SLU
SCIENCE AND EDUCATION FOR SUSTAINABLE LIFE

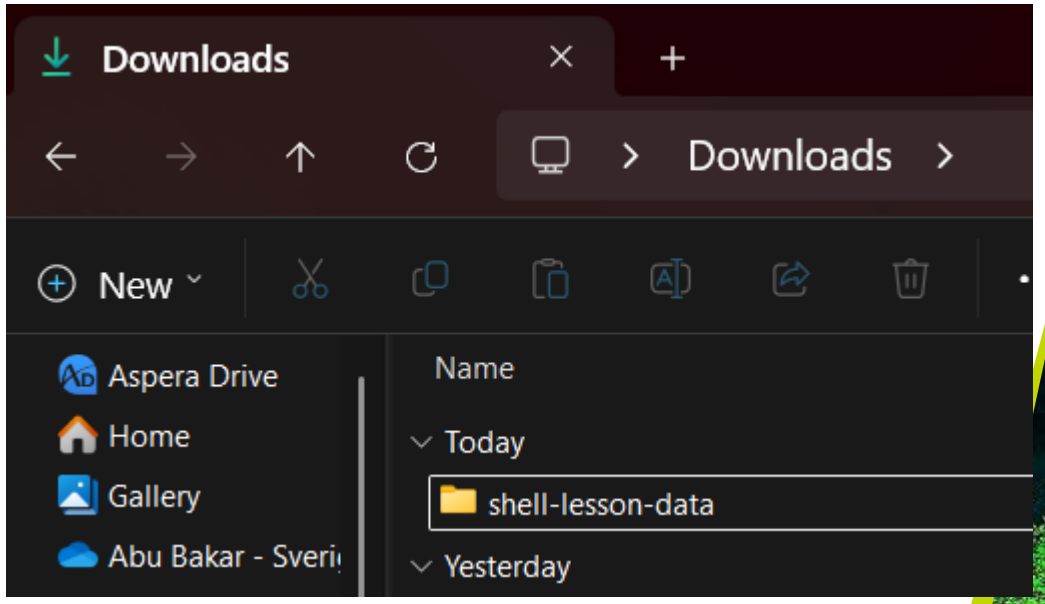
Introduction to Linux & R

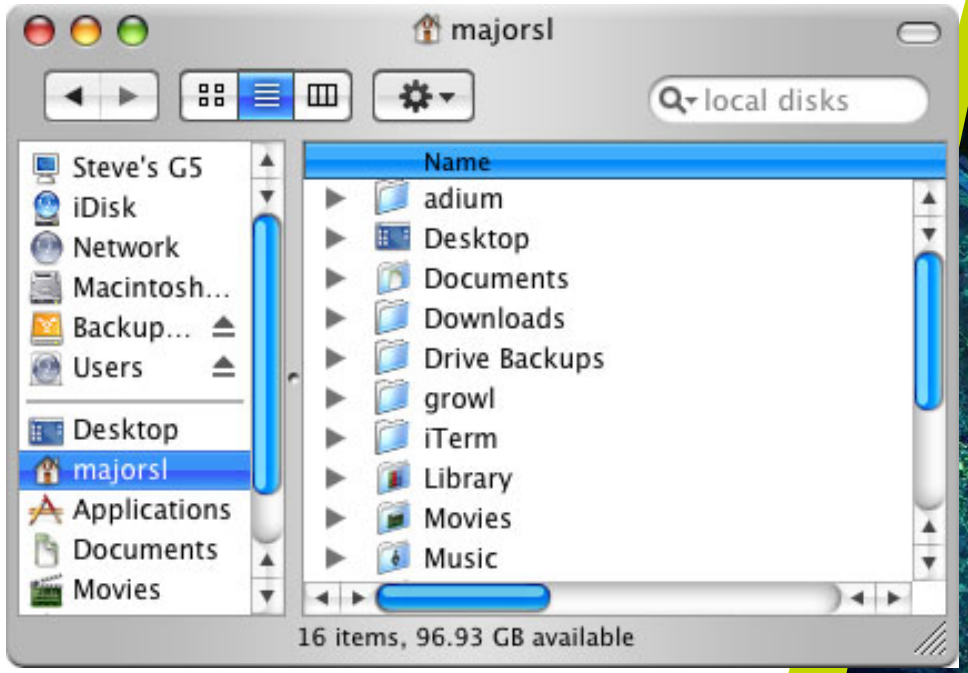
Abu Bakar Siddique

Bioinformatician, SLUBI

Introduction to Linux

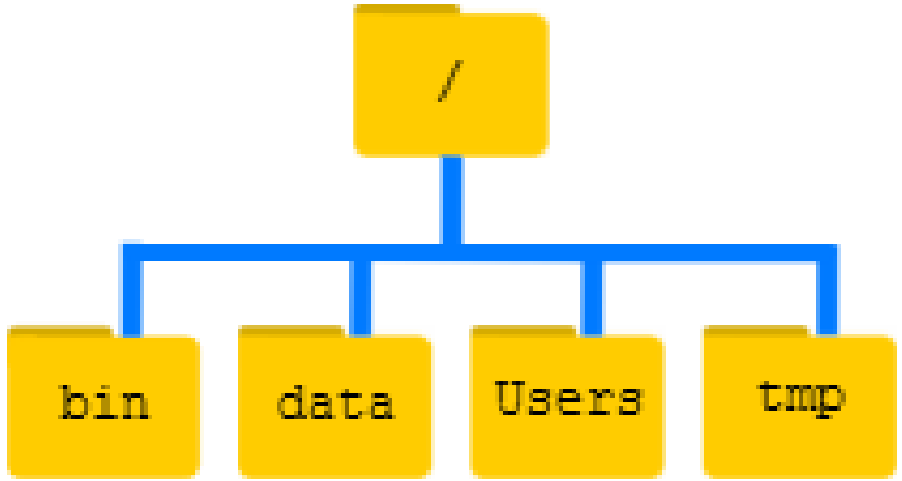
GUI

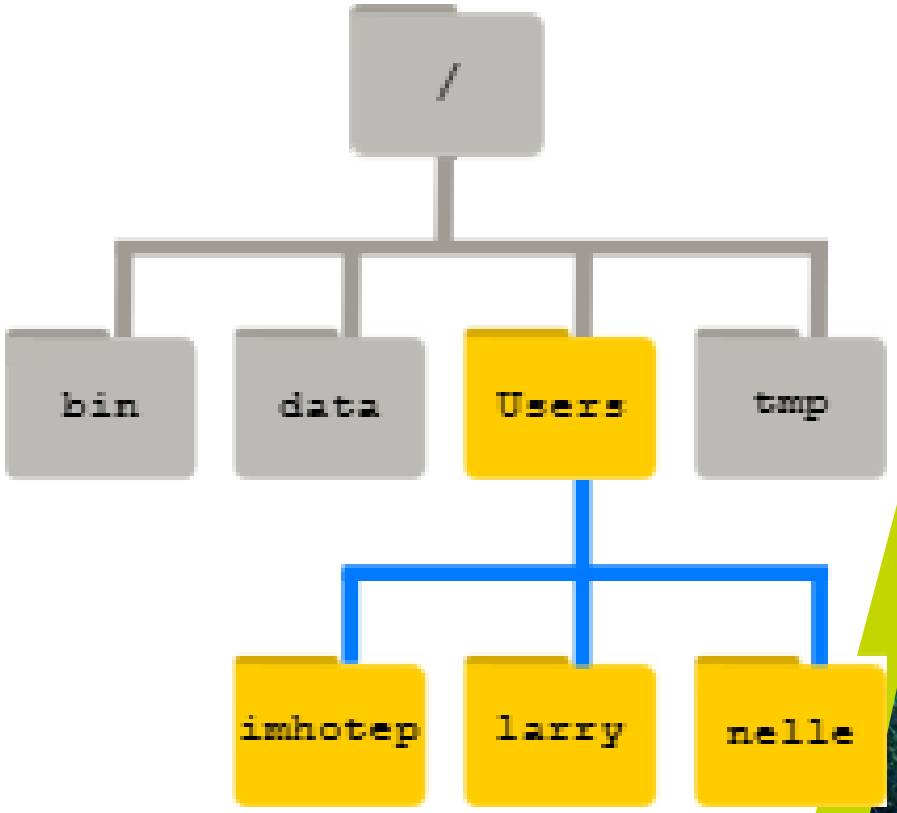

Downloads
Downloads
New
Aspera Drive
Home
Gallery
Abu Bakar - Sveri
Name
Today
shell-lesson-data
Yesterday


majorsl
local disks
Steve's G5
iDisk
Network
Macintosh...
Backup...
Users
Desktop
majorsl
Applications
Documents
Movies
Name
adium
Desktop
Documents
Downloads
Drive Backups
growl
iTerm
Library
Movies
Music
16 items, 96.93 GB available

CLI (Linux)

```
abuslu@vbsg-asiddique
(base) abuslu@vbsg-asiddique:/$ ls
bin   init    lost+found  root  sys         wslOLKdLM
boot  lib     media       run   tmp         wslPbmDDl
dev   lib32   mnt         sbin  usr         wsleOlmKM
etc   lib64   opt         snap  var         wslfAACGl
home  libx32  proc        srv   wslBcLjOM   wslnAMOEl
(base) abuslu@vbsg-asiddique:/$
```





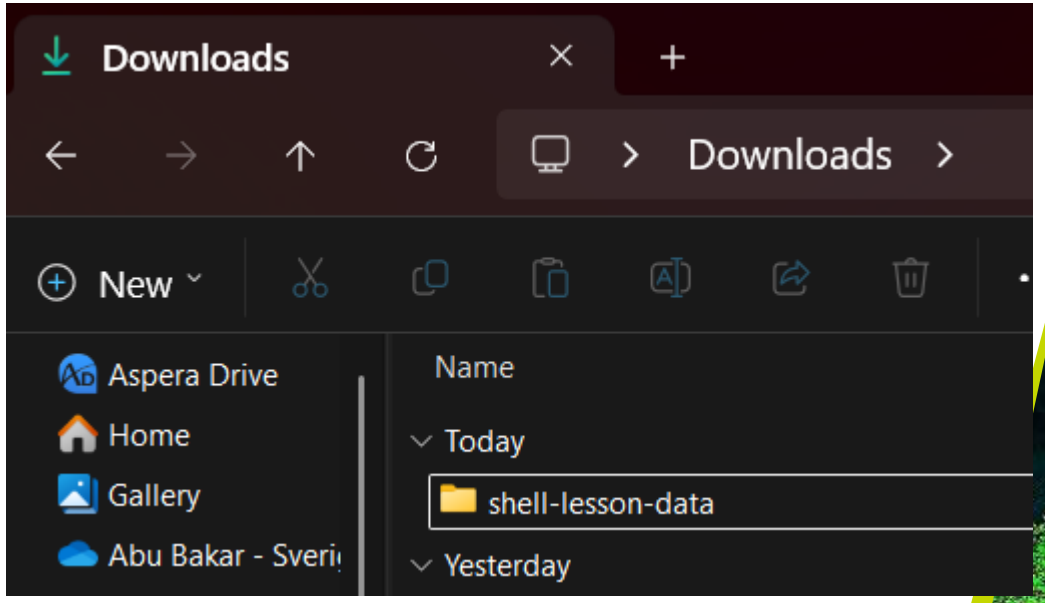
GUI vs Linux

Graphical **U**ser **I**nterface (**GUI**):

- **visual** representations

Command **L**ine **I**nterface (**CLI**): **WSL, Terminal**, LINUX, Ubuntu, UNIX etc

- **text** representations
- ***shell*** (or bash) is a program that allows you to control **CLI** with keyboard



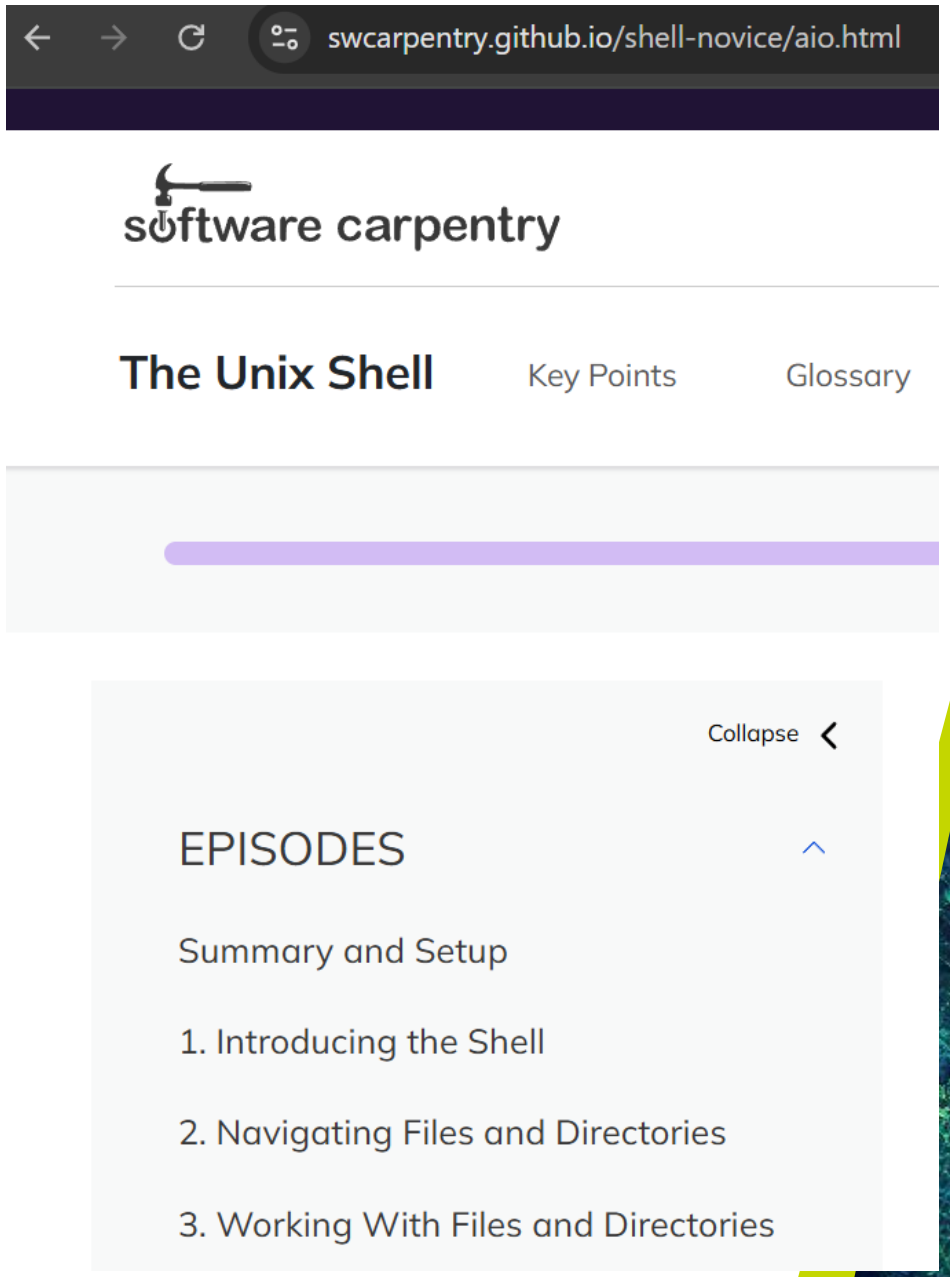
Why Linux?

- A **powerful** command-line interface (CLI)
- Built for **multi-user systems**.
- An enormous amount of **software is free**ly available
- **Bioinformatics tools** can only be used **through a CLI**

- *shell* makes your work **less boring**
- *shell* makes your work **less error-prone**
- *shell* makes your work more **reproducible**
- Computationally **intensive job**
- Remote computer or **cloud computing**

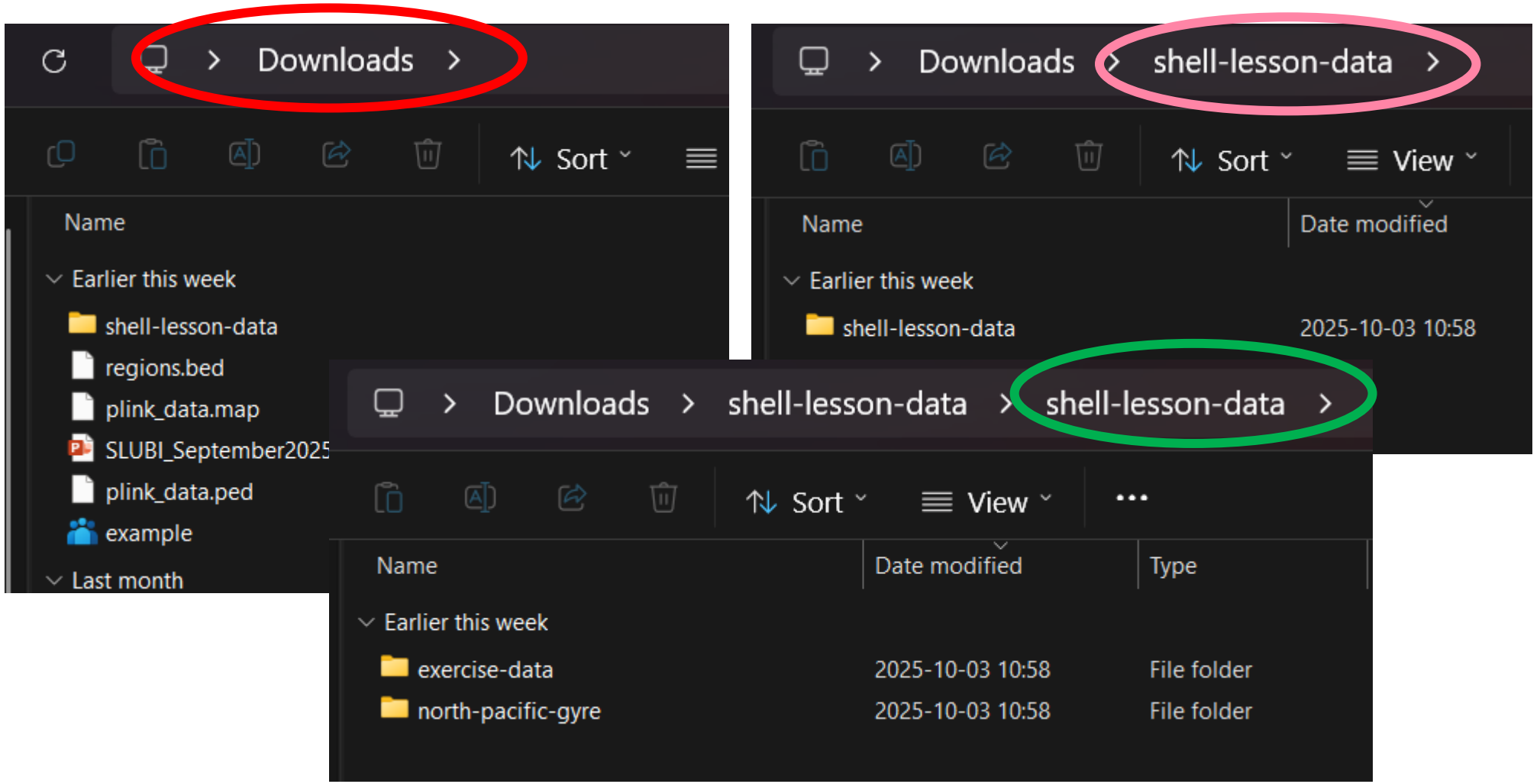
Full Tutorial

- The Unix shell



Navigating your file system

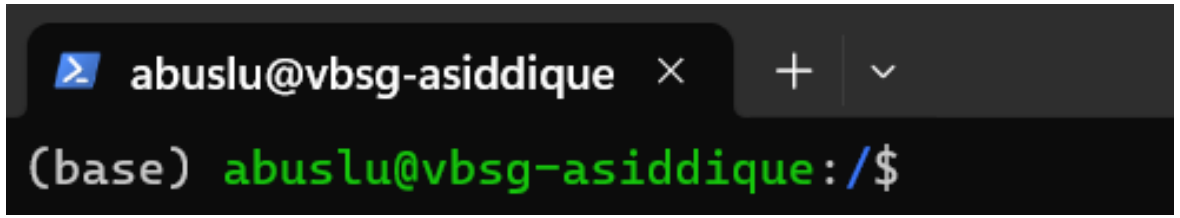
Navigating your file system



Navigating your file system

Linux way

- `cd` – change working directory

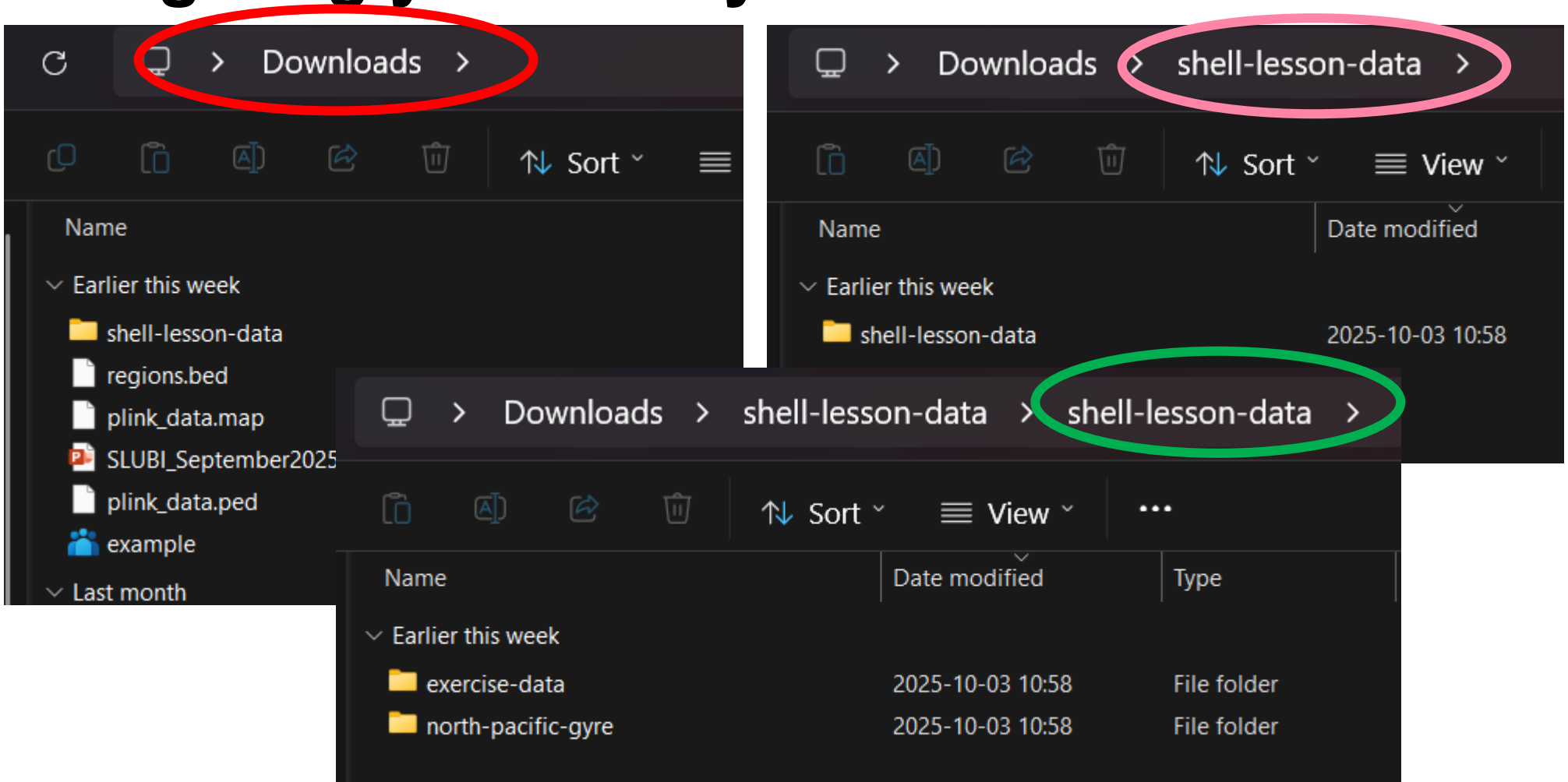



```
(base) abuslu@vbsg-asiddique:~$ cd /mnt/c/Users/auue0001/Downloads
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads$
```

```
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads$ cd shell-lesson-data/
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads/shell-lesson-data$
```

```
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads/shell-lesson-data$ cd shell-lesson-data/
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads/shell-lesson-data/shell-lesson-data$
```

Navigating your file system



Navigating your file system

- `ls` – list content of directory

```
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads/shell-lesson-data/shell-lesson-data$ ls
exercise-data  north-pacific-gyre
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads/shell-lesson-data/shell-lesson-data$
```

```
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads$ ls
 1748541649-takbir-muna-family-dev-10-steps-imam-ajmal-masroor.png
 1750352364-Imam-Ajmal-2.jpg
 1750352364-Imam-Ajmal-2.png
 2023_AMRI_SLUBI_Practical_considerations_and_available_tools_for_big_data_science_abp.pptx
'2025-08-26 NOC Secretariat meeting summary.jpg'
 223256388_invoice.pdf
'3.daily temperature 7 and 18 stataion-20250904T073127Z-1-001'
'3.daily temperature 7 and 18 stataion-20250904T073127Z-1-001.zip'
```

Navigating your file system

- `pwd` – present working directory

```
(base) abuslu@vbsg-asiddique:~$ pwd
/home/abuslu
```

```
(base) abuslu@vbsg-asiddique:/mnt/c/Users/auue0001/Downloads/shell-lesson-data/shell-lesson-data$ pwd
/mnt/c/Users/auue0001/Downloads/shell-lesson-data/shell-lesson-data
```

Absolute/Relative path

- Absolute paths
  - Starts with a /
  - Defined **from the root** of the file system

```
/mnt/c/Users/auue0001/Downloads/shell-lesson-data/shell-lesson-data/
```

  - Like an **address**:

    Sweden, Uppsala, Ultuna, Biocentre, 4th floor, room A436

Absolute/Relative path

- Relative paths
  - **Does not start with a /**
  - **Relative** to where you are when using it

```
exercise-data/numbers.txt
```

```
../exercise-data/numbers.txt
```

  - Like a **direction** **(from Campus ultuna):**
    ‘straight, then turn left, 4th floor, right, then left’

Summary

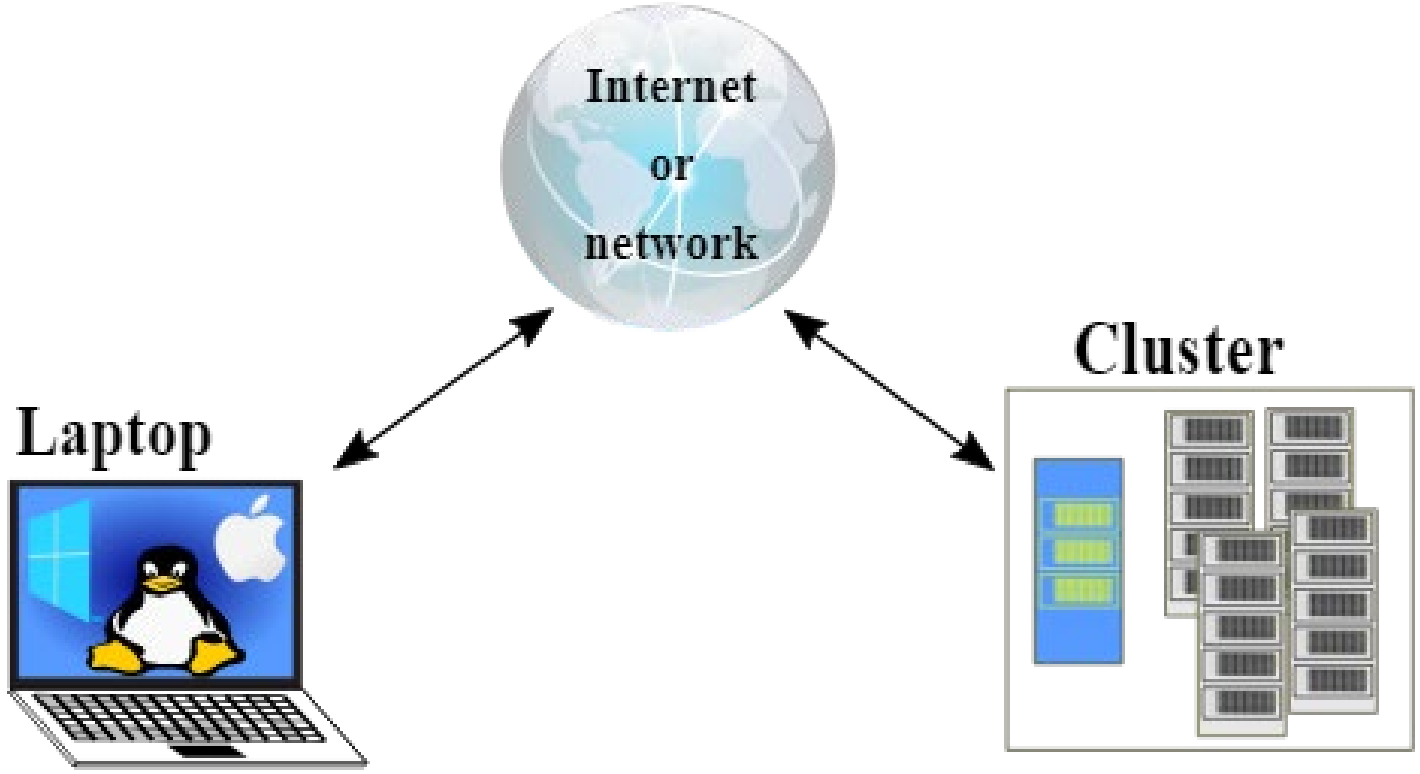
- `ls` – list content of directory
- `cd` – change working directory
- `pwd` – print working directory
- absolute/relative paths

More practice *(after lunch)*

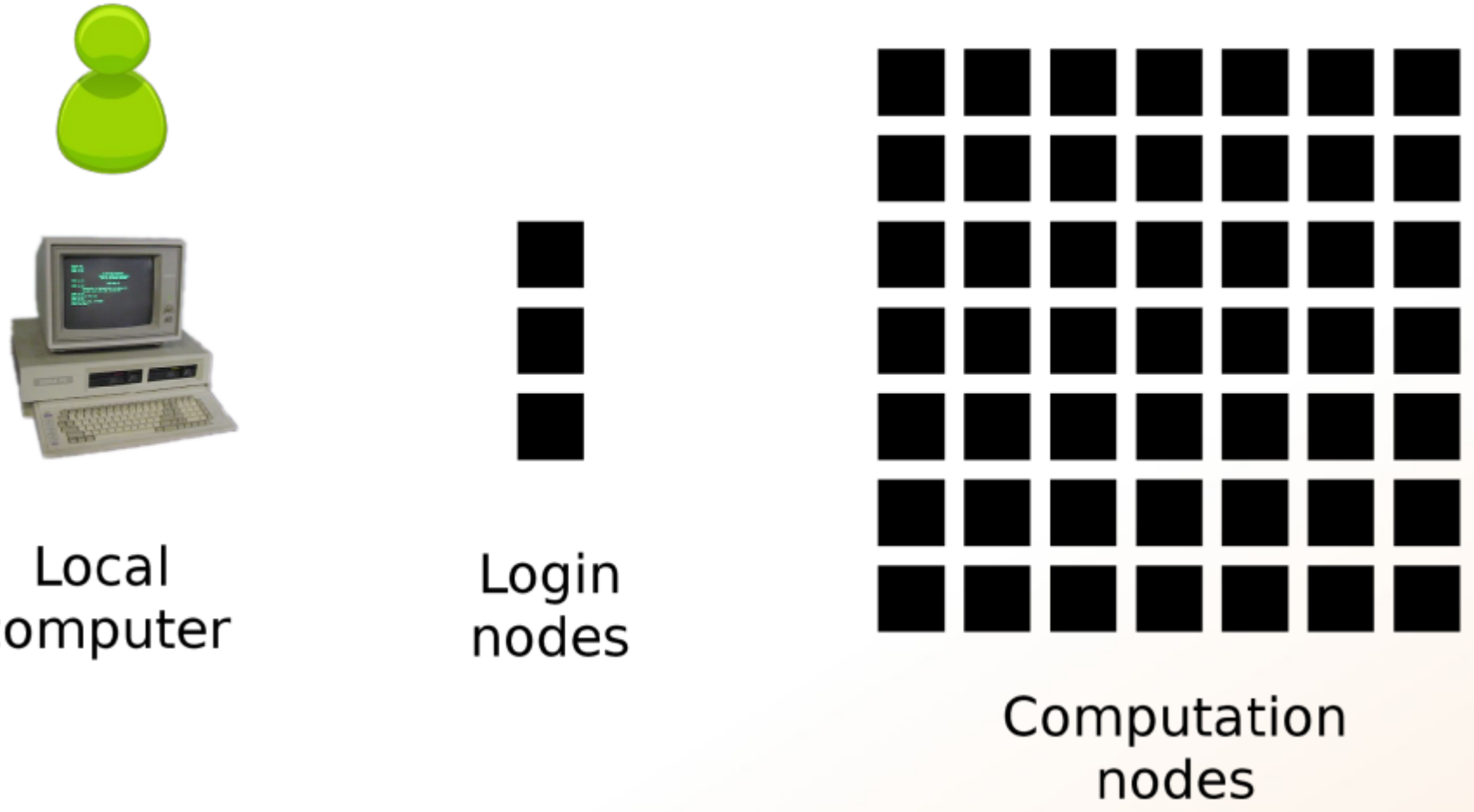
- ls – list content of directory
- cd – change working directory
- pwd – print working directory
- absolute/relative paths
- **mkdir – make a directory**
- **cp – copy a file**
- **mv – move a file**
- less – view a file
- nano – view and edit a file
- rm – remove a file
- head / tail
- **grep**
- find
- wildcards
- tab completion – use it!!!
- top – see active processes
- man – manual pages or anycommand --help
- exit – Log out current terminal

30 minutes…

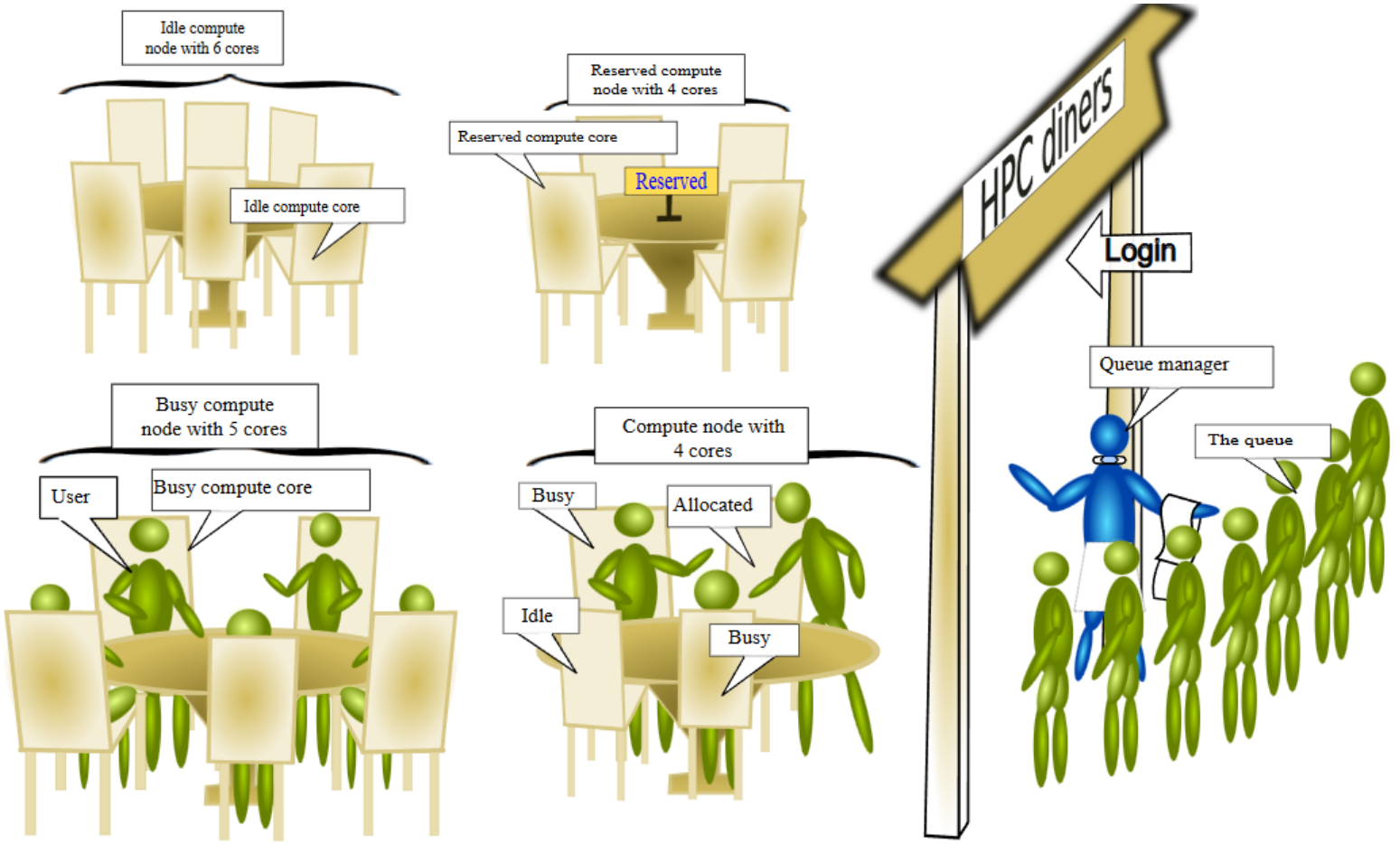
Computer cluster



Computer cluster



Computer cluster



CC-BY-2.0, sabryr

Online materials

- Unix/Linux
- Data Carpentry datacarpentry/shell-genomics | website
- Teaching materials at the Harvard Chan Bioinformatics Core hbctraining/Intro-to-shell-flipped | website
- Ted Laderas laderast/bash_for_bioinformatics | website
- SIB Swiss Institute of Bioinformatics sib-swiss/unix-first-steps-training
- Software Carpentry swcarpentry/shell-novice | website
- The Carpentries Lab carpentries-lab/metagenomics-shell | website
- UC Davis Bioinformatics Core Training Page ucdavis-bioinformatics-training/2023-September-Introduction-to-the-Command-Line-for-Bioinformatics

Navigating your file system

- Online terminal
- https://www.terminaltemple.com/
- https://copy.sh/v86/?profile=linux26&ref=itsfoss.com
- 17 Best Online Linux Terminals and Online Bash Editors

Introduction to basic R

Contents

- About R
- R vs Rstudio
- Timeline
- Setting a project or working directory
- Interacting with R
- Packages
- Assign values or objects
- Data types
- Data structures
- How to import and export data or results
- Housekeeping

Basic about R

R is ...

- a programming language
- a programming platform (= environment + interpreter)
- a software project driven by the core team and the community
- a very powerful tool for statistical computing
- a very powerful computational tool in general

Yet ...

- it is very elegant
- it becomes more and more feature-rich

R is not ...

- a tool to replace a statistician
- the very best programming language
- the most elegant programming solution
- the most efficient programming language

R: Engine	RStudio: Dashboard

R

- Programming language
- For data analysis and graphics
- Refers to both the language and the software that interprets it's scripts
- Free and open source

RStudio

- User interface for working with R
- Wrapper around the R language
- Extends what R can do and facilitates writing R code
- Free and open source

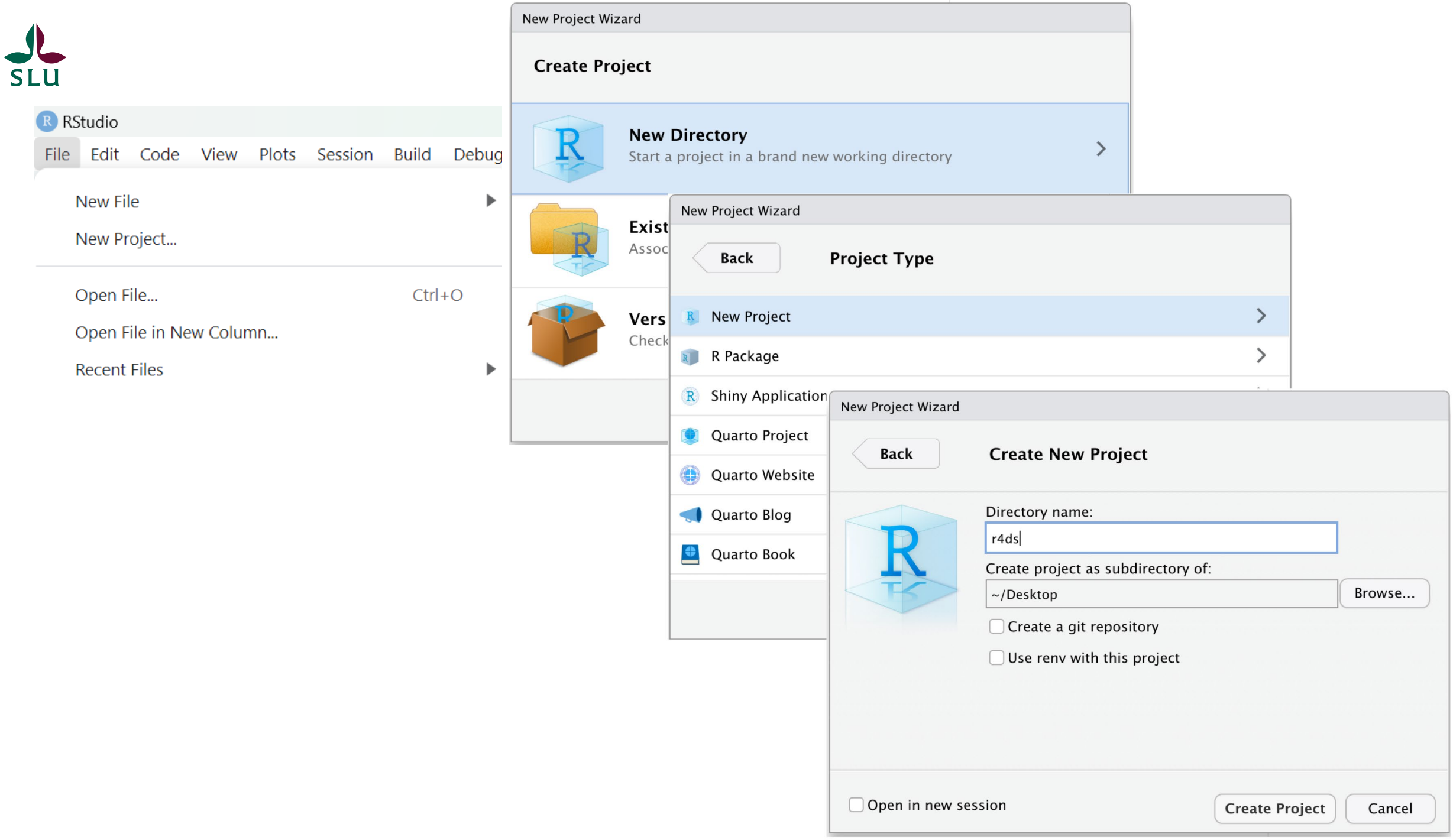
Timeline

- ca. 1992 — conceived by Robert Gentleman and Ross Ihaka (R&R) at the University of Auckland, NZ as a tool for **teaching statistics**
- 1994 — initial version
- 2000 — stable version

- 2011 — RStudio, first release by J.J. Allaire

- ca. 2017 — Tidyverse by Hadley Wickham

Setting up a project

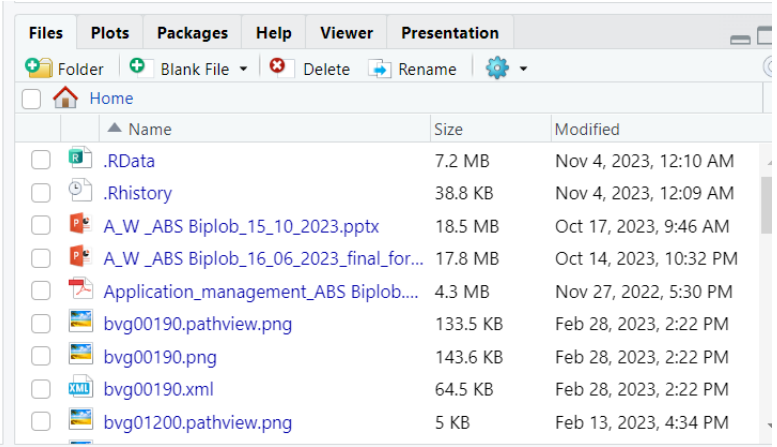

RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
New File
New Project...
Open File... Ctrl+O
Open File in New Column...
Recent Files
New Project Wizard
Create Project
New Directory
Start a project in a brand new working directory
Exist
Assoc
Vers
Check
New Project Wizard
Back
Project Type
New Project
R Package
Shiny Application
Quarto Project
Quarto Website
Quarto Blog
Quarto Book
New Project Wizard
Back
Create New Project
Directory name:
r4ds
Create project as subdirectory of:
~/Desktop
Browse...
Create a git repository
Use renv with this project
Open in new session
Create Project
Cancel

The working directory

- Where R will look for and save files
- Check working directory with the getwd() , setwd() functions

```
> getwd()
[1] "C:/Users/auue0001/OneDrive - Sveriges lantbruksuniversitet/Dokument"
```

```
> setwd("C:/Users/auue0001/OneDrive - Sveriges lantbruksuniversitet/Dokument")
```



Suggested subdirectories

- data/ for raw data and intermediate datasets
- data_output/ modified versions of raw data
- documents/ outlines, drafts, other text
- fig_output/ graphics generated by scripts
- scripts/ R scripts for different analyses or plotting

```
# create subdirectory
dir.create(path = “path_to_working_directory/data_output")
```

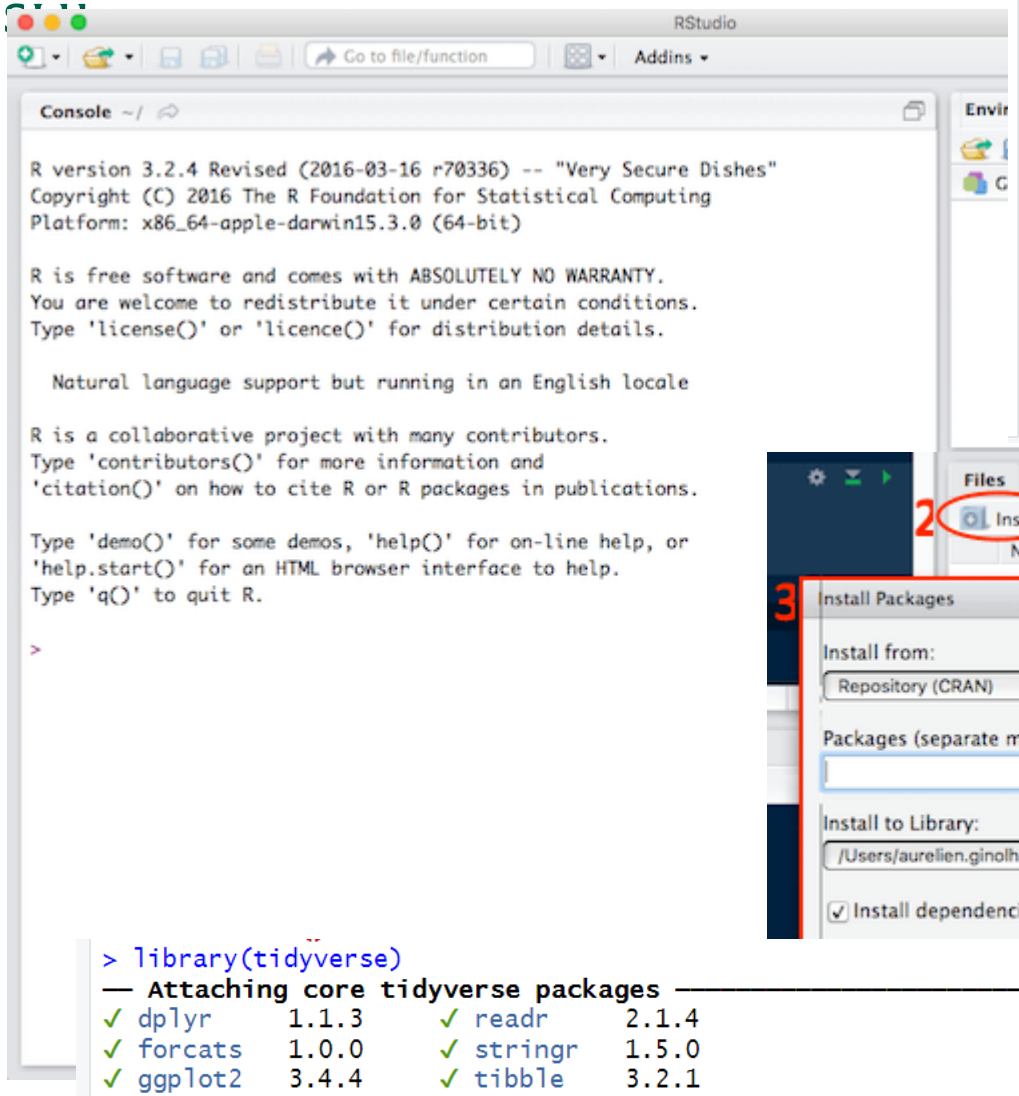
Interacting with R

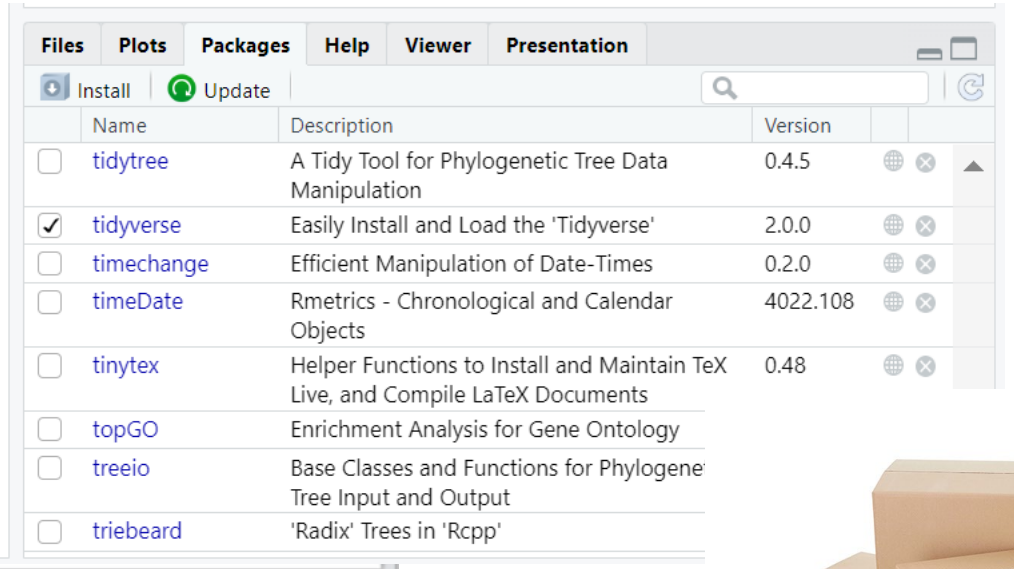
Packages

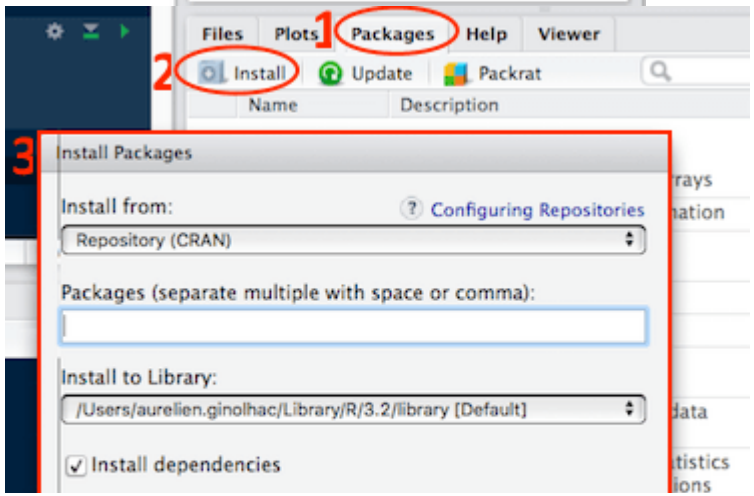
- developed by the community
- cover several very diverse areas of science/life
- uniformly structured and documented
- organized in repositories:
  - CRAN

```
install.packages("tidyverse")

library("tidyverse")
```







```
> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.3     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.4     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
```

Packages

- developed by the community
- cover several very diverse areas of science/life
- uniformly structured and documented
- organized in repositories:
  - CRAN
  - Bioconductor
  - R-Forge
  - GitHub

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.18")
```

Assign values to objects

Assign values to objects

```
p <- 3
```

```
x <- c(1, 6, 8)
```

```
y <- c("car", "truck")
```

Scalars, vectors and matrices

Types of vectors and objects

- **Scalar** (0 dimention):
  p <- 3

- **Vector** (arrays -1 dimention) :
  x <- c(1, 6, 8)
  y <- (“car”, “truck”)

```
> vec1 = c(1,4,6,8,10)
> vec1
[1]  1  4  6  8 10
> vec1[5]
[1] 10
```

- **Matrix** (2 dimension):

```
mat=matrix(data=c(9,2,3,4,5,6),ncol=3)
> mat
     [,1] [,2] [,3]
[1,]    9    3    5
[2,]    2    4    6
```

Types of vectors and objects

- **character**: y <- (“car”, “truck”)
- **numeric**: x <- c(1.1, 6.2, 8.4)
- **integer**: z <- 2
- **logical**: TRUE, FALSE
- **complex**: 1+4i (complex numbers with real and imaginary parts)

R has many functions to examine features of vectors and other objects:,

•class() - what kind of object is it (high-level)?

•typeof() - what is the object’s data type (low-level)?

•length() - how long is it? What about two dimensional objects?

•attributes() - does it have any metadata?

Structures

Data structures

- Matrix
- Data frames
- Factors (r assign a level for each values)
- Arrays
- Lists

```
> x <- 1:12

> # Create a 3 x 4 array from the vector
> my_array <- array(values, dim = c(3, 4))

> # Print the array
> print(my_array)
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> values
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

```
mat=matrix(data=c(9,2,3,4,5,6),ncol=3)
> mat
     [,1] [,2] [,3]
[1,]    9    3    5
[2,]    2    4    6
```

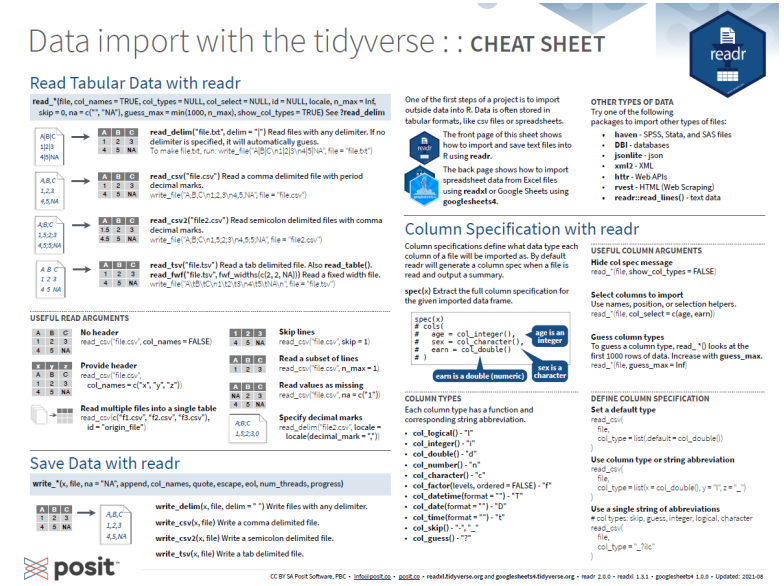
```
> t = data.frame(x = c(11,12,14),
 y = c(19,20,21), z = c(10,9,7))
> t
   x y z
1 11 19 10
2 12 20 9
3 14 21 7
```

```
> L = list(one=1, two=c(1,2),
 five=seq(0, 1, length=5))
> L
$one
[1] 1
$two
[1] 1 2
```

Getting a dataset in R

Ways to get data in R or in RStudio

- **Manually:**
  - data.frame() function in Base R,
  - or the tibble() function in the tidyverse.
- **Import it from a file:**
  - Tabular data: CSV, TSV (read.table() or read_delim() function or readr package which contains read_csv())
  - Excel: XLSX (xlsx package)
  - Google sheets: (googlesheets package)



A|B|C
1|2|3
4|5|NA

A	B	C
1	2	3
4	5	NA

read_delim("file.txt", delim = "|") Read files with any delimiter. If no delimiter is specified, it will automatically guess.
To make file.txt, run: write_file("A|B|C\n1|2|3\n4|5|NA", file = "file.txt")

A,B,C
1,2,3
4,5,NA

A	B	C
1	2	3
4	5	NA

read_csv("file.csv") Read a comma delimited file with period decimal marks.
write_file("A,B,C\n1,2,3\n4,5,NA", file = "file.csv")

A;B;C
1,5;2;3
4,5;5;NA

A	B	C
1.5	2	3
4.5	5	NA

read_csv2("file2.csv") Read semicolon delimited files with comma decimal marks.
write_file("A;B;C\n1,5;2;3\n4,5;5;NA", file = "file2.csv")

A B C
1 2 3
4 5 NA

A	B	C
1	2	3
4	5	NA

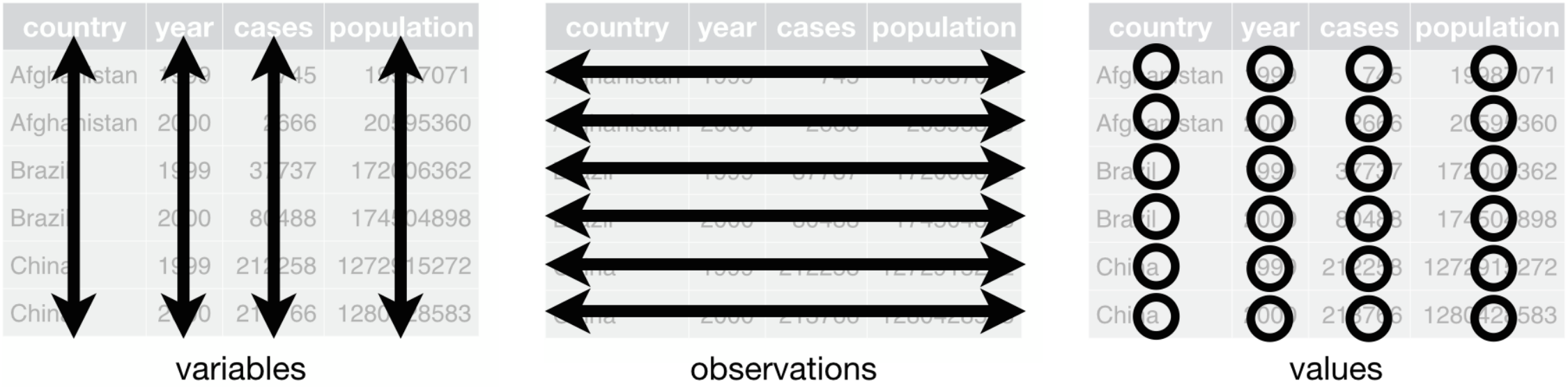
read_tsv("file.tsv") Read a tab delimited file. Also **read_table()**.
read_fwf("file.tsv", fwf_widths(c(2, 2, NA))) Read a fixed width file.
write_file("A\tB\tC\n1\t2\t3\n4\t5\tNA\n", file = "file.tsv")

```
Student ID,Full Name,favourite.food,mealPlan,AGE
1,Sunil Huffmann,Strawberry yoghurt,Lunch only,4
2,Barclay Lynn,French fries,Lunch only,5
3,Jayendra Lyne,N/A,Breakfast and lunch,7
4,Leon Rossini,Anchovies,Lunch only,
5,Chidiegwu Dunkel,Pizza,Breakfast and lunch,five
6,Güvenç Attila,Ice cream,Lunch only,6
```

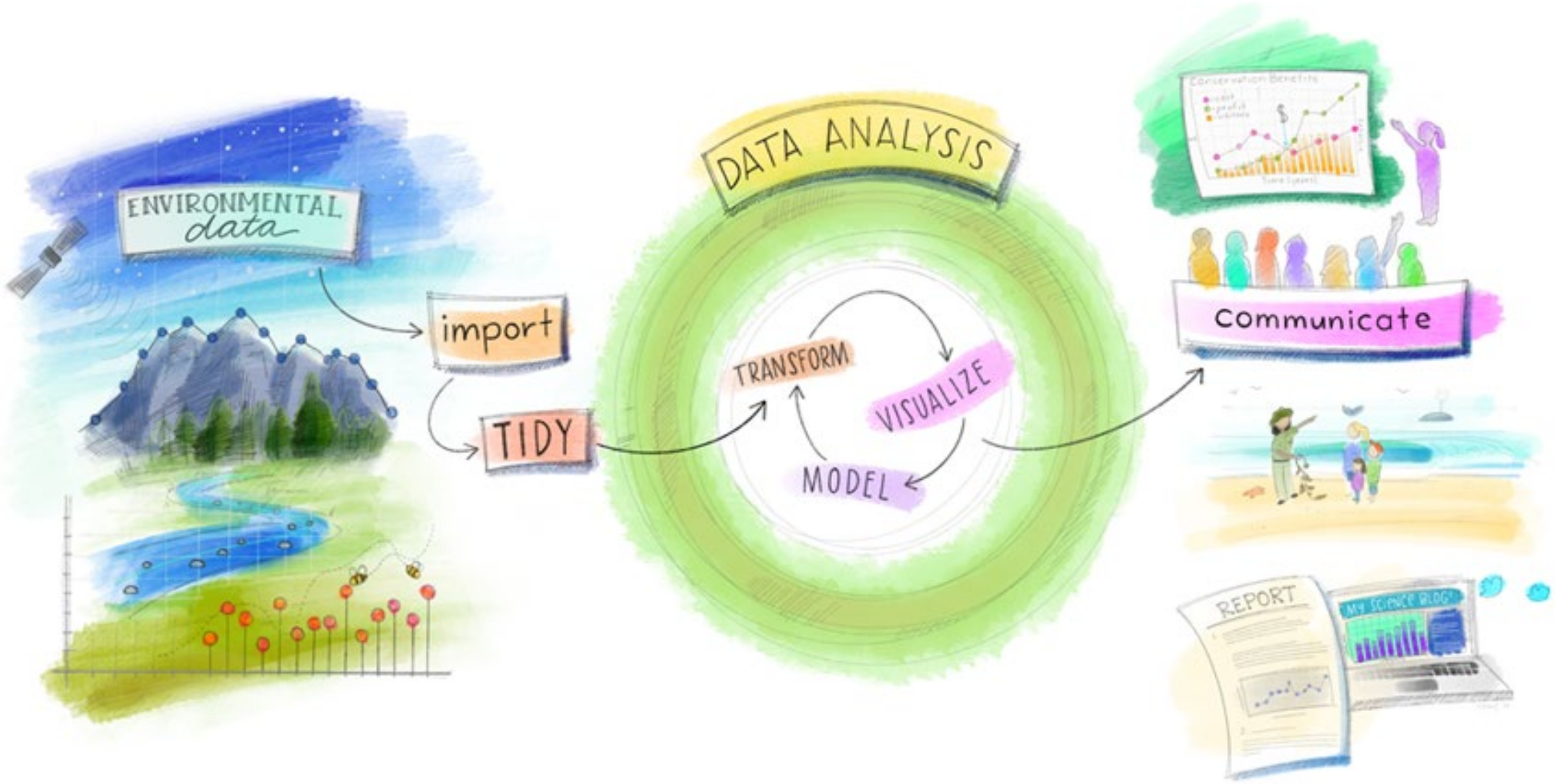
Table 7.1 shows a representation of the same data as a table.

Table 7.1: Data from the students.csv file as a table.

Student ID	Full Name	favourite.food	mealPlan	AGE
1	Sunil Huffmann	Strawberry yoghurt	Lunch only	4
2	Barclay Lynn	French fries	Lunch only	5
3	Jayendra Lyne	N/A	Breakfast and lunch	7
4	Leon Rossini	Anchovies	Lunch only	NA
5	Chidiegwu Dunkel	Pizza	Breakfast and lunch	five
6	Güvenç Attila	Ice cream	Lunch only	6


country
year
cases
population
variables
observations
values

Processes


ENVIRONMENTAL data
import
TIDY
DATA ANALYSIS
TRANSFORM
VISUALIZE
MODEL
Communicate
REPORT

How to export or save as results or data

```
write.csv(df , file = “path/to/your/saving/folder/df.csv")
```

```
writexl::write_xlsx (test_df, path = "C:/Users/your_username/test_df.xlsx",
col_names = TRUE, format_headers = TRUE)
```

How to export or save as plot

```
dev.print (device=jpeg, file="path/to/your/saving/plots/figure_1a.jpg",
width=par("din")[1]*300, res=300, quality=100)
```

```
ggsave("path/to/your/saving/plots/figure_1a.png", plot = p1, bg ="white")
```

!! Practical session (afternoon session): read a data file into Rstudio!!

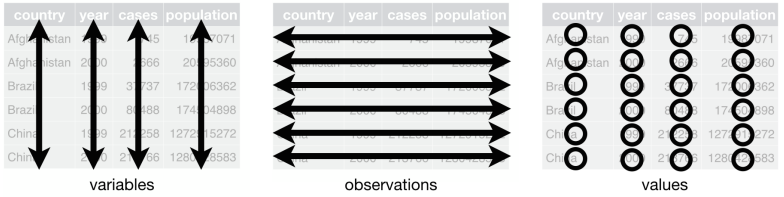
Seeking help

- RStudio help interface
  - ?sum
  - help.search(), with term in “” inside parentheses
- Google or chatgpt “R <task>”
- When asking others
  - Use correct words
  - Reduce to reproducible example
  - Always include output of sessionInfo() function

Good housekeeping strategies for scripts

- Comments, Structure:
  - Use #
  - Outline
- Consistent Naming Conventions:
  - Use consistent naming conventions for variables, functions, and objects
  - Avoid duplicating code. create a function or use a loop
- Version Control:
  - Git. Platforms like GitHub or GitLab repositories.
- Imports and Dependencies:
  - List all package imports at the top of your script
  - call rhistory: sessionInfo()
- File Organization:
  - Separate your R scripts, data, documentation, and output files into logical folders.
- Reproducibility:
  - Use Quarto, R Markdown Documents:
  - Documentation Files:
  - Create README files

Filtering of genotype data



- Subset data, metadata
  - Based on “rows (observations)” , “variable” or “values”
  - Filter
- Missing data (“NA”, stat, summary)

```
height_sub_12 <- subset(data, height<12)

data_filter_na <- data[!is.na(data$height),]
```

- `mutate()` adds new variables that are functions of existing variables
- `select()` picks variables based on their names.
- `filter()` picks cases based on their values.
- `summarise()` reduces multiple values down to a single summary.
- `arrange()` changes the ordering of the rows.

Questions?!

Acknowledgements

- NBIS(RaukR)
- Shannon Alexandra Sheridan