Advanced Sequencing Technologies & Applications

http://meetings.cshl.edu/courses.html

Cold Spring Harbor Laboratory

Introduction to cloud computing

Kelsy Cotto, Obi Griffith, Malachi Griffith,
Alex Wagner, Jason Walker
Advanced Sequencing Technologies & Applications
November 6- 18, 2018

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
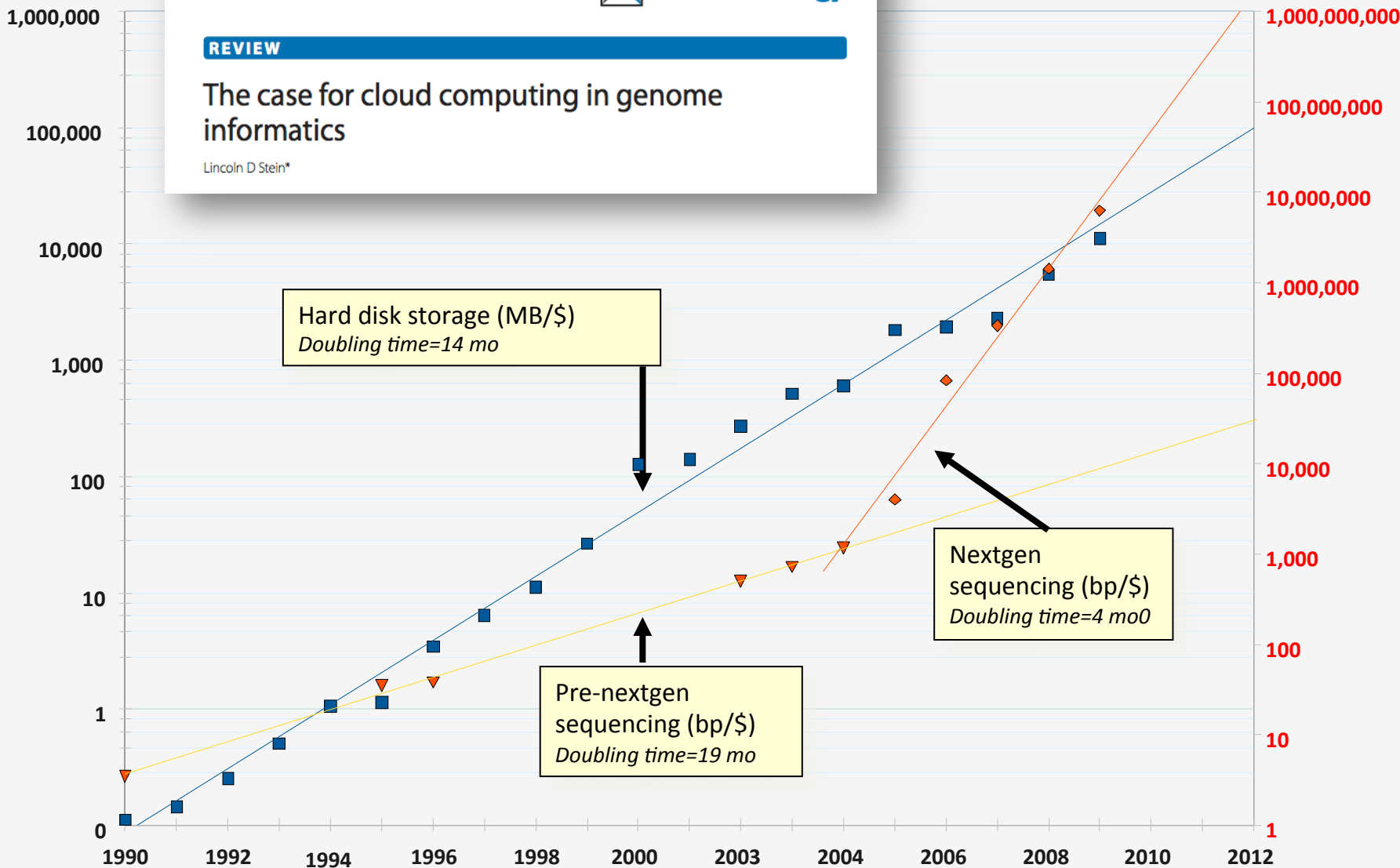
Disk Storage (Mbytes/$)

DNA Sequencing (bp/$)

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*

Hard disk storage (MB/$)
Doubling time=14 mo

Nextgen sequencing (bp/$)
Doubling time=4 mo0

Pre-nextgen sequencing (bp/$)
Doubling time=19 mo

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?

- 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/
- 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
- 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

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.

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 14 regions: (US East (N.Virginia), US East (Ohio), US West (Oregon), US West (N. California), Canada (Central), EU (Ireland), EU (Frankfurt), EU (London), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Seoul), Asia Pacific (Tokyo), Asia Pacific (Mumbai) 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.

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?

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
Linux RHEL SLES Windows Windows with SQL Standard Windows with SQL Web Windows with SQL Enterprise					
Region: US West (Oregon)					
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://github.com/griffithlab/rnaseq_tutorial/wiki/Intro-to-AWS-Cloud-Computing

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
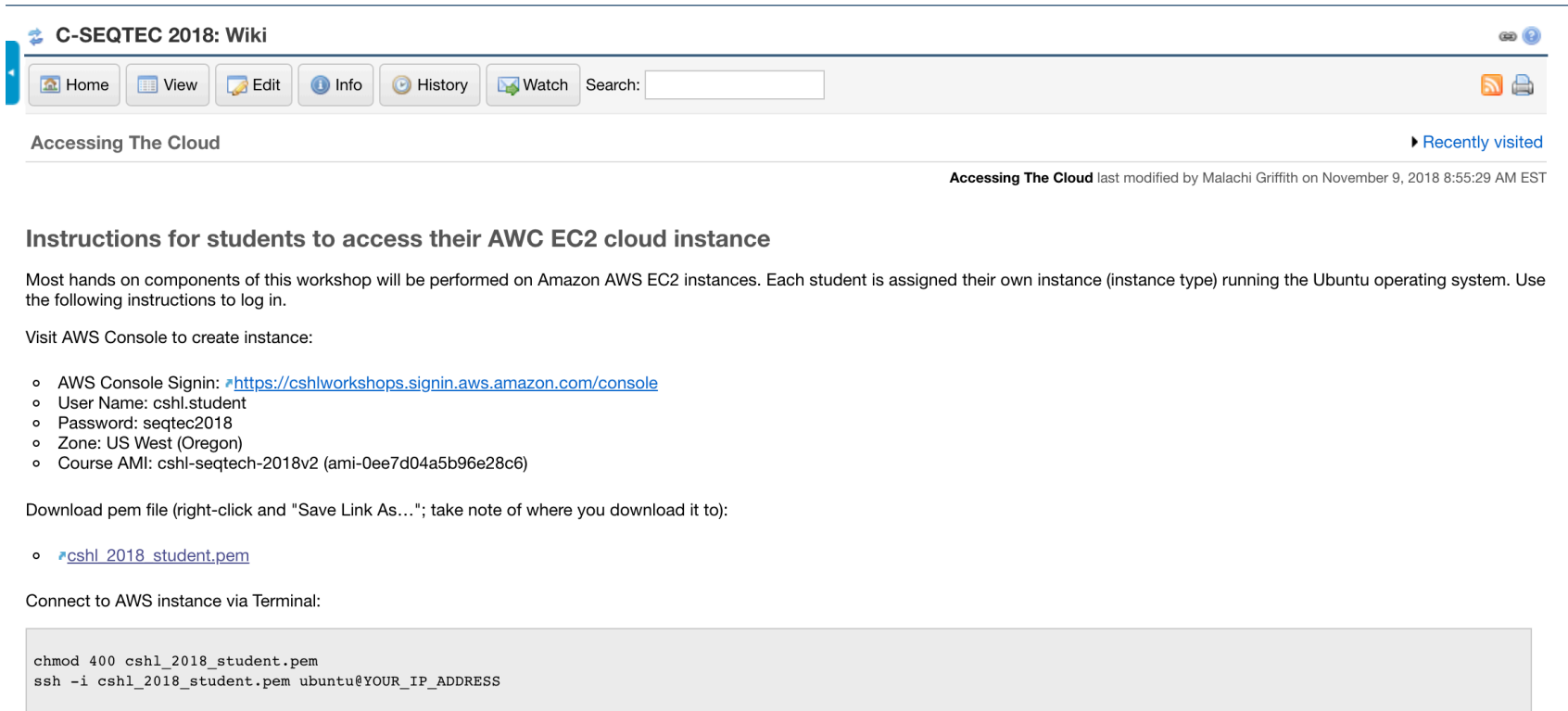
Things we have set up:

- 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, "Accessing the cloud" page

C-SEQTEC 2018: Wiki

Home | View | Edit | Info | History | Watch | Search:

Accessing The Cloud

▶ Recently visited

Accessing The Cloud last modified by Malachi Griffith on November 9, 2018 8:55:29 AM EST

Instructions for students to access their AWC EC2 cloud instance

Most hands on components of this workshop will be performed on Amazon AWS EC2 instances. Each student is assigned their own instance (instance type) running the Ubuntu operating system. Use the following instructions to log in.

Visit AWS Console to create instance:

- AWS Console Signin: https://cshlworkshops.signin.aws.amazon.com/console
- User Name: cshl.student
- Password: seqtec2018
- Zone: US West (Oregon)
- Course AMI: cshl-seqtech-2018v2 (ami-0ee7d04a5b96e28c6)

Download pem file (right-click and "Save Link As…"; take note of where you download it to):

- cshl_2018_student.pem

Connect to AWS instance via Terminal:

```
chmod 400 cshl_2018_student.pem
ssh -i cshl_2018_student.pem ubuntu@YOUR_IP_ADDRESS
```

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 2018 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_2018_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)

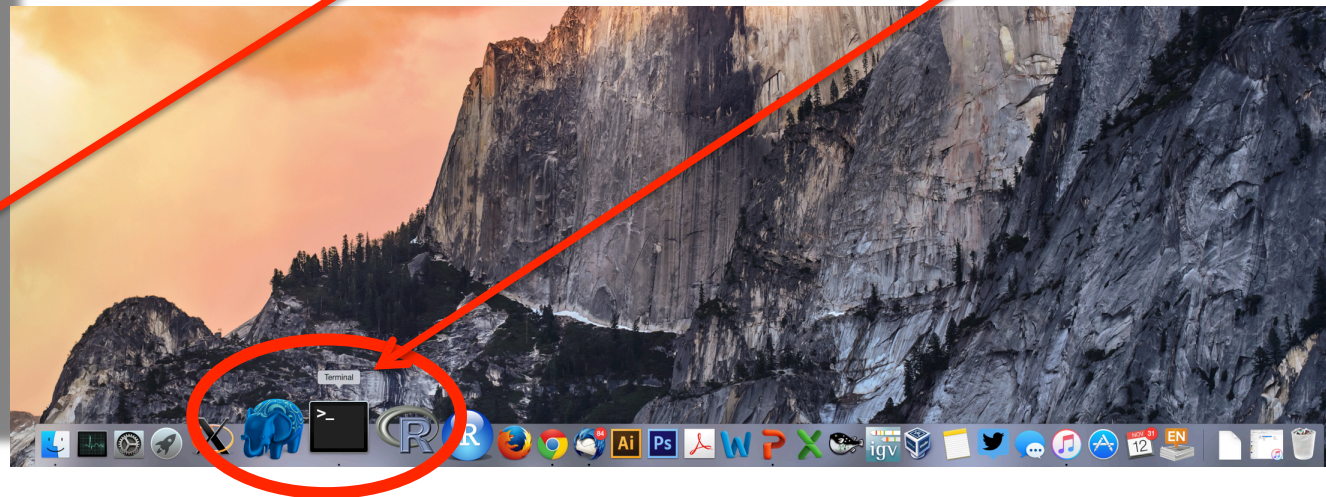# Take note of your Public DNS/IP and the instructions on changing permissions for the key file (Note, we will login as ubuntu NOT root)

Opening a 'terminal session' on a Mac

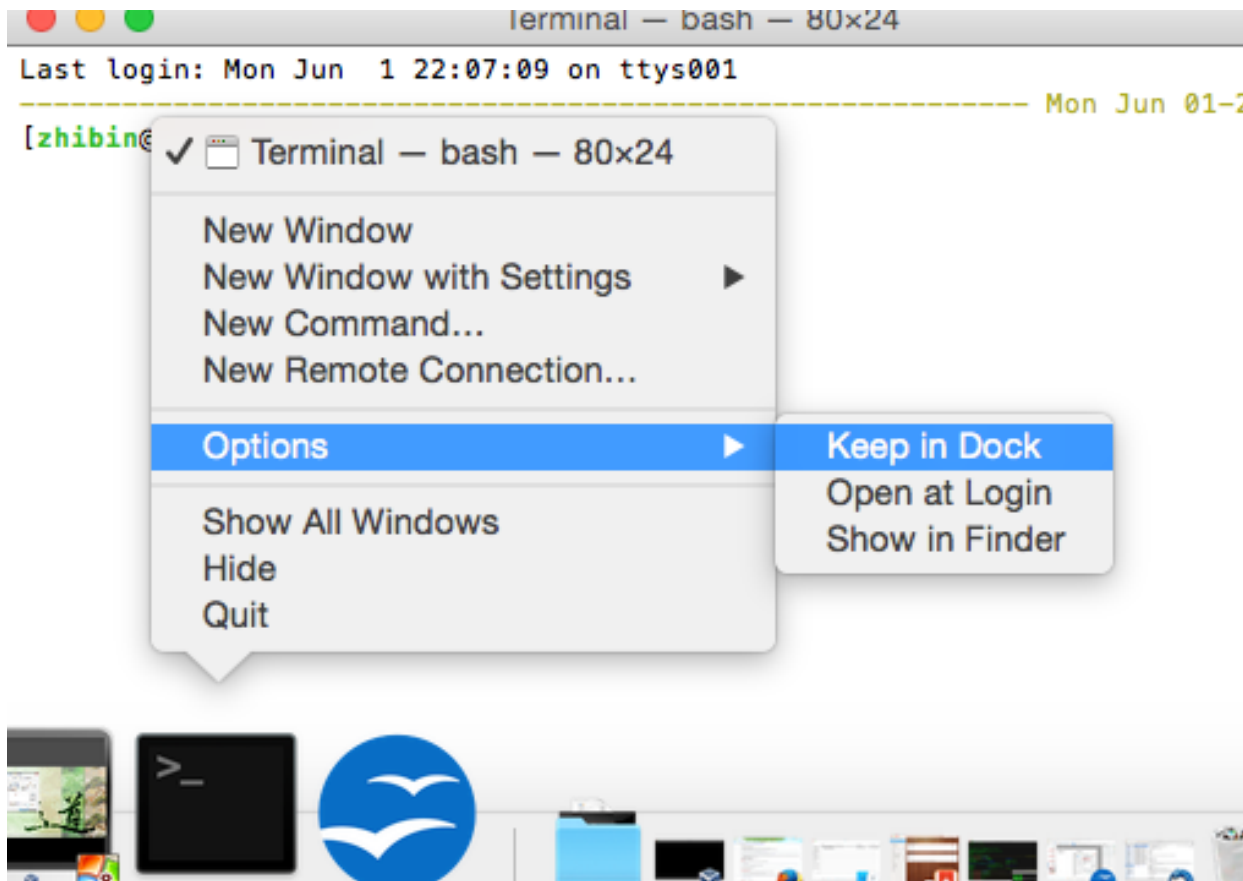In a Finder window
'Applications' -> 'Utilities' -> 'Terminal'
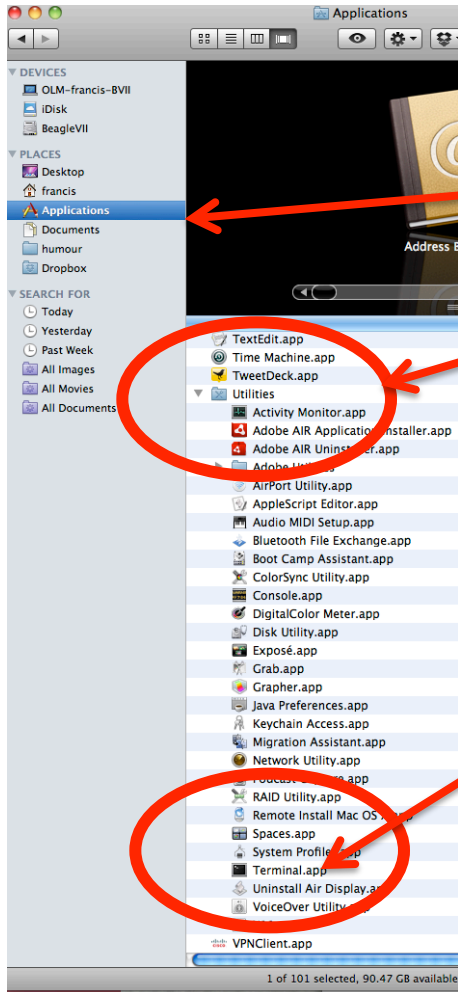
Or on your dock

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

Download .pem file to cshl directory from course wiki

Viewing the 'key' file once downloaded

cat cshl_2018_student.pem

```
GSC-Systemss-MacBook-Pro-3:cshl kcotto$ cat cshl_2018_student.pem
-----BEGIN RSA PRIVATE KEY-----
MIIEpAIBAAKCAQEA2Tgc+HFIf9S1YofCDwMHOC7S/o+t7sr/X660u7P5KhE22EQuiHFVtku4BkPM
oEEUYBtCwBR5dJvoqSoMBfd7/PQnLhPdWuVRVc4XiPZGhsq4E9jZFodzDS9aPAq1aRbHFxbjTCoi
clsKWqcdQQ2g5vLC0P4kqBRnB+KPicuGjPrN5ZrMefZcm57eYTmF9BAWoL67xh+vnNghKH82eXUj
fvk93PuzOKYKAKm5qzkpFTTSjKH5ip264JX44Kr8iNJeycGMpKxD2iZ4CPeiq/w1Zh8PWRfDzlnP
/PzKtrpWW+rkbZ/+hJIe2y8q7IbzUvKCrVZsG1QHfwUzUAZUw5DZ+QIDAQABAoIBAQCJqLsLY9d7
U5Uj/T8pflTL8Ti8JNgotjNDMxfMwNDOLRyh0AvhYlncbn4QUlVS1guQoJYb3p3eHZRD6Ka3/VBx
TVWrKOGmYh+6jiJyJQkrETu5MDrbwTlTiMfZtzuiftHrOxEPSsCL0lBshO3QNU/cLMrz4avfbRai
c51dyULbAkghWD/FHQzsgQ7KqdM6l5pLy1u2JvQ0wdlR6emEEB9OOZfDrSBZzmBGeoJd1EkepEtg
2Lgyjs4CLMxKIiqTMalY2fCJ3Xm9II5slOryGoo8lvujKSICJM2zyVIbVgEfwcolZlO20i97IcjU
GSqjDr0/EGu2iCi+8NvQ+u1R0JsVAoGBAO70xzZ6k02yLRi2Pq7ll3D67M9YaG1PeqOrE3M1nwmm
5uAWhmcBAGfGeLH8myJDBmcqtfocsCDyggkI9Ept5UmwxlZ1Ecq8C+Bsj/aDL/jvY+7AxfpK+GIU
L00kXA9FaTICeOVukFDyR4zIlP4sy68iFcKAwIn+tDHyewPi+7BzAoGBAOi2bbdvEDRfNBO9bi1g
qo5Kbo92oXfEwEB55/qHj1hABj+W7hU2V4tavNG1w2RJ9y36Zp0S8QsY1LyiaaZ9i4Zow27rJG1N
2L3B+JeF3ib1JHb94aLiFPsi492k4Wco1QHRsk5dXZykCpwWrkTImgzQErtAQMO35DXG5zWz6gzj
AoGAUqHU7fLvU3PlAmGrRYHhn9E16PHSktJkP4reW/3T9FuCBUXkzmPbRwesHyiSYQ06iByN92LK
SWpzKzqsd7yPCK+XtrgfIfNrS/1ev0lYYoYuEZcaH1ajlA8U1RhWOXWwTfnq4rQcNdHCil2moErP
WM25ppwteMrad1c50PtCPB8CgYAkK0Z+PircmPd1Ifof32hi2Tq53YxXGyNuYgQmMi+bXcga1hCH
B30v0ZiLV57Y1Y4vfT24Lhx8xw05jFyzXcvzxhCqCq4Mtw66340Nr+bJU0ypmegZ8t2j8tZLura1
2+jPcZIvD9/e2rgzbmsDkdIvr59ZrzDhupEjsph5Bk+YVQKBgQC6vNLisnJDX+dbLB6My2D1ZSJ+
paubgw7fi78PsAeozNLWX8iHzBBEuDcxVU7TD8fsFIttBEipUERCMAQRxQOKjrdxRHTc4G7DcDwG
vQ6O1EwNyPK0Mn8FzEGBb9WT7bk5Ox9iNox2YYzpRwZdTSCpFaxwErihSaf+VetF2+lijQ==
-----END RSA PRIVATE KEY-----GSC-Systemss-MacBook-Pro-3:cshl 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_2018_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 …)
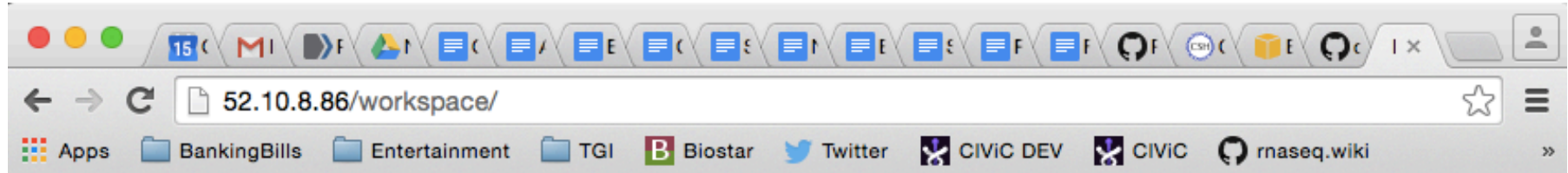
So, when you have:

chmod 400 <file name>

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

Logging into your instance

Mac/Linux

cd ~/cshl

chmod 400 cshl_2018_student.pem

ssh -i cshl_2018_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!

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

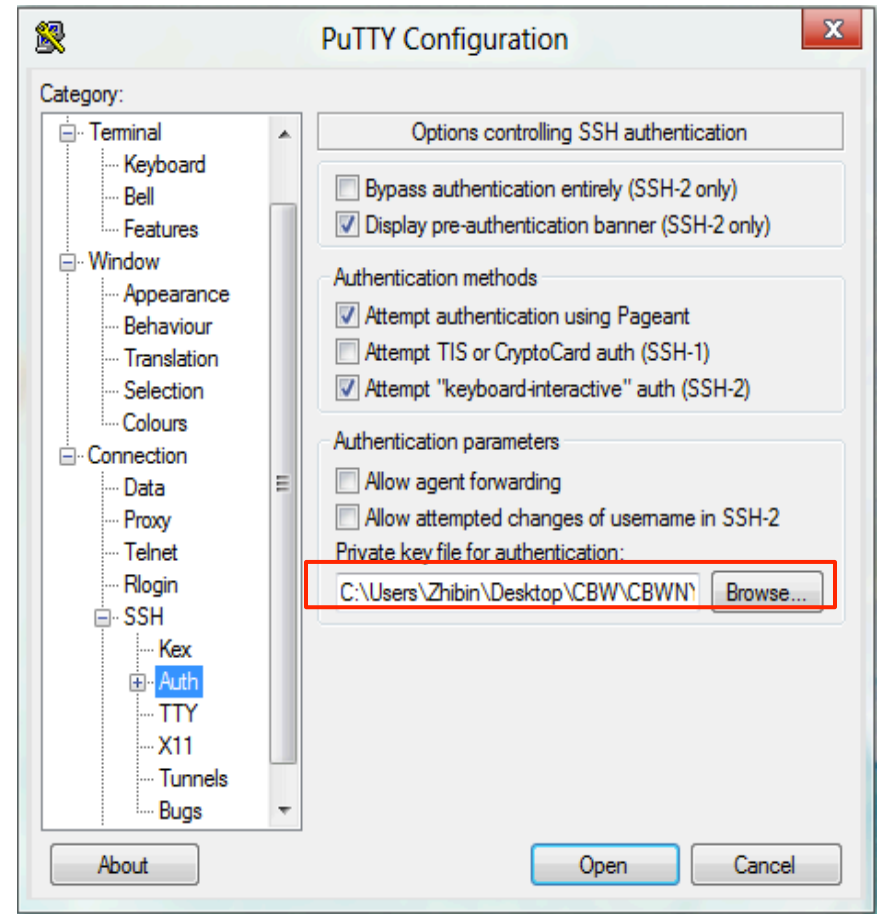# Break

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
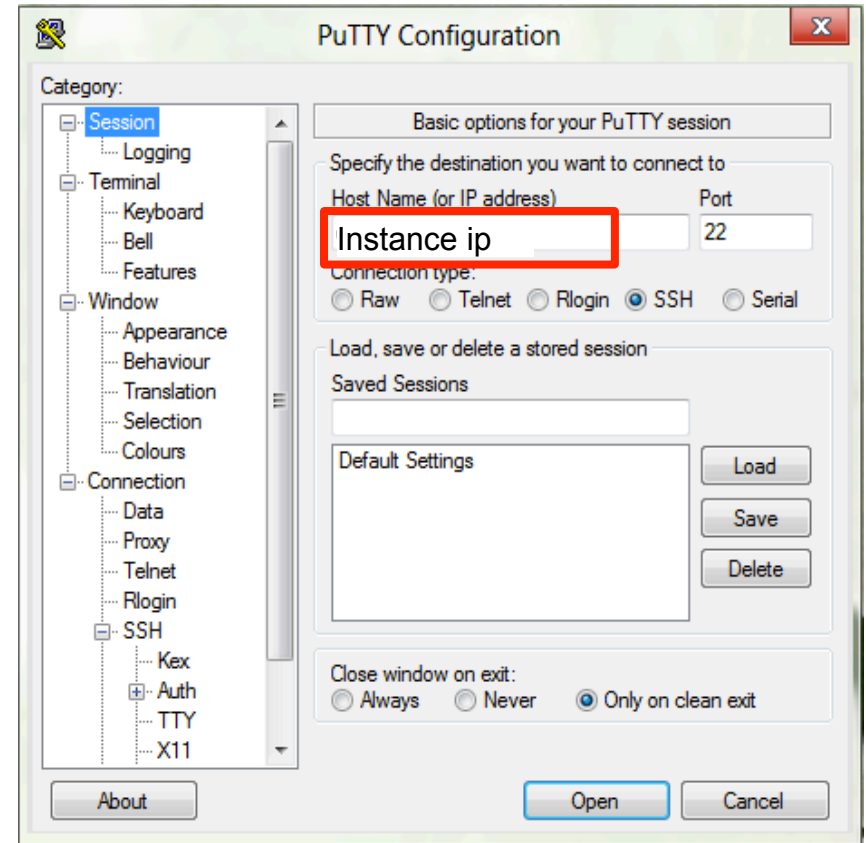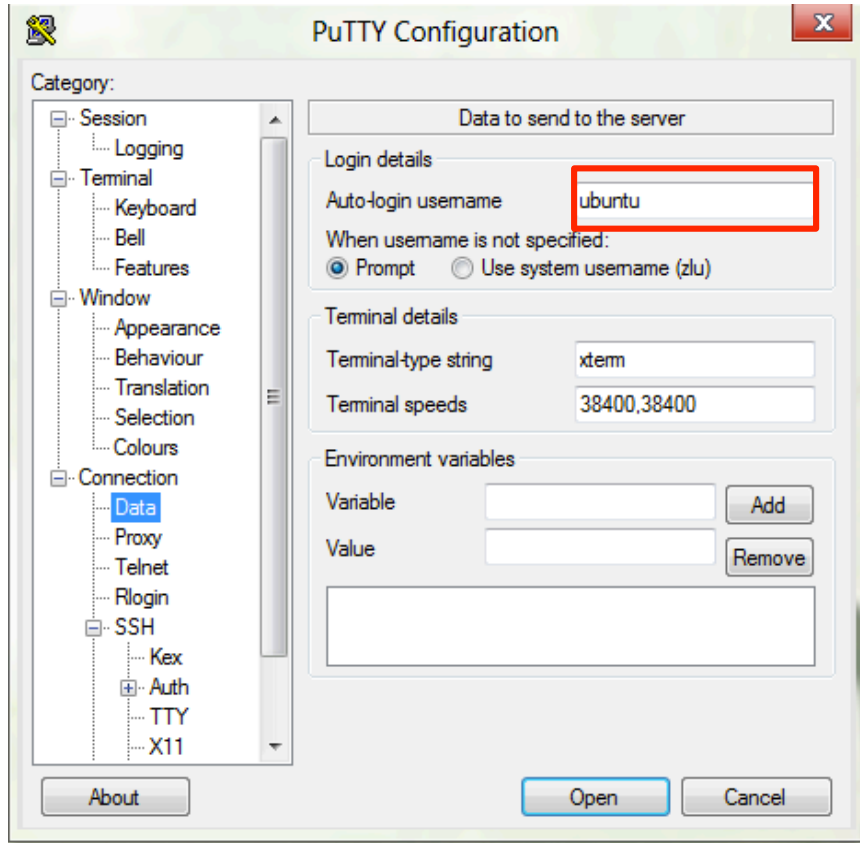
Logging into your instance (Windows)
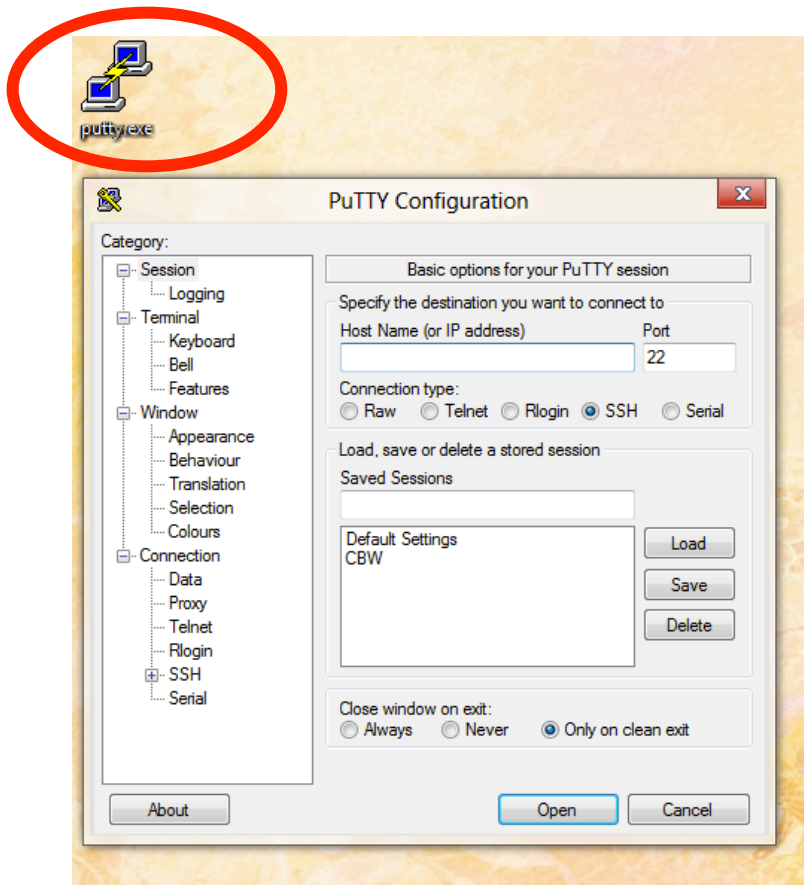
Open PuTTY

Browse to the cshl_2018_student.ppk file
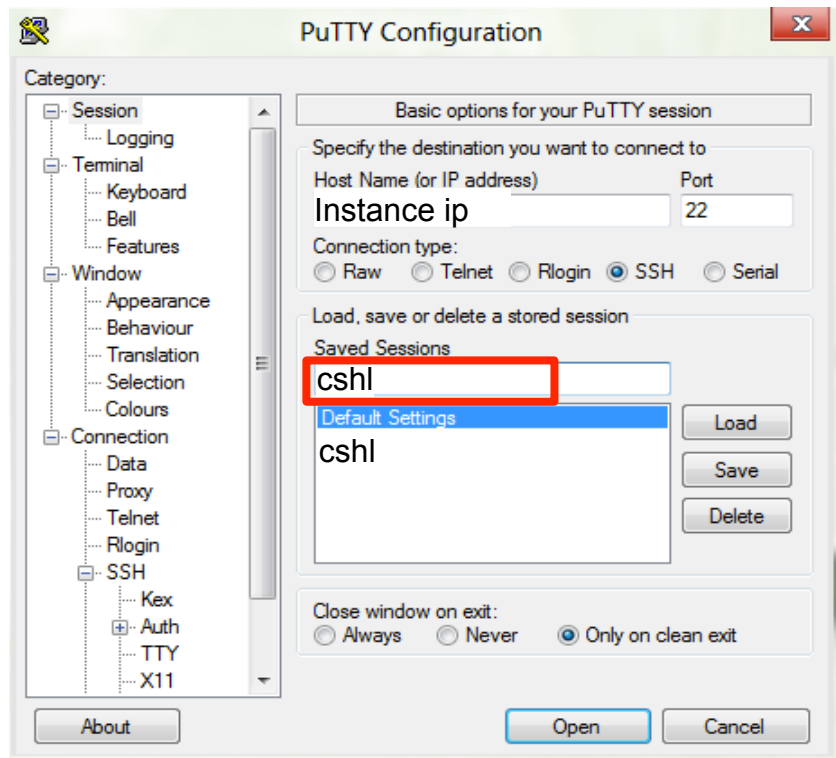
Logging into your instance (Windows)

Enter the user name 'ubuntu'   Enter the host name

Logging into your instance (Windows)
Open PuTTY

Break