Introduction to R Programming and Collaborative Science

Pranav Pandit

BVSc & AH, MPVM, PhD

Department of Population Health and Reproduction,
School of Veterinary Medicine, UC Davis

Slides Courtesy: Dr. Nistara Randhawa

Learning Objectives

- Introduction to R Programming
  - Basics of R syntax and data types
- Project management
  - Setting up Jupyter and Anaconda Environments
- Collaborative and Reproducible Science
  - Best Practices for collaborative research and reproducibility

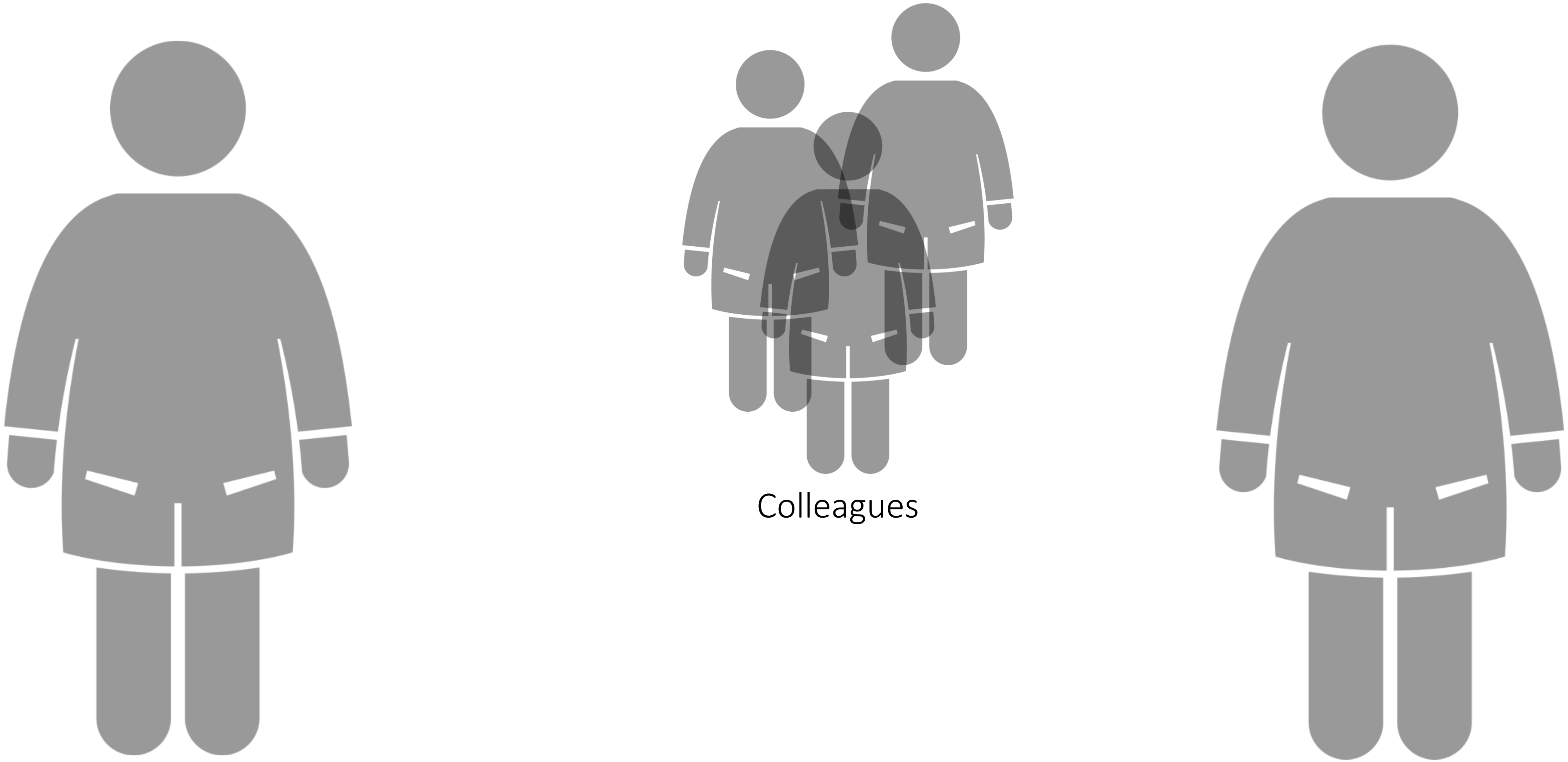
Data is valuable

- Great effort is put in to collect data systematically
- Hard work, meticulous planning, and recourses are put in to collect the data

Data management

- For safe storage and sharing
- Generating reproducible research

Collaborate effectively



You

Colleagues

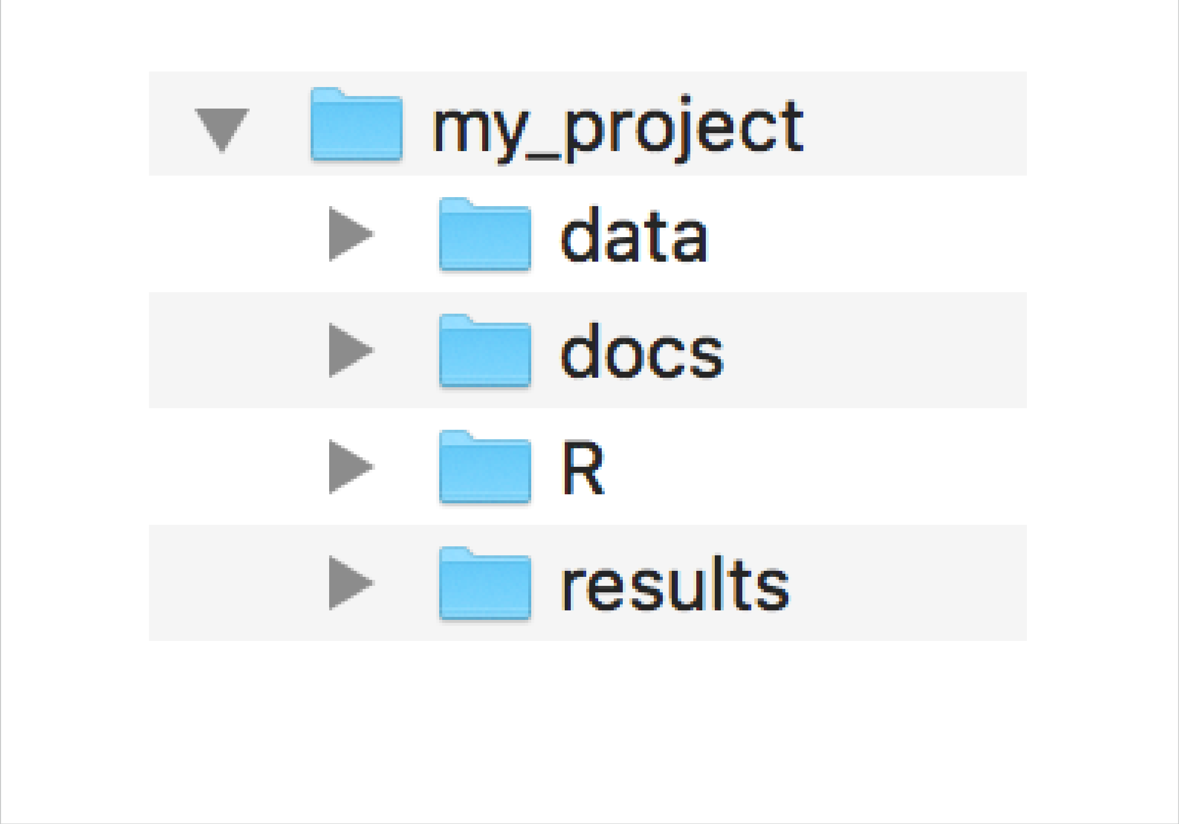
Future you

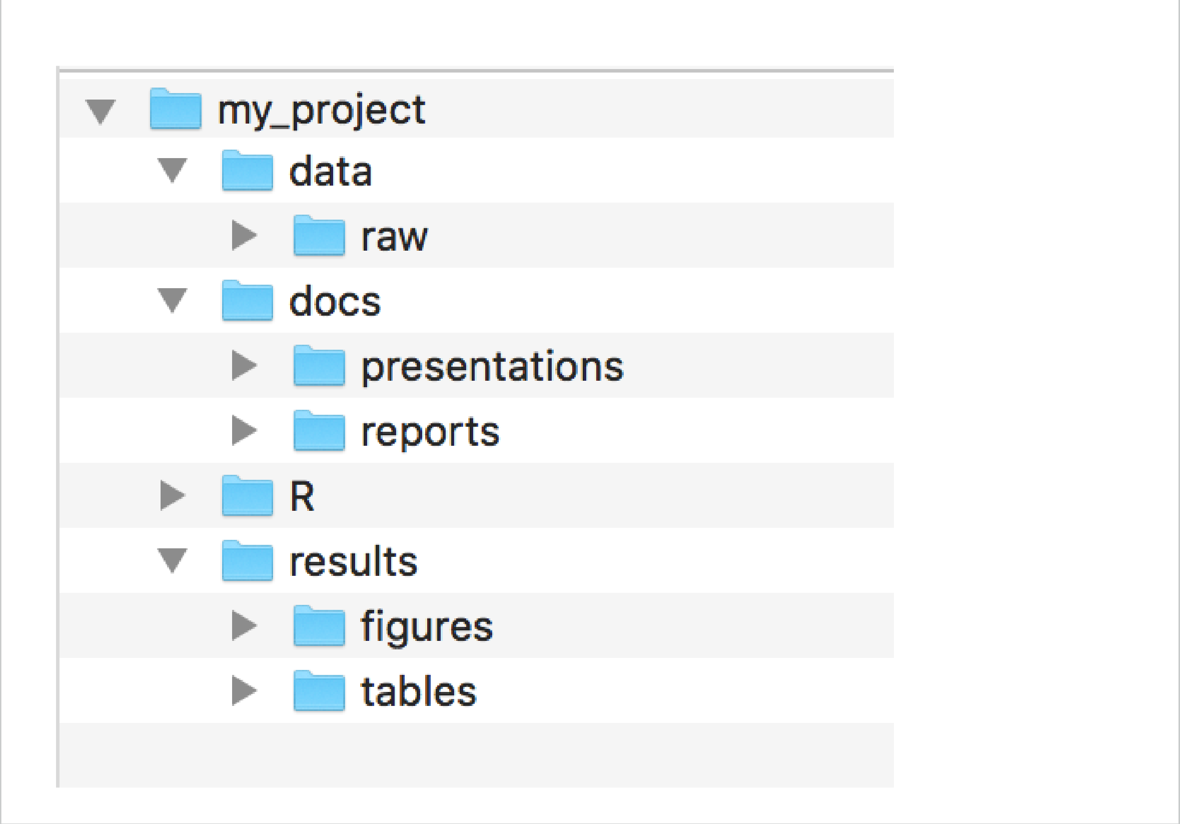
Data management

- Organize your project folder
- Protect your raw data
- Name your files effectively
- Track your project’s changes
- Backup your project

Project folder

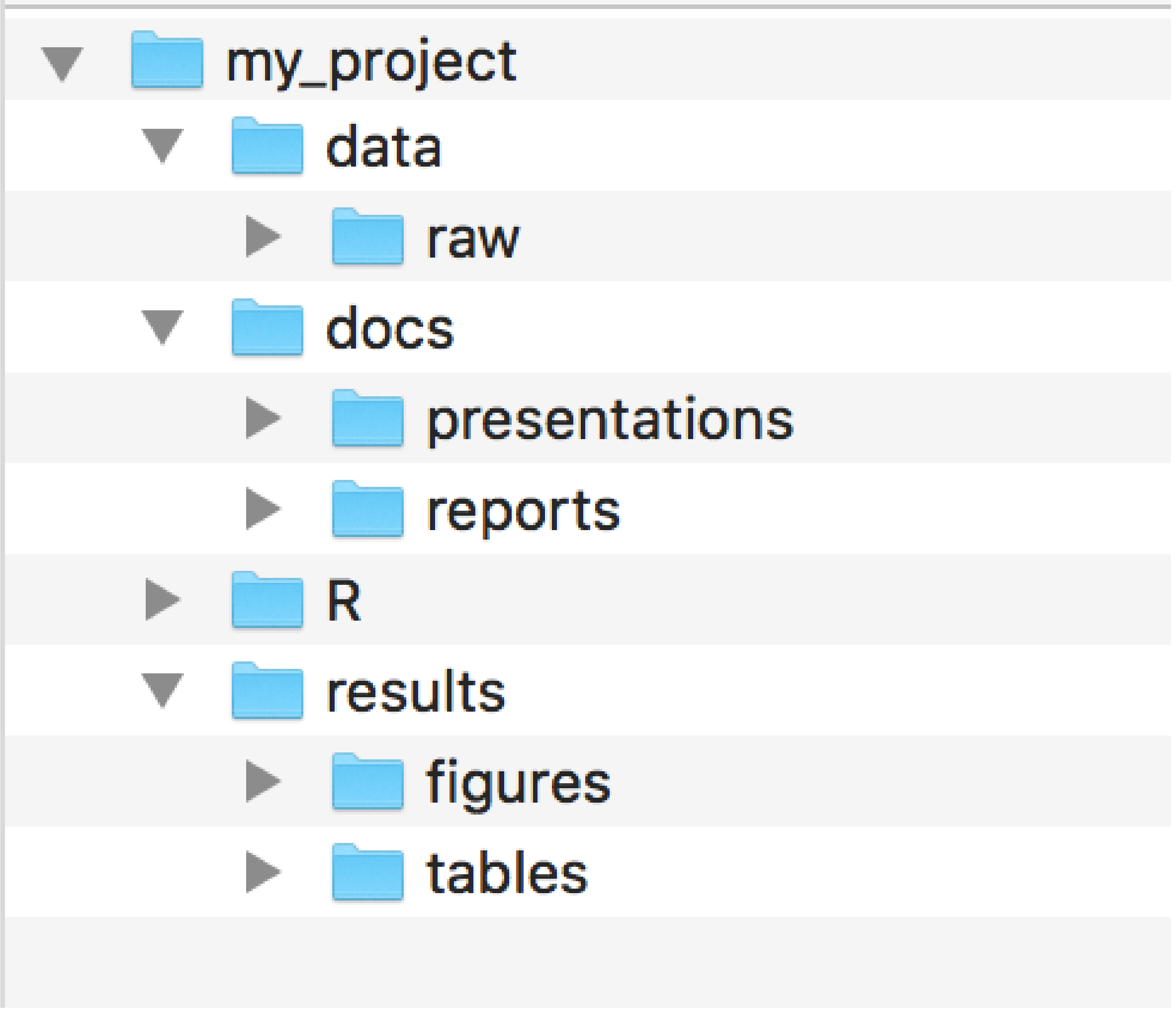
- AnimalsSampled_Export_May16_0438.csv
  Site_Exoort_May18_2353.csv





Protect your raw data

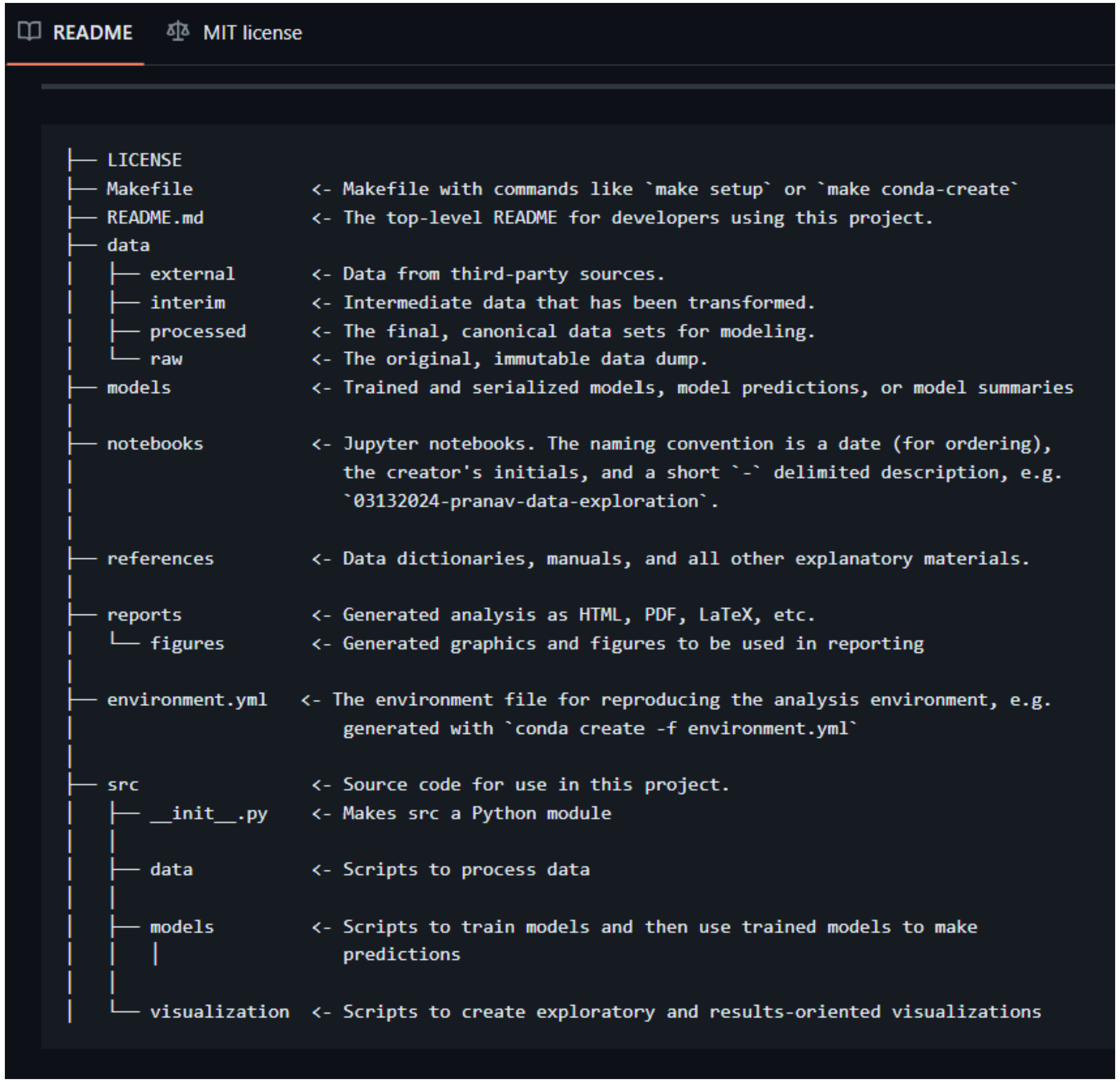
- Do not edit raw data directly
- Copy and worn with it so the original data is not modified



```
my_project
    data
        raw
    docs
        presentations
        reports
    R
    results
        figures
        tables
```

Version Control you data and code

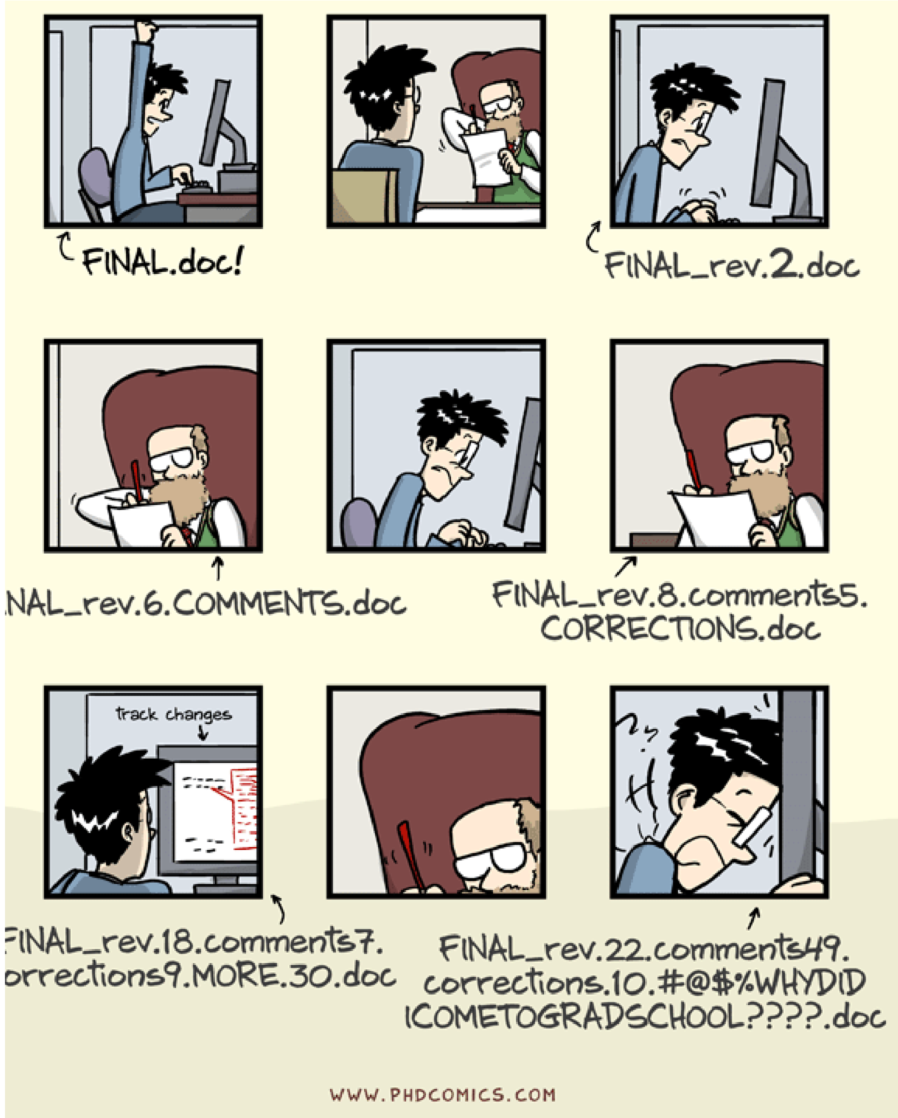
- Git: https://github.com/
- Drive/Box/Dropbox
- Local external hard-disk



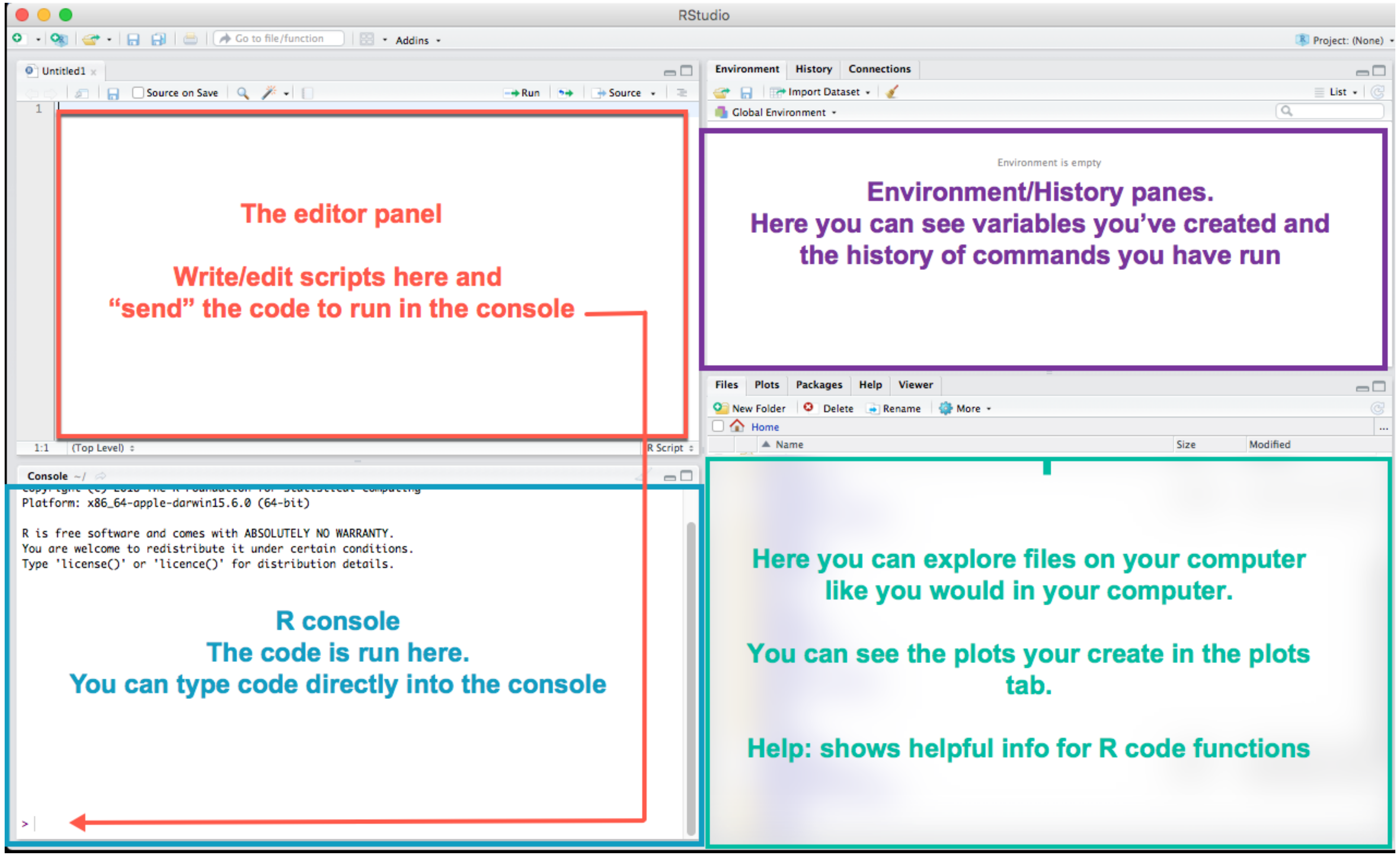
```
├── LICENSE
├── Makefile           <- Makefile with commands like `make setup` or `make conda-create`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third-party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. The naming convention is a date (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `03132024-pranav-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── environment.yml   <- The environment file for reproducing the analysis environment, e.g.
│                         generated with `conda create -f environment.yml`
│
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
│   │
│   ├── data           <- Scripts to process data
│   │
│   ├── models         <- Scripts to train models and then use trained models to make
│   │   │                 predictions
│   │
│   └── visualization  <- Scripts to create exploratory and results-oriented visualizations
```

Naming files

- Descriptive
- Consistent
- Human readable
- Machine readable



RStudio



The editor panel

Write/edit scripts here and "send" the code to run in the console

Environment/History panes.
Here you can see variables you've created and the history of commands you have run

```
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

>
```

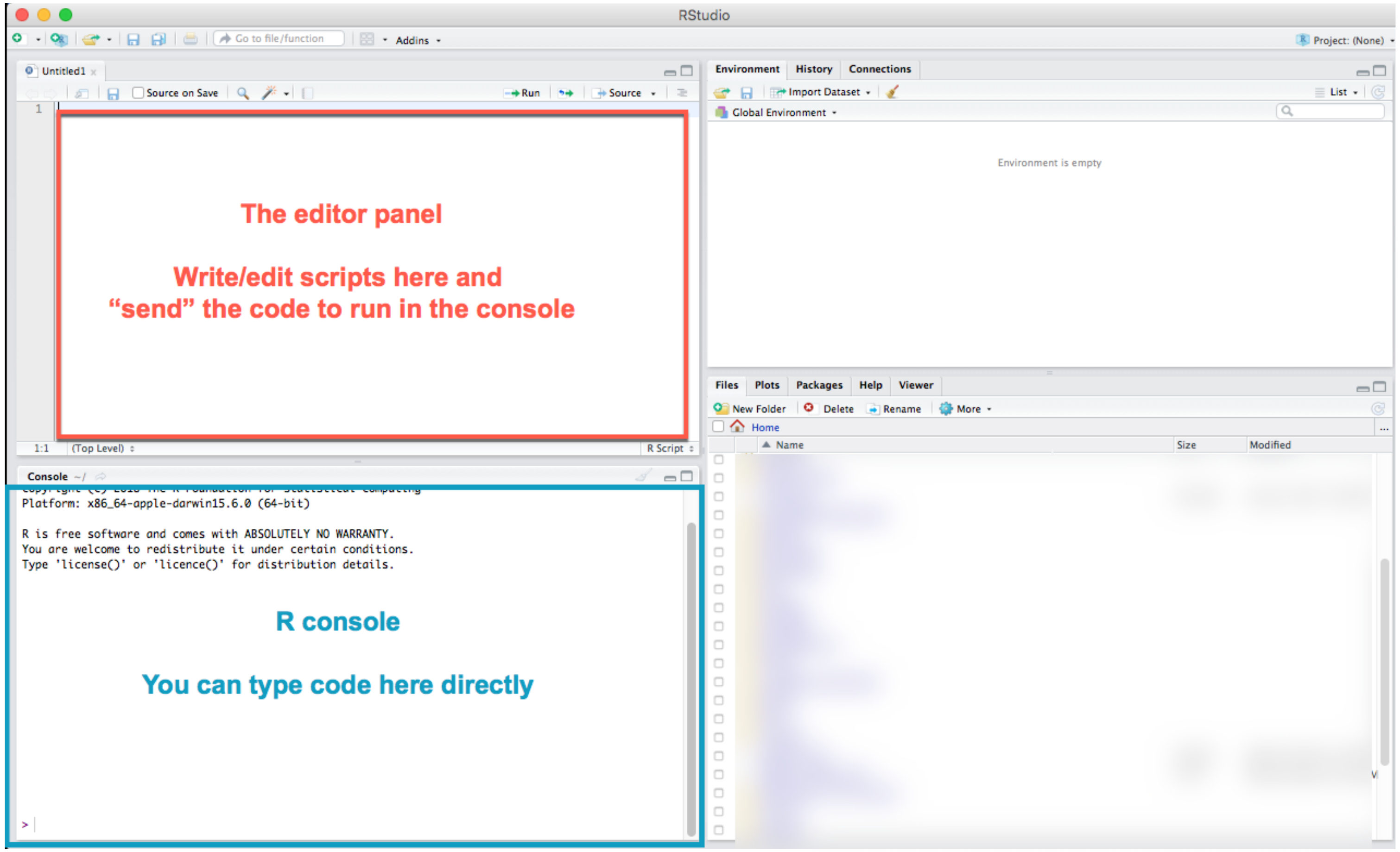
R console
The code is run here.
You can type code directly into the console

Here you can explore files on your computer like you would in your computer.

You can see the plots your create in the plots tab.

Help: shows helpful info for R code functions

RStudio



The editor panel

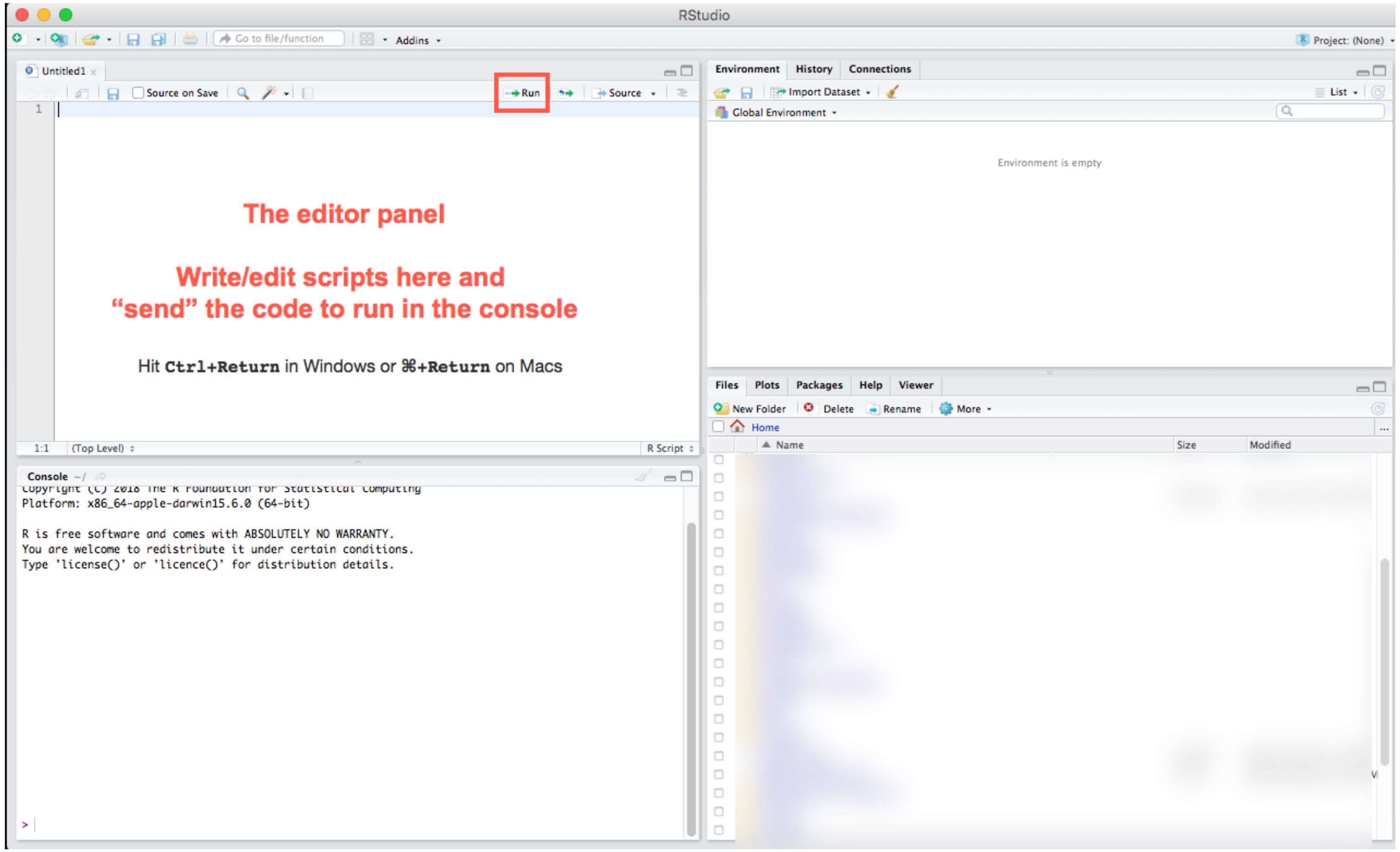
Write/edit scripts here and “send” the code to run in the console

```
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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 console

You can type code here directly

RStudio



The editor panel

Write/edit scripts here and “send” the code to run in the console

Hit **`Ctrl+Return`** in Windows or **`⌘+Return`** on Macs

```
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

>
```

Important data management and wrangling concepts in R

- Refer to the code demonstration

“*Introduction to R.html*”