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

- **Google** is your friend, and so are **BioStar**, **StackExchange**, **ResearchGate**

People often have the same question as you do

Basic bioinformatics files

For any sequencing projects

- 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
```

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

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

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

Quality score: Each character represent a score

```
@ERR506076.1 1 length=100
TAAGAATTAATGTATTGAGCTACACGTAATGTGATGTGCAACTCAATCAGATGAGTGAAATTGCCCAGAAACAAATCACCAAAGAAGGTACA
TATATGTA
+ERR506076.1 1 length=100
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJJJJJJJJJJJJJJHIJJJJJJJJJJJJJJJJJJJJJHHHHFFFFFEAEEE
FEEEEEEE
@ERR506076.2 2 length=100
GCCGAAATTAGTGTTGACGGTCCGTTAAGAATTAATGTATTGATACACGTAATGTGATGTGCAACTCAATCAGATGAGTGAAATTGCCCAGA
AACAAATC
+ERR506076.2 2 length=100
BCCFFFFFHHHFHIJJIJJJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJIIJJJCHIJIHIIJJJJJHHHHHFFFFFFCEEEEEEDDDDDDD
DDDDDDDD
... (and many more lines follow)
```

FASTA file

schistosoma_mansoni.PRJEA36577.WBPS13.genomic.fa

file name:
.fa or .fasta

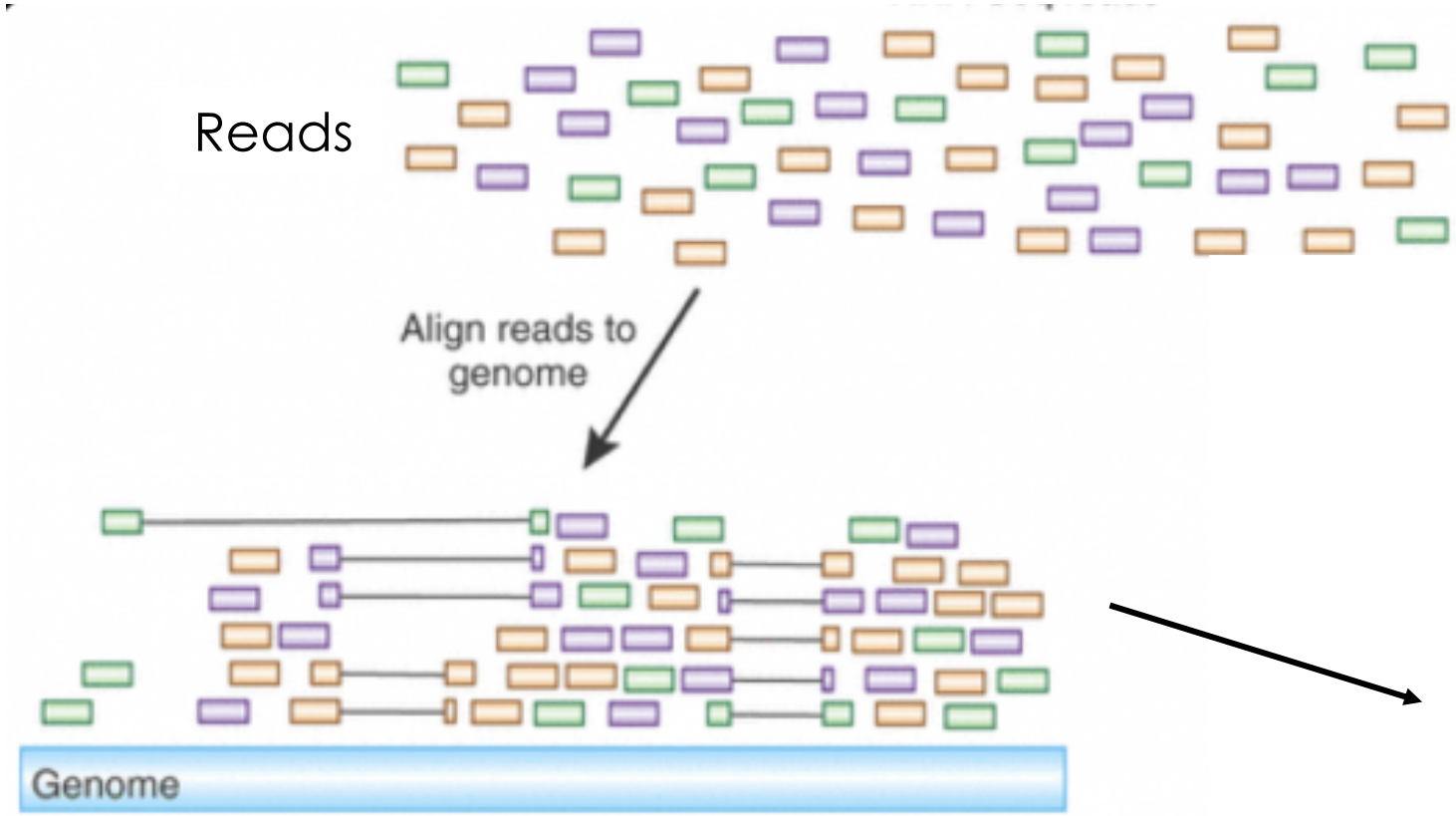
Header line: Information about the sequence. For genomic sequence, this is often contig ID or chromosome number. It can also contain other information such as contig length, species name, source of the information

```
>SM_V7_1 length=88881357
TGATAGTTAGTCATATGAAAGCATCATTAGTAAACCACATTGCTTATTATATTGAACAGT
TACATCTGGCTTATTATACAAAGAGAAAACCATACTATTCATACTATTCTCTTTTTGATC
TTCTCAATCTTCTGTTGTTAGATATTCTATTCCTTGCTCACCATATATACTACTTATGTC
AATATAAGTAGCTCACACCACACTACTACTTCAACTACTACTACTACTGCTACTGCTACT
GTTGTGAACAGAACACGACTGTTGGACAATCGAATCTAATTAAGCAAACATTACAAACTA

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

SAM and BAM files

sequence alignment/map, and binary version



- Genome analysis
- Transcriptome analysis

Modified from: https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/read-mapping-or

SAM and BAM files

sequence alignment/map, and binary version

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

```
1:497:R:-272+13M17D24M 113    chr1 497  37        37M       15        100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-==9;>>>>>=>>>>>>>>>>>>=>>>>>>>>>>   XT:A:U    NM:i:0
SM:i:37   AM:i:0    X0:i:1    X1:i:0    XM:i:0    XO:i:0    XG:i:0    MD:Z:37
19:20389:F:275+18M2D19M        99       chr1      17644     0         37M       =         17919
314       TATGACTGCTAATAATACCTACACATGTTAGAACCAT   >>>>>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9
RG:Z:UM0098:1       XT:A:R    NM:i:0    SM:i:0    AM:i:0    X0:i:4    X1:i:0    XM:i:0    XO:i:0
XG:i:0    MD:Z:37
19:20389:F:275+18M2D19M        147      chr1      17919     0         18M2D19M  =
17644     -314      GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT   ;44999;499<8<8<<<8<<>><<<<><7<;<<<>><<
XT:A:R    NM:i:2    SM:i:0    AM:i:0    X0:i:4    X1:i:0    XM:i:0    XO:i:1    XG:i:2    MD:Z:
18^CA19
9:21597+10M2I25M:R:-209        83       chr1      21678     0         8M2I27M   =
21469     -244      CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT   <;9<<5><<<<><<<><>><<><>><9><<>>>9>>><>
XT:A:R    NM:i:2    SM:i:0    AM:i:0    X0:i:5    X1:i:0    XM:i:0    XO:i:1    XG:i:2    MD:Z:35
```

https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/slides.html#1

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
```

https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40

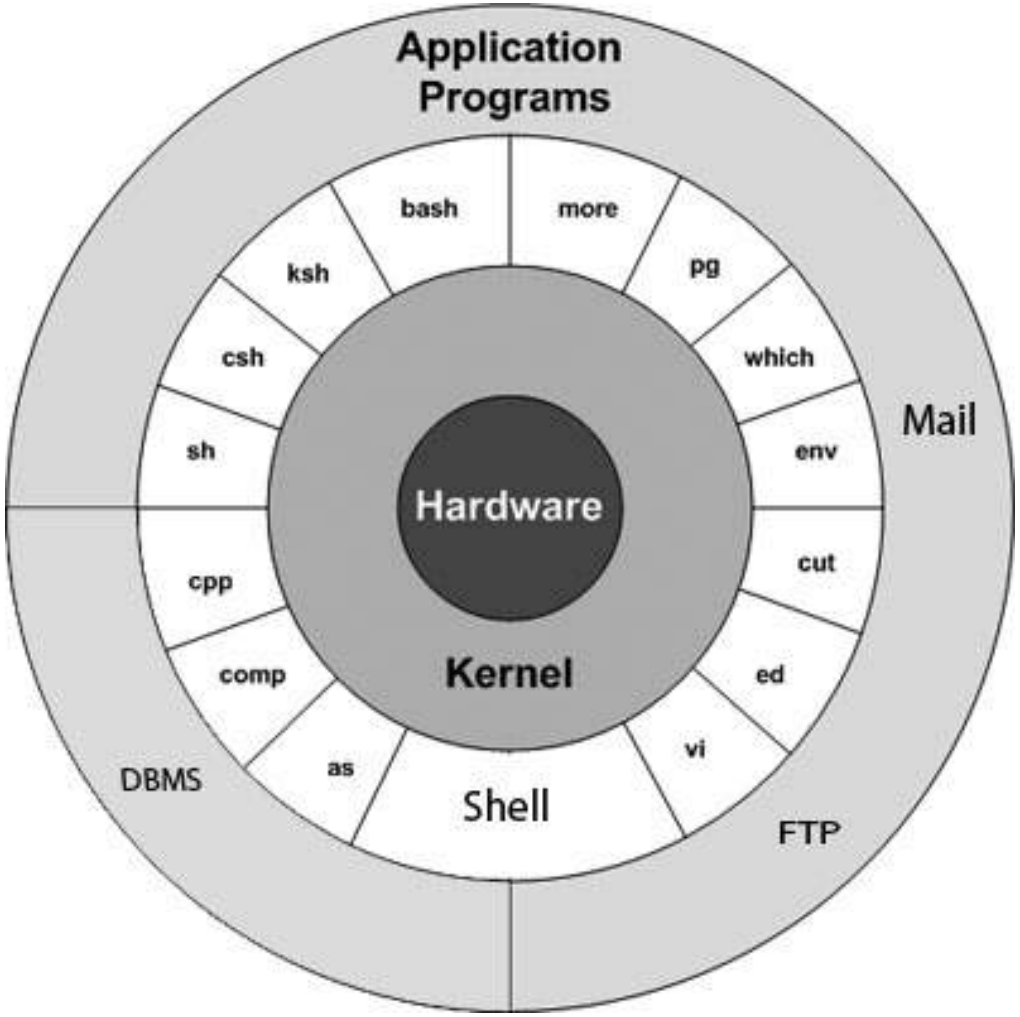
Working with bioinformatics files

- Mostly text files in specific format
- Can be over many gigabytes
- Can have millions of lines

- ... we won’t work manually

Unix

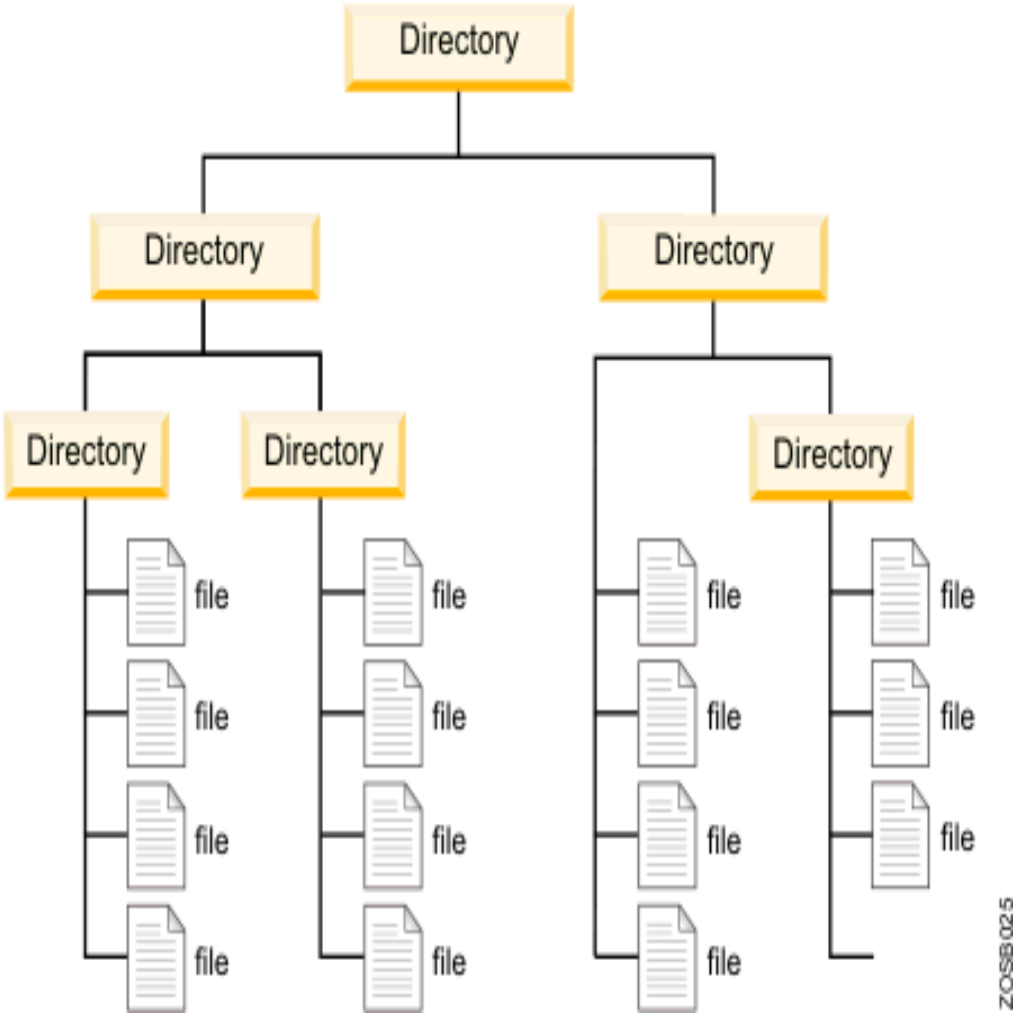
Operating system that underlies a lot of bioinformatics work, and system administration



https://www.tutorialspoint.com/unix/unix-getting-started.htm

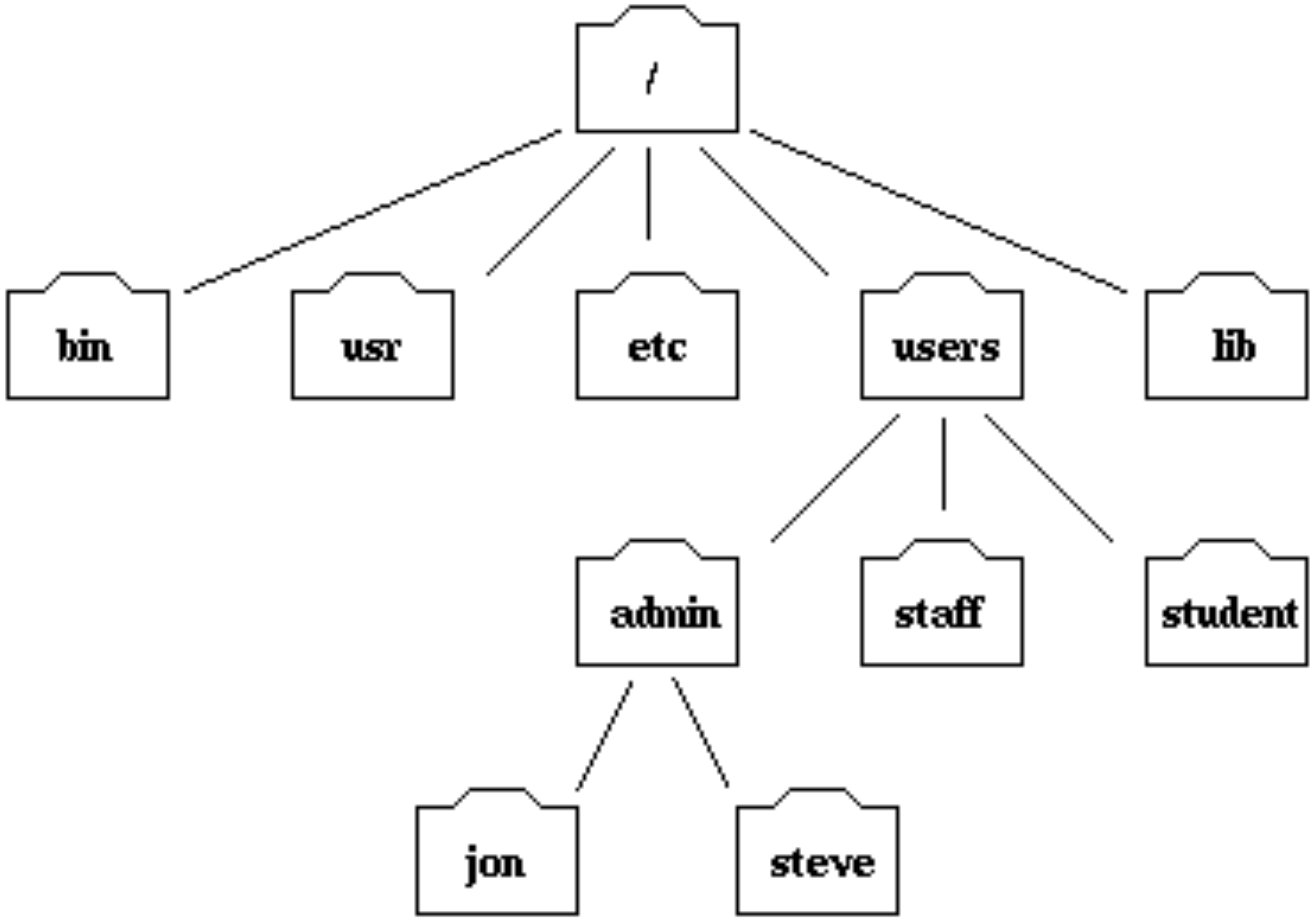
Unix

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



https://www.tutorialspoint.com/unix/unix-getting-started.htm

Unix file system



Part of the filesystem tree

https://rabernat.github.io/research_computing_2018/intro-to-unix.html

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

File permission (ls -l)

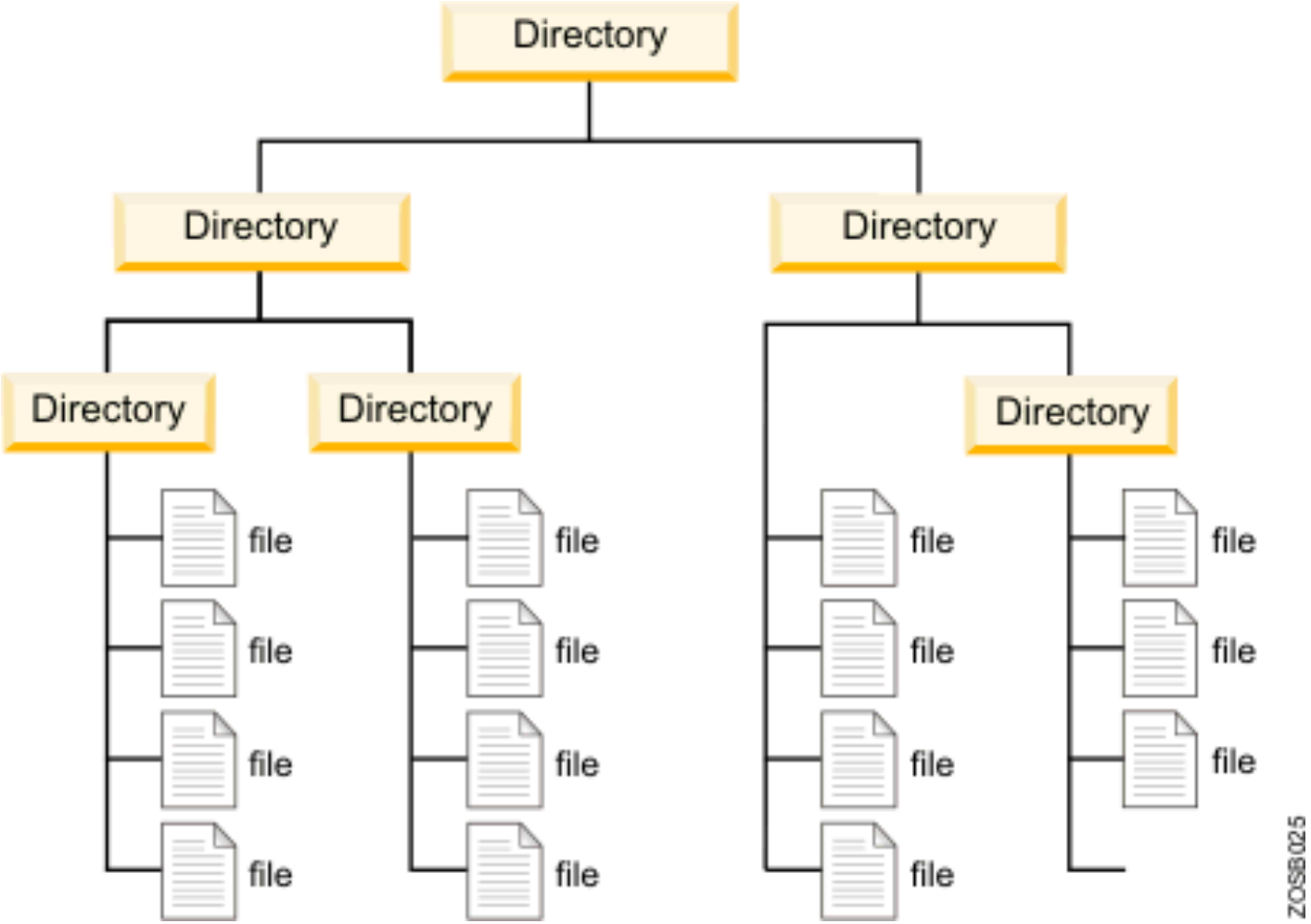
-rwxrw-r--, this means the line displayed is:

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

Read	Write	Execute
4	2	1

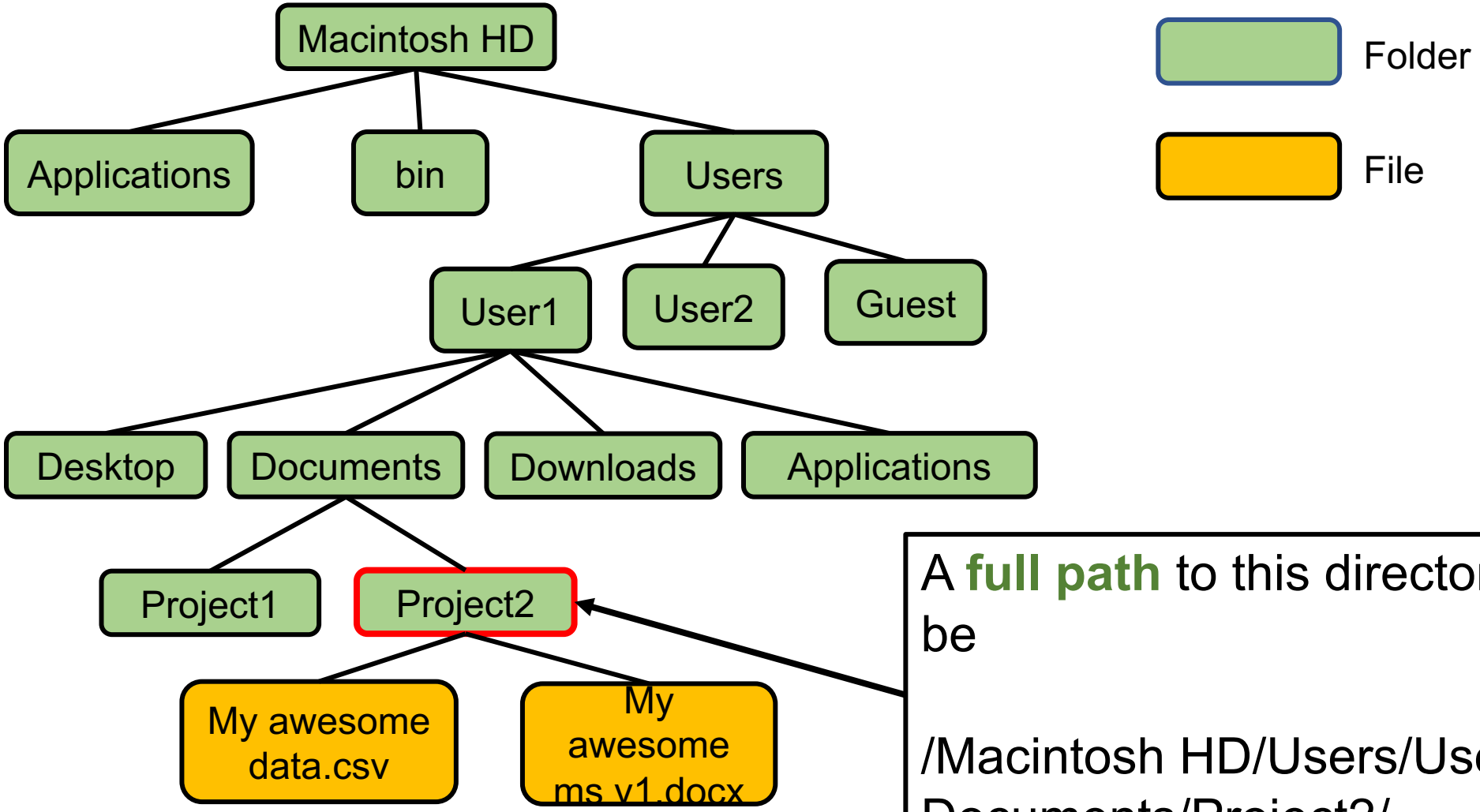
Computer file system

How computer organize files



https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconcepts_177.htm

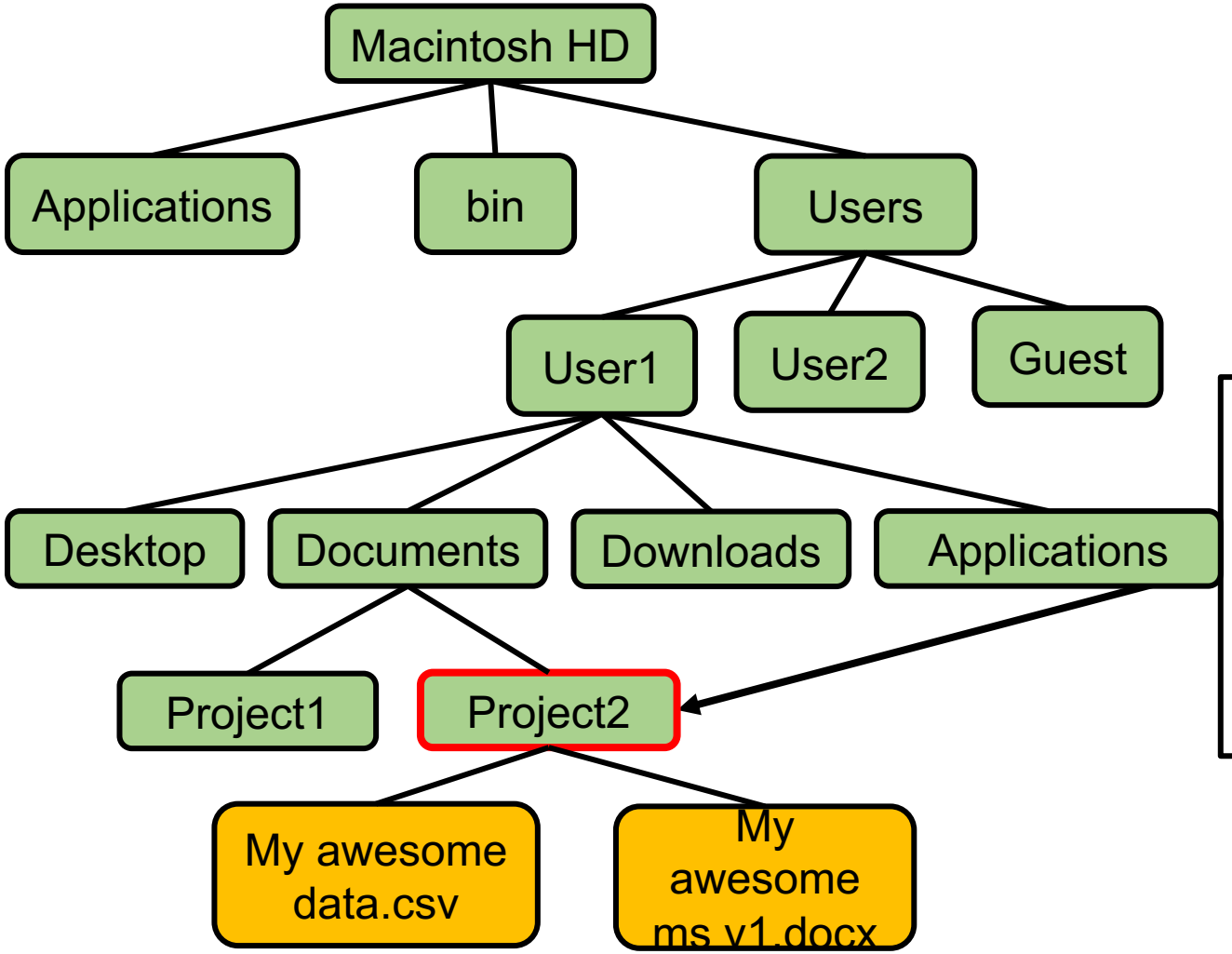
Mac file system (simplified)



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

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

Mac file system (simplified)



Folder

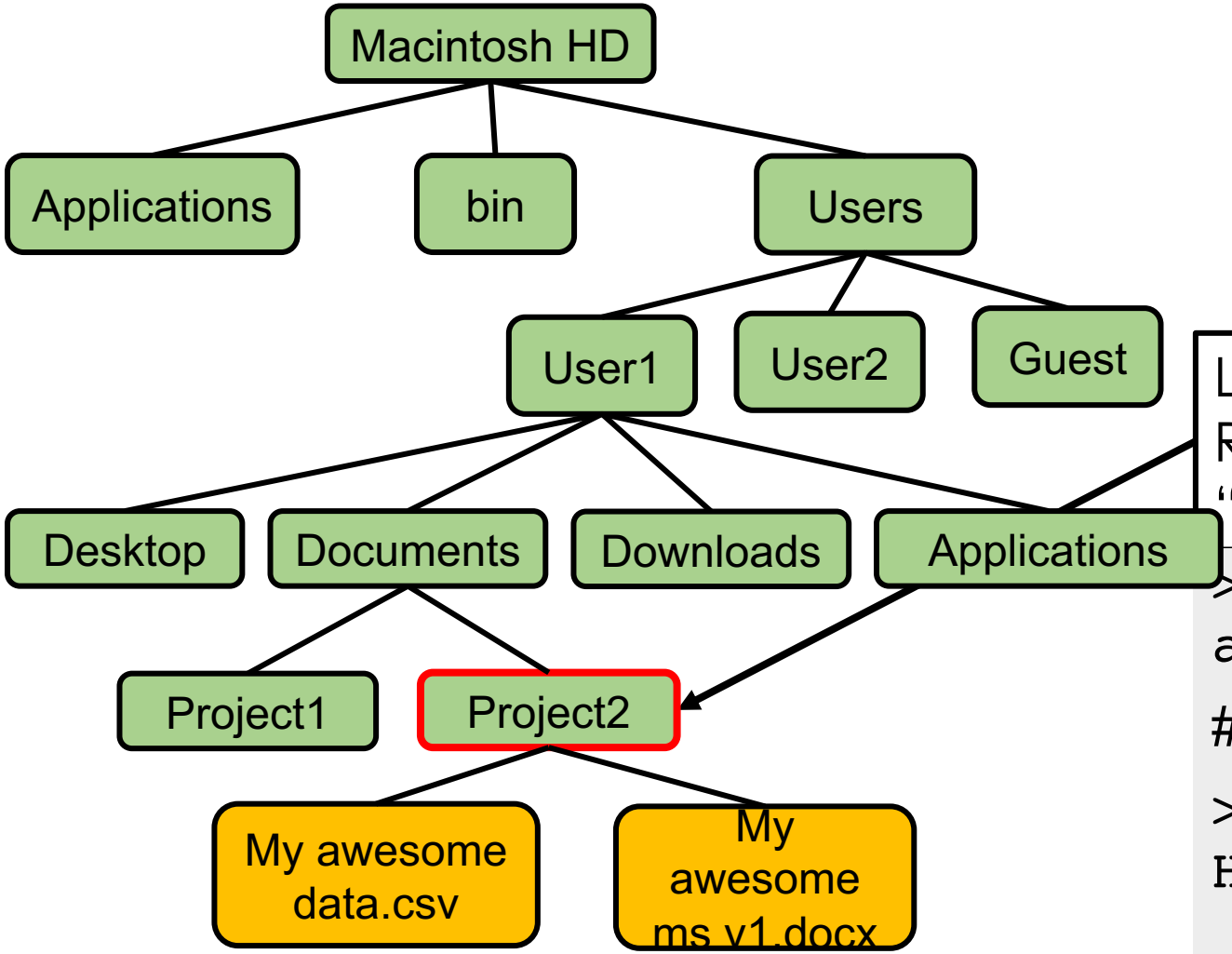
File

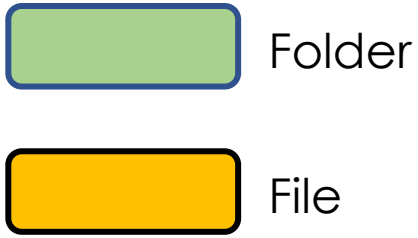
Let's say you are inside Project1 directory, your **relative path** to Project2 would be

`../Project2/`

.. means going up 1 step

Mac file system (simplified)





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")
```

If you want revision/extension on Unix

- http://www.ee.surrey.ac.uk/Teaching/Unix/
- http://swcarpentry.github.io/shell-novice/
- http://jnmaloof.github.io/BIS180L_web/2019/04/02/2Just_Enough_Unix/
- Look for Unix cheatsheet(s)