

# Introduction

Daniel Arribas-Bel & Thomas de Graaff

September 5, 2014

# Introduction

# Why this workshop?

- ▶ In the *social sciences* few attention to what tools to use (and why they make sense)
- ▶ Increasing *need* for/in openness & transparency
  - ▶ from journals, universities and governments
  - ▶ increase in cooperation (over wider distances)
  - ▶ access to your own files
  - ▶ make yourself more visible
- ▶ Why we want to give this workshop
  - ▶ intrinsic interest
  - ▶ our goal: pre-conferences workshops / courses

# What we want (and don't want) with this workshop

- ▶ We are mostly interested in the principles behind a good open (scientific) workflow, aware of the facts that
  - ▶ there is no final, optimal, set of workflow tools
  - ▶ investment is very, very costly
- ▶ However, being a practical workshop we do
  - ▶ work with a specific set of tools (markdown, R, RStudio, git) which
  - ▶ enables us *in this workshop* to make a paper reproducible and open

## Related work

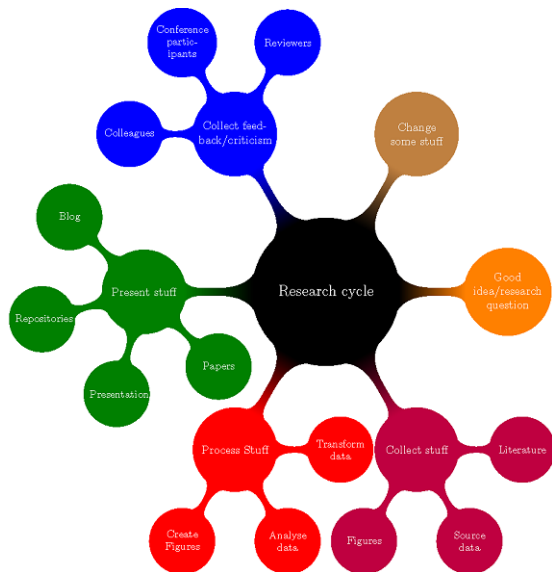
- ▶ Inspired by Kieran Healey's (associate professor in sociology) work: Choosing your Workflow Applications
- ▶ Courses for reproducible research seems to pop up everywhere (but mostly in datascience courses):
  - ▶ Datascience course: <https://www.coursera.org/>
  - ▶ Tools for Reproducible Research  
<http://kbroman.org/Tools4RR/>

# Workflow

# Open?

- ▶ Workflow: *Progression of steps (tasks, events, interactions) that comprise a work process, involve two or more persons, and create or add value to the organization's activities* (BusinessDictionary)
- ▶ Open workflow: One that enhances *transparency, collaboration* and *reproducibility*

# Research cycle





# Why bother about a workflow or tools?

- ▶ Good scientific practice: *document how you have achieved your results*; this ensures
  - ▶ Reproducibility
  - ▶ Transparency
  - ▶ Modularity
  - ▶ Portability (across systems and users)
  - ▶ Efficiency
  - ▶ Self-sanity

# Why should it be open?

- ▶ Open Science
- ▶ Reproducibility
- ▶ Transparency
- ▶ Modularity
- ▶ Portability (across systems and users)
- ▶ Efficiency
- ▶ Visibility

# When should I adopt an open reproducible workflow?

- ▶ The sooner the better
- ▶ But think twice about which one (switching is costly)
- ▶ Start one step at a time

*A journey of a thousand miles begins with a single step*

Lao-tzu

# Reproducibility

# In general

*In science consensus is irrelevant. What is relevant is reproducible results. The greatest scientists in history are great precisely because they broke with the consensus (Michael Crichton)*

# In computation science:

*The data and code used to make a finding are available and they are sufficient for an independent researcher to recreate the finding (Peng, 2011)*

- ▶ Literature programming (Donald E. Knuth, 1984):
  - ▶ weaving of **code**, **documentation** and **output** (articles, presentations, websites)

# In the social sciences?

- ▶ Complete reproducibility often not feasible
  - ▶ qualitative research
  - ▶ proprietary data (?)
- ▶ but you can come a long way, especially with
  - ▶ theoretical work
  - ▶ quantitative (e.g., statistical or simulation) work
- ▶ Goal should be more to make your research as reproducible as *possible*

# Code, documentation and output

1. Synonyms
2. All based on text files
3. Encompasses almost anything
  - ▶ data itself
  - ▶ set of commands for data cleaning and statistical analysis
  - ▶ database with references
  - ▶ transcript of interviews
  - ▶ text for articles, presentations or websites
4. Only output is displayed/interpreted differently (e.g., in a browser or pdf viewer)



# Our goal (not being ambitious)

What we want is that with *one single* command we

- ▶ read in and transform our data
- ▶ run the analysis
- ▶ create output (tables and figures)
- ▶ combine output with text and references
- ▶ create presentation material (paper, slides, webpages) and
- ▶ publish presentation material on an open repository

This all under a full fledged versioning control system

# Tools for reproducibility

- ▶ Markup languages
  - ▶ Markdown
  - ▶ LaTeX
  - ▶ HTML
- ▶ Terminal tools (GNU make, diff, pandoc)
- ▶ Versioning system (Git & VCN)
- ▶ Reference manager (bibdesk/Mendeley)

# Tools for reproducibility (cnt.)

- ▶ Statistical software (pure command line driven): Python and R
- ▶ Environments
  - ▶ R and RStudio environment
  - ▶ Python and iPython notebook environment
  - ▶ Python and Sumatra
  - ▶ Emacs org mode

# Tools for openness

- ▶ Repositories:
  - ▶ Github (host webpages as well)
  - ▶ Bitbucket
- ▶ R packages <http://cran.r-project.org/>
- ▶ iPython notebook viewer <http://nbviewer.ipython.org/>

# Examples

## *Reproducible Research with R and RStudio* Book1

- ▶ `https://github.com/christophergandrud/Rep-Res-Book`

Amsterdam paper example using ipython notebook:

- ▶ `http://darribas.org/buzz_adam`

# What we use in this workshop

1. R and RStudio (with Yihui Xie's `knitr` package)
2. Markdown language
3. Bibdesk/Mendeley
4. Git and Github
5. GNU make

Only implicitly we make use of LaTeX, BibTeX, HTML and pandoc  
(all under the hood of RStudio)

# Schedule

# Schedule Day 1 - Friday Sept. 5th

## ▶ **[9am-12am]** Introduction

- ▶ Concepts behind open workflows/Overview of tools
- ▶ Install session

[Lunch]

## ▶ **[1pm-3pm]** Version control and task automation

- ▶ Terminal/git/make

[Break]

## ▶ **[3:30pm-5:30pm]** Typesetting

- ▶ Markdown/LaTeX/bibtex/pandoc/RStudio

[Diner]

- ▶ Location and time: To be announced



## Schedule Day 2 - Saturday Sept. 6th.

- ▶ **[9am-11am]** Data analysis

- ▶ R

[break]

- ▶ **[11:30am-1pm]** Publishing

- ▶ Slides
- ▶ Publishing on GitHub
- ▶ Other publication channels

[Lunch]

In conclusion

## Loose ends...

- ▶ Questions?

This workshop is financially supported by FOSTER.

