Good Data Practices

Data4ML

Summer 2022

It’s All About Data

- Good Machine Learning starts with good datasets
- What makes a dataset ‘good’
  - Reproducible
  - ‘FAIR’ – Finable Accessible Interoperable and Reusable
- How to plan out your data taking
- What makes a data set ‘AI-Ready’

Reproducible research… what is it good for?

A published data analysis is reproducible if the analytic data sets and the computer code used to create the data analysis are made available to others for independent study and analysis.

This definition is sufficiently vague that it ultimately raises more questions than it answers. What is an "analytic data set"? What does it mean to be "available"? What is included with the "computer code"?

Peng and Hicks, 2021, *Reproducible Research: A retrospective*

Open science: a new framework for research

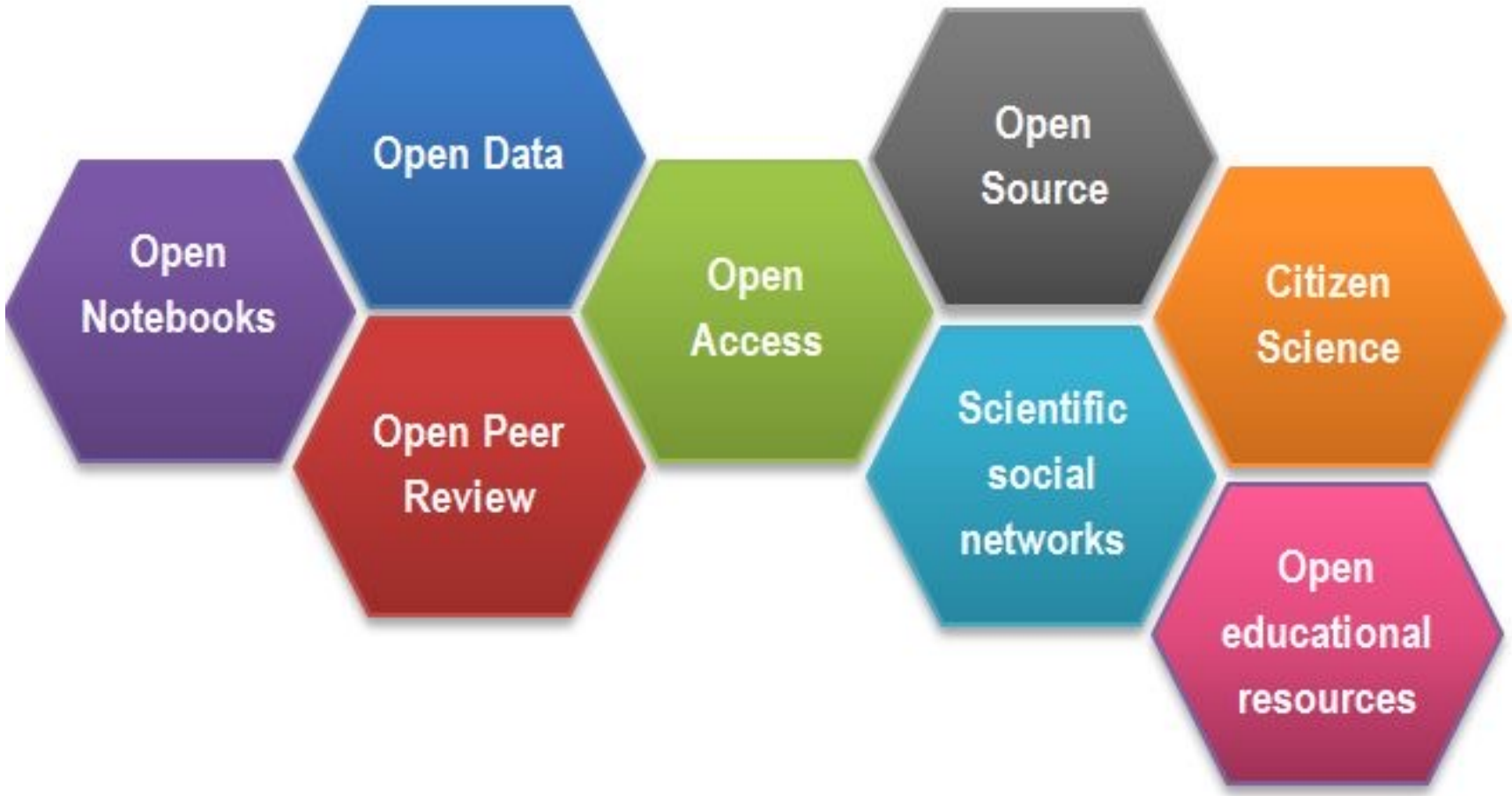


Image credit G. de la Fuente

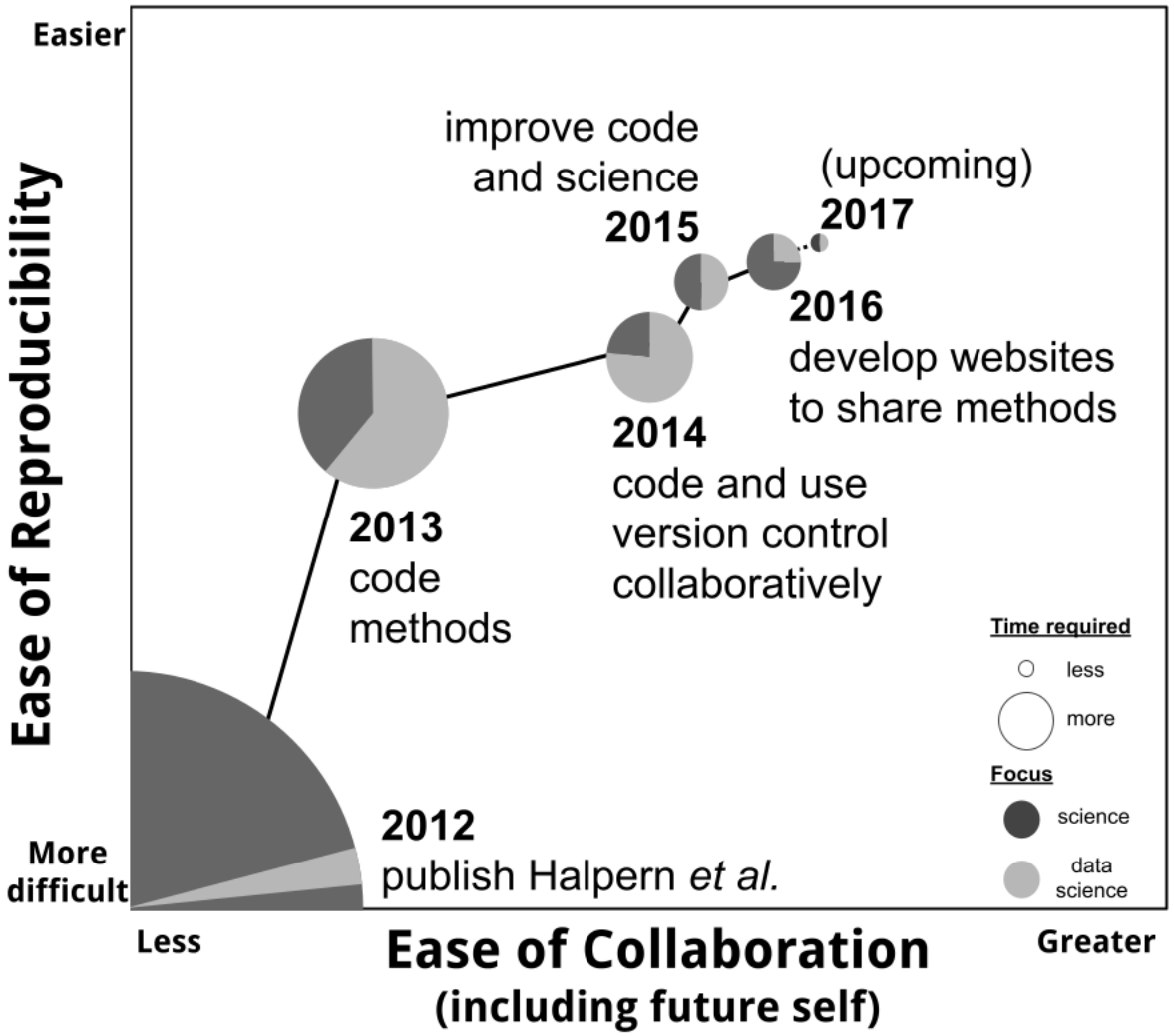
Findable Accessible Interoperable Reusable

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

Global Indigenous Data Alliance

https://www.gida-global.org/

Reproducibility enhances collaboration



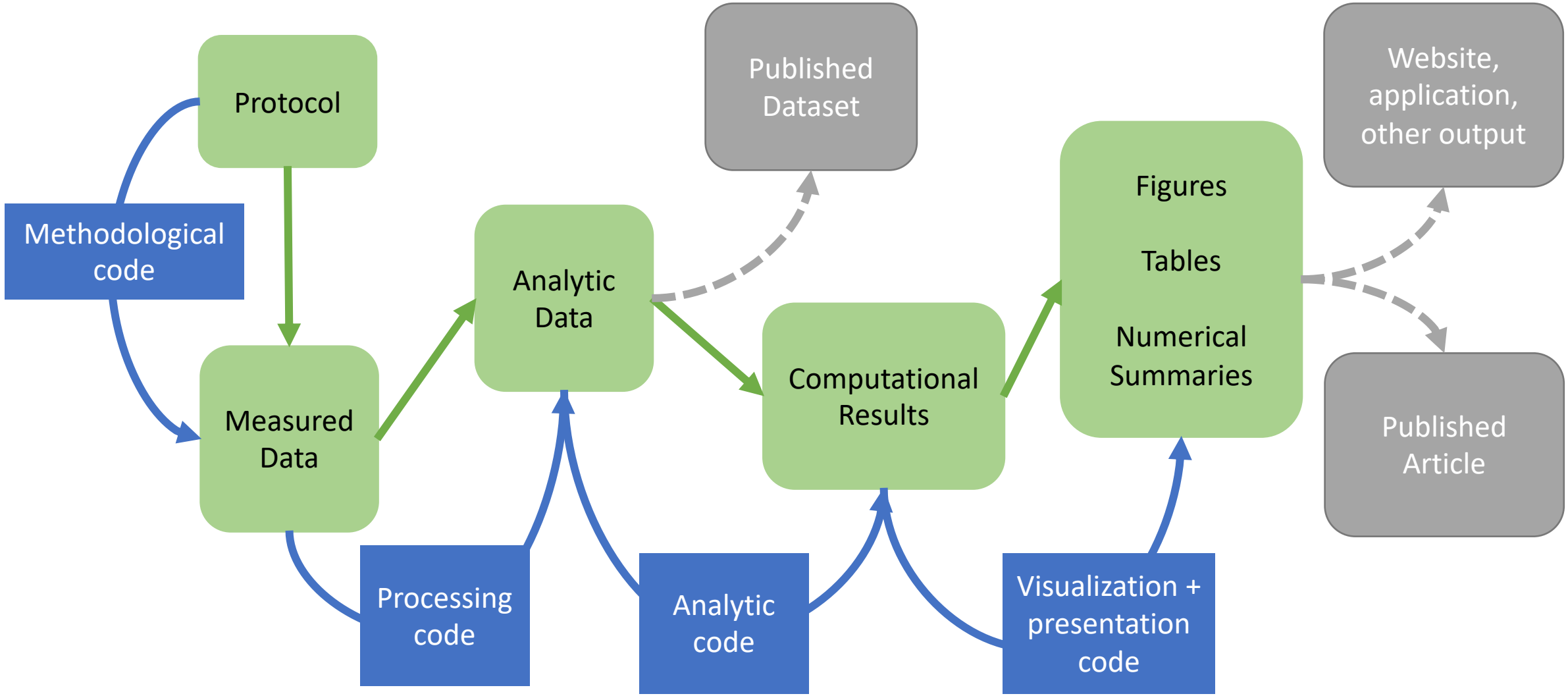
nature ecology & evolution
PERSPECTIVE
PUBLISHED: 23 MAY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0160

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
- Be consistent + reduce decision fatigue

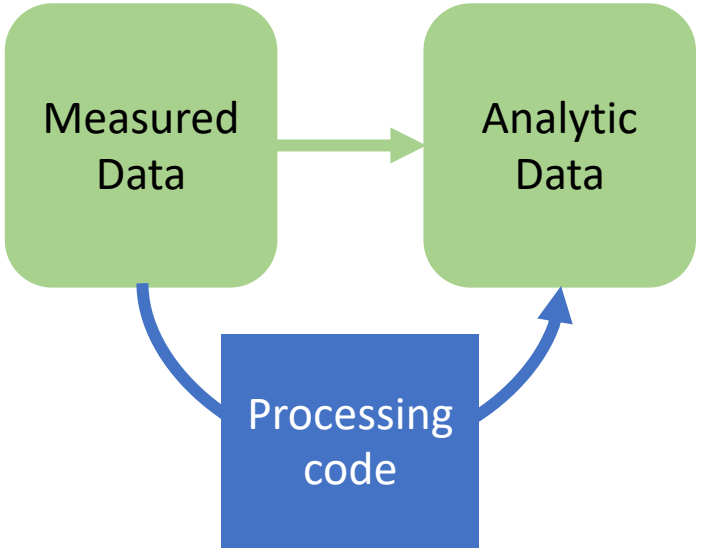
Collaboration exercise

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

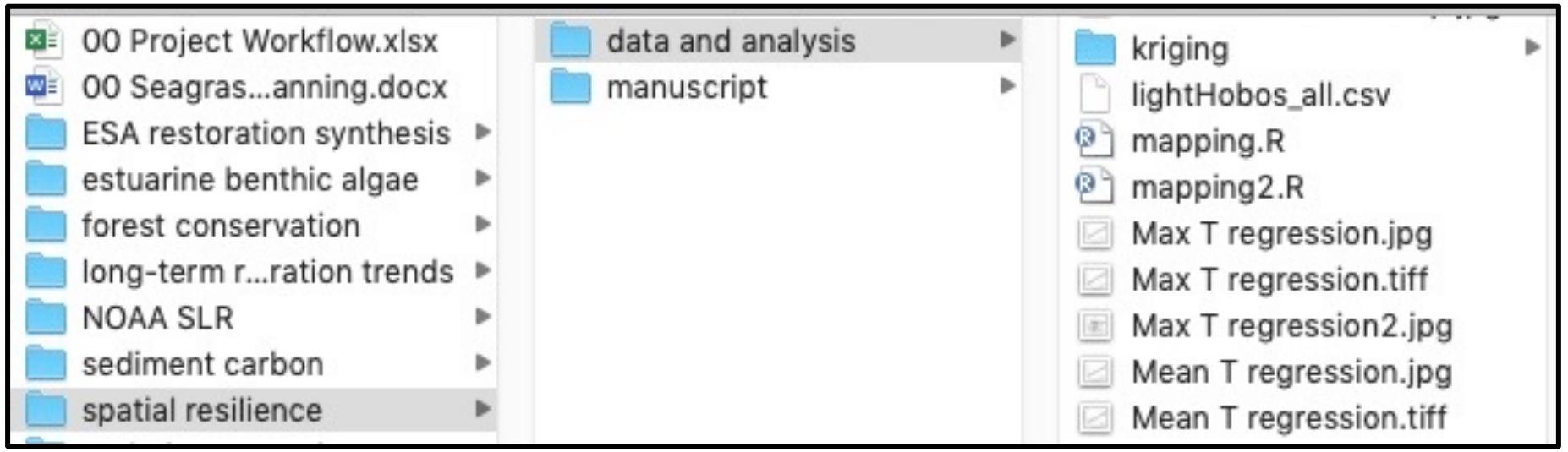
- Separate ‘workflow’ from ‘product’
  - Workflow = personal choices
  - Product = elements you want to reproduce
- Avoid hard-wiring your workflow into your product
- Organize work into ‘projects’



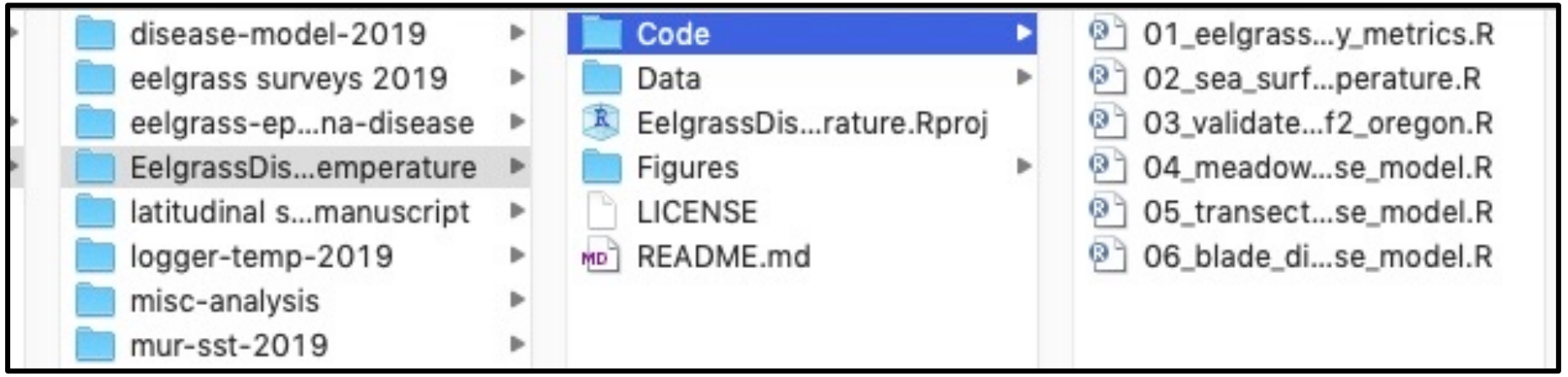
https://www.tidyverse.org/blog/2017/12/workflow-vs-script/
https://rstats.wtf/project-oriented-workflow.html

File and project organization

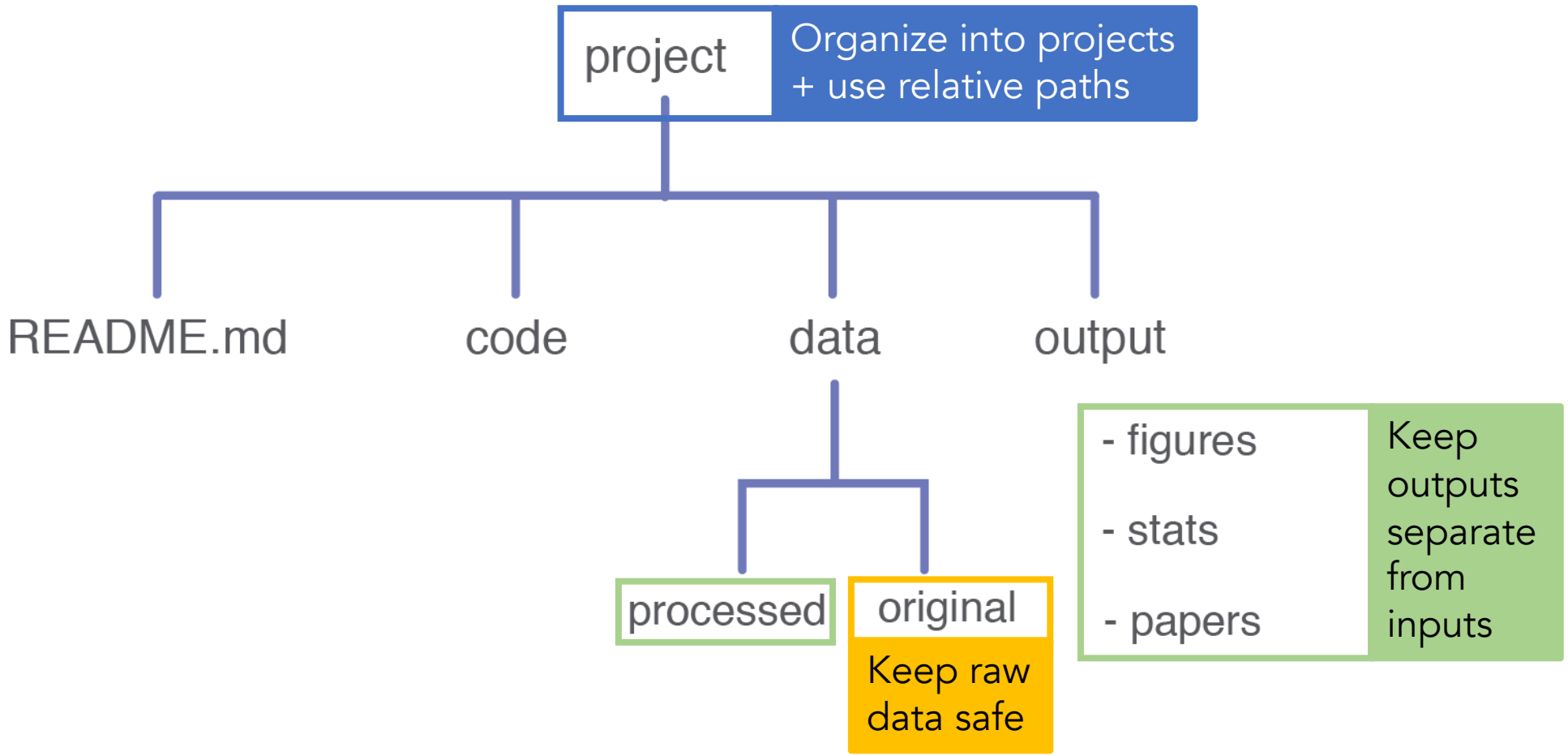
How can file organization enhance your research workflow?



VS



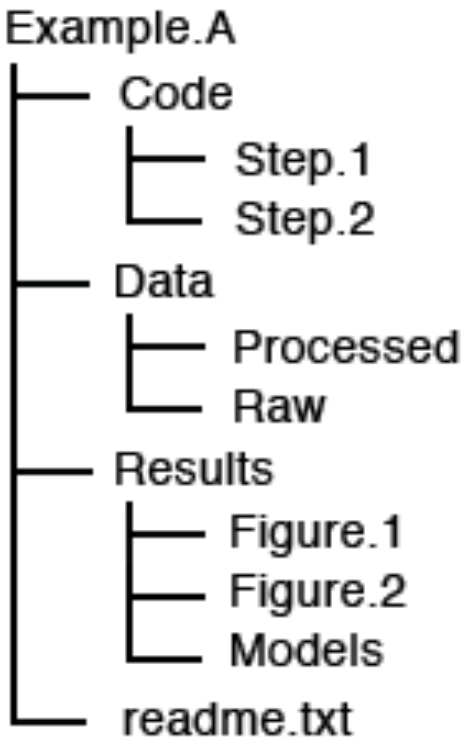
Best practices for project structure



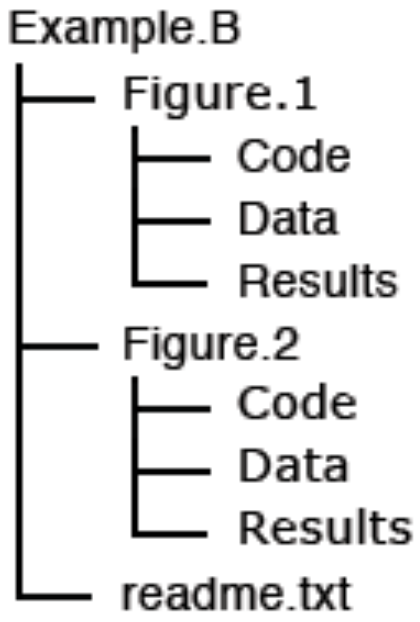
Adapted from https://reproducible-science-curriculum.github.io/

No one way to organize your research

A) Organized by File type



B) Organized by Analysis



- Decide what works for you!
- Aim for consistency
- Automate?

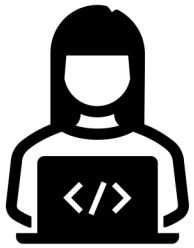
Adapted from Helsinki University Library

Best practices for file and folder naming

- Machine readable

- Avoid spaces, special characters
- Deliberate_delimiters

- Human readable



- CamelCase
- more_deliberate-delimiters

- Works well 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)?

Adapted from MIT Libraries

Backups and Permissions

- It’s always a good idea to keep a copy of your data as a backup
  - Hardware can fail
  - Running the wrong command can accidently delete data
- It’s also a good idea to protect yourself from accidental code
  - All files have permissions that decide who can do what
    - ***read*** permissions let you look at file
    - ***write*** permissions let you change a file
    - ***execute*** permissions let you run a file
  - It’s a good idea to protect your raw data by removing all permissions but read
    - prevents accidentally changing the raw data

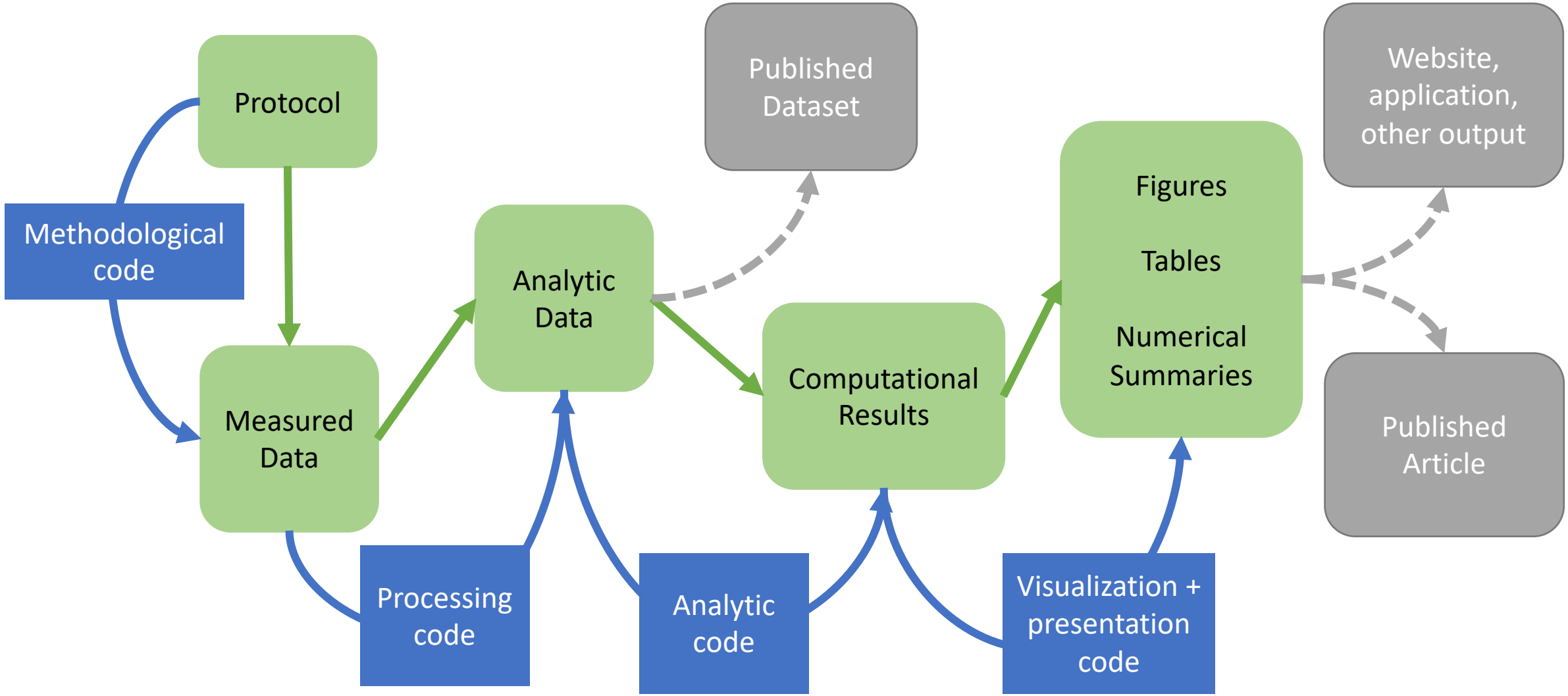
What Makes a Dataset AI Ready?

- Clear Data Splits
  - Designate some examples as training data
  - Designate some examples as testing data
- Structured Data is ‘Tidy’ - https://vita.had.co.nz/papers/tidy-data.pdf
  - Every column is a variable.
  - Every row is an observation.
  - Every cell is a single value.
- The Dataset is Clean
  - No missing values or ‘Nans’
  - No ‘bad’ data
  - Consistent as Possible

Why do we split datasets?

- Machine Learning Algorithms learn from data
  - Often powerful enough to ‘memorize’ or overfit data
- You’re training ML models to work with new data
  - Memorizing training data isn’t good enough
  - Normally called generalization
    - We want models to learning something from the training data the applies to future data as well
- How do we know if the model generalizes?
  - We check by splitting off some test data
  - In contests the testing labels aren’t published
  - In research settings we use testing data to compare models
  - In both cases it’s useful for the dataset creator to decide, so it’s consistent moving forward

Where do you make your data AI ready?



Adapted from R. Peng

Example Layouts

- Example 1
  - Code
    - Step.1.clean
    - train_model
  - Data
    - Raw
    - Processed
      - **ML_Ready**
        - **Training.csv**
        - **Testing.csv**
  - Results

- Example 1
  - Code
    - Step.1.clean
    - train_model
  - Data
    - Raw
    - Processed
      - **ML_Ready**
        - **Training**
          - labels.csv
          - Images
            - Img1.png
            - Img2.png
            - …
        - **Testing**
          - labels.csv
          - Images
            - Img100.png
            - Img201.png
            - …
  - Results

- Example 2
  - Data
    - Raw
    - Processed
      - **ML_Ready**
        - **Training**
          - **Cats**
            - Img1.png
            - Img2.png
            - …
          - **Dogs**
            - Img1.png
            - Img2.png
        - **Testing**

…same layout as Training

Example Differences between Reproducible and AI-Ready

Blurry Image results from Bad Calibration

- Useful for your future-self to debug
- Not useful for training an ML model

An instrument is broken, so you're missing a variable

- Useful if you don't need that variable
- Not useful for training an ML model that needs that variable

Contaminated Sequences

- Useful for your future-self to debug
- Not useful for training an ML model

You switch units on an instrument

- Useful it's still the same data
- Not useful for training an ML model you must convert to consistent units in advance

When taking data consider recording whether your confident if that data should be used for science or if there is a possibility something went wrong. This can help you clean your data for ML projects, can True/False, or more fine grained red/yellow/green.

Sometimes it’s ok to break the rules!

- It’s often OK to break the rules when make AI-Ready Datasets. Just make sure you know why
- Examples:
  - Some ML algorithms can handle missing values
    - Most can not, so make sure to check
  - There are other ways of validating ML models besides a training and test split
    - Cross validation - K-Folds, Leave one out
    - These are often computationally expensive, so generally they work with small models and small datasets where you want to use as much data as possible for training

Next Steps

- Were going to practice moving files and setting up folders on the command line
- We'll get an existing dataset to be "AI-Ready"
- Running and training ML algorithm will start here.