ERSA-WooW: Introduction

Thomas de Graaff

August 23, 2016

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 & transparancy
  - from journals, universities and governments
  - increase in cooperation (over wider distances)
  - access to your own files
  - make yourself more visible

What I want (and don't want) with this workshop

- 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 reproducable and open

How we do it

- Every session start with some introductionary slides
- Then some assignment is given
  - use with some tool
  - try to figure it out for yourself

Related work

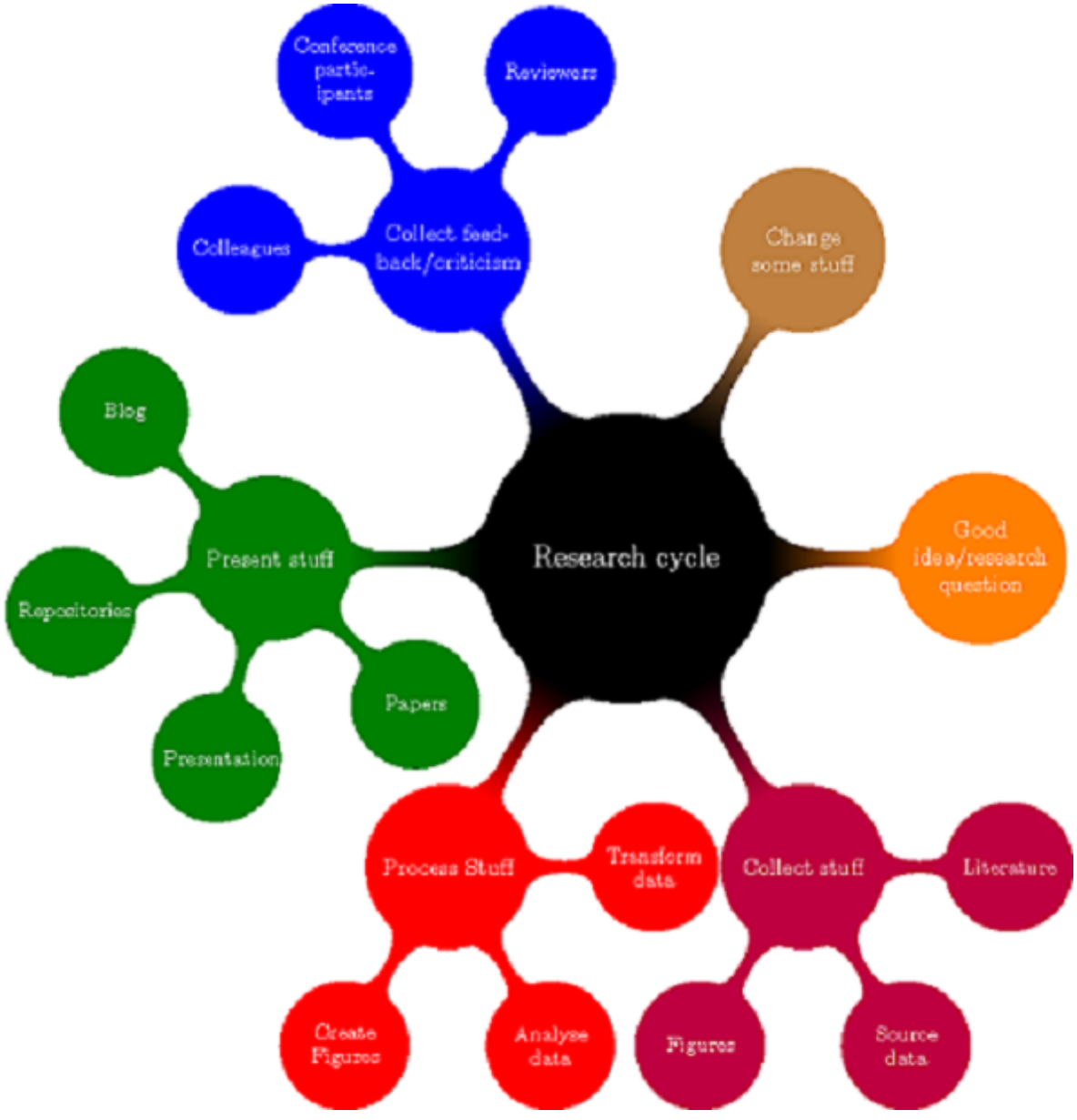
- Inspired by Kieran Healey's (associate professor in sociology) work: Choosing your Workflow Applications
- Courses for reproducable research seems to pop up everywhere (but mostly in datascience courses):
  - Datascience course: `https://www.coursera.org/`
  - Tools for Reproducible Research `http://kbroman.org/Tools4RR/`

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 reproducable 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

Reproducability

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 reproducability 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 reproducable *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 aticles, presentations or websites
4. Only output is displayed/interpreted differently (e.g., in a browser or pdf viewer)

Tools for reproducability

- Markup lanaguages
  - Markdown
- Versioning system (Git)
- Online repository (GitHub)
- Terminal tools (diff, pandoc) RStudio GitHub DeskTop

Only implicitly we make use of LaTeXHTML and pandoc (all under the hood of RStudio)

Schedule

Schedule