Open source & data sharing

Or . . . how to stop worrying and leverage your research

Thomas de Graaff

June 23, 2020

Department of Spatial Economics

Open source—free as in beer, not in time!

How to draw an Owl.

"A fun and creative guide for beginners"

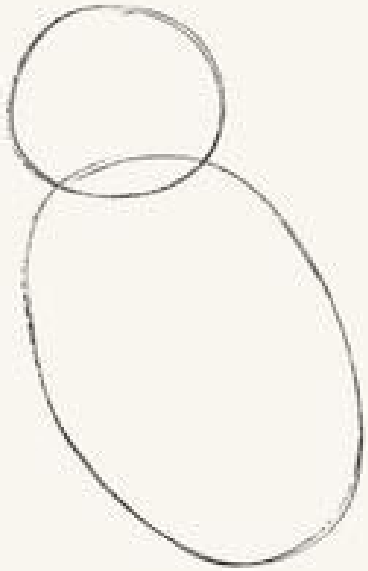

Fig 1. Draw two circles

Fig 2. Draw the rest of the damn Owl

What I advocate

To make available online and for anyone to access, assess, reuse & contribute to:

- All your data. And, if not possible
  - metadata
  - subset
  - simulated/scrambled data
- All text files, including scripts, figures, working paper

What I advocate

To make available online and for anyone to access, assess, reuse & contribute to:

- All your data. And, if not possible
  - metadata
  - subset
  - simulated/scrambled data
- All text files, including scripts, figures, working paper

So, why 'give' it away?

- The greater good: transparency & reproducibility
  - avoid "additional results available upon request"
- Private gains:
  - higher visibility (others will actually use your research)
  - incentives on working 'tidy'

How can I do that?

- Using a versioning system and an online platform
  - To share, distribute and cooperate on data
- Currently, Git and GitHub is the default (but Bitbucket, etc.)
- Complete version history with time stamps

How can I do that?

- Using a versioning system and an online platform
  - To share, distribute and cooperate on data
- Currently, Git and GitHub is the default (but Bitbucket, etc.)
- Complete version history with time stamps
- You work on a piece (data) and using Git commit to your own local repository
- Then you push to an open online repository (GitHub)
- Which now serves as your (ugly) website as well
  - Actually people (well, me) blog using GitHub as a free server using Jekyll or Hugo (blogdown)

Applications for Git/GitHub

- GUI applications
  - GitHub desktop, GitKraken, ScourgeTree
- Built-in most proper editors:
  - Sublime, Emacs, Overleaf, RStudio, etc.
- Command line (most powerful, but . . .)

Package it!

Instead of just providing the data you can wrap it in an (R-)package

- allows to give examples of the code and data
- easier loading of the data
- distribution again via GitHub or CRAN

Many R-package contain (access) to economic data

- `AER` package—all textbook data (California test score data...)
- `IMFData`
- `wbstats`—World bank data