Reproducible scientific workflows in R

Introduction to the {targets} package

Ernest Guevarra

2022-02-14

Outline

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

Concepts on scientific workflows

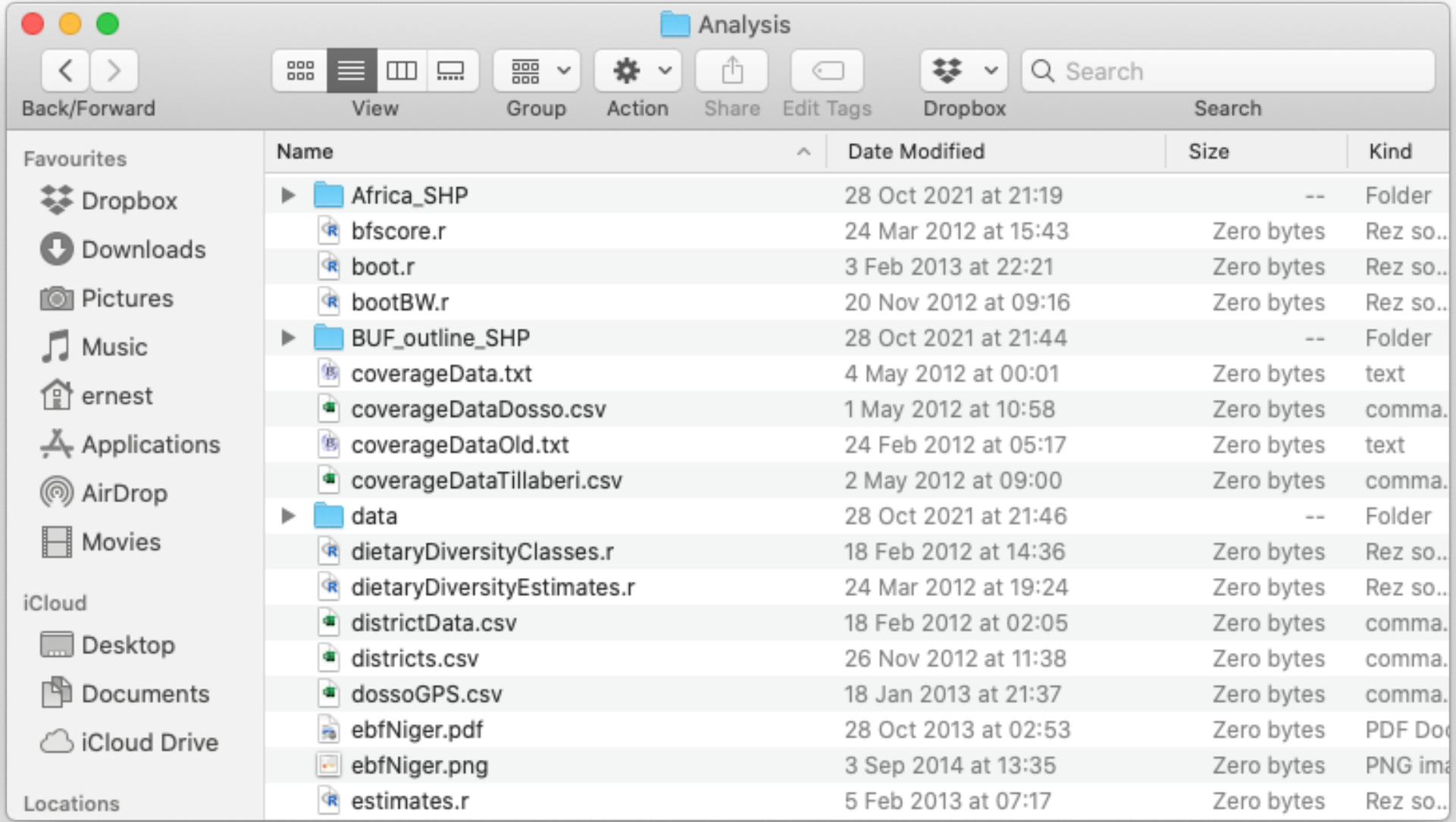
Concept #1: Reproducibility, reproducibility, reproducibility!

Keith Baggerly, via @kwbroman tweet

Concept #2: Organisation

File organization and naming are powerful weapons against chaos.

@JennyBryan


Analysis
Back/Forward
View
Group
Action
Share
Edit Tags
Dropbox
Search
Search
Favourites
Dropbox
Downloads
Pictures
Music
ernest
Applications
AirDrop
Movies
iCloud
Desktop
Documents
iCloud Drive
Locations
Name
Date Modified
Size
Kind
Africa_SHP 28 Oct 2021 at 21:19 -- Folder
bfscore.r 24 Mar 2012 at 15:43 Zero bytes Rez so..
boot.r 3 Feb 2013 at 22:21 Zero bytes Rez so..
bootBW.r 20 Nov 2012 at 09:16 Zero bytes Rez so..
BUF_outline_SHP 28 Oct 2021 at 21:44 -- Folder
coverageData.txt 4 May 2012 at 00:01 Zero bytes text
coverageDataDosso.csv 1 May 2012 at 10:58 Zero bytes comma.
coverageDataOld.txt 24 Feb 2012 at 05:17 Zero bytes text
coverageDataTillaberi.csv 2 May 2012 at 09:00 Zero bytes comma.
data 28 Oct 2021 at 21:46 -- Folder
dietaryDiversityClasses.r 18 Feb 2012 at 14:36 Zero bytes Rez so..
dietaryDiversityEstimates.r 24 Mar 2012 at 19:24 Zero bytes Rez so..
districtData.csv 18 Feb 2012 at 02:05 Zero bytes comma.
districts.csv 26 Nov 2012 at 11:38 Zero bytes comma.
dossoGPS.csv 18 Jan 2013 at 21:37 Zero bytes comma.
ebfNiger.pdf 28 Oct 2013 at 02:53 Zero bytes PDF Doc
ebfNiger.png 3 Sep 2014 at 13:35 Zero bytes PNG ima
estimates.r 5 Feb 2013 at 07:17 Zero bytes Rez so..

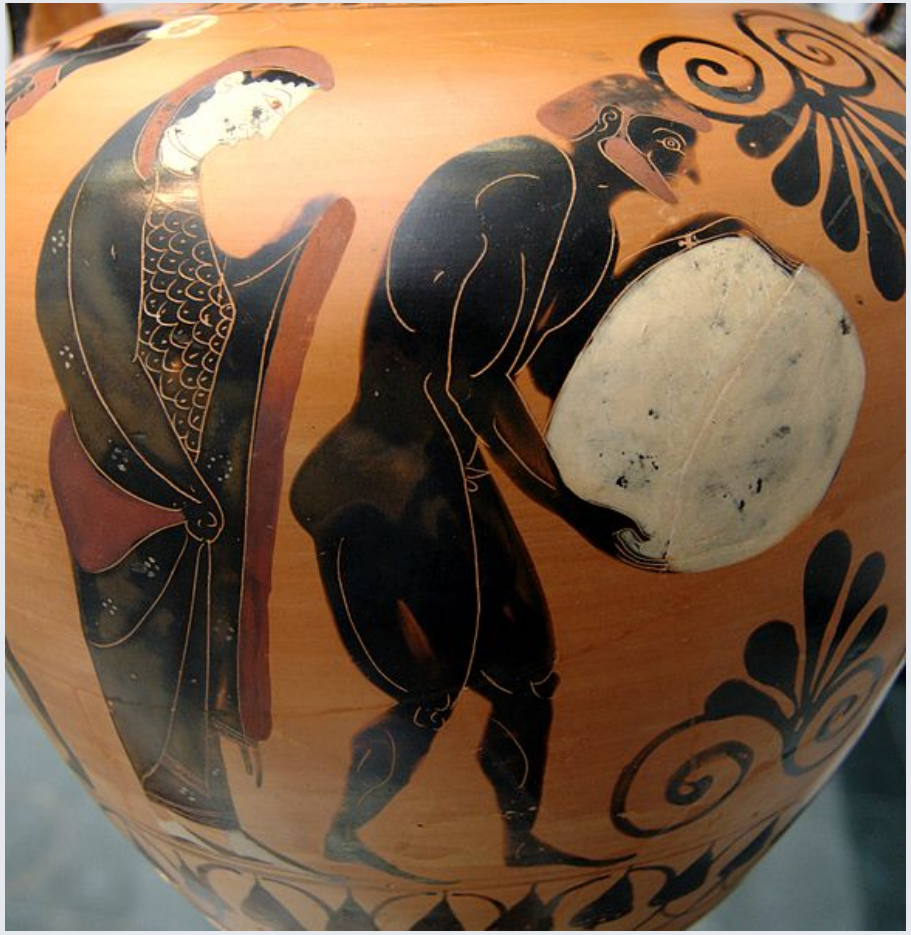
Concept #3: DRY - Don't repeat yourself

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

Lemony Snicket

```r
# Overlay maps of Niger and Nigeria to clean-up the map
par(new=TRUE)
plot(nigeria, axes = FALSE, xlim = mapXLimits, ylim = mapYLimits, border = "white", col = "white")

par(new = TRUE)
plot(boundaries, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.5, border = "black")

par(new = TRUE)
plot(n1, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.25, col = "blue")

par(new = TRUE)
plot(n4, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.25, col = "blue")

par(new = TRUE)
plot(n5, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.25, col = "blue")

par(new = TRUE)
plot(n6, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.25, col = "blue")

par(new = TRUE)
plot(n6.27, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.25, col = "blue")

par(new = TRUE)
plot(n7, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.25, col = "blue")

par(new = TRUE)
plot(n14, axes = FALSE, xlim = mapXLimits, ylim=mapYLimits, lwd = 0.25, col = "blue")
```

Sisyphean loop



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

Concept #3: Frequency reduces difficulty

@martinfowler, via @JennyBryan tweet

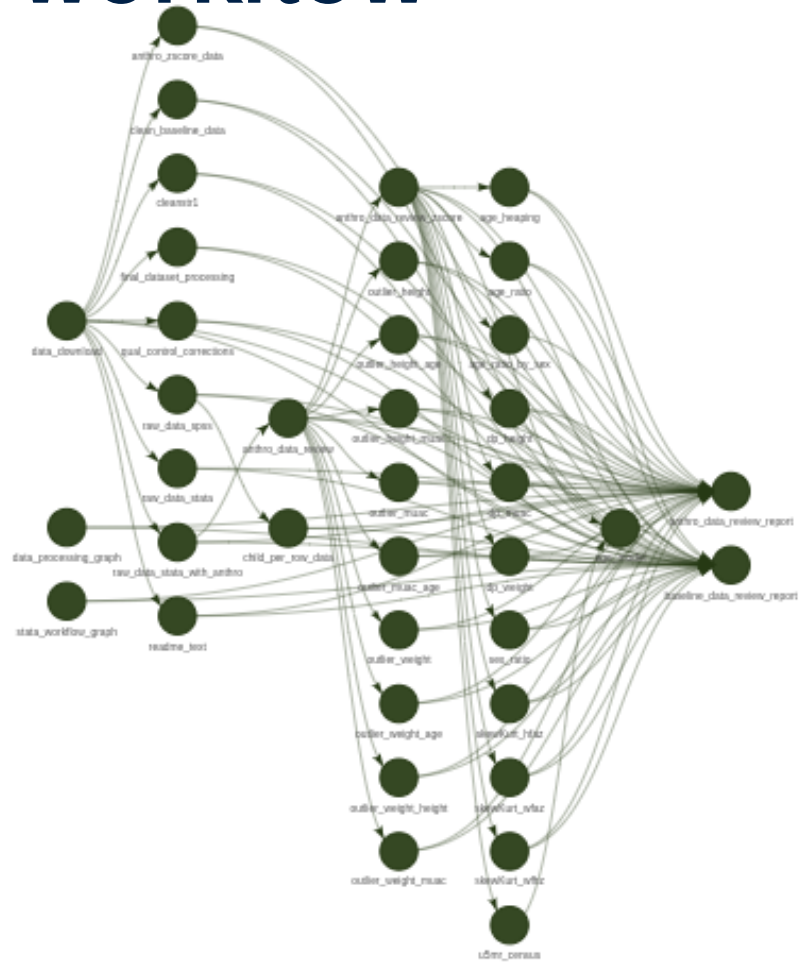
The {targets} package

- 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

`{targets}` script file

{targets} workflow



Questions?

Practical session

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://OxfordIHTM.github.io/open-reproducible-science/session4.html

PDF version of slides can be downloaded at https://OxfordIHTM.github.io/open-reproducible-science/pdf/session4-reproducible-scientific-workflows.pdf

R scripts for slides available here