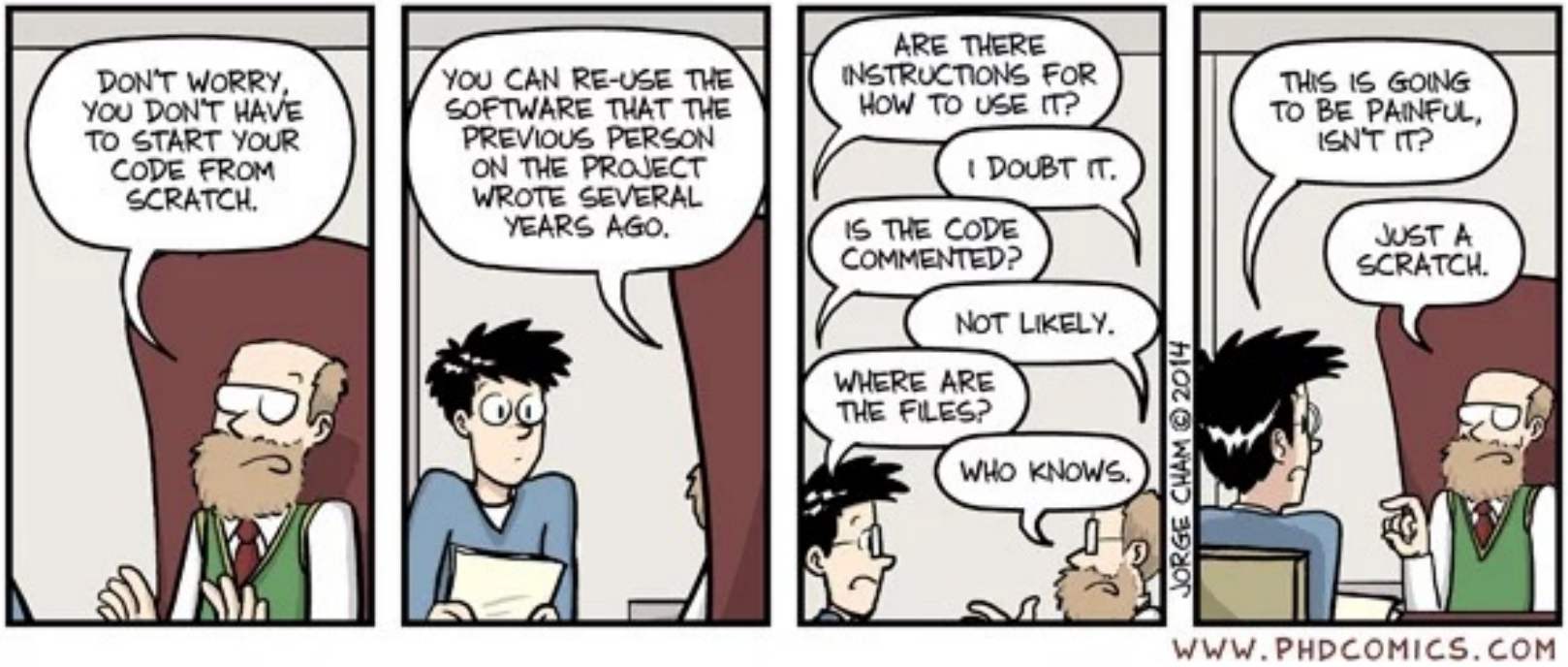
Reproducible Research Workflows

Data Science Initiative Workshop Series

Fall 2021



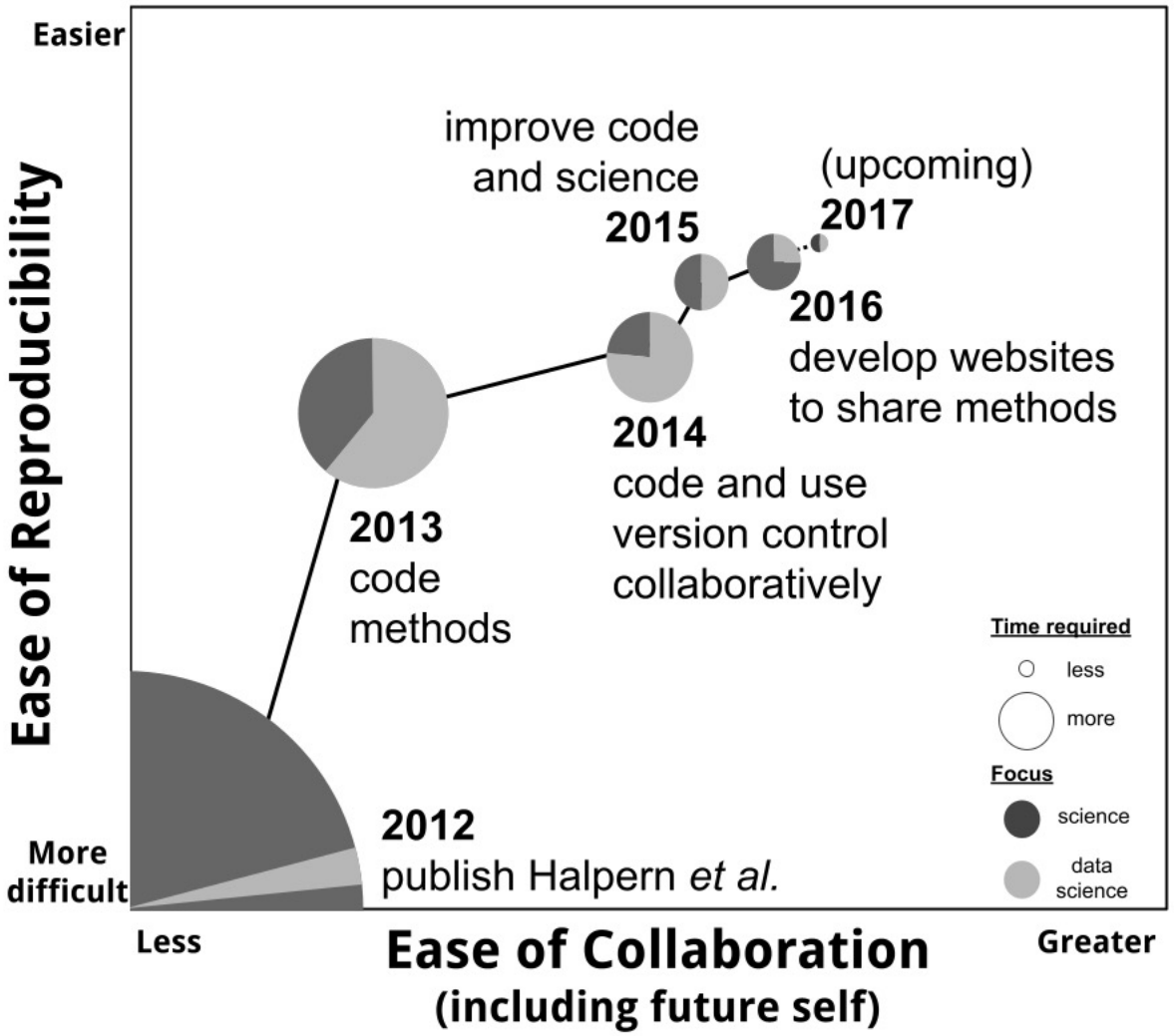
Reproducible research… what is it good for?

FAIR data

- Findable
- Accessible
- Interoperable
- Reusable

https://www.go-fair.org/fair-principles/

Open science



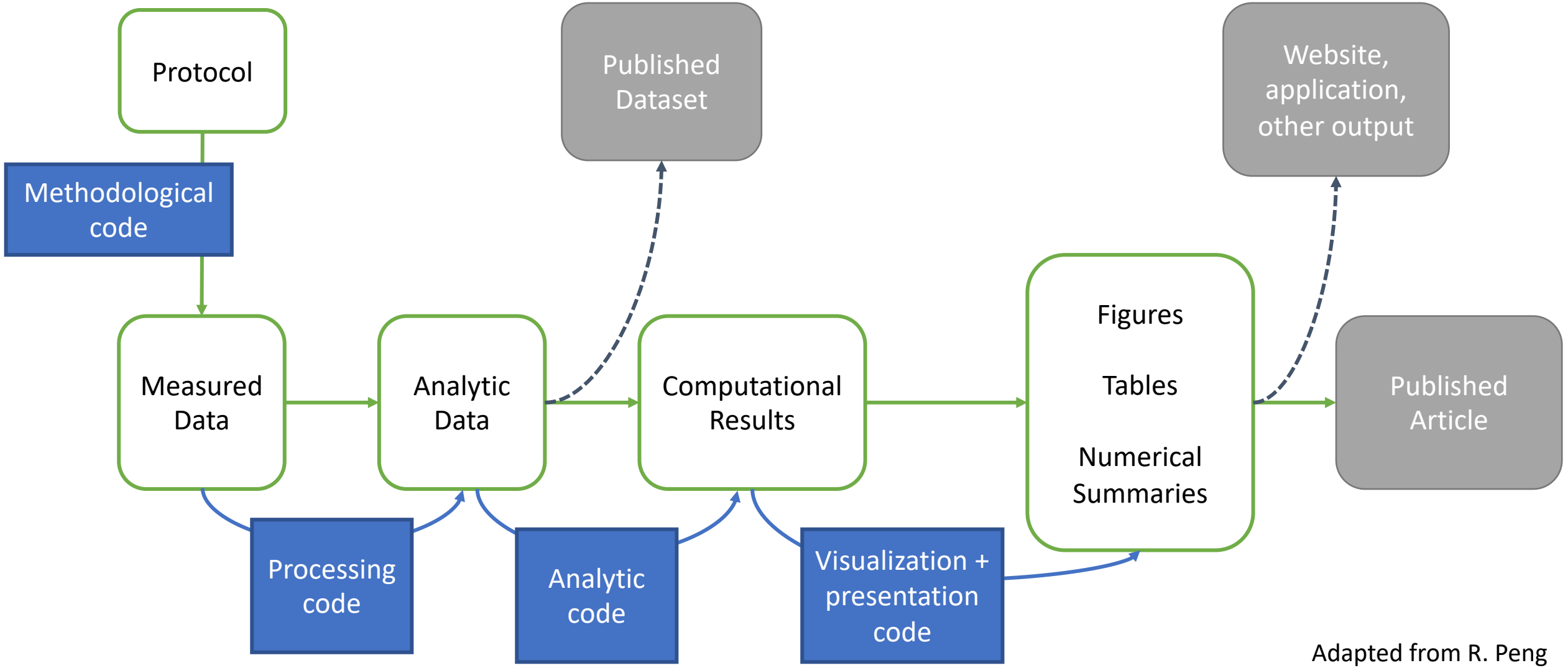
nature ecology & evolution
PERSPECTIVE
PUBLISHED: 23 MAY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0160

Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes[1]*, Benjamin D. Best[2], Courtney Scarborough[1], Jamie C. Afflerbach[1], Melanie R. Frazier[1], Casey C. O'Hara[1], Ning Jiang[1] and Benjamin S. Halpern[1,3,4]

http://ohi-science.org/betterscienceinlesstime/

Research workflow



Adapted from R. Peng

How can we build a reproducible workflow?

Tools

- Version control (e.g. Git)
- Transparent collaboration (e.g. GitHub)
- Documentation
- Data repositories

Practices

- Think about the whole workflow
- Avoid doing things by hand
- Use best practices for coding
- Don’t save output

Collaboration exercise

1. Diagram (part of) your research workflow
2. Identify collaborators who contribute at different steps
3. Pick one step (e.g. moving from raw to processed data)
    1. What access do your collaborators need to data, analysis, or products at this step? How do they contribute?
    2. How do you maintain reproducibility with these collaborators at this step?

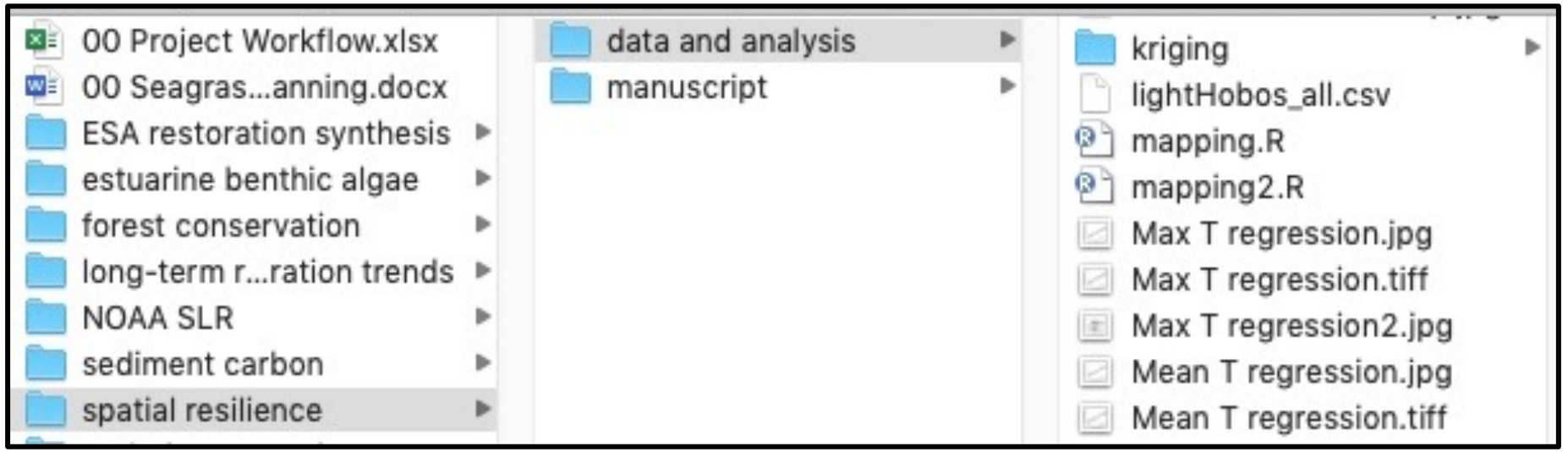
Don’t forget to include your future self as a collaborator!

Project-oriented workflows

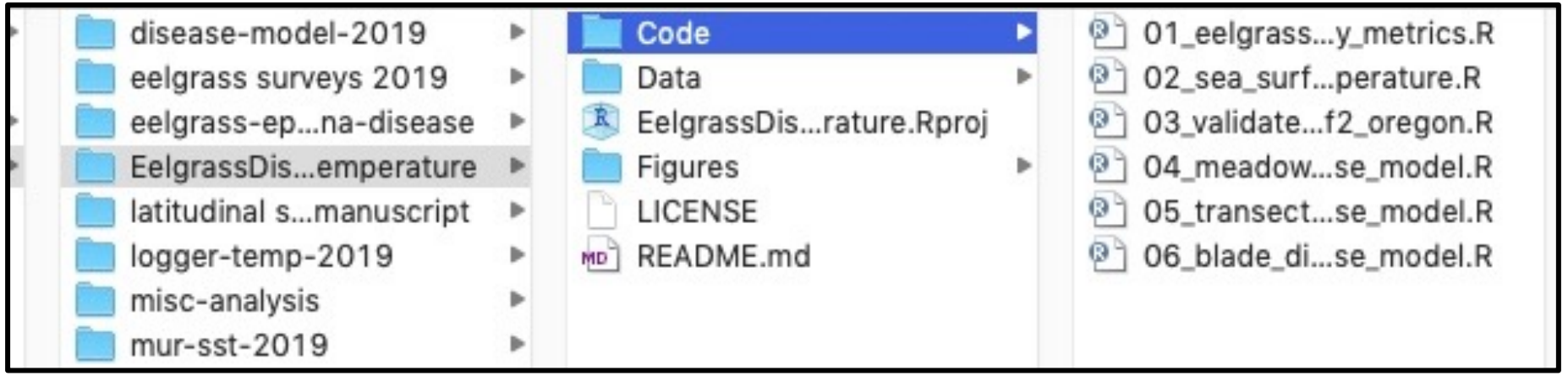
- Separate ‘workflow’ from ‘product’
  - Workflow = things you do because of personal taste and habit, such as the choice of editor for writing code, file paths/directory structure
  - Product = your raw data, the code to turn your raw data into results, your results
- Avoid hard-wiring anything about your workflow into your product
- Many more details from Jenny Bryan
  - https://www.tidyverse.org/blog/2017/12/workflow-vs-script/
  - https://rstats.wtf/project-oriented-workflow.html

File and project organization

How can file organization enhance your research workflow?

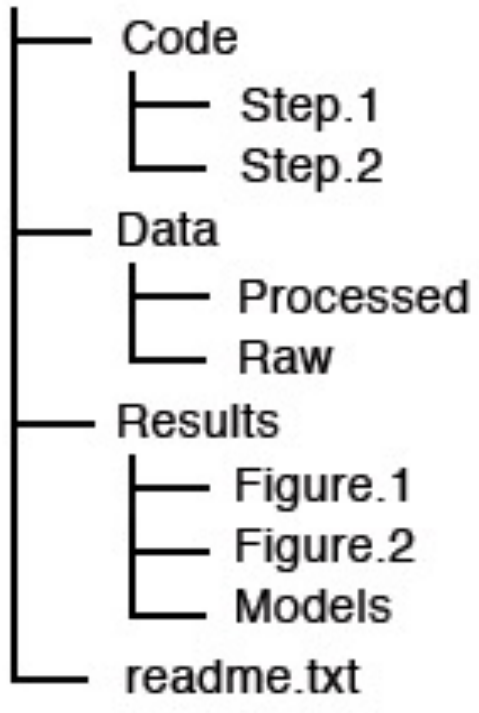


vs

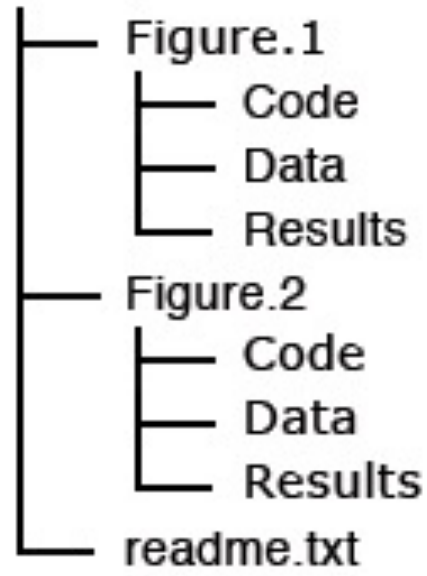


There’s no ‘right’ way to organize your research

A) Organized by File type



B) Organized by Analysis



Best practices for file and folder naming

- Machine readable
  - Avoid spaces, special characters
  - Deliberate_delimiters
- Human readable
  - CamelCase
  - more_deliberate-delimiters
- Works with default ordering
  - 01_first_script
  - 10_tenth_script
  - 2002-09-06_data.csv
  - 2004-06-09_data.csv

Adapted from J. Bryan

File Organization Exercise

1. Consider the files for (one of) your research projects. Diagram or screenshot your directory structure.

What works and what doesn't work about this structure?

Who else might need access to these files?

2. Assess your naming scheme for the files related to this project.

What kinds of files do you create and in what formats?

What are the unique characteristics of these files? E.g. date created, experiment number, investigator, location

Use the unique identifiers to draft file names

3. Create a systemic folder hierarchy

How can you group the individual files into folders?

Can you improve the directory structure to address the needs you identified in (1)?