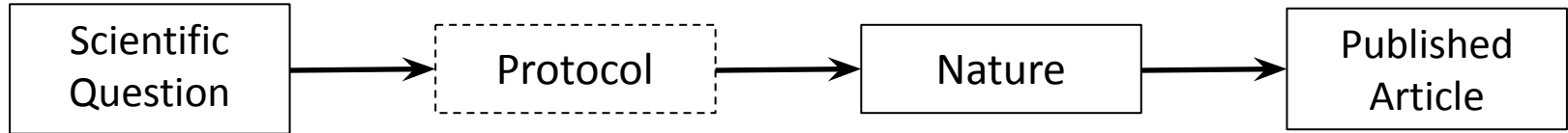


Data Wrangling in R

Reproducible Research

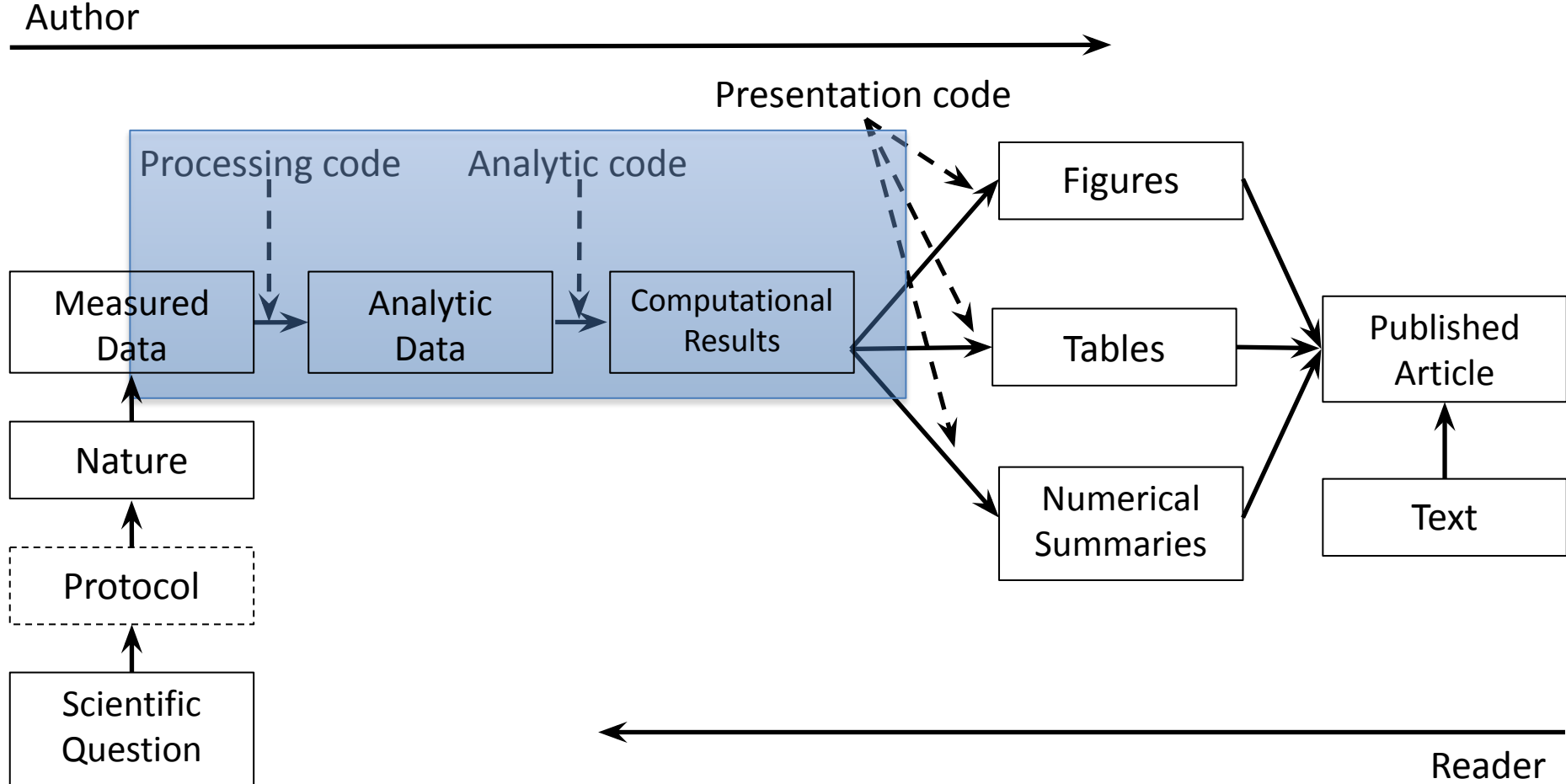
What is Reproducible Research?

Author



Reader

What is Reproducible Research?



Uh...who cares?

FIXING SCIENCE

Most science research findings are false. Here's how we can change that



All

Images

Videos

News

Shopping

More

Settings

Tools

About 176,000,000 results (0.43 seconds)

Most Scientific Findings Are Wrong or Useless - Reason.com

reason.com/archives/2016/08/26/most-scientific-results-are-wrong-or-useless

Aug 26, 2016 - ScientistYanlevDreamstime Yanlev/Dreamstime"Science, the pride of modernity, our one source of objective knowledge, is in deep trouble.

PLOS Medicine: Why Most Published Research Findings Are False

journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

by JPA Ioannidis - 2005 - Cited by 4846 - Related articles

Aug 30, 2005 - Moreover, for many current **scientific** fields, claimed research findings ... Citation: Ioannidis JPA (2005) Why **Most** Published Research Findings Are False. what might have gone **wrong** with their data, analyses, and results.

Is Most Published Research Wrong? - YouTube

<https://www.youtube.com/watch?v=42QuXLuch3Q>

Aug 11, 2016 - Uploaded by Veritasium

Why **Most** Published Research Findings Are False: The problem with the approach to **science** is that ...

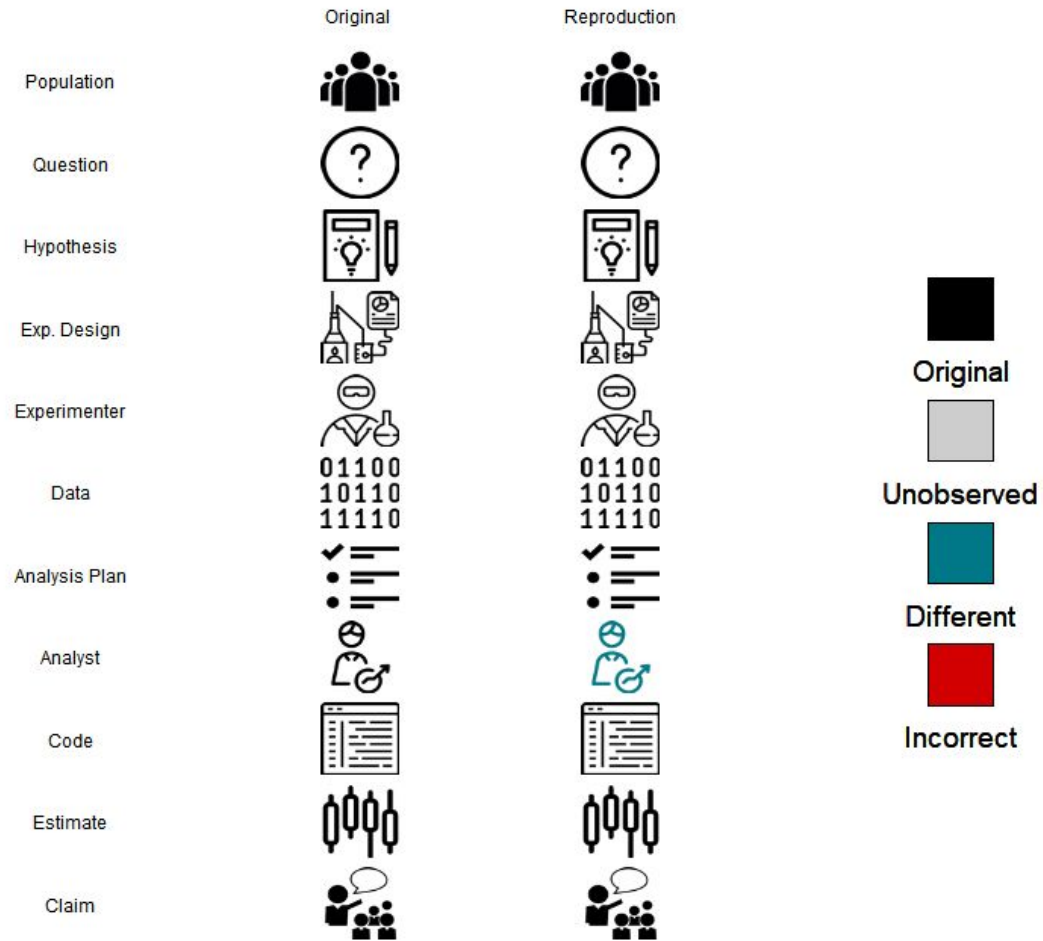
Believe It Or Not, Most Published Research Findings Are Probably ...

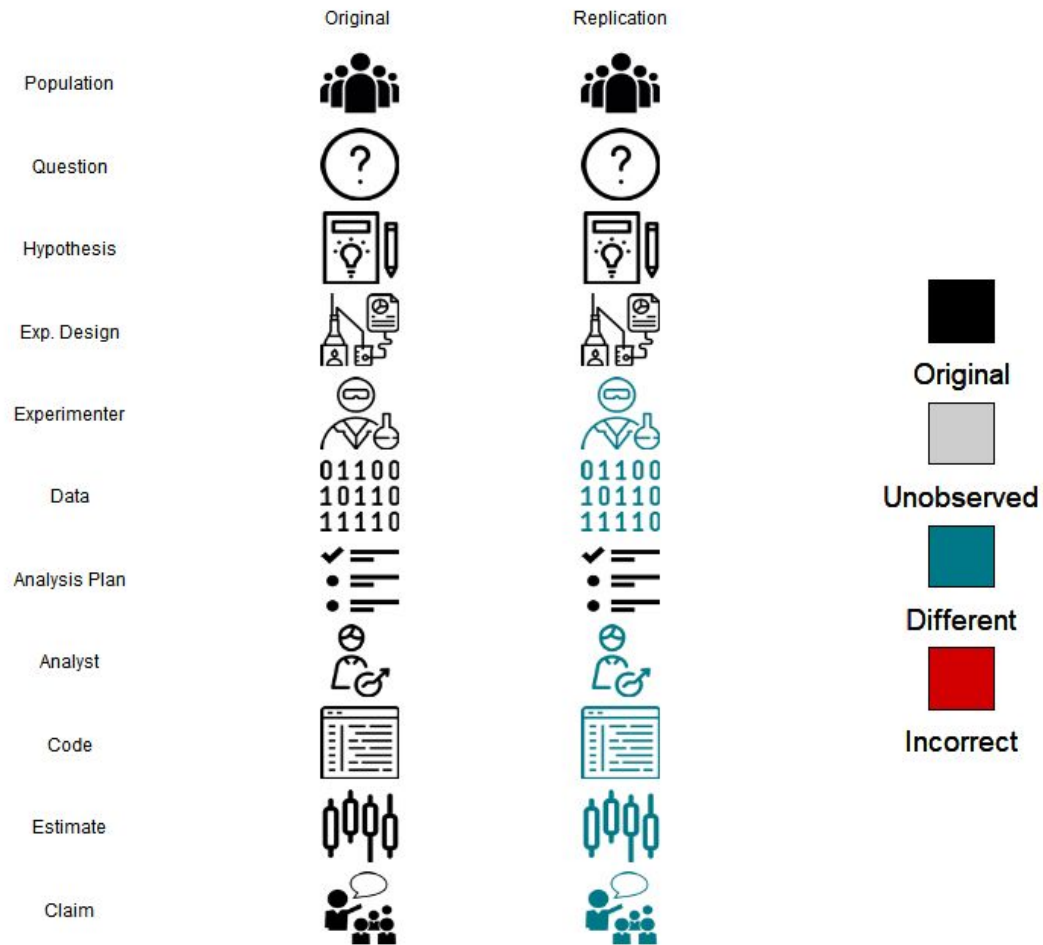
bigthink.com/.../believe-it-or-not-most-published-research-findings-are-probably-fals...

Ten years ago, a researcher claimed **most** published research findings are false; ... of the Internet has worked wonders for the public's access to **science**, but this ... the case, experiments are underpowered,

176,000,000!

Reproduce & replicate





When human harm isn't involved

A few things that would reduce stress around reproducibility/replicability in science

POWER POSE STUDY

👤 Jeff Leek 📅 2017/11/21

I was listening to the Effort Report Episode on [The Messy Execution of Reproducible Research](#) where they were discussing the piece about [Amy Cuddy in the New York Times](#). I think both the article and the podcast did a good job of discussing the nuances of the importance of reproducibility and the challenges of the social interactions around this topic. After listening to the podcast I realized that I see a lot of posts about reproducibility/replicability, but many of them are focused on the technical side. So I started to think about compiling a list of more cultural things we can do to reduce the stress/pressure around the reproducibility crisis.

I'm sure others have pointed these out in other places but I am procrastinating writing something else so I'm writing these down while I'm thinking about them :).

1. **We can define what we mean by “reproduce” and “replicate”** Different fields have different definitions of the words *reproduce* and *replicate*. If you are publishing a new study we now have an [R package](#) that you can use to create figures that show what changed and what was the same between the original study and your new work. Defining concretely what was the same and different will reduce some of the miscommunication about what a reproducibility/replicability study means.

When human harm could happen

From the article:

Cancer trial errors revealed

2006 Anil Potti, a cancer geneticist at Duke University in Durham, North Carolina, and others file patent applications on the idea of using gene-expression data to predict sensitivity to cancer drugs. Potti is first author on a paper in *Nature Medicine*¹.

2007 Potti is last author on a paper in the *Journal of Clinical Oncology (JCO)*². Duke begins three clinical trials to test Potti's predictors in patients with breast or lung cancer.

SEPTEMBER 2009 Keith Baggerly and Kevin Coombes, statisticians at the University of Texas M. D. Anderson Cancer Centre in Houston, publish a paper in *Annals of Applied Statistics*³ stating that they could not replicate Potti's claims. Duke suspends the trials and asks a review panel to investigate.

NOVEMBER 2009 Potti places data underlying the *JCO* paper online. Baggerly writes to Sally Kornbluth, Duke vice-dean for research, and Michael Cuffe, Duke vice-president for medical affairs, to point out differences from raw data.

DECEMBER 2009 An unredacted copy of the report by Duke's review panel, later obtained by *Nature*, shows that the panel replicated Potti's claims using his data, but were unaware that those data contained discrepancies.

JANUARY 2010 Duke restarts clinical trials.

JULY 2010 *The Cancer Letter* reveals that Potti made false claims about his CV. Trials are suspended and an investigation begins. Harold Varmus, director of the National Cancer Institute in Bethesda, Maryland, asks the Institute of Medicine to review Duke's trials.

NOVEMBER 2010 *JCO* paper is retracted. Duke closes the trials permanently. Potti resigns.

DECEMBER 2010 Institute of Medicine study begins, but will now focus more generally on criteria for genomics predictor.

JANUARY 2011 *Nature Medicine* paper is retracted.

<http://www.nature.com/news/2011/110111/full/469139a/box/1.html>

FRAUDULENT/
MISCONDUCT
IN A CLINICAL
TRIAL

1. Code + documentation
2. Versions of software
3. Data provenance



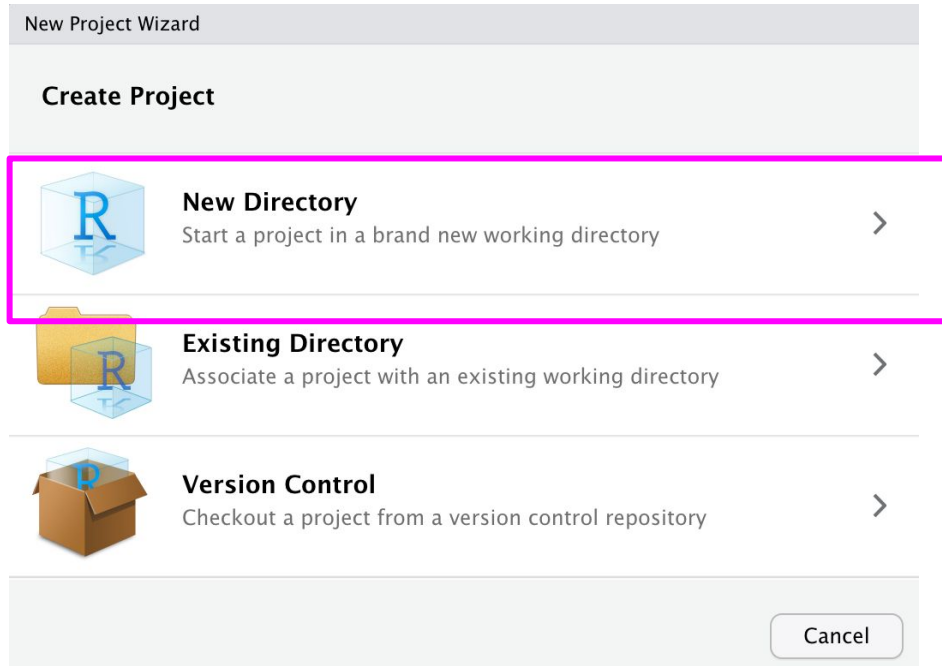
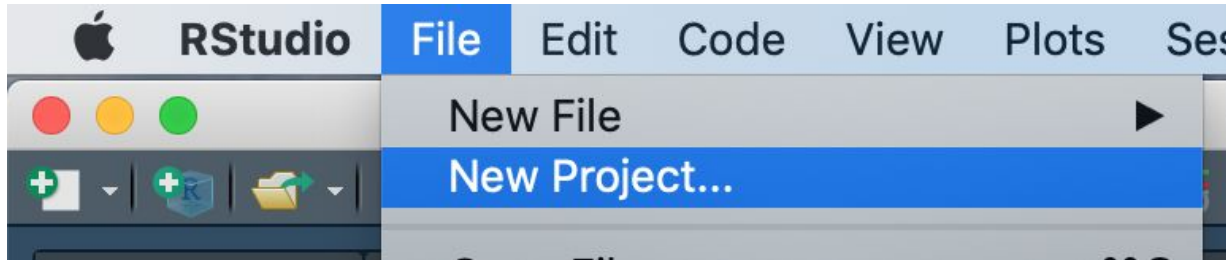
<https://tenor.com/view/struggle-cant-move-over-it-hard-no-gif-4734482>

Your closest collaborator is
you six months ago, but you
don't reply to emails

- Karl Broman

(http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf)

RStudio Projects



Back

Project Type



New Project



Back

Create New Project



Directory name:

SISBID

Create project as subdirectory of:

~/Desktop

Browse...

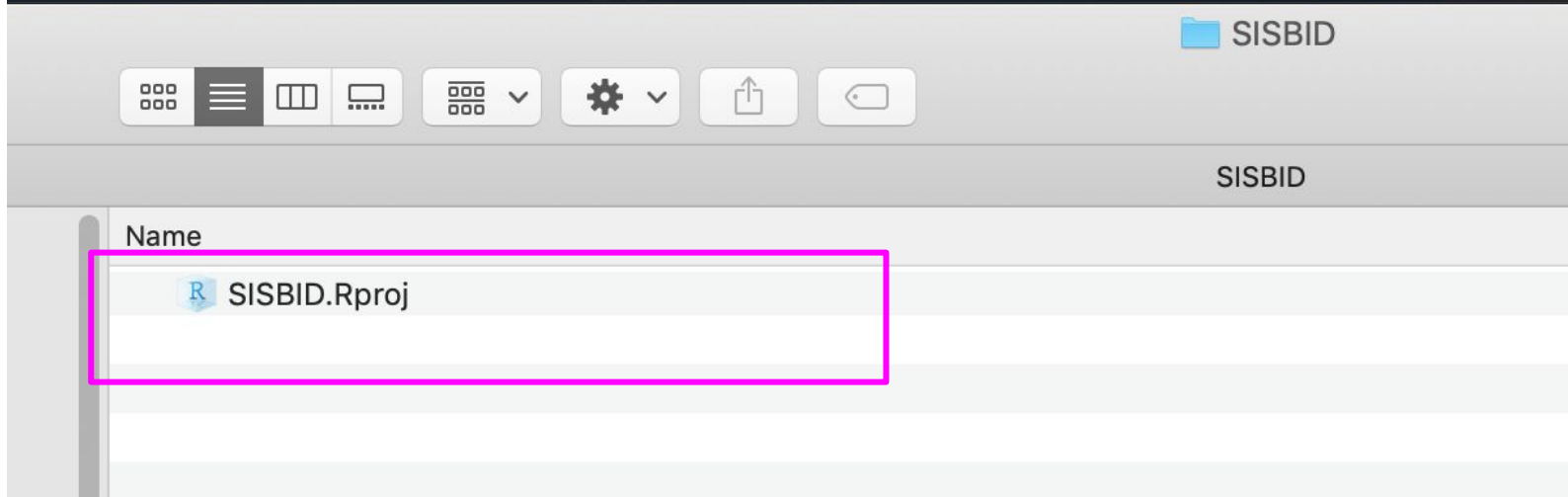
Create a git repository

Use renv with this project

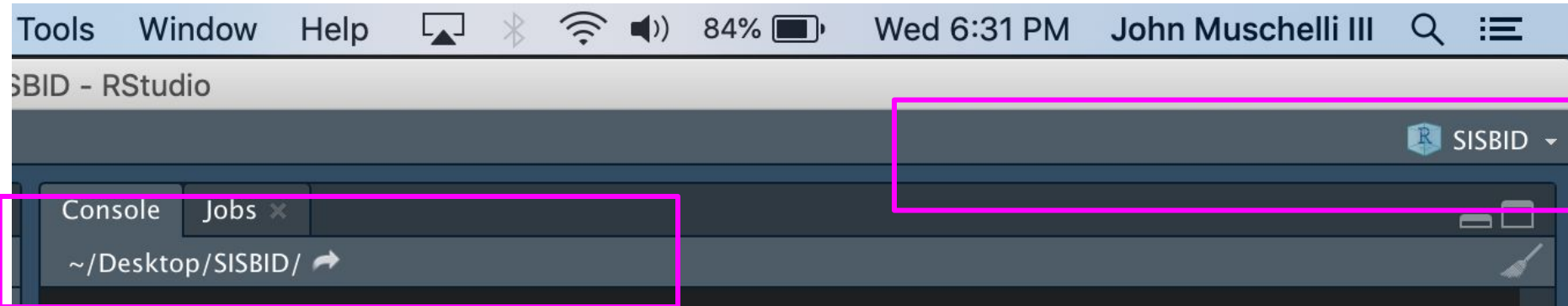
Open in new session

Create Project

Cancel



Double click on the Rproj file - opens RStudio



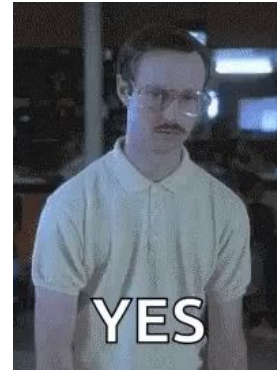
RStudio Projects

- Rproj - Open to directory of Project
- Everything is **relative** to working directory
- Zip up whole folder and send to someone else

```
read_csv("C:/terrible/path/john/blah.csv")
```



```
read_csv("data/blah.csv")
```



Couldn't you just re-run that
code with the
[latest/different/best]
parameters?

- Every collaborator/PI

The magic of Markdown

- bullets
- **bold**
- *italics*
- [links](https://google.com)
- or run inline `r code`



- bullets
- **bold**
- *italics*
- links
- or run inline `r code`

https://rmarkdown.rstudio.com/authoring_basics.html

```
# Introduction
```

Here is some background you need to know:

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam iaculis enim ut enim viverra molestie. In lacinia aliquet urna, nec vulputate quam congue et. Maecenas porta mauris sem, nec laoreet sapien tincidunt non. Integer sit amet consequat neque, non iaculis ligula.

```
# Hypothesis
```

Pellentesque molestie erat nec elit efficitur, sit amet sodales erat viverra. Mauris sed commodo eros, ac volutpat sem. Morbi convallis leo et dui cursus, eu suscipit turpis efficitur.

```
# Section 1 code and results
```

First I will run this.

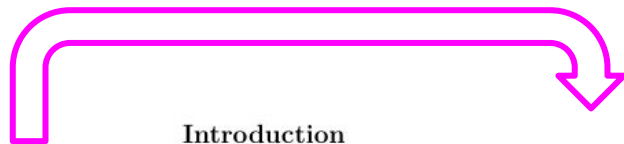
```
```{r}
print("Hello world")
print("Yup, this is important")
```
```

The output of which is consistent with my hypothesis.

```
# conclusion
```

I can move on to the next part of my project

.Rmd
document



PDF, HTML or Word
document

Introduction

Here is some background you need to know:

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam iaculis enim ut enim viverra molestie. In lacinia aliquet urna, nec vulputate quam congue et. Maecenas porta mauris sem, nec laoreet sapien tincidunt non. Integer sit amet consequat neque, non iaculis ligula.

Hypothesis

Pellentesque molestie erat nec elit efficitur, sit amet sodales erat viverra. Mauris sed commodo eros, ac volutpat sem. Morbi convallis leo et dui cursus, eu suscipit turpis efficitur.

Section 1 code and results

First I will run this.

```
print("Hello world")

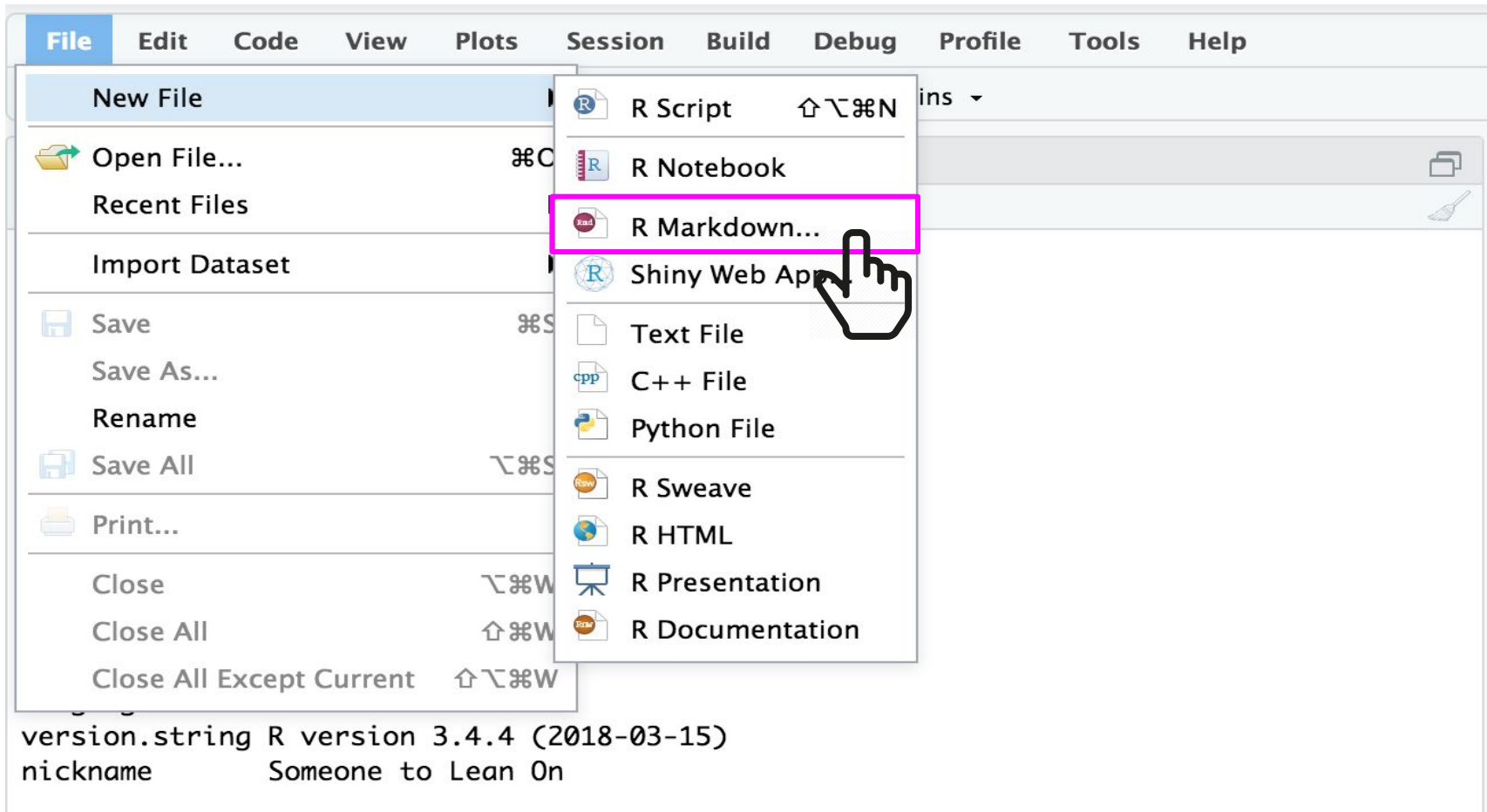
## [1] "Hello world"
print("Yup, this is important")
```

```
## [1] "Yup, this is important"
```

The output of which is consistent with my hypothesis.

Conclusion

I can move on to the next part of my project



Install Required Packages



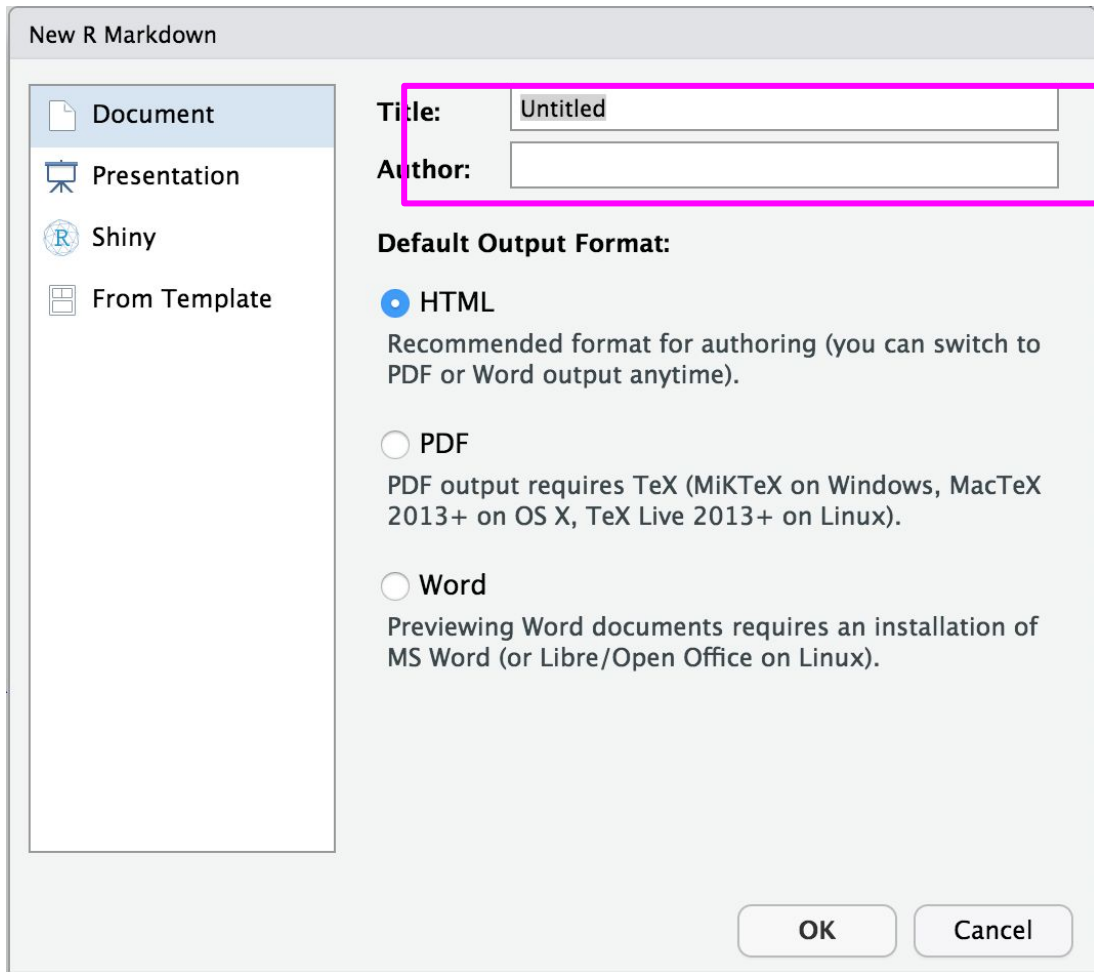
Creating R Markdown documents requires updated versions of the following packages: evaluate, highr, markdown, yaml, htmltools, caTools, bitops, knitr, jsonlite, base64enc, rprojroot, rmarkdown.

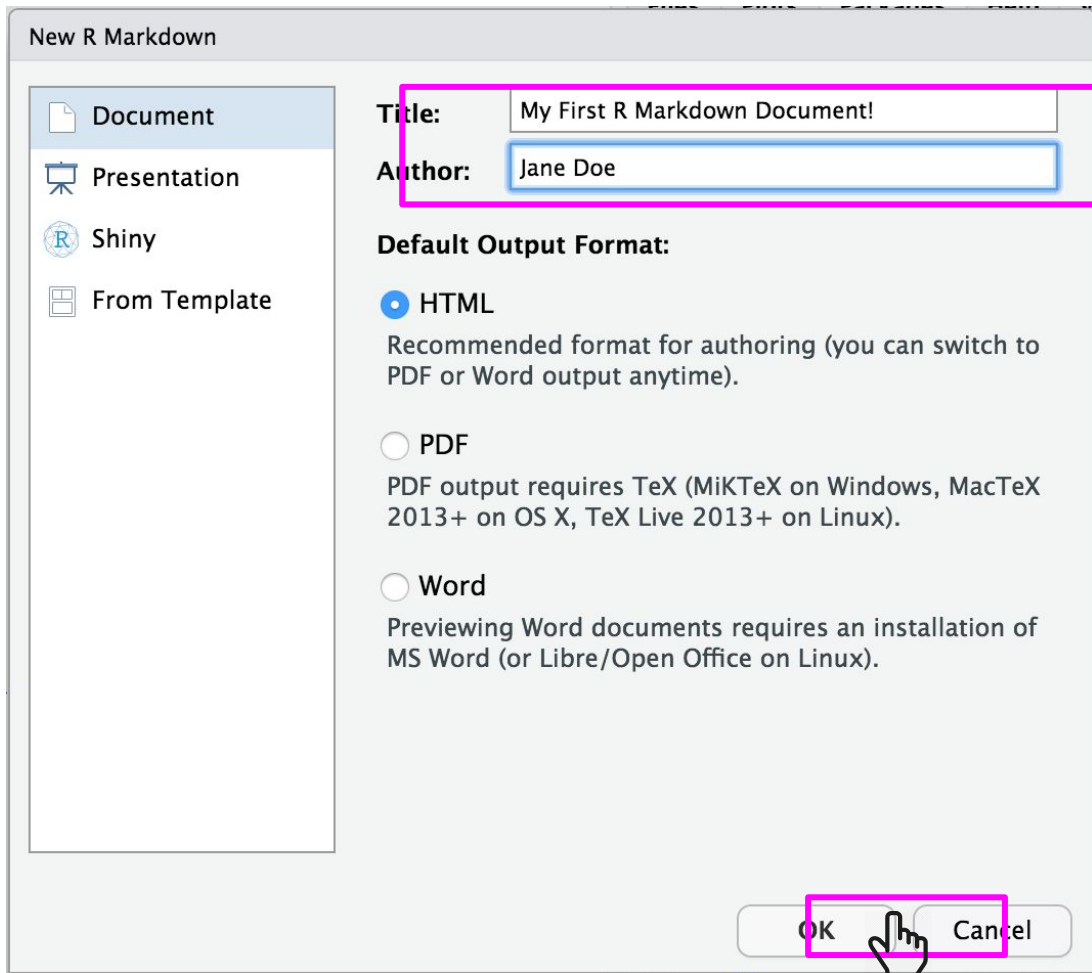
Do you want to install these packages now?

Yes

No







```
1 ---
2 title: "My First R Markdown Document!"
3 author: "Jane Doe"
4 date: "5/31/2018"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word
15 documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
16
17 When you click the Knit button a document will be generated that includes both content as well as the
18 output of any embedded R code chunks within the document. You can embed an R code chunk like this:
19
20 ```{r cars}
21 summary(cars)
22 ```
23
24 ## Including Plots
25
26 You can also embed plots, for example:
27
28 ```{r pressure, echo=FALSE}
29 plot(pressure)
30 ```
```

YAML - not markdown,
Spaces matter, usually goes
option: value

TEXT

CODE CHUNK

```
1 ---
2 title: "My First R Markdown Document!"
3 author: "Jane Doe"
4 date: "5/31/2018"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word
15 documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
16
17 When you click the Knit button a document will be generated that includes both content as well as the
18 output of any embedded R code chunks within the document. You can embed an R code chunk like this:
19
20 ```{r cars}
21 summary(cars)
22 ```
23
24 ## Including Plots
25
26 You can also embed plots, for example:
27
28 ```{r pressure, echo=FALSE}
29 plot(pressure)
30 ```
```

2:1 # My First R Markdown Document! R Markdown

Save File - Untitled1

File name:

> / > cloud > project



..



.Rhistory

0 B

May 30, 2018, 12:34 PM



project.Rproj

205 B

May 31, 2018, 3:35 PM

New Folder

Save

Cancel



My First R Markdown Document!

Jane Doe

5/31/2018

R Markdown

Text section rendered as formatted text

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

Code rendered as the input code AND the output of running the code chunk

```
##      speed      dist
## Min.   : 4.0    Min.   : 2.00
## 1st Qu.:12.0    1st Qu.: 26.00
## Median :15.0    Median : 36.00
## Mean   :15.4    Mean    : 42.98
## 3rd Qu.:19.0    3rd Qu.: 56.00
## Max.   :25.0    Max.    :120.00
```

Including Plots

You can also embed plots, for example:

test_document.Rmd x

Insert Run

```
1 ---
2 title: "My First R Markdown Document!"
3 author: "Jane Doe"
4 date: "5/31/2018"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for
15 authoring HTML, PDF, and MS Word documents. For more details on using R Markdown
16 see <http://rmarkdown.rstudio.com>.
```

17:1 R Markdown R Markdown

Console Terminal x

/cloud/project/

> |

Environment History Connections

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

/ > cloud > project

| | Name | Size | Modified |
|--------------------------|--------------------|----------|------------------------|
| | .. | | |
| <input type="checkbox"/> | .Rhistory | 0 B | May 30, 2018, 12:34 PM |
| <input type="checkbox"/> | project.Rproj | 205 B | Jun 1, 2018, 8:33 AM |
| <input type="checkbox"/> | test_document.html | 741.2 KB | May 31, 2018, 3:57 PM |
| <input type="checkbox"/> | test_document.Rmd | 854 B | May 31, 2018, 3:57 PM |

```
... {r eval}
... eval=
```

eval = FALSE will show but not run the code

```
... {r ec}
... echo=
```

echo = FALSE will not show but run the code

Using both set to FALSE will hide the code and will not run the code!

The default is TRUE - do not need to write it out if you want code to be shown and run.

Biiconductor_intro.Rmd x Data_Subsetting_Part1.R x Data_Cleaning.Rmd x

Knit on Save | ABC | Knit | Run | Outline

Environment History Connection
R | Global Environment | 99 MiB

```
60
61 ```{r}
62 p = problems(ufo)
63 p
64 ```
65
66 ## Let's just drop those problematic rows for now.
67
68 Though you would usually want to check them!
69
70 ```{r}
71 ufo = ufo[-p$row,] # brackets can also be used for subsetting
72 ```
```

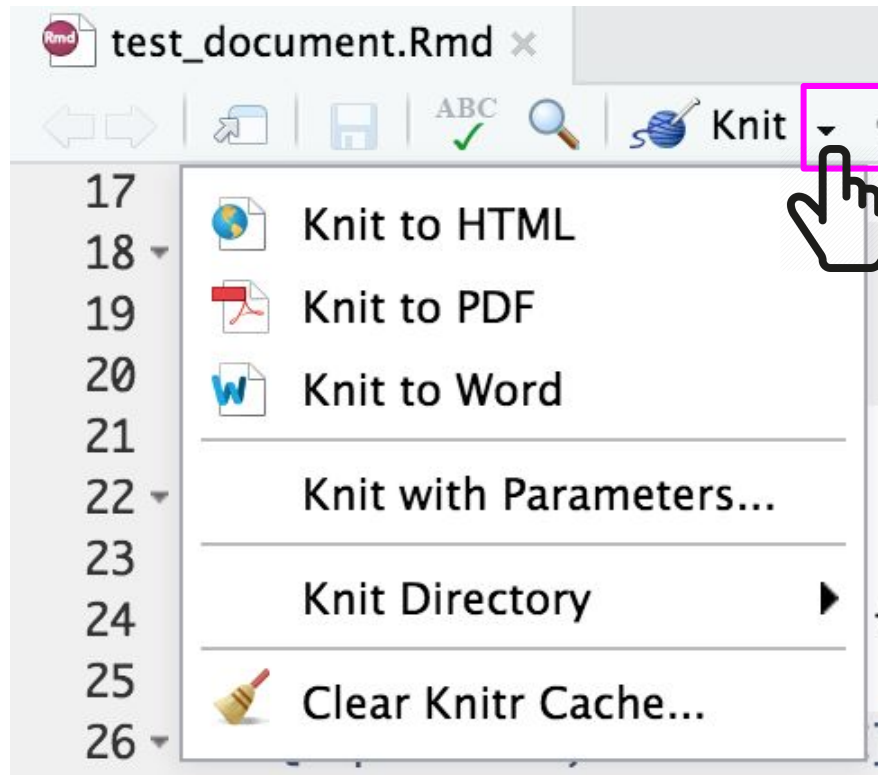
Chunk Name: Unnamed chunk
Output: (Use document default)
 Show warnings
 Show messages
 Use paged tables
 Use custom figure size
[? Chunk options](#) Revert Apply

Can Also influence how a plot is displayed with a chunk!

```
fig.align = "center",  
fig.height = 3, fig.width = 6
```

```
```{r pressure, echo=FALSE, fig.width=6, fig.height=3}  
plot(pressure)
```
```





<https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts>

Rendering in R

```
library(rmarkdown)  
render("Untitled.Rmd")
```

Session information - what's loaded?

```
devtools::session_info()  
# comment-for specific package  
devtools::session_info("pkg")
```


Rmarkdown lab

Download the file from

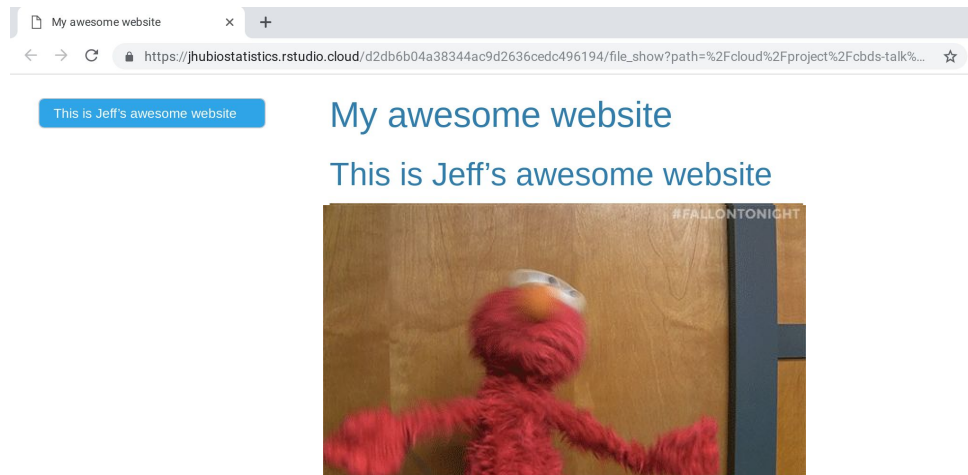
<https://github.com/SISBID/data-wrangling/raw/gh-pages/labs/rmarkdown-lab.Rmd>

Barely scratching the surface

rmarkdown

```
---  
title: "My awesome website"  
output:  
  html_document:  
    toc: true  
    toc_float: true  
    theme: cerulean  
---  
# This is Jeff's awesome website  
  

```



flexdashboard

```
title: "How does your BMI measure up?"  
output: flexdashboard::flex_dashboard  
runtime: shiny  
---
```

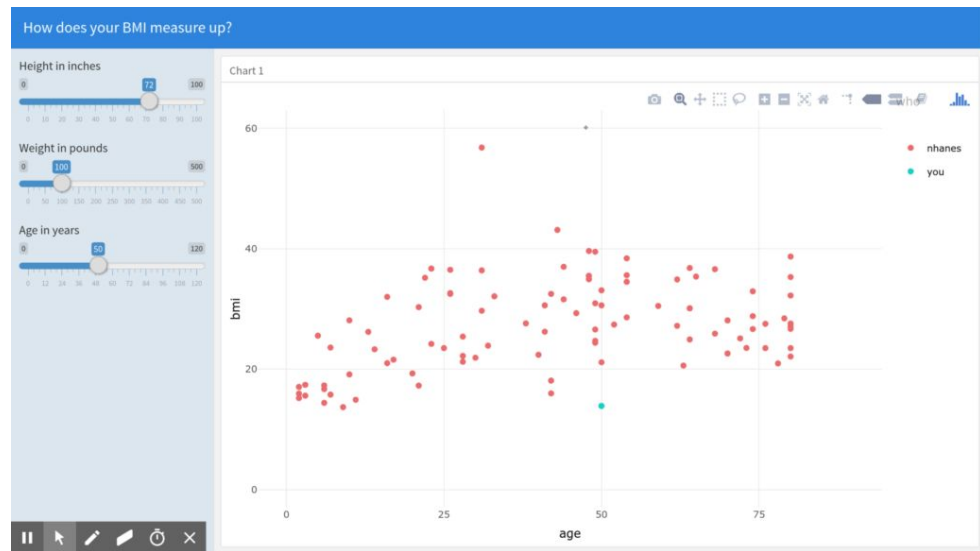
Inputs {.sidebar}

```
```{r}  
library(flexdashboard); library(NHANES); library(plotly);library(dplyr)
sliderInput("height", "Height in inches",0,100,72)
sliderInput("weight", "Weight in pounds",0,500,100)
sliderInput("age", "Age in years",0,120,50)
...
Column

```

### Chart 1

```
```{r}  
nhanes = sample_n(NHANES,100)  
renderPlotly({  
  df = data.frame(bmi = c(nhanes$BMI,input$weight*0.45/(input$height*0.025)^2),  
                 age = c(nhanes$Age,input$age),  
                 who = c(rep("nhanes",100),"you"))  
  ggplotly(ggplot(df) +  
    geom_point(aes(x=age,y=bmi,color=who)) +  
    scale_x_continuous(limits=c(0,90)) +  
    scale_y_continuous(limits=c(0,60)) +  
    theme_minimal()  
  )  
})  
...`
```



Downloading data
reproducibly

Finding and creating files

```
getwd() # get working directory
```

```
setwd("data") # set
```

```
file.exists("data")
```

```
dir.create("data")
```

```
list.files("data")
```

Putting it together

```
if (!file.exists("data")) {  
    dir.create("data")  
}
```

Finding and creating files

```
file.exists("data")
```

```
dir.create("data")
```

```
list.files("data")
```



```
fileUrl <-  
"https://data.baltimorecity.gov/api/views/dz54-2aru/rows.csv?accessType=DOWNLOAD"  
  
curl::curl_download(fileUrl,  
  destfile="./data/cameras.csv")  
list.files("./data")  
dateDownloaded <- date()  
dateDownloaded
```