

Introduction to the {targets} package

Reproducible Scientific Workflows in R

Ernest Guevarra

2024-02-05

Masons, when they start upon a building,
Are careful to test out the scaffolding;

Make sure that planks won't slip at busy points,
Secure all ladders, tighten bolted joints.

And yet all this comes down when the job's done
Showing off walls of sure and solid stone.

So if, my dear, there sometimes seem to be
Old bridges breaking between you and me

Never fear. We may let the scaffolds fall
Confident that we have built our wall.

- "Scaffolding" by Seamus Heaney, 1939-2013

Outline

- Concepts on scientific workflows
- The `{targets}` package
- Practical session

Concepts on scientific workflows

**Concept #1: Reproducibility, reproducibility,
reproducibility!**

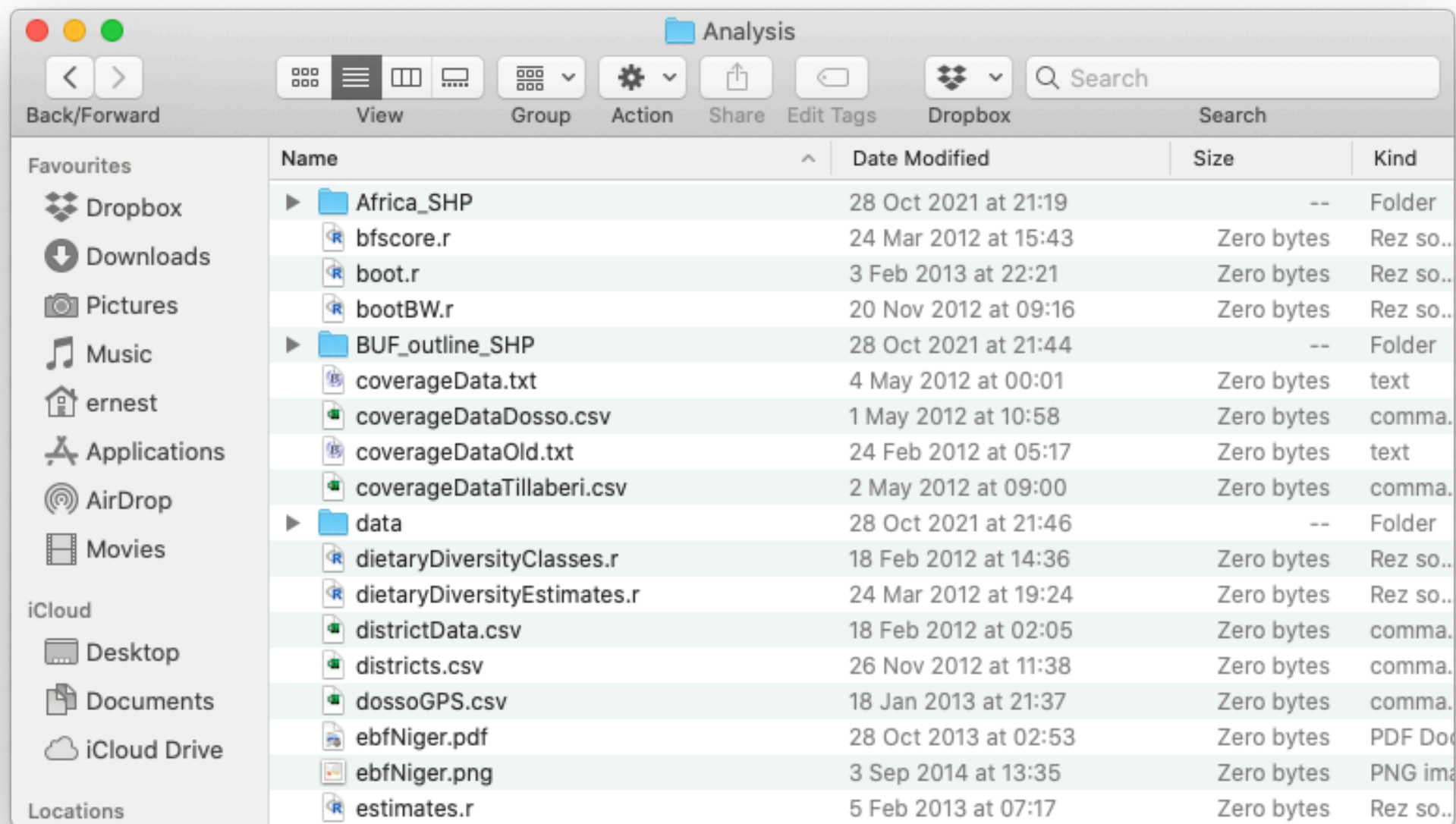
**Most important tool for Reproducible Research
is the mindset, when starting, taht the end
product will be reproducible.**

Keith Baggerly, via [@kwbroman](#) tweet

Concept #2: Organisation

File organization and naming are powerful weapons against chaos.

@JennyBryan



Concept #3: DRY - Don't repeat yourself

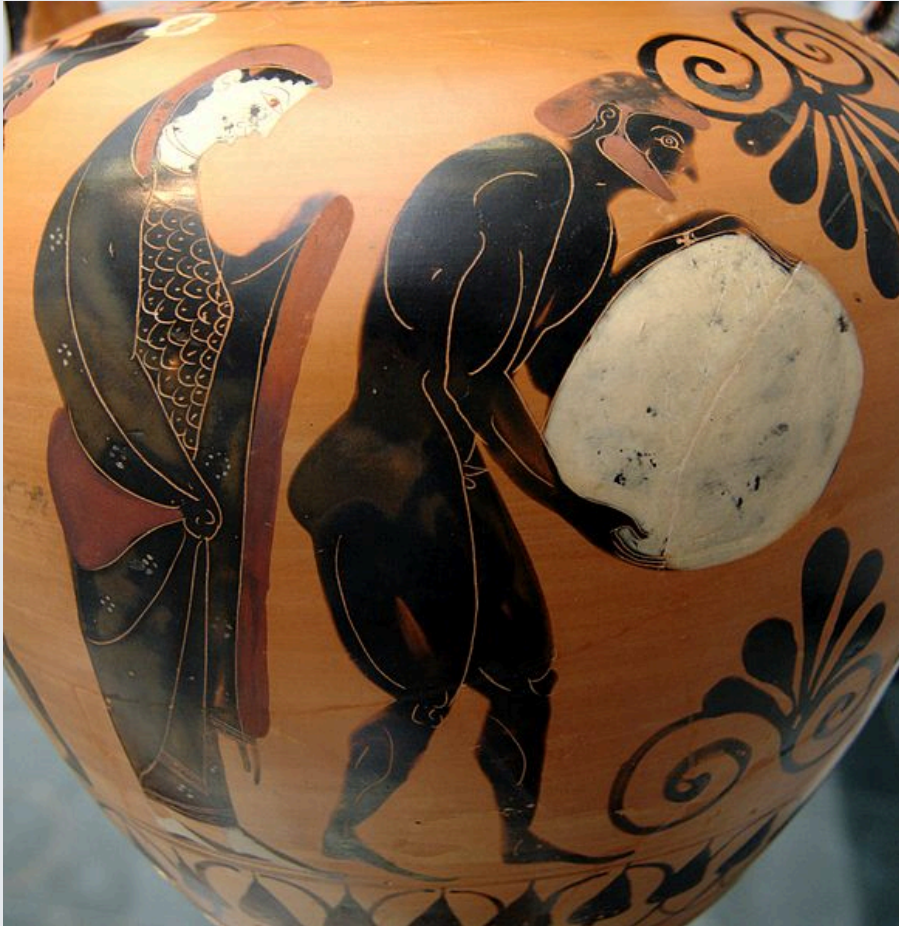
**Don't repeat yourself. It's not only repetitive,
it's redundant, and people have heard it before.**

Lemony Snicket

"You should consider writing a function whenever you've copied and pasted a block of code more than twice (i.e. you now have three copies of the same code)"

Challenges with scientific workflows

Sisyphean loop



1. Launch the code.
2. Wait while it runs.
3. Discover an issue.
4. Restart from scratch.

The {targets} package



- a pipeline toolkit for Statistics and data science in R
- maintain a reproducible workflow without repeating yourself
- learns how your workflow fits together
- skips costly runtime for tasks that are already up-to-date
- runs only the necessary computation
- supports implicit parallel computing
- abstracts files as R objects
- shows tangible evidence that the results match the underlying code and data

{targets} file organisation

ernestguevarra update README

File	Description	Commit Time
.git-crypt	Add 1 git-crypt collaborator	5 days ago
.github/workflows	rename workflows	2 days ago
R	complete data review and anthro review	8 hours ago
auth	encrypt and authenticate with drive	3 days ago
data	add .gitkeep	3 days ago
images	update README	12 minutes ago
outputs	Create .gitkeep	2 days ago
renv	Initial commit	5 days ago
reports	add notes to data review	7 hours ago
.Rprofile	encrypt and authenticate with drive	3 days ago
.env	encrypt and authenticate with drive	3 days ago
.gitattributes	encrypt and authenticate with drive	3 days ago
.gitignore	encrypt and authenticate with drive	3 days ago
README.Rmd	update README	12 minutes ago
README.md	update README	12 minutes ago
_targets.R	complete data review and anthro review	8 hours ago
mozambique-baseline-review.Rproj	encrypt and authenticate with drive	3 days ago
packages.R	complete data review and anthro review	8 hours ago
renv.lock	complete data review and anthro review	8 hours ago

README.md

Improving nutrition status for under 5 children in Zambezia and Nampula baseline survey data review workflow

repo status Active test review workflow passing run review workflow passing

This repository is a template for a `docker`-containerised, `{targets}`-based, `{renv}`-enabled `R` workflow for the baseline survey data review of the improving nutrition status for under 5 children in Zambezia and Nampula.

- this is a typical file structure with user-defined components of any project-oriented workflow
- the `_targets.R` file, however, is special and specific to a `{targets}` workflow - it is the target script file
- the target script file should be in the project's root directory

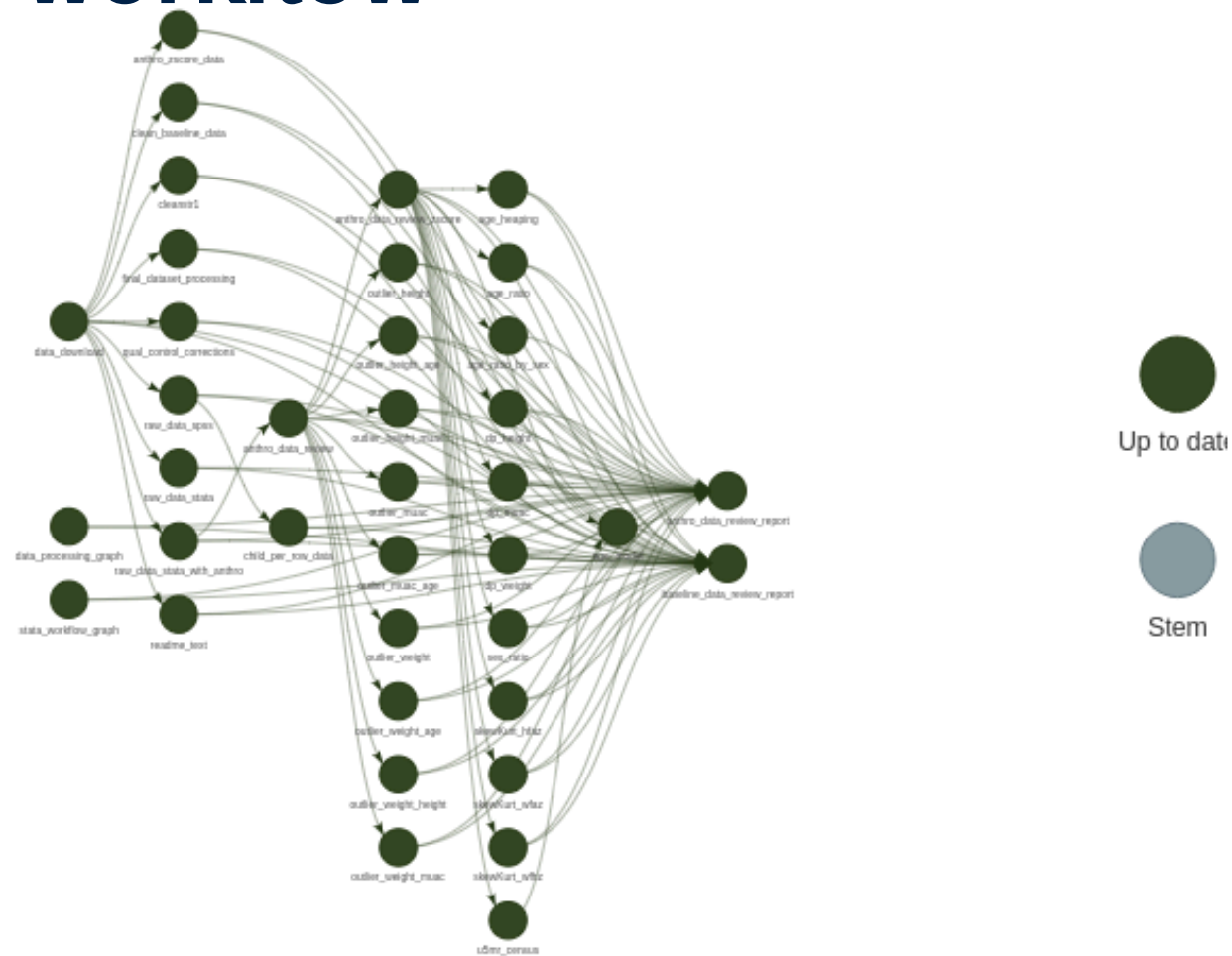
{targets} script file



```
1 #####
2 #
3 # Project build script
4 #
5 #####
6
7 # Load packages (in packages.R) and load project-specific functions in R folder
8 suppressPackageStartupMessages(source("packages.R"))
9 for (f in list.files(here::here("R"), full.names = TRUE)) source(f)
10
11
12 # Set build options -----
13
14
15
16 # Groups of targets -----
17
18 ## Read raw data
19 raw_data <- tar_plan(
20   data_download = download_all_baseline_data(overwrite = TRUE),
21   raw_data_spss = read_spss_data(file_list = data_download),
22   raw_data_stata = read_stata_data(
23     file_list = data_download,
24     filename = "survey_fin.dta"
25   ),
26   anthro_zscore_data = read_csv_data(
27     file_list = data_download, filename = "who_anthrofin_zscore.csv"
28   ),
29   raw_data_stata_with_anthro = read_stata_data(
30     file_list = data_download,
31     filename = "survey_plus_who_fin.dta"
32   ),
33   clean_baseline_data = read_csv_data(
34     file_list = data_download,
35     filename = "Zambia and Nampula baseline survey dataset.csv",
36     fileEncoding = "latin1"
37   )
38 )
39
40 ## Read associated text and code information
41 text_data <- tar_plan(
42   readme_text = read_text_data(file_list = data_download, filename = ".txt"),
43   cleanstr1 = read_text_data(
44     file_list = data_download, filename = "cleanstr1", widths = 150),
45   qual_control_corrections = read_text_data(
```

- Load the packages required (1)
- Load custom functions (1)
- Define individual targets - intermediate step of the workflow (2)
- End with a list of targets objects

{targets} workflow



Questions?

Practical session

We will all continue to go through Exercise #1 in the
Practical R for Epidemiologists book

Questions?

Thank you!

Slides can be viewed at <https://oxford-ihtm.io/open-reproducible-science/session8.html>

PDF version of slides can be downloaded at <https://oxford-ihtm.io/open-reproducible-science/pdf/session8-reproducible-scientific-workflows.pdf>

R scripts for slides available [here](#)