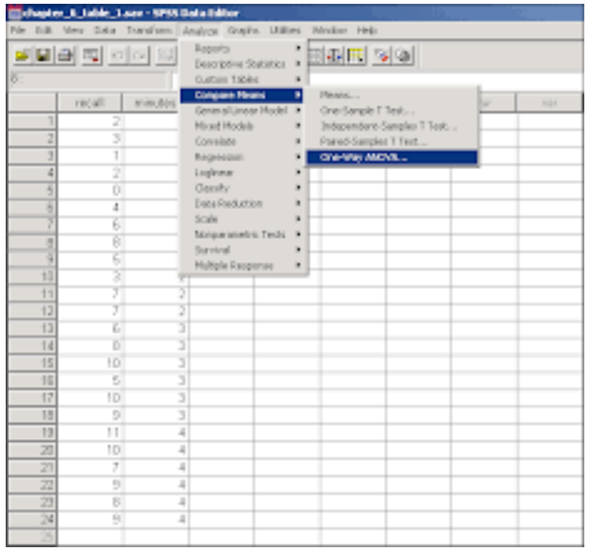
Leveraging Data-Driven Documents for Reproducible Projects

Jacob T.N. Young
jacob.young.1@asu.edu
https://jacobtnyoung.github.io/
Associate Professor - School of Criminology and Criminal Justice
Director of Curriculum - Institute for Social Science Research
Arizona State University

Are you one of these offenders?

The “point and clicker”?



The “cluttered folder person”?

The “task repeater”?

The “disordered desktop’r”?

It’s time to stop your offending!

Create Reproducible Projects!

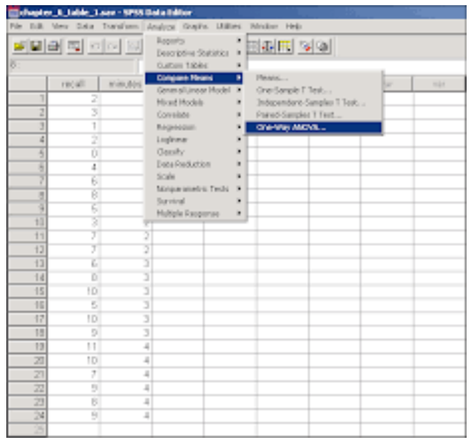
What is a “Reproducible Project”?

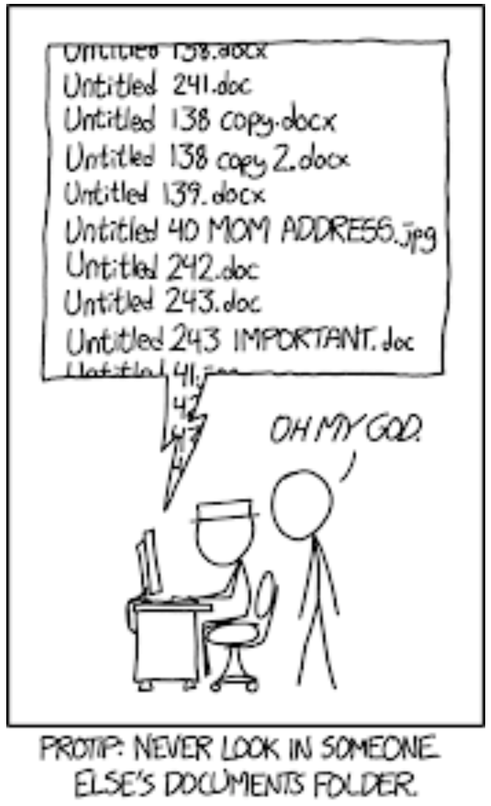
- Basically, a project you can “reproduce” (duh)
  - Using the **same** data and the **same** analysis, you get the **same** results (*not* the same as replication)

- Documented Workflow:
  - Reproducible projects thoroughly document your workflow, including data collection, data preprocessing, analysis, and visualization, in a clear and organized manner.

What does a reproducible project “look like”?

What a reproducible project *does not* "look like"?

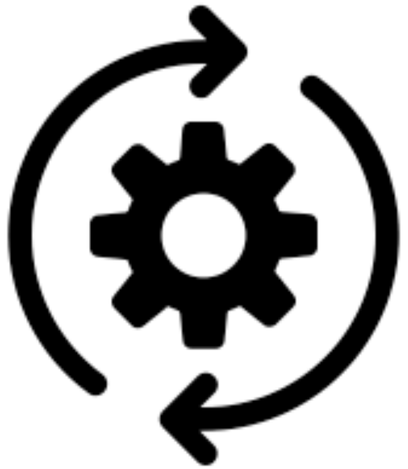




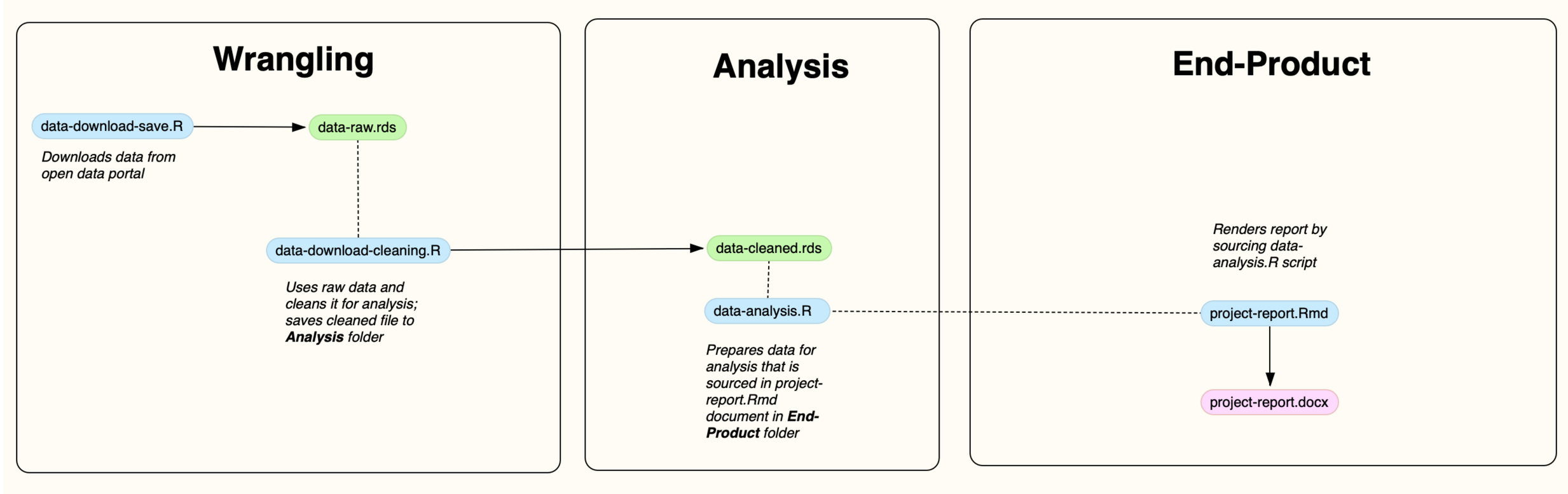
Raw data

county	year	county_ra	v001_raw	v001_nur	v001_dev	v001_clo	v001_chi	v001_race	v001_race	v001_ra
United States	2019		6900.63	3745538	9.08E+08	6890.349	8910.912			
Alabama	2019		9917.233	80440	13636816	9815.191	10019.27			
Autauga County	2019	1	8824.057	815	156132	7935.326	9712.788	10471.25		8706.6
Baldwin County	2019	1	7224.632	2827	576456	6794.128	7655.136	10042.47	3086.606	7277.7
Barbour County	2019	1	9586.165	451	72222	8200.118	10972.21	11332.56		7309.6
Bibb County	2019	1	11783.54	445	63653	10158.98	13407.11	14812.54		11327.
Blount County	2019	1	10908.1	1050	161107	9895.583	11920.62		5619.645	11336.
Bullock County	2019	1	12066.91	205	29066	9519.704	14614.12	14436.54		6638.6
Butler County	2019	1	14018.61	393	54804	11976.43	16060.78	15793.58		12797.
Calhoun County	2019	1	12217.76	2333	321406	11484.59	12950.93	12955.02		12447.
Chambers County	2019	1	11273.17	691	93770	9968.759	12597.58	10774.82		11794.
Cherokee County	2019	1	11294.41	575	71014	9767.023	12821.8	16385.25		10965.
Chilton County	2019	1	10831.92	850	123617	9733.842	11929.99	12000.44		11110.
Choctaw County	2019	1	12047.19	273	35489	9740.13	14354.24	12215.47		12163.
Clarke County	2019	1	9688.374	412	67037	8207.626	11189.12	12447.32		7394.5
Clay County	2019	1	9666.041	271	36916	7821.948	11510.13	12628.16		9302.0
Cleburne County	2019	1	11985.56	334	41389	9955.806	14015.5			
Coffee County	2019	1	7933.13	743	143859	7085.258	8781.002	10333.51	6033.742	7652.6
Colbert County	2019	1	10332.48	1016	149595	9321.505	11343.46	11604.94		10389
Conecuh County	2019	1	10635.37	272	34170	8546.834	12723.9	11100.61		10019.

All the stuff you do with the raw data

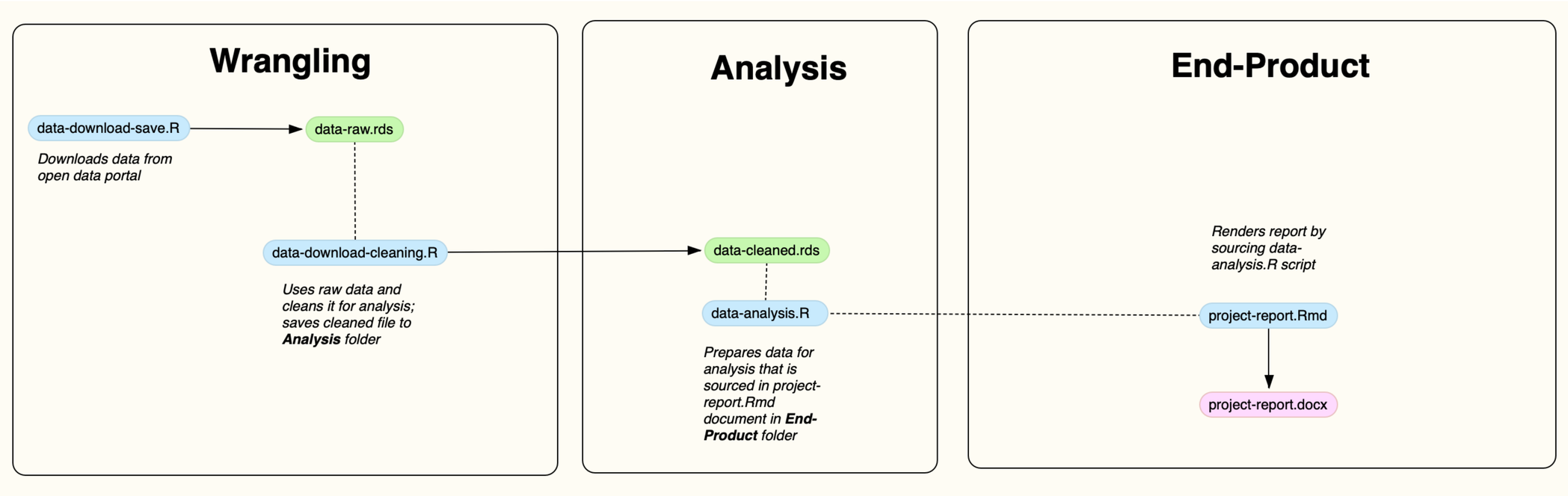


The output of the analysis

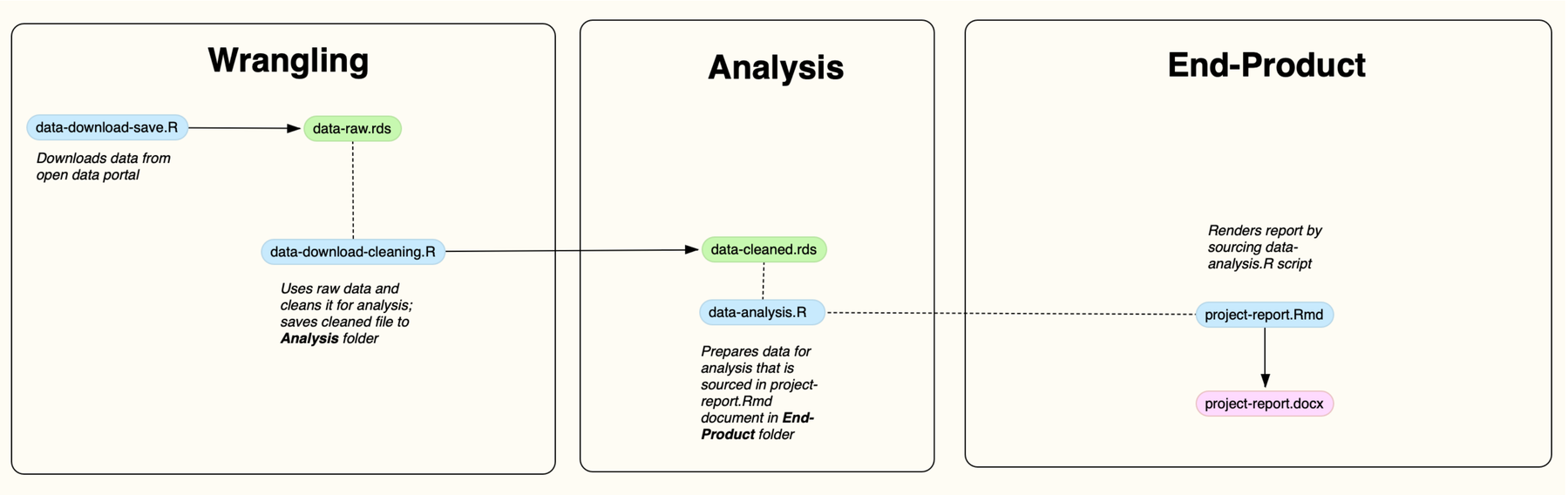

Wrangling
data-download-save.R
data-raw.rds
Downloads data from open data portal
data-download-cleaning.R
Uses raw data and cleans it for analysis; saves cleaned file to Analysis folder
Analysis
data-cleaned.rds
data-analysis.R
Prepares data for analysis that is sourced in project-report.Rmd document in End-Product folder
End-Product
Renders report by sourcing data-analysis.R script
project-report.Rmd
project-report.docx

Project Workflow Map

- What do maps do?
  - They show us where things are and how you get to those things.

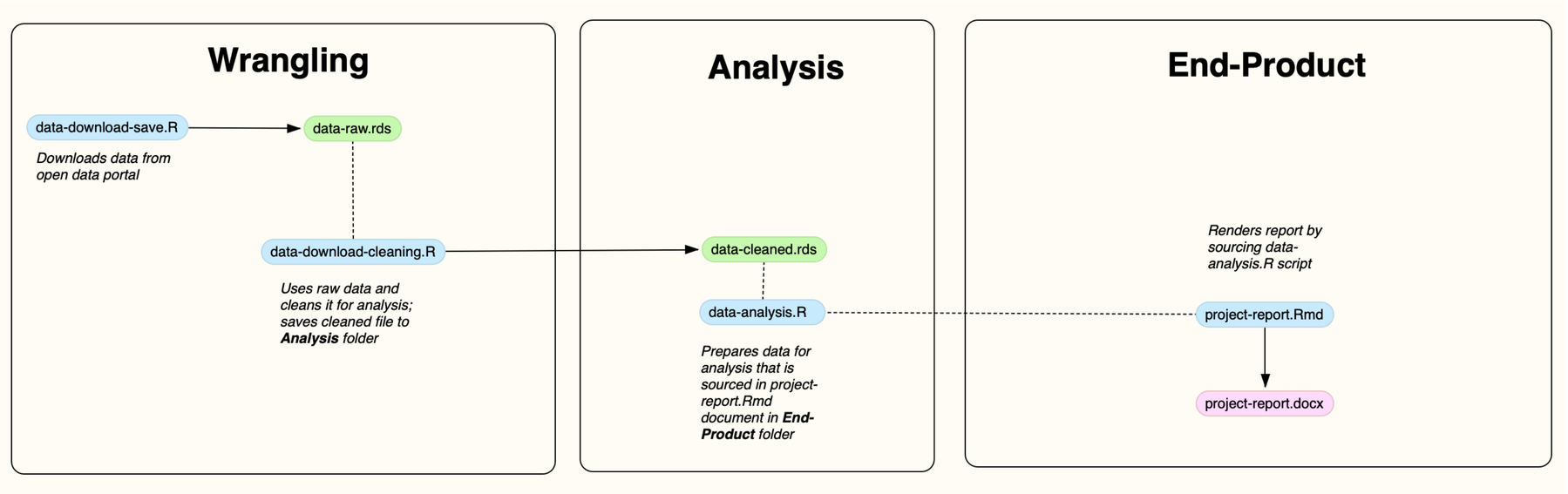


Project Workflow Map: Key Features



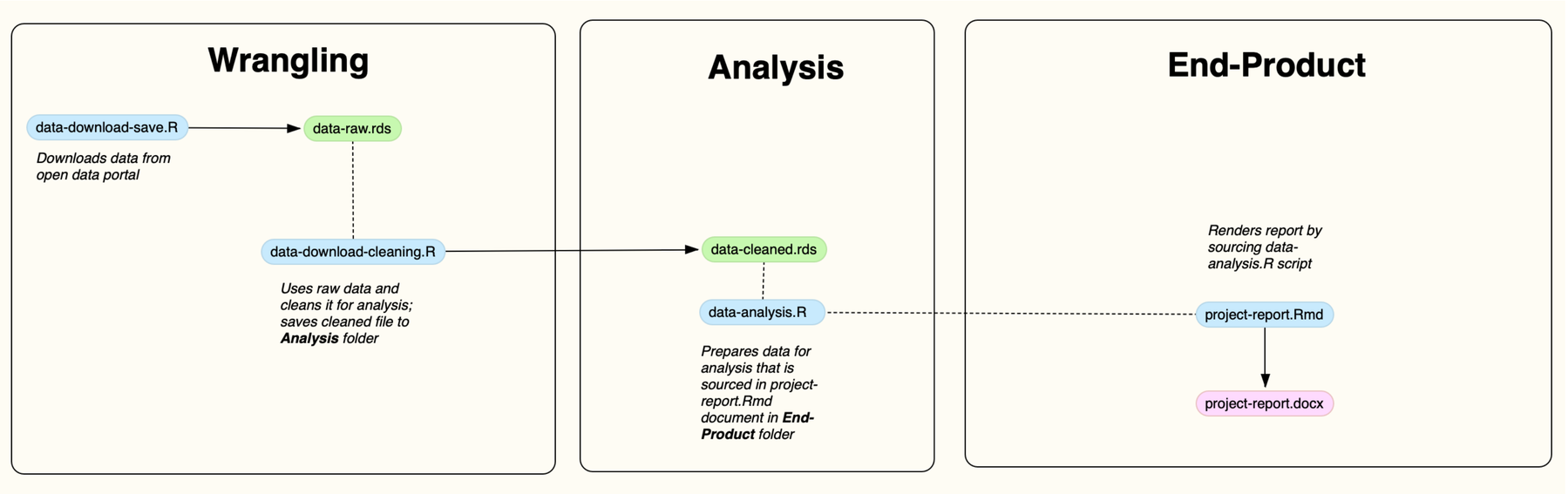
Project Workflow Map: Key Features

- **The three major segments of the project (these are the black boxes)**
- Instructions for generating outputs (the blue bubbles)
- The data outputs of the instructions (the green bubbles)
- The document outputs of the instructions (the purple bubbles)



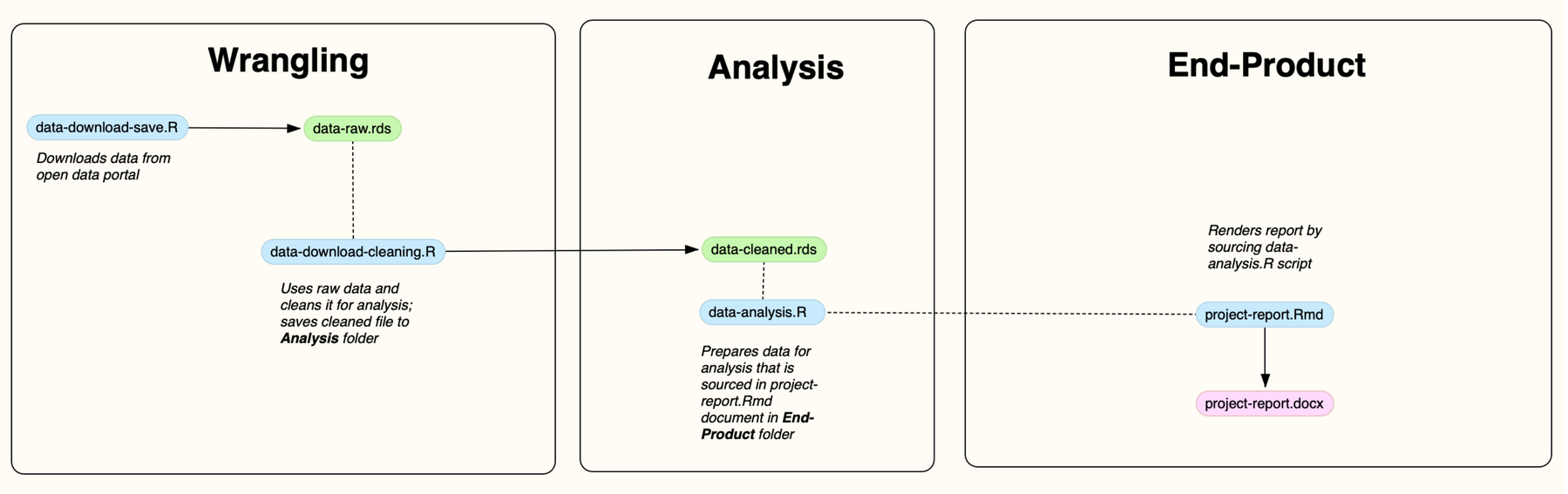
Project Workflow Map: Key Features

- The three major segments of the project (these are the black boxes)
- **Instructions for generating outputs (the blue bubbles)**
- The data outputs of the instructions (the green bubbles)
- The document outputs of the instructions (the purple bubbles)



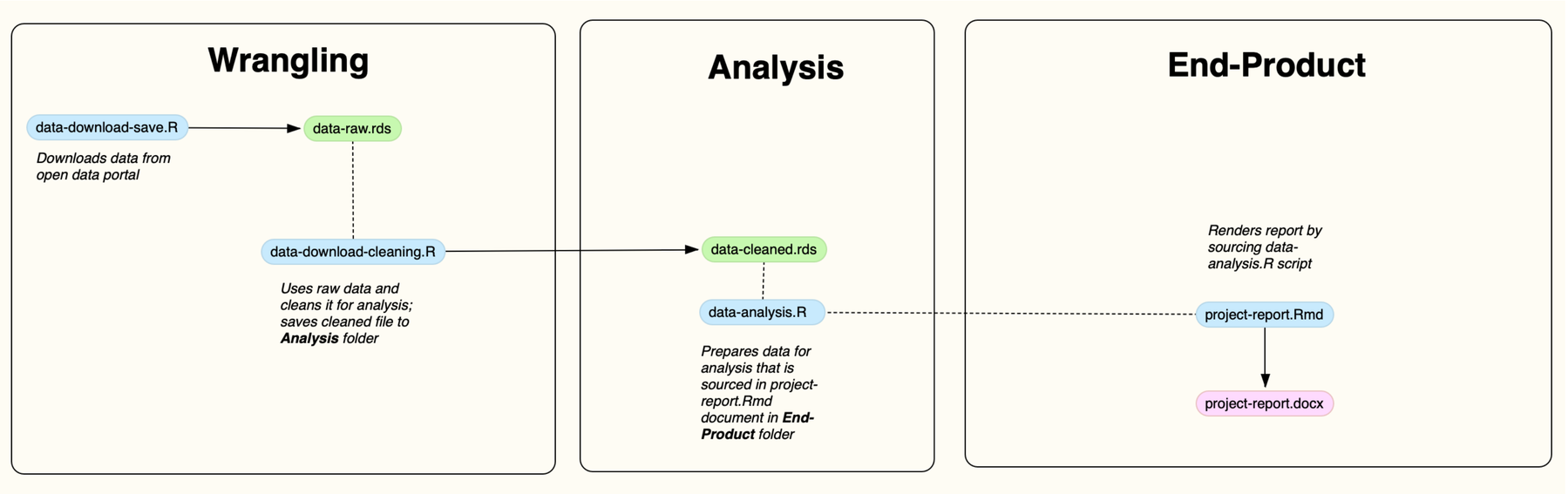
Project Workflow Map: Key Features

- The three major segments of the project (these are the black boxes)
- Instructions for generating outputs (the blue bubbles)
- **The data outputs of the instructions (the green bubbles)**
- The document outputs of the instructions (the purple bubbles)



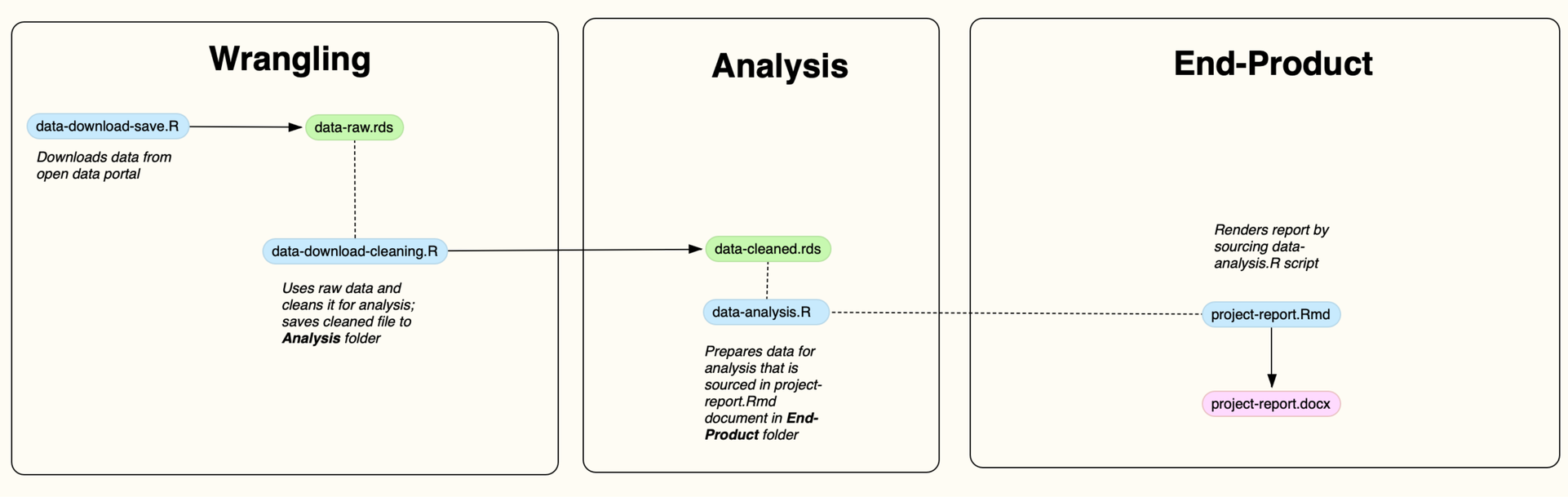
Project Workflow Map: Key Features

- The three major segments of the project (these are the black boxes)
- Instructions for generating outputs (the blue bubbles)
- The data outputs of the instructions (the green bubbles)
- **The document outputs of the instructions (the purple bubbles)**



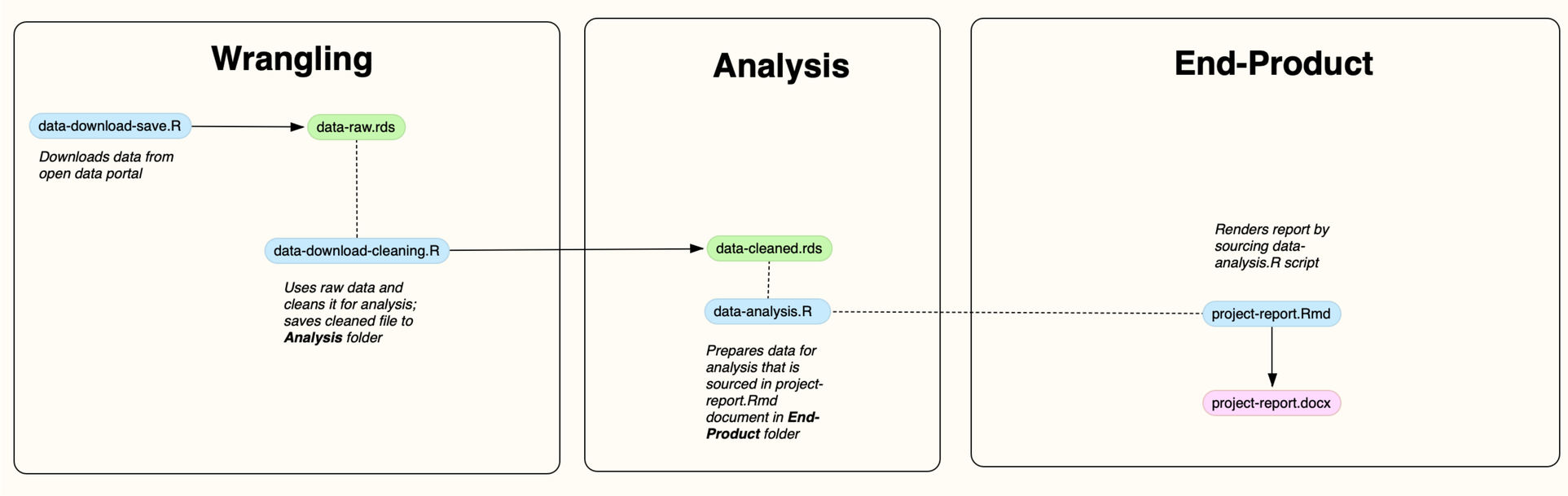
Project Workflow Map

- What is the **route** of this map?



Project Workflow Map

- What is the **route** of this map? The map shows us how to start with a raw data file and work all the way through to a project report and a project presentation.
    - All of the elements of reproducible research are shown in the figure.



Why?

- Learning Best Practices: teaches best practices in data management, code organization, and documentation.
- Transparency and Accountability: Promotes transparency by allowing others to scrutinize and replicate your work.
- Skill Development: Develop technical skills such as coding, data manipulation, and data visualization, which are transferable to various career paths.
- Efficiency: Streamlines the research process by reducing the time spent on troubleshooting and data management.

How?: Use Dynamic Documents!

- RMardown
  - A dynamic document format that combines the power of R programming language and Markdown syntax.
  - It allows you to integrate code, text, and visualizations into a single document. RMarkdown documents can be easily converted to various output formats, such as PDF, HTML, Word, and more.

- An example (with hyperlinks):
  - A manuscript
  - and the RMarkdown file to create the manuscript

Features of RMarkdown

- Code Integration: Embed R code chunks within your document for data analysis and visualization.
- Reproducibility: Ensures that your results are reproducible since readers can see the code and its output.
- Flexibility: Supports various output formats, making it suitable for academic papers, reports, presentations, and more.
- Easy to Learn: Requires minimal coding knowledge, making it accessible to a wide range of users.
- Collaboration: Facilitates collaborative work by combining code, text, and results in one file.
- Customization: Allows you to customize document styles and formatting.

Getting Started

- Download R: https://cran.rstudio.com/
- Download RStudio: https://posit.co/download/rstudio-desktop/

- Now, open RStudio and work through this RMardown tutorial

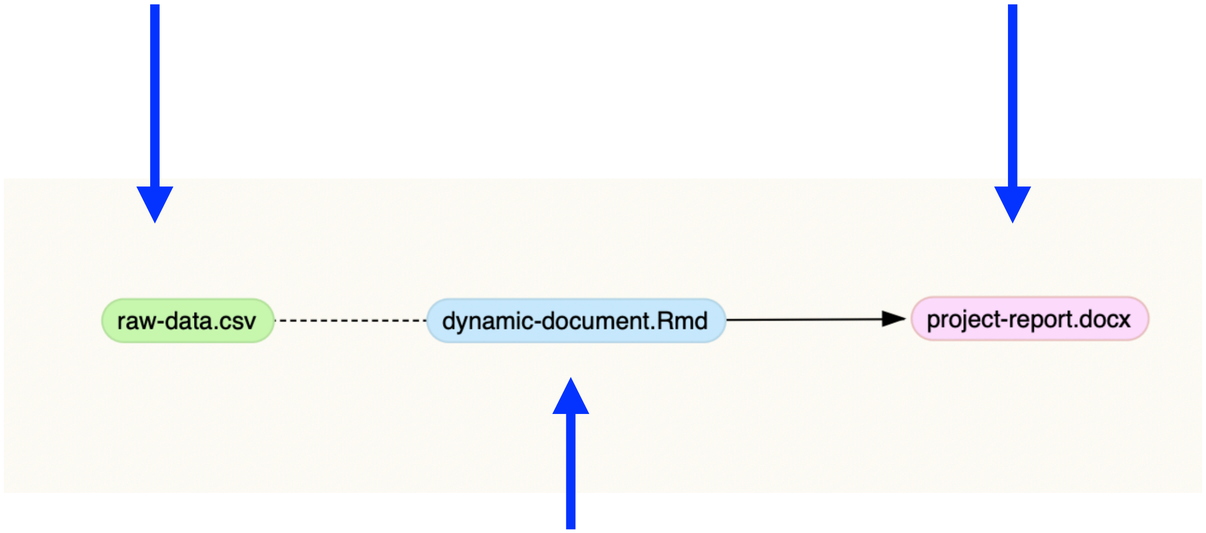
Exercise

- We want to do a simple example showing how we go from a raw data file to a project report.
  - First, think about what the steps are.
    - What needs to be done?
    - What are the stops on the road from raw data to final report?
  - Now, let’s start building that workflow!

Project Workflow Map

A raw data file

The project report as a Word document



Our RMarkdown file that will take the raw data and make the report

Let’s do it!

Exercise

- Suppose we received a new sheet of data, found an error in the original, or had some reason to rerun it...

Final Points

- Building workflows requires lots of planning.
  - But, they can be revised...THIS WILL HAPPEN!

- Teamwork makes the dream work!
  - Version control through Github

Thank you!!!

(*Please remember to submit feedback for the workshop!*)

Jacob T.N. Young
jacob.young.1@asu.edu
https://jacobtnyoung.github.io/
Associate Professor - School of Criminology and Criminal Justice
Director of Curriculum - Institute for Social Science Research
Arizona State University