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

Learning Objectives

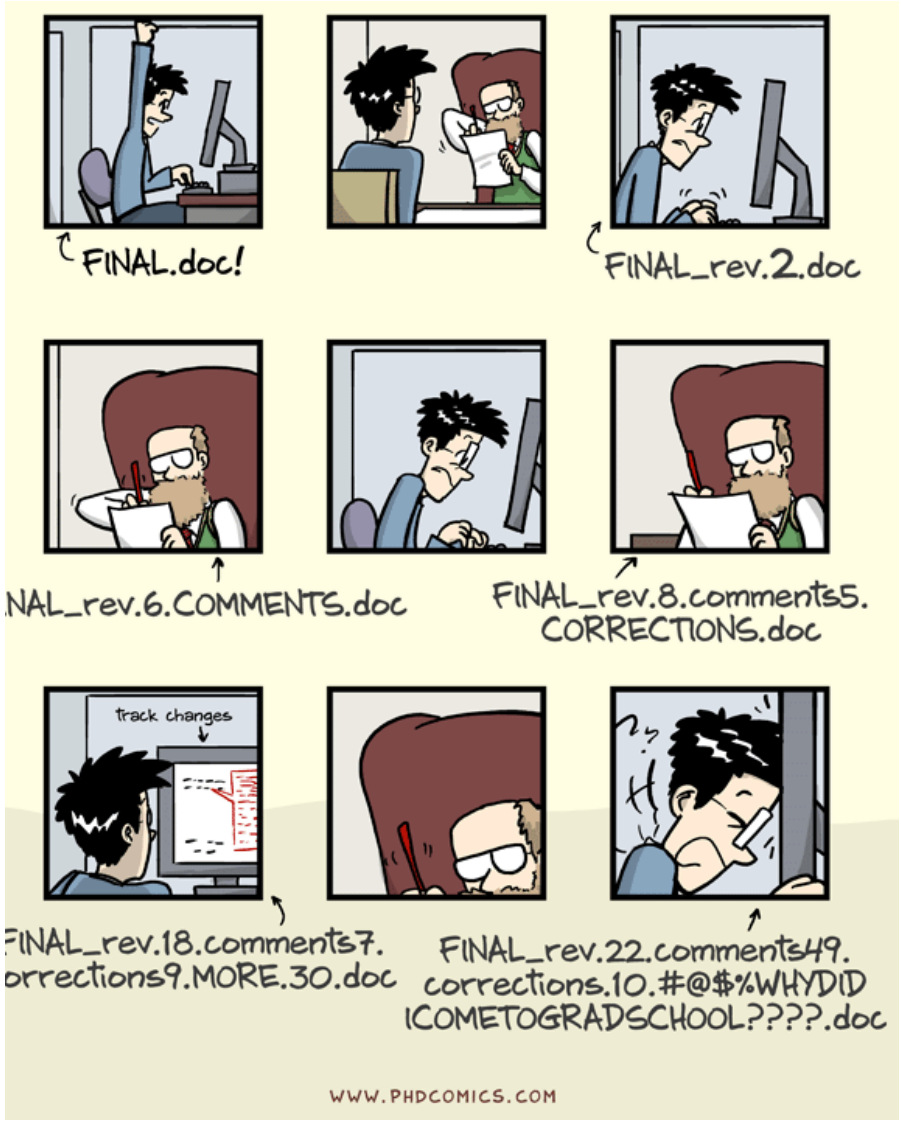
- Introduction to R Programming
  - Basics of R syntax and data types
- Project management
  - Setting up Google Colab and coding environments
- Collaborative and Reproducible Science
  - Best Practices for collaborative research and reproducibility

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



BULLETIN
ECOLOGICAL SOCIETY OF AMERICA

Commentaries | Open Access

A Beginner's Guide to Conducting Reproducible Research

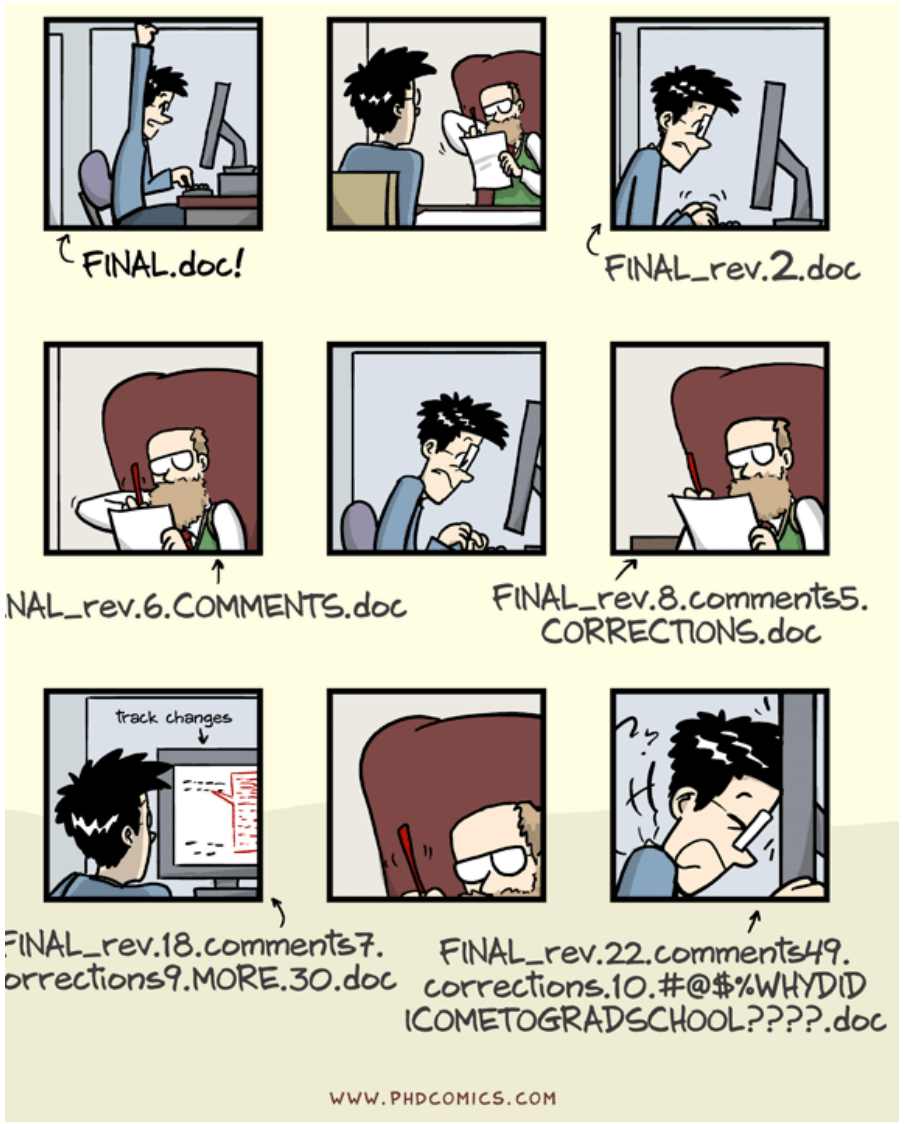
Jesse M. Alston, Jessica A. Rick

First published: 15 January 2021 | https://doi.org/10.1002/bes2.1801 | Citations: 46

SECTIONS PDF TOOLS SHARE

Data management

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



Anatomy of a Working Directory

- Every R project should start with an .Rproj file (rstudio).
- Keep raw data immutable — never overwrite it.
- Modularize code by task.
- Store outputs (plots, summaries) separately.

README MIT license

```
├── 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 Conventions

- Use snake_case or kebab-case (no spaces).
- Prefix with order if sequential: 01_, 02_, 03_.
- Include date (YYYYMMDD) or version (v1.0).
- Example: 20251020_west_nile_cleaning.R or bird_migration_2024_raw.csv.

README MIT license

```
├── 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 Conventions

- e.g., camelCase or snake_case to name objects
- embedding meaningful information in object names
- using “_mat” as a suffix :: matrices
- “_df” :: denote data frames

- Descriptive
- Consistent
- Human readable
- Machine readable

README MIT license

```
├── 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
```

Data Management and Backup

- Store raw and derived data separately.
- Maintain a data dictionary (data_description.xlsx).
- Keep metadata about data source and processing.
- Backup options: Box, Google Drive, GitHub, or rsync/Backblaze.

README MIT license

```
├── 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
```

Version Control you data and code

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

README MIT license

```
├── 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
```

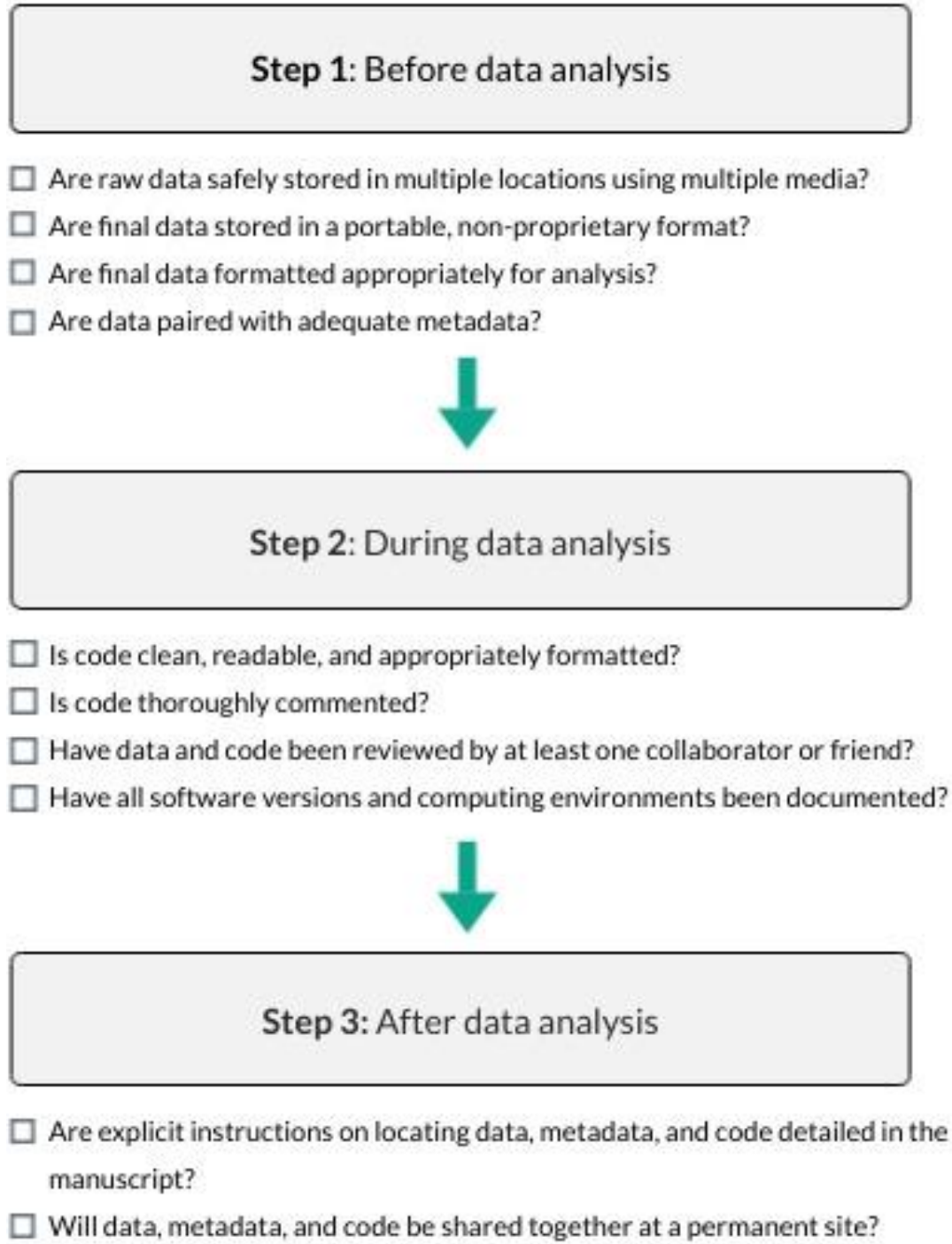
BULLETIN
ECOLOGICAL SOCIETY OF AMERICA

Commentaries | Open Access

A Beginner's Guide to Conducting Reproducible Research

Jesse M. Alston ✉ Jessica A. Rick

First published: 15 January 2021 | https://doi.org/10.1002/bes2.1801 | Citations: 46



Important data management and wrangling concepts in R

- Refer to the code demonstration

"*Introduction to R.html*"