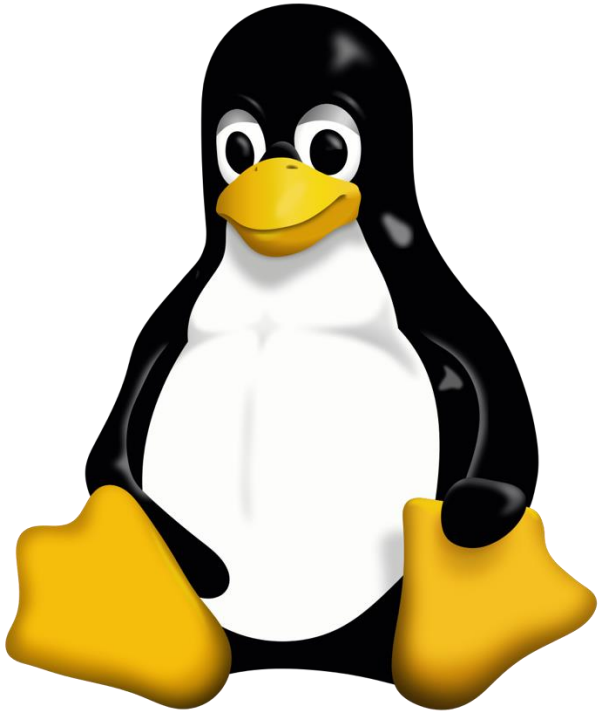
An introduction to Linux OS



A M Mahedi Hasan, PhD | CBiol MRSB | FHEA
Senior Research Associate (Computational Biology)
Treatment Resistance Group
UCL Cancer Institute

Why Linux OS

Linux OS is a
- stable,
- multi-user
- multi-tasking system
- for servers, desktops, and laptops

- The majority of bioinformatics programs and packages are developed on the Linux OS

- Linux was released as free open source software, with its underlying source code publicly available, freely distributed, and freely modified

- particularly suited to working with large text files

Installation of programs on Linux OS

- Installation on Windows computer (.exe or .msi)
- .rpm or .deb installation files can be downloaded from the internet Software Manager and install the software.
- Only download tools from the official website of the company or organisation!
- DEB

  Debian distribution and its derivatives (e.g. Ubuntu, Mint, ...)

- RPM

  Red Hat Package Manager (e.g. Fedora, CentOS, openSUSE, ...)

Installation of programs on Linux OS

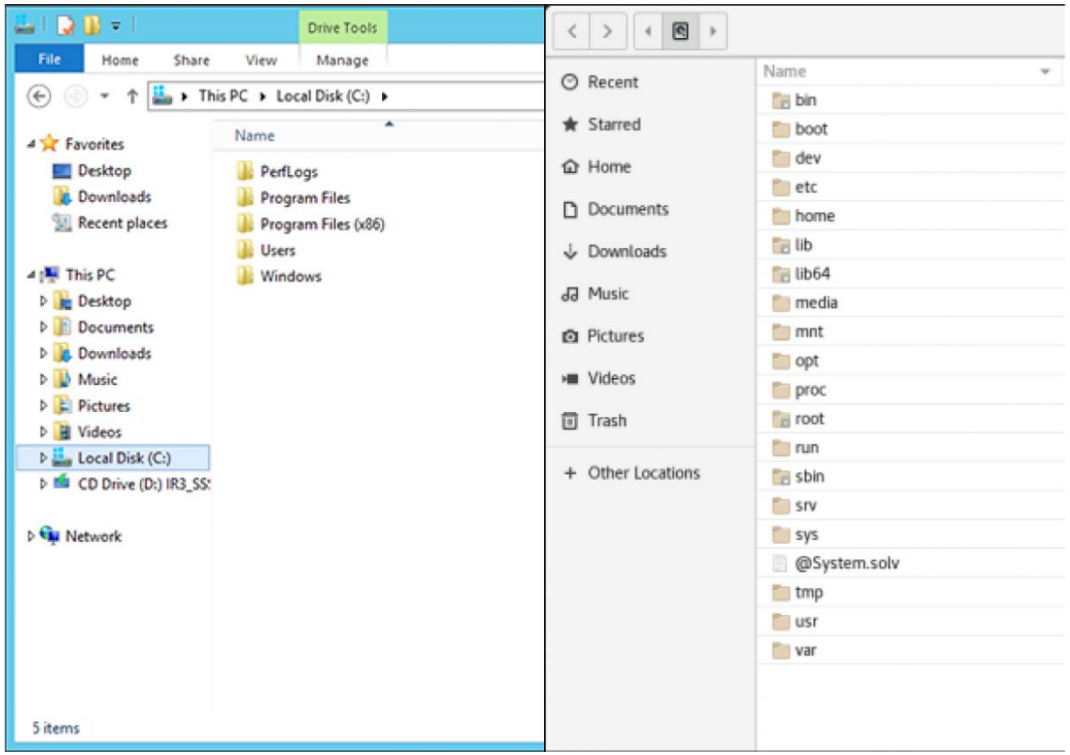
- Dependencies

Software should perform a specific task (stable) and avoid redundant code by reusing other software code. This makes software dependent of one another and therefore creates dependencies between packages.

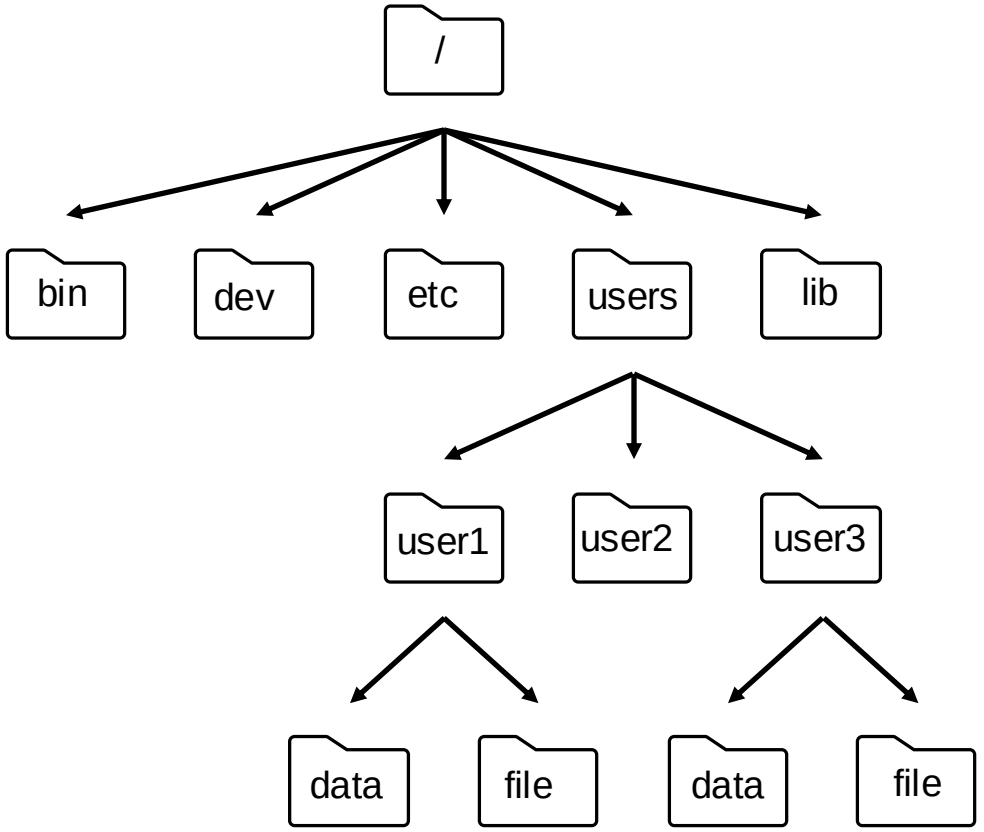
- How do I know?

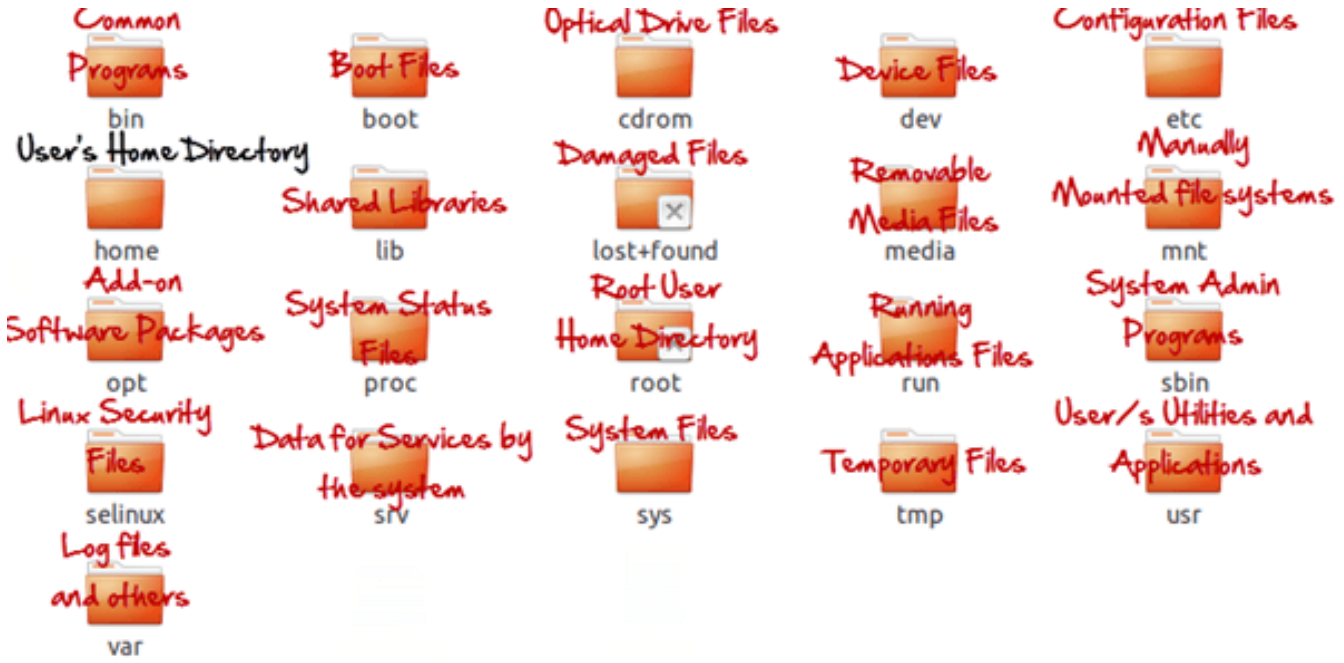
Dependencies need to be co-installed with the software (if not been installed already). When installation software from .rpm or .deb then this will be taken care of. The Software Manager will search for repositories for these dependencies!

Windows OS vs Linux OS

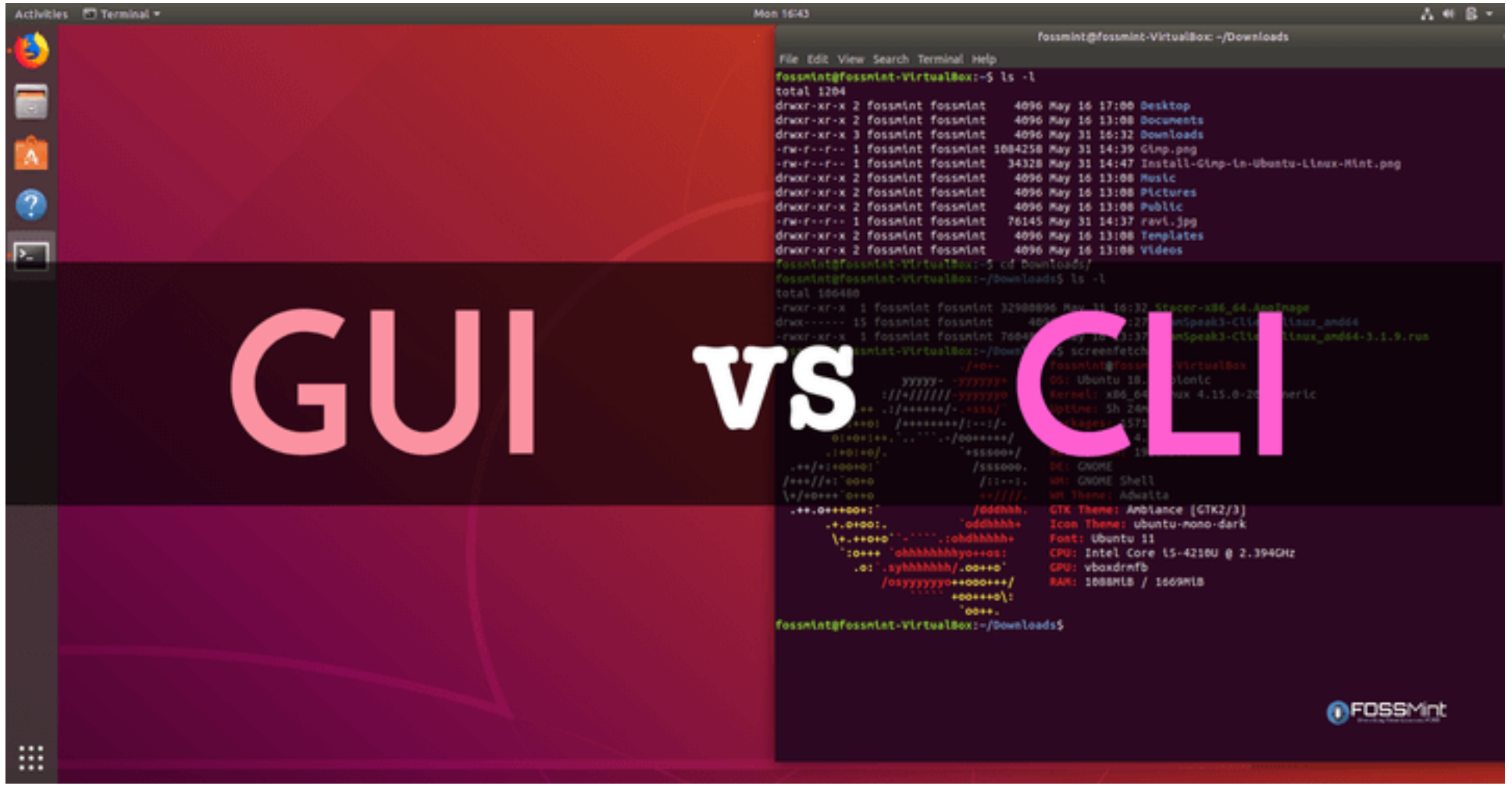


Linux file-system





Graphical vs Command-line interface



Basic Linux commands:

- `pwd` to know the current / working directory
- `cd` to change the directory
- `ls` lists all the contents (directories and files) of a directory.
- `mkdir` creates a new directory
- `touch` creates a new file
- `cp` to copy a file or directory
- `mv` to move / rename a file or directory to a new place
- `rm` to remove / delete a file or directory

Basic Linux commands:

Exercise 01:

- Create a directory “test_dir”.
- Create a file “test_file.txt” and move it to “test_dir”.
- Move to “test_dir”, copy the “test_file.txt” and paste it to its parent directory under the new name “test_file2.txt”
- Remove “test_dir” from /data directory.

Basic Linux commands:

- `vi` or `vim` is to view and edit all the contents of a file
- `cat` is to view all the contents of a file
- `less` is to view a portion of a file
- `more` is to view a portion of a file
- `head` to view from the top of the file
- `tail` to view from the bottom of the file
- `sort` is to sort the content of the files
- `uniq` is to remove duplicates of the content of a file
- `wc` is for word count
- `wget` is for downloading a file from online

Basic Linux commands:

Exercise 02:

- Download a file from the following location:

https://zenodo.org/records/14625079/files/filtered_Cosmic_GenomeScreensMutant.vcf

This is a subsampled Genome Screen Mutation vcf file which contains the mutations in several important genes.

- Open the file with **less** command. Look at the structure of the file. Get out of the file.
- Print first 9 lines of the file.
- Print last 6 lines of the file
- How many lines are there in that file?

Basic Linux commands:

- `grep` is to find/search for a specific word or pattern in a file
- `*` is a wild card
- `^` represents the beginning of a line
- `$` represents the end of a line
- `sed` is to manipulate the content of a file at different levels
- `>` is to redirect the contents or output of an operation
- `>>` is to append the content
- `zip, gzip, tar` to compress files
- `unzip, gunzip, tar` to decompress files.

Basic Linux commands:

Exercise 03:

- How many Mutations are there in the vcf file?
- Print out the mutation information of *TP53* gene?
- How many different genes can you find in the vcf file?
- How many entries are there for BRCA genes?
- Advanced question: How many unique mutations are there in *TP53*?

The end