



eric@krtecek:~\$

Lab Introduction and an Introduction to Unix

Anna Karnkowska, Eric Salomaki

Many slides courtesy
Dr. Sophie Shaw

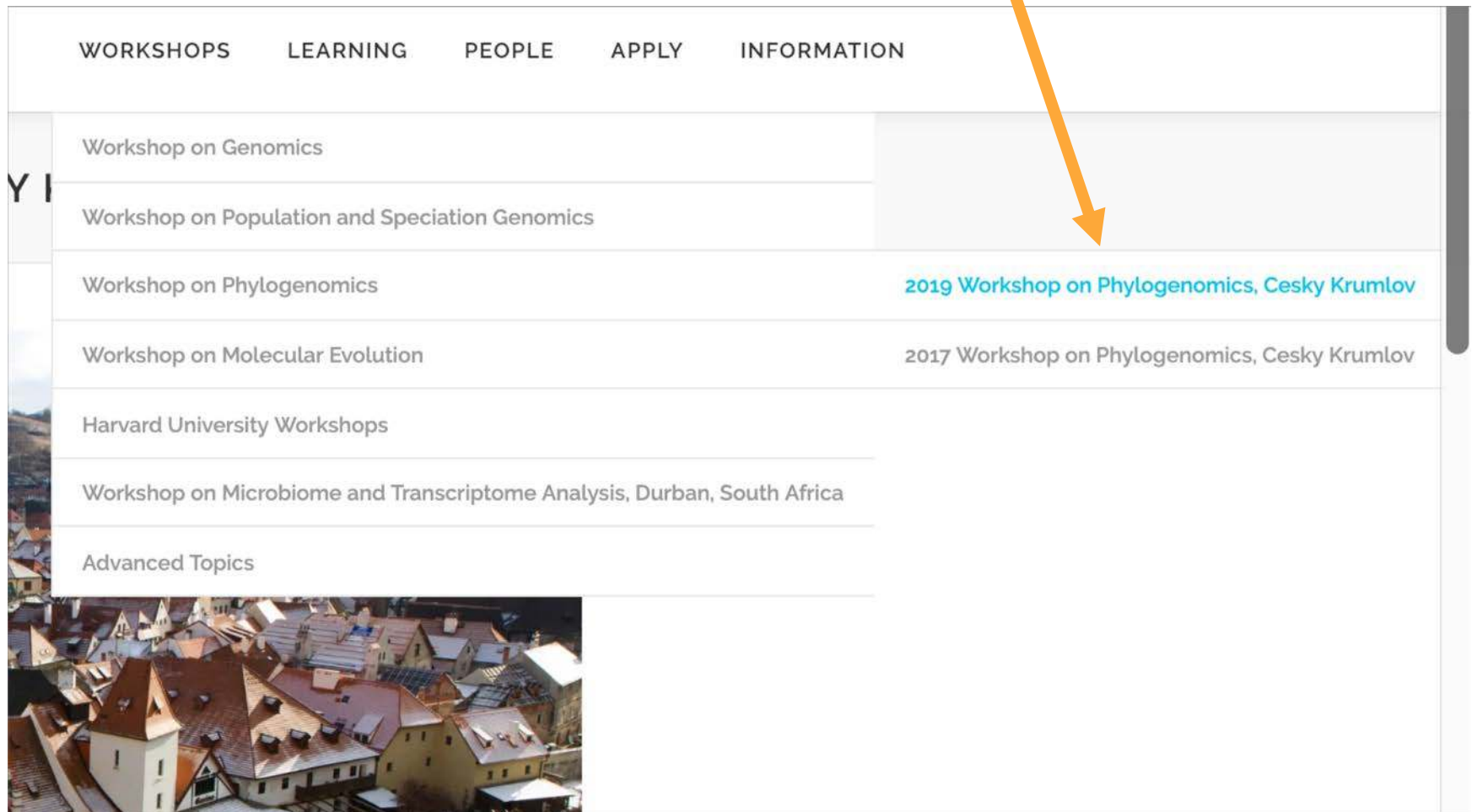


Overview

- Introduction
- Using the Amazon Machine Images (AMI)
- Introduction to Unix

These Slides!

<http://evomics.org/workshops/2019-workshop-on-phylogenomics-cesky-krumlov/>



The screenshot shows the 'WORKSHOPS' section of the evomics.org website. The navigation bar includes 'WORKSHOPS', 'LEARNING', 'PEOPLE', 'APPLY', and 'INFORMATION'. The workshop list is as follows:

Workshop Name	Location / Year
Workshop on Genomics	
Workshop on Population and Speciation Genomics	
Workshop on Phylogenomics	2019 Workshop on Phylogenomics, Cesky Krumlov
Workshop on Molecular Evolution	2017 Workshop on Phylogenomics, Cesky Krumlov
Harvard University Workshops	
Workshop on Microbiome and Transcriptome Analysis, Durban, South Africa	
Advanced Topics	

An orange arrow points from the URL above to the '2019 Workshop on Phylogenomics, Cesky Krumlov' entry. At the bottom left, there is a small image of a town with red-tiled roofs.

What We're Going To Do

21 Jan	9a – 12p	Scott Handley & Toni Gabaldón	Introduction and Orientation	Theater
	2p – 5p	Workshop Team	Lab Introduction	House of Prelate
	7p – 10p	Everyone	Scientific Speed Networking	TBD
22 Jan	9a – 12p	Antonis Rokas	Introduction to Phylogenomics	Theater
	2p – 5p	Workshop Team	Alignment and Alignment Trimming	House of Prelate
	7p – 10p	Workshop Team	Tree Visualization	House of Prelate
23 Jan	9a – 12p	Toni Gabaldón	Introduction to Phylogenetics, and Orthology and Paralogy	Theater
	2p – 5p	Toni Gabaldón	Orthology and Paralogy Prediction lab	House of Prelate
	7p – 10p	Workshop Team	Partitioning and Concatentation Laboratory	House of Prelate

Background on Unix, Cloud computing, and in working in an instance
Introduction to Terminal and Unix

What We're Going To Do

21 Jan	9a – 12p	Scott Handley & Toni Gabaldón	Introduction and Orientation	Theater
	2p – 5p	Workshop Team	Lab Introduction	House of Prelate
	7p – 10p	Everyone	Scientific Speed Networking	TBD
22 Jan	9a – 12p	Antonis Rokas	Introduction to Phylogenomics	Theater
	2p – 5p	Workshop Team	Alignment and Alignment Trimming	House of Prelate
	7p – 10p	Workshop Team	Tree Visualization	House of Prelate
23 Jan	9a – 12p	Toni Gabaldón	Introduction to Phylogenetics, and Orthology and Paralogy	Theater
	2p – 5p	Toni Gabaldón	Orthology and Paralogy Prediction lab	House of Prelate
	7p – 10p	Workshop Team	Partitioning and Concatentation Laboratory	House of Prelate

How and why to build a multiple sequence alignment

Is your alignment correct?

How to trim your alignment for phylogenetics

What We're Going To Do

21 Jan	9a – 12p	Scott Handley & Toni Gabaldón	Introduction and Orientation	Theater
	2p – 5p	Workshop Team	Lab Introduction	House of Prelate
	7p – 10p	Everyone	Scientific Speed Networking	TBD
22 Jan	9a – 12p	Antonis Rokas	Introduction to Phylogenomics	Theater
	2p – 5p	Workshop Team	Alignment and Alignment Trimming	House of Prelate
	7p – 10p	Workshop Team	Tree Visualization	House of Prelate
23 Jan	9a – 12p	Toni Gabaldón	Introduction to Phylogenetics, and Orthology and Paralogy	Theater
	2p – 5p	Toni Gabaldón	Orthology and Paralogy Prediction lab	House of Prelate
	7p – 10p	Workshop Team	Partitioning and Concatentation Laboratory	House of Prelate

Important concepts for making your data accessible

Exploring tools for looking at your tree(s)

Web Based

Software

Toolkits

What We're Going To Do

21 Jan	9a – 12p	Scott Handley & Toni Gabaldón	Introduction and Orientation	Theater
	2p – 5p	Workshop Team	Lab Introduction	House of Prelate
	7p – 10p	Everyone	Scientific Speed Networking	TBD
22 Jan	9a – 12p	Antonis Rokas	Introduction to Phylogenomics	Theater
	2p – 5p	Workshop Team	Alignment and Alignment Trimming	House of Prelate
	7p – 10p	Workshop Team	Tree Visualization	House of Prelate
23 Jan	9a – 12p	Toni Gabaldón	Introduction to Phylogenetics, and Orthology and Paralogy	Theater
	2p – 5p	Toni Gabaldón	Orthology and Paralogy Prediction lab	House of Prelate
	7p – 10p	Workshop Team	Partitioning and Concatentation Laboratory	House of Prelate

What is Partitioning and the why/when/how to do it
Tools for partitioning your data
How to concatenate your data

Good Workshop Practice

- PowerPoint interspersed with Challenges
- Ask us lots of questions!
- Work together
- Take breaks
- Use a cheat sheet, google, etc.
- Have Fun!

File Commands	System Info
ls - directory listing	date - show the current date and time
ls -al - formatted listing with hidden files	cal - show this month's calendar
cd <i>dir</i> - change directory to <i>dir</i>	uptime - show current uptime
cd - change to home	w - display who is online
pwd - show current directory	whoami - who you are logged in as
mkdir <i>dir</i> - create a directory <i>dir</i>	finger <i>user</i> - display information about <i>user</i>
rm <i>file</i> - delete <i>file</i>	uname -a - show kernel information
rm -r <i>dir</i> - delete directory <i>dir</i>	cat /proc/cpuinfo - cpu information
rm -f <i>file</i> - force remove <i>file</i>	cat /proc/meminfo - memory information
rm -rf <i>dir</i> - force remove directory <i>dir</i> *	man <i>command</i> - show the manual for <i>command</i>
cp <i>file1 file2</i> - copy <i>file1</i> to <i>file2</i>	df - show disk usage
cp -r <i>dir1 dir2</i> - copy <i>dir1</i> to <i>dir2</i> ; create <i>dir2</i> if it doesn't exist	du - show directory space usage
mv <i>file1 file2</i> - rename or move <i>file1</i> to <i>file2</i>	free - show memory and swap usage
if <i>file2</i> is an existing directory, moves <i>file1</i> into directory <i>file2</i>	whereis <i>app</i> - show possible locations of <i>app</i>
ln -s <i>file link</i> - create symbolic link <i>link</i> to <i>file</i>	which <i>app</i> - show which <i>app</i> will be run by default
touch <i>file</i> - create or update <i>file</i>	Compression
cat > <i>file</i> - places standard input into <i>file</i>	tar cf <i>file.tar files</i> - create a tar named <i>file.tar</i> containing <i>files</i>
more <i>file</i> - output the contents of <i>file</i>	tar xf <i>file.tar</i> - extract the files from <i>file.tar</i>
head <i>file</i> - output the first 10 lines of <i>file</i>	tar czf <i>file.tar.gz files</i> - create a tar with Gzip compression
tail <i>file</i> - output the last 10 lines of <i>file</i>	tar xzf <i>file.tar.gz</i> - extract a tar using Gzip
tail -f <i>file</i> - output the contents of <i>file</i> as it grows, starting with the last 10 lines	tar cjf <i>file.tar.bz2</i> - create a tar with Bzip2 compression
Process Management	tar xjf <i>file.tar.bz2</i> - extract a tar using Bzip2
ps - display your currently active processes	gzip <i>file</i> - compresses <i>file</i> and renames it to <i>file.gz</i>
top - display all running processes	gzip -d <i>file.gz</i> - decompresses <i>file.gz</i> back to <i>file</i>
kill <i>pid</i> - kill process id <i>pid</i>	Network
killall <i>proc</i> - kill all processes named <i>proc</i> *	ping <i>host</i> - ping <i>host</i> and output results
bg - lists stopped or background jobs; resume a stopped job in the background	whois <i>domain</i> - get whois information for <i>domain</i>
fg - brings the most recent job to foreground	dig <i>domain</i> - get DNS information for <i>domain</i>
fg <i>n</i> - brings job <i>n</i> to the foreground	dig -x <i>host</i> - reverse lookup <i>host</i>
File Permissions	wget <i>file</i> - download <i>file</i>
chmod <i>octal file</i> - change the permissions of <i>file</i> to <i>octal</i> , which can be found separately for user, group, and world by adding:	wget -c <i>file</i> - continue a stopped download
4 - read (r) 2 - write (w) 1 - execute (x)	Installation
Examples:	Install from source:
chmod 777 - read, write, execute for all	./configure
chmod 755 - rwx for owner, rx for group and world	make
For more options, see man chmod.	make install
SSH	dpkg -i <i>pkg.deb</i> - install a package (Debian)
ssh <i>user@host</i> - connect to <i>host</i> as <i>user</i>	rpm -Uvh <i>pkg.rpm</i> - install a package (RPM)
ssh -p <i>port user@host</i> - connect to <i>host</i> on port <i>port</i> as <i>user</i>	Shortcuts
ssh-copy-id <i>user@host</i> - add your key to <i>host</i> for <i>user</i> to enable a keyed or passwordless login	Ctrl+C - halts the current command
Searching	Ctrl+Z - stops the current command, resume with fg in the foreground or bg in the background
grep <i>pattern files</i> - search for <i>pattern</i> in <i>files</i>	Ctrl+D - log out of current session, similar to exit
grep -r <i>pattern dir</i> - search recursively for <i>pattern</i> in <i>dir</i>	Ctrl+W - erases one word in the current line
<i>command</i> grep <i>pattern</i> - search for <i>pattern</i> in the output of <i>command</i>	Ctrl+U - erases the whole line
locate <i>file</i> - find all instances of <i>file</i>	Ctrl+R - type to bring up a recent command
	!! - repeats the last command
	exit - log out of current session
	* use with extreme caution.



Watching vs Doing



Watch and Listen

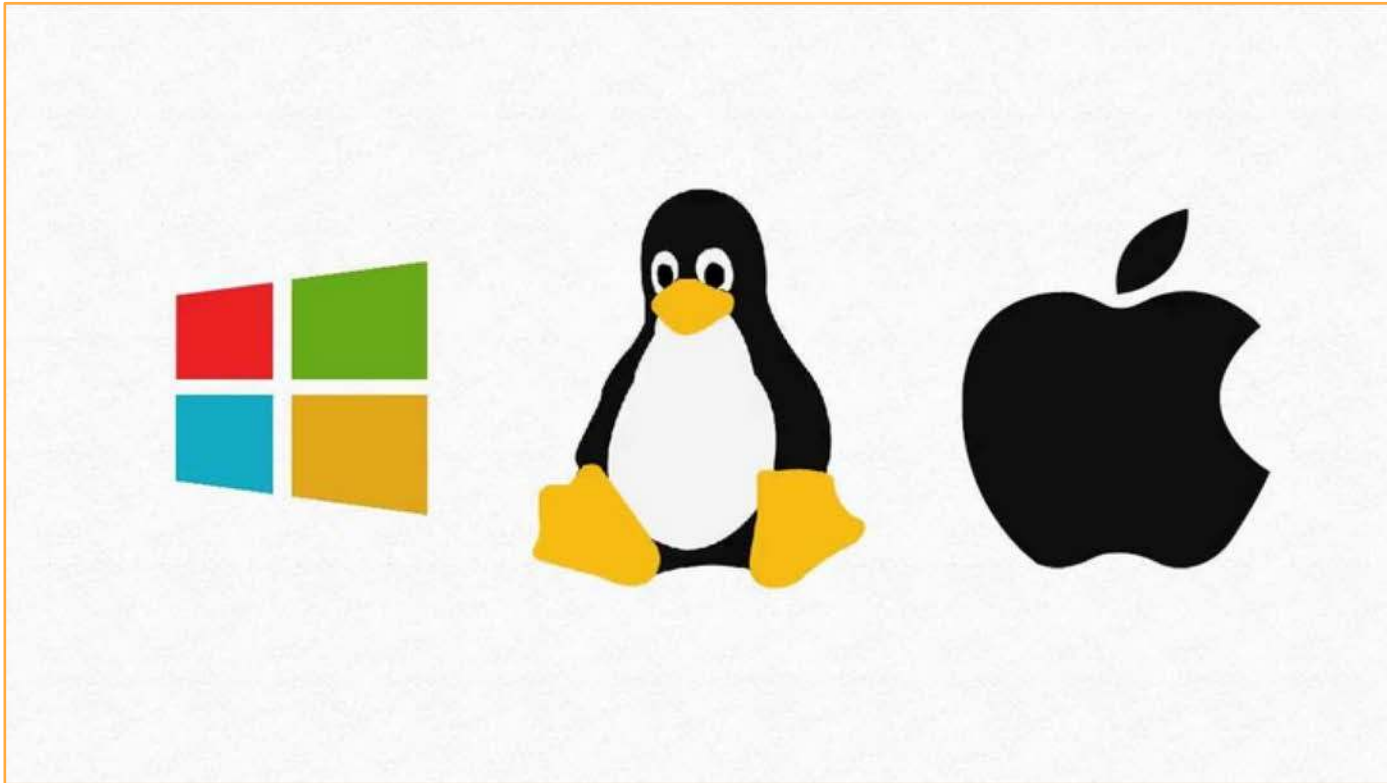


Try it for yourself

What is Unix?



- Operating System





Why Unix?

- Bioinformatics software designed to run on Unix platforms.
- Large amounts of data.
- Much faster than Windows PC.



Beast2

Bayesian evolutionary analysis by sampling trees

IQ-TREE

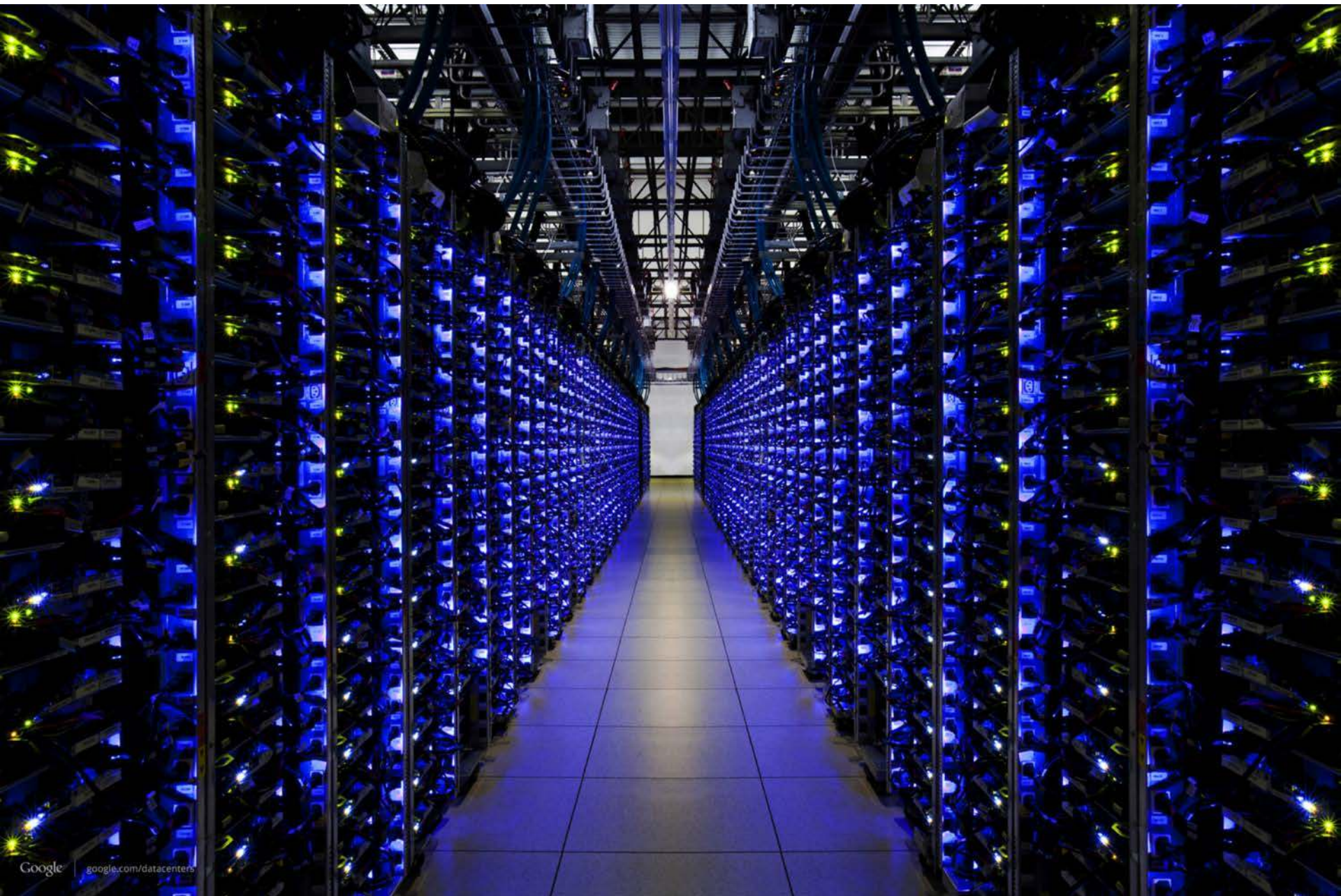
Efficient software for phylogenomic inference

How Can We Use Unix?



- Linux computers or servers
- Computer clusters
- The cloud
 - What we're going to use this week





Cloud Computing Solutions



Google Cloud Platform

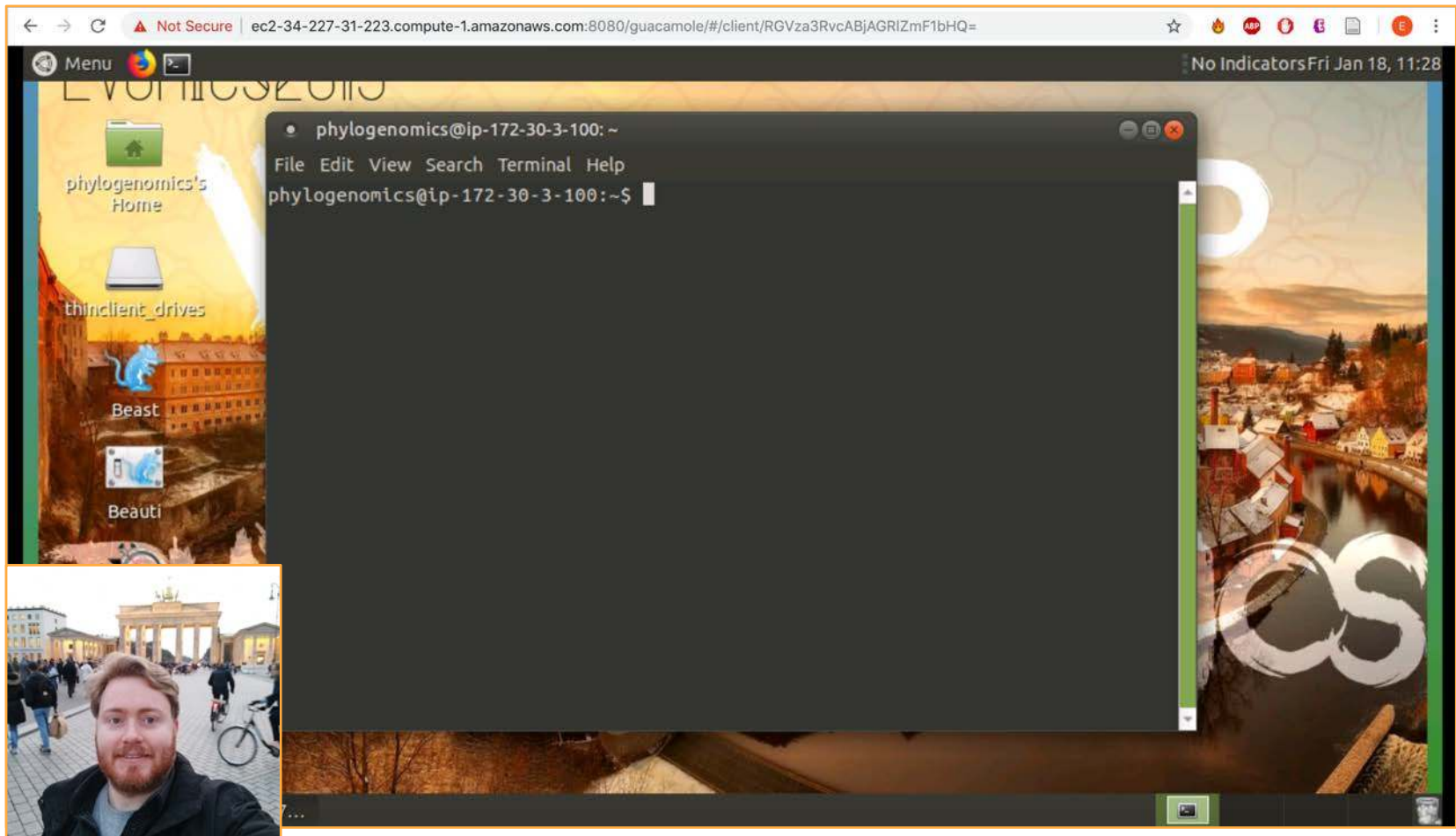




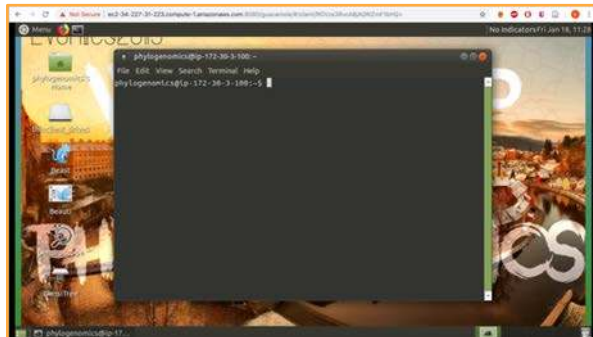
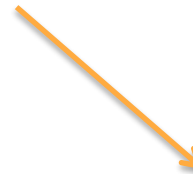
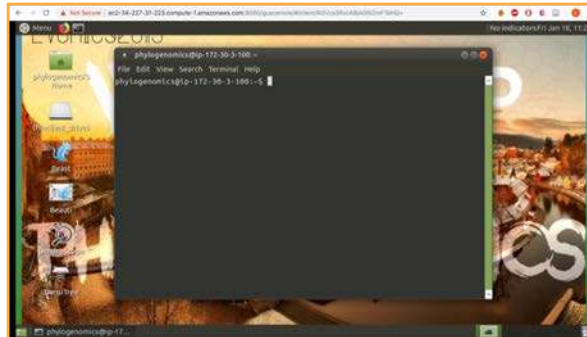
AWS “Availability Zones” and Data Centres

How it Works

AMI (“Amazon Machine Image”)
Base computer with all data and software



How it Works



Own copy of the AMI = Instance
(Virtual Machine or VM)

Terminology

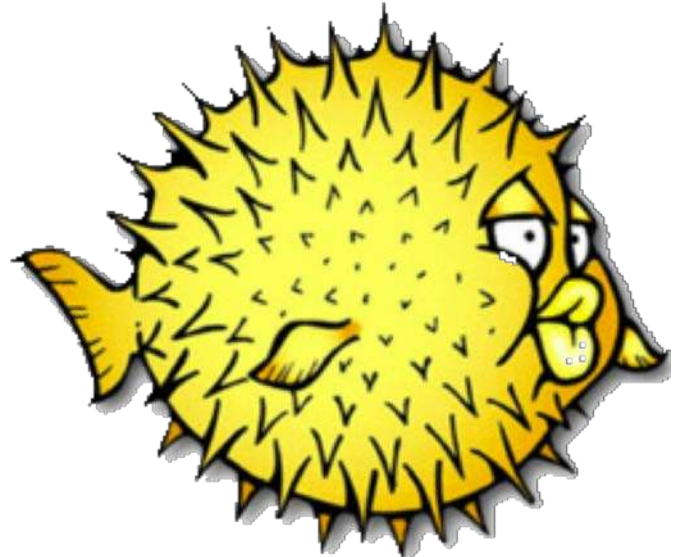


- Creating an instance – *buying a brand new computer with software already installed.*
- Starting an instance – *turning that computer on.*
- Stopping an instance – *turning that computer off.*
- Terminating an instance – *setting that computer on fire and throwing it out of the window.*

Connecting to Your Instance



Remote Desktop
Software
e.g. Guacamole or
X2Go



Secure Shell –
“SSH”
e.g. SSH or PuTTY



<http://evomics.org/workshops/2019-workshop-on-phylogenomics-cesky-krumlov/>

Anna Karnkowska, University of Warsaw

Antonis Rokas, Vanderbilt University

Miscellaneous

- [Instance Addresses – Check Daily](#)
 - [Faculty Lunches – Sign Up!](#)
 - [T-shirt Competition](#)
-

SCHEDULE

Week 1: 20-26 January 2019



Workshop on Phylogenomics 2019 Instance List

File Edit View Insert Format Data Tools Add-ons Help

100% \$ % .0 .00 123 Arial

fx

	A	B	C	D	E
1	Name	Public Domain			
2	Eric	ec2-34-227-31-223.compute-1.amazonaws.com			
3	Marina	ec2-3-82-175-48.compute-1.amazonaws.com			
4	Jacob	ec2-3-83-102-107.compute-1.amazonaws.com			
5	Ania	ec2-52-91-1-245.compute-1.amazonaws.com			



Find your name and copy
your public domain

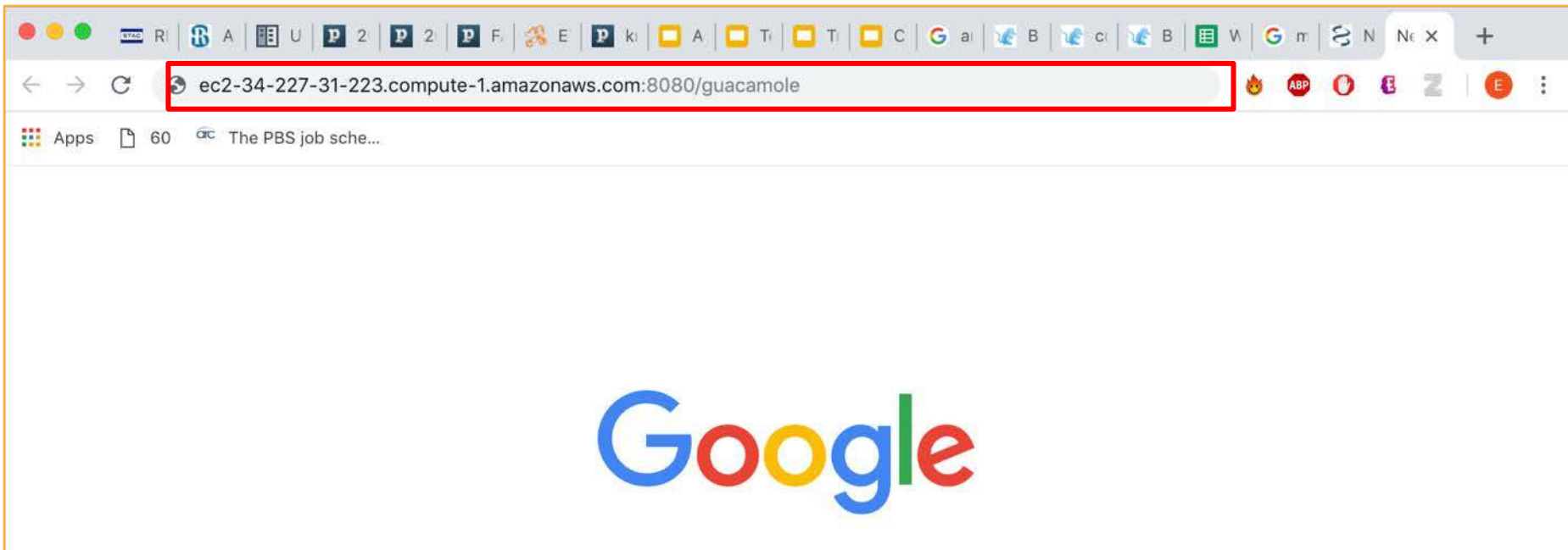
Open your internet browser
(e.g. Google Chrome)



Paste the public domain followed by
:8080/guacamole

Example:

ec2-34-227-31-223.compute-1.amazonaws.com:8080/guacamole



Enter the username
“phylogenomics” and password



APACHE GUACAMOLE

phylogenomics|

Password

Login

Select Desktop

ALL CONNECTIONS

-  Desktop
-  Terminal



Enter the username
“phylogenomics” and
password again

Login to ip-172-30-3-212



Just connecting

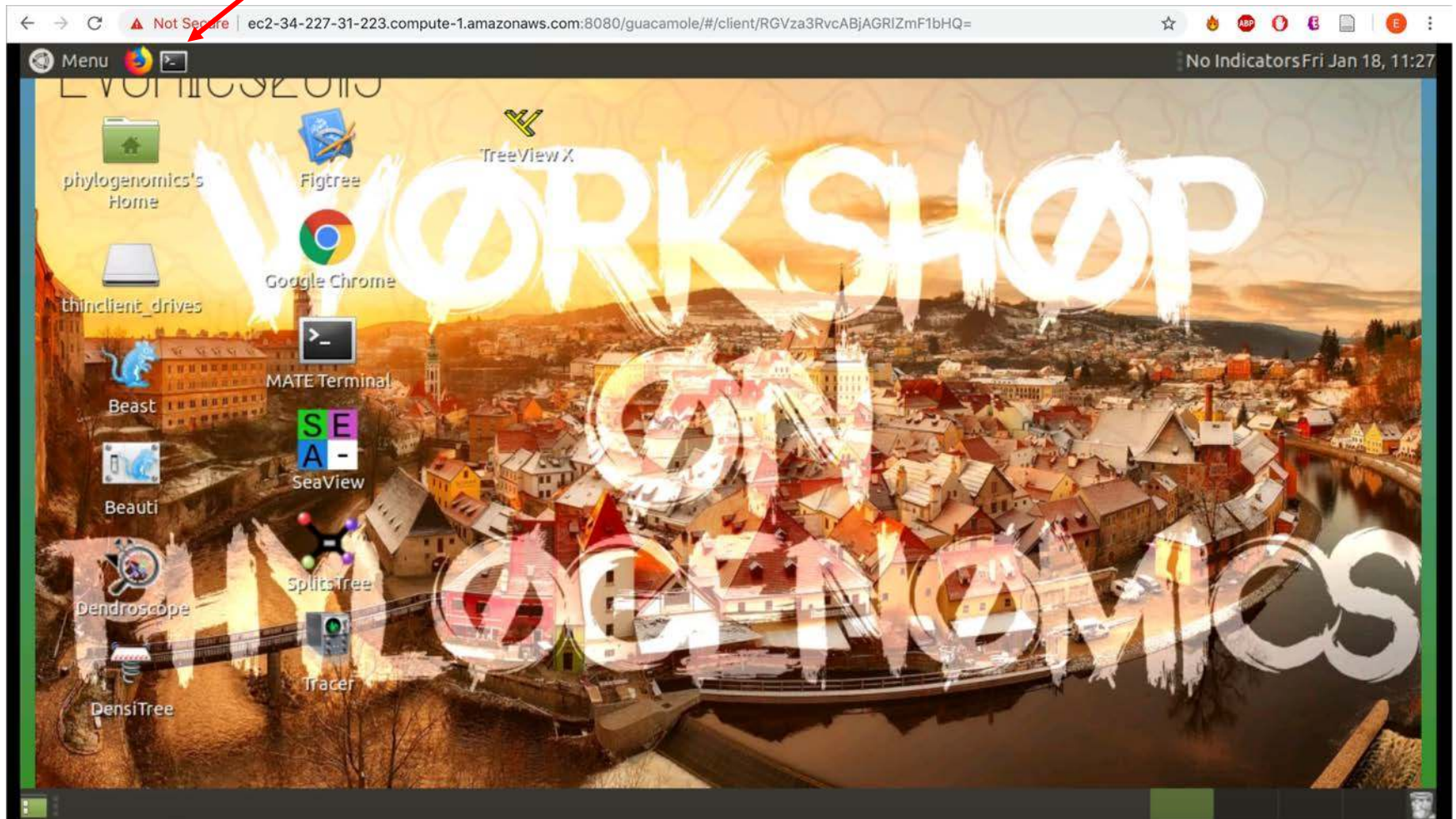
Session

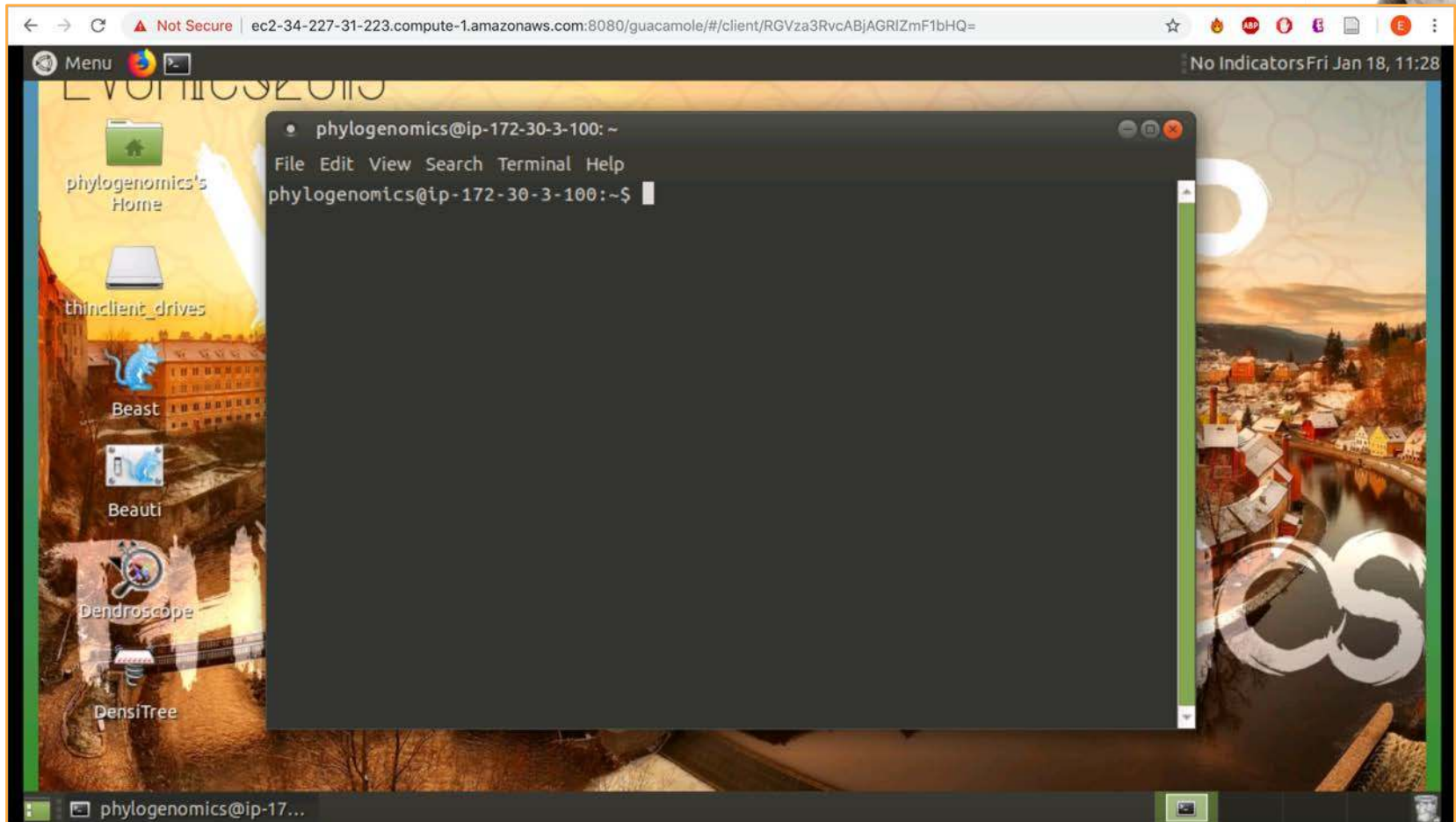
username

password

OK Cancel

Open a Terminal Window using this icon





You're now connected and you're ready to learn some Unix!

But First...

The domain address will change every day after we stop and re-start the instances.

Each morning, you will need to return to the “Instance List” webpage, retrieve your new address and log in again



Workshop on Phylogenomics 2019 Instance List					
File Edit View Insert Format Data Tools Add-ons Help					
100% \$ % .0 .00 123 Arial					
fx					
	A	B	C	D	E
1	Name	Public Domain			
2	Eric	ec2-34-227-31-223.compute-1.amazonaws.com			
3	Marina	ec2-3-82-175-48.compute-1.amazonaws.com			
4	Jacob	ec2-3-83-102-107.compute-1.amazonaws.com			
5	Ania	ec2-52-91-1-245.compute-1.amazonaws.com			

Copying and Pasting

AVOID COPYING AND PASTING WHEREVER POSSIBLE!



But if you do need to...

Press Ctrl+Alt+Shift

Paste the text into the box with right click -> Paste

Press Ctrl+Alt+Shift again

You can now paste into the instance using right click

A screenshot of a web browser displaying a Guacamole interface. The browser's address bar shows a URL starting with 'ec2-34-227-31-223.compute-1.amazonaws.com'. The interface is divided into two main sections. On the left, there is a 'Clipboard' section with a text box containing the placeholder text 'TEXT HERE WILL BE PASTED INTO THE INSTANCE'. Below this is an 'Input method' section with three radio buttons: 'None' (selected), 'Text input', and 'On-screen keyboard'. The 'Text input' option has a small keyboard icon next to it. On the right, there is a large rectangular area showing a remote desktop session. The desktop background is a scenic view of a town with a river and a bridge. Overlaid on this image are large, stylized, semi-transparent letters that spell out 'WORKSHOP' and 'PHYLOGENOMICS'. The top right corner of the remote desktop window shows the text 'No Indicators Fri Jan 18, 11:42'.



Your final task before we get started!

Make sure that typing tilde (~), backslash (\), pipe (|), and carat (^) in the terminal works.

Google search to find these on your computer if you don't know where they are.



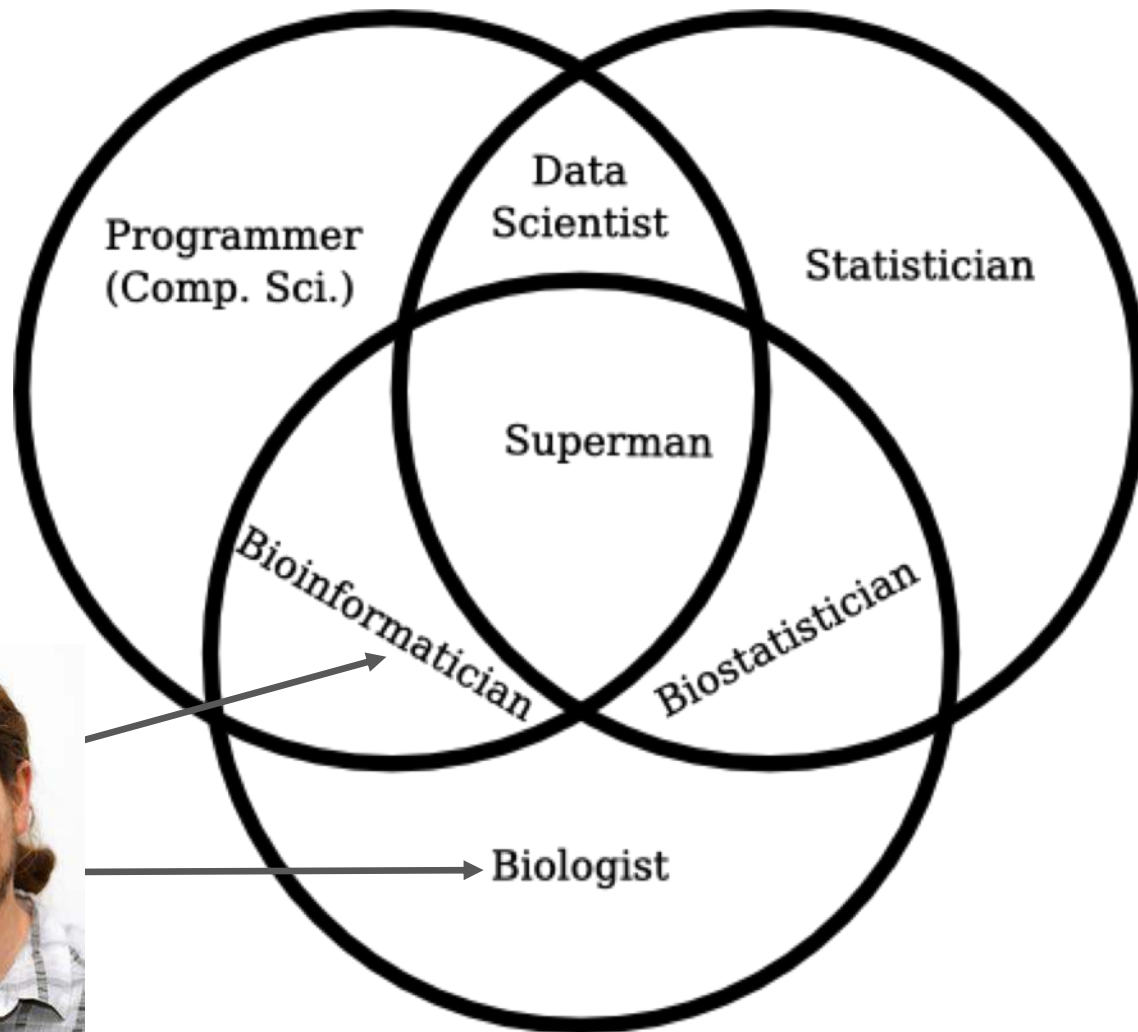


Introduction to Unix

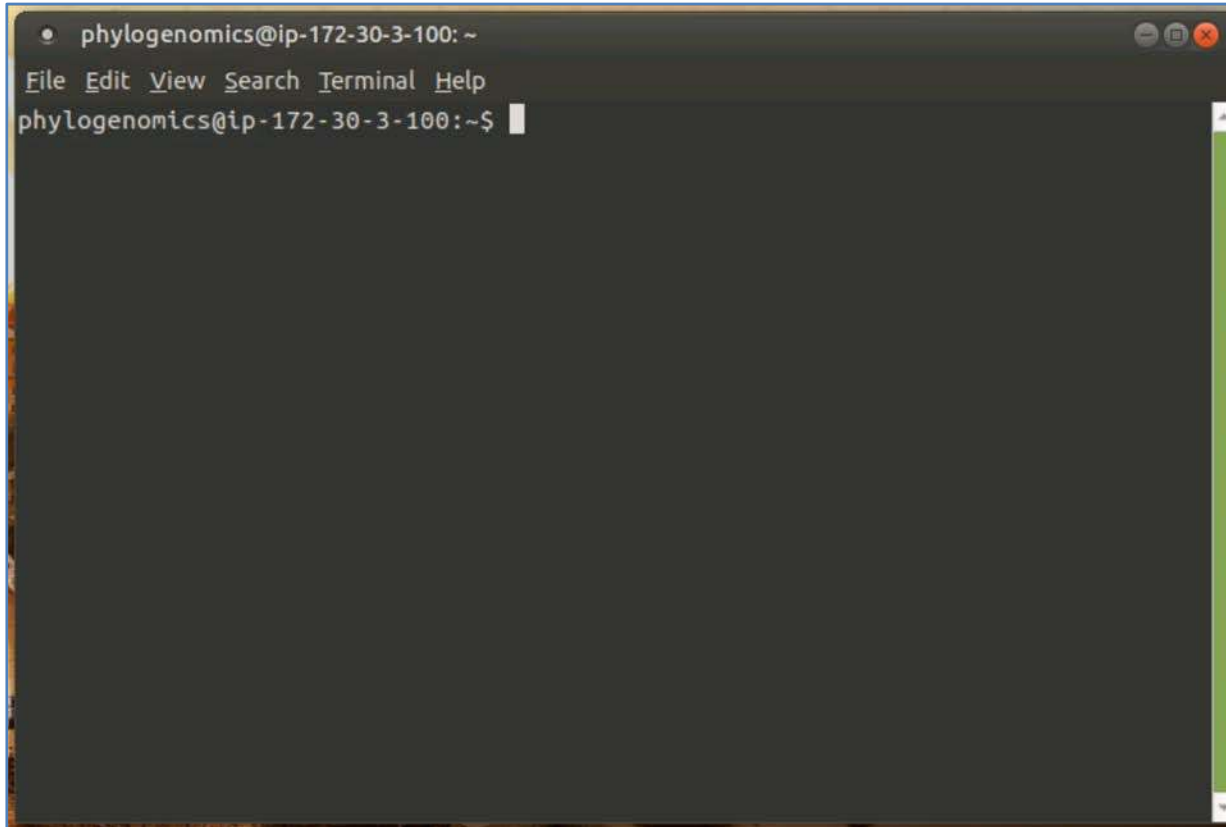
Eric Salomaki



- BA, The Evergreen State College, 2005
- MS, Ohio University, 2012
 - Molecular systematics of freshwater red algae
- PhD, University of Rhode Island, 2017
 - Red algal parasite evolution
- Postdoc, University of Rhode Island, 2017-18
 - Functional ecology of microeukaryotes in the North Atlantic
- Postdoc, Institute of Parasitology, CAS 2018-Present
 - Impact of transitioning to and from parasitism on genome evolution



The Terminal



The Command Line, The Shell, The Prompt

Where you see this "\$" followed by text, I want you to type the text on your command line

Location is Important



First Task – Where am I?

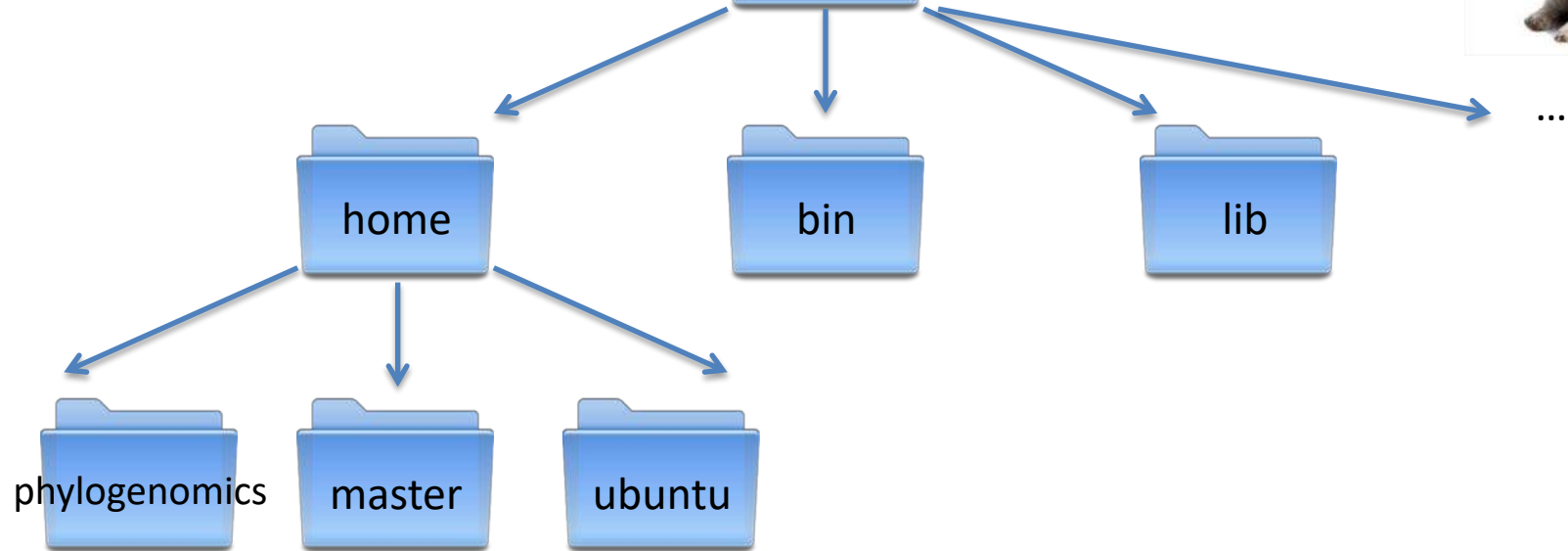
```
~$ pwd
```

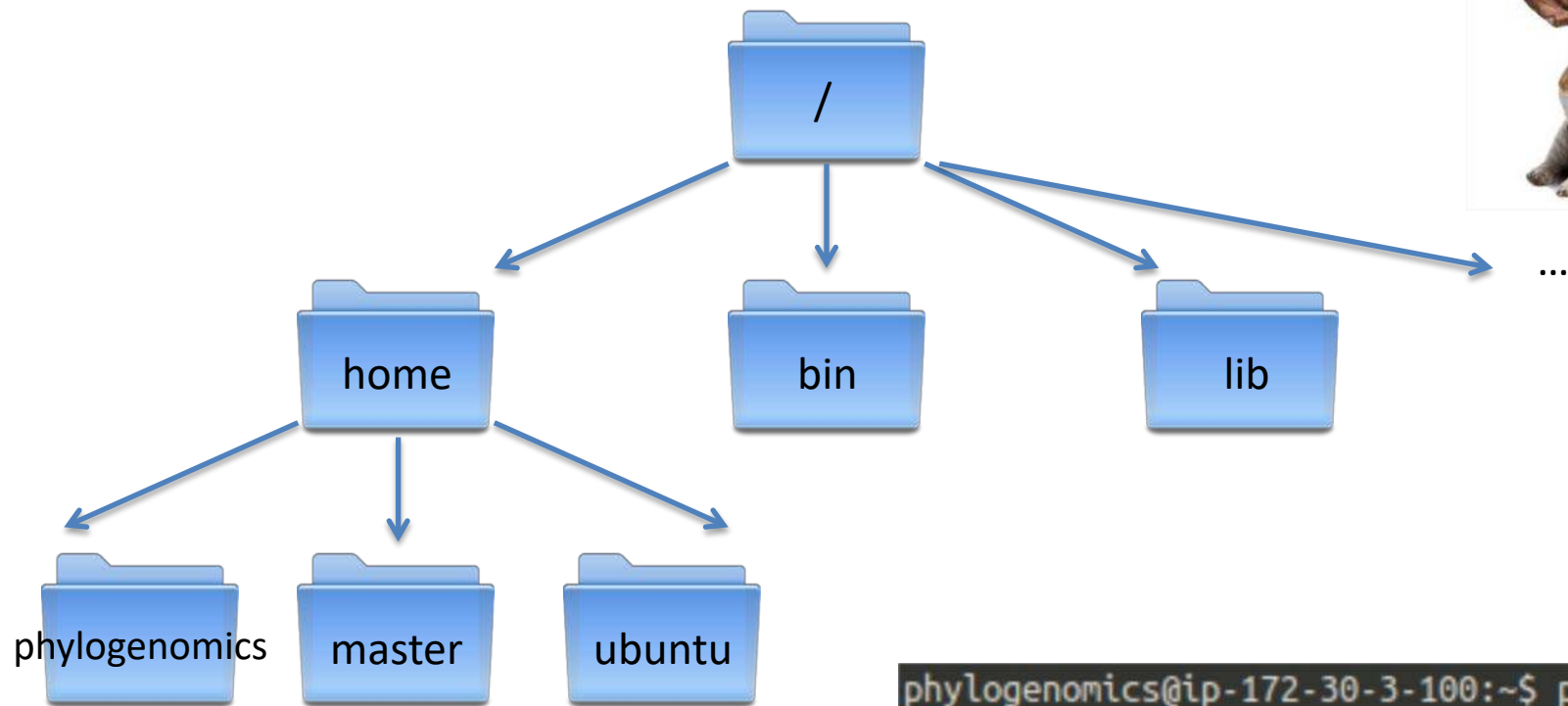
```
phylogenomics@ip-172-30-3-100:~$ pwd  
/home/phylogenomics  
phylogenomics@ip-172-30-3-100:~$
```

This is your “present working directory”.

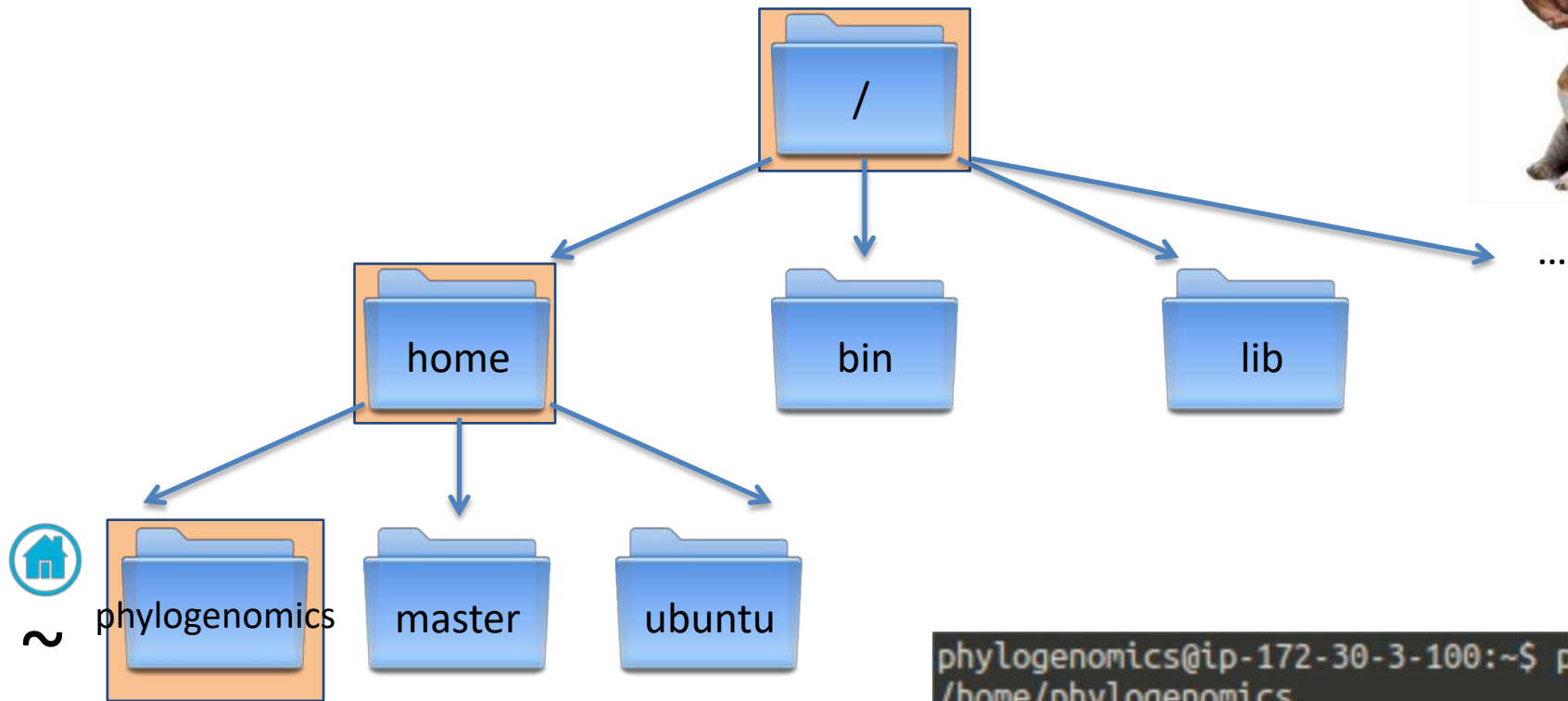


= ROOT





```
phylogenomics@ip-172-30-3-100:~$ pwd  
/home/phylogenomics  
phylogenomics@ip-172-30-3-100:~$
```



```
phylogenomics@ip-172-30-3-100:~$ pwd
/home/phylogenomics
phylogenomics@ip-172-30-3-100:~$
```

This location is also known as your Home Directory

Tilde is shorthand for Home ~

Now let's create some directories and files



Make a directory

```
~$ mkdir ./Data
```

Change into this directory

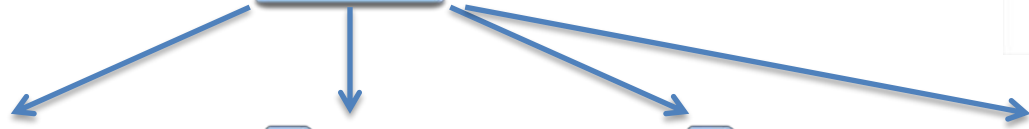
```
~$ cd Data
```

Now what is your present working directory?

NOTE! Directory names (and file names for the matter) can not contain spaces. Underscores are often used instead if you want to separate words.



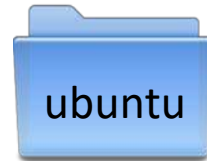
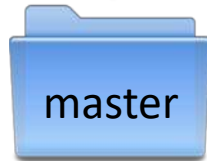
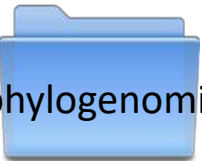
...



master

ubuntu

phylogenomics



Data

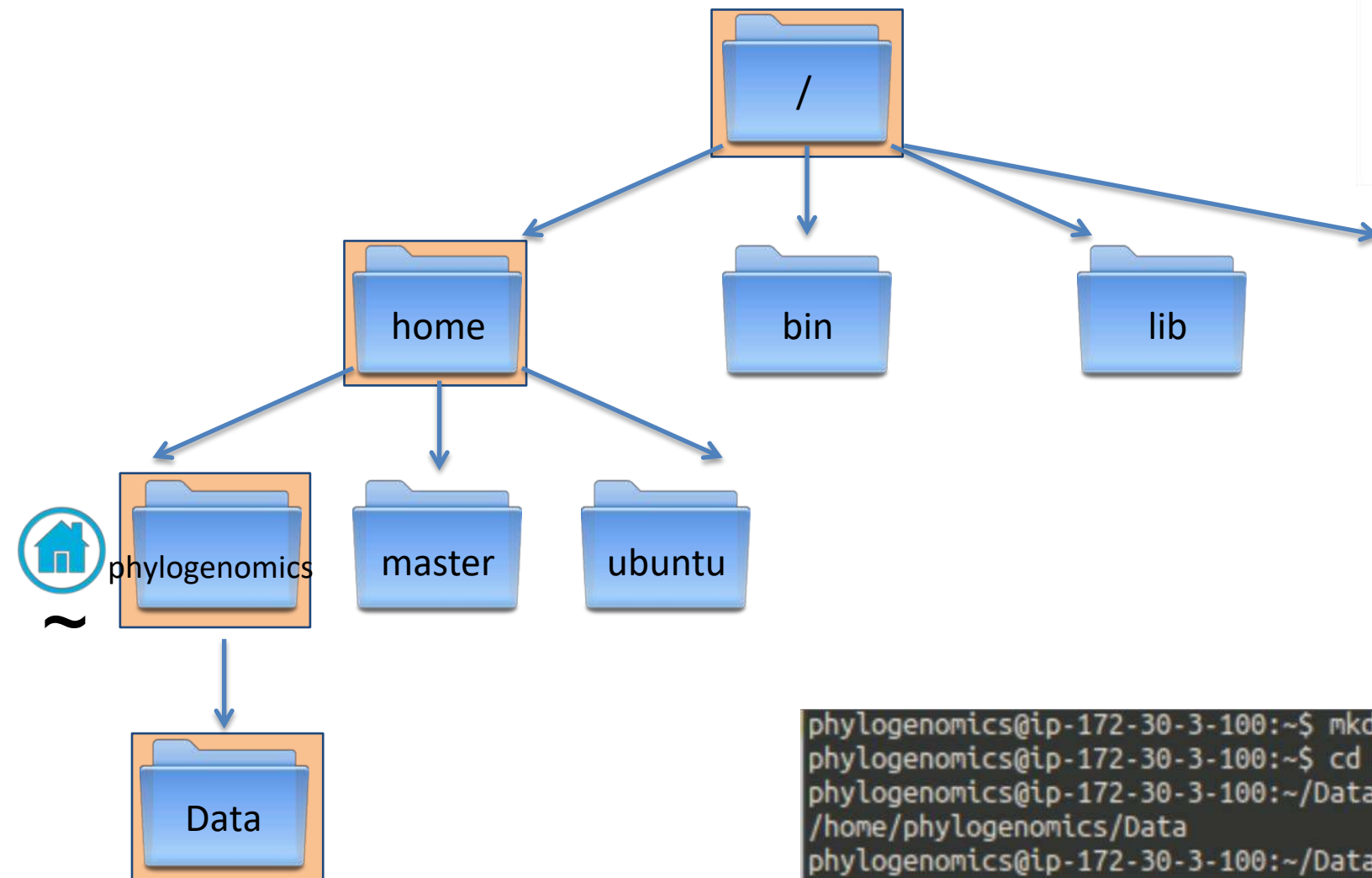


~





...



```
phylogenomics@ip-172-30-3-100:~$ mkdir ./Data
phylogenomics@ip-172-30-3-100:~$ cd Data
phylogenomics@ip-172-30-3-100:~/Data$ pwd
/home/phylogenomics/Data
phylogenomics@ip-172-30-3-100:~/Data$
```

Now let's create some directories and files



Make an empty file

```
~/Data$ touch rags
```

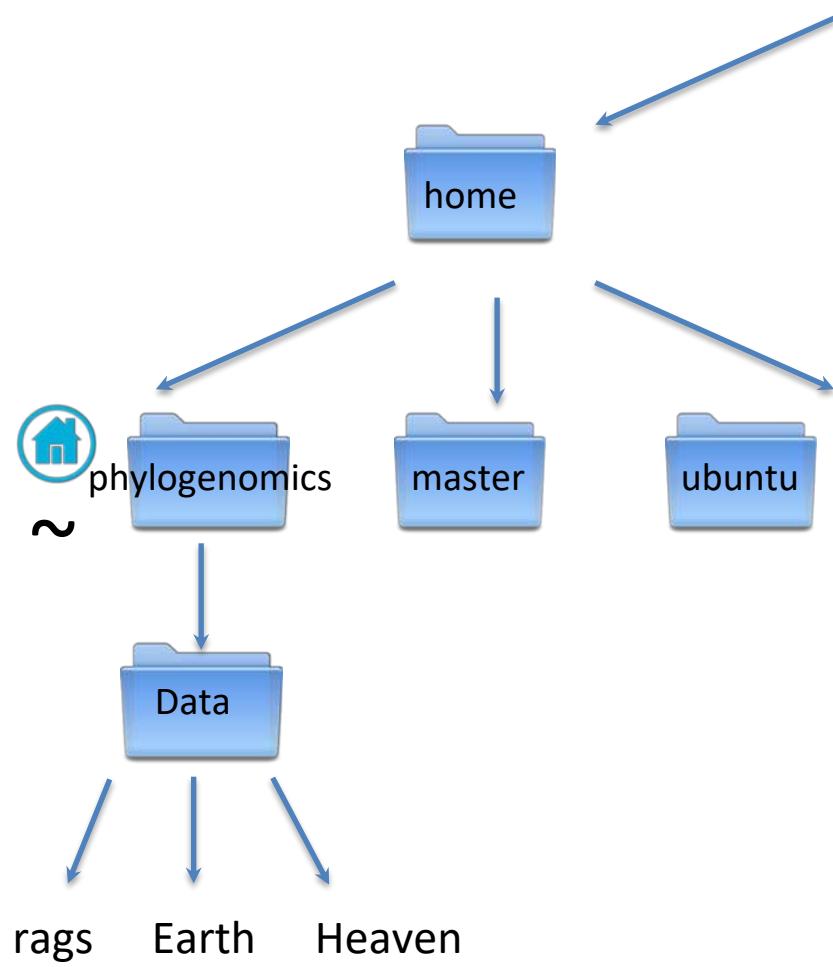
And another two

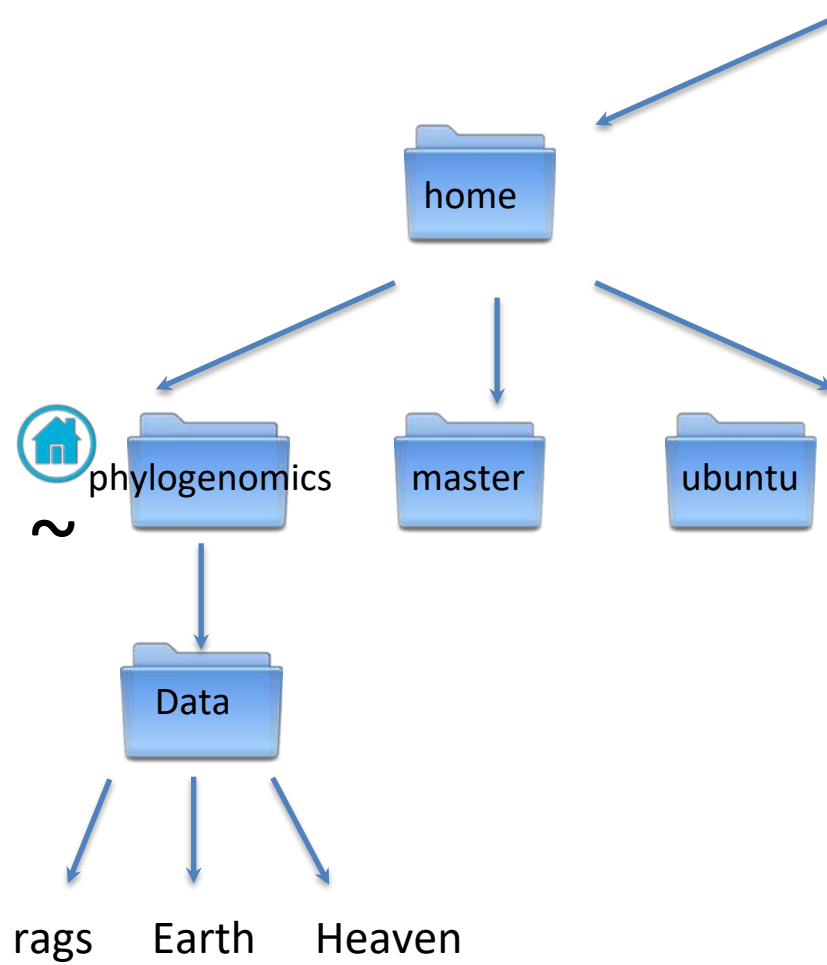
```
~/Data$ touch Heaven Earth
```

Now let's list the contents of the current directory (Data)

```
~/Data$ ls
```

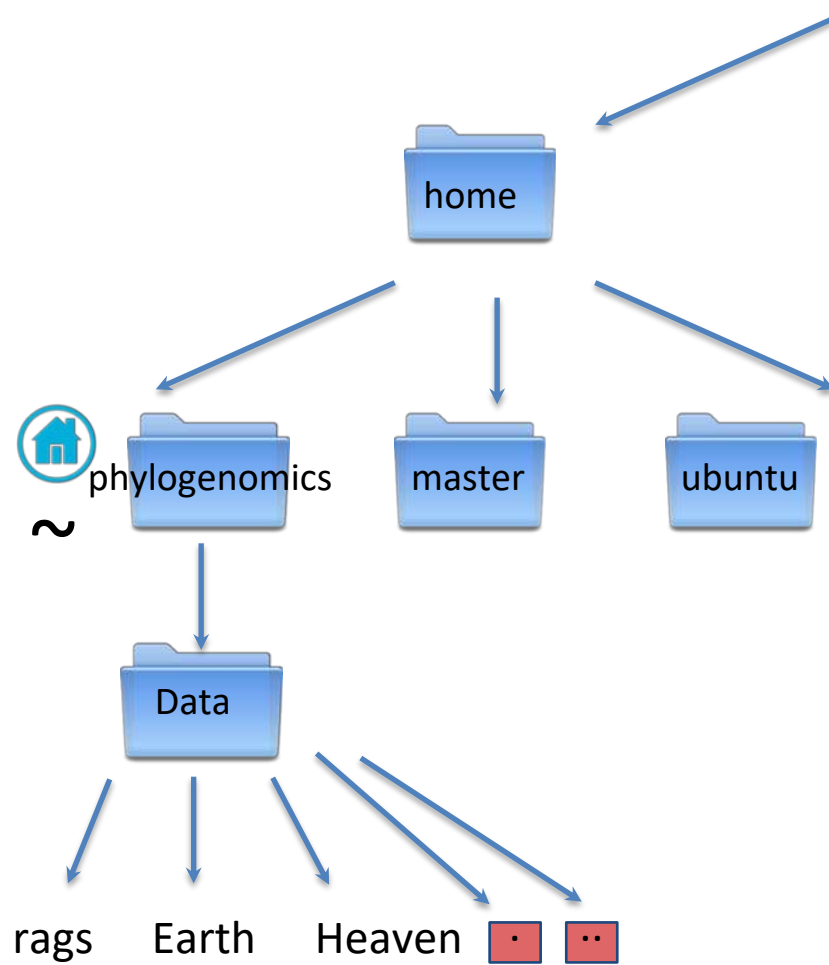
```
phylogenomics@ip-172-30-3-100:~/Data$ touch rags
phylogenomics@ip-172-30-3-100:~/Data$ touch Heaven Earth
phylogenomics@ip-172-30-3-100:~/Data$ ls
Earth Heaven rags
```





Now list ALL of the files

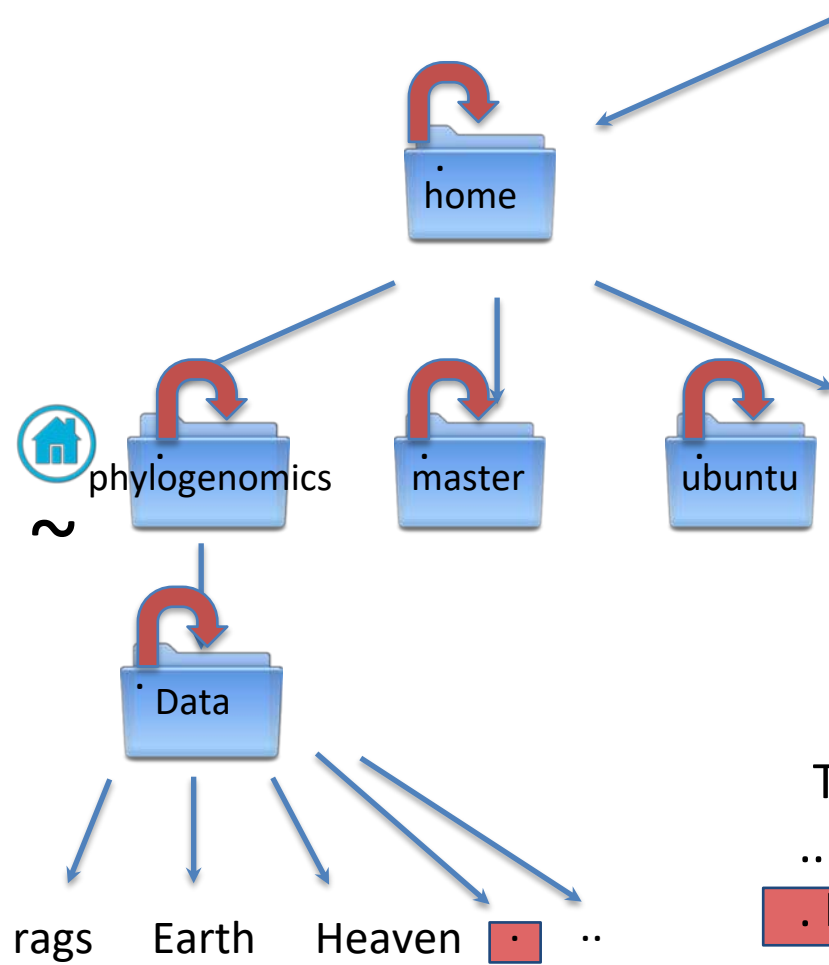
```
~/Data$ ls -a
```



Now list ALL of the files

```
~/Data$ ls -a
```

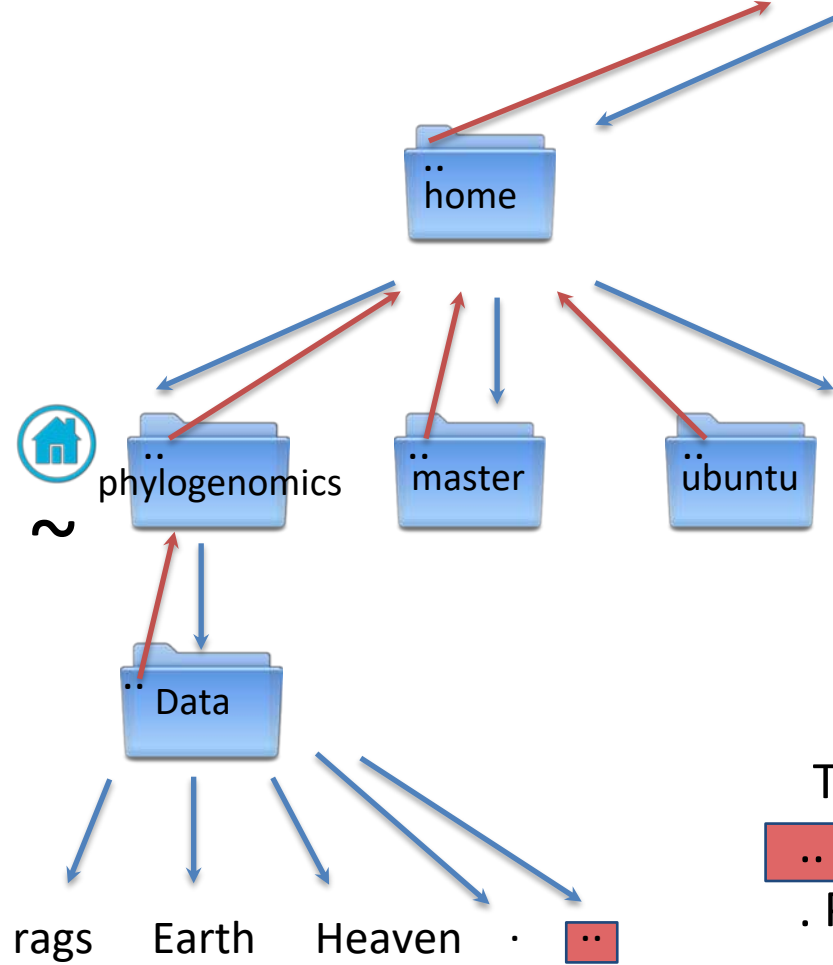
```
phylogenomics@ip-172-30-3-100:~/Data$ ls -a
.  ..  Earth  Heaven  rags
phylogenomics@ip-172-30-3-100:~/Data$
```

These special files are in every directory

.. Points to one directory above

. Points to the current directory



These special files are in every directory

.. Points to one directory above

. Points to the current directory

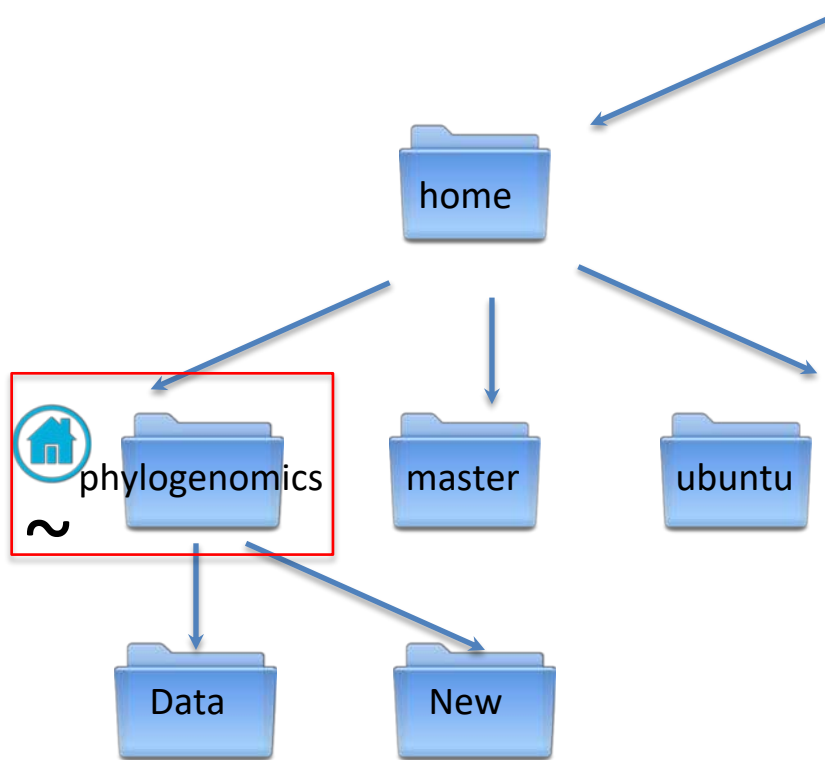


. and .. are used for specifying location

Whenever you do anything on Unix
(move around, move a file, rename a file etc...)
You have to tell the system where that thing is
using a path

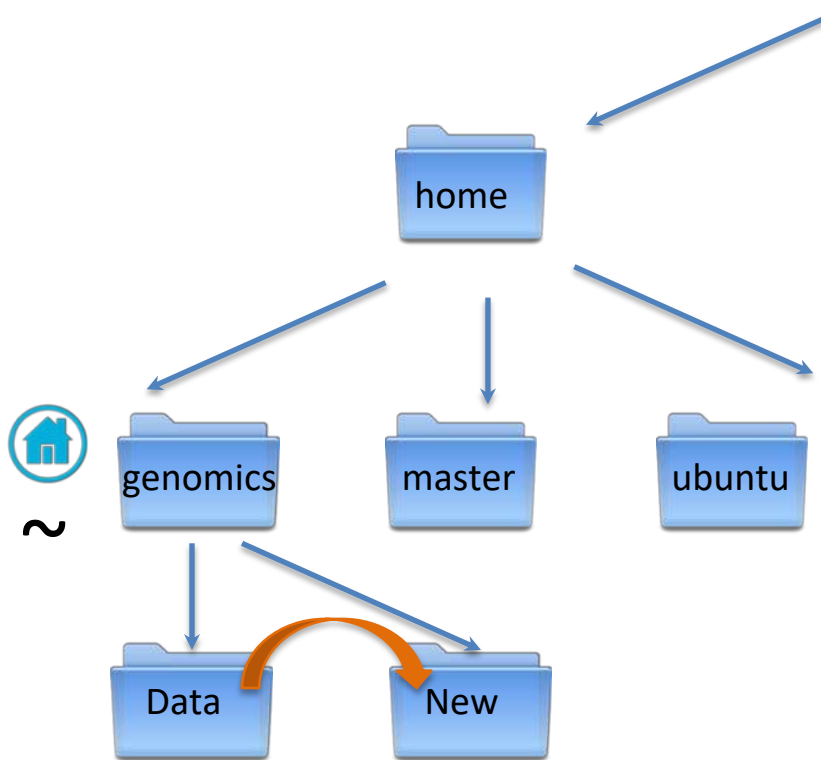
. and .. are part of RELATIVE paths





Create a directory called New
within the phylogenomics
directory using the
RELATIVE PATH

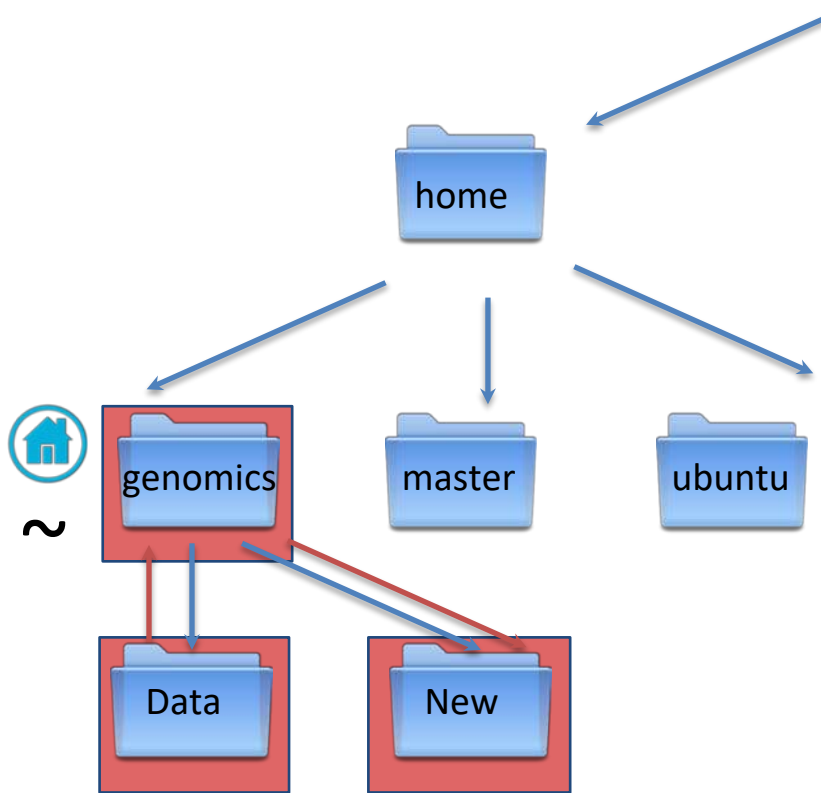
```
~/Data$ mkdir ../New
```



Move from Data to New

RELATIVE PATH

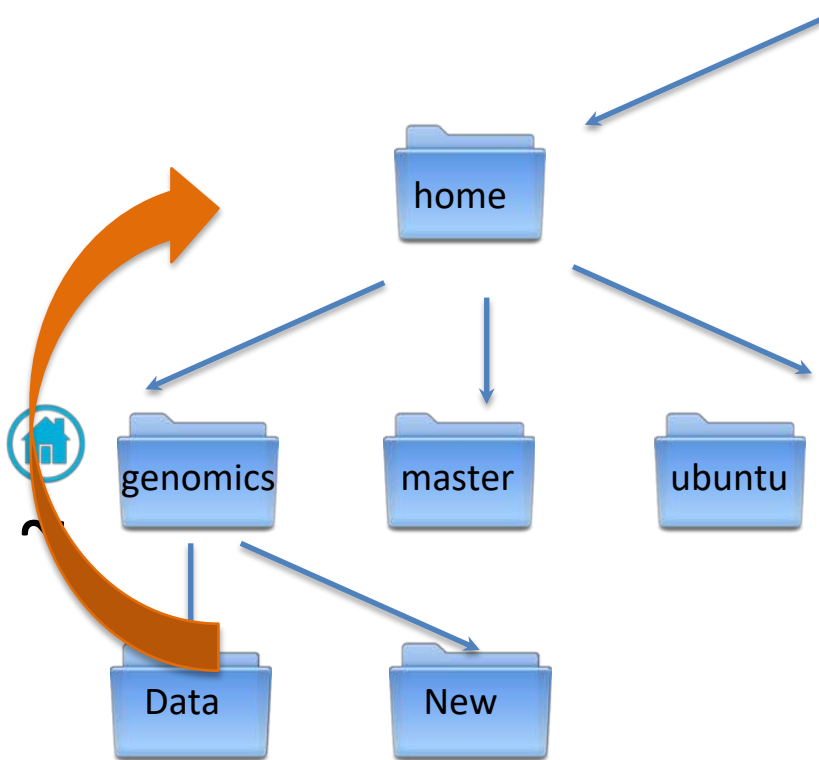
```
~/Data$ cd ../New
```



Moving from Data to New

RELATIVE PATH

```
phylogenomics@ip-172-30-3-100:~/Data$ cd ../New  
phylogenomics@ip-172-30-3-100:~/New$ pwd  
/home/phylogenomics/New
```



Moving from New to home

RELATIVE PATH

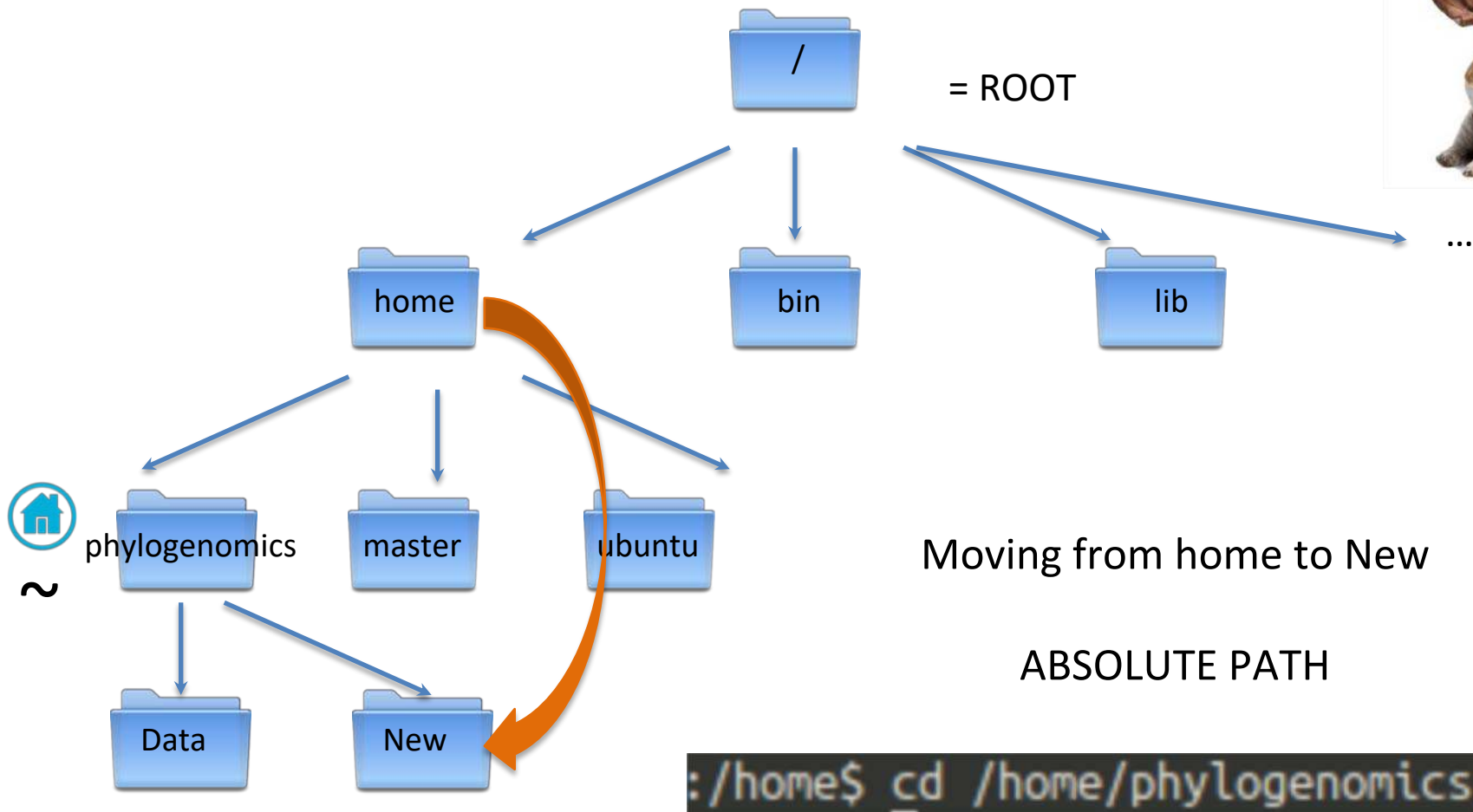
```
:~/New$ cd ../../
```

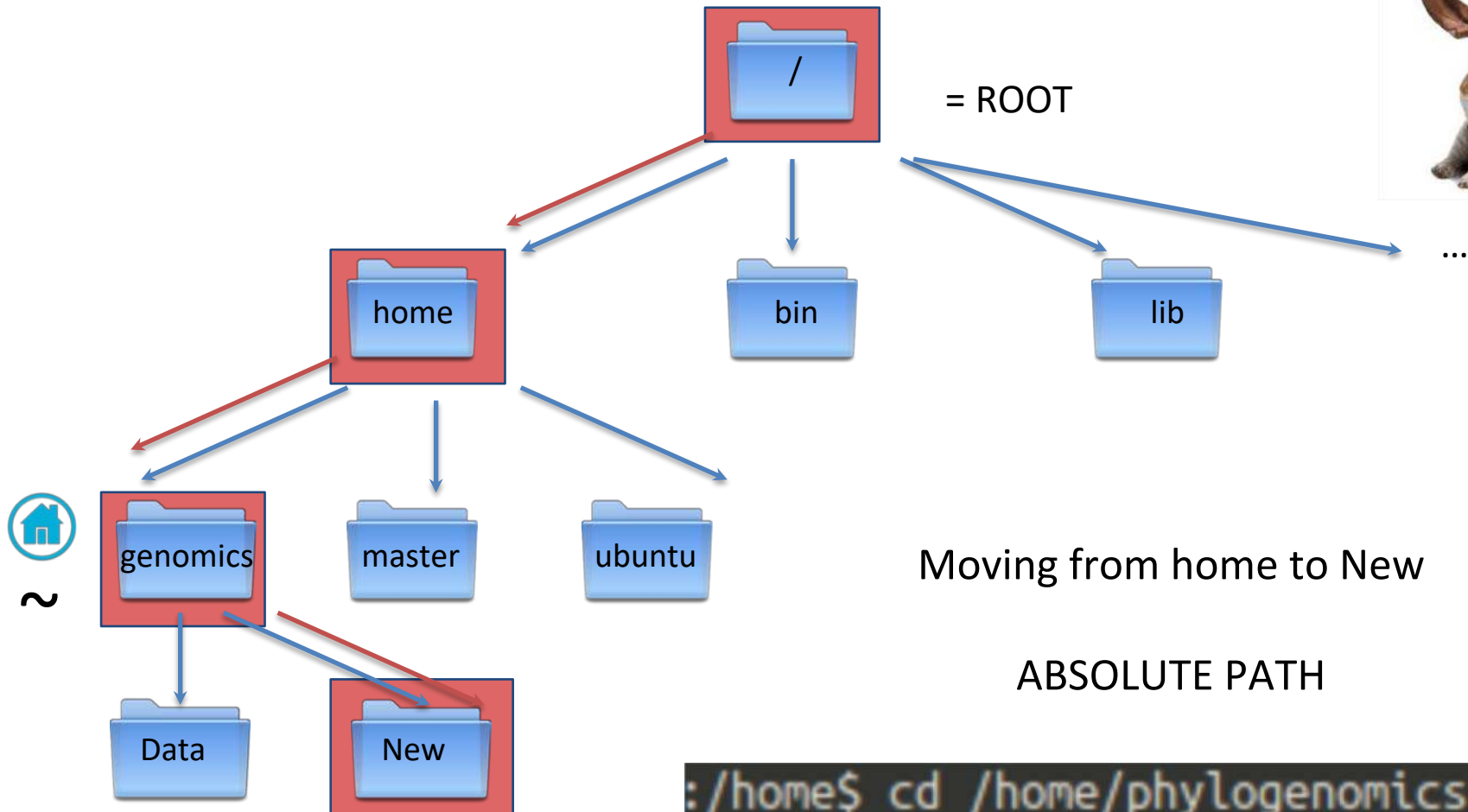



Relative paths will always change depending on your location.

The alternative is ABSOLUTE paths. These always start from root and will never change.



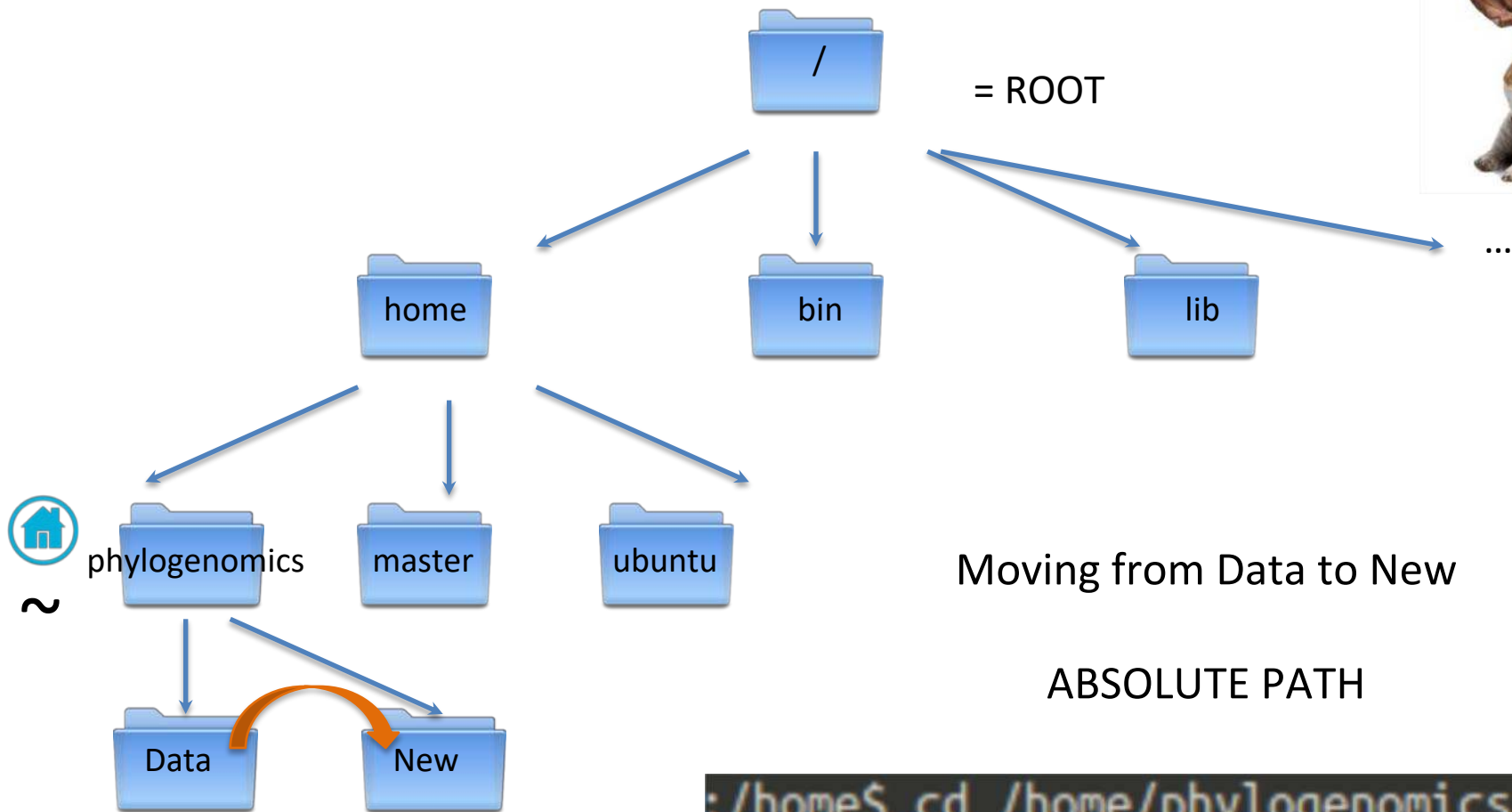




Moving from home to New

ABSOLUTE PATH

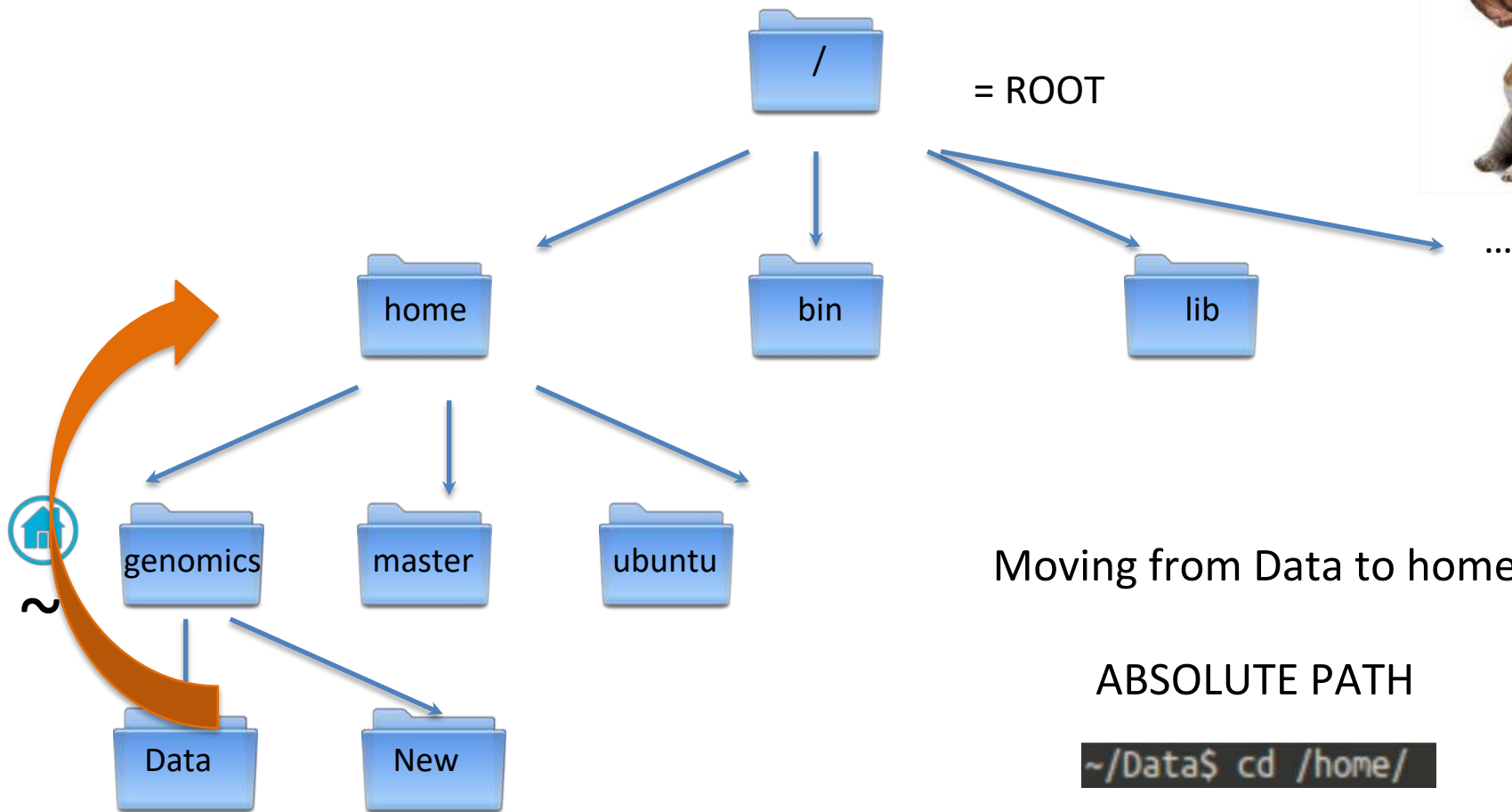
```
:/home$ cd /home/phylogenomics/New/
```

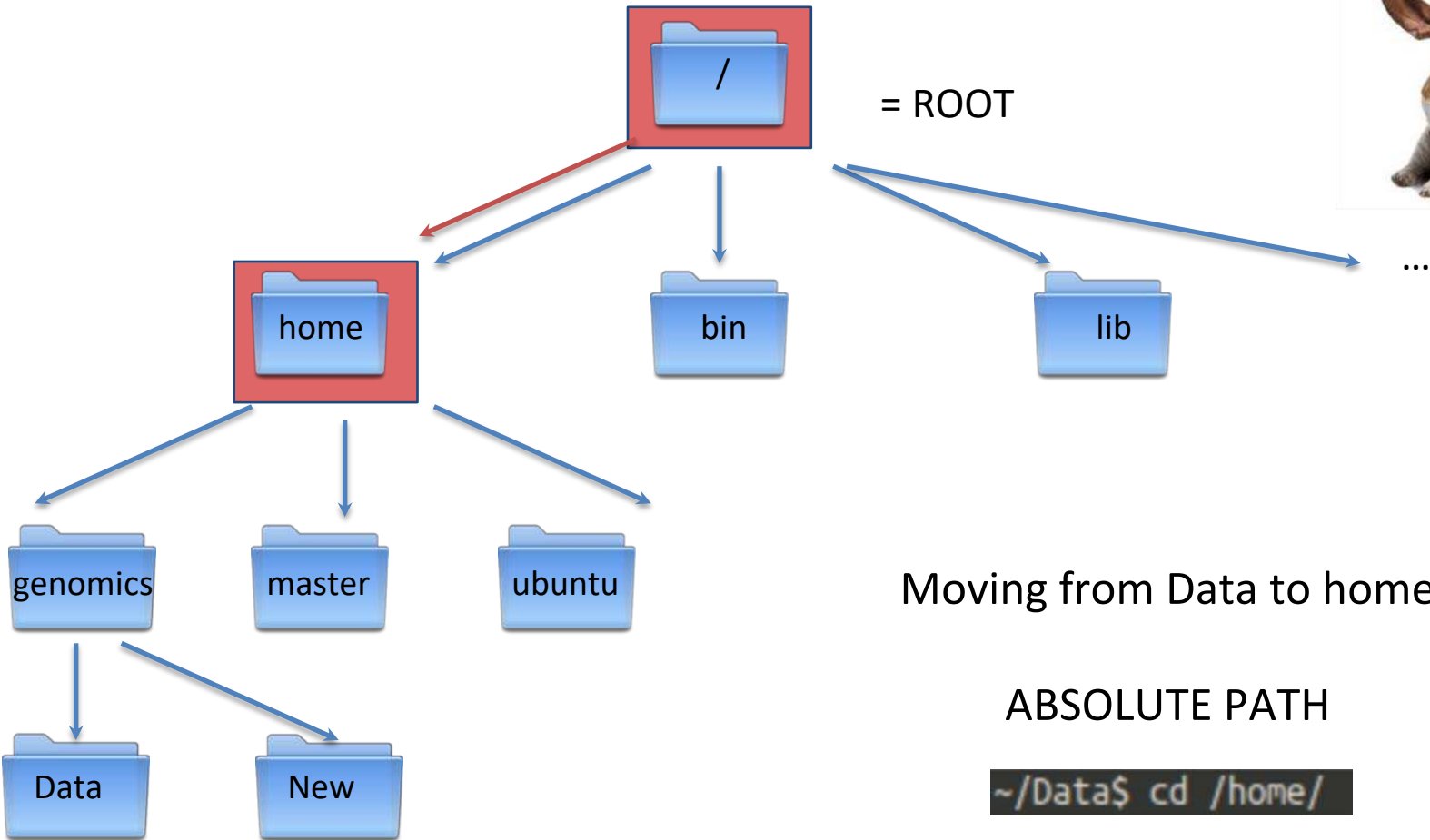


Moving from Data to New

ABSOLUTE PATH

```
:/home$ cd /home/phylogenomics/New/
```





Moving from Data to home

ABSOLUTE PATH

```
~/Data$ cd /home/
```


A Note About Dot!



. means “In your pwd ...”

This command means “List everything that’s in the pwd”

```
$ ls ./
```

This command means “List everything that’s in the pwd within a sub-directory called Data”

```
$ ls ./Data/
```

In most cases, people don’t use ./ at the beginning of a path. As long as the file/directory is within your pwd, the command will work.

```
$ ls ./Data/
```

=

```
$ ls Data
```

Let's put this to practice



Where am I right now? (Should be the Data directory)

```
$ pwd
```

Change to the directory above

```
$ cd ../
```

Let's list the contents of the Data directory

```
$ ls ./Data/
```

CHALLENGE 1!

1. Move into the Data directory and list the contents of your home directory
2. In Data, make a new directory and move into this location
3. From this new directory, move into your home directory IN ONE COMMAND and check your location

Work smarter, not harder!



Tab complete is a nice trick to save you typing paths

For this example we are going to list everything in the directory
/home/phylogenomics/workshop_materials/

Start by typing:

```
~$ ls /
```

Followed by tab twice quickly

```
phylogenomics@ip-172-30-3-100:~$ ls /  
bin    home      lib64      opt        sbin    tmp        vmlinuz.old  
boot   initrd.img lost+found  proc       snap    usr  
dev    initrd.img.old media      root       srv     var  
etc    lib        mnt       run        sys     vmlinuz  
phylogenomics@ip-172-30-3-100:~$
```

This shows the contents of the root directory

Work smarter, not harder!



Now type:

```
$ ls /h
```

Followed by tab once. The path to the /home/ directory has filled in.

```
$ ls /home/
```

Now type:

```
$ ls /home/p
```

Followed by tab once. The path to the /home/phylogenomics/ directory has filled in.

```
$ ls /home/phylogenomics
```

Finally type:

```
$ ls /home/phylogenomics/w
```

Followed by tab once to finish the path, and then enter.
You've now listed that directory contents.

Tab complete will fill in paths, save you time in typing
and prevent typos!

Work smarter, not harder!



Two more tricks for less typing!

* Represents any character

For example:

```
$ ls /home/phylogenomics/*.txt
```

Will list everything in my home directory ending .txt

The up arrow can be used to re-run commands

Press your up arrow and see

If you want all of these commands listed, simply type

```
$ history
```

Any Questions???



Binary programs



These are all programs installed on the Unix machine.

They can be found in /bin

```
$ ls /bin
```

```
phylogenomics@ip-172-30-3-100:~/Data$ ls /bin
bash          bzmre         fgconsole    lesspipe     netcat        ping           stty          uncompress
btrfs         cat           fgrep        ln            netstat       ping4          su            unicode_start
btrfs-debug-tree  chacl        findmnt      loadkeys     networkctl    ping6          sync          vdir
btrfs-find-root  chgrp        fsck.btrfs  login        nisdomainname plymouth       systemctl    wdctl
btrfs-image      chmod        fuser       loginctl     ntfs-3g       ps            systemd      which
btrfs-map-logical chown        fusermount  lowntfs-3g   ntfs-3g.probe pwd           systemd-ask-password whiptail
btrfs-select-super chvt         getfacl     ls           ntfs-3g.probe rbash         systemd-escape ypsdomainname
btrfs-zero-log   cp           grep         lsblk        ntfscluster   readlink      systemd-hwdb  zcat
btrfsck         cpio         gunzip       lsmod        ntfsicmp      red           systemd-inhibit zcmp
btrfstune       dash         gzexe       mkdir        ntfsfallocate rm            systemd-machine-id-setup zdiff
bunzip2         date         gzip        mkfs.btrfs   ntfsfix       rmdir         systemd-notify zegrep
busybox         dd           hciconfig   mknod        ntfsinfo      rnano         systemd-sysusers zfgrep
bzipcat        df           hostname    mktemp       ntfsls        run-parts     systemd-tmpfiles zforce
bzipcmp        dir          ip          more         ntfsmove      sed           systemd-tty-ask-password-agent zgrep
bzdiff         dmesg        journalctl  mount        ntfsrecover   setfacl       tar          zless
bzegrep        dnsdomainname kbd_mode    mountpoint   ntfssecaudit setfont       tempfile     zmore
bzexe          domainname   kill        mt           ntfsstruncate setupcon      touch        znew
bzfgrep        dumpkeys    kmod        mt-gnu       ntfsusermap   sh            true
bzgrep         echo         less        mv           ntfswipe      sh.distrib    udevadm
bzip2          ed           lessecho    nano         open          sleep         unlockmgr_server
bzip2recover    egrep       lessfile    nc           openvt        ss            umount
bzless         false       lesskey     nc.openbsd   pidof         static-sh     uname
```

These include pwd, mkdir, ls ...

Every binary program has a manual



To view the manual page, type man followed by the name of the program

```
$ man <PROGRAM>
```

Open the manual page for ls

```
$ man ls
```

Scroll through (enter) and find the options for:

long listing format (**-l**), human-readable sizes (**-h**) and sort by modification time (**-t**)

Exit the manual page (type q) and give these ls options a go in your Data directory

```
$ ls -l ./Data
```

OR

```
$ ls -t ./Data
```

```
$ ls -l -h ./Data
```

OR

```
$ ls -lh ./Data
```

```
$ ls -l -h -t ./Data
```

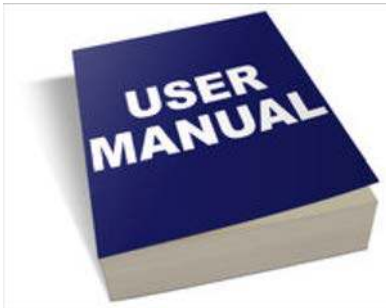
OR

```
$ ls -lht ./Data
```

```
$ ls -l -h -t -a ./Data
```

OR

```
$ ls -lhata ./Data
```



PATH



The computer needs to know where a program is so that it will run

The PATH environment variable is a list of locations your computer looks for programs

You can either provide the path to the program you want to run

```
$ /usr/bin/mkdir
```

PATH



The computer needs to know where a program is so that it will run

The PATH environment variable is a list of locations your computer looks for programs

You can either provide the path to the program you want to run

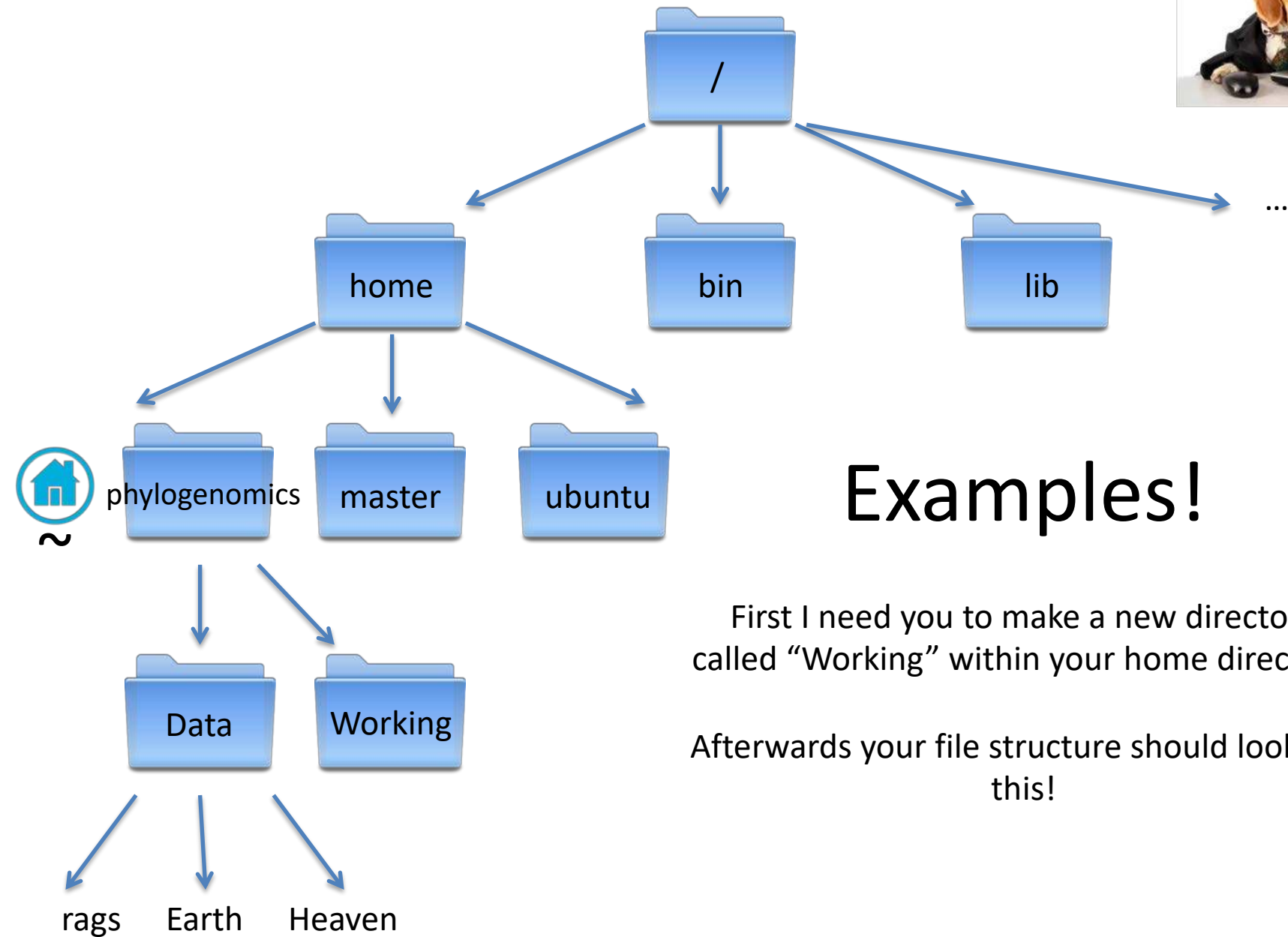
```
$ /usr/bin/mkdir
```

Or make sure the program is in your PATH environment variable

To view locations in your PATH environment variable:

```
$ echo $PATH
```

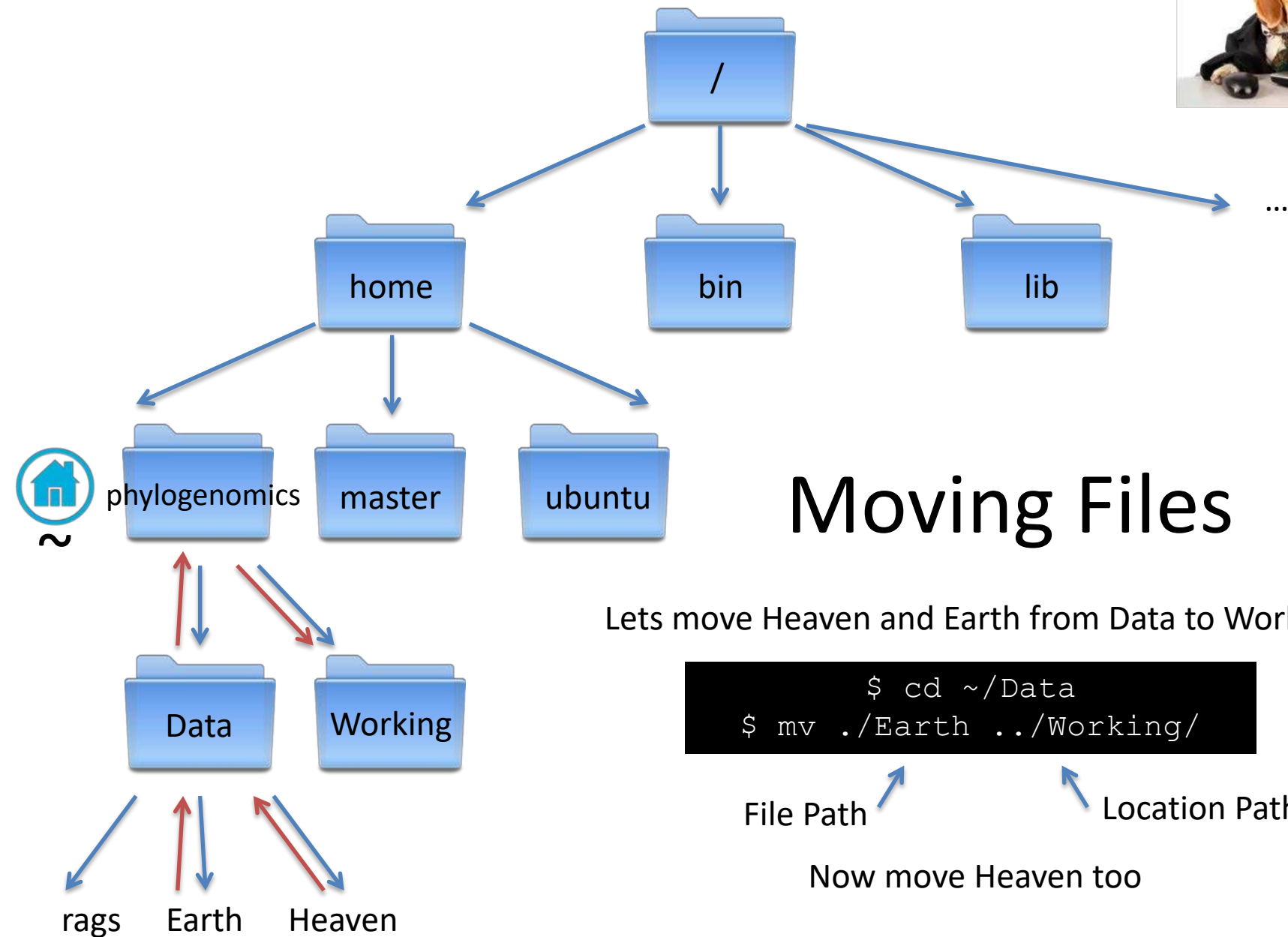
There are ways to add new locations to your PATH, but that is for another time



Examples!

First I need you to make a new directory called “Working” within your home directory.

Afterwards your file structure should look like this!



Moving Files

Lets move Heaven and Earth from Data to Working

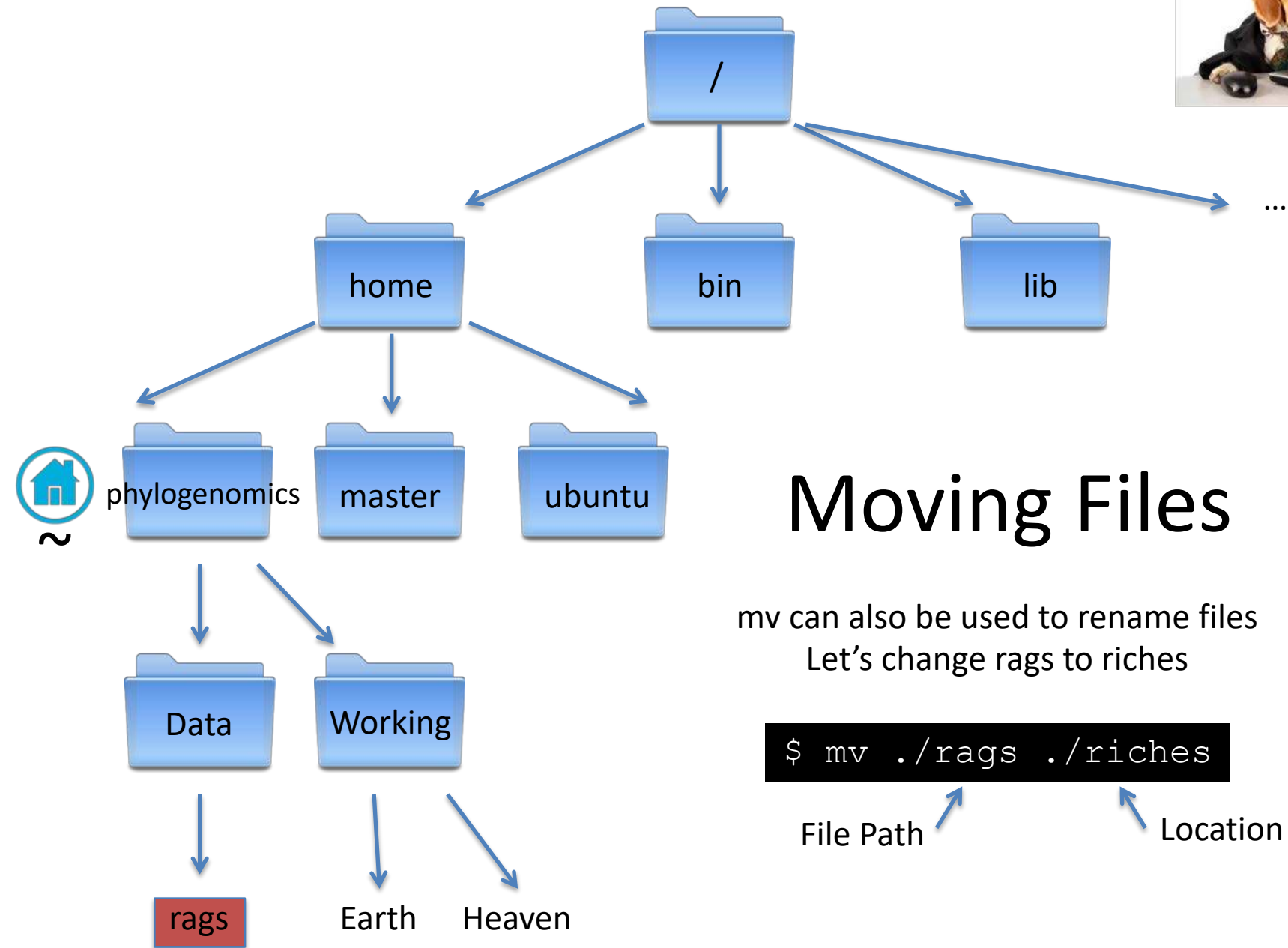
```
$ cd ~/Data  
$ mv ./Earth ../Working/
```

File Path

Location Path

Now move Heaven too

REMEMBER TAB COMPLETE!



Moving Files

`mv` can also be used to rename files
Let's change rags to riches

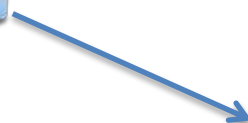
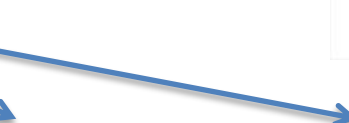
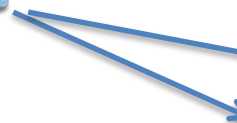
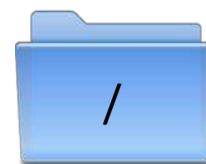
```
$ mv ./rags ./riches
```

File Path

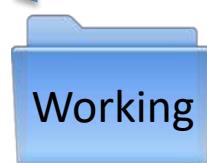
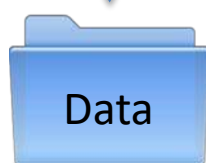
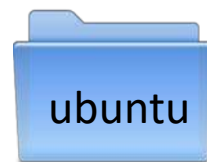
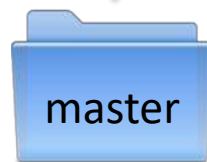
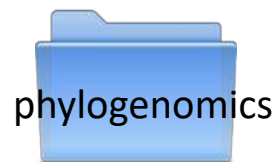
Location Path



...



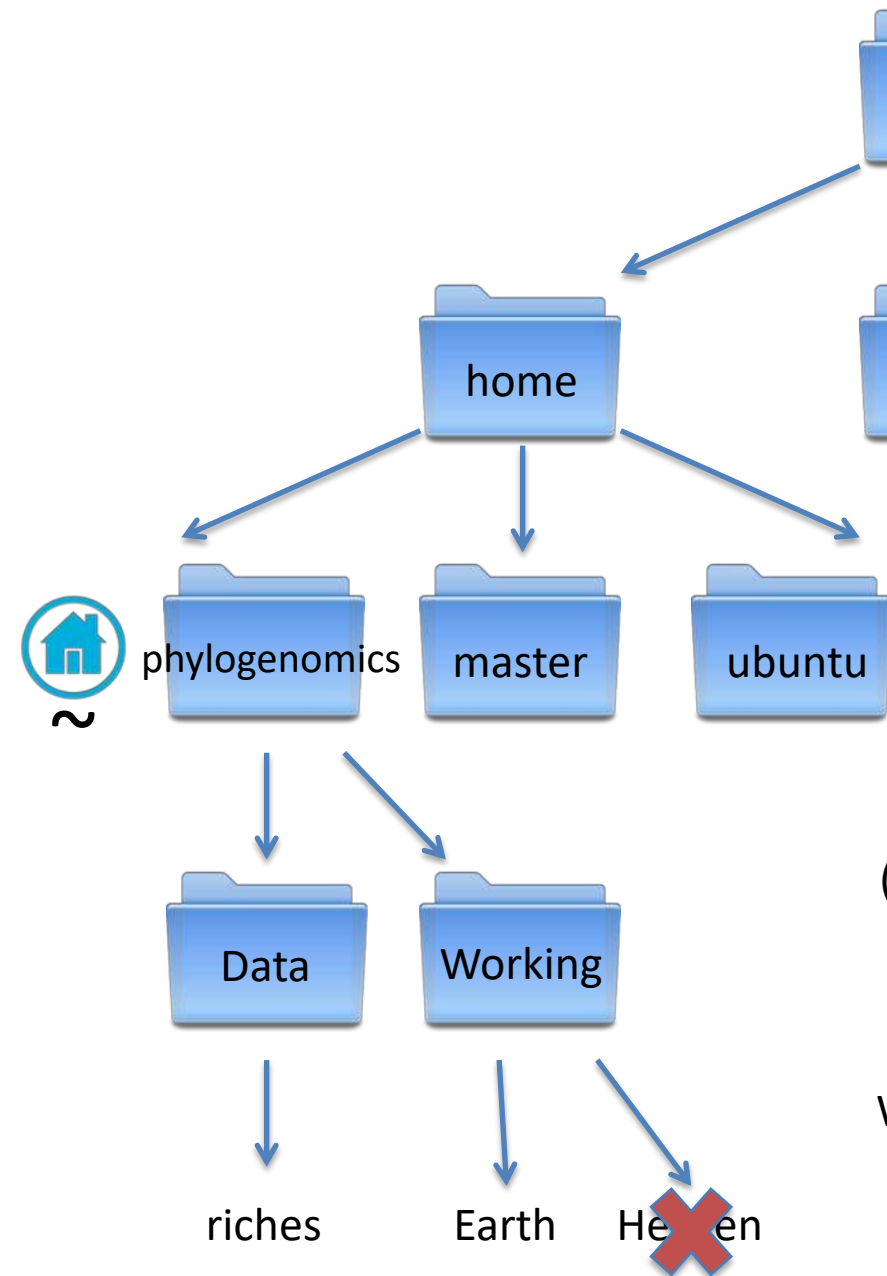
~



riches

Earth

Heaven

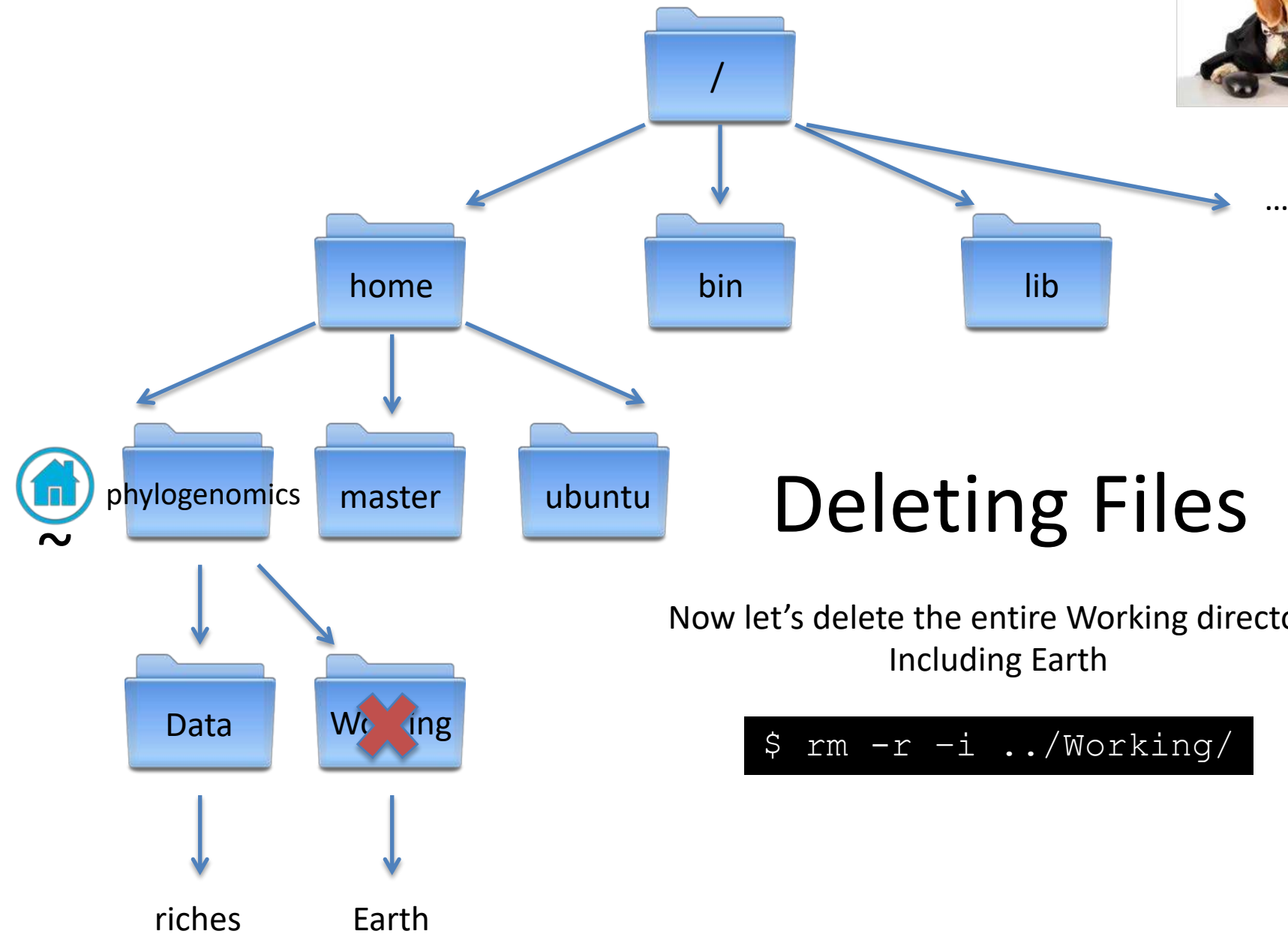


Deleting Files

Now let's delete Heaven
(Check your present working directory is Data)

```
$ rm -i ../Working/Heaven
```

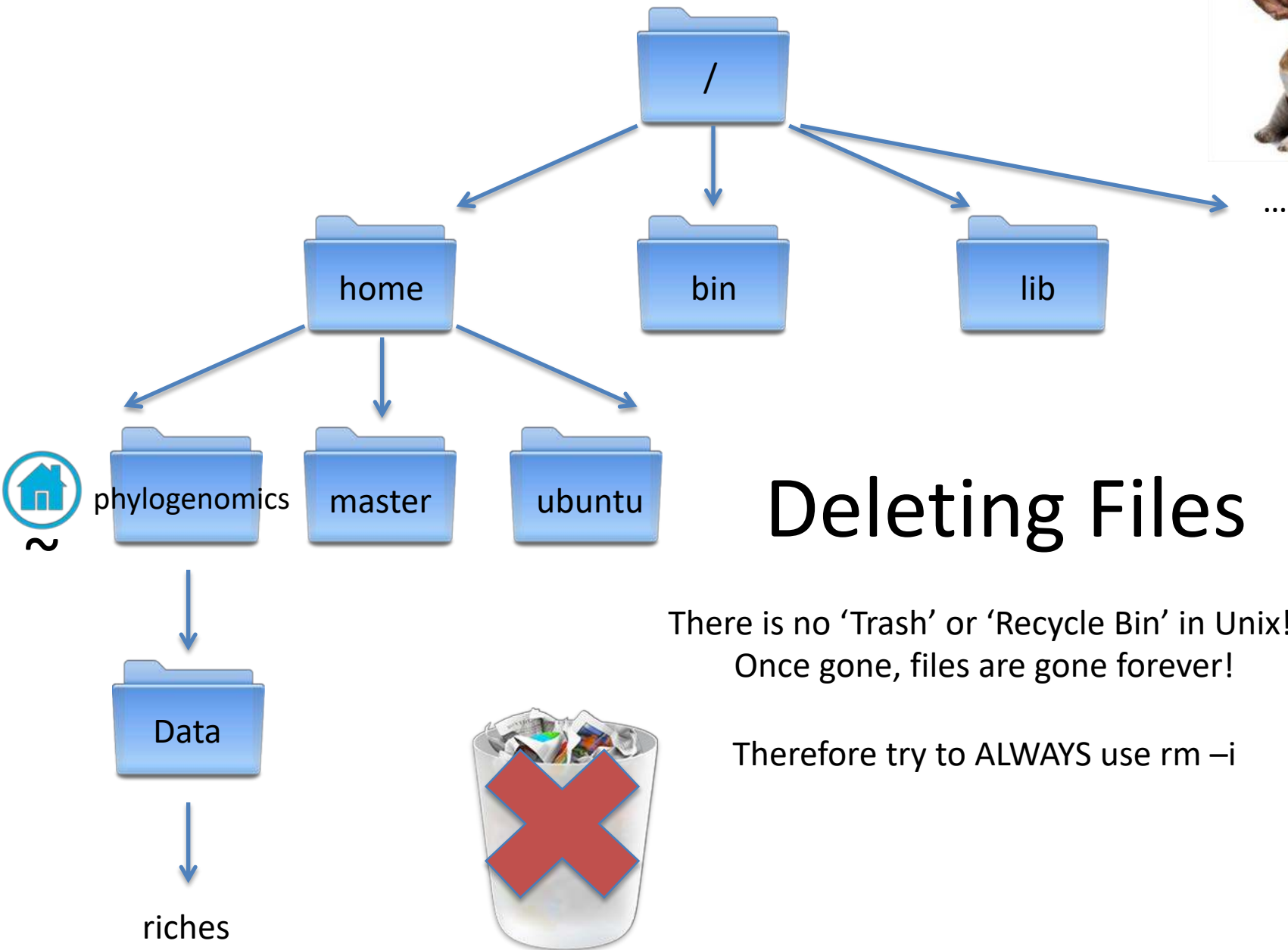
When prompted type y for yes and press enter



Deleting Files

Now let's delete the entire Working directory
Including Earth

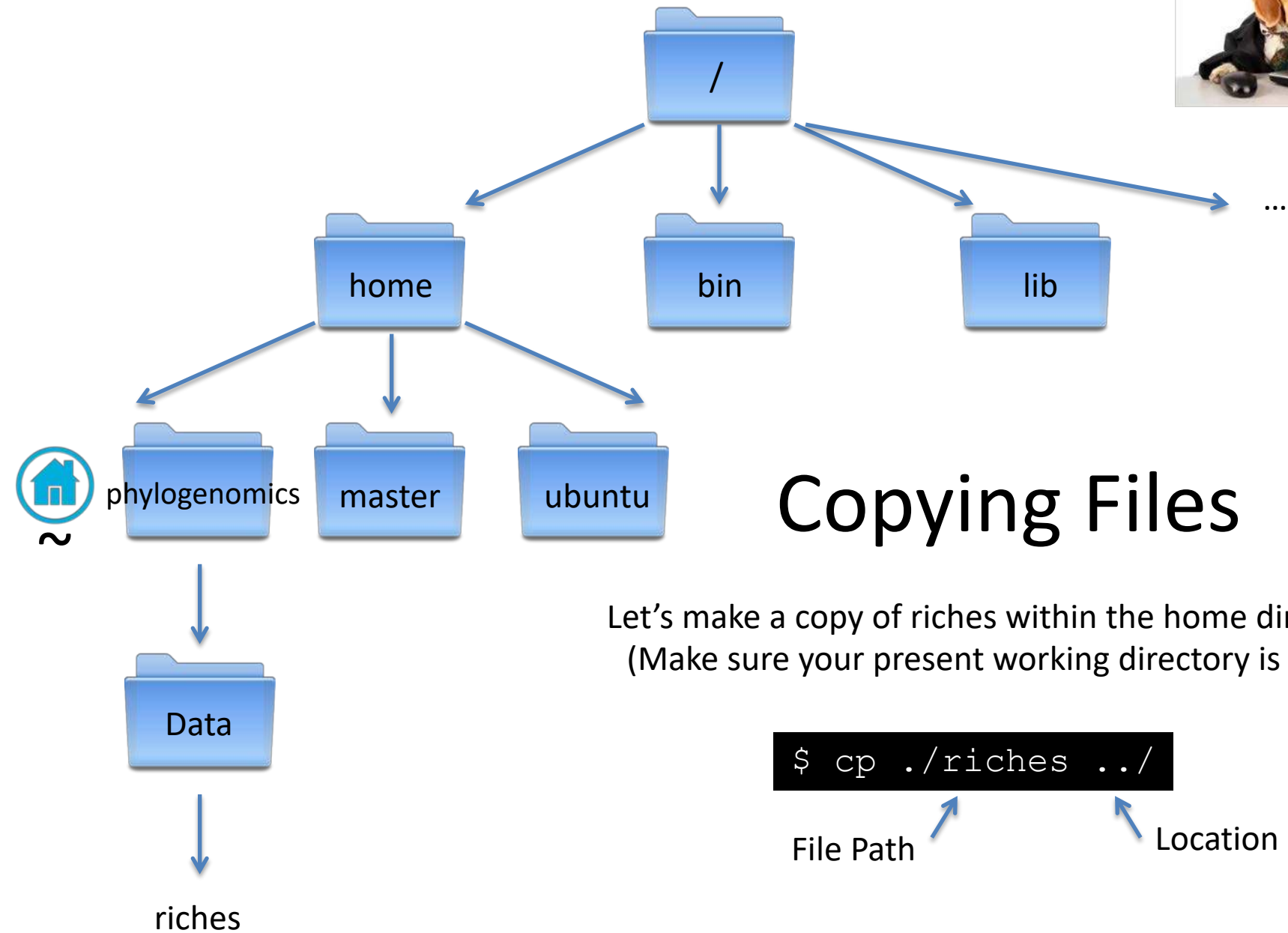
```
$ rm -r -i ../Working/
```

Deleting Files

There is no 'Trash' or 'Recycle Bin' in Unix!
Once gone, files are gone forever!

Therefore try to ALWAYS use `rm -i`



Copying Files

Let's make a copy of riches within the home directory
(Make sure your present working directory is Data)

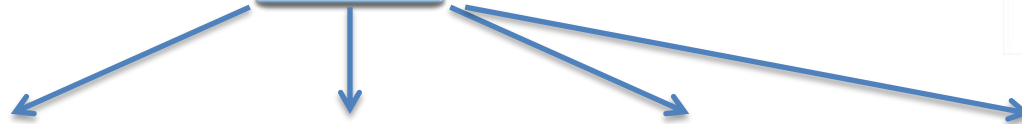
```
$ cp ./riches ../
```

File Path

Location Path



...



master

ubuntu

phylogenomics

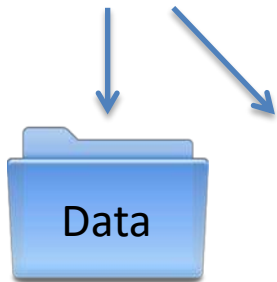
riches

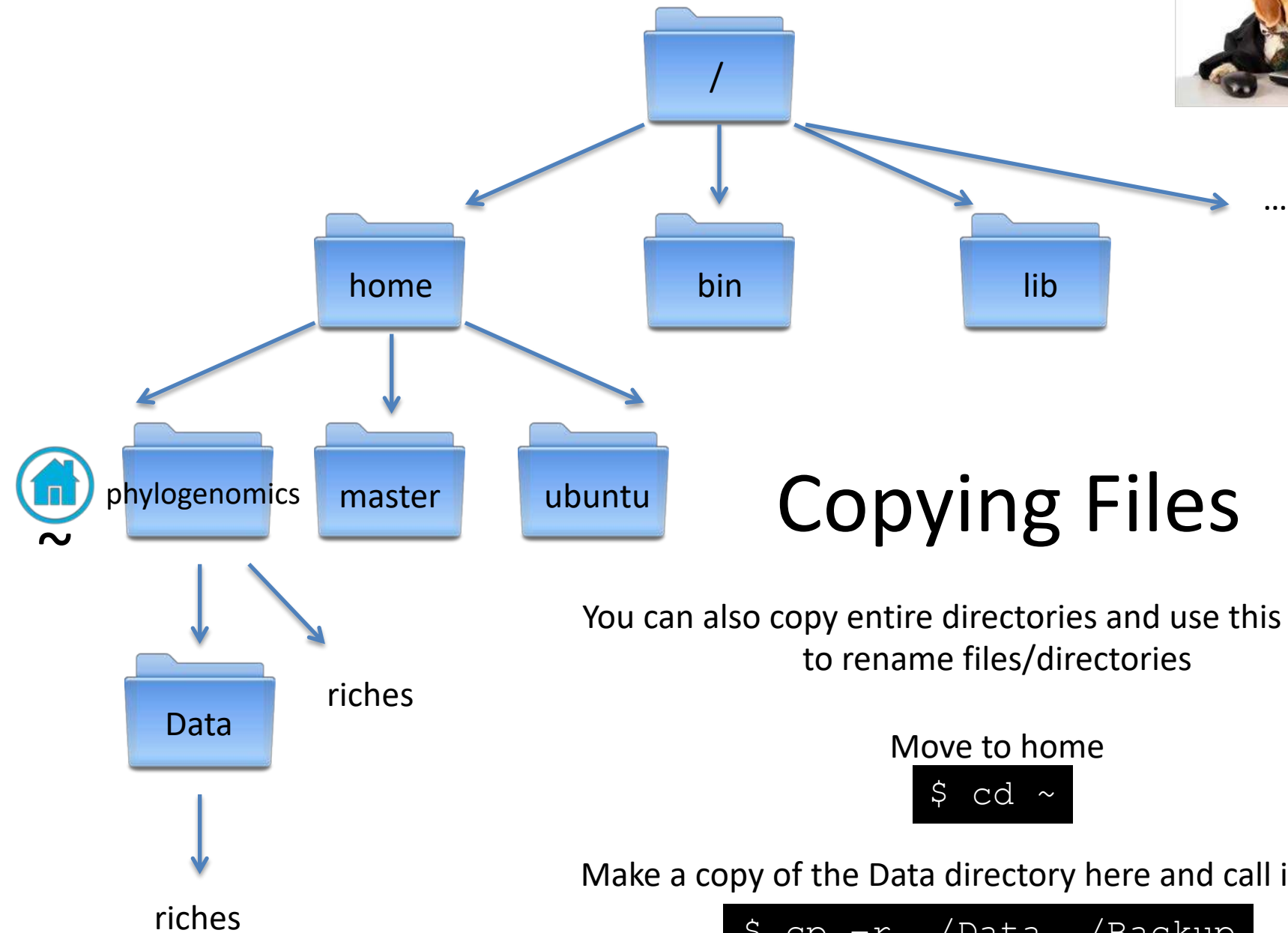
Data

riches



~





Copying Files

You can also copy entire directories and use this function to rename files/directories

Move to home

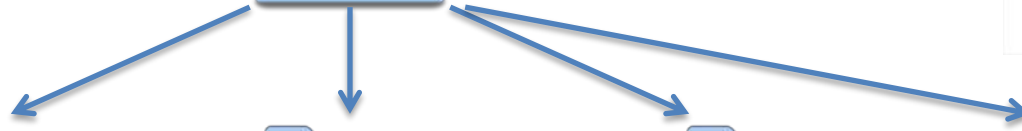
```
$ cd ~
```

Make a copy of the Data directory here and call it Backup

```
$ cp -r ./Data ./Backup
```



...



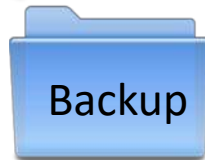
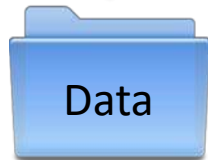
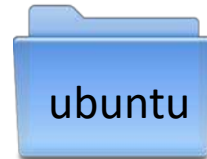
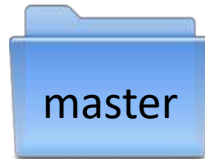
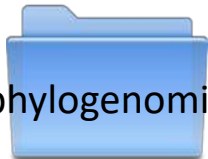
master

ubuntu

phylogenomics



~



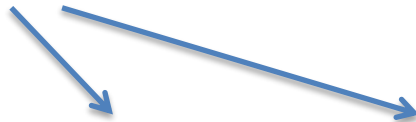
riches

Data

Backup

riches

riches



Typical File Sizes



On the Illumina NextSeq
3,000,000 reads = 1 Gb

But typically your experimental design/sequencing strategies will result
in much more than 3,000,000 reads.

You may have different patients, different locations, replicates etc...

The size of the sequencing data file can easily become 100s of Gb

(or even bigger depending on the sequencer used)

Archived/Compressed Files



Commonly, people will archive directories and compress large files so that they are easier to store or share. Here's an example:

sequences.tar.gz

.tar – means that it is a tape archived directory

.gz – means that it is gzipped file

These can be used alone or in combination

To uncompress

A Tar Archive

```
$ tar -xvf <filename>
```

(x = extract, v = verbose, f = all files)

A Gzipped file

```
$ gunzip <filename>
```

A Gzipped Tar archive

```
$ tar -xzvf <filename>
```

Challenge 2!



1. Change to the workshop_materials directory at the following path:

```
~/workshop_materials/unix
```

You should find a compressed directory:

```
Sequences.tar.gz
```

2. Make a copy of this file in the Backup directory you created earlier

3. Un archive the original directory

4. Unzip the read files

5. Rename the unarchived files – sequence_1.fq and sequence_2.fq

6. Delete the original .tar file

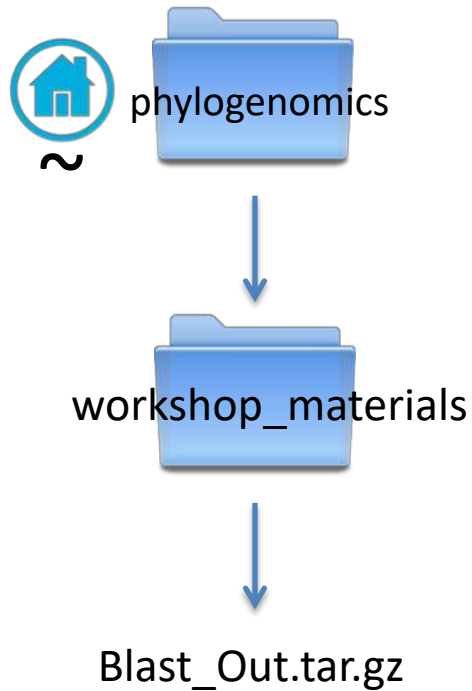
```
tar          gunzip
cp           mv
rm -i       mkdir
cd
```


YEP. IT'S QUESTION TIME



AGAIN

memegenerator.net



Navigate to the workshop_materials directory

```
$ cd ~/workshop_materials
```

Unarchive the Blast_Out.tar.gz

```
$ tar -xzf Blast_Out.tar.gz
```

```
phylogenomics@ip-172-30-3-100:~/workshop_materials$ tar -xzf Blast_Out.tar.gz
Blast_Out/
Blast_Out/LP_blast_seqs.fna
Blast_Out/CP_blast_seqs.fna
phylogenomics@ip-172-30-3-100:~/workshop_materials$
```



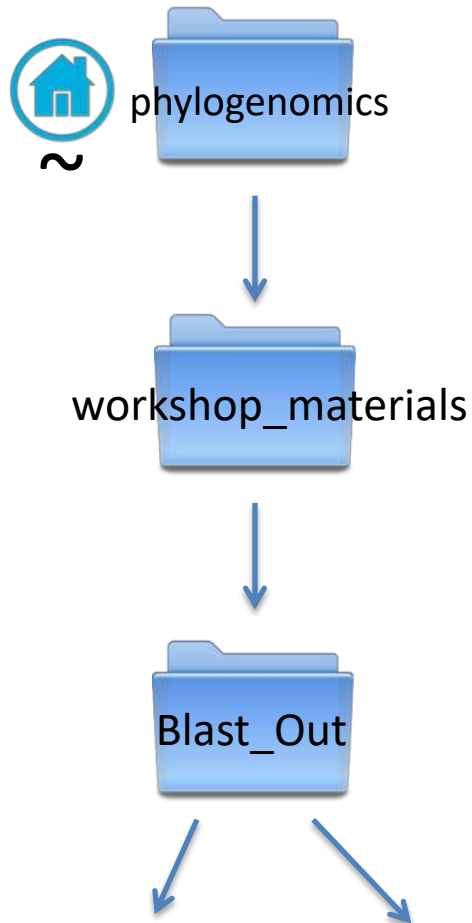
View the first 10 lines of a file

```
$ head CP_blast_seqs.fna
```

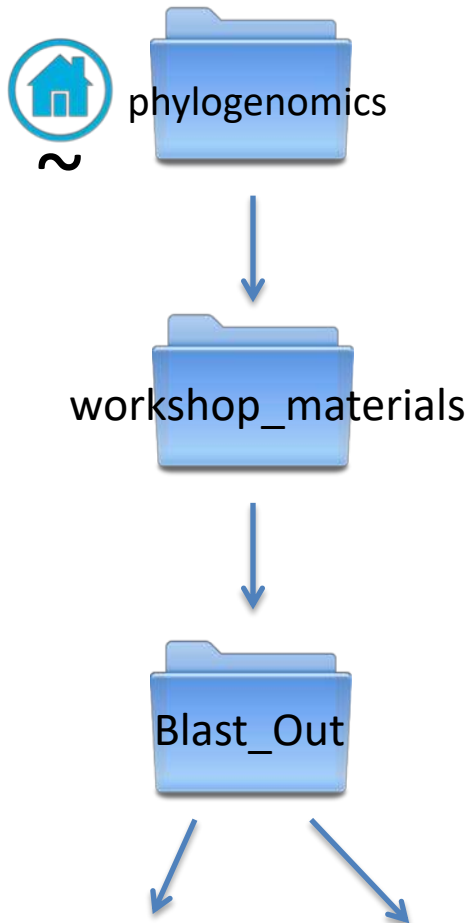
```
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ head CP_blast_seqs.fna
>AY617126.1 Choreocolax polysiphoniae 18S ribosomal RNA gene, partial sequence
CCGAATTCGTCGACAACCTGGTTGATCCTGCCACTGGTATATGCTTGTCTCAAAGACTAAGCCATGCAAGTGCCTGTATG
AGTGTTTATACAACGAACTGCGAATGGCTCGGTAACAGCAATAATTTATTCGGGGTAATACTACTCGGATAACCGT
AGTAAAGCTAGAGCTAATACGTGCAGCGAAGGCGAGAAACGTCTCGTGGTACTAAGTATTAGGTACAAACCAATATTTT
TGGTGATTCATACTTATCTTCTGATCGCAGGAACTGCGACGCCCGAATAAATTTCTGACCTATCAACTATGATGGTAA
GGTAGGGTCTTACCATGGTTGTGACGGGTAACGGACCGTGGGTGCGGGATTCCGGAGAGGGAGCCTGAGAAACGGCTACC
ACATCCAAGGAAGGCAGCAGGCGCGCAACTTACCCAATCCGAACCTCGGGAGGTAGTGACAATACATATCAATAGGGGGA
CCAATGGTCTTTCTAATTGGAATGAATACAATGTAACACATTAATGAGGAACCAGCAGAGGGCAAGTCTGGTGCCAGCA
GCCGCGGTAATTCCAGCTCTGTAAGCGTATACCAAAGTTGTTGCAGTTAAACGCTCGTAGTCGGATTTTGGCGTGTTTC
TGTTGATGCAGCATTTTGCAGTTGACGCAGGAGCCCTTTGTGGAAGGGGTGTGGTGGTACTTTATTGTATTGCTGCAT
```

To view the first 30 lines of the file

```
$ head - 30 CP_blast_seqs.fna
```



CP_Blast_seqs.fna LP_Blast_seqs.fna



View the last 10 lines of a file

```
$ tail CP_blast_seqs.fna
```

```
salami:workshop_files ericsalomaki$ tail CP_blast_seqs.fna
TCGGCACCCTCCGGGAAACCAAAGTGTGTTGCTTTCTGGGGGGAGTATGGTCGCAAGTCTGAAACTTAAAGGAATTGACGG
AAGGGCATCACCGGGTGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGTCCAGACATAATAAGGA
TTGACAGATTGAGAGCTCTTTCTTGATTCTATGGTTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGTGATCTGTCTGGT
TAATTCCGTTAACGAGCGAGACCTGGGCGTGCTAAATAGGGTATGTTATTCTTGTAATAGCAGTTTTGCCCTTCTTAGAC
GGACTGTAGGCTTTTAGTCTATGGAAGCTCCAGGCAATAACAGGTCTGAGATGCCCTTAGATGTTCTGGGCCGCACGCGT
GCTACACTGAACGGTTCAATGGGTGAGGTAGTGCGAAAGCATGACCCAATCTCTAAATCCGTTTCGTGCTGGGGATAGAAG
CTTGCAATTCTCCTTCTTGAACGAGGAATACCTTGTAAGCGCGAGTCATCATCTCGCGCTGAATACGTCCCTGCCCTTG
TACACACCGCCCGTCGCTCCTACCGATTGAGTGGTCCGGTGAGGCTTTGGGAGGGCTGGATGGATTGTGCTTCACAACT
ATTCGGCCTGAACTTAGTCAAACCTTACTACTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCAGA
AGGATCAAGCTAA
```

Now view the last 30 lines of the file

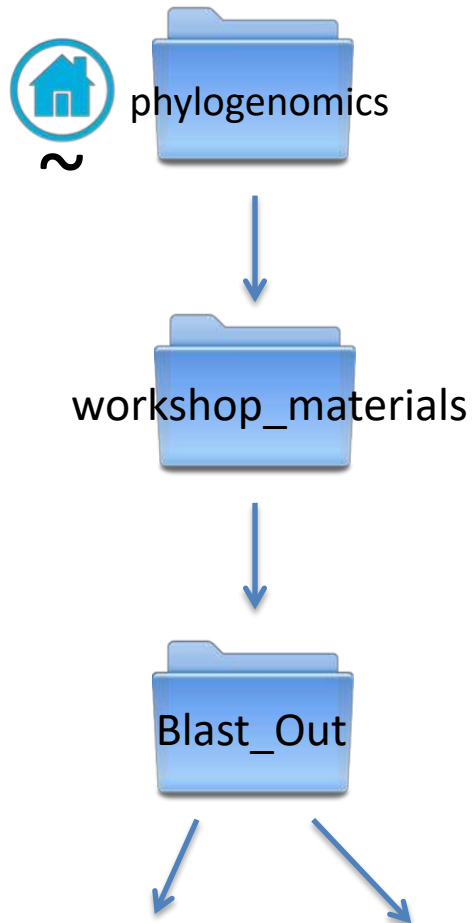
```
$ tail -30 CP_blast_seqs.fna
```

CP_Blast_seqs.fna LP_Blast_seqs.fna



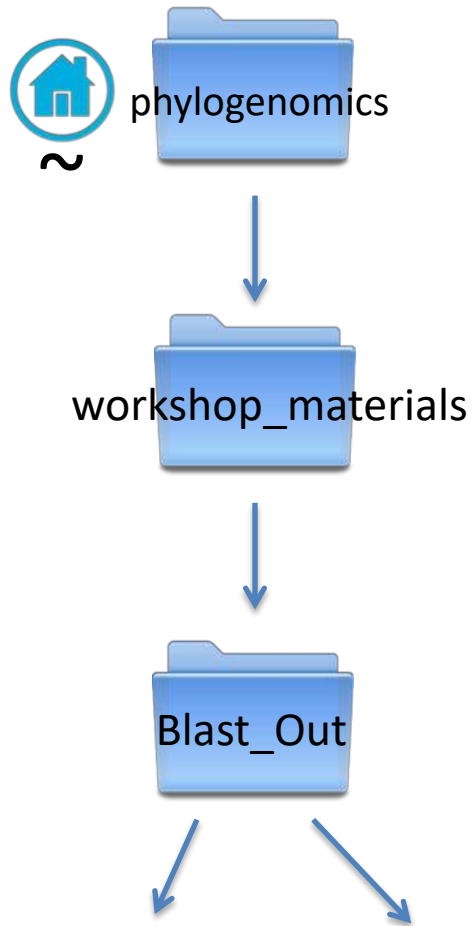
Print the entire file

```
$ cat CP_blast_seqs.fna
```



```
GGTGAAATCTTAGATCCTTGGGAAGACAAACAGCTGCGAAAGCGTCTGCCAAGGACGTTTTTCATTGATCAAGAACGAAAG
TAAGGGGATCGAAGACGATCAGATACCGTCGTAGTCTTTACTATAAACTATGTGATCTGGGGATTAGCCGAGGCCAATTT
GGTTCGGTTAGCACCCCTCCGGGAAACCAAAGATTTTGCATTCTGGGGGAGTATGGTCGCAAGTCTGAAACTTAAAGGAA
TTGACGGAAGGGCATCACCGGTGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACTTACCAGGTCCAGACATA
ATAAGGATTGACAGATTGAGAGCTCTTTCTTGATTCTATGGTTGGTGGTCATGCCCGTTCTTAGTTGGTGGAGTGATCT
GTCTGGTTAATTCCGTTAACGAGCGAGACCTGGGCGTGCTAAATAGGGTTCATTATCTTGAATAGTGGTTCAACTCTT
CTTAGACGGACTGTAGGCTTTTAGTCTATGGAAGCTCCAGGCAATAACAGGTCTGAGATGCCCTTAGATGTTCTGGGCCG
CACGCGTGCTACACTGAACGGTTCAATGGGTGAGGTAGTCCGAAAGCATGACCTAATCTCGAAATCCGTTCTGTCTGGGG
ATAGAAGCTTGCAATTATCCTTCTTGAACGAGGAATACCTTGTAGGCGTGAGTCATCATCTCACGCCGAATACGTCCCTG
CCCTTTGTACACACCGCCCGTCTGCTCCTACCGATTGAGTGTTCCGGTGAGGCTTTGGGAGAACTGGGTGGATTGTGTTCC
ACAACTGCTCGGTTTGAACCTTAGTCAAACCTTATCACTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAAC
CTGCAGAAGCATCAAGCT
>AY237287.1 Osmundaria spiralis 18S ribosomal RNA gene, partial sequence
CCCCACCTGGTTGATCCTGCCAGTGGTATATGCTTGCTCAAAGACTAAGCCATGCAAGTGCCTGTATGAGTGTTTATA
CAACGAACTGCGAATGGCTCGGTAAACAGCAATAATTTCTTCGGGGTAATACTACTCGGATAACCGTAGTAAAGCTA
GAGCTAATACGTGCTACAAAGGCGAGATCGCTCTCGTGGTACTAAGTATTAGGTACAAGCCAGTTTTTCTGGTGATTTCAT
ACTTATCTTCTGATCGCACTCTGTGCGACGCCCGATCAAATTTCTGACCTATCAACTATGATGGTAAGGTAGTGTCTTA
CCATGGTTGTGACGGGTAAACGACCGTGGGTGCGGGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCANNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNACCCAATCCGAACCTTCGGGAGGTAGTGACAAAAAATCAATAGGGGGACCGATGGTCTTC
CTAATTGGAATGAAAACAATGTAAACACCTTATTGAGGAACCAGCAGAGGGCAAGTCTGGTGCCAGCAGCCGCGTAATT
CCAGCTCTGTAAGCGTATACCAAAGTTGTTGCAGTTAAACGCTCGTAGTCGGATTTTGGCGTGCTGCAGAACTCCG
TAAGGAGGTTGATGTAGCTCCGCCTTTGTGGAGGGGGGTGTGGTGGTACTTTATTGTATTGCTGCATTTCCACCAGTGT
TACTGTGAAAAAATTAGAGTGTTCAAAGCGAGCGATTGCCATGAATACATTAGCATGGAATAATAGAATAGGACCTGGTC
CTGTTTTATTGGTTTGTGGTATCAGGTAATGATTAAGAGGGACGGTCGGGGGCAATTTGATTCCAGCGCTAGAGGTGAA
ATTCTTGGATTGCTGGAAGACAAACAGCTGCGAAAGCGTCTGCCAAGGACGTTTTTCATTGATCAAGAACGAAAGTAAAGG
GATCGAAGACGATCAGATACCGTCGTAGTCTTTACTATAAACGATGAGGACTGGGGATCGGATGTATGCTTATGGTACTT
TCGGCACCTCCGGGAAACCAAAGTGTGCTTTCTGGGGGAGTATGGTCGCAAGTCTGAAACTTAAAGGAATTGACGG
AAGGCATCACCGGTGTGGAGCTGCGGCTTAATTTGACTCAACACGGGAAACTTACCAGGTCCAGACATAATAAGGA
TTGACAGATTGAGAGCTCTTTCTTGATTCTATGGTTGGTGGTCATGCCCGTTCTTAGTTGGTGGAGTGATCTGTCTGGT
TAATTCGTTAACGAGCGAGACCTGGGCGTGCTAAATAGGGTATGTTATTCTTGAATAGCAGTTTTTGCCCTTCTTAGAC
GGACTGTAGGCTTTTAGTCTATGGAAGCTCCAGGCAATAACAGGTCTGAGATGCCCTTAGATGTTCTGGGCCGACGCGT
GCTACACTGAACGGTTCAATGGGTGAGGTAGTGCGAAAGCATGACCAATCTCAAATCCGTTCTGCTGCTGGGGATAGAAG
CTTGCAATTCTCCTTCTTGAACGAGGAATACCTTGAAGCGCGAGTCATCATCTCGCGCTGAATACGTCCCTGCCCTTTG
TACACACCGCCCGTCTGCTCCTACCGATTGAGTGGTCCGGTGAGGCTTTGGGAGGGCTGGATGGATTGTGCTTCACAACT
ATTGCGCTGAACCTTAGTCAAACCTTACTACTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCAGA
AGGATCAAGCTAA
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$
```

CP_Blast_seqs.fna LP_Blast_seqs.fna



Many files are too long to meaningfully view in terminal or to edit in a unix text editor

CP_Blast_seqs.fna LP_Blast_seqs.fna



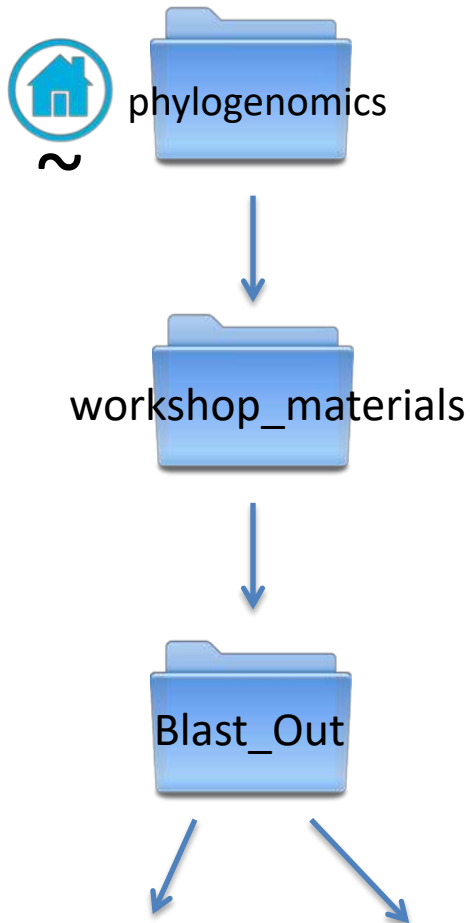
'cat' can also combine multiple files

```
$ cat CP_blast_seqs.fna LP_blast_seqs.fna > All_blast_seqs.fna
```

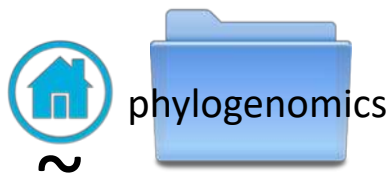
And then count the number of lines in each fasta file
to confirm they have been combined

```
$ wc -l *.fna
```

```
[salami:workshop_files ericsalomaki$ wc -l *.fna
4703 All_blast_seqs.fna
2369 CP_blast_seqs.fna
2334 LP_blast_seqs.fna
9406 total
```



CP_Blast_seqs.fna LP_Blast_seqs.fna



workshop_materials

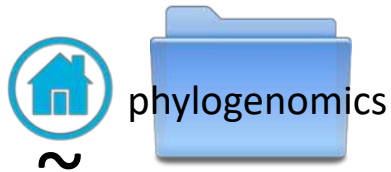


CP_Blast_seqs.fna LP_Blast_seqs.fna

Use 'grep' to print occurrences of a pattern

```
$ grep ">" CP_blast_seqs.fna
```

```
>MG272245.1 Neochondria nidifica voucher UC2026095 18S ribosomal RNA gene, partial sequence
>AF427537.1 Womersleyella setacea small subunit ribosomal RNA gene, complete sequence
>MF093929.1 Gredgaria maugiana isolate PD1230 18S ribosomal RNA gene, partial sequence
>HM560626.1 Digenea simplex voucher Dig sim 18S ribosomal RNA gene, partial sequence
>MG272238.1 Chondria crassicaulis voucher SAP115362 18S ribosomal RNA gene, partial sequence
>AF339897.1 Lenormandia sp. MELU_000065 18S ribosomal RNA gene, complete sequence
>MF093941.1 Periphykon beckeri isolate JH1427 18S ribosomal RNA gene, partial sequence
>MF093947.1 Polysiphonia scopulorum isolate PD0899 18S ribosomal RNA gene, partial sequence
>JX828186.1 Pterosiphonia pennata voucher CH816 small subunit ribosomal RNA gene, partial sequence
>AF427526.1 Boergesenella fruticulosa small subunit ribosomal RNA gene, complete sequence
>MF093935.1 Lophocladia kuetzingii isolate PD1509 18S ribosomal RNA gene, partial sequence
>JX828165.1 Boergesenella thuyoides voucher CH824 small subunit ribosomal RNA gene, partial sequence
>HM582914.1 Osmundaria obtusiloba voucher 03151 small subunit ribosomal RNA gene, partial sequence
>AF427527.1 Enelittosiphonia stimpsonii small subunit ribosomal RNA gene, complete sequence
>JX828192.1 Polysiphonia brodiei voucher CH410 small subunit ribosomal RNA gene, partial sequence
>AF203890.1 Heterocladia australis small subunit ribosomal RNA gene, complete sequence
>KX499570.1 Brongniartella australis isolate PD931 small subunit ribosomal RNA gene, partial sequence
>FJ153773.2 Uncultured eukaryote clone 53 18S ribosomal RNA gene, partial sequence
>AB219915.1 Polysiphonia sp. SNI07 gene for 18S rRNA, complete sequence
>MF093961.1 Vertebrata thuyoides isolate PD0546 18S ribosomal RNA gene, partial sequence
>JX828170.1 Leptosiphonia schousboei voucher CH826 small subunit ribosomal RNA gene, partial sequence
>JX828189.1 Symphyocladia linearis voucher CH419 small subunit ribosomal RNA gene, partial sequence
>JX828171.1 Neosiphonia elongella voucher CH415 small subunit ribosomal RNA gene, partial sequence
>AF427530.1 Polysiphonia fucoides small subunit ribosomal RNA gene, complete sequence
>AF203892.1 Heterocladia umbellata small subunit ribosomal RNA gene, complete sequence
>AB219910.1 Polysiphonia sp. BRI50 gene for 18S rRNA, complete sequence
>AF203885.1 Neosiphonia savatieri small subunit ribosomal RNA gene, complete sequence
>JX828179.1 Polysiphonia atlantica voucher CH1268 small subunit ribosomal RNA gene, partial sequence
>JX828175.1 Neosiphonia yendoi voucher CH420 small subunit ribosomal RNA gene, partial sequence
>JX828181.1 Polysiphonia elongata voucher CH1416 small subunit ribosomal RNA gene, partial sequence
>MF093943.1 Polysiphonia brodiei isolate PD0516 18S ribosomal RNA gene, partial sequence
>AF427534.1 Polysiphonia nigra small subunit ribosomal RNA gene, complete sequence
>JX828168.1 Ceramiales sp. HGC-2012 voucher CH1414 small subunit ribosomal RNA gene, partial sequence
>MF093937.1 Melanothamnus harveyi isolate PD0890 18S ribosomal RNA gene, partial sequence
>AF427529.1 Polysiphonia elongata small subunit ribosomal RNA gene, complete sequence
>AY237287.1 Osmundaria spiralis 18S ribosomal RNA gene, partial sequence
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$
```

Use 'grep' to print occurrences of a pattern

```
$ grep ">" CP_blast_seqs.fna
```



Create a new file of the fasta headers

```
$ grep ">" CP_blast_seqs.fna > CP_blast_headers.txt
```

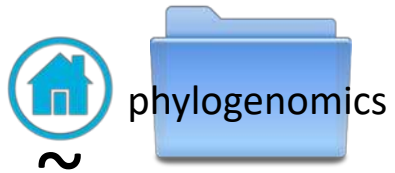


```
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ grep "^>" CP_blast_seqs.fna > CP_blast_headers.txt
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ head CP_blast_headers.txt
>AY617126.1 Choreocolax polysiphoniae 18S ribosomal RNA gene, partial sequence
>AY617142.1 Odonthalia washintoniensis 18S ribosomal RNA gene, partial sequence
>AY617140.1 Neorhodomela larix 18S ribosomal RNA gene, partial sequence
>AY617141.1 Odonthalia floccosa 18S ribosomal RNA gene, partial sequence
>MF093958.1 Thaumattella adunca isolate PD1388 18S ribosomal RNA gene, partial sequence
>MF093926.1 Digenea simplex isolate PD1820 18S ribosomal RNA gene, partial sequence
>JX828176.1 Odonthalia corymbifera voucher OK230 small subunit ribosomal RNA gene, partial sequence
>AY617145.1 Rhodomela confervoides 18S ribosomal RNA gene, partial sequence
>MF093954.1 Rhodomela confervoides isolate PD508 18S ribosomal RNA gene, partial sequence
>L26203.1 Rhodomela confervoides DNA fragment
```

CP_blast_seqs.fna

LP_blast_seqs.fna

CP_blast_headers.txt



Use 'grep' for to count the number of times a pattern occurs

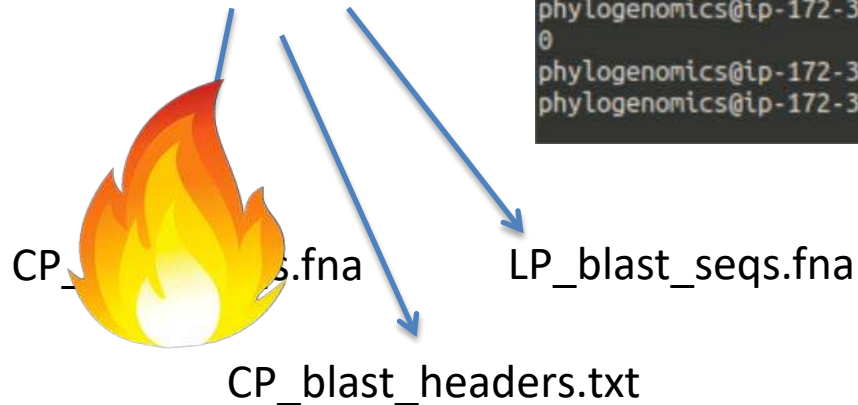
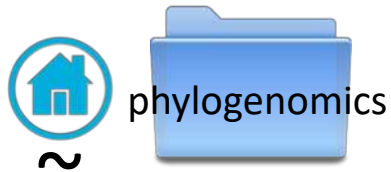
```
$ grep -c ">" CP_Blast_seqs.fna
```

```
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ grep -c "^>" CP_blast_seqs.fna
100
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ grep -c ">" CP_blast_seqs.fna
100
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$
```

CP_blast_seqs.fna

LP_blast_seqs.fna

CP_blast_headers.txt



Quotation marks are important

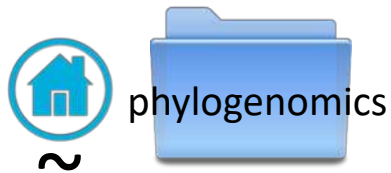
```
$ grep -c ">" CP_blast_seqs.fna
```

```
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ grep -c "^>" CP_blast_seqs.fna
100
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ grep -c ">" CP_blast_seqs.fna
100
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$
```

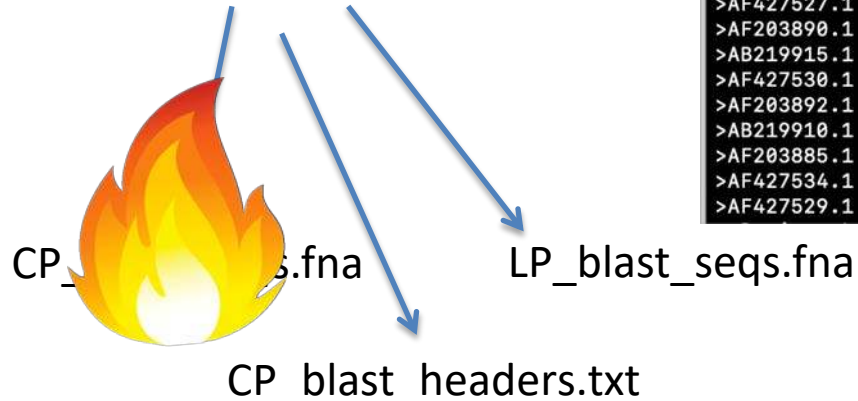
```
$ grep -c > CP_blast_seqs.fna
```

```
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ grep -c > CP_blast_seqs.fna
Usage: grep [OPTION]... PATTERN [FILE]...
Try 'grep --help' for more information.
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ grep -c ">" CP_blast_seqs.fna
0
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ head CP_blast_seqs.fna
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$
```

CP_Blast_seqs.fna is empty



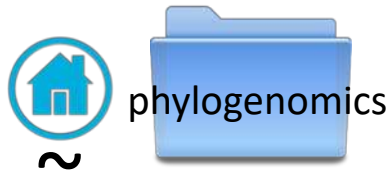
workshop_materials



Search for headers that are not partial sequences

```
$ grep -v "partial" CP_blast_headers.txt
```

```
salami:workshop_files ericsalomaki$ grep -v "partial" CP_blast_headers.txt
>L26203.1 Rhodomela confervoides DNA fragment
>AF339893.1 Lenormandia latifolia 18S ribosomal RNA gene, complete sequence
>AF203897.1 Lenormandia muelleri small subunit ribosomal RNA gene, complete sequence
>AF339899.1 Neurymenia fraxinifolia 18S ribosomal RNA gene, complete sequence
>AF203889.1 Melanamansia mamillaris small subunit ribosomal RNA gene, complete sequence
>AF339896.1 Lenormandia spectabilis 18S ribosomal RNA gene, complete sequence
>AF339901.1 Protokuetszingia australasica 18S ribosomal RNA gene, complete sequence
>AF339895.1 Lenormandia smithiae 18S ribosomal RNA gene, complete sequence
>AF203895.1 Lenormandia prolifera small subunit ribosomal RNA gene, complete sequence
>AF339900.1 Osmundaria prolifera 18S ribosomal RNA gene, complete sequence
>AF203896.1 Micropeuce strobiliferum small subunit ribosomal RNA gene, complete sequence
>AF251513.1 Halopithys incurva 18S ribosomal RNA gene, complete sequence
>AF339892.1 Lenormandia angustifolia 18S ribosomal RNA gene, complete sequence
>AF203894.1 Laurencia filiformis small subunit ribosomal RNA gene, complete sequence
>AF251512.1 Melanamansia glomerata 18S ribosomal RNA gene, complete sequence
>AF203887.1 Murrayella pericladus small subunit ribosomal RNA gene, complete sequence
>AF203886.1 Polysiphonia lanosa small subunit ribosomal RNA gene, complete sequence
>AF339898.1 Lenormandia sp. MELU_000064 18S ribosomal RNA gene, complete sequence
>AF427537.1 Womersleyella setacea small subunit ribosomal RNA gene, complete sequence
>AF339897.1 Lenormandia sp. MELU_000065 18S ribosomal RNA gene, complete sequence
>AF427526.1 Boergesenella fruticulosa small subunit ribosomal RNA gene, complete sequence
>AF427527.1 Enelittosiphonia stimpsonii small subunit ribosomal RNA gene, complete sequence
>AF203890.1 Heterocladia australis small subunit ribosomal RNA gene, complete sequence
>AB219915.1 Polysiphonia sp. SNI07 gene for 18S rRNA, complete sequence
>AF427530.1 Polysiphonia fucoides small subunit ribosomal RNA gene, complete sequence
>AF203892.1 Heterocladia umbellata small subunit ribosomal RNA gene, complete sequence
>AB219910.1 Polysiphonia sp. BRI50 gene for 18S rRNA, complete sequence
>AF203885.1 Neosiphonia savatieri small subunit ribosomal RNA gene, complete sequence
>AF427534.1 Polysiphonia nigra small subunit ribosomal RNA gene, complete sequence
>AF427529.1 Polysiphonia elongata small subunit ribosomal RNA gene, complete sequence
```

Replacing Text in Large files



`sed 's/**FIND**/**REPLACE**/g' filename > output_file`

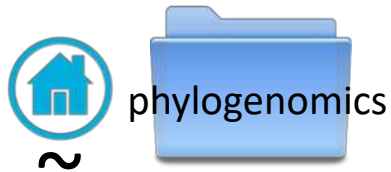
```
$ sed 's/ /_/g' CP_blast_headers.txt
```

```
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ sed 's/ /_/g' CP_blast_headers.txt > CP_blast_headers_clean.txt
phylogenomics@ip-172-30-3-100:~/workshop_materials/Blast_Out$ head CP_blast_headers_clean.txt
>AY617126.1_Choreocolax_polysiphoniae_18S_ribosomal_RNA_gene,_partial_sequence
>AY617142.1_Odonthalia_washingtoniensis_18S_ribosomal_RNA_gene,_partial_sequence
>AY617140.1_Neorhodomela_larix_18S_ribosomal_RNA_gene,_partial_sequence
>AY617141.1_Odonthalia_floccosa_18S_ribosomal_RNA_gene,_partial_sequence
>MF093958.1_Thaumatella_adunca_isolate_PD1388_18S_ribosomal_RNA_gene,_partial_sequence
>MF093926.1_Digenea_simplex_isolate_PD1820_18S_ribosomal_RNA_gene,_partial_sequence
>JX828176.1_Odonthalia_corymbifera_voucher_OK230_small_subunit_ribosomal_RNA_gene,_partial_sequence
>AY617145.1_Rhodomela_confervoides_18S_ribosomal_RNA_gene,_partial_sequence
>MF093954.1_Rhodomela_confervoides_isolate_PD508_18S_ribosomal_RNA_gene,_partial_sequence
>L26203.1_Rhodomela_confervoides_DNA_fragment
```

CP_blast_seqs.fna

LP_blast_seqs.fna

CP_blast_headers.txt



Replacing Text in Large files

Use the `-i` flag to make changes inplace

```
$ sed -i 's/,//g' CP_blast_headers.txt  
$ head CP_blast_headers.txt
```

```
>AY617126.1 Choreocolax polysiphoniae 18S ribosomal RNA gene partial sequence  
>AY617142.1 Odonthalia washintoniensis 18S ribosomal RNA gene partial sequence  
>AY617140.1 Neorhodomela larix 18S ribosomal RNA gene partial sequence  
>AY617141.1 Odonthalia floccosa 18S ribosomal RNA gene partial sequence  
>MF093958.1 Thaumattella adunca isolate PD1388 18S ribosomal RNA gene partial sequence  
>MF093926.1 Digenea simplex isolate PD1820 18S ribosomal RNA gene partial sequence  
>JX828176.1 Odonthalia corymbifera voucher OK230 small subunit ribosomal RNA gene partial sequence  
>AY617145.1 Rhodomela confervoides 18S ribosomal RNA gene partial sequence  
>MF093954.1 Rhodomela confervoides isolate PD508 18S ribosomal RNA gene partial sequence  
>L26203.1 Rhodomela confervoides DNA fragment
```

CP_blast_seqs.fna

LP_blast_seqs.fna

CP_blast_headers.txt

Regular Expressions

Encoding	Modern Equivalent	Pattern Type
.		a single character
.+		one or more characters
.*		zero or more characters
.?		Maybe present
^		first on the line
\$		last on the line
[0-9]	\d	digits
[a-zA-Z]	\w	letters
' '	\s \t	space
{3}		must be exactly 3 characters long
{3,5}		between 3-5 characters long
[ACGT]		a specific set of characters (a class)

Regular Expressions

Use -E with sed and grep to use extended regular expressions

Store pattern in memory using parentheses

Print out only the GenBank accessions

>AY617126.1 Choreocolax polysiphoniae 18S ribosomal RNA gene, partial sequence

1

2

```
$ sed -E 's/(>[A-Z0-9.]+)(.+)/\1/' CP_blast_headers.txt
```

```
[salami:workshop_files ericsalomaki$ sed -E 's/(>[A-Z0-9.]+)(.+)/\1/' CP_blast_headers.txt
>AY617126.1
>AY617142.1
>AY617140.1
>AY617141.1
>MF093958.1
>MF093959.1
```


Regular Expressions

Use -E with sed and grep to use extended regular expressions

Store pattern in memory using parentheses

Print out everything except the GenBank accessions

>AY617126.1 Choreocolax polysiphoniae 18S ribosomal RNA gene, partial sequence

1

2

```
$ sed -E 's/(>[A-Z0-9.]+)(.+)/\2/' CP_blast_headers.txt
```

```
salami:workshop_files ericsalomaki$ sed -E 's/(>[A-Z0-9.]+)(.+)/\2/' CP_blast_headers.txt
Choreocolax polysiphoniae 18S ribosomal RNA gene, partial sequence
Odonthalia washintoniensis 18S ribosomal RNA gene, partial sequence
Neorhodomela larix 18S ribosomal RNA gene, partial sequence
Odonthalia floccosa 18S ribosomal RNA gene, partial sequence
Thaumatella adunca isolate PD1388 18S ribosomal RNA gene, partial sequence
Digenea simplex isolate PD1820 18S ribosomal RNA gene, partial sequence
Odonthalia corymbifera voucher OK230 small subunit ribosomal RNA gene, partial sequence
Rhodomela confervoides 18S ribosomal RNA gene, partial sequence
Rhodomela confervoides isolate PD508 18S ribosomal RNA gene, partial sequence
```

Pipes



Use “|” to string multiple commands together

Combine the fasta files from the blast output and identify how many sequences there are

```
cat CP_blast_seqs.fna LP_blast_seqs.fna | grep ">" | wc -l
```

```
[salami:workshop_files ericsalomaki$ cat CP_blast_seqs.fna LP_blast_seqs.fna | grep ">" | wc -l  
200
```

Pipes



Use the pipe “|” to string multiple commands together

Combine the fasta files from the blast output and identify how many sequences there are

```
cat CP_blast_seqs.fna LP_blast_seqs.fna | grep ">" | wc -l
```

```
[salami:workshop_files ericsalomaki$ cat CP_blast_seqs.fna LP_blast_seqs.fna | grep ">" | wc -l  
200
```

Other built-in unix binaries that are great to use with pipes

‘sort’ – sort lines of text files

‘uniq’ – report or omit repeated lines (only works on a sorted file)



String together many commands to count the number of unique accessions from these blast results

```
$ cat *.fna |
```

 Print out all fasta files

```
$ cat *.fna | grep ">" |
```

 Find all the fasta headers

```
$ cat *.fna | grep ">" | sed -E 's/(>[A-Z0-9.]+)(.+)/\1/' |
```

 Remove everything except accessions

```
$ cat *.fna | grep ">" | sed -E 's/(>[A-Z0-9.]+)(.+)/\1/' | sort |
```

 Sort accessions

```
$ cat *.fna | grep ">" | sed -E 's/(>[A-Z0-9.]+)(.+)/\1/' | sort | uniq |
```

 Keep only unique accessions

```
$ cat *.fna | grep ">" | sed -E 's/(>[A-Z0-9.]+)(.+)/\1/' | sort | uniq | wc -l
```

 Count unique accessions

```
salami:Blast_Out ericsalomaki$ cat *.fna | grep ">" | sed -E 's/(>[A-Z0-9.]+)(.+)/\1/' | sort | uniq | wc -l  
144
```

Using a Loop



Using for loops allow you to apply a command to multiple files at once

```
$ for f in *.fna; do grep -c "ACT" $f ; done
```

Assign variable “f” to all files that end in .fna
Then carry out command on all values of “f”

```
[salami:Blast_Out ericsalomaki$ for f in *.fna; do grep -c "ACT" $f ; done  
1354  
1291
```

I often use loops to make several single gene trees at once

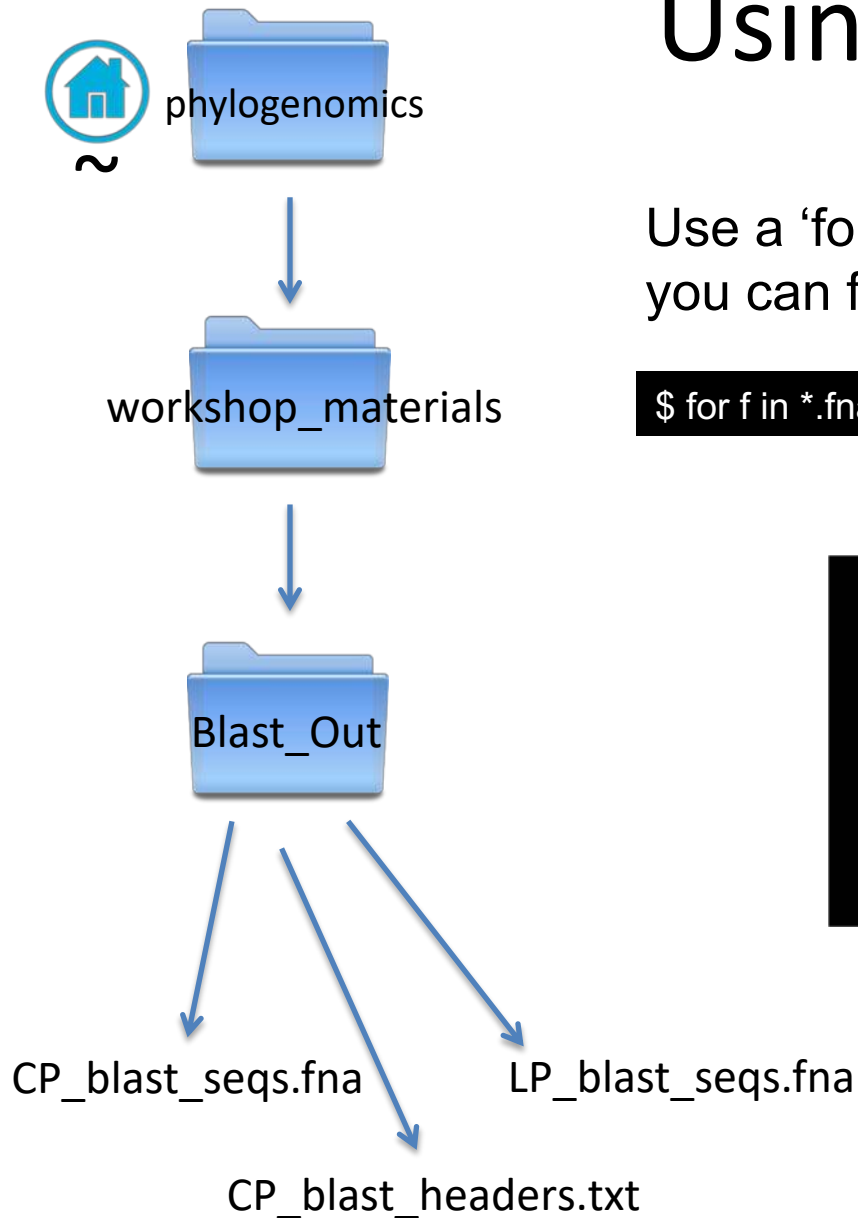


Using a Loop

Use a 'for loop' to find how many times you can find "CAT" in these fasta files

```
$ for f in *.fna; do grep -c "CAT" $f ; done
```

1257
1319



Online Downloads



wget will download to your pwd
Use the `-P` option

```
wget -P /path/to/desired/location/ File_To_Download.txt
```


Online Downloads



Use `wget -P` to download these slides to a new directory
-P will make a new directory if it does not yet exist

```
wget -P ./wget_dir/ http://evomics.org/wp-content/uploads/2019/01/Introduction_to_Unix_Krumlov2019_Jan21_red.pdf
```

http://evomics.org/wp-content/uploads/2019/01/Introduction_to_Unix_Krumlov2019_Jan21_red.pdf

Then use `ls` to see the new directory and then again to look inside the directory to ensure the slides have been downloaded

Final Unix Challenge



Download the challenge and view it in terminal

<http://evomics.org/wp-content/uploads/2019/01/Final Unix Challenge.txt>

Save Your Data Everyday



We will launch new instance tonight so everything you have done today will be gone tomorrow

Use scp or rsync to transfer files

`scp -r username@IP.address:/Location/On/Instance /Location/On/Your/Computer`

```
$ scp -r phyogenomics@ec2-34-227-31-223.compute-1.amazonaws.com:~/workshop_materials/Unix  
~/Location/On/Your/Computer
```

Any questions?

