software carpentry

TGen
June 19-20th, 2017

Instructors:

Nick Banovich
Emily Davenport

Helpers:

Chistophe Legendre
Elizabeth Hutchins
Eric Alsop
Ryan Richholt

Goal:

Learn core skills for doing data analysis effectively, efficiently, and reproducibly.

1. Interacting with your computer on command line (BASH/shell)
2. Programming fundamentals ®
3. Version control (Git)

Do you suffer from any of the following?

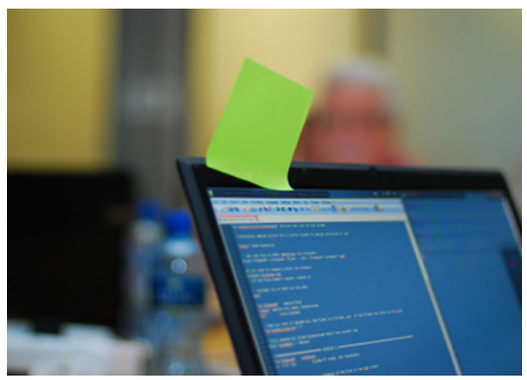
software carpentry

- I usually manage data in excel, but that's caused some errors with dates and I want to learn a different way.
- My advisor insists that we store 50,000 barcodes in a spreadsheet, and something must be done about that.
- I'm having a hard time analyzing microarray, SNP, or multivariate data with Excel and Access.
- I want to use publicly available data, but it's confusing to download it through command line.
- I'm interested in going into industry and companies are asking for data analysis experience.
- I'm trying to reboot my lab's worker to manage data and analysis in a more sustainable way.
- I'm re-entering data over and over again by hand and know there's a better way.
- I'm tired of feeling out of my depth on computation and want to increase my confidence.
- I see other people's figures and wonder if I could generate something like that with my data.

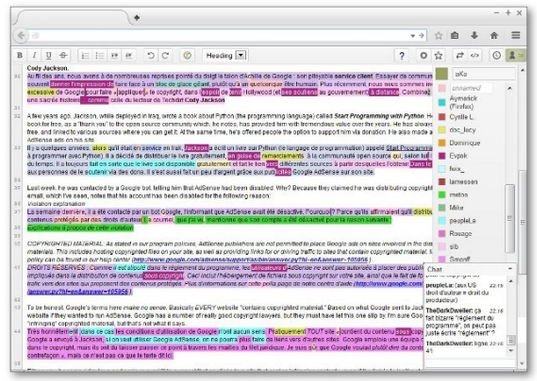
Notes before we start

- Website: https://erdavenport.github.io/2017-06-19-tgen/
- Etherpad: http://pad.software-carpentry.org/2017-06-19-tgen
- Can you see the screen?
- Bathrooms, breaks….
- Getting help: raising hand vs. stickies vs. ether pad

Raise your hand for a question everyone would benefit from.



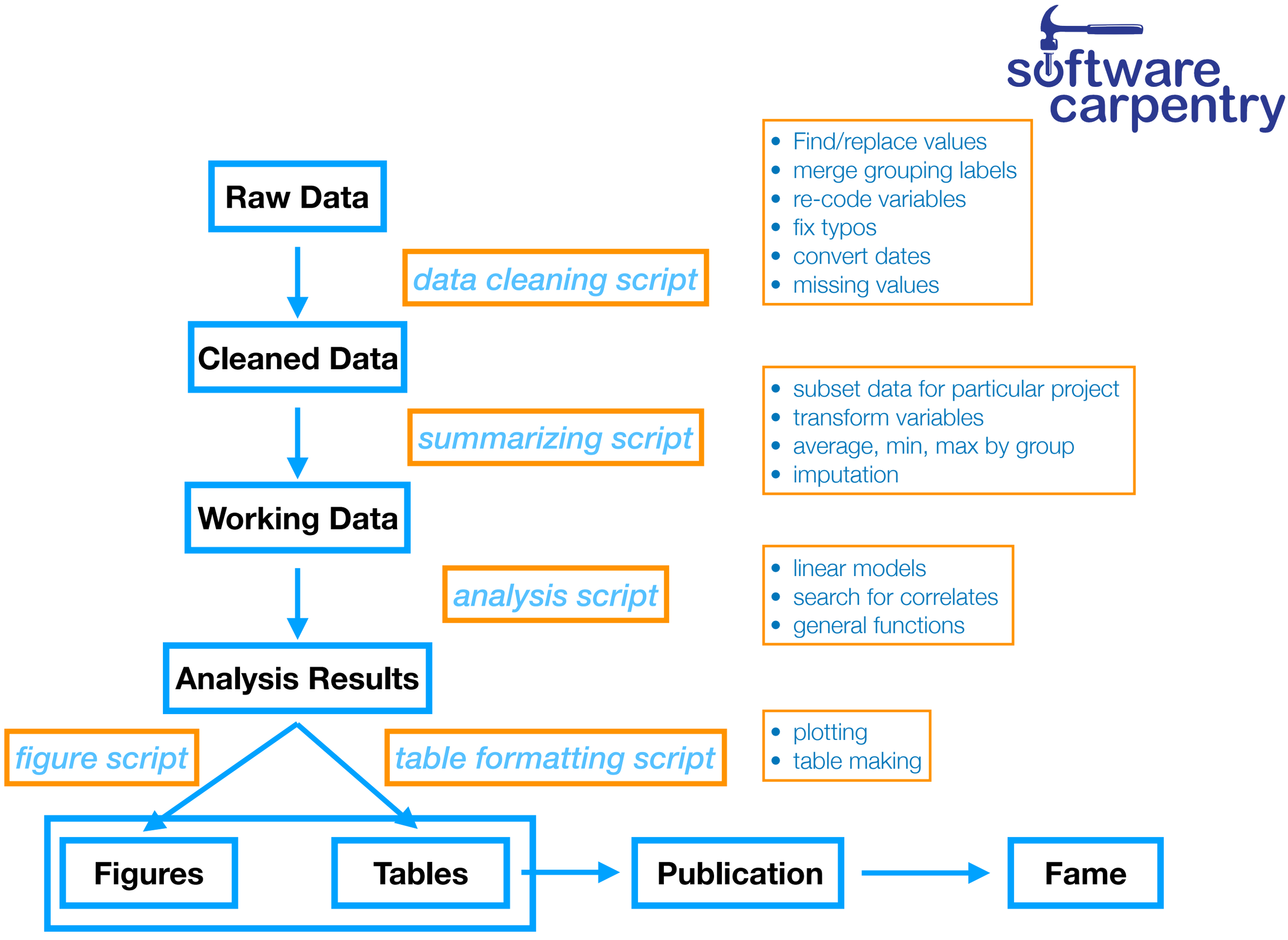
Sticky note when your code doesn't work and you need a helper.

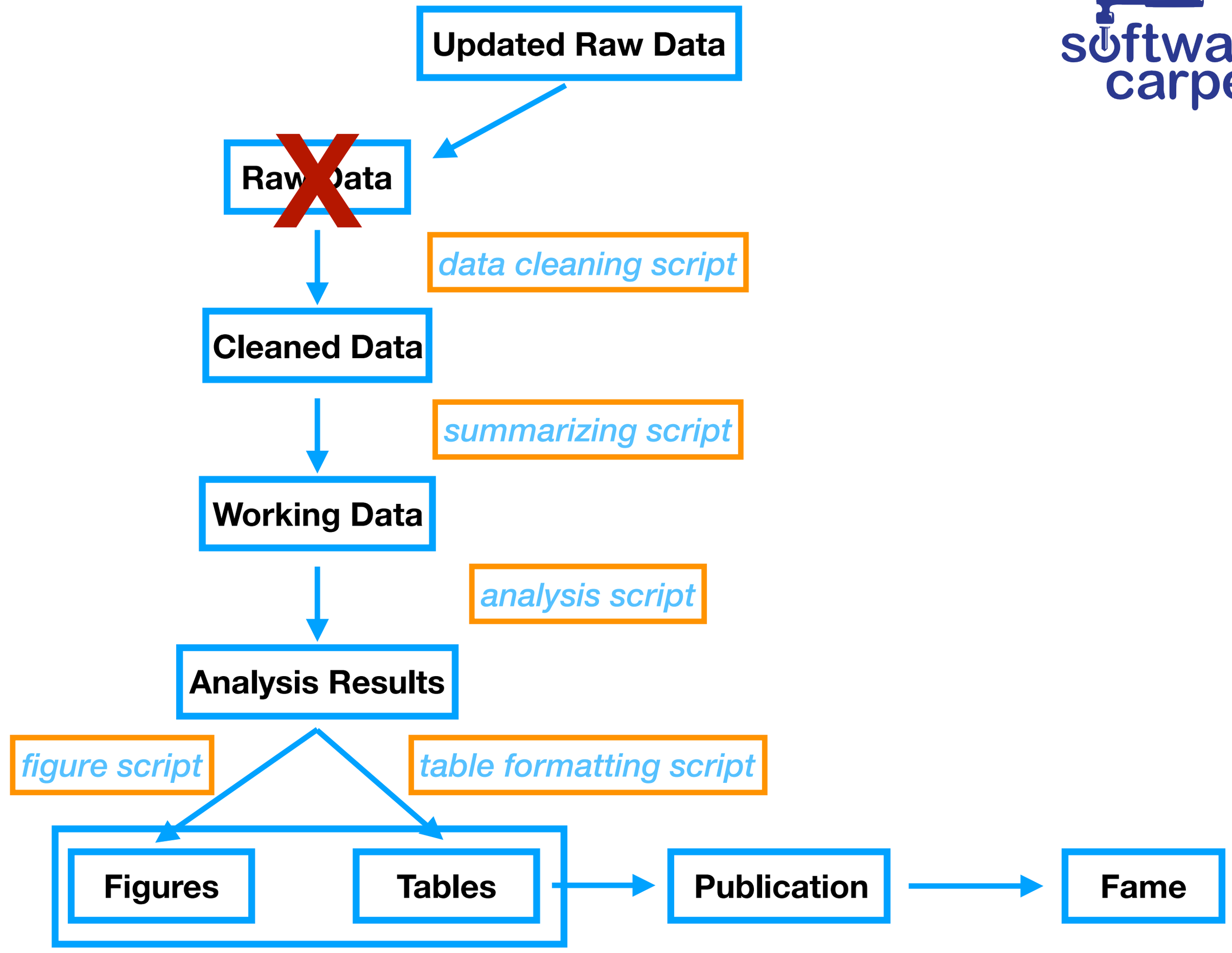


Etherpad for all of the above and for off topic questions.

Reproducible Research

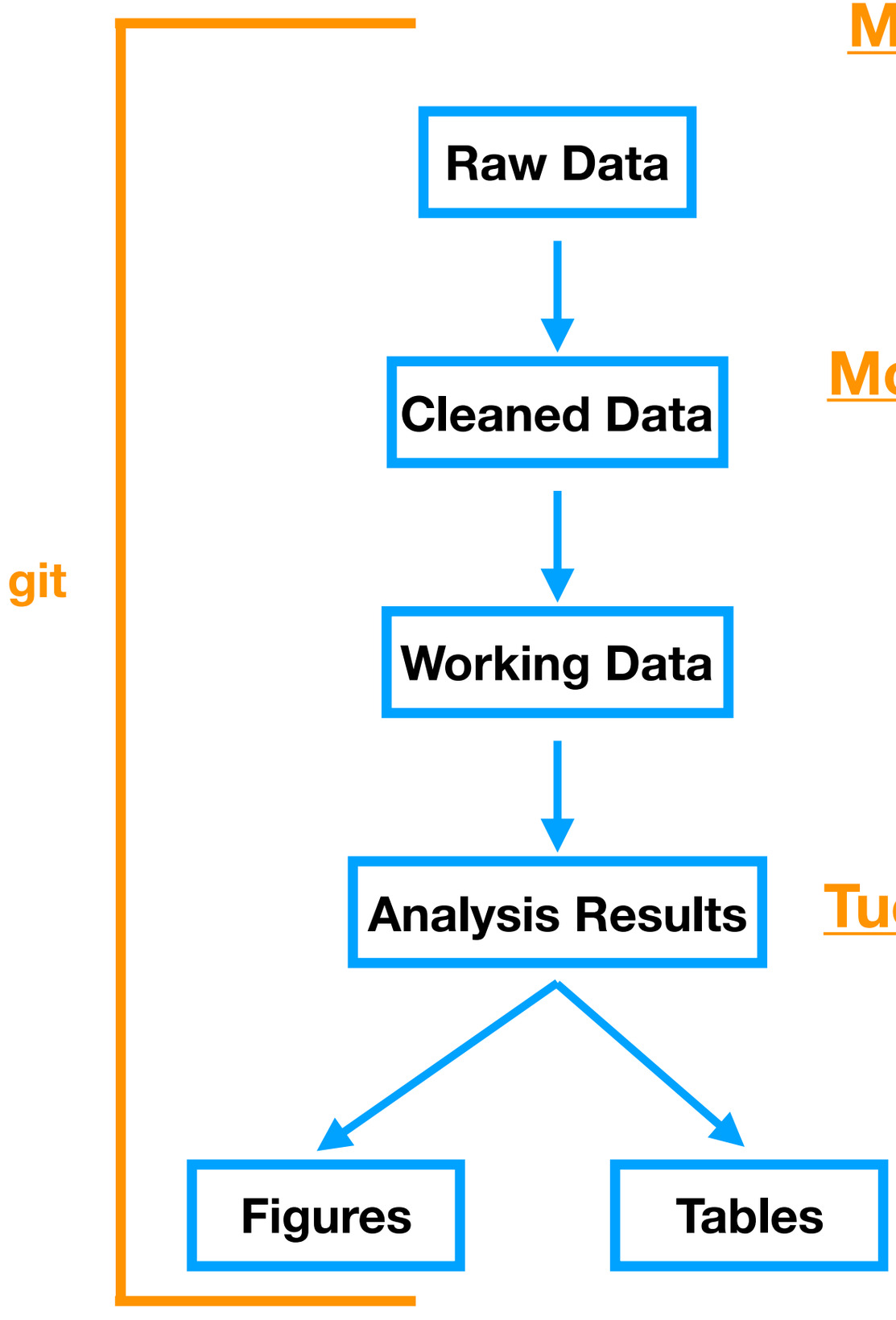
- Well documented and repeatable science.
- Data analysis:
  - Data and analysis can be re-created by anyone
    - Including you in the future!
    - Repeat analysis on updated data.
    - Repeat analysis on similar datasets.
  - Scripted data management and analysis
    - Manages and analyzes
    - Provides a record of what was done
    - Easy to edit and re-run


Raw Data
data cleaning script
Find/replace values
merge grouping labels
re-code variables
fix typos
convert dates
missing values
Cleaned Data
summarizing script
subset data for particular project
transform variables
average, min, max by group
imputation
Working Data
analysis script
linear models
search for correlates
general functions
Analysis Results
figure script
table formatting script
plotting
table making
Figures
Tables
Publication
Fame


Updated Raw Data
Raw Data
data cleaning script
Cleaned Data
summarizing script
Working Data
analysis script
Analysis Results
figure script
table formatting script
Figures
Tables
Publication
Fame

software carpentry

Tuesday morning



Monday morning

BASH/shell

Monday afternoon

Intro to R
R: variables
R: data types
R: loading data
R: subsetting data
R: loops and functions

Tuesday afternoon

R: dplyr
R: ggplot2