Transparent and Reproducible Research with R (Part 2)

And a bit of git/GitHub

Daniel Anderson
Joshua M Rosenberg
April 7, 2019

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

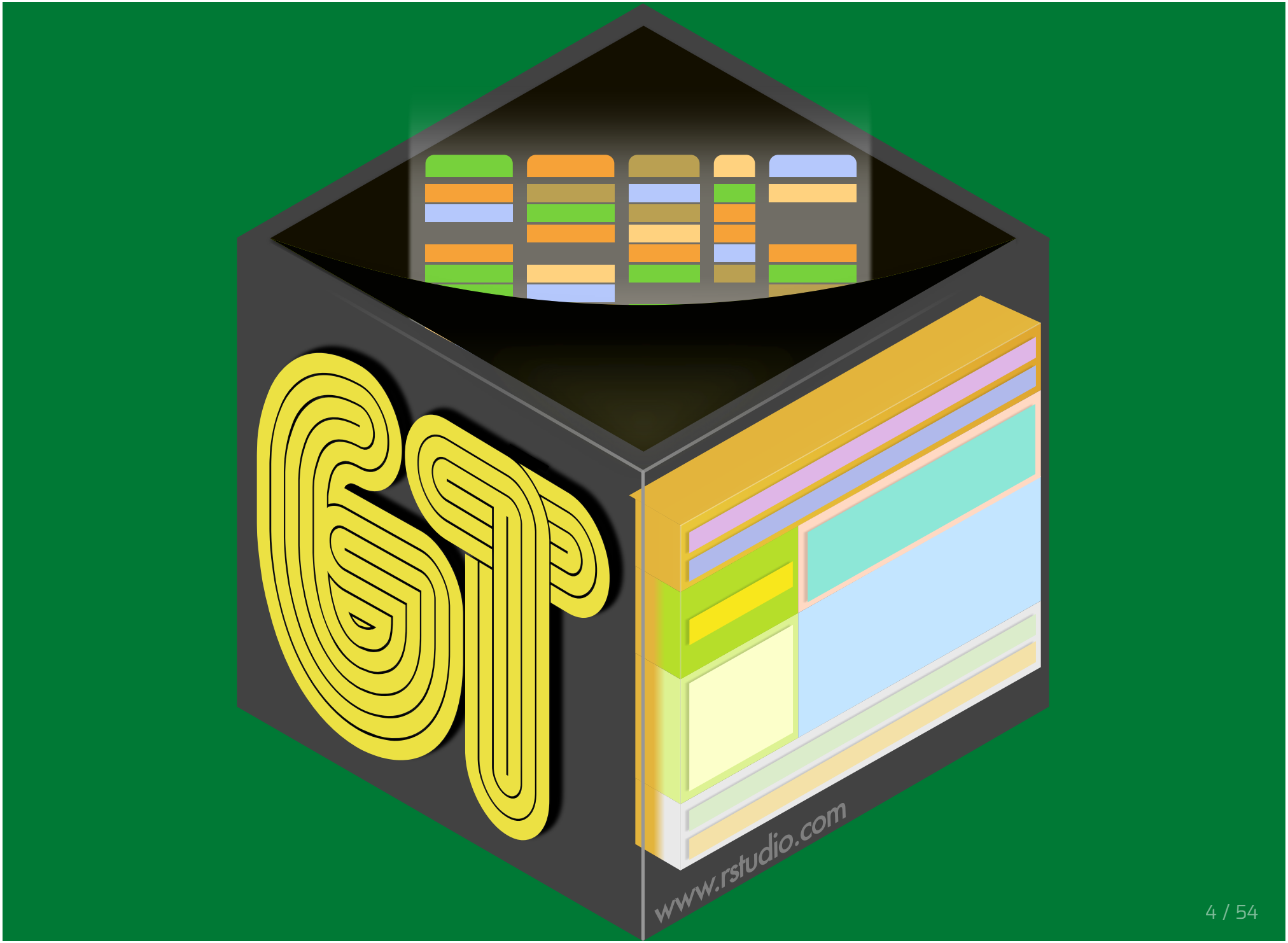
Agenda

Last two hours

- Advanced R Markdown functionality (10:15 - 11:00 minutes)
    - Formatting tables (20 min.)
    - Creating manuscripts to submit for publication (via {papaja}; 25 min.)
- Use of git/GitHub for version control and collaboration (11:00 - 11:45)
    - Introduction to GitHub, RStudio interface, and GitKraken GUI (20 min.)
    - Making changes, committing them, and pushing them to the repository (15 min.)
    - Use of GitHub (and ignoring specific files via `.gitignore`; 10 min.)
- Wrap-up/ideas for next steps/staying in touch (11:45 - 12:00)

R Markdown Tables

10:15 - 10:35am


www.rstudio.com

Overview

- New package (still very actively under development) by RStudio
- Really promising
  - Pipe-oriented
  - Beautiful tables easy
  - Spanner heads/grouping used to be a total pain - not so anymore
  - Renders to HTML/PDF without even thinking about it
- May run into bumps because of the active development

Install

```
remotes::install_github("rstudio/gt")
```

The hard part

- Getting your data in the format you want a table in
- Utilize your `gather`/`spread` skills regularly

```
library(fivethirtyeight)
flying
```

```
## # A tibble: 1,040 x 27
##    respondent_id gender age   height children_under_18 household_income
##            <dbl> <chr>  <ord> <ord>  <lgl>             <ord>
##  1    3436139758 <NA>   <NA>  <NA>   NA                <NA>
##  2    3434278696 Male   30-44 "6'3\… TRUE              <NA>
##  3    3434275578 Male   30-44 "5'8\… FALSE             $100,000 - $149…
##  4    3434268208 Male   30-44 "5'11… FALSE             $0 - $24,999
##  5    3434250245 Male   30-44 "5'7\… FALSE             $50,000 - $99,9…
##  6    3434245875 Male   30-44 "5'9\… TRUE              $25,000 - $49,9…
##  7    3434235351 Male   30-44 "6'2\… TRUE              <NA>
##  8    3434218031 Male   30-44 "6'0\… TRUE              $0 - $24,999
##  9    3434213681 <NA>   <NA>  "6'0\… TRUE              <NA>
## 10    3434172894 Male   30-44 "5'6\… FALSE             $0 - $24,999
## # … with 1,030 more rows, and 21 more variables: education <ord>,
## #   location <chr>, frequency <ord>, recline_frequency <ord>,
## #   recline_obligation <lgl>, recline_rude <ord>, recline_eliminate <lgl>,
```

```
smry <- flying %>%
  count(gender, age, recline_frequency) %>%
  filter(!is.na(age),
         !is.na(recline_frequency)) %>%
  spread(age, n)

smry
```

```
## # A tibble: 10 x 6
##    gender recline_frequency   `18-29` `30-44` `45-60` `> 60`
##    <chr>  <ord>                 <int>   <int>   <int>  <int>
##  1 Female Never                    24      21      19     23
##  2 Female Once in a while          36      25      30     36
##  3 Female About half the time      10      22      18     17
##  4 Female Usually                  13      22      26     28
##  5 Female Always                   10      21      29     12
##  6 Male   Never                    24      17      20     18
##  7 Male   Once in a while          19      39      40     29
##  8 Male   About half the time      11      11      16     11
##  9 Male   Usually                  14      30      15     27
## 10 Male   Always                   11      14      21     14
```

Turn into table

Disclaimer

These all look slightly different on the slides

```
library(gt)
smry %>%
  gt()
```

gender	recline_frequency	18-29	30-44	45-60
Female	Never	24	21	19
Female	Once in a while	36	25	30
Female	About half the time	10	22	18
Female	Usually	13	22	26
Female	Always	10	21	29
Male	Never	24	17	20
Male	Once in a while	19	39	40
Male	About half the time	11	11	16
Male	Usually	14	30	15
Male	Always	11	14	21

Add gender as a grouping variable

```
smry %>%
  group_by(gender) %>%
  gt()
```

recline_frequency	18-29	30-44	45-60	> 60
Female				
Never	24	21	19	23
Once in a while	36	25	30	36
About half the time	10	22	18	17
Usually	13	22	26	28
Always	10	21	29	12
Male				
Never	24	17	20	18
Once in a while	19	39	40	29
About half the time	11	11	16	11
Usually	14	30	15	27
Always	11	14	21	14

Add a spanner head

```
smry %>%
  group_by(gender) %>%
  gt() %>%
  tab_spanner(label = "Age Range",
              columns = vars(`18-29`, `30-44`, `45-60`, `> 60`))
```

recline_frequency	Age Range			
	18-29	30-44	45-60	> 60
Female				
Never	24	21	19	23
Once in a while	36	25	30	36
About half the time	10	22	18	17
Usually	13	22	26	28
Always	10	21	29	12
Male				
Never	24	17	20	18
Once in a while	19	39	40	29
About half the time	11	11	16	11
Usually	14	30	15	27
Always	11	14	21	14

Change column names

```
smry %>%
  group_by(gender) %>%
  gt() %>%
  tab_spanner(label = "Age Range",
              columns = vars(`18-29`, `30-44`, `45-60`, `> 60`)) %>%
  cols_label(recline_frequency = "Recline")
```

Recline	Age Range			
	18-29	30-44	45-60	> 60
Female				
Never	24	21	19	23
Once in a while	36	25	30	36
About half the time	10	22	18	17
Usually	13	22	26	28
Always	10	21	29	12
Male				
Never	24	17	20	18
Once in a while	19	39	40	29
About half the time	11	11	16	11
Usually	14	30	15	27
Always	11	14	21	14

Align columns

```
smry %>%
  group_by(gender) %>%
  gt() %>%
  tab_spanner(label = "Age Range",
              columns = vars(`18-29`, `30-44`, `45-60`, `> 60`)) %>%
  cols_label(recline_frequency = "Recline") %>%
  cols_align(align = "left", columns = vars(recline_frequency))
```

Recline	Age Range			
	18-29	30-44	45-60	> 60
Female				
Never	24	21	19	23
Once in a while	36	25	30	36
About half the time	10	22	18	17
Usually	13	22	26	28
Always	10	21	29	12
Male				
Never	24	17	20	18
Once in a while	19	39	40	29
About half the time	11	11	16	11
Usually	14	30	15	27
Always	11	14	21	14

Add a title

```
smry %>%
  group_by(gender) %>%
  gt() %>%
  tab_spanner(label = "Age Range",
              columns = vars(`18-29`, `30-44`, `45-60`, `> 60`)) %>%
  cols_label(recline_frequency = "Recline") %>%
  cols_align(align = "left", columns = vars(recline_frequency)) %>%
  tab_header(title = "Airline Passengers",
             subtitle = "Leg space is limited, what do you do?")
```

Airline Passengers				
Leg space is limited, what do you do?				
Recline	**Age Range**			
	18-29	**30-44**	**45-60**	**> 60**
Female				
Never	24	21	19	23
Once in a while	36	25	30	36
About half the time	10	22	18	17
Usually	13	22	26	28
Always	10	21	29	12
Male				
Never	24	17	20	18
Once in a while	19	39	40	29
About half the time	11	11	16	11
Usually	14	30	15	27
Always	11	14	21	14

Format columns

```
smry %>%
  mutate_at(vars(`18-29`, `30-44`, `45-60`, `> 60`), ~./100) %>%
  group_by(gender) %>%
  gt() %>%
  tab_spanner(label = "Age Range",
              columns = vars(`18-29`, `30-44`, `45-60`, `> 60`)) %>%
  fmt_percent(vars(`18-29`, `30-44`, `45-60`, `> 60`),
              decimals = 0) %>%
  cols_label(recline_frequency = "Recline") %>%
  cols_align(align = "left", columns = vars(recline_frequency)) %>%
  tab_header(title = "Airline Passengers",
             subtitle = "Leg space is limited, what do you do?")
```

Airline Passengers				
Leg space is limited, what do you do?				
Recline	**Age Range**			
	18-29	**30-44**	**45-60**	**> 60**
Female				
Never	24%	21%	19%	23%
Once in a while	36%	25%	30%	36%
About half the time	10%	22%	18%	17%
Usually	13%	22%	26%	28%
Always	10%	21%	29%	12%
Male				
Never	24%	17%	20%	18%
Once in a while	19%	39%	40%	29%
About half the time	11%	11%	16%	11%
Usually	14%	30%	15%	27%
Always	11%	14%	21%	14%

Add a source note

```
smry %>%
  mutate_at(vars(`18-29`, `30-44`, `45-60`, `> 60`), ~./100) %>%
  group_by(gender) %>%
  gt() %>%
  tab_spanner(label = "Age Range",
              columns = vars(`18-29`, `30-44`, `45-60`, `> 60`)) %>%
  fmt_percent(vars(`18-29`, `30-44`, `45-60`, `> 60`),
              decimals = 0) %>%
  cols_label(recline_frequency = "Recline") %>%
  cols_align(align = "left", columns = vars(recline_frequency)) %>%
  tab_header(title = "Airline Passengers",
             subtitle = "Leg space is limited, what do you do?") %>%
  tab_source_note(source_note = md("Data from [fivethirtyeight](https://five
```

Airline Passengers				
Leg space is limited, what do you do?				
Recline	**Age Range**			
	18-29	**30-44**	**45-60**	**> 60**
Female				
Never	24%	21%	19%	23%
Once in a while	36%	25%	30%	36%
About half the time	10%	22%	18%	17%
Usually	13%	22%	26%	28%
Always	10%	21%	29%	12%
Male				
Never	24%	17%	20%	18%
Once in a while	19%	39%	40%	29%
About half the time	11%	11%	16%	11%
Usually	14%	30%	15%	27%
Always	11%	14%	21%	14%
Data from fivethirtyeight				

Color cells

```
smry %>%
  mutate_at(vars(`18-29`, `30-44`, `45-60`, `> 60`), ~./100) %>%
  group_by(gender) %>%
  gt() %>%
  tab_spanner(label = "Age Range",
              columns = vars(`18-29`, `30-44`, `45-60`, `> 60`)) %>%
  fmt_percent(vars(`18-29`, `30-44`, `45-60`, `> 60`),
              decimals = 0) %>%
  cols_label(recline_frequency = "Recline") %>%
  data_color(vars(`18-29`, `30-44`, `45-60`, `> 60`),
             colors = scales::col_numeric(palette = c(c("#FFFFFF", "#FF0000'
  cols_align(align = "left", columns = vars(recline_frequency)) %>%
  tab_header(title = "Airline Passengers",
             subtitle = "Leg space is limited, what do you do?") %>%
  tab_source_note(source_note = md("Data from [fivethirtyeight](https://five
```

Airline Passengers				
Leg space is limited, what do you do?				
Recline	**Age Range**			
	18-29	**30-44**	**45-60**	**> 60**
Female				
Never	24%	21%	19%	23%
Once in a while	36%	25%	30%	36%
About half the time	10%	22%	18%	17%
Usually	13%	22%	26%	28%
Always	10%	21%	29%	12%
Male				
Never	24%	17%	20%	18%
Once in a while	19%	39%	40%	29%
About half the time	11%	11%	16%	11%
Usually	14%	30%	15%	27%
Always	11%	14%	21%	14%
Data from fivethirtyeight				

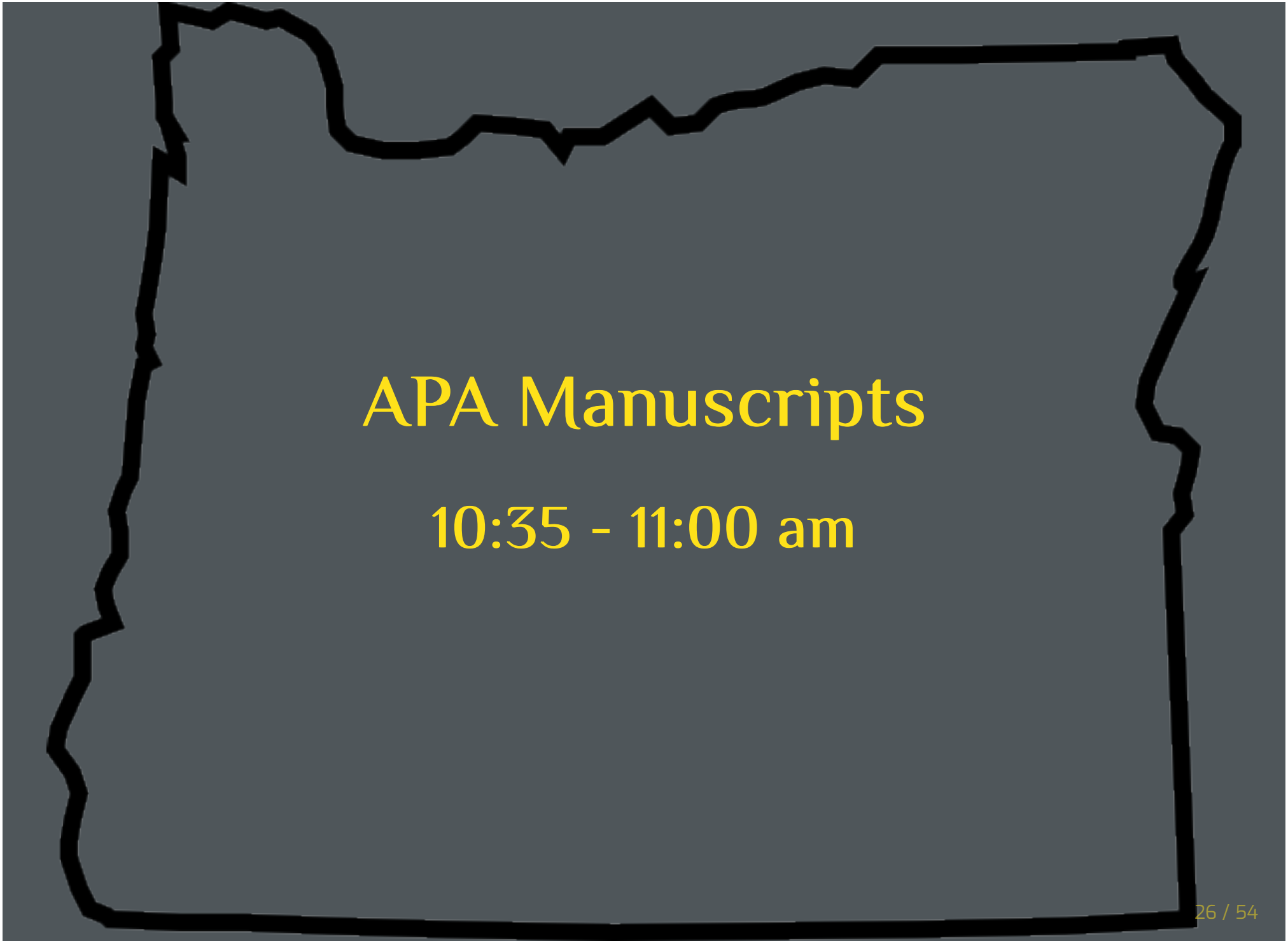
What else?

- Lots more it can do, and lots more in development
- See the website

What else?

- Lots more it can do, and lots more in development
- See the website

gtcars case study is worth going through



APA Manuscripts

10:35 - 11:00 am

{papaja}

Despite having been around for about 4 years {papaja} is still not on CRAN.

More evidence that some of the best packages are not on CRAN.

{papaja}

Despite having been around for about 4 years {papaja} is still not on CRAN.

More evidence that some of the best packages are not on CRAN.

Install with devtools

```
devtools::install_github("crsh/papaja")
```

WIP

The package is seemingly perpetually under development. What does this mean?

WIP

The package is seemingly perpetually under development. What does this mean?

- Re-install regularly.

WIP

The package is seemingly perpetually under development. What does this mean?

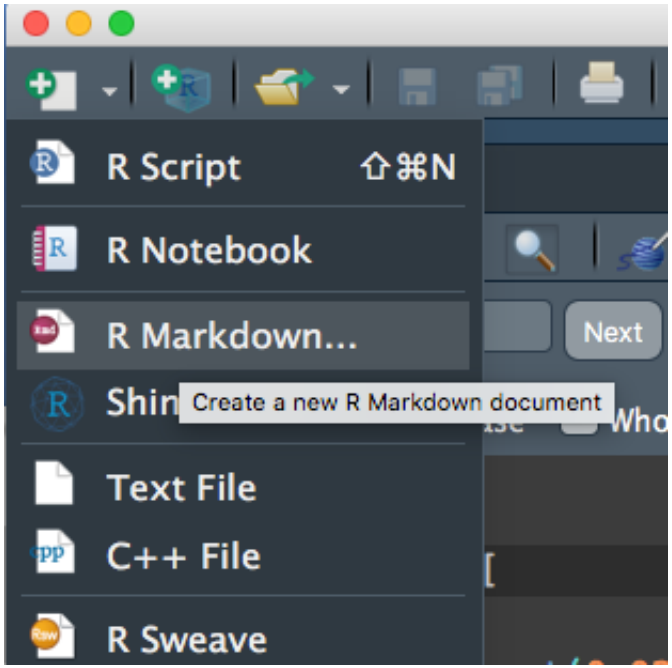
- Re-install regularly.
- Not everything may work perfect - don't worry though, most things do

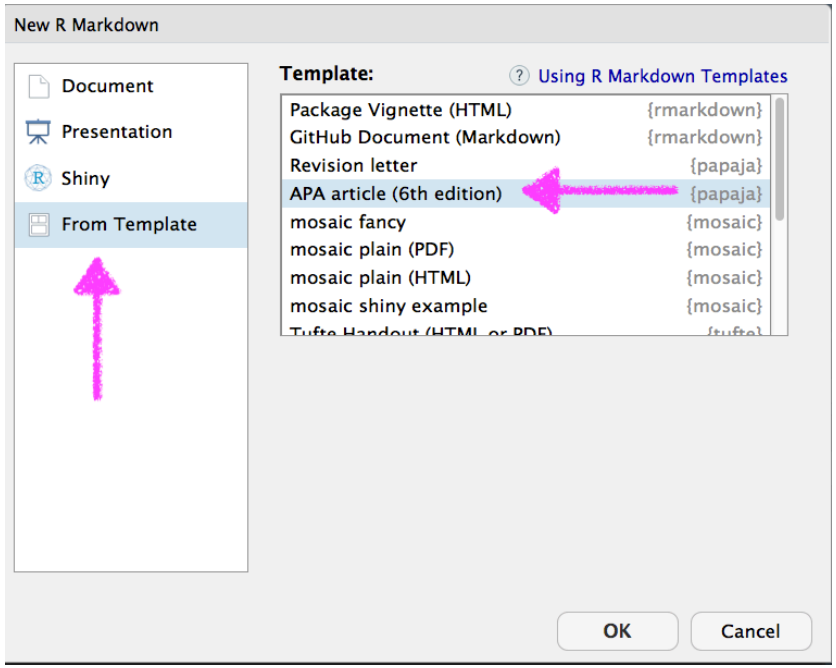
WIP

The package is seemingly perpetually under development. What does this mean?

- Re-install regularly.
- Not everything may work perfect - don't worry though, most things do
- You may want to peruse the current issues
  - If you run into one (and you're sure it's an issue) consider opening one yourself
  - Bonus - the developer is very kind, so even if you open up a silly issue, he's likely to be understanding

Use the template





YAML

A few more options than the default

```yaml
---
title             : "The title"
shorttitle        : "Title"

author:
  - name          : "First Author"
    affiliation   : "1"
    corresponding : yes    # Define only one corresponding author
    address       : "Postal address"
    email         : "my@email.com"
  - name          : "Ernst-August Doelle"
    affiliation   : "1,2"

affiliation:
  - id            : "1"
    institution   : "Wilhelm-Wundt-University"
  - id            : "2"
    institution   : "Konstanz Business School"

authornote: |
  Add complete departmental affiliations for each author here. Each new line herein must be ir

  Enter author note here.

abstract: |
  One or two sentences providing a **basic introduction** to the field,  comprehensible to a s

  Two to three sentences of **more detailed background**, comprehensible  to scientists in rel

  One sentence clearly stating the **general problem** being addressed by  this particular stu

  One sentence summarizing the main result (with the words "**here we show**" or their equival

  Two or three sentences explaining what the **main result** reveals in direct comparison to w

  One or two sentences to put the results into a more **general context**.

  Two or three sentences to provide a **broader perspective**, readily comprehensible to a sci

  <!-- https://tinyurl.com/ybremelq -->

keywords          : "keywords"
wordcount         : "X"

bibliography      : ["r-references.bib"]

floatsintext      : no
figurelist        : no
tablelist         : no
footnotelist      : no
linenumbers       : yes
mask              : no
draft             : no

documentclass     : "apa6"
classoption       : "man"
output            : papaja::apa6_pdf
---
```

First thing - Render!

The title

First Author[1] & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Modifications

- Obvious ones
  - title
  - author & author info
  - abstract
  - keywords

Modifications

- Obvious ones
  - title
  - author & author info
  - abstract
  - keywords
- Less obvious
  - `shorttitle` (running head)
  - `authornote` (can fully delete or modify)
  - `wordcount` (fairly useless at this point imo)
  - `bibliography` (we'll talk more about this momentarily)
  - `linenumbers`
  - floats
  - mask (for blind peer-review)
  - `classoption`

Let's play for a minute!

Modify some of the options on the previous slide.

Specifically, try changing `classoption` from `man` to `jou`. Try other things too.

Add some LaTeX options

```
header-includes:
  - \raggedbottom
  - \setlength{\parskip}{0pt}
```

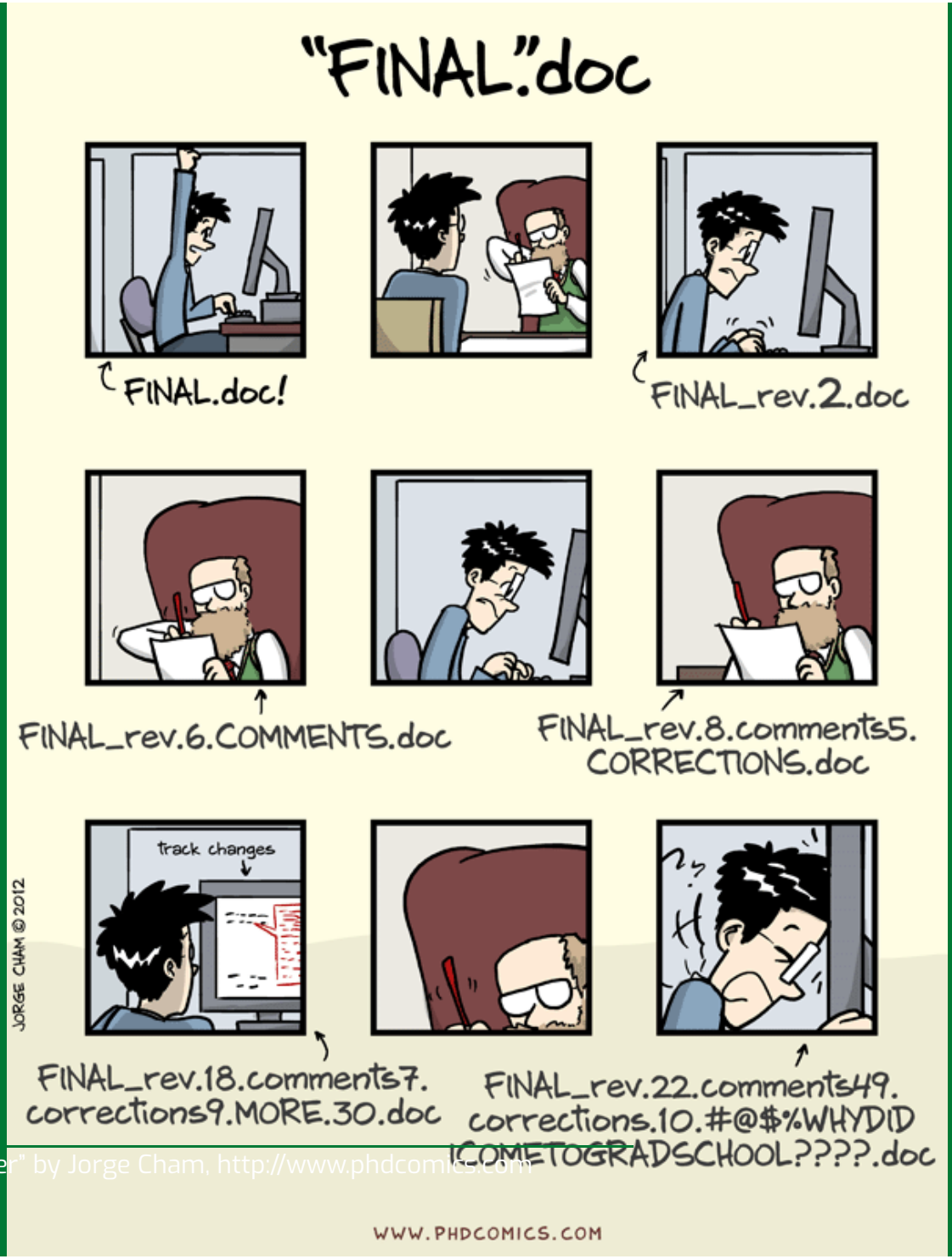
This will help (save you lots of time googling) remove the extra space between paragraphs.

git/GitHub

An Introduction

11:00 - 11:40am

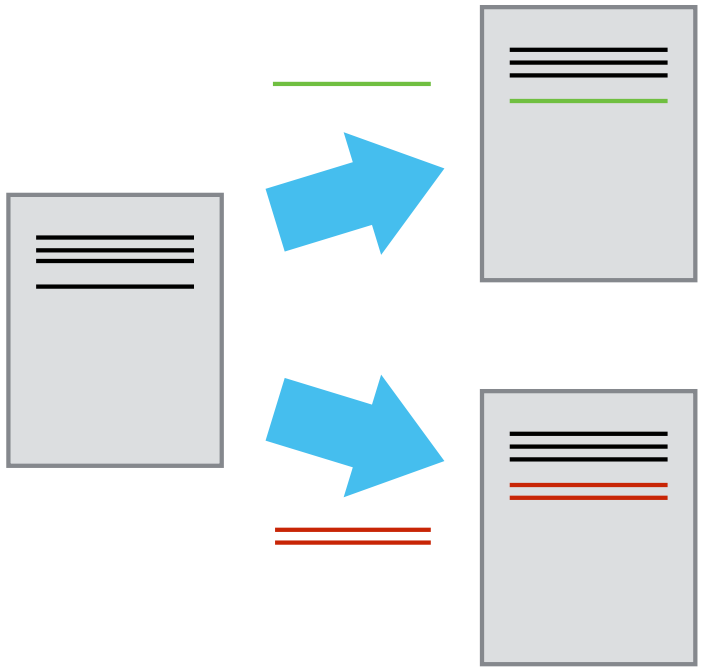
More info can be found here: http://www.datalorax.com/vita/ds/ds1-slides/w4p2/



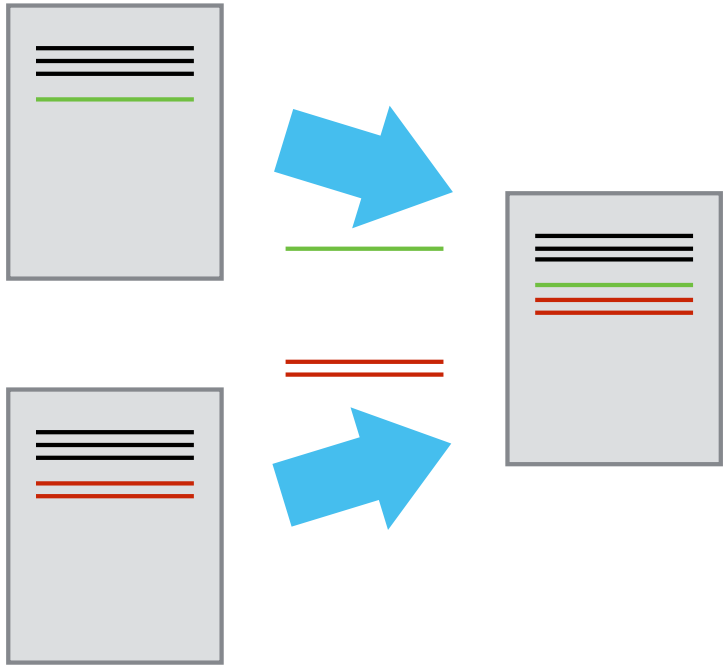
"Piled Higher and Deeper" by Jorge Cham, http://www.phdcomics.com

From swcarpentry

We can think of the changes as separate from the document



This means there are many possible versions of the same document



Unless there are conflicts, two changes from the same document can be merged together

How?

That's what we'll do today!

Some basic terminology

- Version Control System
  - A tool to help us track changes. *git* is one such system.

Some basic terminology

- Version Control System
  - A tool to help us track changes. *git* is one such system.
- Commit
  - Changes that have been made to the file(s)

Some basic terminology

- Version Control System
  - A tool to help us track changes. *git* is one such system.
- Commit
  - Changes that have been made to the file(s)
- Repository (repo)
  - Entire project (including history)

Some basic terminology

- Version Control System
  - A tool to help us track changes. *git* is one such system.
- Commit
  - Changes that have been made to the file(s)
- Repository (repo)
  - Entire project (including history)
- Clone
  - Download files locally

Some basic terminology

- Version Control System
  - A tool to help us track changes. *git* is one such system.
- Commit
  - Changes that have been made to the file(s)
- Repository (repo)
  - Entire project (including history)
- Clone
  - Download files locally
- Pull
  - Get latest changes

Some basic terminology

- Version Control System
  - A tool to help us track changes. *git* is one such system.
- Commit
  - Changes that have been made to the file(s)
- Repository (repo)
  - Entire project (including history)
- Clone
  - Download files locally
- Pull
  - Get latest changes
- Push
  - Send latest commits to remote repo

Demo (30 minutes)

Goal: Get your work from this morning in a GitHub repo!

- Create new repo
- Clone it locally
- Add your files
- Commit
- Push
- Publish

Using GitHub

11:35 - 11:45am

Example

tidyLPA

Ignoring Files

- When we initialized the repo, we started it with a `.gitignore` file
- The `.gitignore` file tells the repo not to track certain files
  - e.g., proprietary data

Wrapping up

11:45am - 12:00pm

Daniel Anderson

- daniela@uoregon.edu
- @datalorax_ Joshua Rosenberg -jmrosenberg@utk.edu -[@jrosenberg6432](https://twitter.com/jrosenberg6432)

Mailing list: rr-in-edu@googlegroups.com

Questions? Ideas? Thank you!

Appendix

Citations

Citations

To include references in your paper, you must:

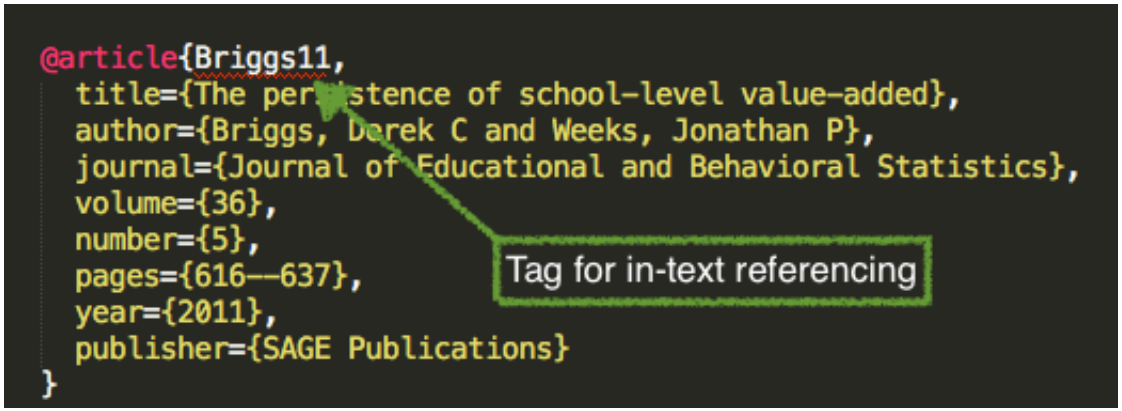
- Create an external .bib file using LaTeX formatting (we'll get to this)
- Include `bibliography: nameOfYourBibFile.bib` in your YAML front matter.
- Refer to the citations in text using `@`

Creating a .bib doc

Creating a .bib doc

```
@article{briggs2011persistence,
  title={The persistence of school-level value-added},
  author={Briggs, Derek C and Weeks, Jonathan P},
  journal={Journal of Educational and Behavioral Statistics},
  volume={36},
  number={5},
  pages={616--637},
  year={2011},
  publisher={SAGE Publications}
}
```

```
@article{briggs2011persistence,
  title={The persistence of school-level value-added},
  author={Briggs, Derek C and Weeks, Jonathan P},
  journal={Journal of Educational and Behavioral Statistics},
  volume={36},
  number={5},
  pages={616--637},
  year={2011},
  publisher={SAGE Publications}
}
```



In text citations

Citation Style	Output
@Briggs11	Briggs and Weeks (2011)
[see @Baldwin2014; @Caruso2000]	(see Baldwin et al. 2014; Caruso 2000)
[@Linn02, p. 9]	(Linn and Haug 2002, 9)
[-@Goldhaber08]	(2008)

Note this is not APA. However, references are included automatically at the end of the document. Include `# References` as the last line of your document to give it a title.

A few real examples

```
A more flexible approach, discussed by @ho_09 and others [@ho_12;
@reardon_15], is to construct a probability-probability (PP) plot and compute
the area under the PP curve (see Figure 1). Note that the PP curve is
```

```
normal distributions separated by standard deviation units [@ho_09]. The *V*
statistic assumes respective normality "where two distributions that may not
be normal may nonetheless be normal with respect to each other, under a shared
transformation" [@ho_09, p. 217]. The assumptions underlying *V* are thus less
```

```
where $ref$ and $foc$ represent the reference and focal groups, respectively
[see @lakens_13]. The magnitude of the effect size is primarily driven by the
```

References

References

Baldwin, Scott A, Zac E Imel, Scott R Braithwaite, and David C Atkins. 2014. "Analyzing Multiple Outcomes in Clinical Research Using Multivariate Multilevel Models." *Journal of Consulting and Clinical Psychology* 82 (5). American Psychological Association: 920.

Briggs, Derek C, and Jonathan P Weeks. 2011. "The Persistence of School-Level Value-Added." *Journal of Educational and Behavioral Statistics* 36 (5). SAGE Publications: 616–37.

Caruso, John C. 2000. "Reliability Generalization of the NEO Personality Scales." *Educational and Psychological Measurement* 60 (2). Sage Publications: 236–54.

Goldhaber, D., and M. Hansen. 2008. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. CPRE Working Paper No. 2008-5, University of Washington." Report.

Linn, R. L., and C. Haug. 2002. "Stability of School-Building Accountability Scores and Gains." Journal Article. *Educational Evaluation and Policy Analysis* 24: 29–36. doi:10.3102/01623737024001029.