# Research compendia as R packages

## Scientific workflows: Tools and Tips 🛠️

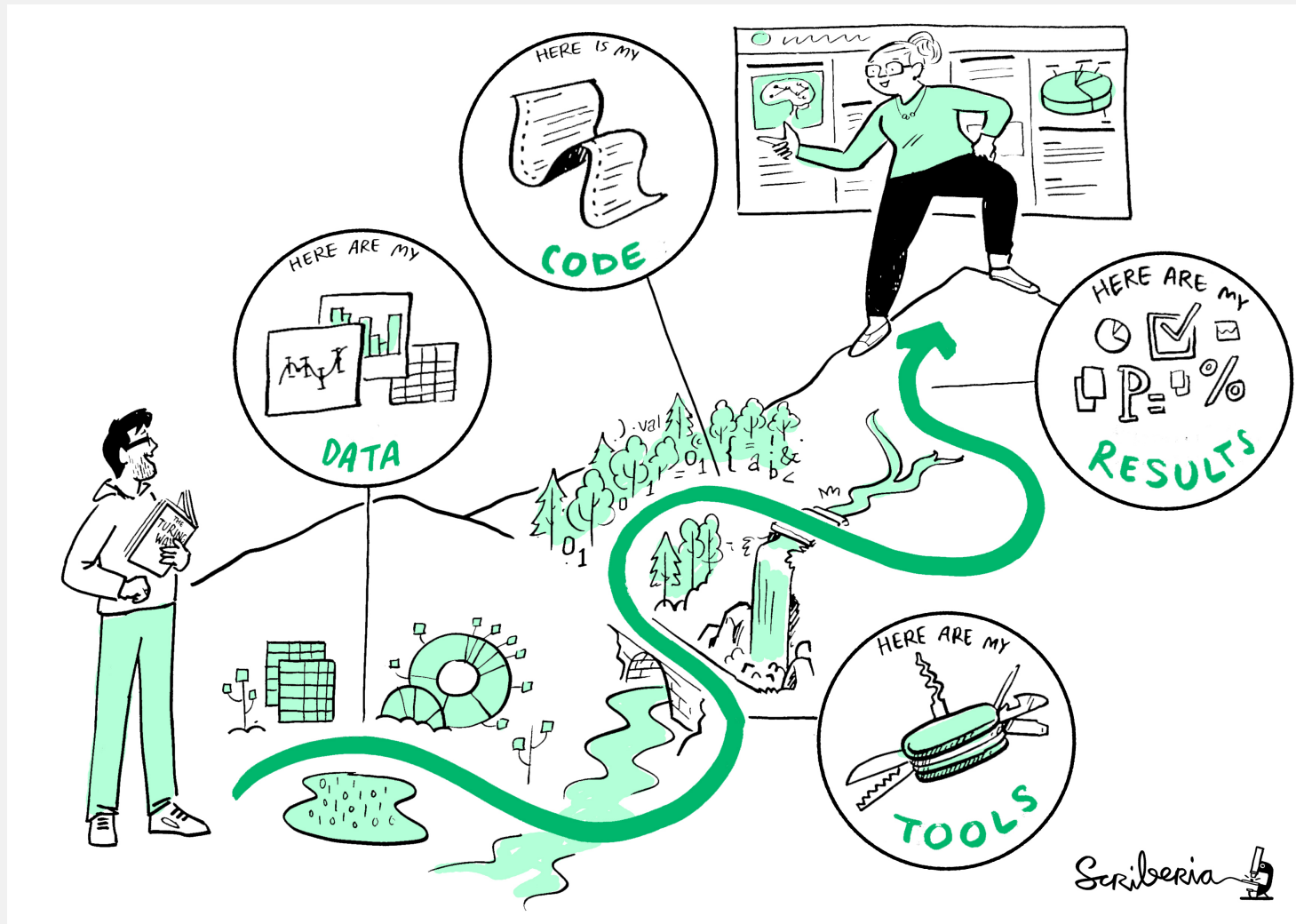2023-07-20

# What is this lecture series?

## Scientific workflows: Tools and Tips 🛠️

📅 Every 3rd Thursday 🕐 4-5 p.m. 📍 Webex

- One topic from the world of scientific workflows

- For topic suggestions send me an email

- If you don't want to miss a lecture

    - Check out the lecture website

    - Subscribe to the mailing list

- Slides provided on Github

# Steps of a scientific project

# Steps of a scientific project

How to properly structure the project?

I want

- 🔁 Reproduciblity (for you and others)
- 🏋️ Reliablity (will it work again?)
- ⚙ Re-usablity (don't re-invent the wheel)
- 🔍 Visiblity (let others see and use your work)

# How? Use a research compendium!

## What is a research compendium?

- Collection of all digital parts of a research project (data + code + text)

> The goal of a research compendium is to provide a **standard and easily recognizable** way for organizing the digital materials of a project to enable others to **inspect, reproduce, and extend** the research.

From Marwick et al. 2018

# Principles for building research compendia
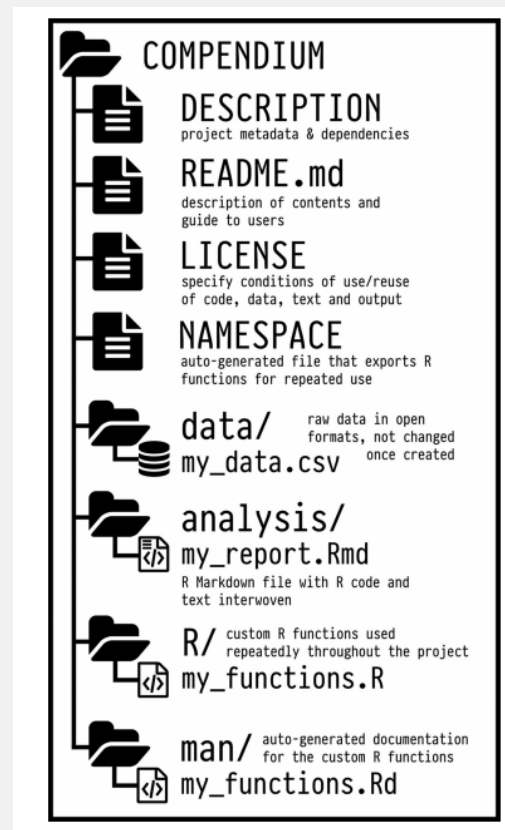
- Stick with the conventions in your field

- Keep **data**, **methods** and **output** separate

- Specify the computational environment

- Key components for sharing the compendium include

  - Licence

  - Version control

  - Metadata

  - Persistent identifier (e.g DOI)

# Examples of different complexities

## Small compendium
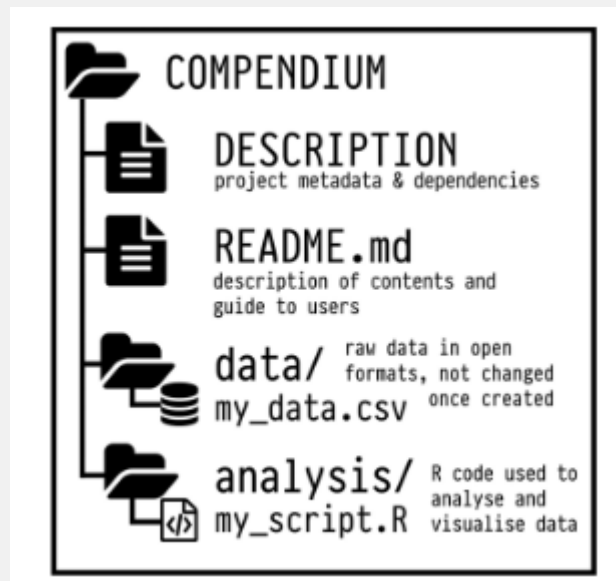


## Medium size compendium



They are R packages!

# R packages as research compendia



Packaging data analytical work reproducibly using R (and friends)

Ben Marwick
University of Washington, University of Wollongong
and
Carl Boettiger
University of California, Berkeley
and
Lincoln Mullen
George Mason University

March 20, 2018
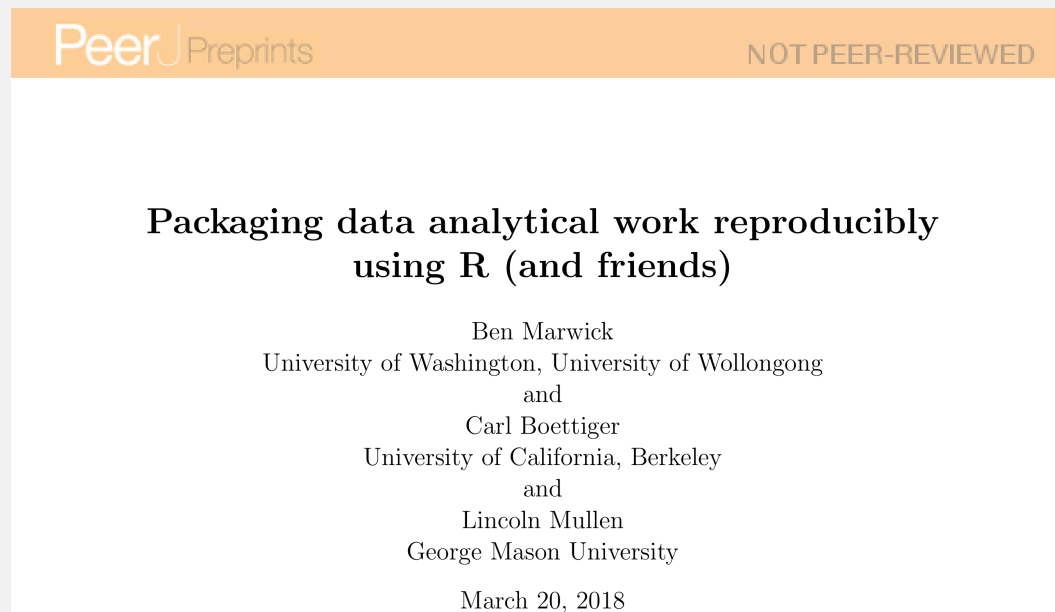
**Basic idea:** Hijack the R package development ecosystem to build a research compendium

Different use cases, e.g.

- Publish code, data and analysis scripts alongside your paper
- Publish a dataset in a way that other people can work with it in R

# Some benefits of R packages

- Benefit from quality control mechanisms built around R packages

- Additional packages around this ecosystem to make your life easier

- Easy documentation

- Easy sharing of data

# Hands-on: Create a research compendium with the R package structure

Find a detailed step-by-step guide on the website

# Summary and Conclusions

It's convenient to have standards you can follow

- R packages provide an helpful development ecosystem that we can hijack for our research compendia

- You can develop your compendium in different ways

  - A package that is meant to be installed with `install.packages()`

  - A compendium that won't be installed but uses quality checks from package ecosystem

- `usethis` is a great workflow package (not only for package development)

# Outlook

- Include large datasets with piggyback

- Automated tests with Github actions

- Unit tests to test your functions using testthat

- Easily connect Github repo to Zenodo to get a DOI

- holepunch package to build a Docker image

The nice thing:

- Also this easy to set up with `usethis` and friends

- They are also documented on the lecture series website

# Next lecture

Summer/Conference break in August and September!

Time for some feedback from you!

Please fill out the questionnaire if you have time (5 mins)

## Topic of next lecture t.b.a.

📅 19th October 🕐 4-5 p.m. 📍 Webex

🔔 Subscribe to the mailing list

✉️ For topic suggestions and/or feedback send me an email

# Thank you for your attention :)

Questions?

# References

- Marwick, Ben, Carl Boettiger, and Lincoln Mullen. 2018. "Packaging Data Analytical Work Reproducibly Using R (and Friends)." American Statistician 72 (1): 80–88. https://doi.org/10.1080/00031305.2017.1375986.

- The Turing Way website is a very useful guide to reproducible data science

- Slides and list of resources for research compendia by Karthik