Reproducibility in research

Health Data Working Group

Aya Mitani

Dalla Lana School of Public Health

2021-11-28

Some definitions

Science

> knowledge about the structure and behaviour of the natural and physical world, based on **facts** that you can prove, for example by experiments (Oxford Learner's Dictionary)

Research

> a careful study of a subject, especially in order to discover new **facts** or information about it (Oxford Learner's Dictionary)

Some definitions

Science

> knowledge about the structure and behaviour of the natural and physical world, based on **facts** that you can prove, for example by experiments (Oxford Learner's Dictionary)

Research

> a careful study of a subject, especially in order to discover new **facts** or information about it (Oxford Learner's Dictionary)

AND to rediscover new ~~facts or information~~ claims about it

Some definitions

Science

> knowledge about the structure and behaviour of the natural and physical world, based on **facts** that you can prove, for example by experiments (Oxford Learner's Dictionary)

Research

> a careful study of a subject, especially in order to discover new **facts** or information about it (Oxford Learner's Dictionary)

AND to rediscover new ~~facts or information~~ claims about it, BUT research is rarely reproduced

> "in the field of cancer research, only about 20–25% or 11% of published studies could be validated or reproduced, and that only about 36% were reproduced in the field of psychology" (Miyakawa, 2020)

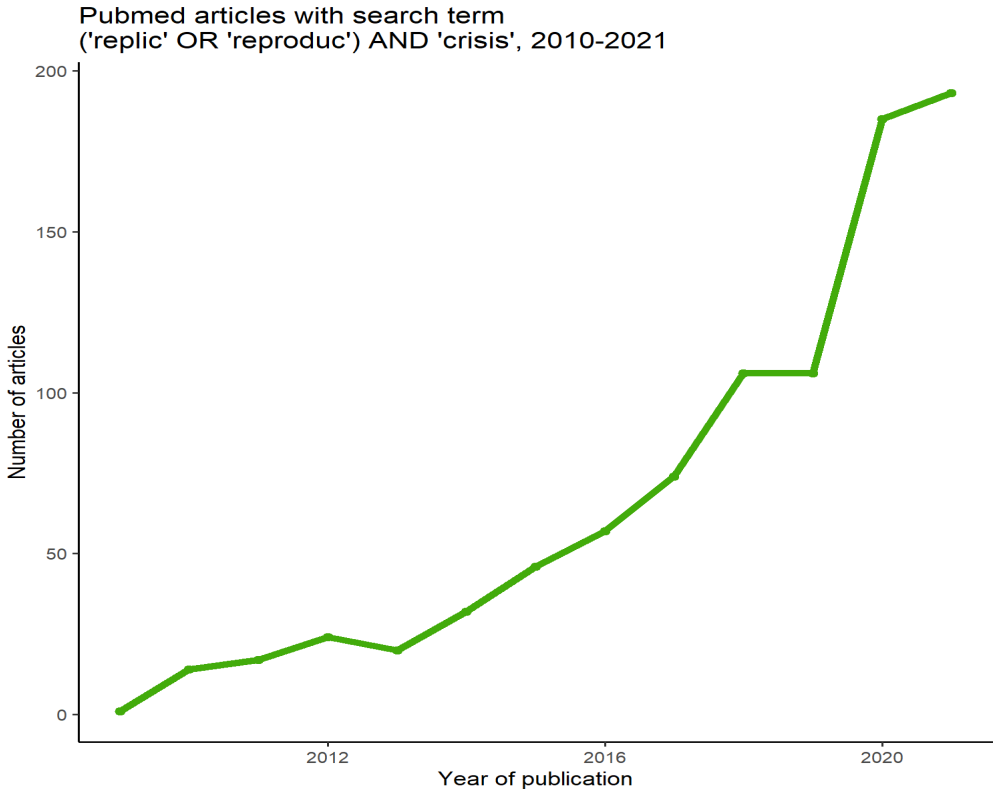
Tsuyoshi Miyakawa. No raw data, no science: another possible source of the reproducibility crisis. Molecular Brain, 13(24), 2020

Reproducibility crisis in research

Factors contributing to crisis

- Absence of replication
- Lack of transparency
- Data is not generalizable
- Poor quality of analysis

> "while our ability to generate data has grown dramatically, our ability to understand them has not developed at the same rate" (Peng, 2015)



Roger Peng. The reproducibility crisis in science: A statistical counterattack. Significance, 12(3):30-32, 2015

Reproducibility crisis in research

More factors contributing to crisis

- Publication bias
- Pressure to publish
- Lack of training

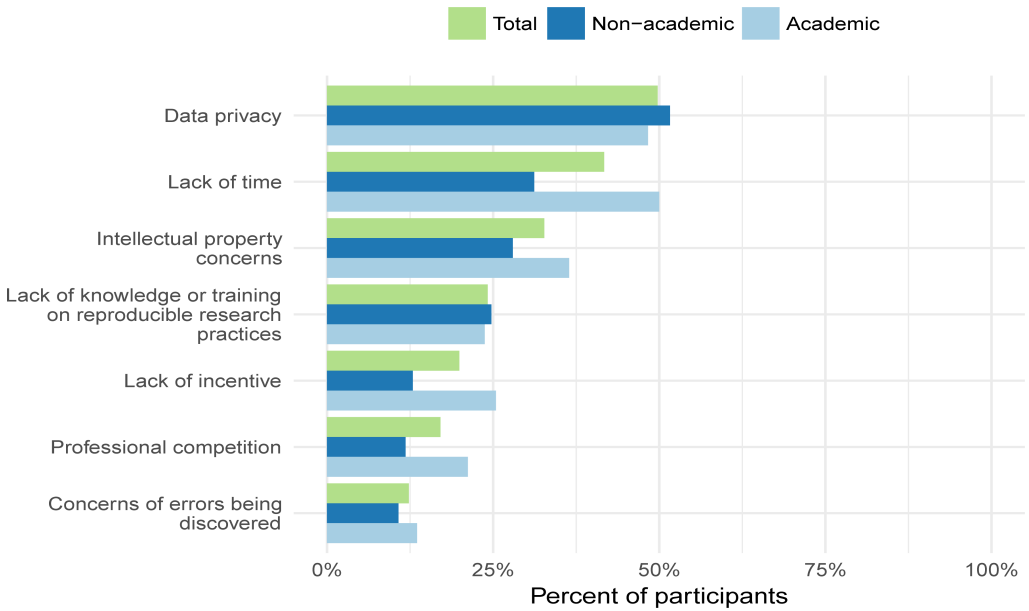


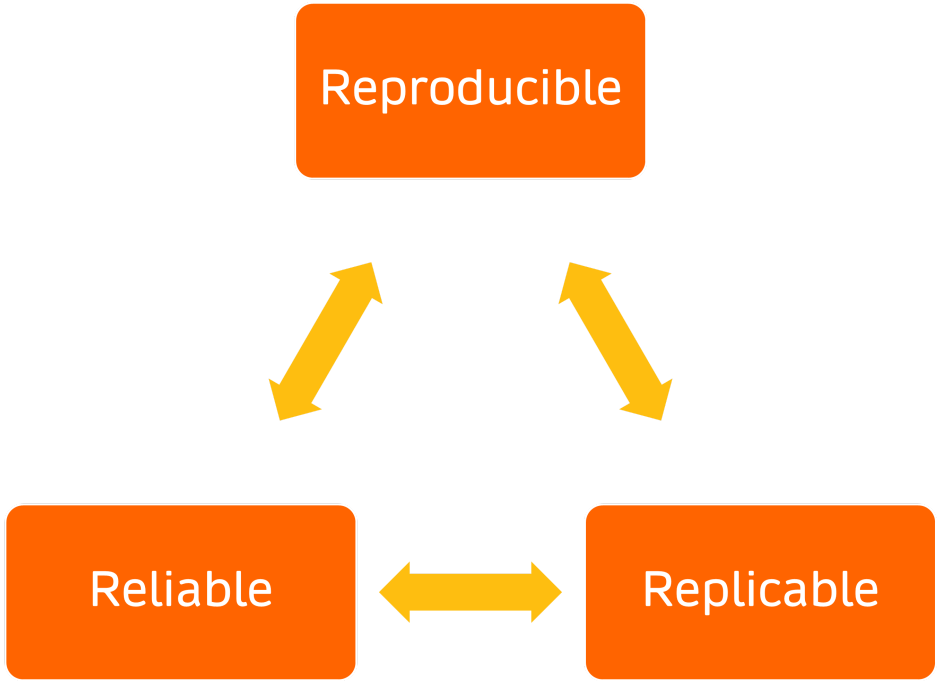
Fig 2. Percent of participants who perceived each of seven barriers to using reproducible research practices (Harris, 2018)

Jenine K. Harris et al. Use of reproducible research practices in public health: A survey of public health analysts. PLOS ONE, 13(9):e0202447, 2018

Reproducibile or replicable?

- Can someone else completely **reproduce** the results, given the data and code?
- Can someone else **replicate** the analysis using different data?

The three R in research



> "While **replication** is the gold standard for confirming evidence, **reproducibility** requires fewer resources and increases **reliability**" (Harris, 2018)
>
> "Reproducible research can still be wrong" (Leek & Peng, 2015)

Jeffrey T. Leek, Roger D. Peng. Reproducible research can still be wrong. PNAS, 112(6):1645-1646, 2015

Why make research reproducible

> It's good to repeat and review what is good twice and thrice over. [Plato]

- For **yourself**
  - Build on your own work effectively and efficiently
  - Higher research impact
  - Produce more reliable research
- For **science**
  - Standard to judge scientific claims
  - Encourage replication
  - Avoid effort duplication
  - Encourage cumulative knowledge development

How is research presented?

How is research **presented**?

- Slideshows
- Journal articles
- Books
- Websites

These are ways to **advertise** your research!

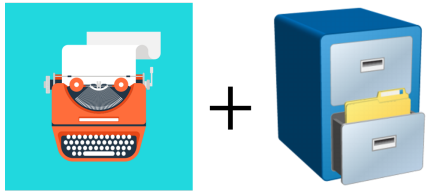
Bridging the gap between research and advertisement

- Your **research** is the

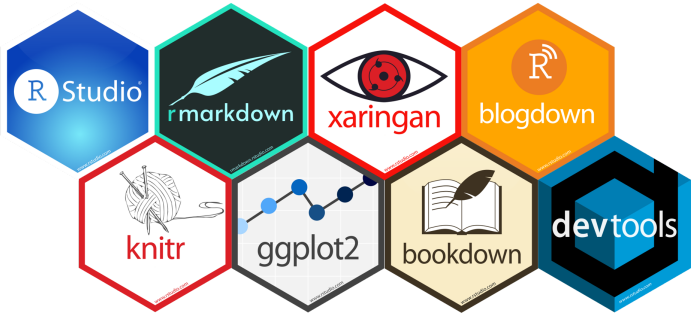
  > "full software environment, code, and data that produced the results" (Donoho, 2010)

- **Research** and **advertisement** should be combined

Then



Now



- ~~"Data and code can be requested from the first author."~~

David L Donoho. An invitation to reproducible computational research. Biostatistics, 11(3):385–388, 2010.

In this presentation

I will cover some basic tips to make your research more reproducible in `R` and `RStudio`

- Set up your **project**
- **Read in** data
- Automate **tables**
- Use **functions** and **loops**
- Develop a **package**
- **Future proofing** your project

Example data

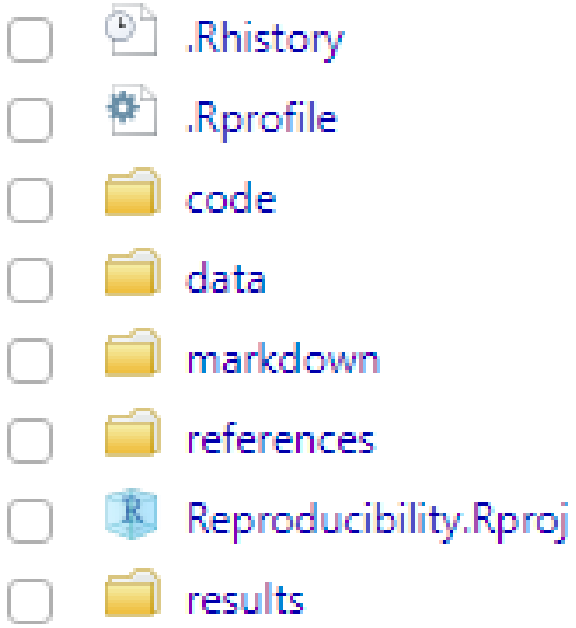
NYC Reported Dog Bites

```
##         DateOfBite                                Breed  Age Gender SpayNeuter
## 1  January 02 2015                    Poodle, Standard    3      M       true
## 2  January 02 2015                                HUSKY <NA>      U      false
## 3  January 02 2015                                 <NA> <NA>      U      false
## 4  January 01 2015 American Pit Bull Terrier/Pit Bull    6      M      false
## 5  January 03 2015 American Pit Bull Terrier/Pit Bull    1      M      false
## 6  January 05 2015 American Pit Bull Terrier/Pit Bull    1      F      false
## 7  January 04 2015                               MORKIE    1      M      false
## 8  January 05 2015                            Chihuahua    1      M      false
## 9  January 04 2015                      PIT BULL MIXED <NA>      M      false
## 10 January 04 2015                                 <NA> <NA>      U      false
## 11 January 02 2015                                 <NA> <NA>      U      false
## 12 January 02 2015          Cocker Spaniel Crossbreed <NA>      U      false
##     Borough ZipCode
## 1  Brooklyn   11238
## 2  Brooklyn   11249
## 3  Brooklyn    <NA>
## 4  Brooklyn   11221
## 5  Brooklyn   11207
```

Project from RStudio

Create your project in RStudio

- File > New Project > ...
- Create sub-folders to **organize** your project
- Separate folders for
  - code
  - data
  - results
  - presentation files
  - other documents



Data gathering

Read in external data

Are you still doing this?

```
setwd("C:\\Users\\ayami\\Documents\\Talks\\Reproducibility\\data\\raw data")
repdata <- read.csv("DOHMH_Dog_Bite_Data.csv", header = TRUE)
```

NOT REPRODUCIBLE!!!!!

- Doesn't work on a different machine
- `C:`, `D:`, or `P:`?
- `/`, `\`, or `\\` ?
- Can't output object to a different folder

Data gathering

`here` package to the rescue!

```
library(here)
```

Directory is set to project root folder

```
here()
```

```
## [1] "C:/Users/ayami/Documents/Talks/Reproducibility"
```

Read in data from a sub-folder

```
repdata <- read.csv(here("data", "raw data", "DOHMH_Dog_Bite_Data.csv"), header = TRUE)
```

Output results to a different sub-folder

```
write.csv(bite, here("data", "dogbite.csv"), row.names = FALSE)
```

Data gathering

My `create_nicedata.R` script

```
#-----------------------------------------------------------------
# Author: Aya Mitani
# Last updated: 2021-11-24
# What: Read in raw dog bite data,
#         remove incidences with missing data,
#         clean breed and age variables,
#         write new clean data
#-----------------------------------------------------------------

biteraw <- read.csv(here("data", "raw data", "DOHMH_Dog_Bite_Data.csv"), header = TRUE, na.strings=c

# create new variables, exclude missing observations, select relevant variables, etc.

write.csv(bite, here("data", "dogbite.csv"), row.names = FALSE) # output new data
```

Data gathering

My analytical data

```r
# read in data
bite <- read.csv(here("data", "dogbite.csv"), header = TRUE)
bite[67:75,]
```

```
##    UniqueID Year                                Breed Breedclean  Age Agenum
## 67      122 2015                      HARRIER/BEAGLE     Beagle 10 M      0
## 68      124 2015   Dachshund, Long Haired Miniature      Other    4      4
## 69      126 2015                 BOXER X W/ PIT BULL   Pit bull    8      8
## 70      128 2015                   Yorkshire Terrier    Terrier    6      6
## 71      130 2015                            Bull dog   Bull dog   8M      0
## 72      132 2015 American Pit Bull Terrier/Pit Bull   Pit bull    4      4
## 73      134 2015        BOXER/RHODESIAN RIDGEBACK X      Other    5      5
## 74      135 2015                Chihuahua Crossbreed  Chihuahua   4M      0
## 75      136 2015                    Poodle, Standard     Poodle    2      2
##    Gender SpayNeuter  Borough ZipCode
## 67      M       true Brooklyn   11235
## 68      M       true Brooklyn   11234
## 69      M       true Brooklyn   11231
## 70      M       true Brooklyn   11233
```

Data analysis

```
bite <- read.csv(here("data", "dogbite.csv"), header = TRUE)
```

I like to create separate files for each step of analysis

Inside `code` folder

- build_models.R
- create_nicedata.R
- figure_barchart_dogbreed.R
- figure_line_yearborough.R
- scrapbook.R
- table_ORof3models.R

Inside `results` folder

- figure_line_yearborough.png
- figure_barchart_dogbreed.png
- table_ORof3models.txt

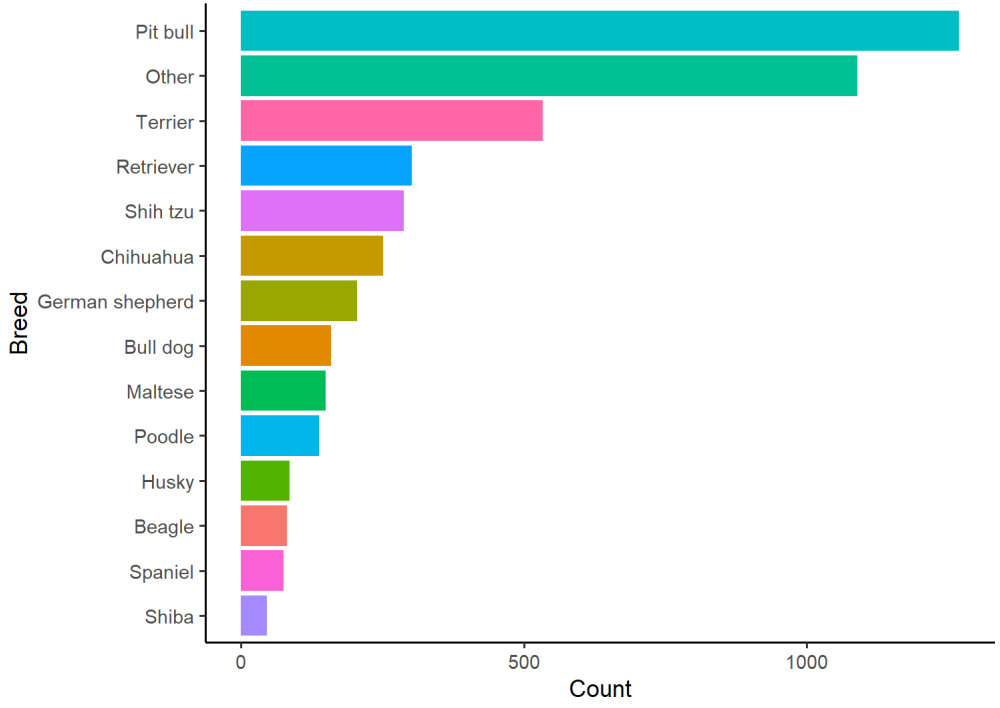
If a task relies on code from a different script,

```
source(here("code", "another_script.R"))
```

Data analysis

Document, document, document...

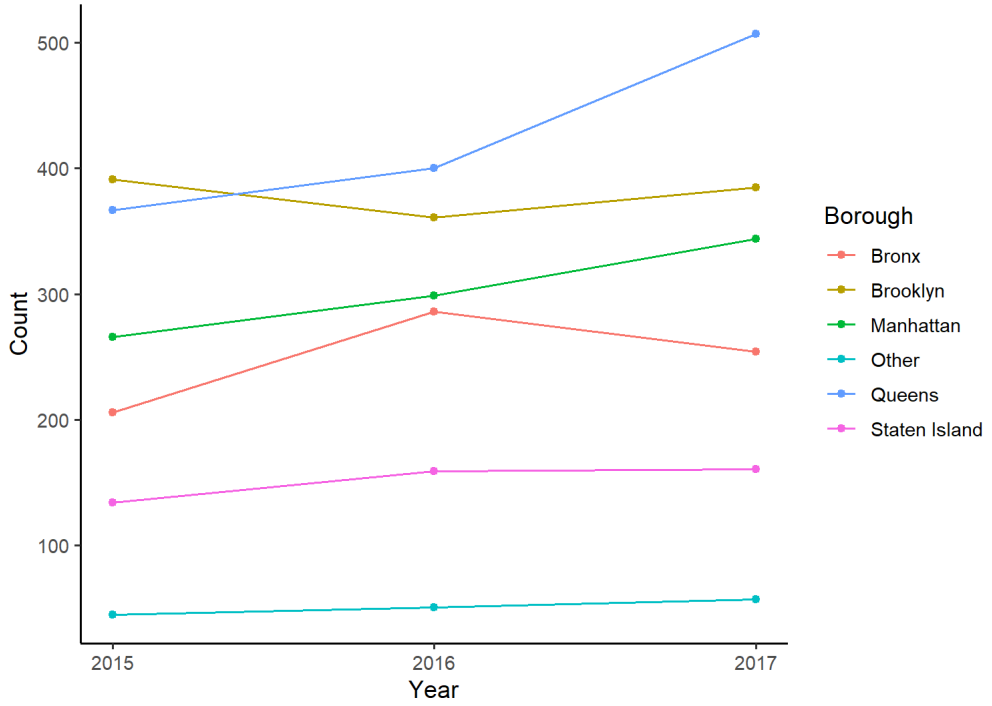
```
bite %>%
  # frequency of bite by breed
  group_by(Breedclean) %>%
  summarise(n=n()) %>%
  ggplot(aes(x = reorder(Breedclean, n), y = n
  # bar chart
  geom_bar(stat = "identity", aes(fill = Breed
  theme_classic(base_size = 12) +
  # remove legend
  theme(legend.position = "none") +
  # name axes
  labs(y = "Count", x = "Breed") +
  # flip the coordinates
  coord_flip()
```



Data analysis

Document, document, document...

```
bite %>%
  # frequency of bite by borough and year
  group_by(Borough, Year) %>%
  summarise(n = n()) %>%
  # line graph
  ggplot(aes(x = Year, y = n, color = Borough)
  geom_point() +
  geom_line() +
  labs(y = "Count", x = "Year") +
  # specify x-axis ticks and labels
  scale_x_continuous(breaks = seq(2015, 2017, 
  theme_classic(base_size = 12)
```



Data analysis

Research question

> Are bites from pit bulls more likely to be by puppies, males, or spayed dogs?

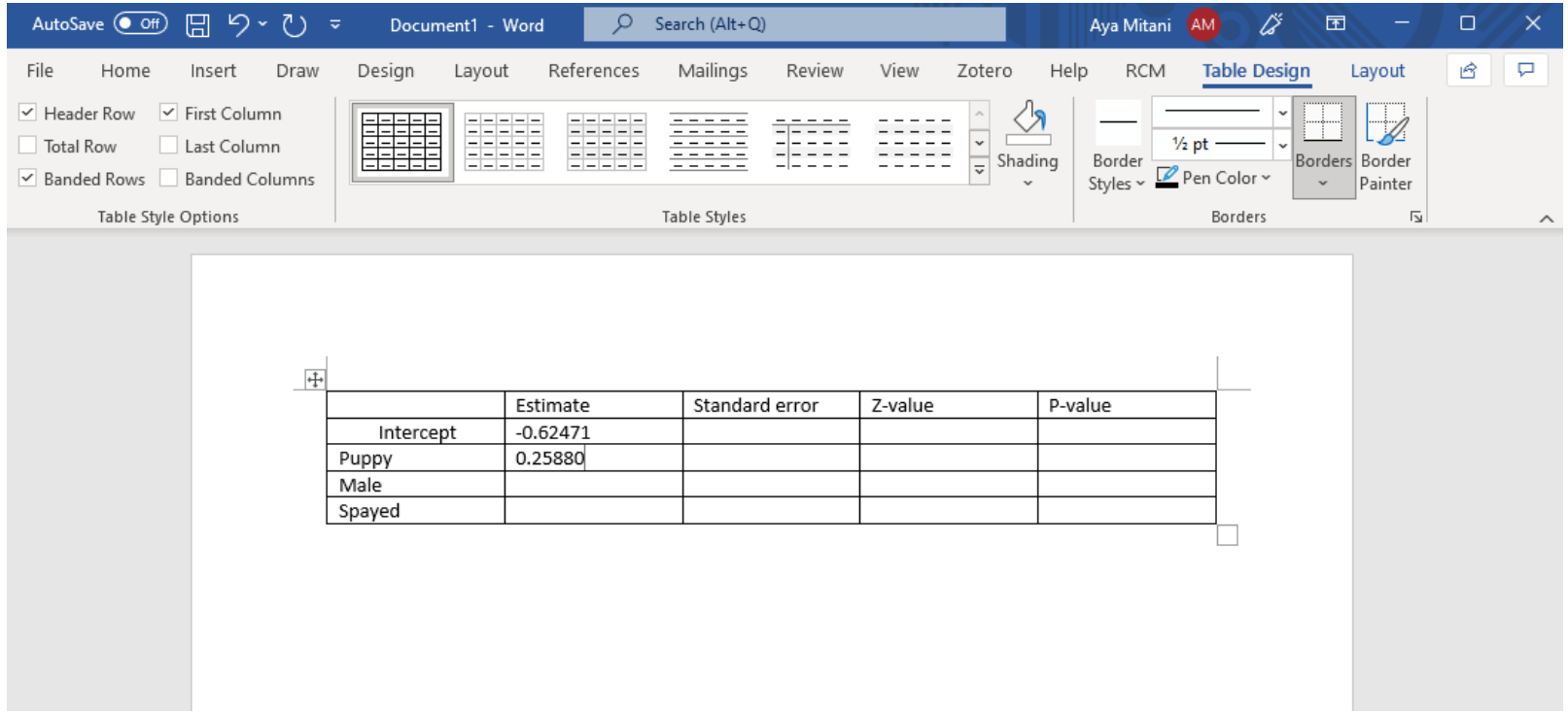
```
bite2 <- bite %>%
  # create new dummy variables
  mutate(Puppy = ifelse(Age < 4, 1, 0),
         Male = ifelse(Gender == "M", 1, 0),
         Spayed = ifelse(SpayNeuter == "true", 1, 0),
         Pitbull = ifelse(Breedclean == "Pit bull", 1, 0))

# logistic regression
glmbite <- glm(Pitbull ~ Puppy + Male + Spayed, data = bite2, family = binomial("logit"))
summary(glmbite)$coef
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.6265040 0.07903863 -7.926555 2.253089e-15
## Puppy        0.2473758 0.06755955  3.661597 2.506478e-04
## Male        -0.2670991 0.07209909 -3.704612 2.117148e-04
## Spayed      -0.6348316 0.06712119 -9.457991 3.139153e-21
```

Are you still doing this?

Copy & paste into Word 😱



	Estimate	Standard error	Z-value	P-value
Intercept	-0.62471			
Puppy	0.25880			
Male				
Spayed				

NOT REPRODUCIBLE!!!!!

`knitr` and `xtable` packages to the rescue!

You wan to

- **minimize** (or eliminate) human error
- use time **efficiently**
- **automate** table creation with updated/new results

```
library(knitr)
```

- `kable()` can create pretty tables from dataframes and matrices
- quick and simple but limited customization ability

```
library(xtable)
```

- `xtable()` converts many R objects into LaTex table code
- Flexible with more customization ability

`knitr` package

Basic `kable()` options

```
sumbite <- summary(glmbite)
class(sumbite$coefficients)
```

```
## [1] "matrix" "array"
```

```
kable(sumbite$coefficients, format = 'html')
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6265040	0.0790386	-7.926555	0.0000000
Puppy	0.2473758	0.0675595	3.661597	0.0002506
Male	-0.2670991	0.0720991	-3.704612	0.0002117
Spayed	-0.6348316	0.0671212	-9.457991	0.0000000

knitr package

More kable() options

```
kable(sumbite$coefficients,
      format = 'html',
      digits = 3,       # specify number of decimal places to show
      col.names = c("Est", "SE", "Z-value", "P-value"), # edit the column names
      align = "c",      # options for alignment are 'l' (left), 'c' (center), 'r' (right) which is t
      caption = "Results from logistic regression analysis") # title of the table
```

Results from logistic regression analysis

	Est	SE	Z-value	P-value
(Intercept)	-0.627	0.079	-7.927	0
Puppy	0.247	0.068	3.662	0
Male	-0.267	0.072	-3.705	0
Spayed	-0.635	0.067	-9.458	0

`knitr` package

`kable()` with LaTex format

```
kable(sumbite$coefficients,
      # use latex format
      format = 'latex',
      # specify number of decimal places to sh
      digits = 3,
      # edit the column names
      col.names = c("Est", "SE", "Z-value", "P-
      # options for alignment are 'l' (left),
      align = "c",
      # title of the table
      caption = "Results from logistic regress
```

\begin{table}

\caption{\label{tab:kable-out}Results from logistic regression analysis} \centering

```
\begin{tabular}[t]{l|c|c|c|c}
\hline
  & Est & SE & Z-value & P-value\
\hline
(Intercept) & -0.627 & 0.079 & -7.927 & 0\
\hline
Puppy & 0.247 & 0.068 & 3.662 & 0\
\hline
Male & -0.267 & 0.072 & -3.705 & 0\
\hline
Spayed & -0.635 & 0.067 & -9.458 & 0\
\hline
\end{tabular}
```

\end{table}

`knitr` package

`kable()` with reStructuredText format

```
kable(sumbite$coefficients,
      format = 'rst',
      digits = 3,       # specify number of decimal places to show
      col.names = c("Est", "SE", "Z-value", "P-value"), # edit the column names
      align = "c",      # options for alignment are 'l' (left), 'c' (center), 'r' (right) which is t
      caption = "Results from logistic regression analysis") # title of the table
```

========== ====== ===== ======= ======= \ Est SE Z-value P-value ========== ====== ===== ======= =======
(Intercept) -0.627 0.079 -7.927 0
Puppy 0.247 0.068 3.662 0
Male -0.267 0.072 -3.705 0
Spayed -0.635 0.067 -9.458 0
========== ====== ===== ======= =======

`xtable` package

Basic `xtable()` with `lm` or `glm` class objects

```
class(glmbite)
```

```
## [1] "glm" "lm"
```

```
xtable(glmbite)
```

```
## % latex table generated in R 4.1.1 by xtable 1.8-4 package
## % Sun Nov 28 21:06:57 2021
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & z value & Pr($>$$|$z$|$) \\
##   \hline
## (Intercept) & -0.6265 & 0.0790 & -7.93 & 0.0000 \\
##   Puppy & 0.2474 & 0.0676 & 3.66 & 0.0003 \\
##   Male & -0.2671 & 0.0721 & -3.70 & 0.0002 \\
##   Spayed & -0.6348 & 0.0671 & -9.46 & 0.0000 \\
##    \hline
## \end{tabular}
## \end{table}
```

`xtable` package

Some `xtable()` customizations

```
table1 <-
xtable(glmbite,
       # title of the table
       caption = "Results from logistic regres
       # label for referencing
       label = "table:logreg",
       # number of decimal places
       digits = 2
       )
print.xtable(table1)
```

- Include in Rmarkdown file with PDF output
- Copy & paste into LaTex editor, Overleaf

```
## % latex table generated in R 4.1.1 by xtable 1.8-4
## % Sun Nov 28 21:06:57 2021
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & z value & Pr($>$$|$z$|$
##   \hline
## (Intercept) & -0.63 & 0.08 & -7.93 & 0.00 \\
##   Puppy & 0.25 & 0.07 & 3.66 & 0.00 \\
##   Male & -0.27 & 0.07 & -3.70 & 0.00 \\
##   Spayed & -0.63 & 0.07 & -9.46 & 0.00 \\
##    \hline
## \end{tabular}
## \caption{Results from logistic regression analysis
## \label{table:logreg}
## \end{table}
```

`xtable` package

`print.table()` has even more options

Save LaTex code as text file

```
print.xtable(table1, file = here("results", "table1.txt"))
```

`xtable` package

`print.table()` has even more options

Create HTML table

```
print.xtable(table1, type = "html", caption.placement = "top")
```

```
## <!-- html table generated in R 4.1.1 by xtable 1.8-4 package -->
## <!-- Sun Nov 28 21:06:57 2021 -->
## <table border=1>
## <caption align="top"> Results from logistic regression analysis </caption>
## <tr> <th>  </th> <th> Estimate </th> <th> Std. Error </th> <th> z value </th> <th> Pr(&gt;|z|) </th>  <
##   <tr> <td align="right"> (Intercept) </td> <td align="right"> -0.63 </td> <td align="right"> 0.08 </td
##   <tr> <td align="right"> Puppy </td> <td align="right"> 0.25 </td> <td align="right"> 0.07 </td> <td a
##   <tr> <td align="right"> Male </td> <td align="right"> -0.27 </td> <td align="right"> 0.07 </td> <td a
##   <tr> <td align="right"> Spayed </td> <td align="right"> -0.63 </td> <td align="right"> 0.07 </td> <td
##    </table>
```

`xtable` package

`print.table()` has even more options

Create HTML table

Results from logistic regression analysis

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-0.63	0.08	-7.93	0.00
Puppy	0.25	0.07	3.66	0.00
Male	-0.27	0.07	-3.70	0.00
Spayed	-0.63	0.07	-9.46	0.00

`texreg` package for multiple models

I want to compare results from these three models

```
# m1: x = Puppy
m1 <- glm(Pitbull ~ Puppy, data = bite2, family = binomial("logit"))
# m2: x = Puppy & Male
m2 <- glm(Pitbull ~ Puppy + Male, data = bite2, family = binomial("logit"))
# m3: x = Puppy & Male & Spayed
m3 <- glm(Pitbull ~ Puppy + Male + Spayed, data = bite2, family = binomial("logit"))
```

`texreg` package for multiple models

`screenreg()` for text output to the R console

```
screenreg(list(m1, 
```

```
##
## ==========================================================
##                   Model 1       Model 2       Model 3
## ----------------------------------------------------------
## (Intercept)         -1.15 ***     -0.97 ***     -0.63 ***
##                     (0.05)        (0.07)        (0.08)
## Puppy                0.29 ***      0.29 ***      0.25 ***
##                     (0.07)        (0.07)        (0.07)
## Male                              -0.26 ***     -0.27 ***
##                                   (0.07)        (0.07)
## Spayed                                          -0.63 ***
##                                                 (0.07)
## ----------------------------------------------------------
## AIC                5449.95       5439.24       5350.44
## BIC                5462.85       5458.59       5376.24
## Log Likelihood    -2722.97      -2716.62      -2671.22
## Deviance           5445.95       5433.24       5342.44
## Num. obs.          4673          4673          4673
## ==========================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

`texreg` package for multiple models

Display 95% confidence intervals

```
screenreg(list(m1, m
          ci.force =
```

```
##
## ==============================================================
##                  Model 1         Model 2         Model 3
## --------------------------------------------------------------
## (Intercept)        -1.15 *         -0.97 *         -0.63 *
##                 [-1.25; -1.05]  [-1.11; -0.83]  [-0.78; -0.47]
## Puppy               0.29 *          0.29 *          0.25 *
##                 [ 0.16;  0.42]  [ 0.16;  0.42]  [ 0.11;  0.38]
## Male                               -0.26 *         -0.27 *
##                                 [-0.40; -0.12]  [-0.41; -0.13]
## Spayed                                             -0.63 *
##                                                 [-0.77; -0.50]
## --------------------------------------------------------------
## AIC              5449.95         5439.24         5350.44
## BIC              5462.85         5458.59         5376.24
## Log Likelihood  -2722.97        -2716.62        -2671.22
## Deviance         5445.95         5433.24         5342.44
## Num. obs.        4673            4673            4673
## ==============================================================
## * Null hypothesis value outside the confidence interval.
```

`texreg` package for multiple models

`texreg()` for LaTex output

```
texreg(list(m1, m2,
       caption = "D
```

```
##
## \begin{table}
## \begin{center}
## \begin{tabular}{l c c c}
## \hline
##  & Model 1 & Model 2 & Model 3 \\
## \hline
## (Intercept)    & $-1.15^{***}$ & $-0.97^{***}$ & $-0.63^{***}$ \\
##                & $(0.05)$      & $(0.07)$      & $(0.08)$      \\
## Puppy          & $0.29^{***}$  & $0.29^{***}$  & $0.25^{***}$  \\
##                & $(0.07)$      & $(0.07)$      & $(0.07)$      \\
## Male           &               & $-0.26^{***}$ & $-0.27^{***}$ \\
##                &               & $(0.07)$      & $(0.07)$      \\
## Spayed         &               &               & $-0.63^{***}$ \\
##                &               &               & $(0.07)$      \\
## \hline
## AIC            & $5449.95$     & $5439.24$     & $5350.44$     \\
## BIC            & $5462.85$     & $5458.59$     & $5376.24$     \\
## Log Likelihood & $-2722.97$    & $-2716.62$    & $-2671.22$    \\
## Deviance       & $5445.95$     & $5433.24$     & $5342.44$     \\
## Num. obs.      & $4673$        & $4673$        & $4673$        \\
```

`texreg` package for multiple models

`htmlreg()` for HTML output

```
htmlreg(list(m1, m2, m3), caption = "Dog bite models")
```

```
## <table class="texreg" style="margin: 10px auto;border-collapse: collapse;border-spacing: 0px;caption-si
## <caption>Dog bite models</caption>
## <thead>
## <tr>
## <th style="padding-left: 5px;padding-right: 5px;"> </th>
## <th style="padding-left: 5px;padding-right: 5px;">Model 1</th>
## <th style="padding-left: 5px;padding-right: 5px;">Model 2</th>
## <th style="padding-left: 5px;padding-right: 5px;">Model 3</th>
## </tr>
## </thead>
## <tbody>
## <tr style="border-top: 1px solid #000000;">
## <td style="padding-left: 5px;padding-right: 5px;">(Intercept)</td>
## <td style="padding-left: 5px;padding-right: 5px;">-1.15<sup>***</sup></td>
## <td style="padding-left: 5px;padding-right: 5px;">-0.97<sup>***</sup></td>
## <td style="padding-left: 5px;padding-right: 5px;">-0.63<sup>***</sup></td>
## </tr>
## <tr>
```

HTML table from `htmlreg()`

	Model 1	Model 2	Model 3
(Intercept)	-1.15***	-0.97***	-0.63***
	(0.05)	(0.07)	(0.08)
Puppy	0.29***	0.29***	0.25***
	(0.07)	(0.07)	(0.07)
Male		-0.26***	-0.27***
		(0.07)	(0.07)
Spayed			-0.63***
			(0.07)
AIC	5449.95	5439.24	5350.44
BIC	5462.85	5458.59	5376.24
Log Likelihood	-2722.97	-2716.62	-2671.22
Deviance	5445.95	5433.24	5342.44
Num. obs.	4673	4673	4673

***p < 0.001; **p < 0.01; *p < 0.05

Dog bite models

Writing R functions

I want to show Odds Ratio (95% CI) for the table

In logistic regression,

$$\text{OR} = \exp(\hat{\beta})$$

```
OR <- exp(beta)
```

$$95\%\ \text{CI} = \exp\left[\hat{\beta} \pm 1.96 \times \text{SE}(\hat{\beta})\right]$$

```
ORlcl <- exp(beta - 1.96 * se_beta)
ORucl <- exp(beta + 1.96 * se_beta)
```

Are you still doing this?

```
# OR for m1
OR_m1 <- exp(summary(m1)$coefficients[,1])
ORlcl_m1 <- exp(summary(m1)$coefficients[,1] - 1.96 * summary(m1)$coefficients[,2])
ORucl_m1 <- exp(summary(m1)$coefficients[,1] + 1.96 * summary(m1)$coefficients[,2])

# OR for m2
OR_m2 <- exp(summary(m2)$coefficients[,1])
ORlcl_m2 <- exp(summary(m2)$coefficients[,1] - 1.96 * summary(m2)$coefficients[,2])
ORucl_m2 <- exp(summary(m2)$coefficients[,1] + 1.96 * summary(m2)$coefficients[,2])

# OR for m3
OR_m3 <- exp(summary(m3)$coefficients[,1])
ORlcl_m3 <- exp(summary(m3)$coefficients[,1] - 1.96 * summary(m3)$coefficients[,2])
ORucl_m3 <- exp(summary(m3)$coefficients[,1] + 1.96 * summary(m3)$coefficients[,2])
```

NOT REPRODUCIBLE!!!

Writing R functions

A function to create OR (95%) with options for α level and number of decimal places

```
# coef is the vector of estimates
# se is the vector of standard errors
# siglevel is the significance (alpha) level
# roundto is the number of decimal places
OR_95CI <- function(coef, se, siglevel, roundto){
  q <- 1 - siglevel / 2
  OR <- exp(coef)
  ORlcl <- exp(coef - qnorm(q) * se)
  ORucl <- exp(coef + qnorm(q) * se)
  ORresult <- paste0(format(round(OR, roundto), nsmall=roundto), " (", format(round(ORlcl, roundto)
  return(ORresult)
}
orout1 <- OR_95CI(summary(m1)$coef[,1], summary(m1)$coef[,2], 0.05, 2)
orout1
```

```
## [1] "0.32 (0.29, 0.35)" "1.34 (1.18, 1.53)"
```

Using loops in R

Lists are very useful in R

```
m <- list()
m[[1]] <- glm(Pitbull ~ Puppy, data = bite2, family = binomial("logit"))
m[[2]] <- glm(Pitbull ~ Puppy + Male, data = bite2, family = binomial("logit"))
m[[3]] <- glm(Pitbull ~ Puppy + Male + Spayed, data = bite2, family = binomial("logit"))
msum <- lapply(m, summary)
msum[[1]]$coef
```

```
##               Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) -1.1509406 0.05055684 -22.765281 1.012824e-114
## Puppy        0.2942141 0.06667320   4.412779  1.020521e-05
```

```
msum[[1]]$coef[,1]
```

```
## (Intercept)       Puppy
##  -1.1509406   0.2942141
```

Using loops in R

Lists are very useful in R

```
orvec <- list()
for(i in 1:3) orvec[[i]] <- OR_95CI(msum[[i]]$coef[,1], msum[[i]]$coef[,2], 0.05, 2)
orvec
```

```
## [[1]]
## [1] "0.32 (0.29, 0.35)" "1.34 (1.18, 1.53)"
##
## [[2]]
## [1] "0.38 (0.33, 0.43)" "1.34 (1.18, 1.53)" "0.77 (0.67, 0.89)"
##
## [[3]]
## [1] "0.53 (0.46, 0.62)" "1.28 (1.12, 1.46)" "0.77 (0.66, 0.88)"
## [4] "0.53 (0.46, 0.60)"
```

Back to `xtable`

```r
varnames <- c("Intercept", "Puppy", "Male", "Spayed") # create vector of variable names
ORout <- data.frame(varnames, c(orvec[[1]], rep(NA, 2)), c(orvec[[2]], NA), orvec[[3]]) # create da
names(ORout) <- c("Variable", "M1", "M2", "M3") # give column names
ORtable <- xtable(ORout[-1,], caption = "OR (95% CI)") # create xtable object but remove Intercept
print(ORtable, include.rownames = FALSE) # print xtable but remove row numbers
```

```
## % latex table generated in R 4.1.1 by xtable 1.8-4 package
## % Sun Nov 28 21:06:58 2021
## \begin{table}[ht]
## \centering
## \begin{tabular}{llll}
##   \hline
## Variable & M1 & M2 & M3 \\
##   \hline
## Puppy & 1.34 (1.18, 1.53) & 1.34 (1.18, 1.53) & 1.28 (1.12, 1.46) \\
##   Male &  & 0.77 (0.67, 0.89) & 0.77 (0.66, 0.88) \\
##   Spayed &  &  & 0.53 (0.46, 0.60) \\
##    \hline
## \end{tabular}
## \caption{OR (95% CI)}
## \end{table}
```

Output of `xtable`

Variable	M1	M2	M3
Puppy	1.34 (1.18, 1.53)	1.34 (1.18, 1.53)	1.28 (1.12, 1.46)
Male		0.77 (0.67, 0.89)	0.77 (0.66, 0.88)
Spayed			0.53 (0.46, 0.60)

Table 2: OR (95% CI)

R function to R package

I want to keep using `OR_95CI()` function for my other projects!

```
OR_95CI <- function(coef, se, siglevel, roundto){
  q <- 1 - siglevel / 2
  OR <- exp(coef)
  ORlcl <- exp(coef - qnorm(q) * se)
  ORucl <- exp(coef + qnorm(q) * se)
  ORresult <- paste0(format(round(OR, roundto), nsmall=roundto), " (", format(round(ORlcl, roundto)
  return(ORresult)
}
```

- Save the R script on my computer and copy & paste it for every project → NOT REPRODUCIBLE!!!
- I have some other related R functions that I often use
- Turn them into a **R package** and save it on my **GitHub** account

Creating R packages

Use `devtools` package

```
library(devtools)
```

Major steps (More details here)

1. Open R Studio
   - New Project > New Directory > R Package > Enter info > Create Project
2. Edit your package
   - Each function should be saved in its own file
   - Write the package description and document functions
   - Include some data
   - Write a vignette
3. Create a new repository in GitHub
   - Repo name = package name
4. Connect to GitHub
5. Pull + Commit + Push
6. Use/share package with `install_github()`

Using my R package

Install and load package from GitHub

```r
library(devtools) # load devtools package
devtools::install_github("ayamitani/oddsratio") # install package from git repo
library(oddsratio) # load package
```

```r
# new logistic regression model
m4 <- glm(Pitbull ~ Agenum, data = bite2, family = binomial("logit"))
# save table of coefficients and standard errors from summary output
m4coef <- summary(m4)$coef
# apply function
OR_95CI(m4coef[,1], m4coef[,2], 0.05, 3)
```

```
## [1] "0.576 (0.518, 0.641)" "0.899 (0.880, 0.919)"
```

Publishing R package on CRAN

- The Comprehensive R Archive Network (CRAN) has ~18,000 R packages
- Package on CRAN can be downloaded with `install.packages()`
- Publishing your R package on CRAN requires a lot more work than making it available on GitHub
- Going through all the necessary steps, your package will be more **accessible** to a wider audience

Publishing R package on CRAN

- Read the CRAN Repository Policy and Checklist for CRAN submissions
- Follow @CRANPolicyWatch on Twitter
- Submit your package using the submission form
- CRAN maintainer will review your package
  - Review can take up to 5 days (check status here)
  - You package may get rejected $\rightarrow$ revise, recheck and resubmit
- Write an article describing your package
  - Your own website
  - Research paper
  - Software journals (Journal of Statistical Software, The R Journal, Journal of Open Source Software)

Future proof your research

- Record your R session information

```
sessionInfo()
```

- Set seed for random generation

```
set.seed()
```

- Save everything as text files (`.txt`)
- Make your code **human** readable
  - `formatR` package is useful

```
library(formatR)
```

- Document, document, document...

Final thoughts

- I'm still learning
- Normalize making research reproducible
  - **Teach** the tools
  - **Train** the new generation of researchers
  - **Practice** reproducibility
- Contribute more to replicating research
  - Change the institutional culture from bottom up!

This presentation is reproducible!

Slides created via the R packages:

xaringan
gadenbuie/xaringanthemer

Very new `sjPlot` package can show odds ratios

`tab_model()` can create very nice tables with OR (95% CI)

```
#tab_model(m1, m2, m3, show.p = FALSE)
```