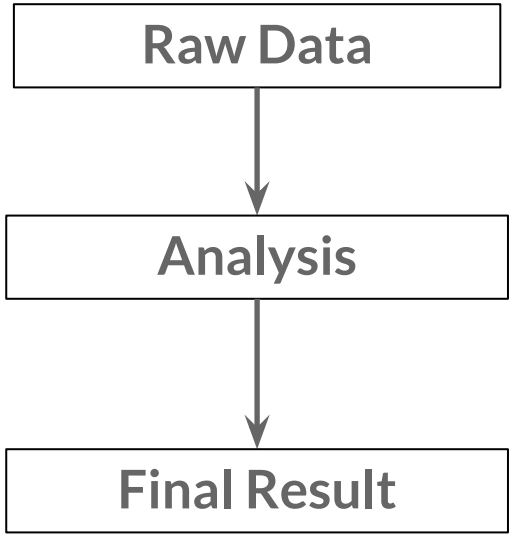
Best Practices for Data Science

The Data Science Process



Raw data gets cleaned and...

Tested	Visualized
Validated	Modeled
Transformed	*and much more*

Following certain guidelines will allow your future self as well as other researchers **to easily replicate and reproduce your project**

Data and Project Management

Basic Project Folder Organization

Admin

- read_me.txt
- task_tracker.xlsx

Inputs

- **Raw Data**
  - raw_data.csv
- **Cleaned Data**
  - cleaned_data.csv

Scripts and Programs

- 0. Import and Clean Data 5-20-2017.py
- 1. Analyze Data 5-20-2017.py
- 2. …
- **!Old Scripts**

Outputs

- results 5-20-2017.csv
- **!Old Results**

Basic Project Folder Organization

Admin
- read_me.txt
- task_tracker.xlsx

Inputs
- **Raw Data**
  - raw_data.csv
- **Cleaned Data**
  - cleaned_data.csv

Scripts and Programs
- 0. Import and Clean Data 5-20-2017.py
- 1. Analyze Data 5-20-2017.py
- 2. …
- **!Old Scripts**

Outputs
- results 5-20-2017.csv
- **!Old Results**

Have a place in your project where you briefly describe the project
Identify who in a team is the owner of a piece, as well as to-do items

Basic Project Folder Organization

Admin
- read_me.txt
- task_tracker.xlsx

Inputs
- **Raw Data**
  - raw_data.csv
- **Cleaned Data**
  - cleaned_data.csv

Scripts and Programs
- 0. Import and Clean Data 5-20-2017.py
- 1. Analyze Data 5-20-2017.py
- 2. …
- !Old Scripts

Outputs
- results 5-20-2017.csv
- !Old Results

Raw data should **never** be altered or overwritten

Instead, create a script that produces the tidy data to use in your analyses

Your tidy data should have:
- Rows as observations, columns as variables
  - A unique key for each observation
  - Intuitive variable names
- "NA" coded for all missing values

Only the absolute necessary transformations should be in your tidy data

Basic Project Folder Organization

Admin
- read_me.txt
- task_tracker.xlsx

Inputs
- **Raw Data**
  - raw_data.csv
- **Cleaned Data**
  - cleaned_data.csv

Scripts and Programs
- 0. Import and Clean Data 5-20-2017.py
- 1. Analyze Data 5-20-2017.py
- 2. …
- **!Old Scripts**

Outputs
- results 5-20-2017.csv
- **!Old Results**

Your scripts and programs should:
- Be named and ordered meaningfully
- Be well-commented (but not *too* commented)
- Implement version control (*e.g.*, by adding dates to filenames; Github) to track changes

Save old work in an "Old" folder in case it may be useful again

Basic Project Folder Organization

Admin
- read_me.txt
- task_tracker.xlsx

Inputs
- **Raw Data**
  - raw_data.csv
- **Cleaned Data**
  - cleaned_data.csv

Scripts and Programs
- 0. Import and Clean Data 5-20-2017.py
- 1. Analyze Data 5-20-2017.py
- 2. …
- **!Old Scripts**

Outputs
- results 5-20-2017.csv
- **!Old Results**

Similarly, your outputs should implement version control to allow you to track changes

Key Takeaway

(Save, organize, and document everything. Your future self and future researchers will thank you)

For More Best Practices

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal: "Good Enough Practices for Scientific Computing". http://github.com/swcarpentry/good-enough-practices-in-scientific-computing/, 2016.