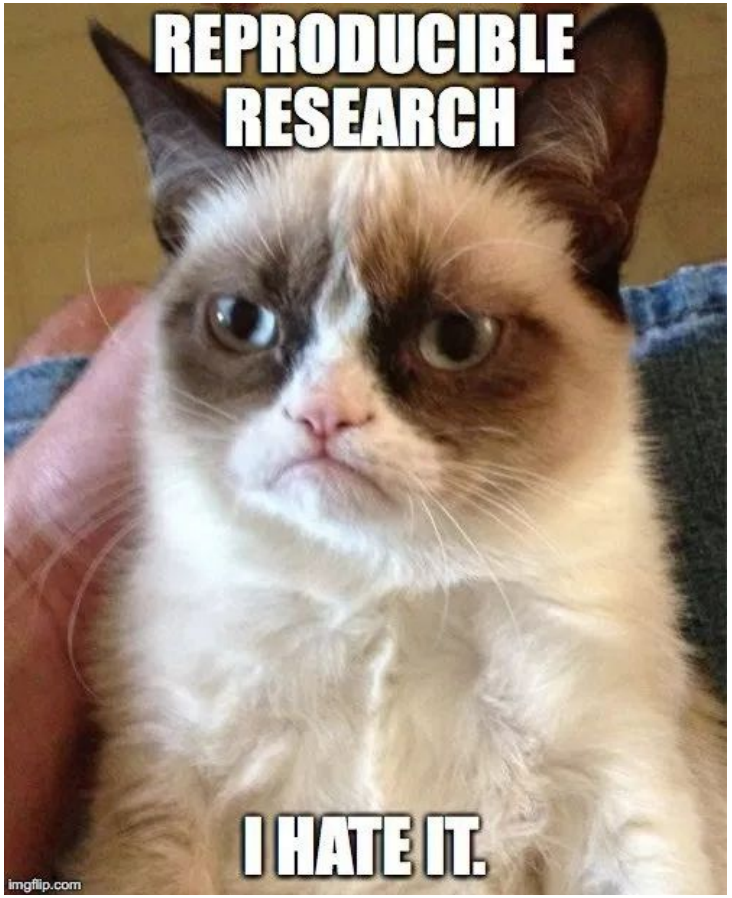
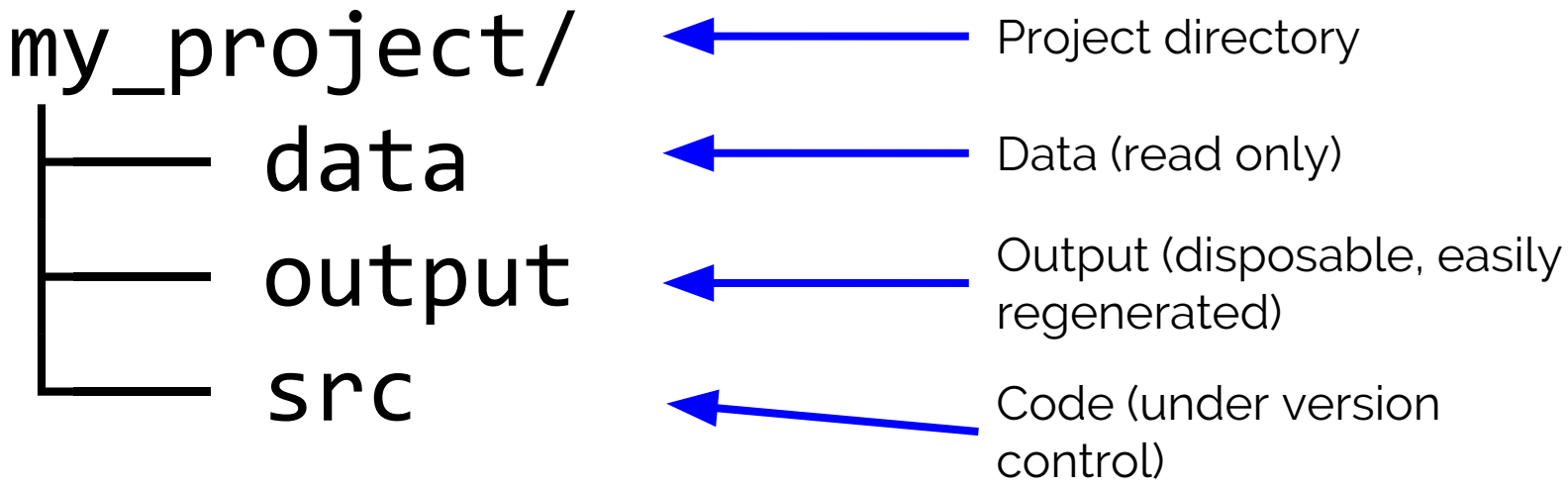
Reproducible Analysis demo

Lecture 10
BIOS 6660, Spring 2019
Instructor: Pam Russell

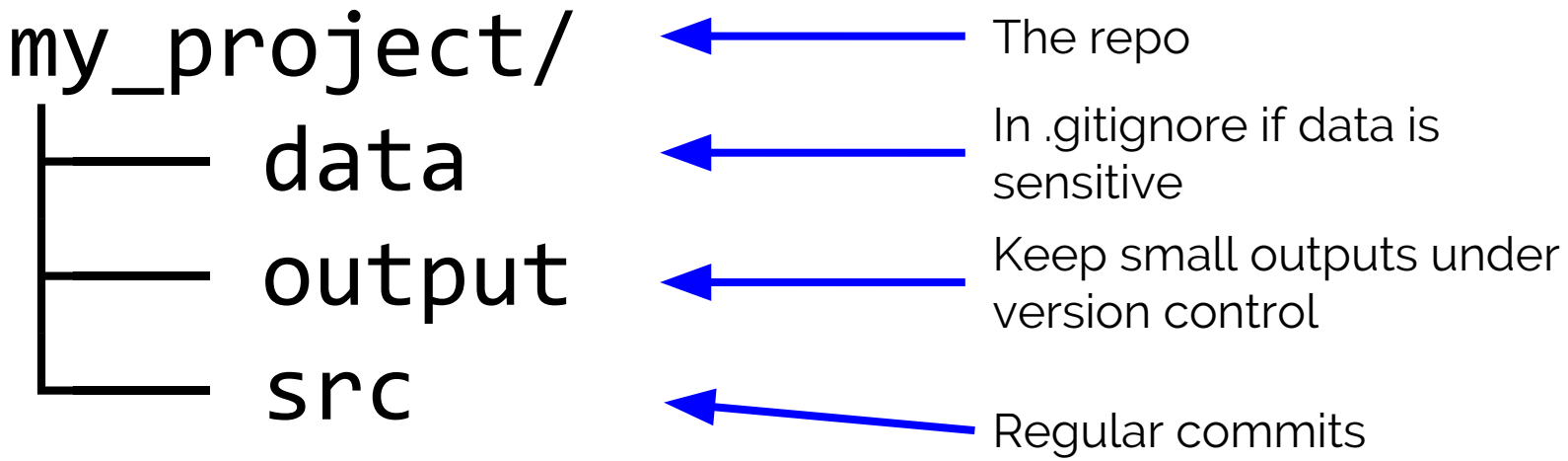


https://www.r-bloggers.com/reproducible-research-write-your-clinical-chemistry-paper-using-r-markdown/

Organization



Version control



New data from project

- Full data management and sharing practices from last week

Public data

- From data repository: document DOI
- From public database: document version
- Paper supplemental data: document paper
- Small dataset with open license: can go on GitHub

- Record a digital fingerprint (more on this later)
- Put in **`data`** directory with documentation in **`README.txt`**
- Remove write permissions from file(s)

R Markdown

We use R Markdown for complete analyses on Thursday and on Homework 5

RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

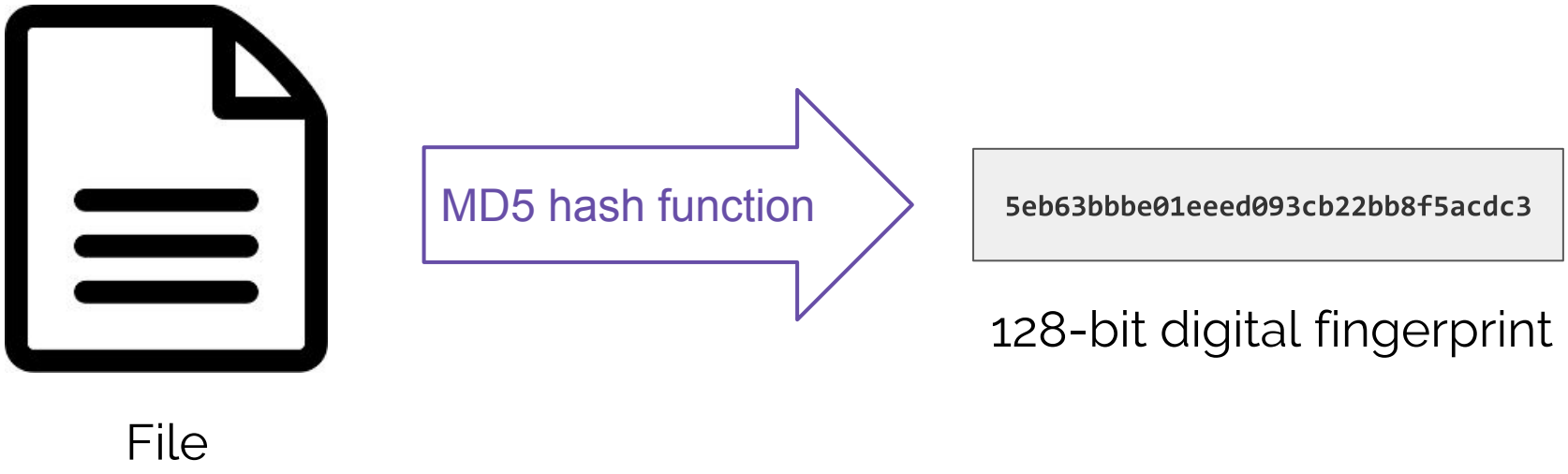
RULE #7—ALWAYS STORE RAW DATA BEHIND PLOTS

RULE #8—GENERATE HIERARCHICAL ANALYSIS OUTPUT, ALLOWING LAYERS OF INCREASING DETAIL TO BE INSPECTED

RULE #9—CONNECT TEXTUAL STATEMENTS TO UNDERLYING RESULTS

Data fingerprint

Capture a digital fingerprint of the data so future users can verify their copy of the data



Work from raw data

Workflow starts with loading raw data

Should always be able to delete any intermediate data and run entire workflow from raw data

RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

RULE #2—AVOID MANUAL DATA MANIPULATION STEPS

Exploratory analysis

Exploratory plots

Helps make decisions about future analysis

Keep under version control

Mostly for you to come back to

RULE #4—VERSION CONTROL ALL CUSTOM SCRIPTS

Main analysis

Funding agency and journal requirements:
Mostly stop at theoretical reproducibility.
No requirement of practical reproducibility.

Make a good faith effort toward practical reproducibility.
Put yourself in the user's shoes!

Main analysis

Basic requirements

- Keep code under version control
- Share repo publicly

“Good faith” requirements:

- Documentation in GitHub README
  - Repo contents
  - Mapping between paper results and scripts
  - How data is imported and moves through pipeline
- Code comments to guide new users

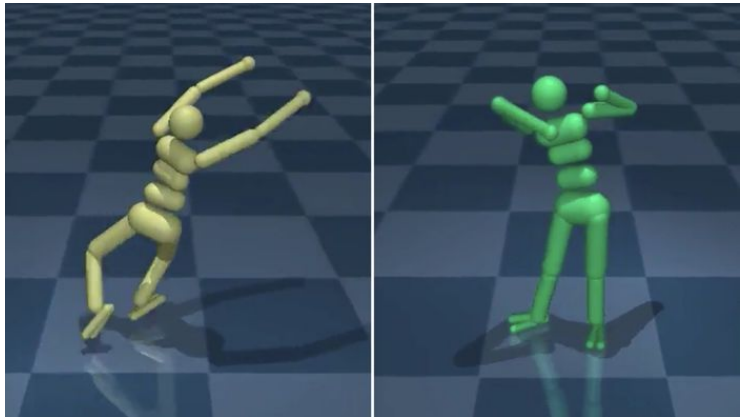
RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

RULE #4—VERSION CONTROL ALL CUSTOM SCRIPTS

RULE #10—PROVIDE PUBLIC ACCESS TO SCRIPTS, RUNS, AND RESULTS

Analyses with randomness

- Any analysis with randomness: machine learning, simulations, …
- Provide pseudo-random number generator with an initial value
- Subsequent runs will get same sequence of “random” numbers
- In R: **`set.seed()`**

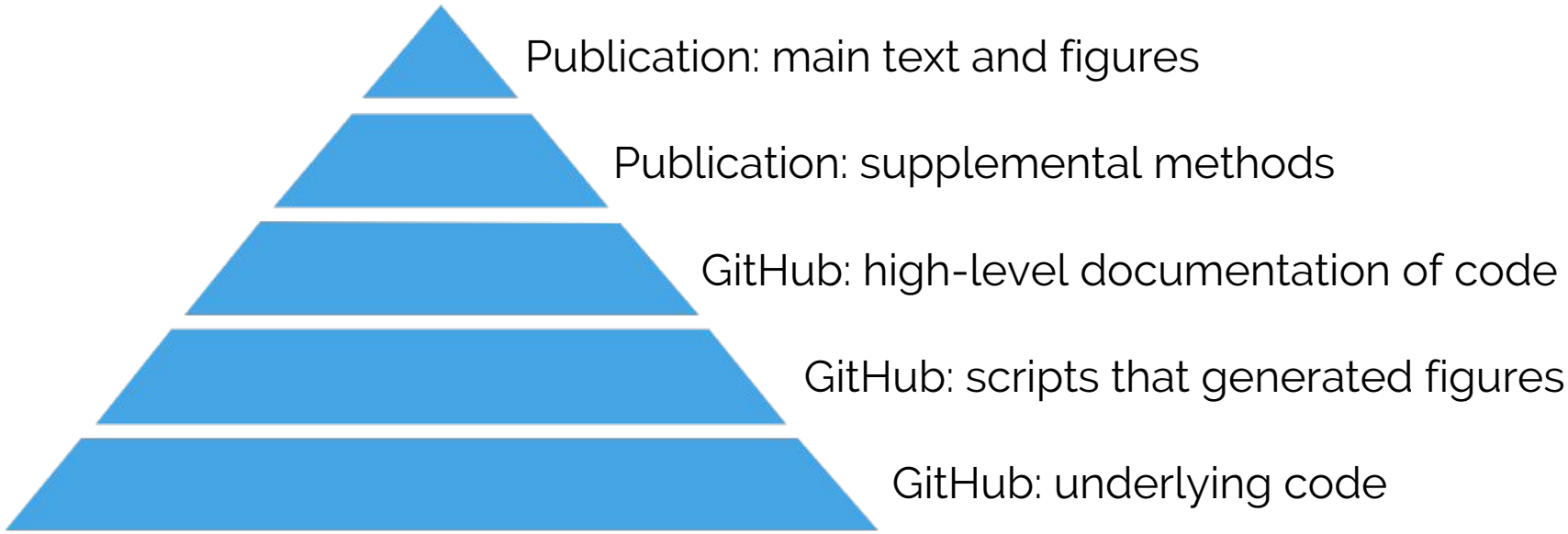


Algorithm learning to walk differently with different initial conditions

https://doi.org/10.1126/science.aat3298

RULE #6—FOR ANALYSES THAT INCLUDE RANDOMNESS, NOTE UNDERLYING RANDOM SEEDS

Output



RULE #8—GENERATE HIERARCHICAL ANALYSIS OUTPUT, ALLOWING LAYERS OF INCREASING DETAIL TO BE INSPECTED

RULE #10—PROVIDE PUBLIC ACCESS TO SCRIPTS, RUNS, AND RESULTS

Software environment

- On Linux
  - Minimum requirement: record all program versions and system information
  - Better: use a container
- In R: **`sessionInfo()`**

RULE #3—ARCHIVE THE EXACT VERSIONS OF ALL EXTERNAL PROGRAMS USED