DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

Cornell University June 2017

Sponsored by Cornell Statistical Consulting Unit

Instructors

- Erika Mudrak (CSCU)
- Lynn Johnson (CSCU)
- Stephen Parry (CSCU)
- David Kent (Food Science)

Assistants

- Emily Davenport (Molecular Biology and Genetics)
- Francoise Vermeylen (CSCU)
- Kevin Packard (CSCU)
- Michael Ko (CSCU)

Goal:
A Data Carpentry workshop teaches the core skills for working with data effectively and reproducibly.

Community driven effort

Staff

- **Executive Director**
  Tracy K. Teal, PhD, Michigan State University
- **Associate Director**
  Erin Becker, PhD
- **Program Coordinator**
  Maneesha Sane
- **Deputy Director of Assessment**
  Kari Jordan, PhD

Steering Committee Members

- Karen Cranston, PhD, Principal Investigator, Open Tree of Life
- Hilmar Lapp, Director of Informatics, Duke Center for Genomic & Computational Biology
- Aleksandra Pawlik, PhD, Training Lead, Software Sustainability Institute
- Karthik Ram, PhD, rOpenSci co-founder, Berkeley Institute for Data Science Fellow
- Ethan White, PhD, Associate Professor, University of Florida

Open source materials

https://github.com/datacarpentry/datacarpentry/

Sentiments on data within the NSF BIO Centers (BEACON, SESYNC, NESCent, iPlant, iDigBio)

- I usually manage data in Excel and it's terrible and I want to do it better.
- I'm organizing GIS data and it's becoming a nightmare.
- 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 public data.
- I work with faculty at undergrad institutions and want to teach data practices, but I need to learn it myself first.
- I'm interested in going in to industry and companies are asking for data analysis experience.
- I'm trying to reboot my lab's workflow 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 have overwhelming amounts of data.
- I'm tired of feeling out of my depth on computation and want to increase my confidence.

Notes before we start

- **Website:** https://emudrak.github.io/2017-06-14-cornell/
  - Will have links to lessons after we go through them

- **Etherpad:** http://pad.software-carpentry.org/2017-06-14-cornell
  - Instructor will update with current code and monitor questions,
- Can you see the screen? Insight…
- Bathrooms, breaks…

Two kinds of questions

Raise your hand for a question that everyone could benefit

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

Reproducible Research

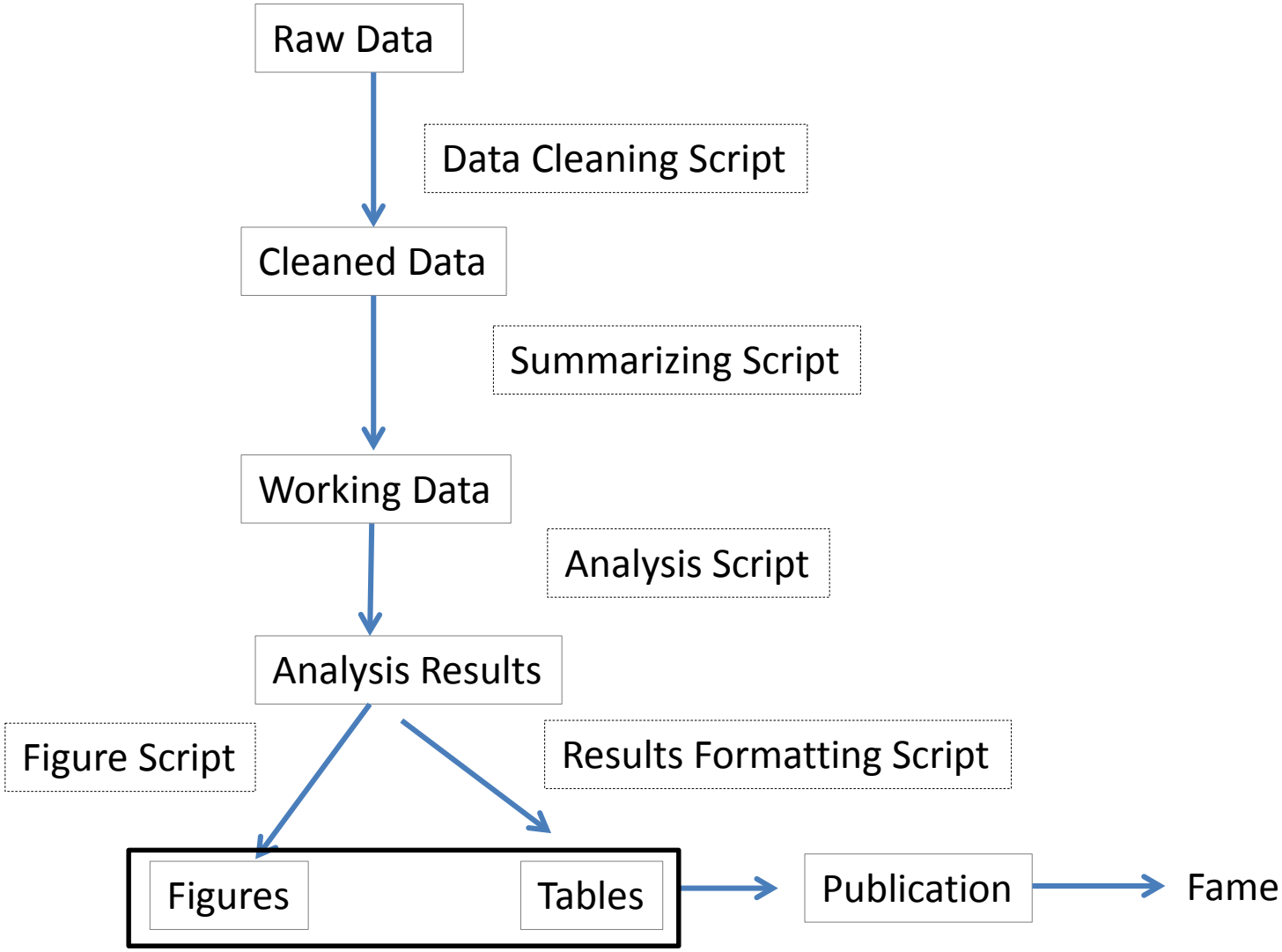
Well documented

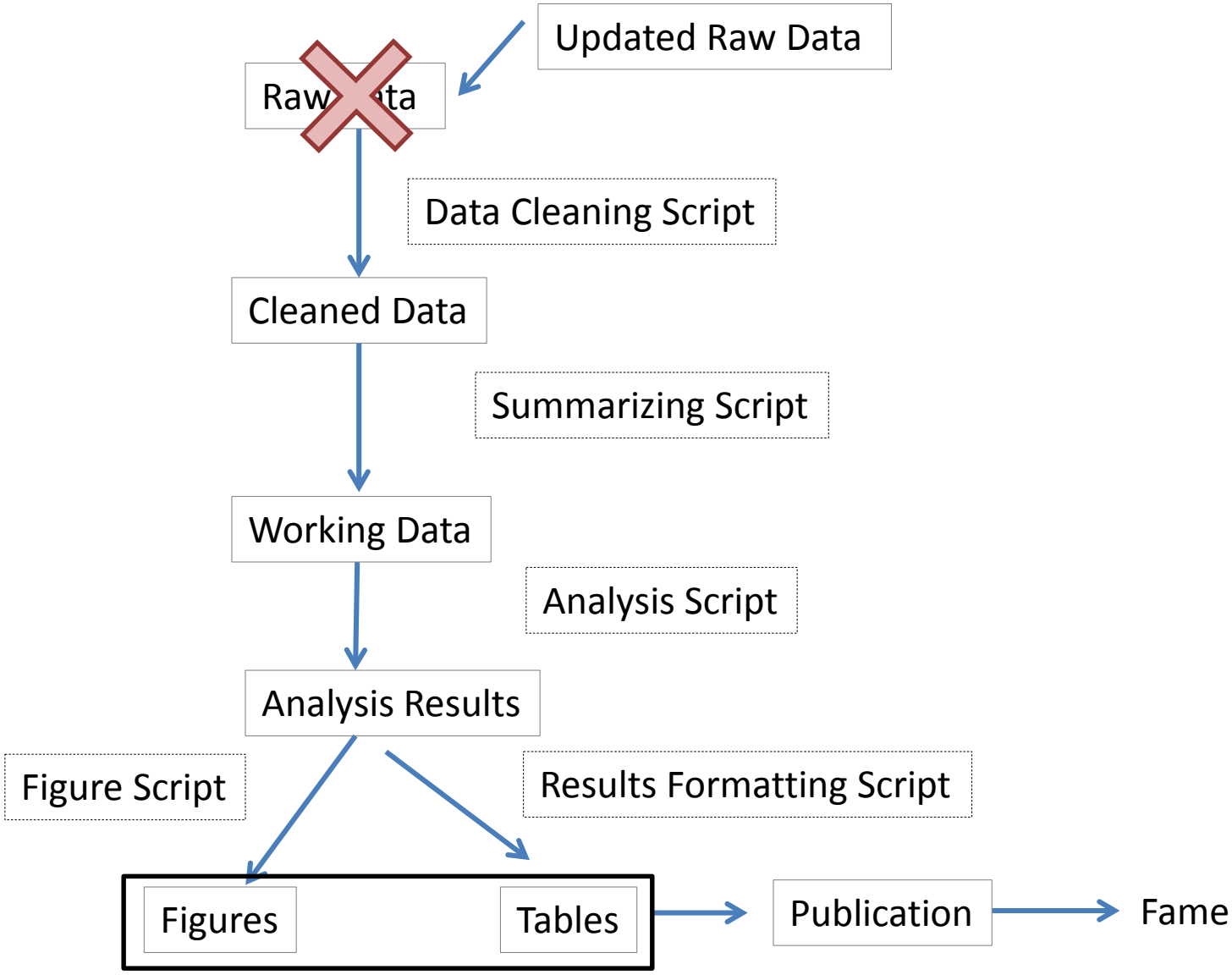
and

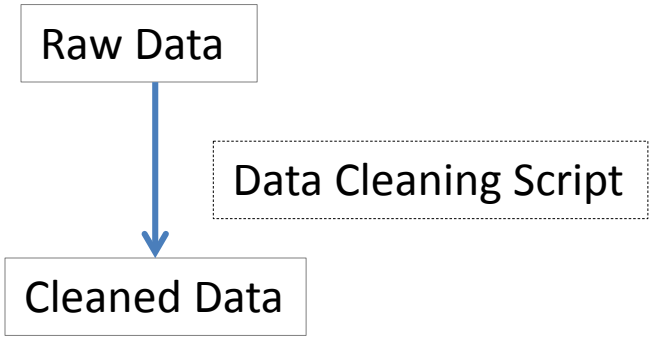
Repeatable

Reproducible Research

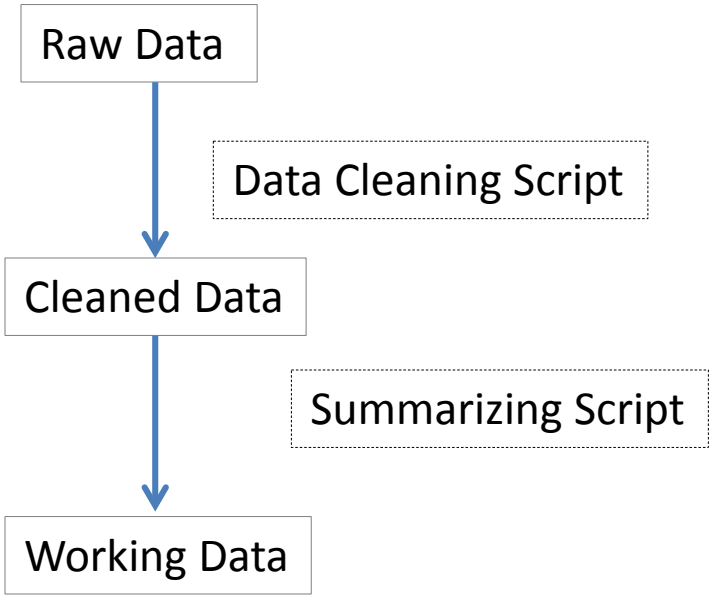
- Data analysis
  - Data and analysis can be re-created by anyone
    - Including you in the future!
    - Repeat analysis on updated data
    - Repeat analyses on similar datasets
  - Scripted data management and analysis
    - Manages and analyzes
    - Provides a record of what was done
    - Easy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


Raw Data
Data Cleaning Script
Cleaned Data
Summarizing Script
Working Data
Analysis Script
Analysis Results
Figure Script
Results Formatting Script
Figures
Tables
Publication
Fame

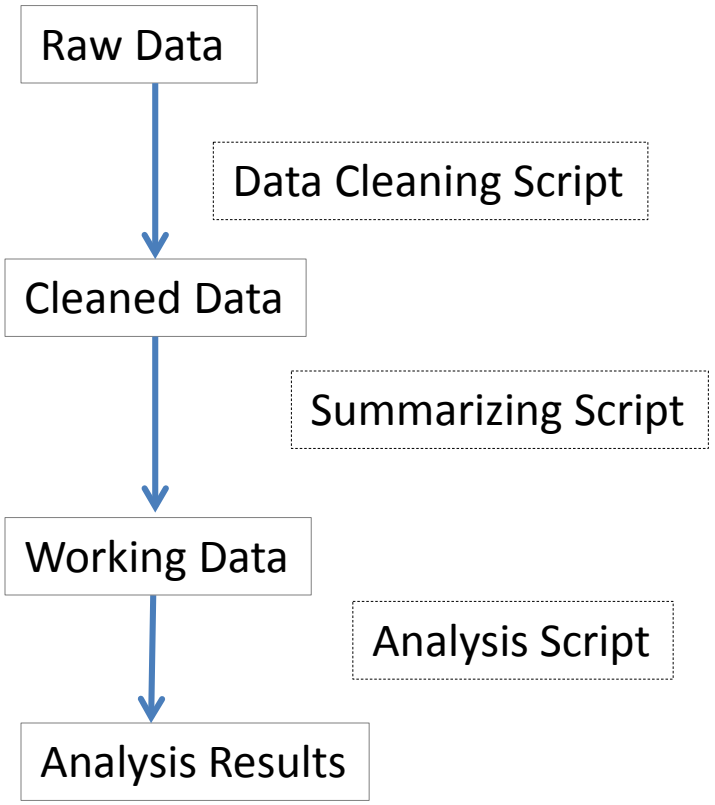

Updated Raw Data
Raw ata
Data Cleaning Script
Cleaned Data
Summarizing Script
Working Data
Analysis Script
Analysis Results
Figure Script
Results Formatting Script
Figures
Tables
Publication
Fame



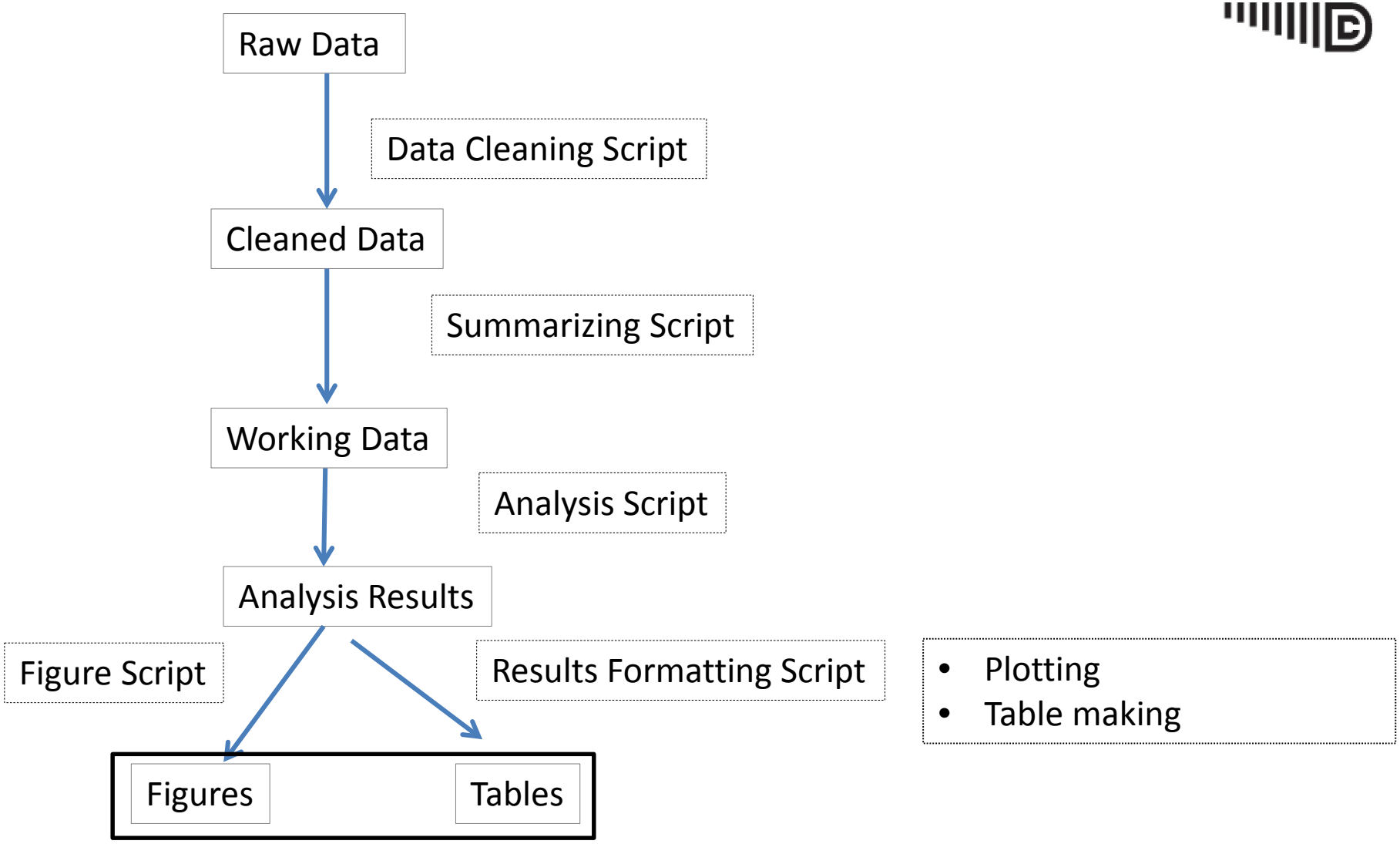
- Univariate & Bivariate EDA
- Find/Replace values
- Merge grouping labels
- Re-code variables
- Fix typos
- Standardize entries
- Convert dates
- Convert variable formats
- Missing values

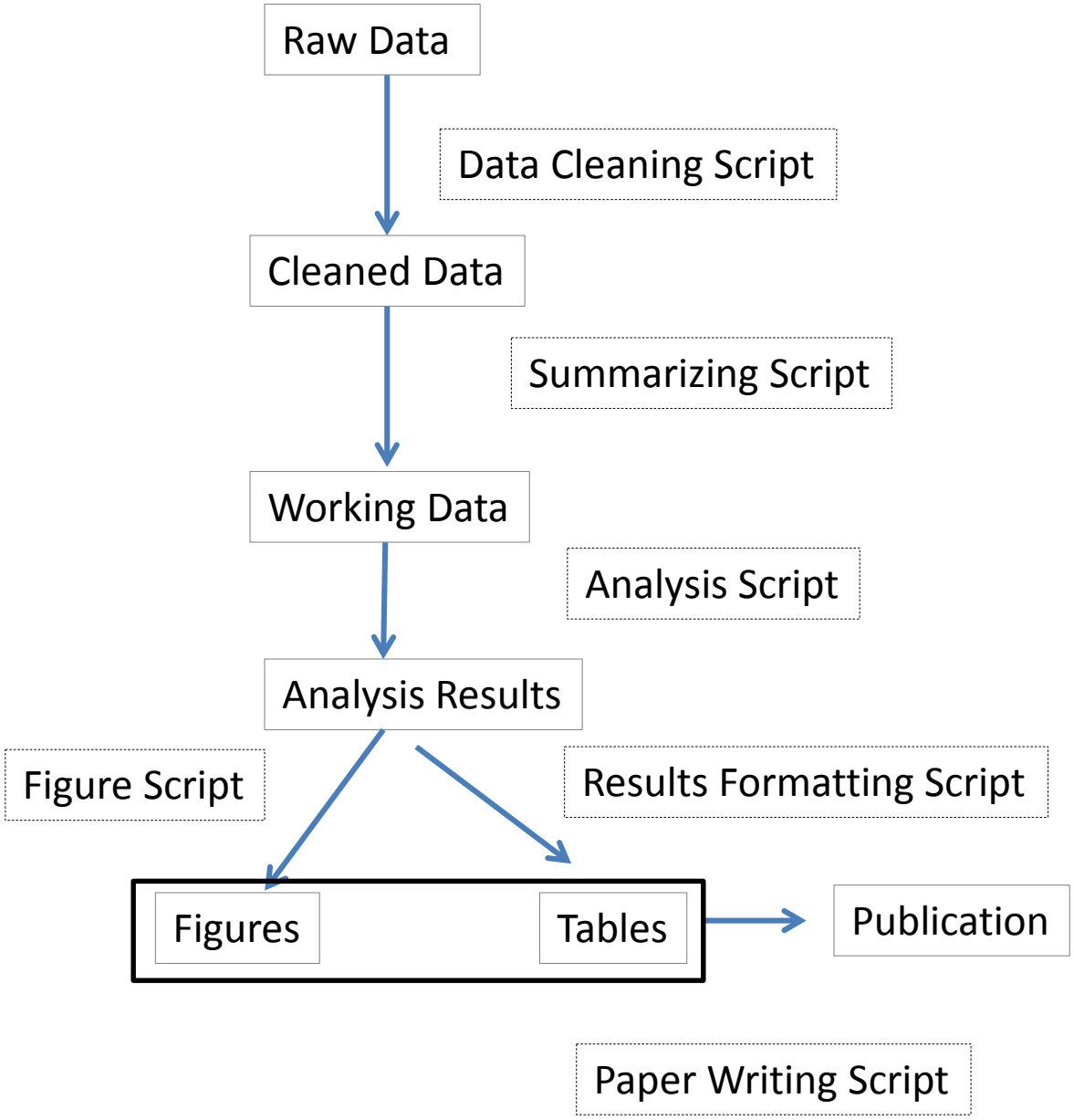


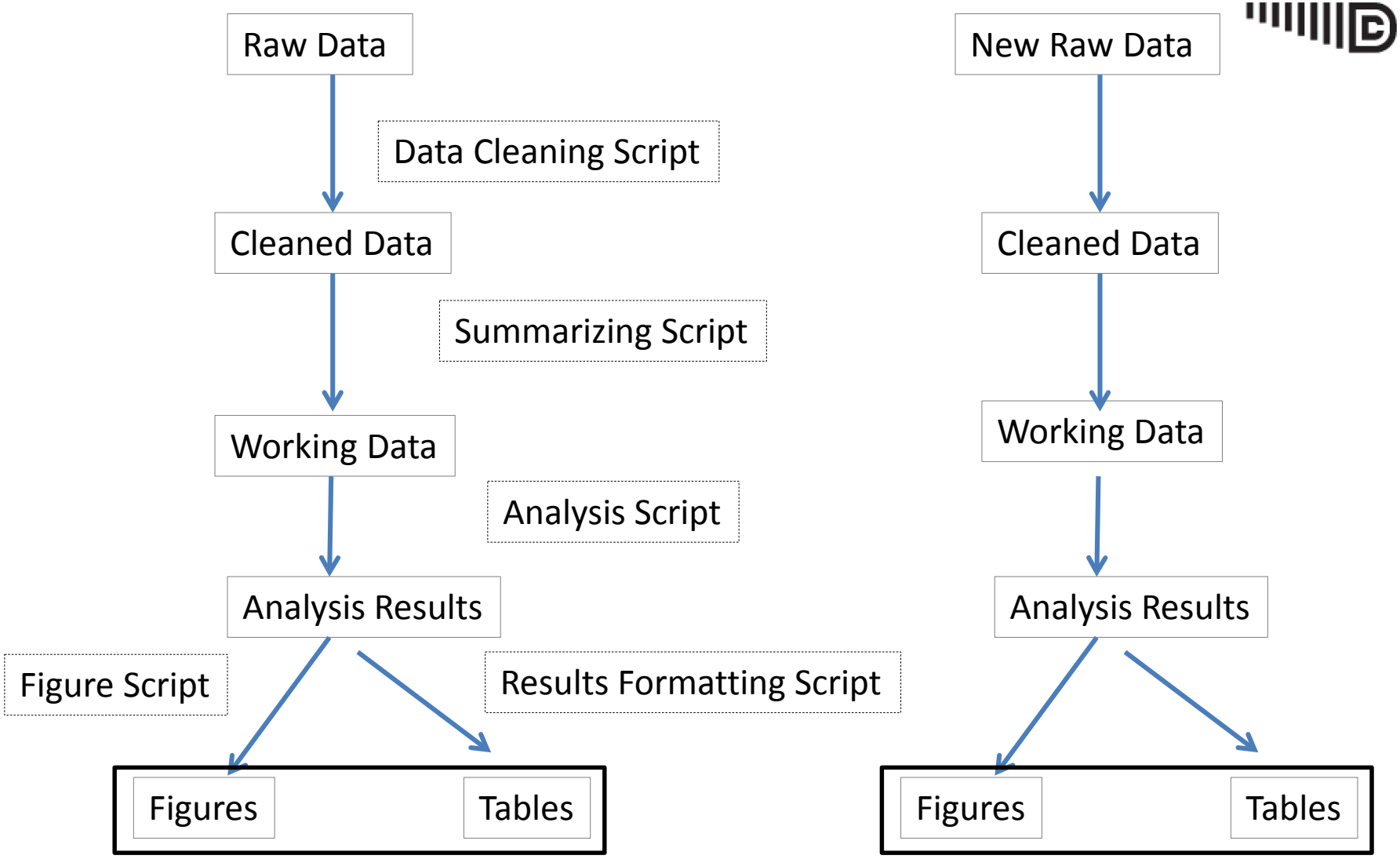
- Subset data for particular project
- Transform variables
- Average, min, max by group
- imputation



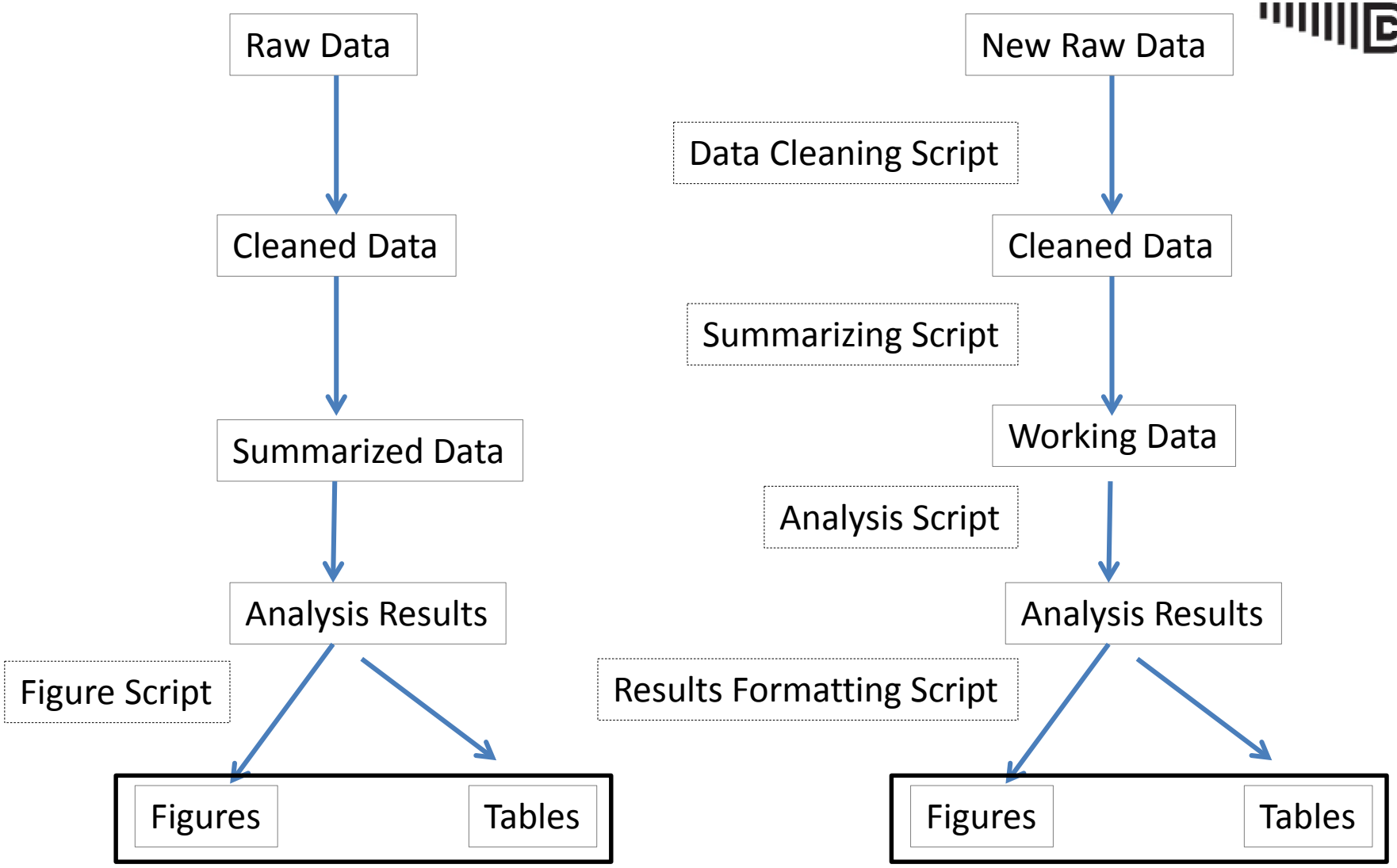
- Linear Models
- Mixed Models
- Search for Correlates
- Loop!
- General Functions

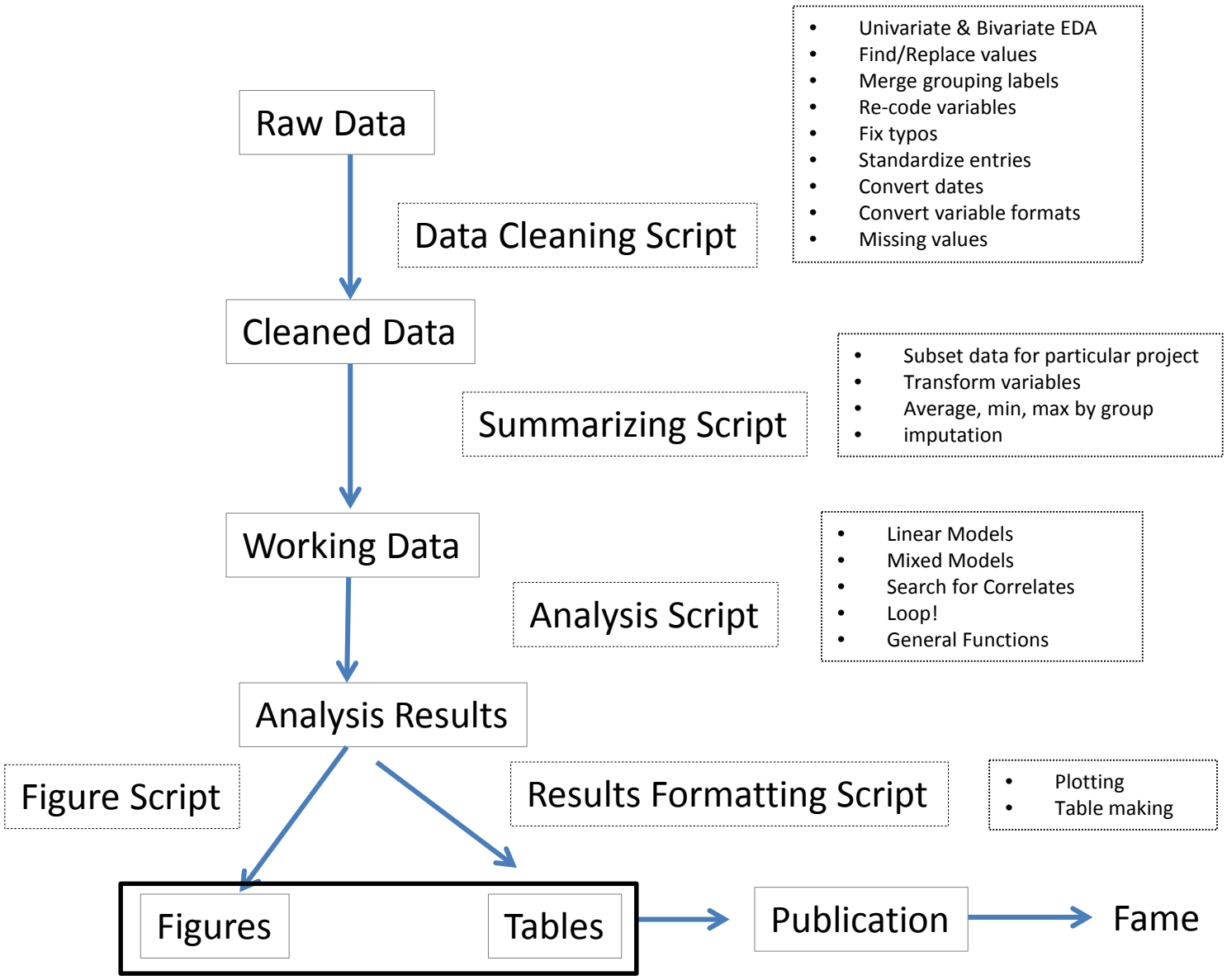

Raw Data
Data Cleaning Script
Cleaned Data
Summarizing Script
Working Data
Analysis Script
Analysis Results
Figure Script
Results Formatting Script
Plotting
Table making
Figures
Tables


Raw Data
Data Cleaning Script
Cleaned Data
Summarizing Script
Working Data
Analysis Script
Analysis Results
Figure Script
Results Formatting Script
Figures
Tables
Publication
Paper Writing Script


Raw Data
Data Cleaning Script
Cleaned Data
Summarizing Script
Working Data
Analysis Script
Analysis Results
Figure Script
Results Formatting Script
Figures
Tables
New Raw Data
Cleaned Data
Working Data
Analysis Results
Figures
Tables

Re-use and edit scripts for new projects




Raw Data
Data Cleaning Script
Univariate & Bivariate EDA
Find/Replace values
Merge grouping labels
Re-code variables
Fix typos
Standardize entries
Convert dates
Convert variable formats
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
Mixed Models
Search for Correlates
Loop!
General Functions
Analysis Results
Figure Script
Results Formatting Script
Plotting
Table making
Figures
Tables
Publication
Fame

Raw Data

Data Cleaning Script

- Univariate & Bivariate EDA
- Find/Replace values
- Merge grouping labels
- Re-code variables
- Fix typos
- Standardize entries
- Convert dates
- Convert variable formats
- Missing values

Summarizing Script

- Subset data for particular project
- Transform variables
- Average, min, max by group
- imputation

Analysis Script

- Linear Models
- Mixed Models
- Search for Correlates
- Loops!
- General Functions

Results Formatting Script

- Plotting
- Table making

Wednesday morning

Excel

OpenRefine

Wednesday Afternoon

R: ggplot

R: dplyr

Thursday Morning

R: loops & functions

R: Rmarkdown, knitr and reports

Thursday Afternoon

Python

DATA CARPENTRY