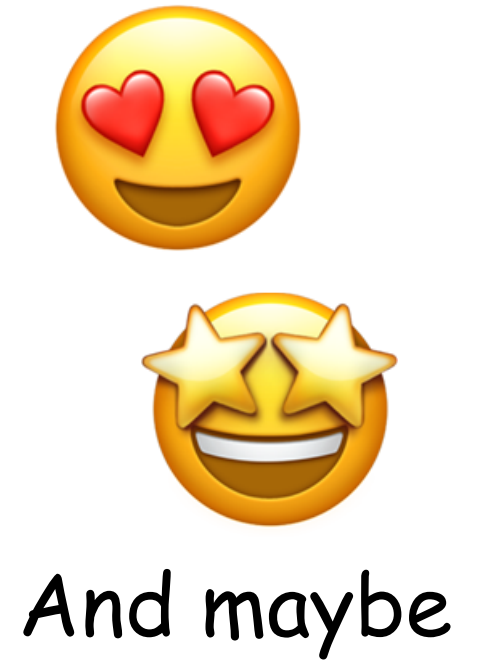
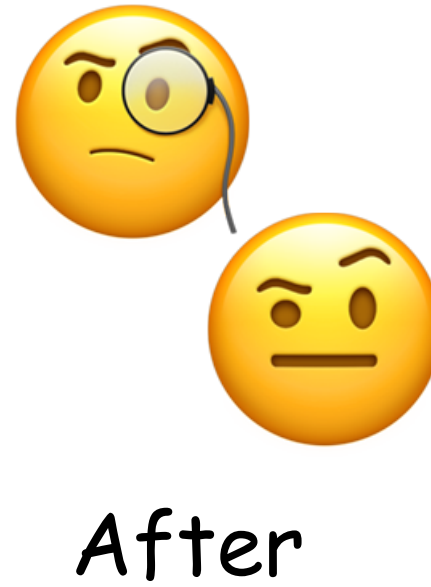
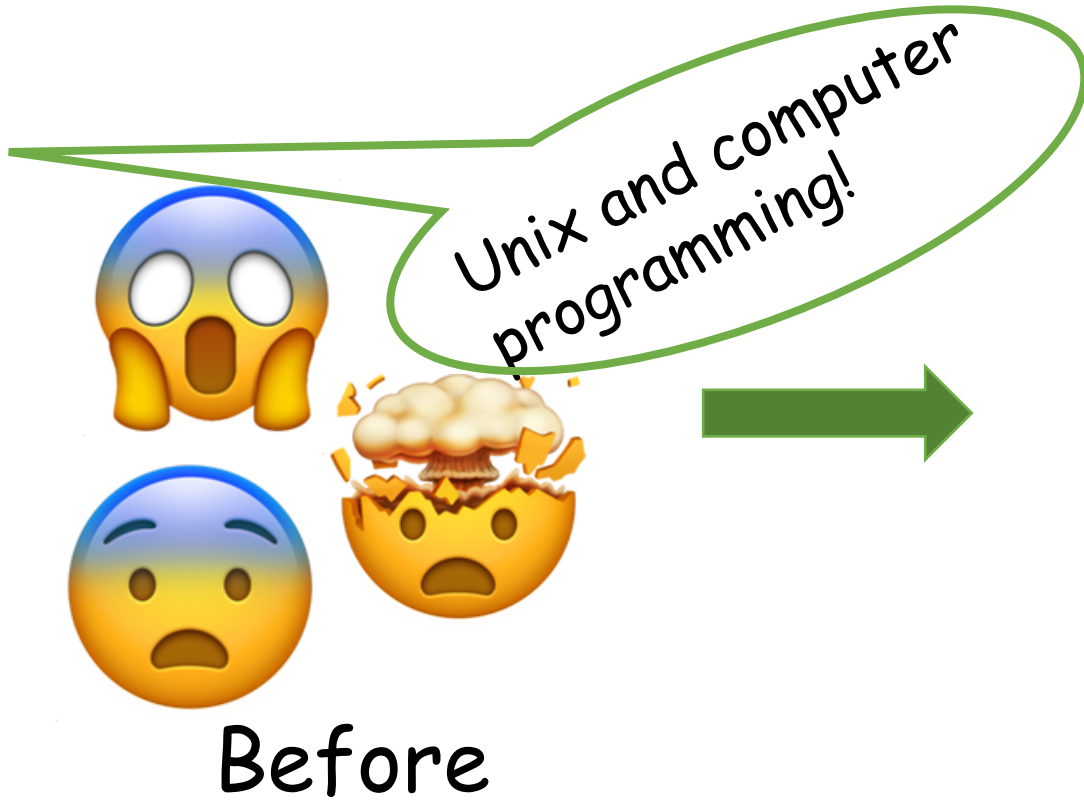


# General concepts of Computer programming

- Programming is a language for talking to computer.
- Language has **grammar** (syntax) and **vocabs** (commands = noun + verb (+ adjective)).
- Computer often reads from top line and go downward.
- Computer will do **EXACTLY** as you tell them, and sometimes will **over-write** things without asking for a confirmation
- Commands and names are **case-sensitive**.
- You can name objects anything you like, but
  - Use **meaningful** names
  - Avoid using function name as object name
  - **Special characters** e.g. # \$ & space are **not allowed** in object names (or it will make life difficult)
- **Google** is your friend, and so are **BioStar**, **StackExchange**, **ResearchGate**

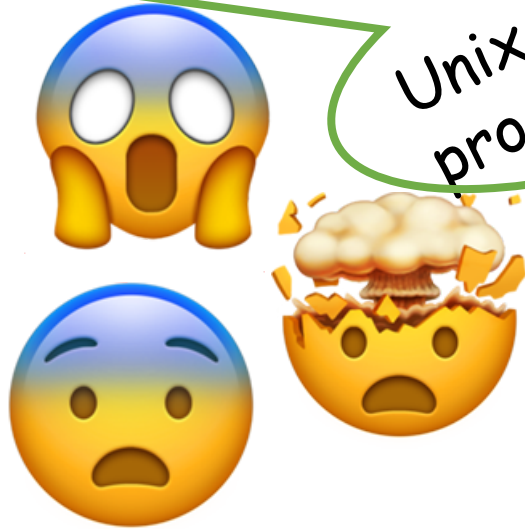
People often have the same question as you do

# What this session is



# What this session is NOT

Unix and computer programming!



Before



After

# Basic bioinformatics files

- For any sequencing projects
  - Or if you work with genome databases, nucleic acid and, protein sequences
- 
- FASTQ .fastq .fq .fq.gz
  - FASTA .fasta .fa .fa.gz .fasta.gz
  - sam, bam, gff, vcf, etc.

<https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/#>

# FASTQ file

ERR506076\_1.fastq ERR506076\_2.fastq

**Header line:** can contain various information about the sequence eg. coordinate on sequencing chip. Here, it shows accession number and read ID.

**File name:** data from paired-end sequencing will have \_1 \_2 at the end of their file name. These numbers correspond to reads from each end of the same sequenced fragment

Another header line for the same sequence. Sometimes there will be only the plus (+) sign.

**Quality score:** Each character represent a score

[illegible]

# FASTA file

schistosoma\_mansoni.PRJEA36577.WBPS13.genomic.fa

**file name:**  
.fa or .fasta

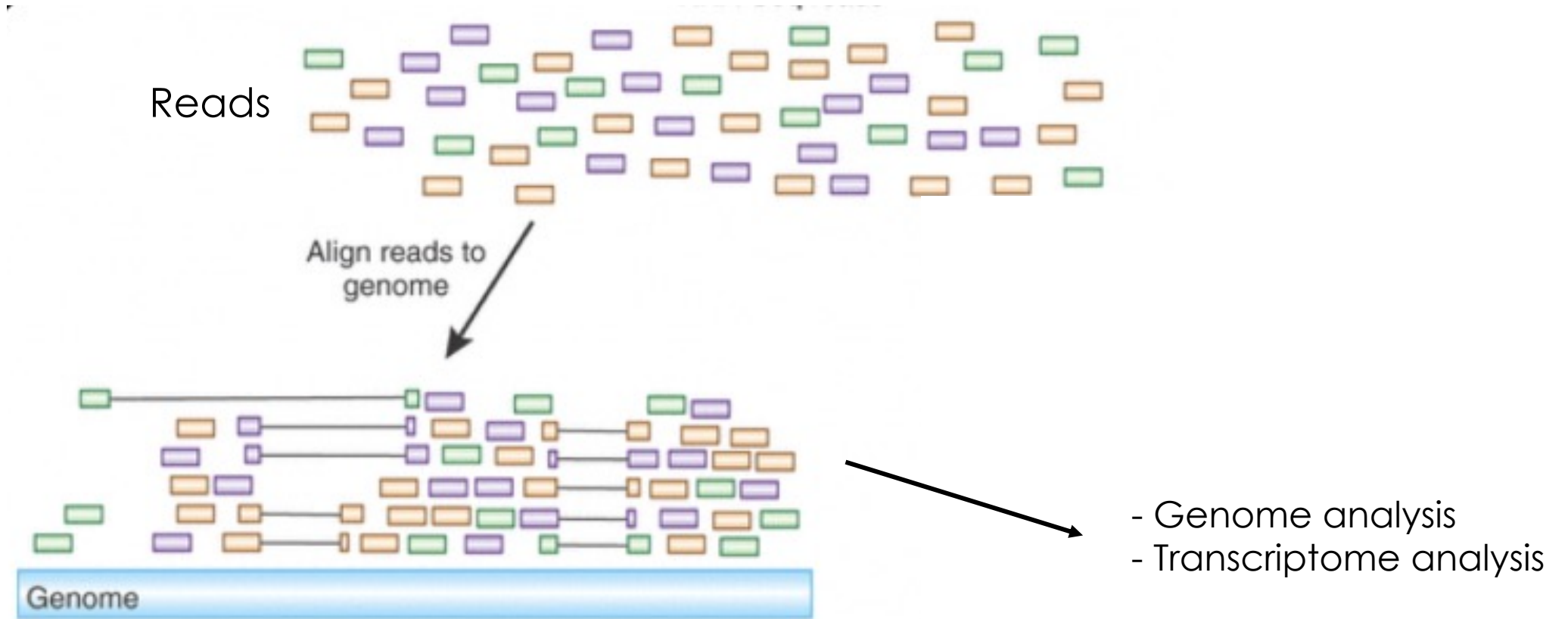
**Header line:** Information about the sequence. For genomic sequence, this is often contig ID or chromosome number. It can also contain other information such as contig length, species name, source of the information

```
>SM_V7_1 length=88881357
TGATAGTTAGTCATATGAAAGCATCATTAGTAAACCACATTGCTTATTATATTGAACAGT
TACATCTGGCTTATTATACAAAGAGAAAACCATACTATTTCATACTATTCTCTTTTTTGATC
TTCTCAATCTTCTGTTGTTAGATATTCTATTCCTTGCTCACCATATATACTACTTATGTC
AATATAAGTAGCTCACACCACACTACTACTTCAACTACTACTACTGCTACTGCTACT
GTTGTGAACAGAACACGACTGTTGGACAATCGAATCTAATTAAGCAAACATTACAAACTA

... (and many more lines follow)
```

# SAM and BAM files

sequence alignment/map, and binary version





# **SAM and BAM files**

**sequence alignment/map, and binary version**

Information about the mapping e.g. read ID,  
mapped position, mapping score

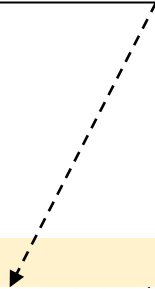
[illegible]



# GFF files

## Genome annotation

Location on the genome, what feature(s) has it been annotated for, evidence, etc.



Schisto_mansoni.Chr_3.unplaced.SC_0083	chado	CDS	852574	852785	.	-	0	ID=Smp_138770.1:exon:18;Parent=Smp_138770.1;isObsolete=false;timelastmodified=17.10.2011+10:47:01+BST
Schisto_mansoni.Chr_3.unplaced.SC_0083	chado	CDS	854903	854990	.	-	0	ID=Smp_138770.1:exon:17;Parent=Smp_138770.1;isObsolete=false;timelastmodified=17.10.2011+10:47:01+BST
Schisto_mansoni.Chr_3.unplaced.SC_0083	chado	CDS	857133	857186	.	-	0	ID=Smp_138770.1:exon:16;Parent=Smp_138770.1;isObsolete=false;timelastmodified=17.10.2011+10:47:01+BST
Schisto_mansoni.Chr_3.unplaced.SC_0083	chado	CDS	860632	860763	.	-	0	ID=Smp_138770.1:exon:15;Parent=Smp_138770.1;isObsolete=false;timelastmodified=17.10.2011+10:47:01+BST
Schisto_mansoni.Chr_3.unplaced.SC_0083	chado	CDS	861448	861573	.	-	0	ID=Smp_138770.1:exon:14;Parent=Smp_138770.1;isObsolete=false;timelastmodified=17.10.2011+10:47:01+BST

# VCF files

## Variant calling format

Position of genome, sequence on reference genome, sequence in mapped reads, other information

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
.....more header lines.....
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51
1|0:48:8:51,51 1/1:43:5:,...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3
0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51
0/0:61:2
20 1234567 microsat1 GTCT G,GTACT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2
1/1:40:3
```

# Working with bioinformatics files

- Mostly text files in specific format
- Can be over many gigabytes in size
- Can have millions of lines
- ... we won't work manually

# Unix

- **What is Unix?**

- An operating system (like Windows or OSX)
- Provides a way to interact with the computer
- Many versions, including Linux, which is a free version of Unix

- **Why use it?**

- Output of biological research is often large text files
- Suitable for working with these files and very powerful
- Saves time
- Widely used by the scientific community
- Robust and stable

# Unix is another (computer) language

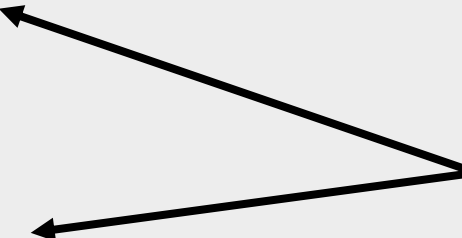
- Languages have **syntax (grammar)**
- Languages have **nouns**
- Languages have **verbs**

# Unix is another (computer) language

- Languages have **syntax (grammar)**

Unix command line normally look like this...

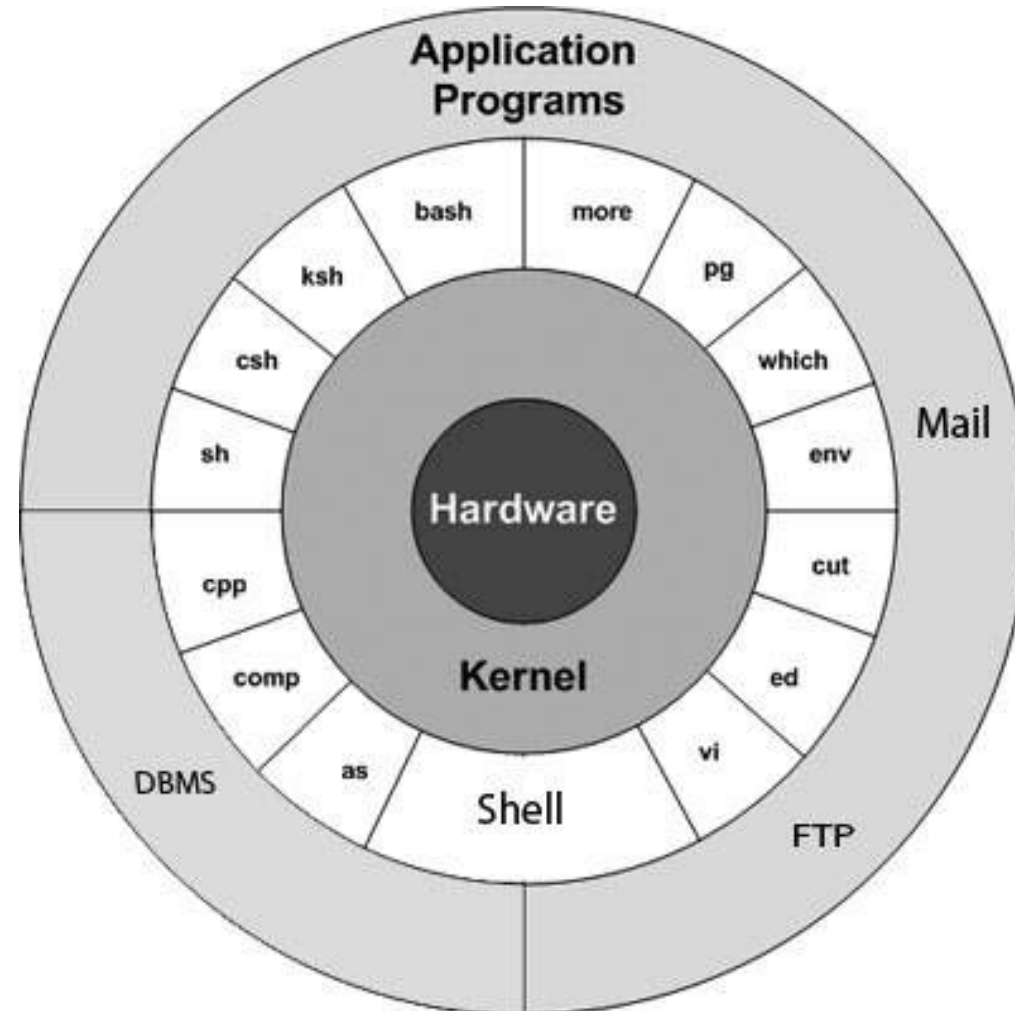
```
$ some_command filename
$ some_command location
$ some_command -i -a -x -n5 filename
$ some_command filename anotherfile
$ some_command filename > newfilename
```



**command (verb)** follows by  
**filename or location (noun)**  
or additional information  
**(arguments/options)**

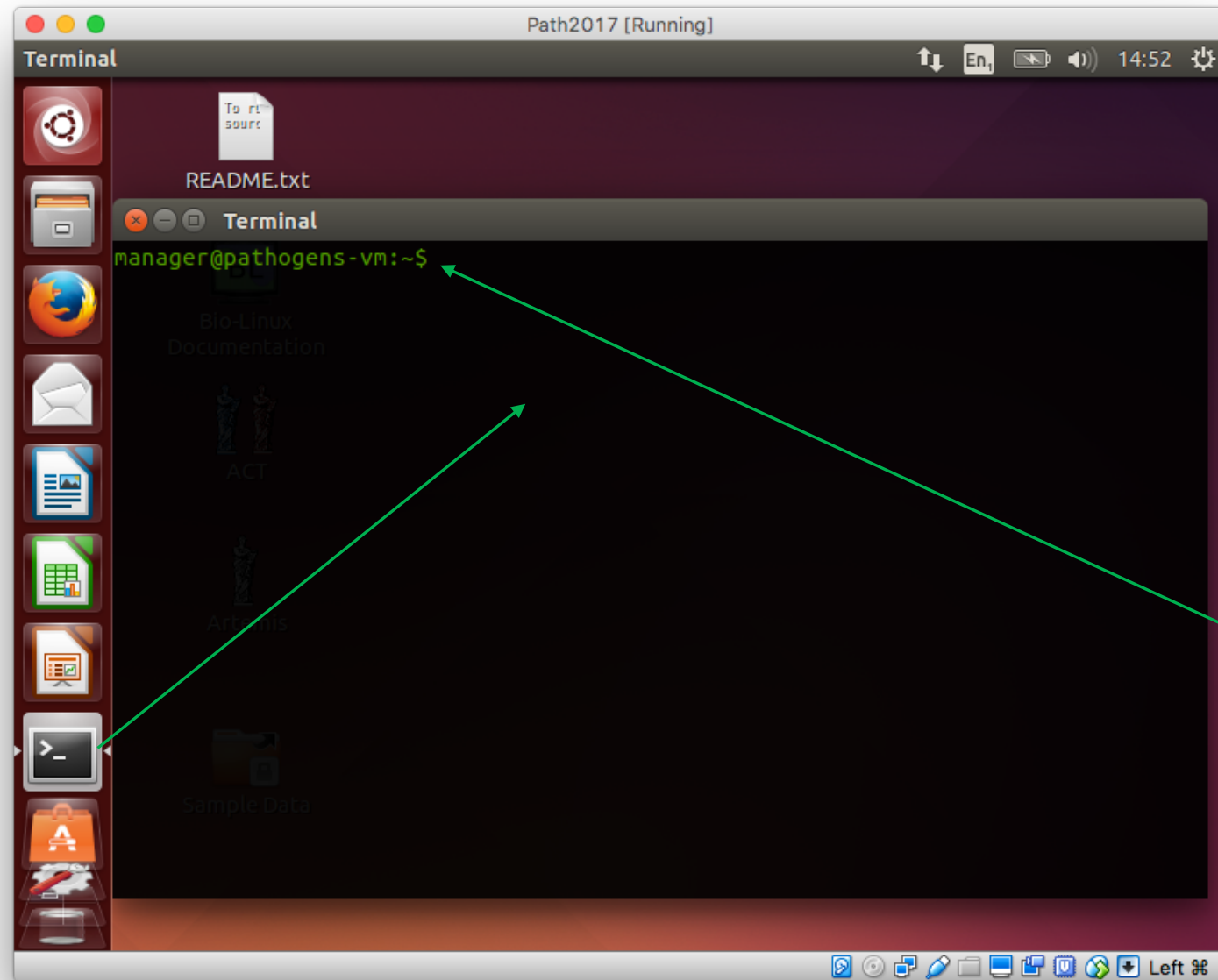
# Unix

Operating system that underlies a lot of bioinformatics work, and system administration





# Terminals and the Command Line



UNIX prompt

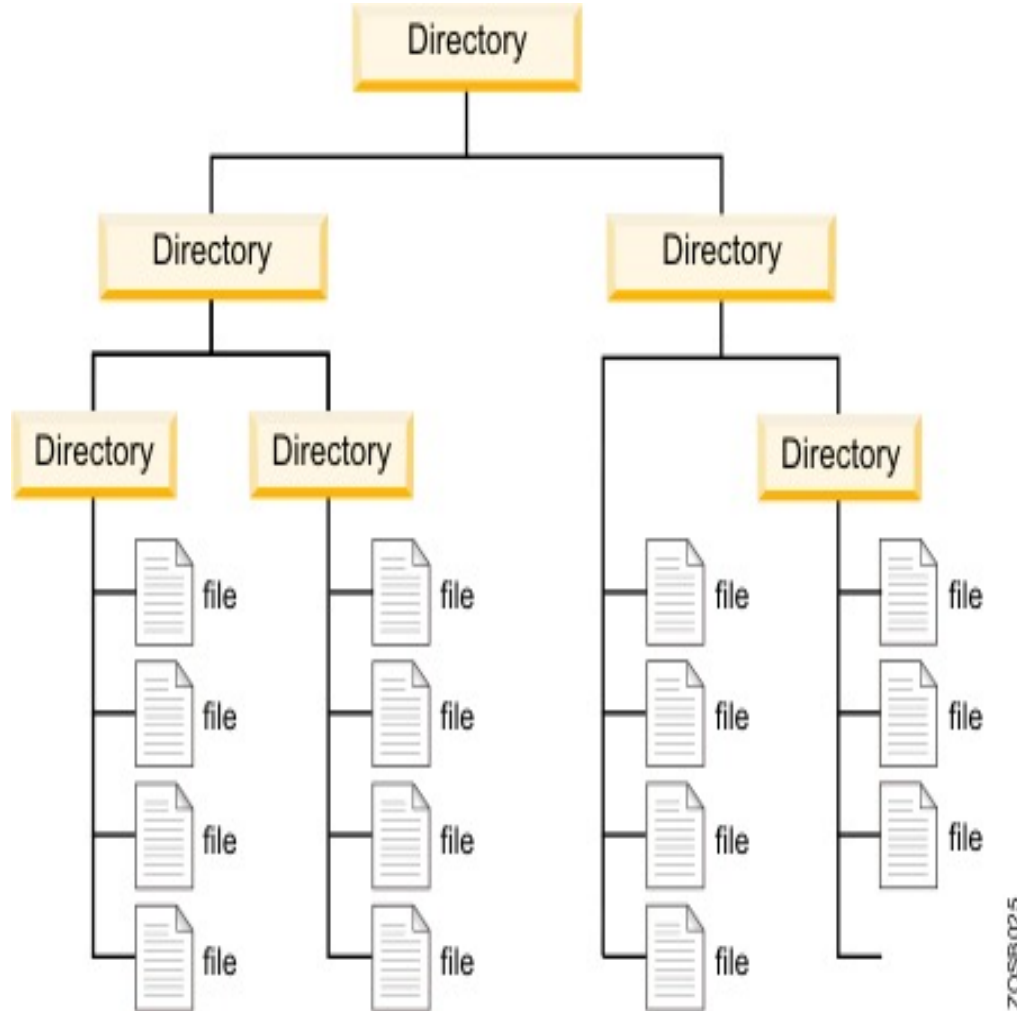
# Unix Commands

Command	What it does
ls	List the contents of the current directory
cd	Change directory
mv	Moves a file
cp	Copies a file
rm	Deletes a file
less	Displays the contents of a file
head	Displays the first ten lines of a file
tail	Displays the last ten lines of a file
cat	Concatenates files together
pwd	Prints working directory
mkdir	Makes a new directory

# Computer file system

Folder is called “directory” in programming world

You need to tell the computer the EXACT location of the file

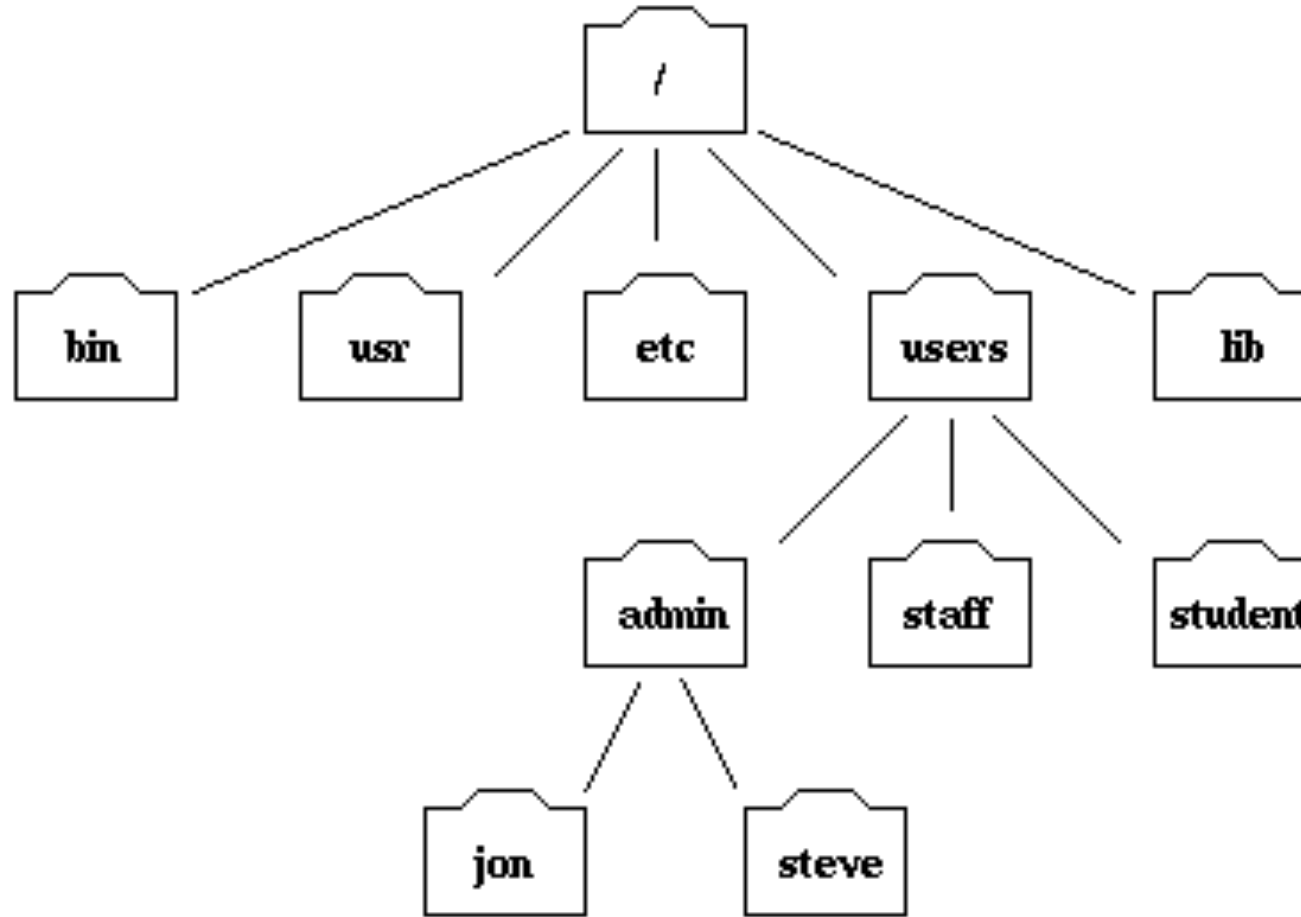


File not found

There is no such file

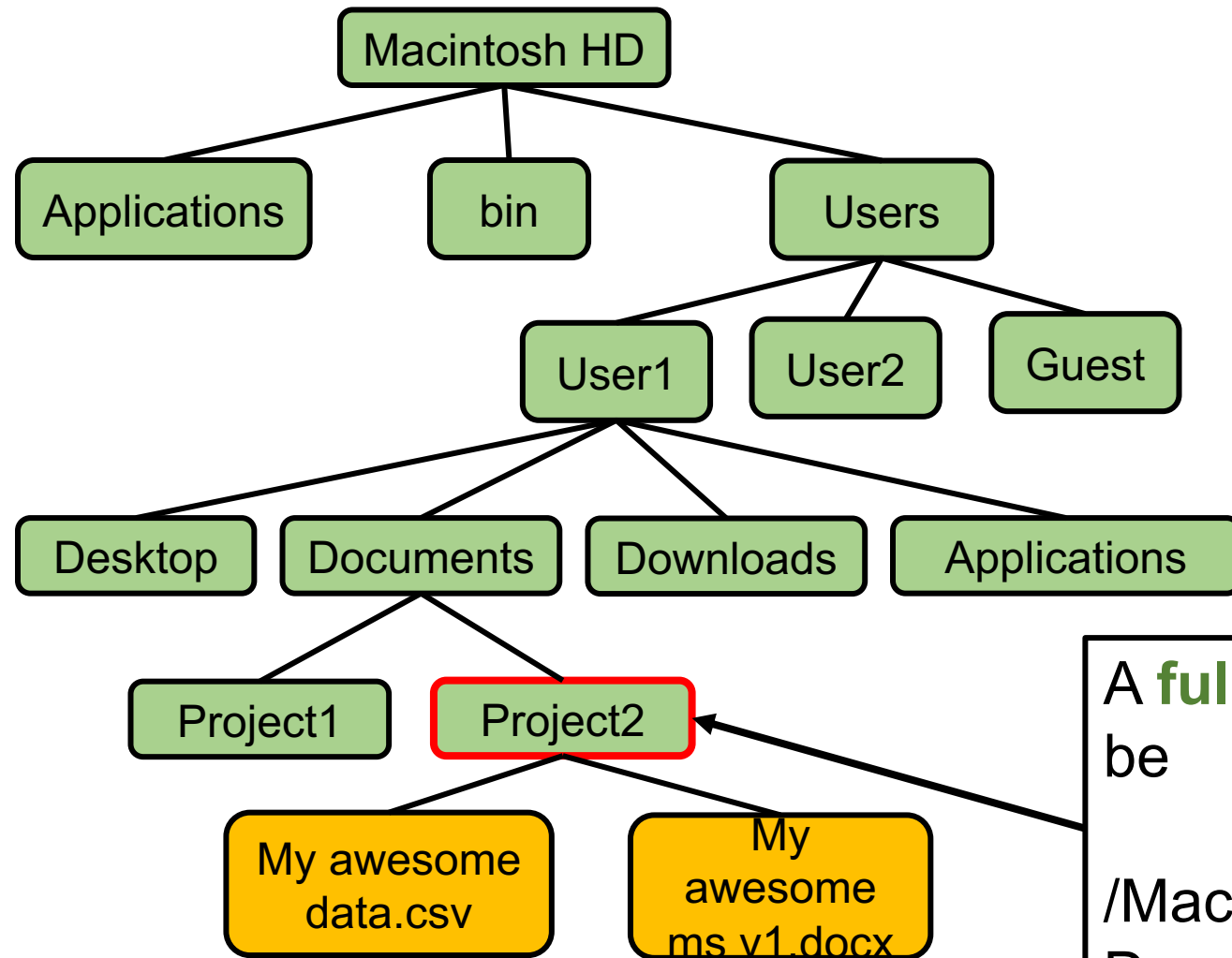
These are very common mistake.  
Make sure you get the location  
correct + the spelling  
**\*use TAB key\***

# Unix file system



**Part of the filesystem tree**

# Mac file system (simplified)



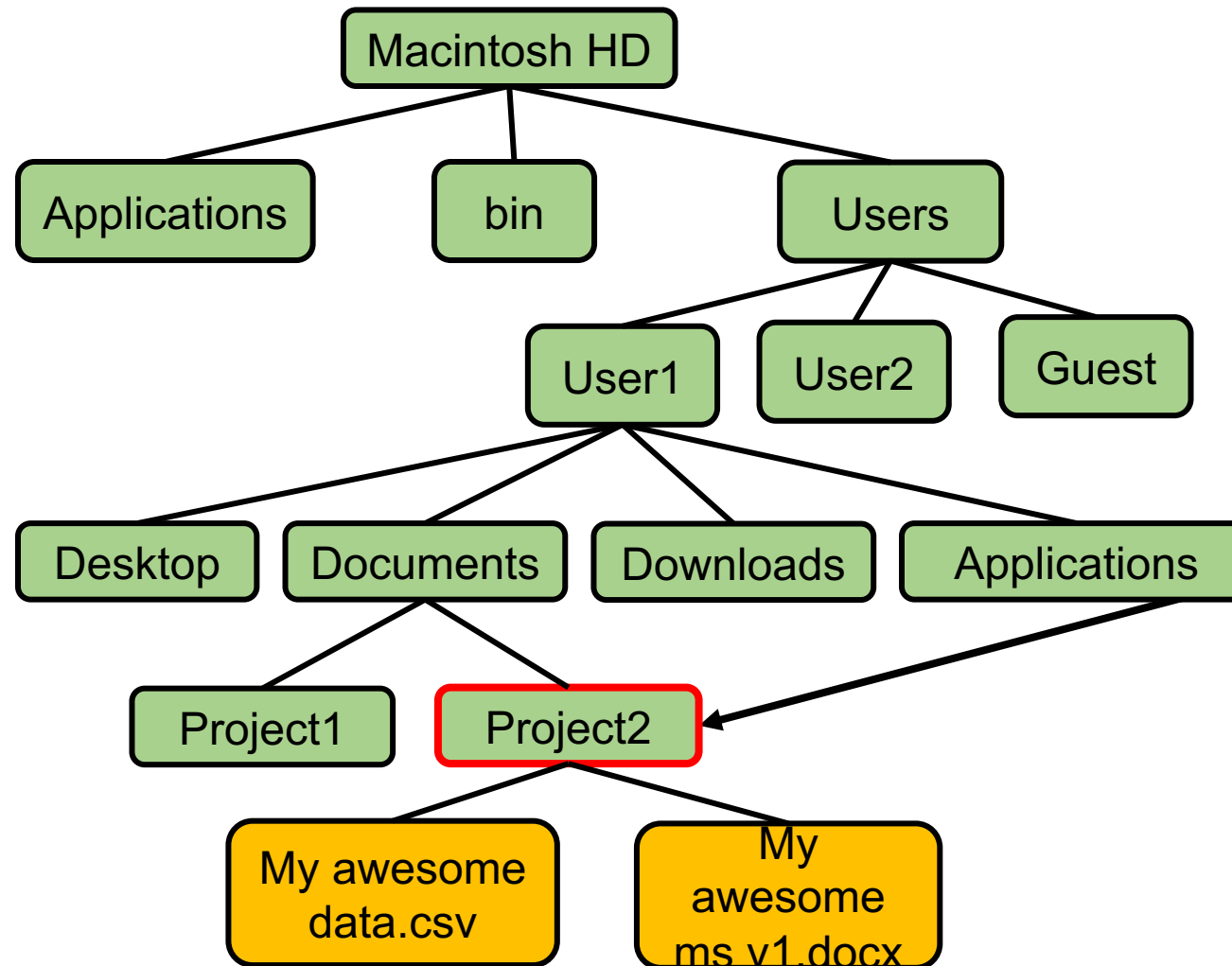
 Folder

 File

A **full path** to this directory would be

/Macintosh HD/Users/User1/  
Documents/Project2/

# Mac file system (simplified)



 Folder

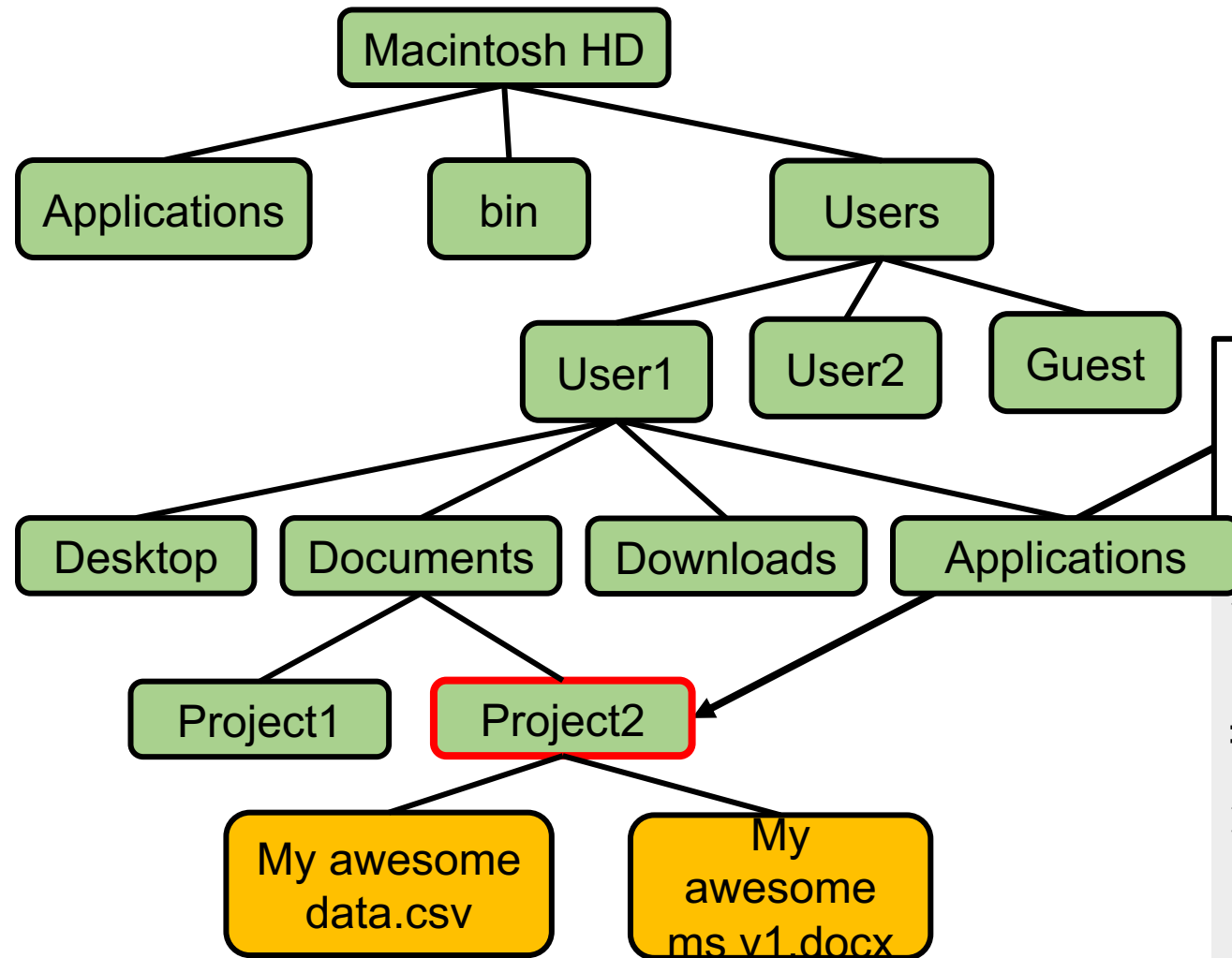
 File

Let's say you are inside Project1 directory, your **relative path** to Project2 would be

`../Project2/`

`..` means going up 1 step

# Mac file system (simplified)



 Folder

 File

Let's say you are inside Project1 in R, but you want to import the file "My awesome data.csv" which is

```
> read.csv("../Project2/My  
awesome data.csv")
```

# or

```
> read.csv("/Macintosh  
HD/  
_____/My awesome data.csv")
```



# Common Unix commands

## Navigating

pwd

ls

cd

find

.

..

# Common Unix commands

## Look into files

more

less

head

tail

cat (2 functions)

# Common Unix commands

## Extracting information from/about files

grep

wc

cut

sort

uniq

# Common Unix commands

## Creating/moving/copying/deleting files

touch

mkdir

mv

cp

**rm**

**rmdir**

# Common Unix commands

## Editing files

vim

chmod (and file permission)

cat

paste

>

>>

# Common Unix commands

## Special characters and wildcard characters

	output from command on the left become input for command on the right
*	any number of characters
?	one character
[]	specify a range of characters
{ }	specify a list of terms, separated by commas
!	exclude this range of characters

## And more complex (fancy/useful) commands

sed	string editor
awk	pattern search and more

# Common Unix commands

## Other useful things

history

man

which

tar

gunzip

diff/gdiff

command line options (- ...)

## **TAB and double TAB**

"arrow up"      press arrow up to get old commands

CTRL a      Go to start of line

CTRL e      Go to end of line



# Reminders

- Unix is case sensitive
  - *LS* is not the same as *ls*
- Spaces must go between a command and the values passed to a command
  - *mkdir newdir* will work
  - *mkdirnewdir* will give an error
- Spelling must be correct. Mistyping a command or file name will give an error
- Use **TAB** key
- Don't fear the command line!

# File permission (ls -l)

-**rw**x**rw**-**r**--, this means the line displayed is:

- a regular file (displayed as -)
- readable, writable and executable by owner (rw**x**)
- readable, writable, but not executable by group (rw-)
- readable but not writable or executable by other (r--)

Read	Write	Execute
4	2	1

# If you want revision/extension on Unix

- Introduction to Linux for Bioinformatics - Online Course - FutureLearn

<https://www.futurelearn.com/courses/linux-for-bioinformatics>

- UNIX / Linux Tutorial for Beginners

<http://www.ee.surrey.ac.uk/Teaching/Unix/>

- The Unix Shell

<https://swcarpentry.github.io/shell-novice/>

- A Primer for Computational Biology – Open Textbook

<https://open.oregonstate.education/computationalbiology/>

+ Look for Unix cheatsheet(s)