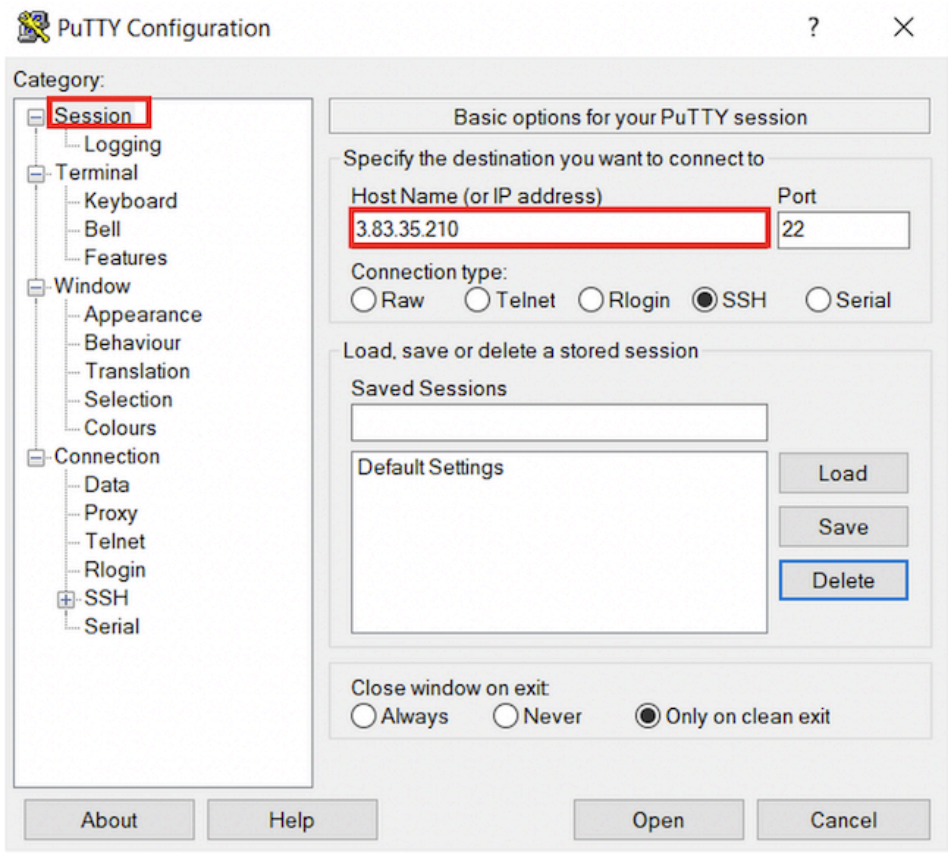


# Logging into your instance (Windows)

## Open PuTTY

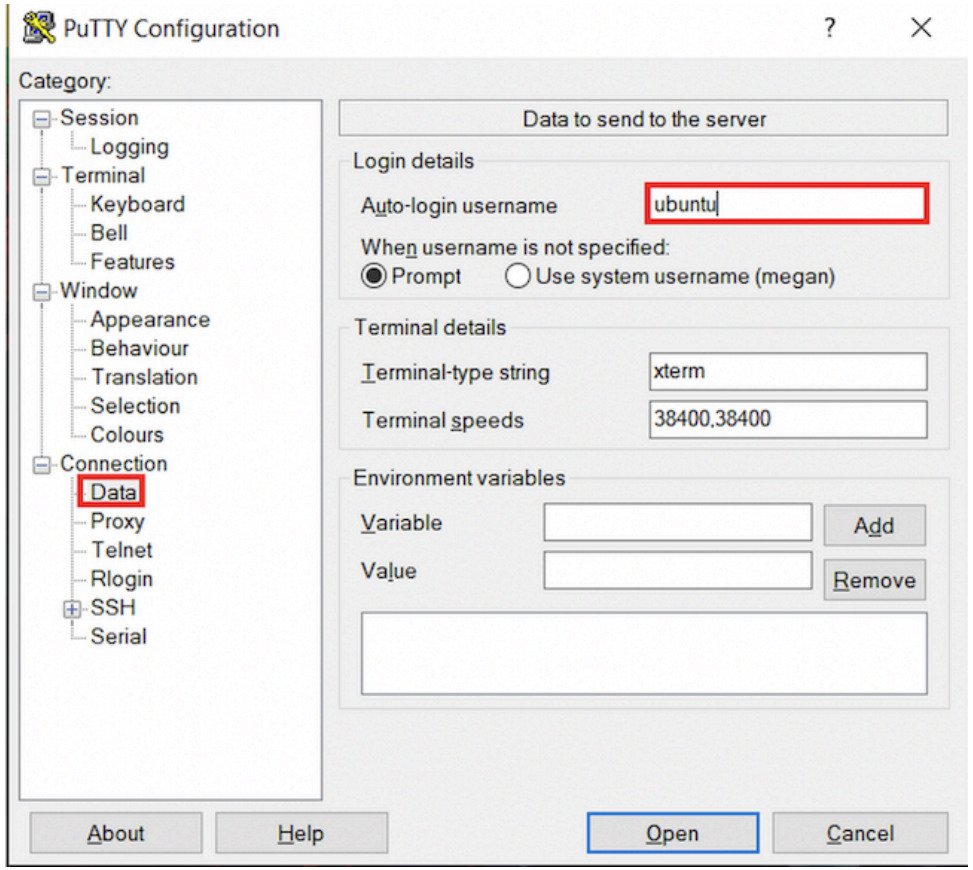



## Enter the Host Name (IP address)

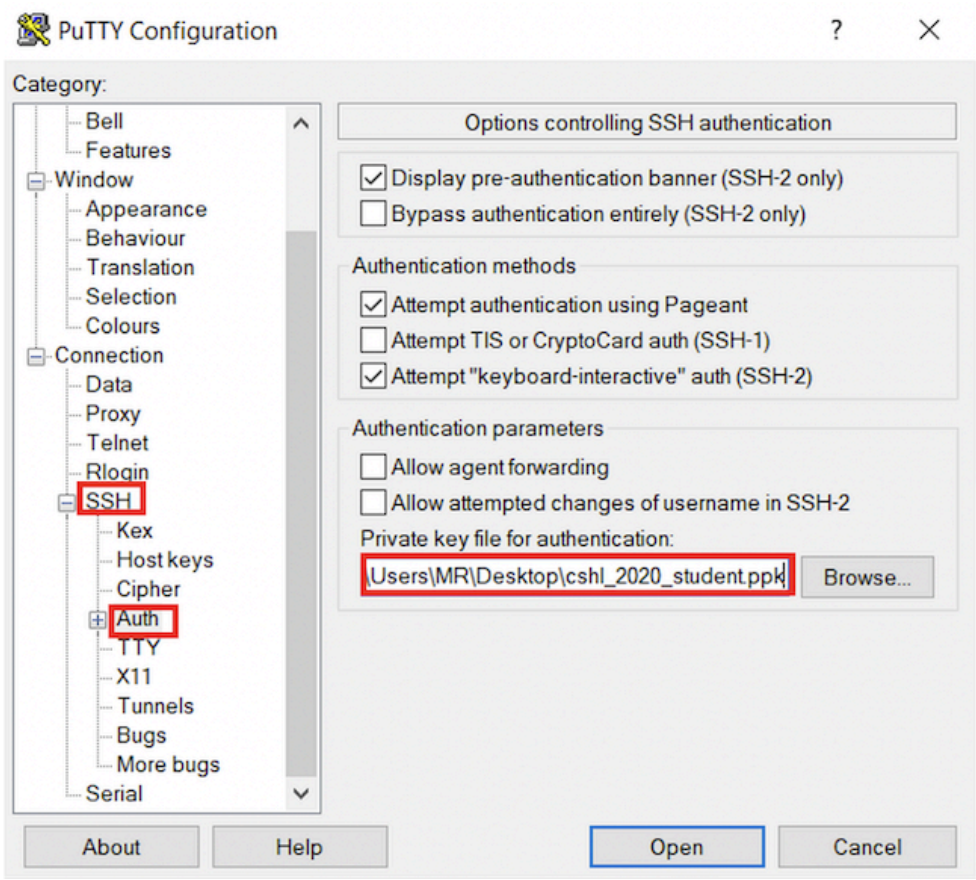


# Logging into your instance (Windows)

Choose Connection -> Data
Enter the username ‘ubuntu’

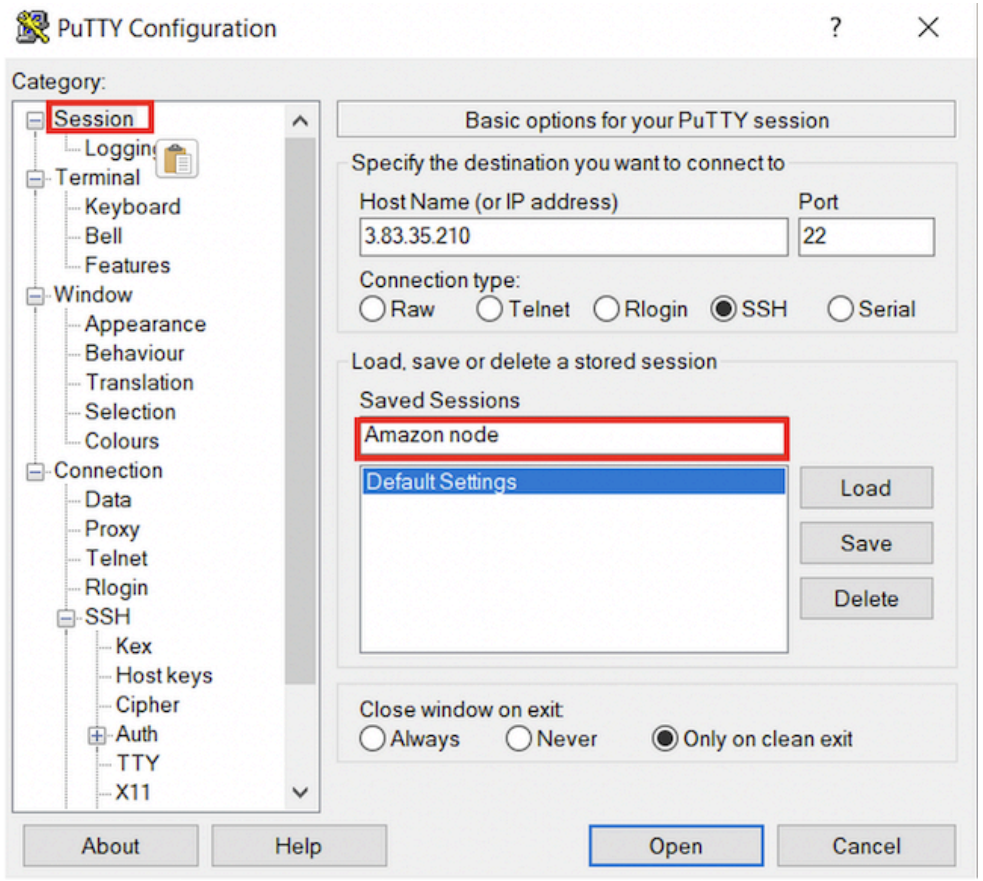



Choose SSH -> Auth
Browse to Private key (ppk) file

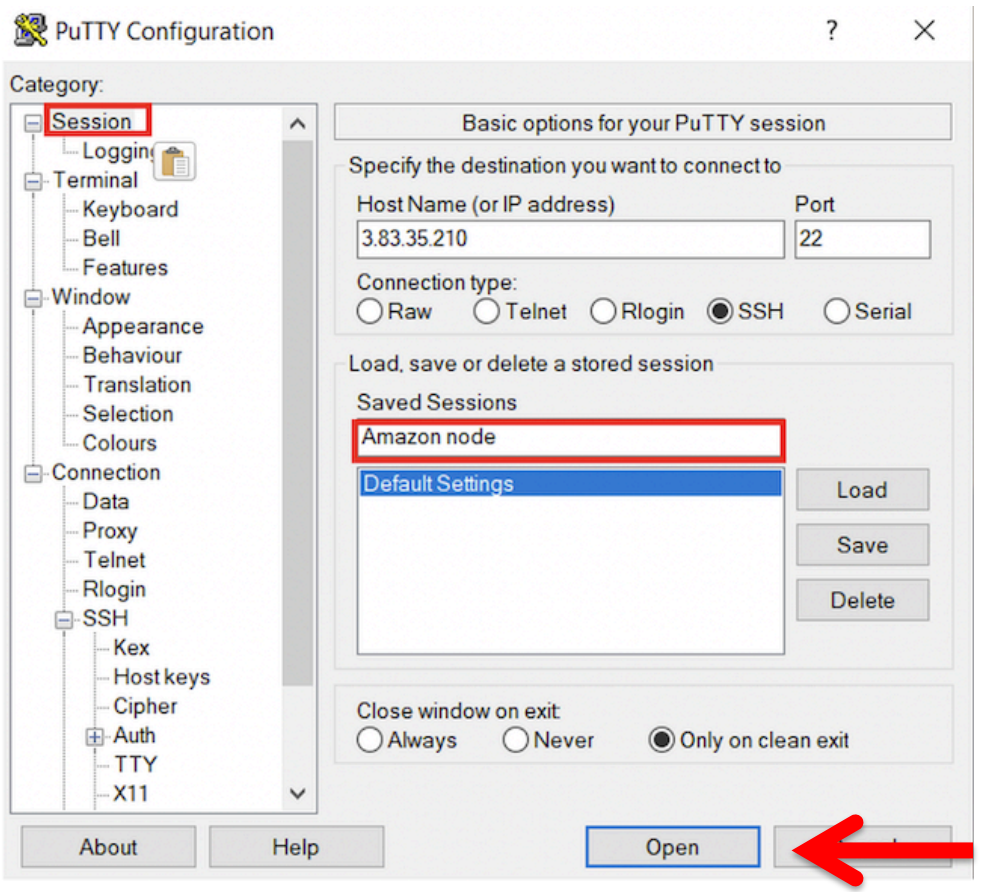


# Logging into your instance (Windows)

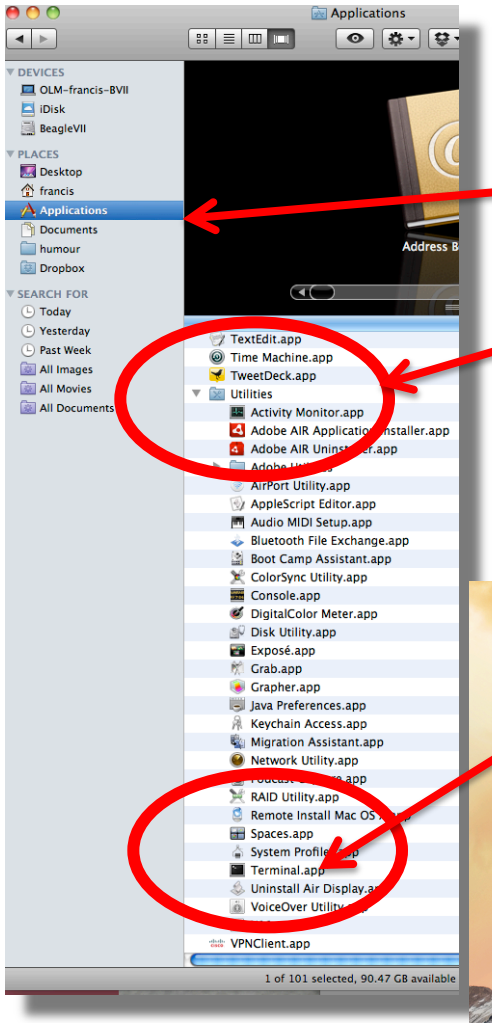
Choose Session
Save your session as ”Amazon Node”




Double-click saved “Amazon Node” session OR
Select ”Amazon Node” session and click Open



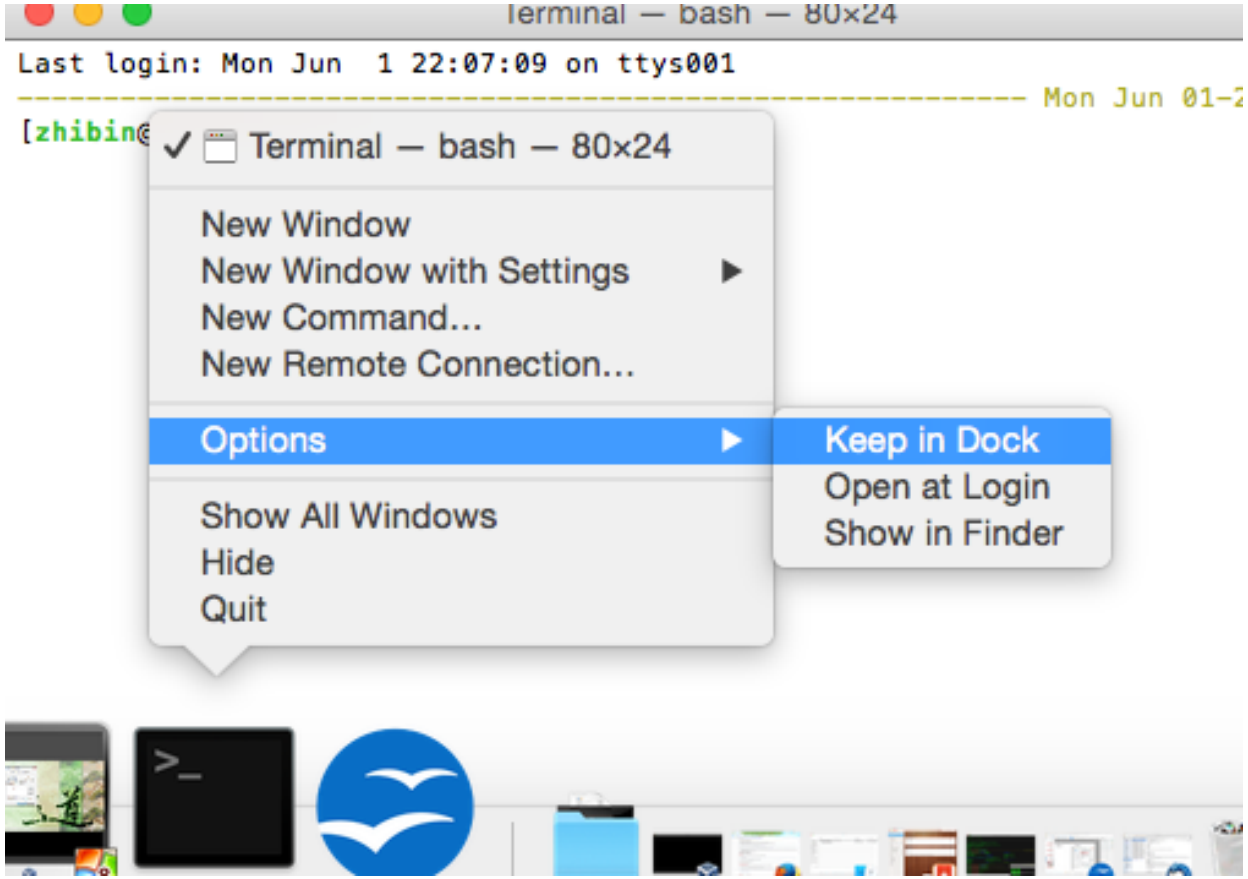
# Logging into your instance (Mac)




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 the course SSH key file

- NOTE for Mac users.  You will need to use a ".pem" file
- **NOTE for Windows Users**.  You will need to use a ".ppk" file instead.
  - This is created from the ".pem" file.
  - https://aws.amazon.com/premiumsupport/knowledge-center/convert-pem-file-into-ppk/
- The SSH key file will be used to securely login to your student instance on the cloud

**Save the pem/ppk file you received via email/slack to your new cshl folder**

# Viewing the ‘key’ file once downloaded

cat cshl_2020_student.pem

```
Kelsys-MacBook-Pro:keys kcotto$ cat cshl_2019_student.pem
-----BEGIN RSA PRIVATE KEY-----
MIIEpAIBAAKCAQEAgtGtpnqERtEu/SCmeF2r1HMESMaoOfEJiAwQwk2/SNXK8izr0IH0zTVvmE1D
VUwWq7pkvhhJh05pDb+2U9HiZe3sxLv3S1NrkATYF/NsrpwB+q1vwqzGW9sQ6uj45RWrPkiZlsaj
TQZmyFRu+tlJTRU3hQDqA0MRWTx1WxvOgFzuZy/qb+DALuFQsInrEKnijrwdLmd6usaBTvhc0gFS
B9oEelH0bZHJTZFw/wP+Z0uZq0Ujir7QwOLTM45QH/L0dBdUl3k/mBeezOOyvnKMwj8E4Xi0rQ0t
hHtQ7F9iSILK8OW1rRH0qwxwt9ycEH1JtNMQmUTifOvE2XJ6lO6chQIDAQABAoIBABG7P/FHu/Qp
WFgg+89myuqR6GvA2X55CFSzZfYgOaQyrj5jDleFtdu2uXiISG8gUBZYvlzxx82aOC0P5jO4SBqO
xD/qRlukY/jyXyPn77w/ExmaNoLJjl1W9RUSHOJYLIzVpFPGes3u5zGSGDTSDNhw3sSdWhq1FX3l
7vY5b6UAQgahXfOdpGFxt6P6qb/BKFQFsThXk7GXMyS/kr4w7ZlhPWHrMSu2UBdS/a/beAVN76qU
E/1OHR3OoAuqghusZabpbX7MJlO0VcviAQgeF8Z+xf2uugDEAKut1PxWO+yvGM4SpZZOskFZz6YI
pBnX4ELWPhCeMzq4Ml8QY6ptR2UCgYEAuc6YfNlymg24NJu8PANx8navTi5OWYuXWXrjOmrLr3SR
/XY+w26cgipM+K5eQfXSr4Yb8BQKjRktMzBzf5nKdaX4pzYIquQH02BOHDhTooHAhkbTMTmKukv4
oEW06wcEE6RQifw4xbMEnfQfHJBI21am+jwI8Xb7idwMG4pU/nsCgYEAtD0x4bNC1X3A5by5OuYO
WXrtgQszCXYcbkrpbjRET12f9hgz9MRMHY/xH/XGvMutZSFV2rCZRwd7lm+QWGadk/MQSOkouzW6
gSasyjFq+MKCkqYnS3/JTbx8yrLZmzl0LtX6pwmwgOZy8aJjYHo9a2/EI8Tjh2dOSxeadIRVYP8C
gYATOiXww1Vdu+dj/7TDLqYCtdHOVAxJX/utI9QQ3yoIryuh+bWmFvEIvAmIGXyyQZRyoZwgIS4A
PNHO3+bEa+69wbzlhksiK5g8GKgISVdlC4rZZXB5ehgTmWV7IgJ89y/SF4G/Ityo30K0ohALh597
NcvNEzzqruTja1IIMvTKMwKBgQCkM+QP1TqcOTbVlfvClviXuJBLsiJLClmYeZL0nZVmIMusbhxX
b8ZQYGSyUz09nulXaulGlQDvXvfO89CzWLlSomxBoHlFJQvGwa9FfYQRIVPHuqut8rs4oPGn0QzC
h7M7QCJcr00oAcrSLLkQmgz+phIw7BzFrO39J4HFiRInjQKBgQCtdEvcbtyk8Jh4WH3z0wpkc43f
U8DZhZwjRQpGWlD8CPj9RgRnE4+1PCH6s/RLQf7SiE1ZjX/0UdOWPEvrOj5sVjy0IujohRbty0CM
oqWeSeUb1sLogRvMrTfCEpl/rz3GpoQlSC/5s6XvjnnKK8RN8s7MseLuuJ63T/CRBpIs8A==
-----END RSA PRIVATE KEY-----Kelsys-MacBook-Pro:keys kcotto$
```

# Changing file permissions of your ‘key’ file (Mac/Linux)

ls -l (long listing)

-rw-r--r--@ 1 kcotto staff 1696 Nov 9 09:19 cshl_2020_student.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>**

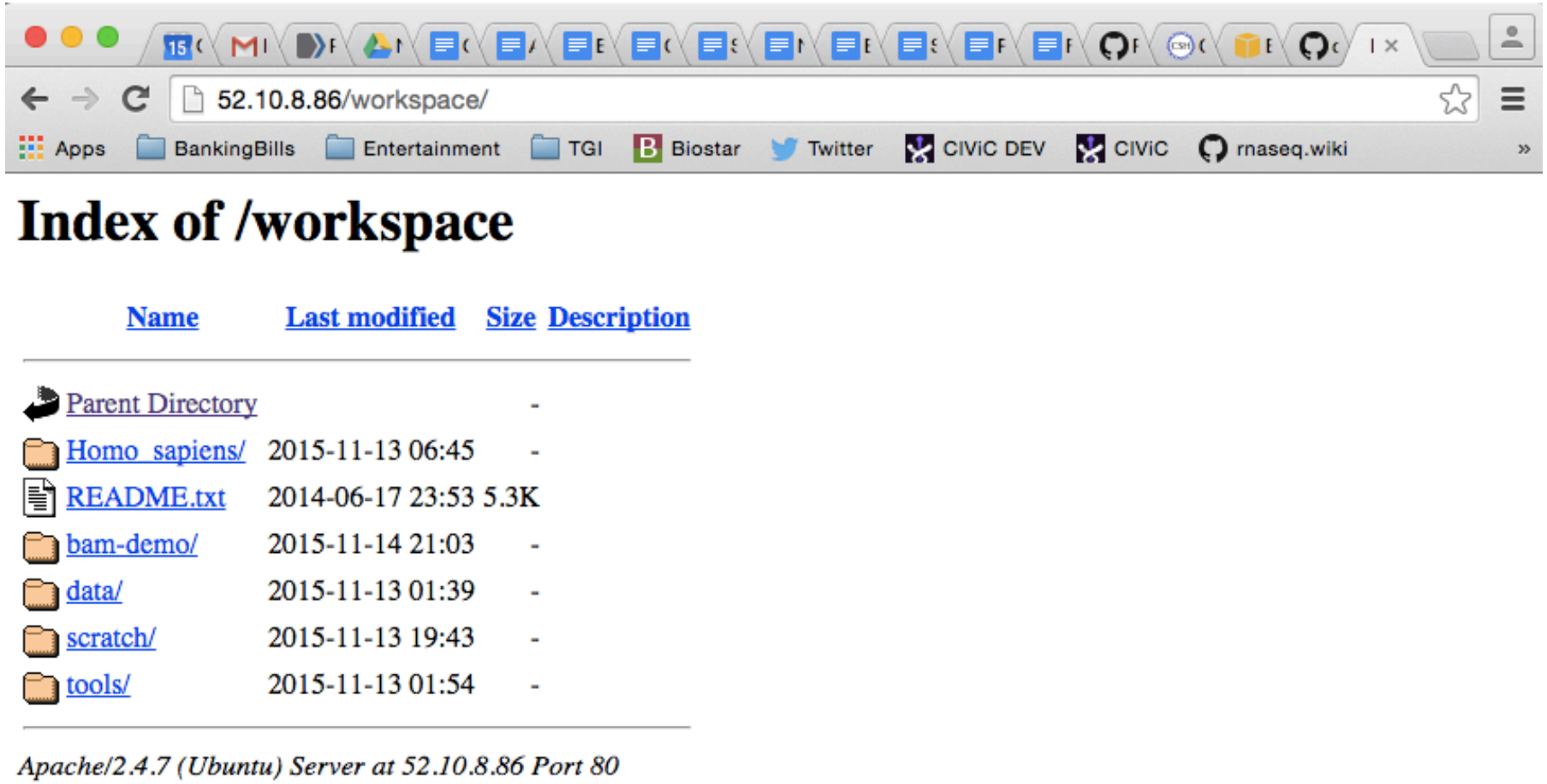
It is “r” for the the file owner **only**

# Logging into your instance

**Mac/Linux**

```
cd ~/cshl
chmod 400 cshl_2020_student.pem
ssh -i cshl_2020_student.pem ubuntu@[YOUR PUBLIC IP]
```

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




http://[YOUR PUBLIC DNS OR IP]/

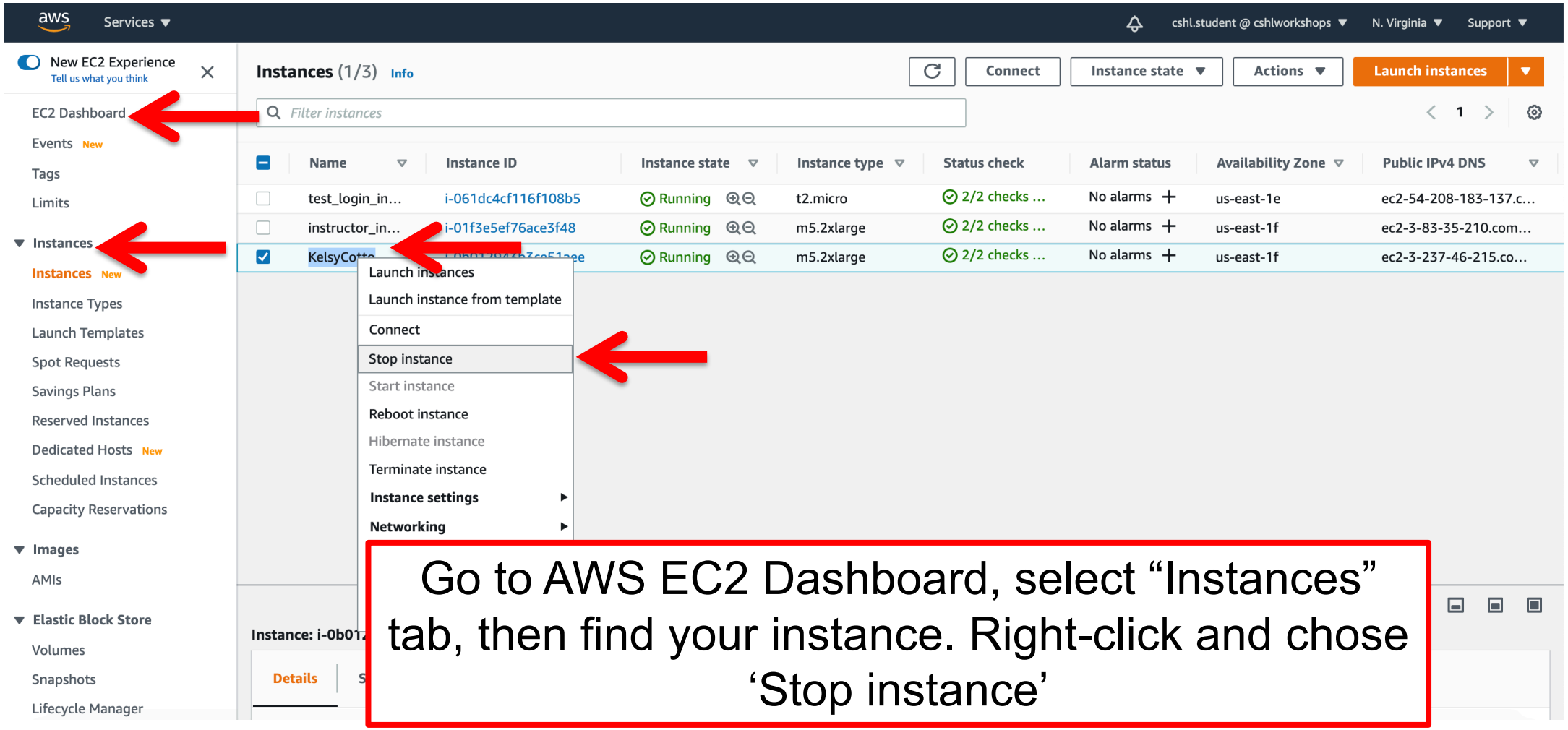
# Logging out of your instance

**Mac/Linux – simply type exit**

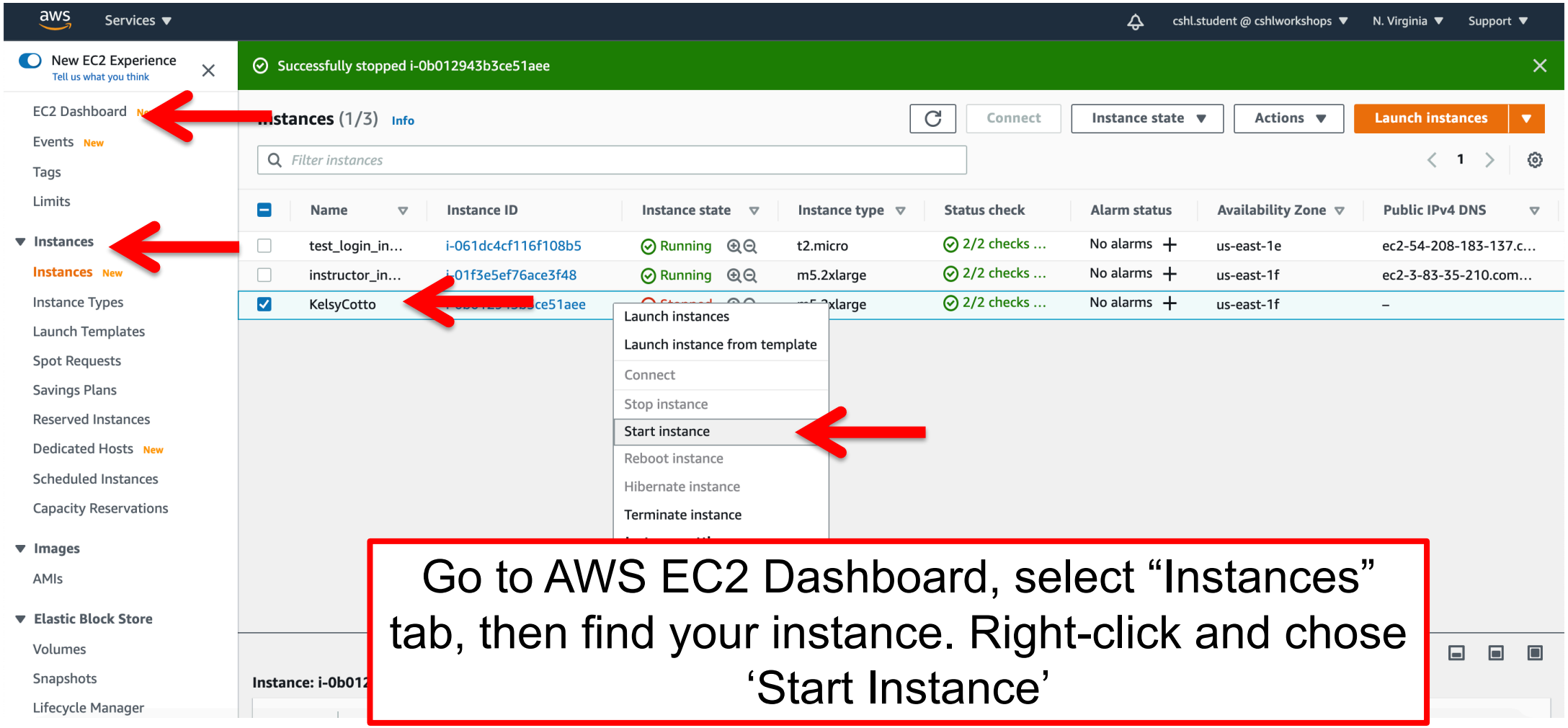
```
exit
```

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!

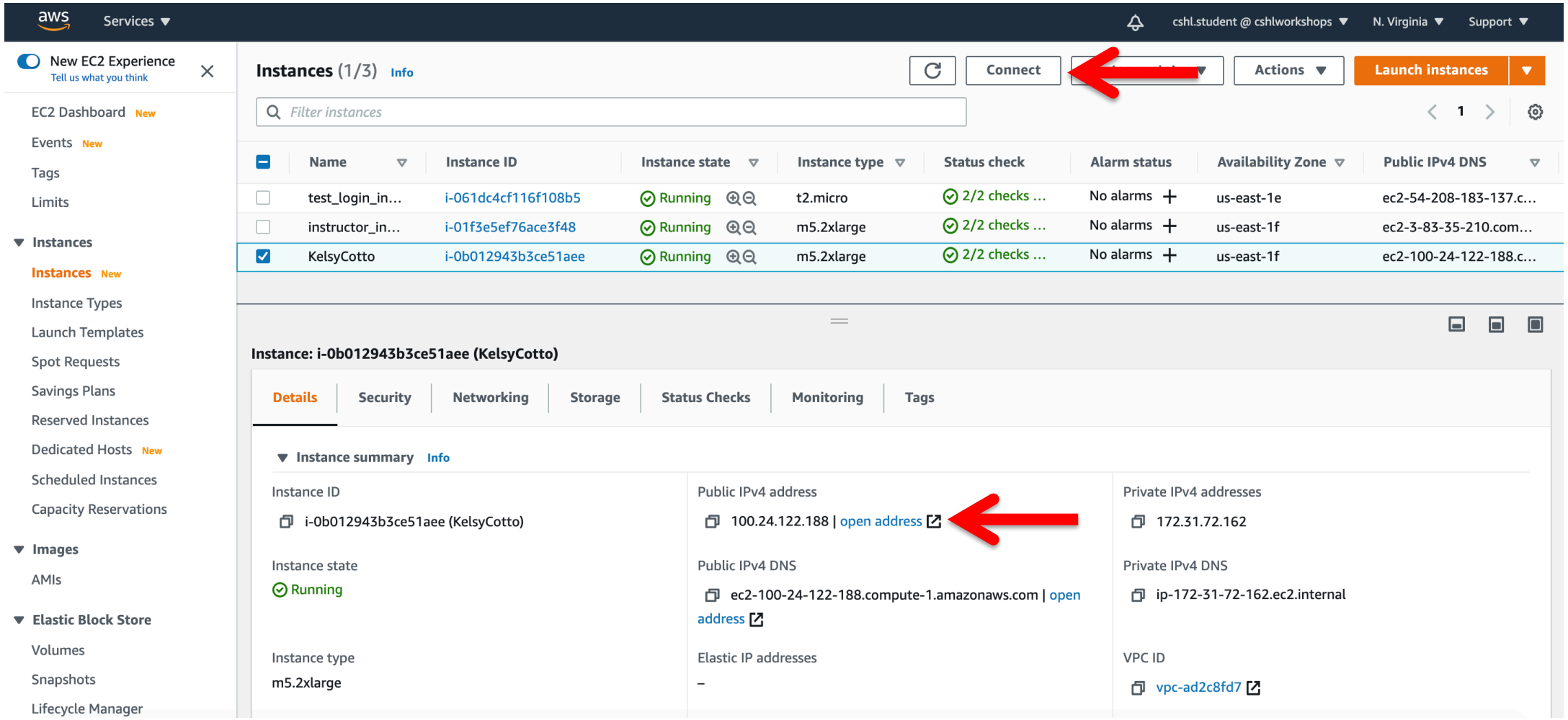


# Next morning, you can “Start” your instance again




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

**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 laptop/pc 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

# Break

# Key AWS concepts and terminology

- **HDD** - Hard disk drive. A particular type of storage hardware that is generally cheaper and larger but slower than SSD. HDD drives are traditional hard drives that access data on a spinning magnetic disk.
- **Ephemeral storage** - Also known as Instance Store storage. Data storage associated with an EC2 instance that is local to the host computer. This storage does not persist when the instance is stopped or terminated. In other words, anything you store in this way will be lost if the system is stopped or terminated. Instance store volumes may be backed by SSD or HDD devices.

# What is a Region?

- An AWS Region is set of compute resources that Amazon maintains (like the Data Center image shown before)
- Each Region corresponds to a physical warehouse of compute hardware (computers, storage, networking, etc.).
- At the time of writing there are 22 regions: (US East (N.Virginia), US East (Ohio), US West (Oregon), US West (N. California), GovCloud (US-West), GovCloud (US-East), Canada (Montreal), EU (Ireland), EU (Frankfurt), EU (London), EU (Paris), EU (Milan), EU (Stockholm), Middle East (Bahrain), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Seoul), Asia Pacific (Tokyo), Asia Pacific (Mumbai), Asia Pacific (Hong Kong), Asia Pacific (Beijing), and South America (Sao Paulo).
- When you are logged into the AWS EC2 console, you are always operating in one of these regions.

# What is a Region?

- Current region shown in the upper right corner of console
- It is important to pay attention to what region you are using for several reasons.
  - When you create an EC2 instance (EBS volume, etc) in one region you won't see it in another region.
  - The cost to use many AWS resources varies by region.
  - The region may influence network performance when you are accessing the instance, especially if you need to transfer large amounts of data in or out.
  - Billing is tracked separately for each region
  - Generally you should choose a region that is close to you or your users. But cost is also a consideration.