International Statistical Genetics Workshop

IBG Boulder, Colorado

2024 Edition

Intro to Unix / R / computing

8AM - Monday 4th March

José J. Morosoli

Elizabeth Prom-Wormley

with thanks to Sarah Medland

Please answer the survey before we start!

https://forms.office.com/e/ibLW7WNNec

Welcome!

https://forms.office.com/e/ibLW7WNNec

Getting the most out of the workshop:

- ► Ask questions!
- ► Try to sit next to someone you don’t already know.
- ► Work with someone with different skillset and experience level.
- ► You will have access to your files after you leave.
- ► Come to the social functions.
- ► Do ask questions!!
  - ✓ In person, email, or on the forum.

https://isgw-forum.colorado.edu/

https://isgw-forum.colorado.edu

A diverse community!

https://forms.office.com/e/ibLW7WNNec

Diverse skillsets, backgrounds, research focus, experience levels… and timezones!

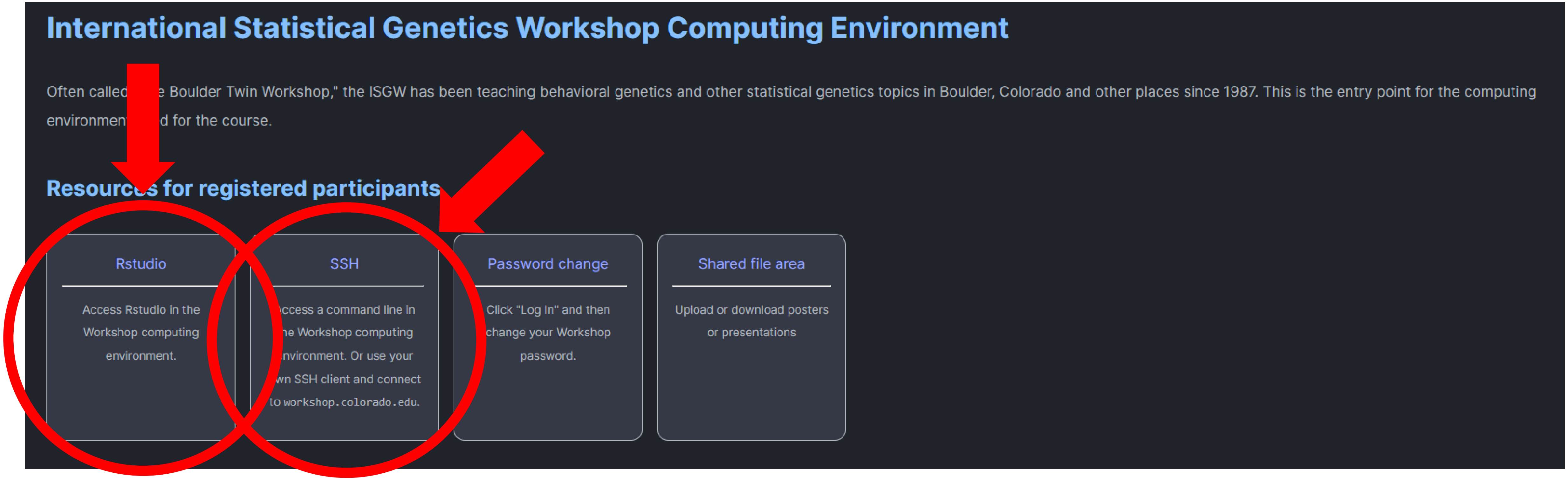
José J. Morosoli

Be open, be kind, be respectful. We are all learners.

Survey results!

Getting started

- https://workshop.colorado.edu/



International Statistical Genetics Workshop Computing Environment

Often calle e Boulder Twin Workshop," the ISGW has been teaching behavioral genetics and other statistical genetics topics in Boulder, Colorado and other places since 1987. This is the entry point for the computing environmen d for the course.

Resourc s for registered participants

Rstudio

Access Rstudio in the Workshop computing environment.

SSH

ccess a command line in ne Workshop computing nvironment. Or use your wn SSH client and connect to `workshop.colorado.edu`.

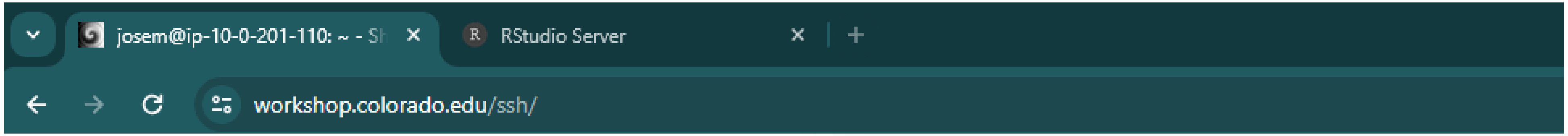
Password change

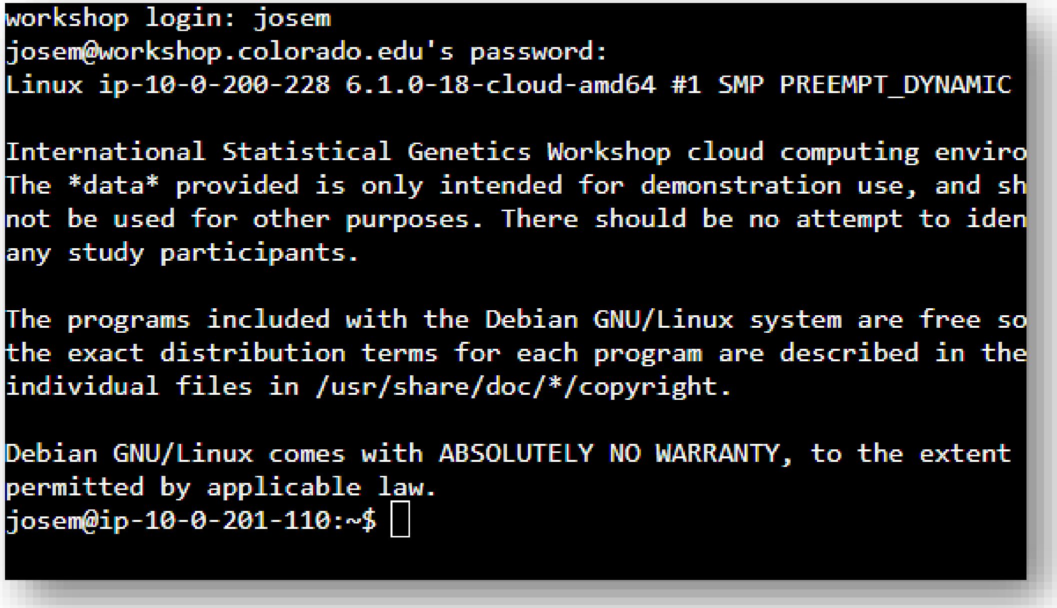
Click "Log In" and then change your Workshop password.

Shared file area

Upload or download posters or presentations

Getting started





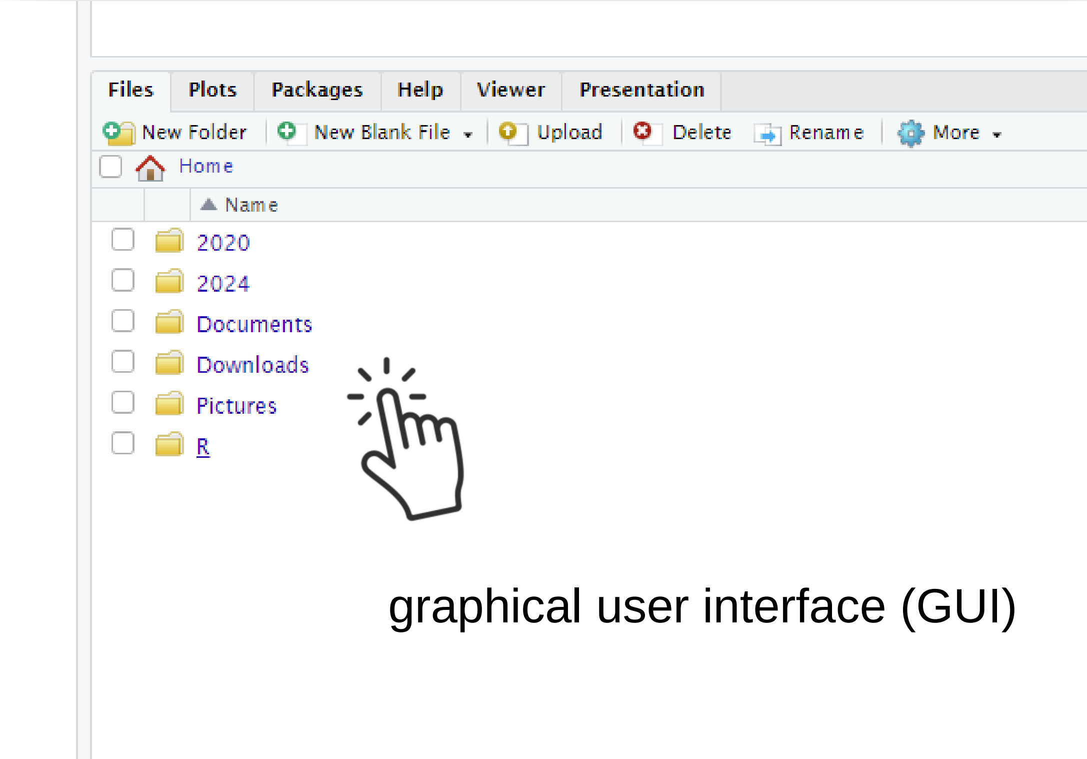
```
workshop login: josem
josem@workshop.colorado.edu's password:
Linux ip-10-0-200-228 6.1.0-18-cloud-amd64 #1 SMP PREEMPT_DYNAMIC

International Statistical Genetics Workshop cloud computing enviro
The *data* provided is only intended for demonstration use, and sh
not be used for other purposes. There should be no attempt to iden
any study participants.

The programs included with the Debian GNU/Linux system are free so
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
josem@ip-10-0-201-110:~$ 
```

command-line interface (CLI)



graphical user interface (GUI)

Why Unix? Why?

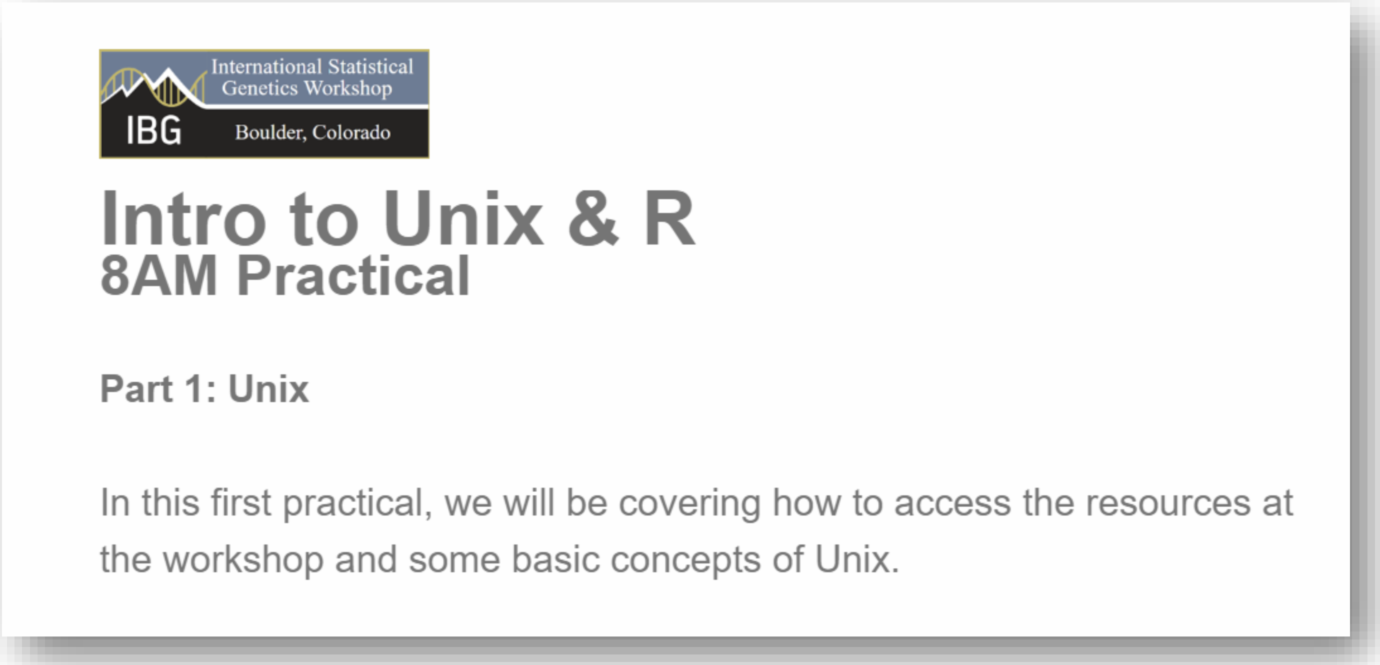
- ✓ It allows you to automate tasks.
- ✓ Replicability: one script, multiple re-runs.
- ✓ It is more efficient (i.e., fast), scalable and stable than other systems, not to mention open source (e.g., Windows, MacOS).
- ✓ Big data (e.g., genetic data) is usually stored and analyzed in high-performance computing (HPCs) environments, which for multiple reasons (see above) are based on this language.

Getting started

- Open the first exercise:

  https://qualtrics.ucl.ac.uk/jfe/form/SV_0pHnuiW6juZ9ezs

- Open the SSH tab and log in with your username.

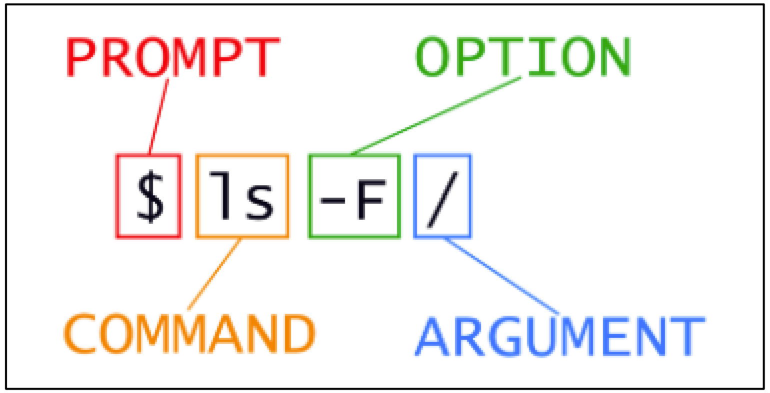


Intro to Unix & R
8AM Practical

Part 1: Unix

In this first practical, we will be covering how to access the resources at the workshop and some basic concepts of Unix.

Intro to Unix: Glossary

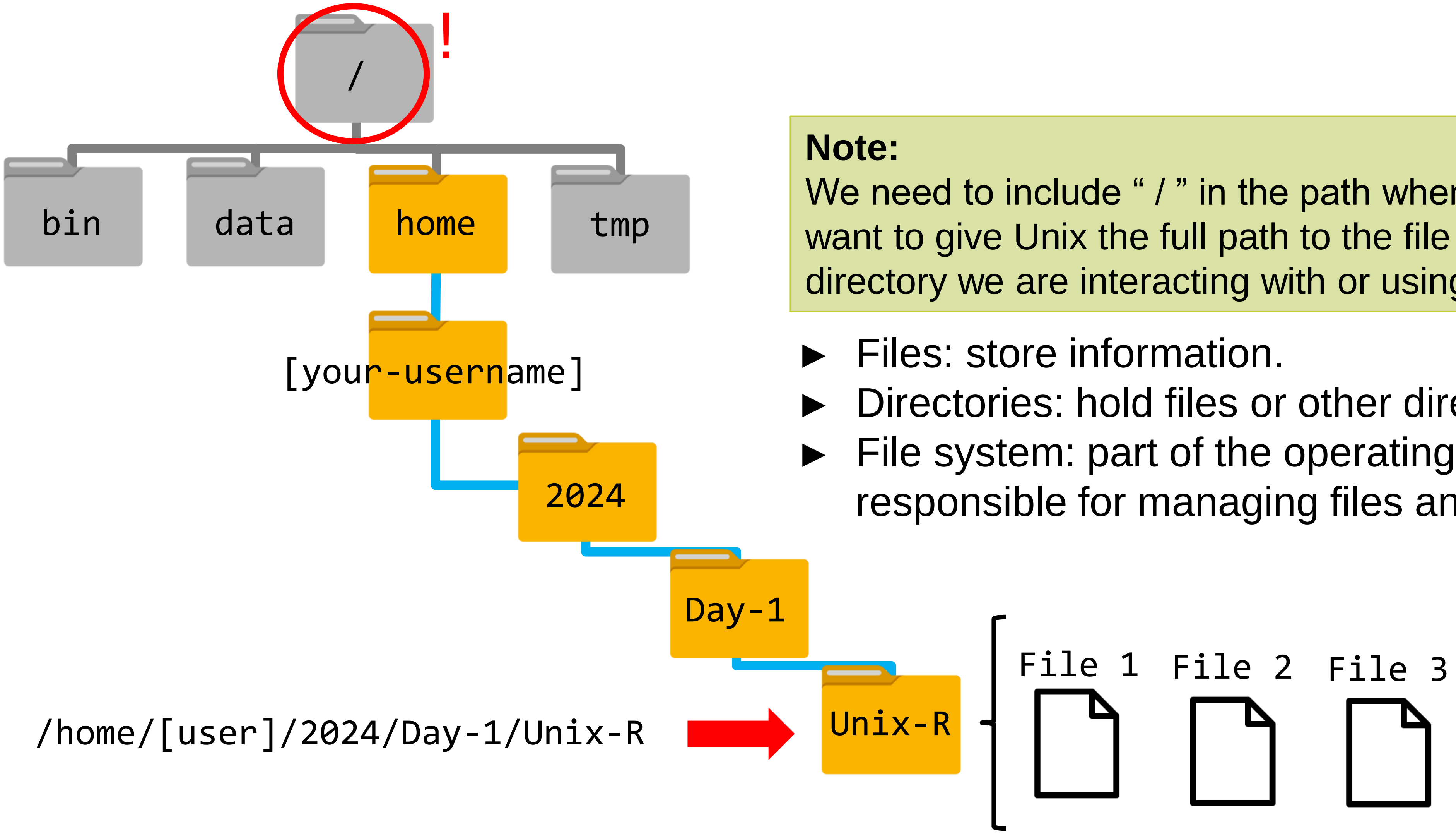

PROMPT
OPTION

```
$ ls -F /
```

COMMAND
ARGUMENT

- **Graphical User Interface (GUI):** platform to interact with a computer that involves point and clicking and using menus.
- **Command Line Interface (CLI):** text-based platform where you can input text commands to interact with a computer.
- **Unix**: an operative system (just like Windows or MacOS).
- **Shell**: program where users can type commands.
- **Bash**: most popular shell in Unix.
- **Prompt**: symbol that indicates that the shell is waiting for input.
- **Command**: Pre-defined “words” that tell the system what to do, they can be modified using **options** and sometimes require **arguments** to indicate what files to operate on, the paths to find them, etc.

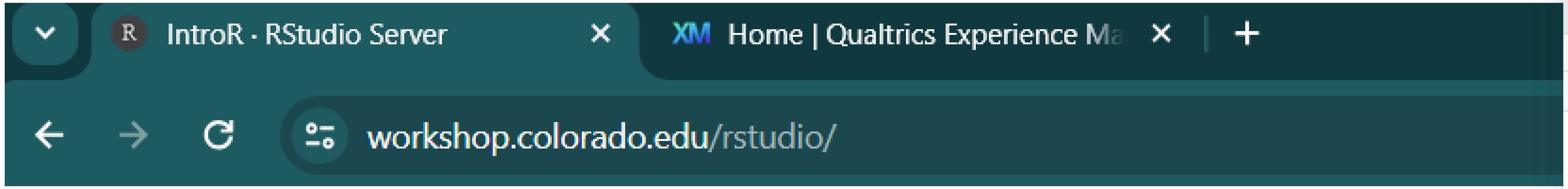
Setting up our home directory



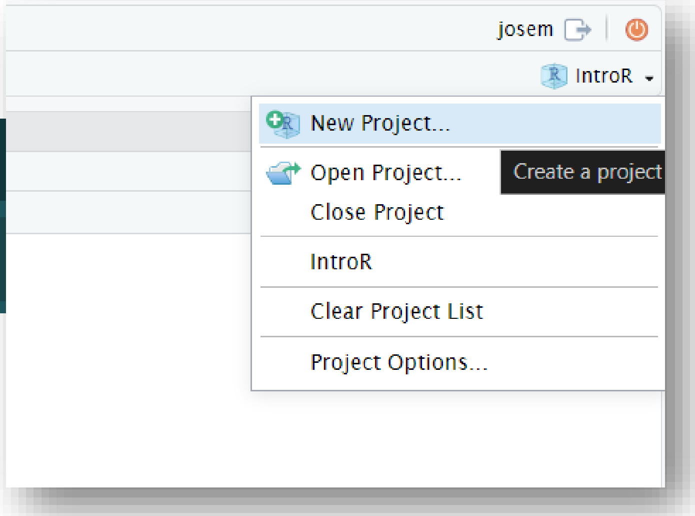
Note:
We need to include “ / ” in the path whenever we want to give Unix the full path to the file or directory we are interacting with or using.

- Files: store information.
- Directories: hold files or other directories.
- File system: part of the operating system responsible for managing files and directories

Next stop: R



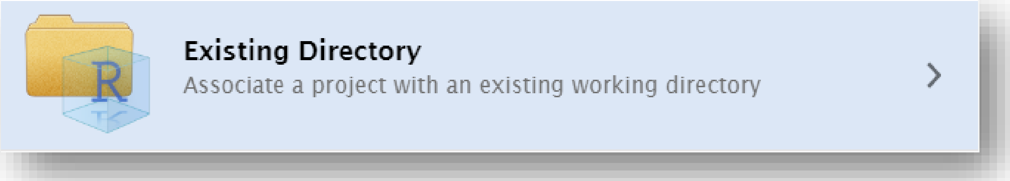
workshop.colorado.edu/rstudio/



- Go to the next Qualtrics link:

  https://qualtrics.ucl.ac.uk/jfe/form/SV_6FGfSGwaABJQo1E

- Create a R project:
  - Creates/finds a working directory for you.
  - Remembers its location.
  - Makes it easier to resume work after a break.



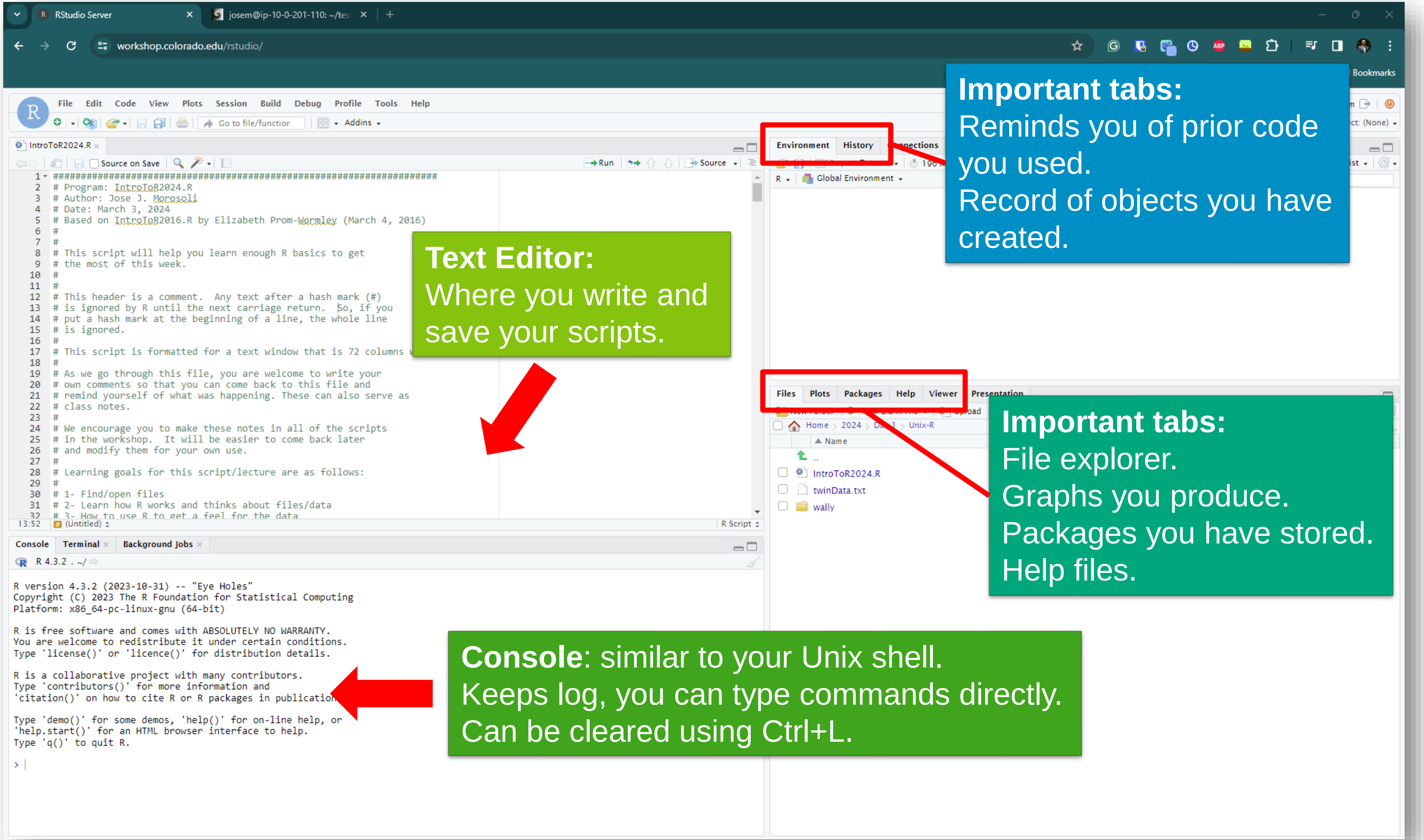
- Click on IntroToR2024.R to open it.

Directory: ~/2024/Day-1/Unix-R

Similarities between Unix and R

- ✓Unix and R offer command-line interfaces.
- ✓Both support scripting for automation.
- ✓Active communities and extensive package.
- ✓Open-source.
- ✓Customizable.
- ✓Similar commands but not the same!
  - E.g., `pwd` vs `getwd()`

DeepAI (b. 2016)
Unix r programming languages are friends, 2024
Digital work


RStudio Server
workshop.colorado.edu/rstudio/
File Edit Code View Plots Session Build Debug Profile Tools Help
IntroToR2024.R
Source on Save
Run
Source
Environment
History
Global Environment
Files Plots Packages Help Viewer Presentation
Home 2024 Unix-R
Name
IntroToR2024.R
twinData.txt
wally
Console Terminal Background Jobs
R 4.3.2 . ~/
R Script

```
########################################################################
# Program: IntroToR2024.R
# Author: Jose J. Morosoli
# Date: March 3, 2024
# Based on IntroToR2016.R by Elizabeth Prom-Wormley (March 4, 2016)
#
#
# This script will help you learn enough R basics to get
# the most of this week.
#
#
# This header is a comment.  Any text after a hash mark (#)
# is ignored by R until the next carriage return.  So, if you
# put a hash mark at the beginning of a line, the whole line
# is ignored.
#
# This script is formatted for a text window that is 72 columns w
#
# As we go through this file, you are welcome to write your
# own comments so that you can come back to this file and
# remind yourself of what was happening. These can also serve as
# class notes.
#
# We encourage you to make these notes in all of the scripts
# in the workshop.  It will be easier to come back later
# and modify them for your own use.
#
# Learning goals for this script/lecture are as follows:
#
# 1- Find/open files
# 2- Learn how R works and thinks about files/data
# 3- How to use R to get a feel for the data
```

```
R version 4.3.2 (2023-10-31) -- "Eye Holes"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publication

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Important tabs:
Reminds you of prior code you used.
Record of objects you have created.

Text Editor:
Where you write and save your scripts.

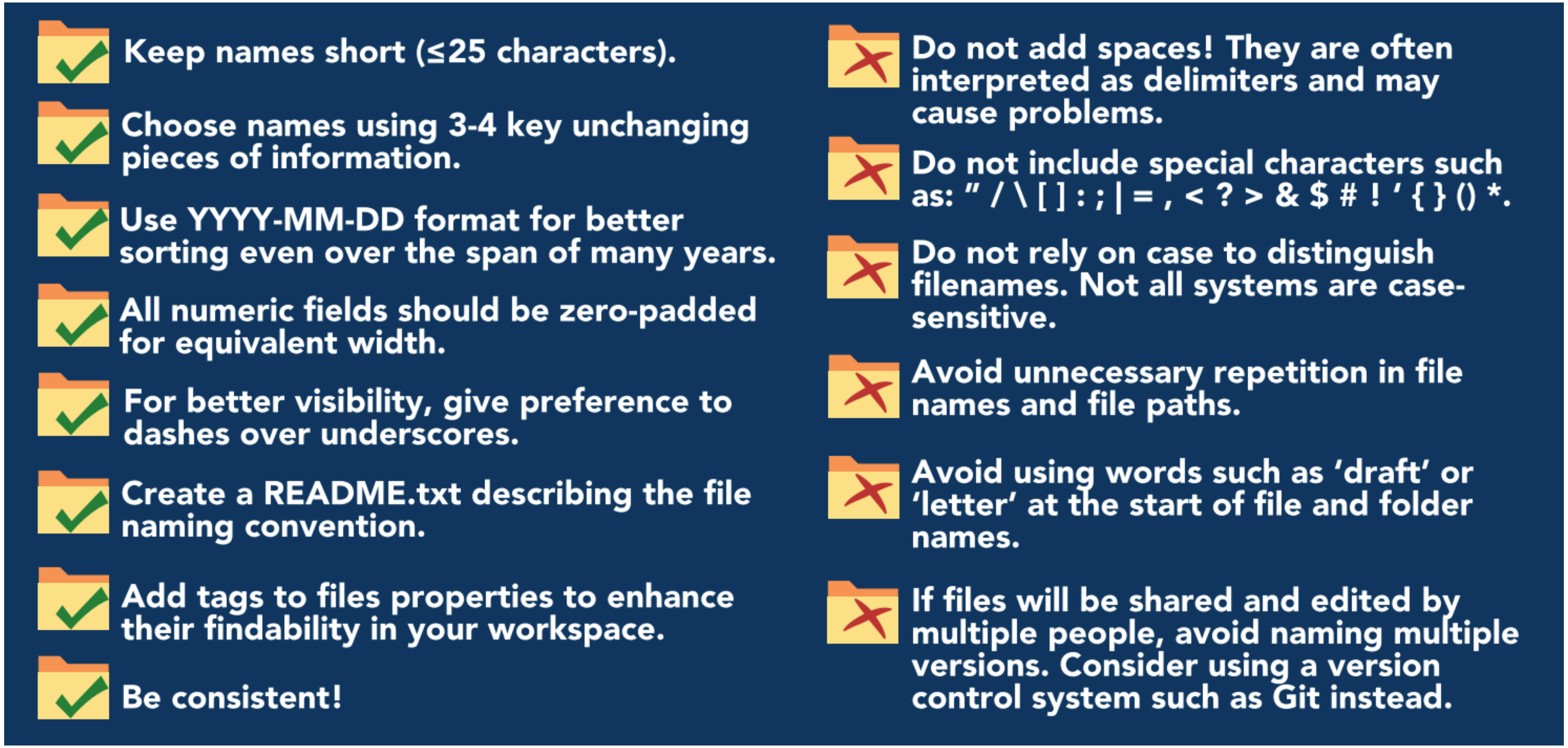
Important tabs:
File explorer.
Graphs you produce.
Packages you have stored.
Help files.

Console: similar to your Unix shell.
Keeps log, you can type commands directly.
Can be cleared using Ctrl+L.

Resources

- Software carpentry: https://software-carpentry.org/lessons/index.html
- https://stackoverflow.com/ and https://unix.stackexchange.com/
- UNIX cheatsheet in `/home/josem/2024/Day-1/Unix-R`.
- Specific resources for R:
  1. R for SAS and SPSS Users: https://science.nature.nps.gov/im/datamgmt/statistics/R/documents/R_for_SAS_SPSS_users.pdf
  2. An R Style Guide: http://google-styleguide.googlecode.com/svn/trunk/Rguide.xml
  3. Rseek: A search engine for all things R related (http://www.rseek.org)
  4. R-Bloggers.
  5. Quick-R's section on graphics: http://www.statmethods.net/advgraphs/parameters.html
  6. More information on data frames: http://www.r-bloggers.com/exploratory-data-analysis-useful-r-functions-for-exploring-a-data-frame.
  7. Details on how to develop your own package: http://r-pkgs.had.co.nz

R Code – Best practices



- ✓ Keep names short (≤25 characters).
- ✓ Choose names using 3-4 key unchanging pieces of information.
- ✓ Use YYYY-MM-DD format for better sorting even over the span of many years.
- ✓ All numeric fields should be zero-padded for equivalent width.
- ✓ For better visibility, give preference to dashes over underscores.
- ✓ Create a README.txt describing the file naming convention.
- ✓ Add tags to files properties to enhance their findability in your workspace.
- ✓ Be consistent!

- ✗ Do not add spaces! They are often interpreted as delimiters and may cause problems.
- ✗ Do not include special characters such as: ” / \ [ ] : ; | = , < ? > & $ # ! ‘ { } () *.
- ✗ Do not rely on case to distinguish filenames. Not all systems are case-sensitive.
- ✗ Avoid unnecessary repetition in file names and file paths.
- ✗ Avoid using words such as ‘draft’ or ‘letter’ at the start of file and folder names.
- ✗ If files will be shared and edited by multiple people, avoid naming multiple versions. Consider using a version control system such as Git instead.

https://www.r-bloggers.com/2018/09/r-code-best-practices/

https://www.library.ucsb.edu/sites/default/files/dls-n01-2021-filenaming.pdf