

Where to Start?

Open Science for Early Career Researchers

Daisy Lei*, Sai Koneru, Hollie Mullin*

*Department of Psychology, College of Information
Sciences and Technology (IST)

dul261@psu.edu, sdk96@psu.edu, ham5439@psu.edu



Introduction



About Us



Daisy Lei



Sai Koneru



Hollie Mullin

Motivation for Open Science Practices

Collaboration

Reproducibility



Efficiency

Quality

Stages of Research



1.

Ideation and Research Proposal

- What is a pre-registration/registered report?
- Open Science Framework ([OSF](#))
- Some journals require pre-registrations.
 - You can check journal requirements [here](#).
 - See slides from this bootcamp's [Pre-registration workshop](#).
- Can use Quarto to write pre-registrations. What is [Quarto](#)?
 - This is an [example](#) of OSF pre-reg using Quarto.
 - See my pre-reg template for Quarto [here](#).
 - See slides from this bootcamp's [R Markdown and Quarto workshop](#).



1.

Ideation and Research Proposal cont.

```

1 ---
2 title: "Establishing neurotrauma: toward c
3 bibliography: new_pap
4 author:
5   - name: Hollie A. C. Mullin
6     orcid: 0000-0003-4730-1807
7     email: ham5439@psu.edu
8     affiliations:
9       - name: The Pennsylvania State Univer
10

```

- > Study Information
- > Sampling Plan
- > Variables
- > Design Plan
- > Analysis Plan
- Scripts (optional)
- Other
- References

Indices

Reliability will be measured using the intraclass correlation coefficient (ICC). ICCs are the proportion of total measured variance (e.g., variability between people, sessions, etc.) that can be attributed to variability between people (Noble, Scheinost, and Constable 2019). Within-session reliability will be defined as the mean ICC value between the back-to-back resting-state runs within the same scanning session, within the same individual. Between-session reliability will be defined as the mean ICC between the back-to-back resting-state runs over the two-year time period, within the same individual. We will utilize the ICC (3,1) by Shrout and Fleiss (1979). The between-subjects mean square is represented by BMS , EMS represents error mean square, and k is the number of raters or scanning sessions. See formula below:

$$ICC(3, 1) = \frac{BMS - EMS}{BMS + (k - 1)EMS}$$

Correlation matrices, which include Pearson correlation coefficients describing the relationship between each resting-state brain region, will be Fisher r-to-z transformed for each subject.

Graph theory metrics are described below:

- Degree: The number of brain regions that the current region is connected to. These connections are also known as edges.
- Clustering Coefficient: The proportion of connected brain regions across all neighboring regions. This is the fraction of a region's neighbors that are neighbors of each other. The clustering coefficient is synonymous with the term local efficiency.



2.

Project Organization

- Project documentation
 - Readme: file folder structure, file naming conventions, project description, experimental scripts description, methods & protocols
- Project access: who (private, collaborators, public), when (all times, upon publication)
- Project storage (hard drives and cloud storage) - both!

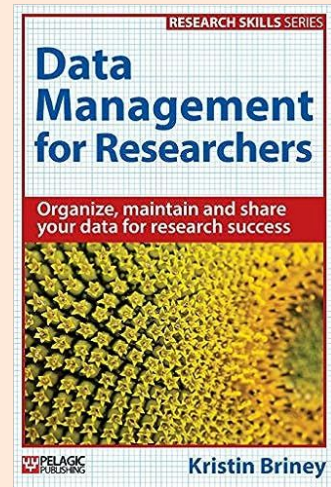
<http://osf.io>



3.

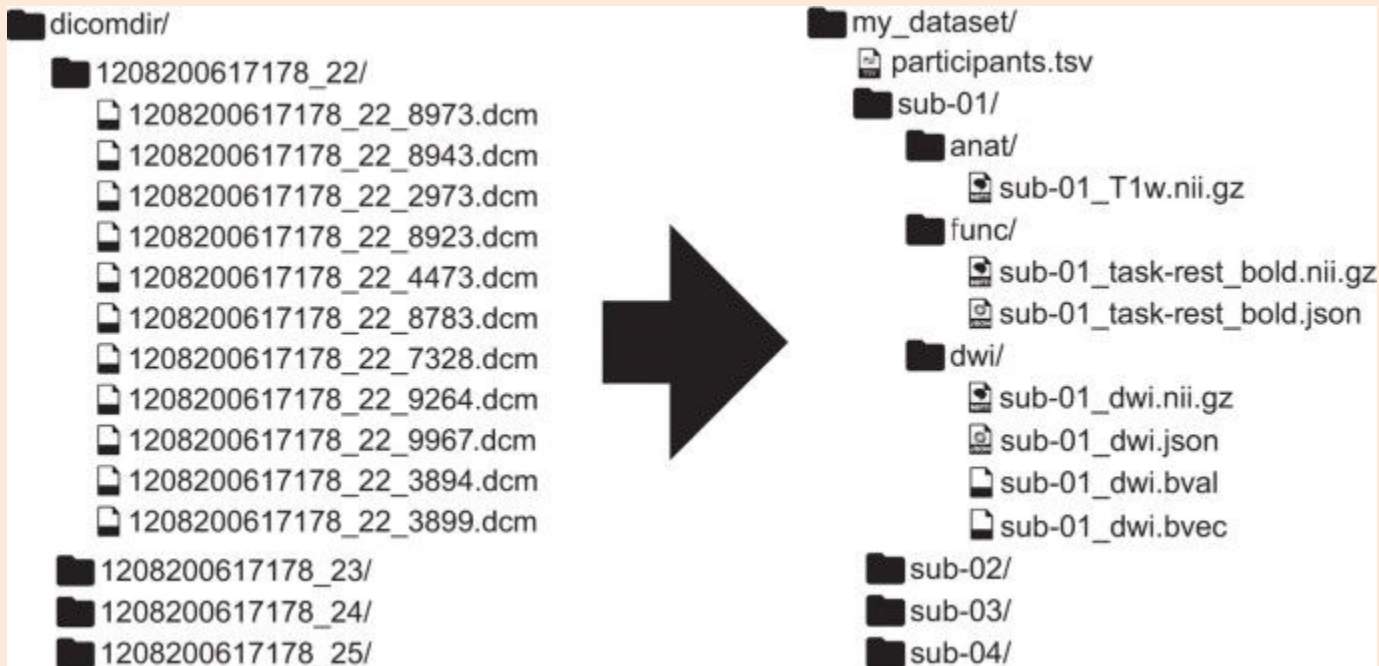
Data Management & Organization

- Readme documentation
 - Files manifest, folder hierarchy, file description, variable description (data dictionary), version info, license info
- Original, local copy, hard drive backup, cloud storage copy
- Are there certain practices used in your lab or in your field?
 - <https://guides.libraries.psu.edu/DMP>
 - Google! Ask around!
- Relevant talks/bootcamp workshops: ['Good Enough' Practices](#), [Data management: Policies](#), [Data management: Practicalities](#)



3.

Data Management and Organization cont.



4.

Data Analysis

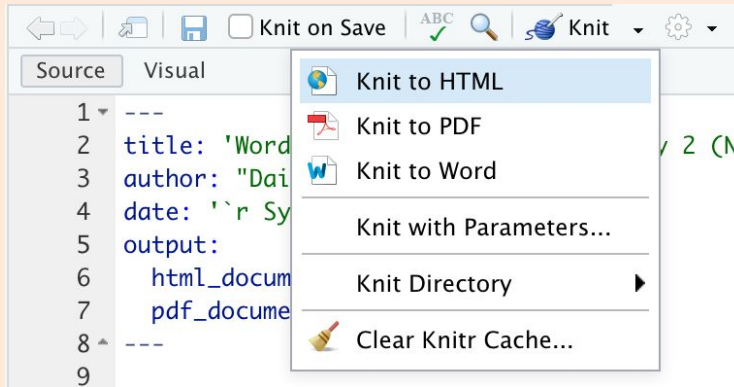
- Steps towards open data analysis
 - Tidy data format
 - Use open file formats, open source software
 - Minimize manual manipulation
 - Clear documentation of decision you make when cleaning, processing, & analysing the data
 - Comments throughout code
 - Readme of your analysis pipeline: scripts description, order of use, version of your software & packages, change log etc.



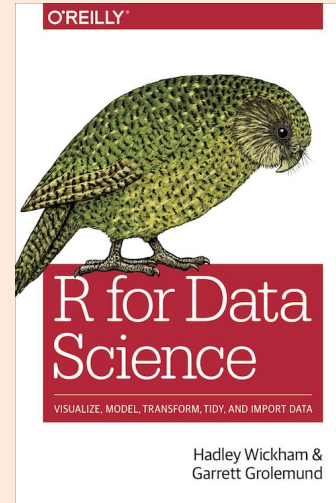
4.

Data Analysis

- Software: R/R Studio (R Markdown), Jupyter notebooks, Matlab, etc.



- Relevant bootcamp talks/workshops
 - ['Good Enough' Practices](#)
 - [Intro to R Markdown & Quarto](#)
 - [Intro to Jupyter notebooks](#)



<https://r4ds.had.co.nz/>



5.

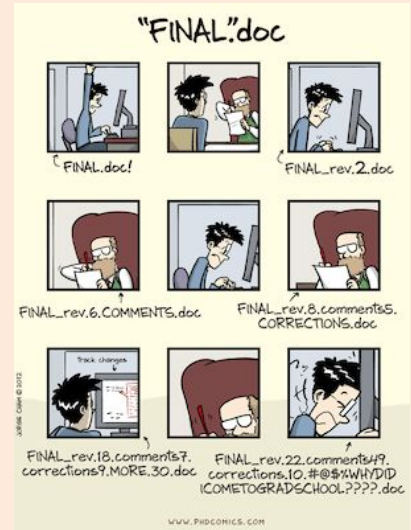
Code Sharing

Why?

- Shareable code from the beginning can save you time
- Sharing your code improves reproducibility
- Makes your code reusable, robust

Version controlling: git

- Helps in collaboration
- File versioning
- Create backups if using a remote repository like GitHub
- Allows you to experiment with your code using branches



source:

http://phdcomics.com/comics/archive_print.php?comicid=153

1



5.

Code Sharing

Folder structure

Distinguish folder types, name them accordingly

- Read-only: data, metadata
- Human-generated: code, paper, documentation
- Project-generated: clean data, figures, models

<https://asciinema.org/a/244658>

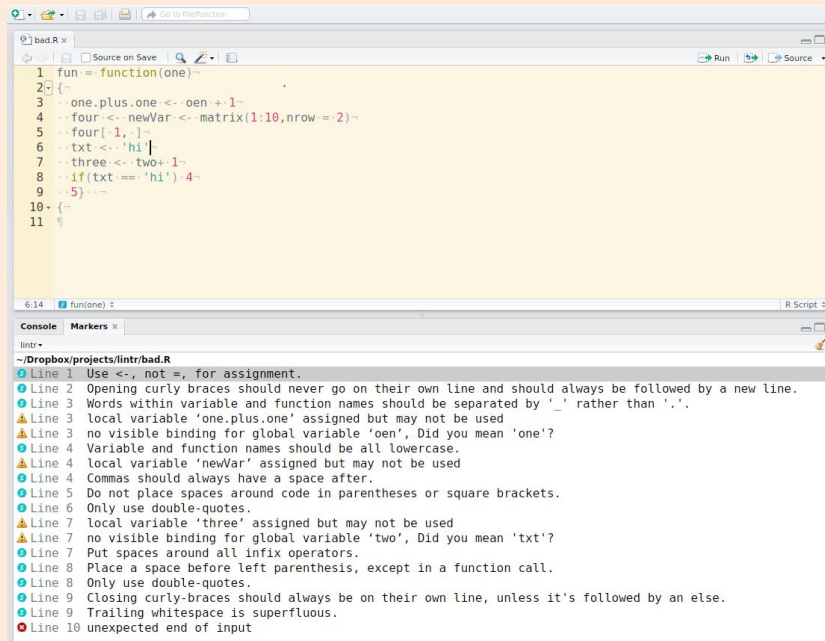


5.

Code Sharing

Code style

- Use consistent coding style e.g. PEP8, flake8 for Python or linter for R
- Improve readability with comments and descriptive naming for functions and variables (not too long though)



```
1 fun = function(one) {
2 {
3   one.plus.one <- oen + 1
4   four <- newVar <- matrix(1:10, nrow = 2)
5   four[1, ]
6   txt <- 'hi'
7   three <- two + 1
8   if(txt == 'hi') 4
9   5
10 }
11 }
```

6:14 fun(one) R Script

lint -

~/Dropbox/projects/linter/bad.R

- Line 1 Use <-, not =, for assignment.
- Line 2 Opening curly braces should never go on their own line and should always be followed by a new line.
- Line 3 Words within variable and function names should be separated by '_' rather than '.'.
- Line 3 local variable 'one.plus.one' assigned but may not be used
- Line 3 no visible binding for global variable 'oen', Did you mean 'one'?
- Line 4 Variable and function names should be all lowercase.
- Line 4 local variable 'newVar' assigned but may not be used
- Line 4 Commas should always have a space after.
- Line 5 Do not place spaces around code in parentheses or square brackets.
- Line 6 Only use double-quotes.
- Line 7 local variable 'three' assigned but may not be used
- Line 7 no visible binding for global variable 'two', Did you mean 'txt'?
- Line 7 Put spaces around all infix operators.
- Line 8 Place a space before left parenthesis, except in a function call.
- Line 8 Only use double-quotes.
- Line 9 Closing curly-braces should always be on their own line, unless it's followed by an else.
- Line 9 Trailing whitespace is superfluous.
- Line 10 unexpected end of input



5.

Code Sharing

Archiving and sharing

- Document your project using readme files
- Simplify code execution by creating easy to use high level scripts
- Document code dependencies
- Share code using platforms such as Github
- Make code citable by creating identifier using services like Zonedo/Figshare

*caution: secrets/privacy



5.

Code Sharing

Choose a license for your code

MIT License

A short and simple permissive license with conditions only requiring preservation of copyright and license notices. Licensed works, modifications, and larger works may be distributed under different terms and without source code.

Permissions

- Commercial use
- Distribution
- Modification
- Private use

Conditions

- License and copyright notice

Limitations

- Liability
- Warranty

<https://choosealicense.com/>



5.

Publication

- Preprint
 - [arXiv](#), [bioRxiv](#), [PsyArXiv](#), [SSRN](#), <https://osf.io/preprints/>, etc.
- Open access (OA)
 - OA-journals vs hybrid journals (APC Discounts for PSU Authors!)
- Postprint
 - Author-formatted accepted manuscript, not publisher-formatted
- Resource to look up a journal/publisher's OA policy:
 - Sherpa Romeo (<https://sherpa.ac.uk/romeo/>)
- Relevant bootcamp datablitz: Open Access at Penn State,



5.

Publication cont

Data sharing, material sharing, code sharing

- Considerations
 - General repositories (github, OSF) or field-specific repositories ([Databray](#), [OpenNeuro](#), [IRIS](#)), at PSU ([Data Commons](#), [ScholarSphere](#), [QDR](#))
 - Local & regional laws governing certain kinds of data
 - Data privacy
- Resource
 - Mayer (2018): [Practical Tips for Ethical Data Sharing](#)
 - Relevant bootcamp workshop: [Data Sharing](#)



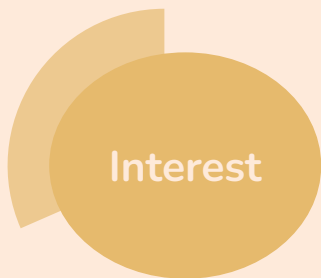
Barriers to Open Science



Challenges to Open Science in Early Career



Takes time to learn and implement open science



Your PI and others in your lab might not be interested



A pre-reg, data management, etc. can feel limiting



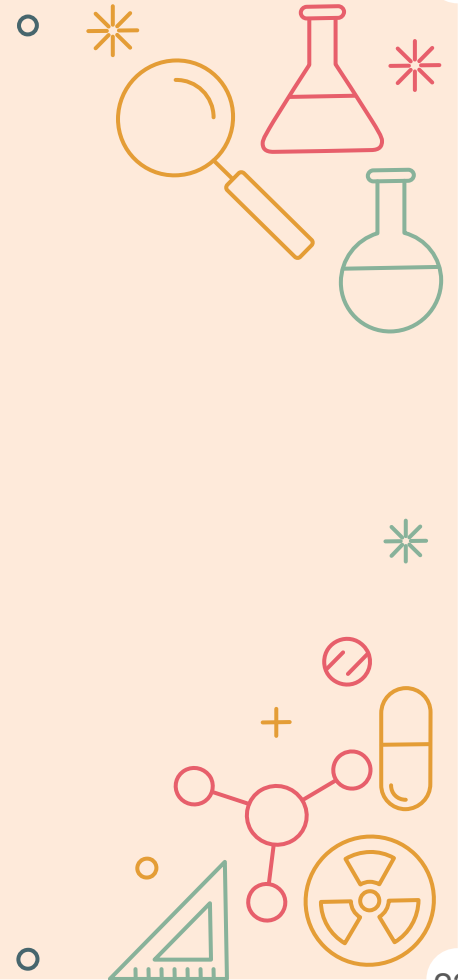
Might seem easier to keep things the way they are

Is Open Science Worth it?

Yes!



Resources



Questions?

Please take our feedback survey!



Slides created by:

Slidesgo

