

# Statistical Computing: The Course

Biostatistics 140.776

Roger D. Peng

# About Me

- Outdoor air pollution and health
- Air pollution epidemiology
- Ambient air quality standards
- Time series, spatial statistics, hierarchical modeling



# About Me

- Indoor air pollution and health
- Panel studies in vulnerable groups (COPD, asthma)
- Environmental interventions and clinical trials
- Longitudinal data, causal inference, mediation

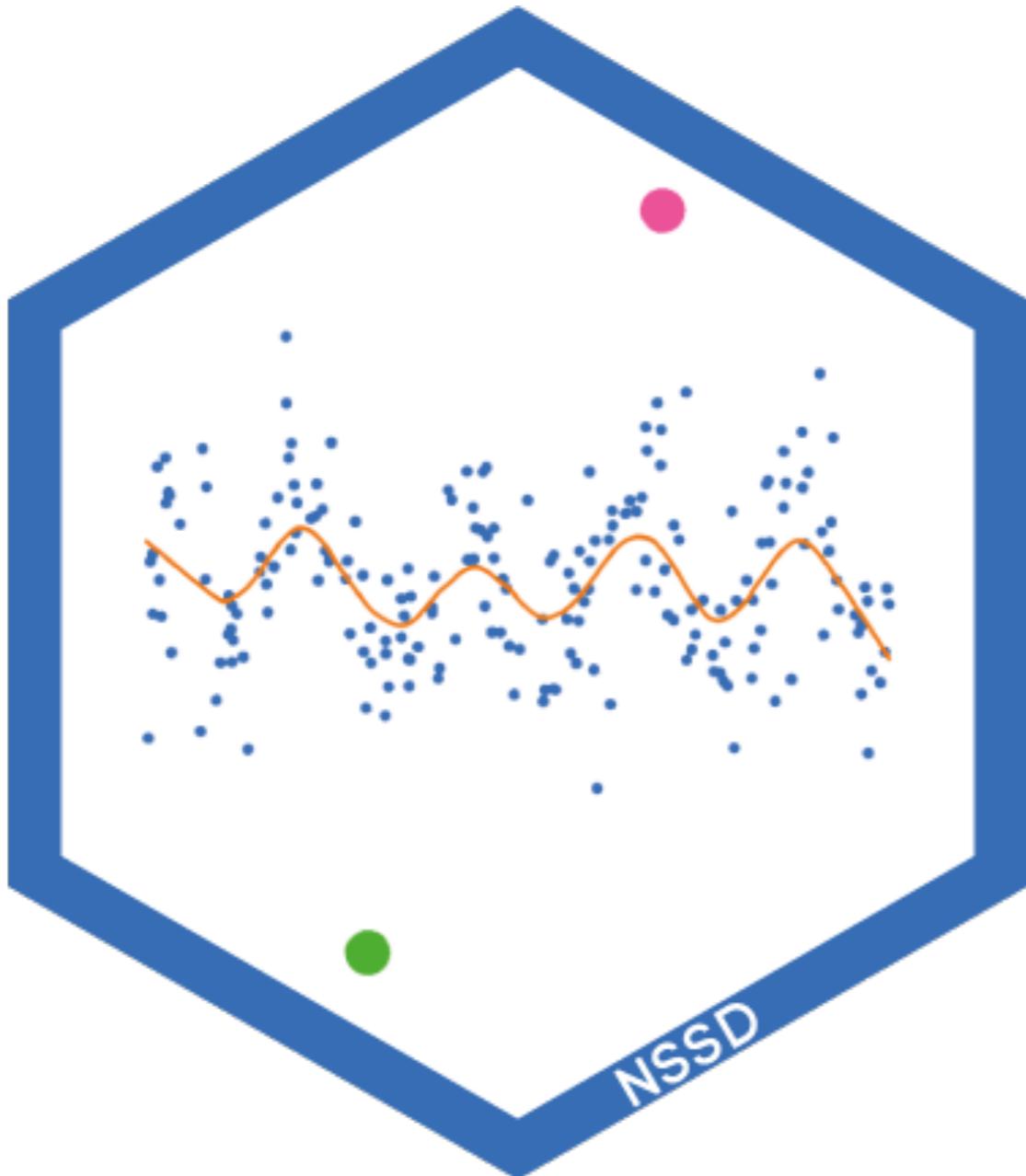




# simplystats

[simplystatistics.org](http://simplystatistics.org)

@simplystats



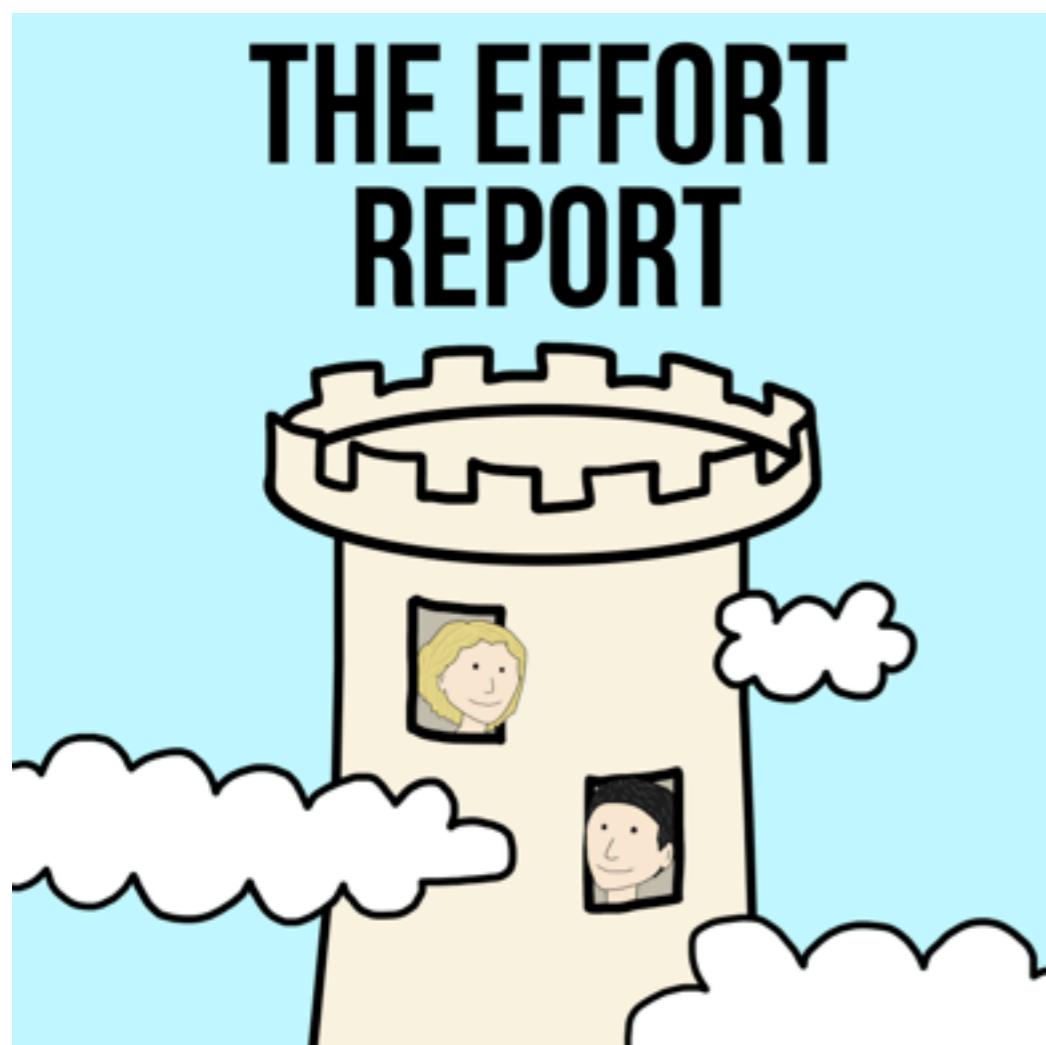
# Not So Standard Deviations

(with Hilary Parker of Stitch Fix)



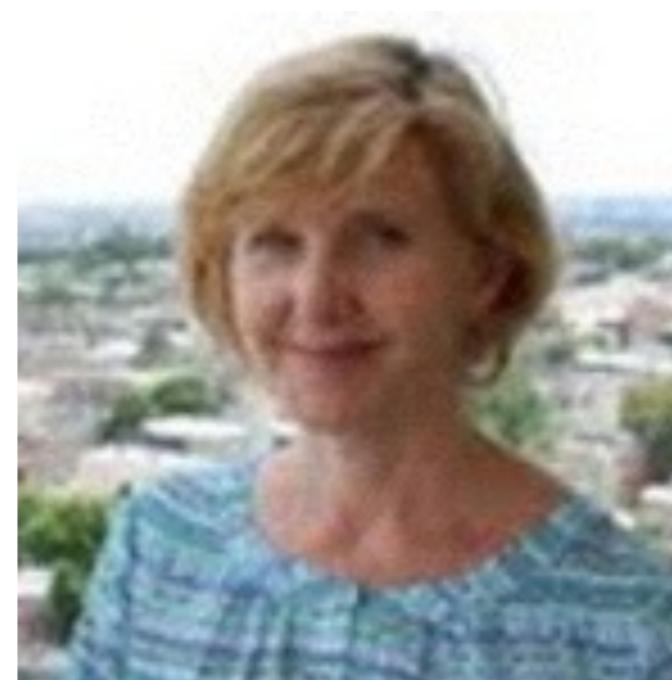
<https://soundcloud.com/nssd-podcast>

Subscribe in iTunes: <https://goo.gl/ZhWYbd>



# The Effort Report

(with Elizabeth Matsui of JHU)



<http://effortreport.libsyn.com>

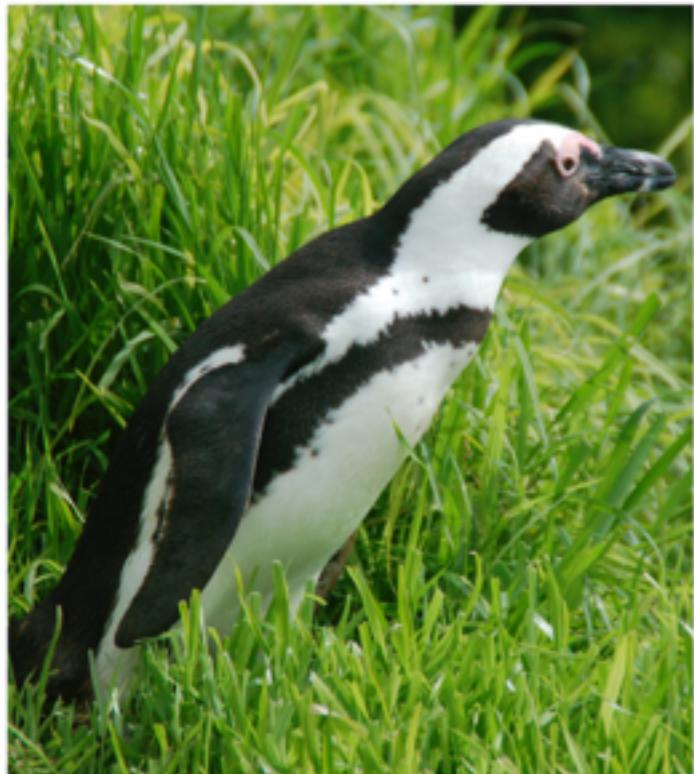
Subscribe in iTunes: <http://goo.gl/hz3AHL>

# Course Logistics

- **Instructor:** Roger D. Peng (Dept. of Biostatistics)
- **Web Site:** [rdpeng.github.io/Biostat776](http://rdpeng.github.io/Biostat776)
- **Meets:** T/Th 1:30—2:50pm in W2008
- **Office Hour:** Wednesday 12—1pm (E3535), or email me, or if my door is open, come in!
- **NOTE:** Lecture on September 15 is **cancelled**; will be recorded instead (I'll send a reminder)

# Textbooks

R Programming  
for Data Science



Roger D. Peng

[leanpub.com/rprogramming](http://leanpub.com/rprogramming)

Exploratory Data  
Analysis with R



Roger D. Peng

[leanpub.com/exdata](http://leanpub.com/exdata)

Report Writing for  
Data Science in R



**Roger D. Peng**

[leanpub.com/reportwriting](http://leanpub.com/reportwriting)

# Leanpub



96,483  
READERS 182  
PAGES

## R Programming for Data Science

 Roger D. Peng

This book brings the fundamentals of R programming to you, using the same material developed as part of the industry-leading Johns Hopkins Data Science Specialization. The skills taught in this book will lay the foundation for you to begin your journey learning data science. See the packages below to obtain datasets, R code files, and video lectures. Printed copies of this book are available through Lulu.

### R Programming for Data Science



Roger D. Peng

UPDATED 28 DAYS AGO

Edit

 ENGLISH  PDF  EPUB  MOBI  APP

FREE! \$20.00  
MINIMUM SUGGESTED

You pay (USD) 

\$ 20.00

Author earns 

\$ 17.50

Add Ebook to Cart

# Textbooks

## R Programming for Research

*Colorado State University, ERHS 535*

*Brooke Anderson and Rachel Severson*

[geanders.github.io/RProgrammingForResearch/](http://geanders.github.io/RProgrammingForResearch/)

## Mastering Software Development in R

*Roger D. Peng, Sean Kross, and Brooke Anderson*

2016-08-31

[rdpeng.github.io/RProgDA/](http://rdpeng.github.io/RProgDA/)

# Grading

- No exams!
- **Three** homeworks requiring programming in R and some basic data analysis
- Each homework counts equally (1/3)
- **Homeworks submitted via Courseplus dropbox**

# Software

- R (of course)
- Make sure you have the **latest version** installed
- Obtain R from <https://cran.rstudio.com>
- Various R packages
- You can use whatever you want with respect to Mac, Windows, Linux....

# CRAN



[CRAN](#)  
[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)

[About R](#)  
[R Homepage](#)  
[The R Journal](#)

[Software](#)  
[R Sources](#)  
[R Binaries](#)  
[Packages](#)  
[Other](#)

[Documentation](#)  
[Manuals](#)  
[FAQs](#)  
[Contributed](#)

## The Comprehensive R Archive Network

### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows** and **Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

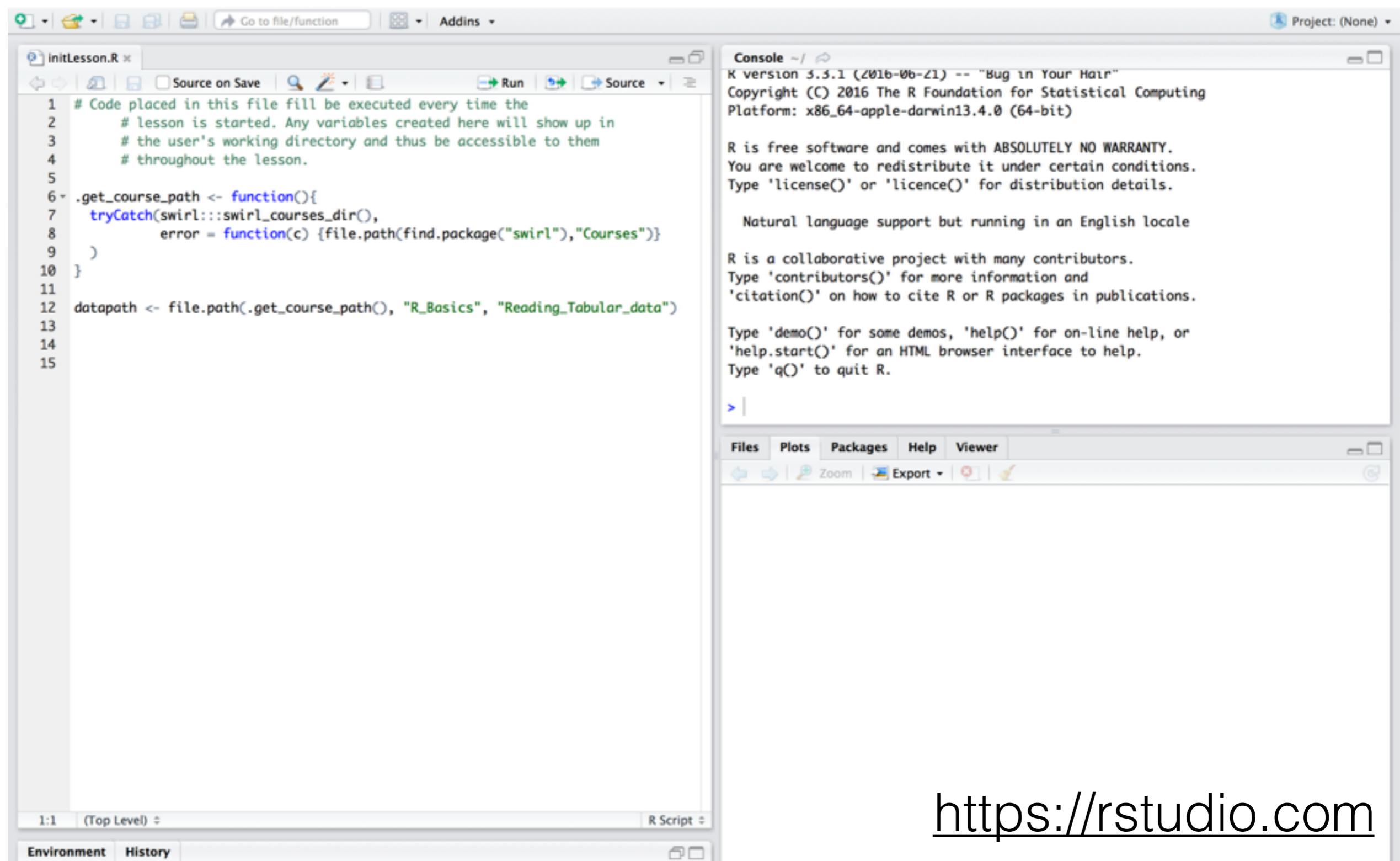
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

### Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Tuesday 2016-06-21, Bug in Your Hair) [R-3.3.1.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

# RStudio IDE



# RStudio

The screenshot shows the RStudio homepage. At the top left is the RStudio logo. A red arrow points from the logo to the "Products" menu item. The "Products" menu is highlighted and circled in red, showing options like RStudio, Shiny, R Packages, RStudio Server Pro, Shiny Server Pro, and shinyapps.io. Below the menu, there's a large banner with text about new features and RStudio Server, along with "Learn More" and "Download a 45 day evaluation" buttons. On the right side, there's a preview of the RStudio IDE interface, including a code editor with R code, a console window, and a map visualization.

RStudio

Products Resources Pricing About Us Blog Q

New Features RStudio Server

Learn More

Download a 45 day evaluation

RStudio

Shiny

R Packages

RStudio Server Pro

Shiny Server Pro

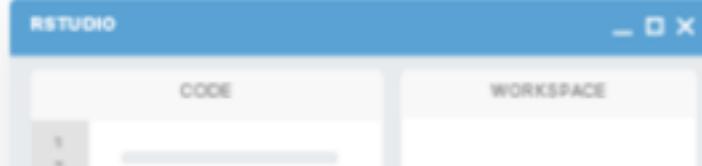
shinyapps.io

```
data.url <- url("http://data.cityofchicago.org/api/views/41jzn-s7e5/rows.csv")  
data <- read.csv(data.url, header = TRUE) # takes a minute...  
# drop columns with missing values  
data <- na.omit(data)  
# drop columns with missing values  
data <- dropNA(data)  
# convert categorical variables to factors  
data <- as.factor(data$RISK)  
# convert categorical variables to factors  
data <- as.factor(data$ZIP)  
# convert categorical variables to factors  
data <- as.factor(data$CITY)  
# convert categorical variables to factors  
data <- as.factor(data$TAXCLASS)  
# convert categorical variables to factors  
data <- as.factor(data$PROPERTYCLASS)  
# convert categorical variables to factors  
data <- as.factor(data$BUILDINGCLASS)  
# convert categorical variables to factors  
data <- as.factor(data$STRUCTURECLASS)  
# convert categorical variables to factors  
data <- as.factor(data$USECLASS)  
# convert categorical variables to factors  
data <- as.factor(data$ASSESSMENTCLASS)  
# convert categorical variables to factors  
data <- as.factor(data$ASSESSMENTTYPE)  
# convert categorical variables to factors  
data <- as.factor(data$ASSESSMENTMETHOD)  
# convert categorical variables to factors  
data <- as.factor(data$ASSESSMENTTYPE)  
# convert categorical variables to factors  
data <- as.factor(data$ASSESSMENTMETHOD)  
# convert categorical variables to factors  
data <- as.factor(data$ASSESSMENTTYPE)  
# convert categorical variables to factors  
data <- as.factor(data$ASSESSMENTMETHOD)
```

Console - /Radar/

```
> data1 <- subset(data, risk %in% c("Risk 1 (High)", "Risk 2 (Medium)", "Risk 3 (Low)"))  
> data1$risk <- droplevels(data1$risk)  
>  
> data1 <- data1[1:50]  
> library(leaflet)  
> leaflet(data1) %>%  
+   addTiles() %>%  
+   addMarkers(lat = -latitude, lng = -longitude)
```

ZIP explorer



# RStudio

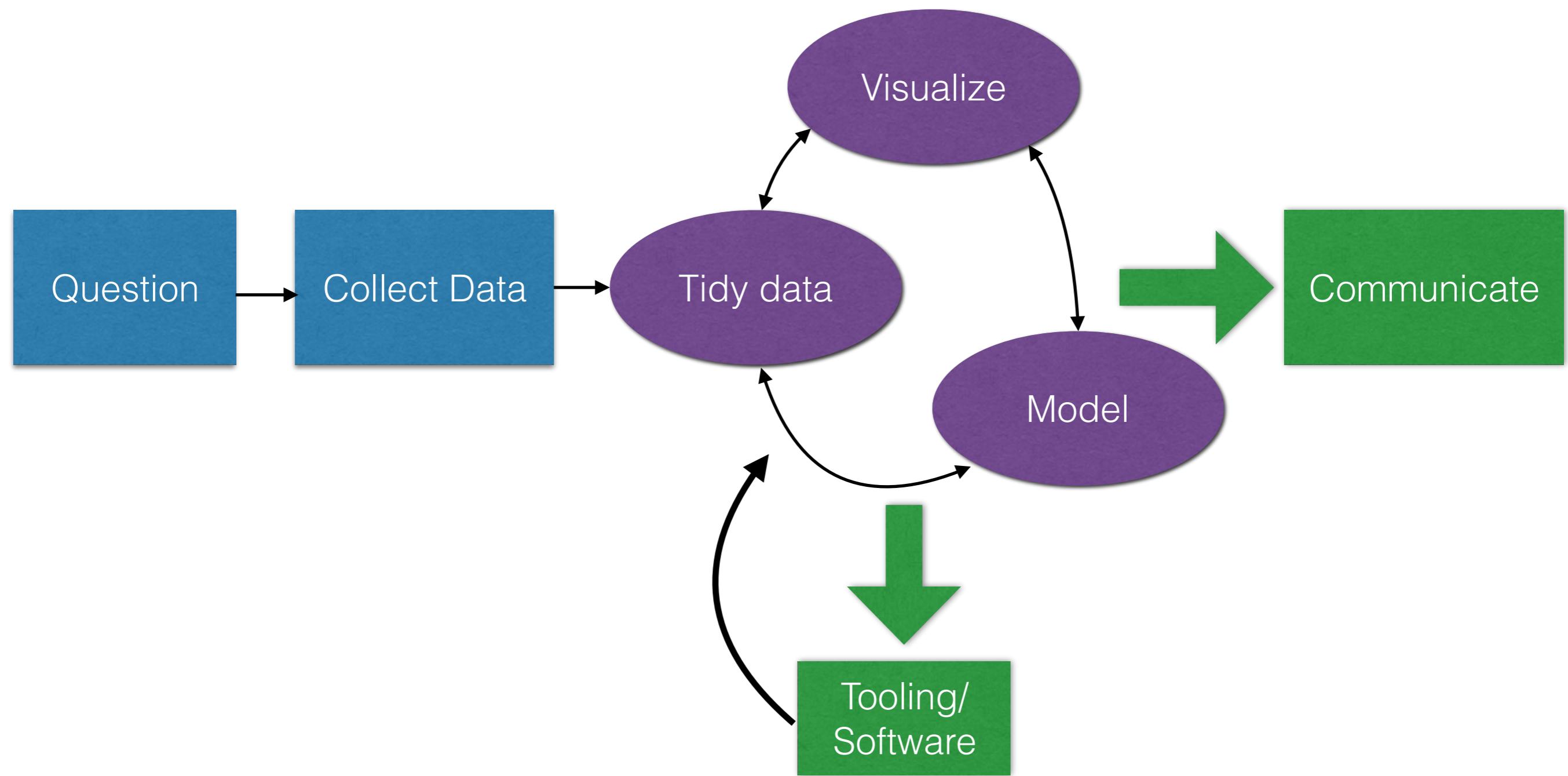
## Choose Your Version of RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. [Learn More](#)

	RStudio Desktop (Free License)	RStudio Desktop (Commercial License)	RStudio Server (Free License)	RStudio Server Pro (Commercial License)
Integrated Development Environment for R	✓	✓	✓	✓
Priority support		✓		✓
Access via Web Browser			✓	✓
Enterprise Security and Access Controls				✓
Project Sharing				✓

# Intermission

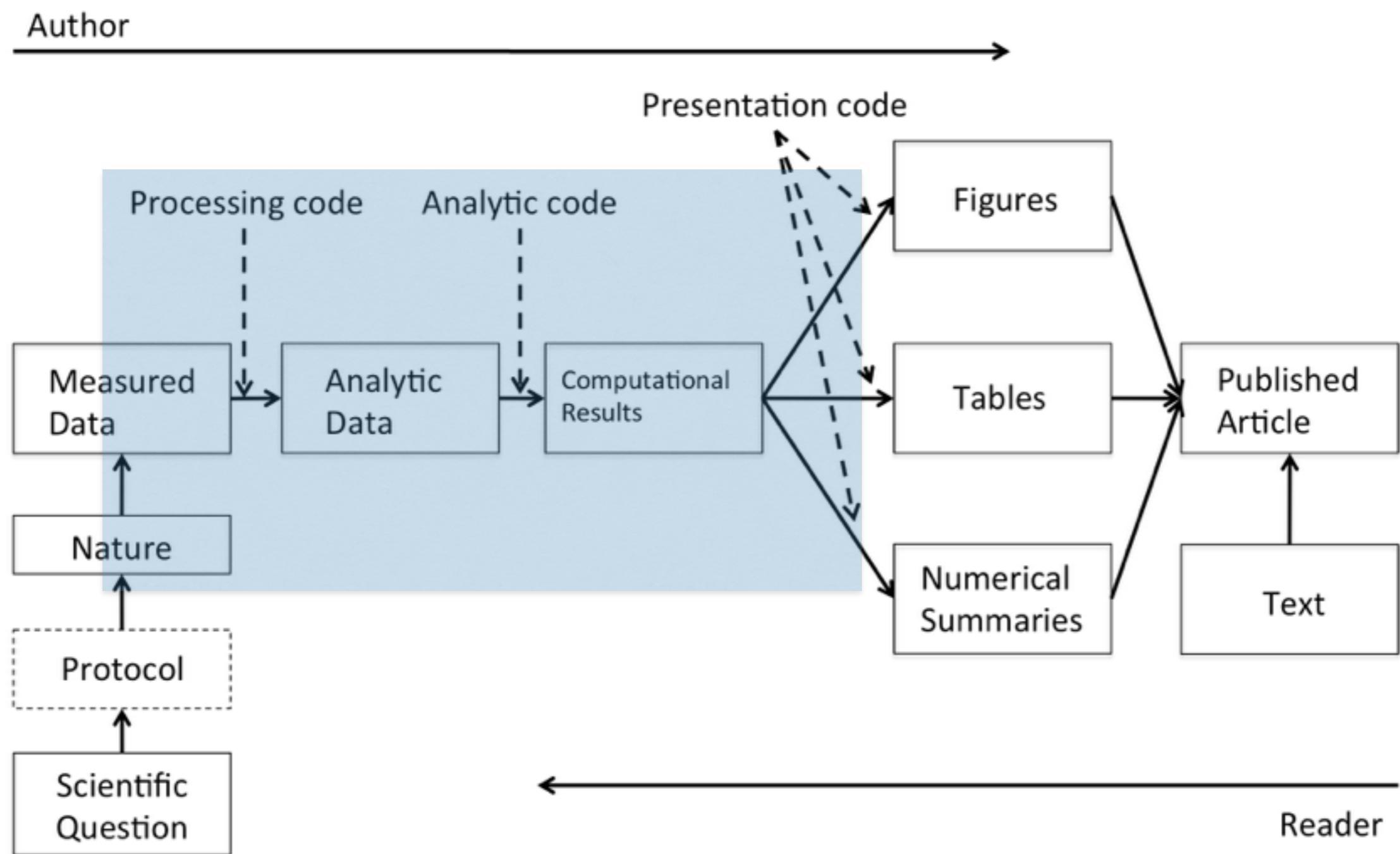
# Data Science Workflow



# Major Themes

- Reproducible research
- Data management and manipulation, tidy data
- Data visualization and communication
- Programming with R
- Products and tooling

# Reproducible Research?



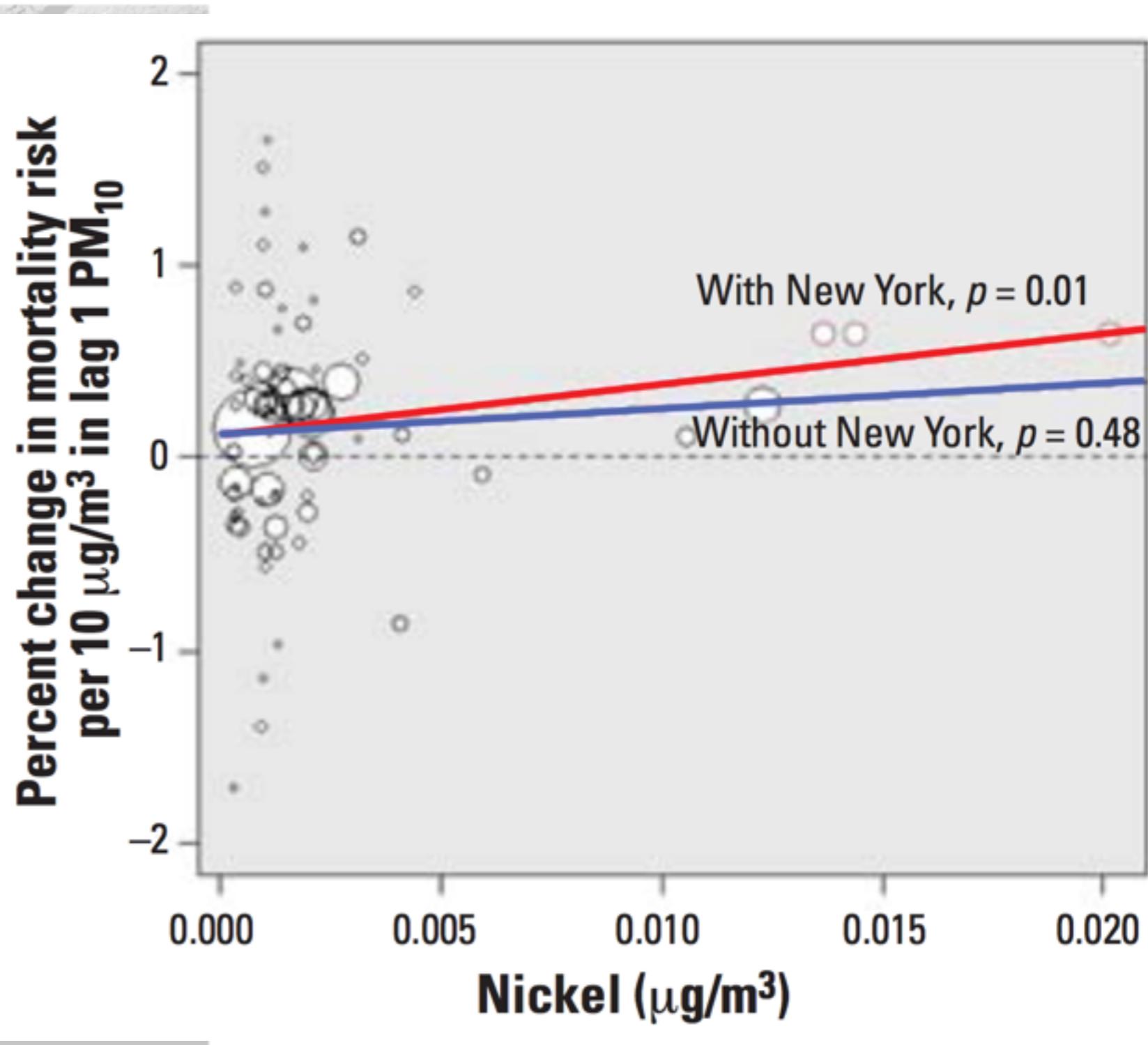
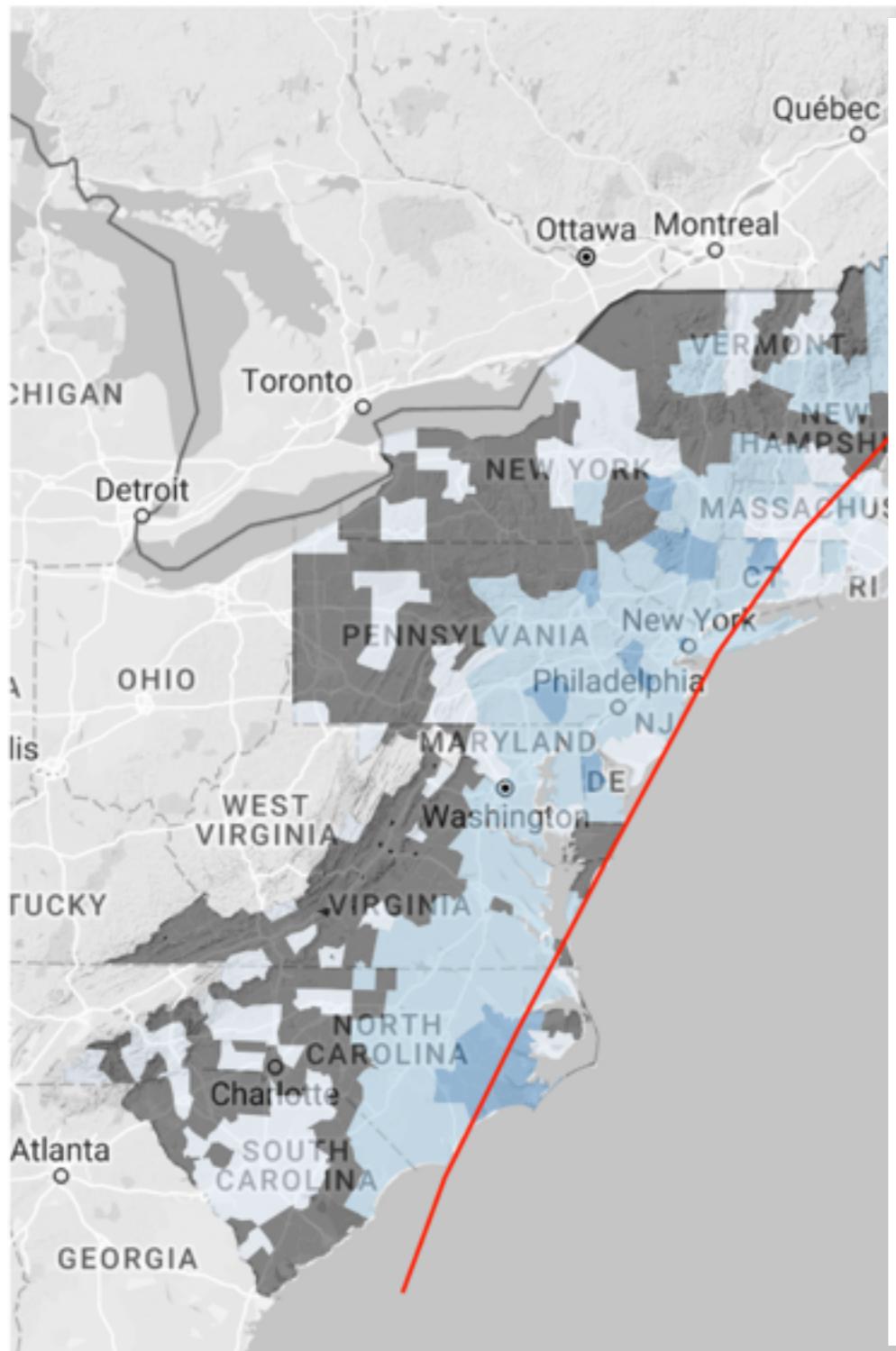
# Tidying Data

# Tidying Data

```
## Source: local data frame [280 x 7]

##           row row_info to_date value          header1          header2
##   (int)     (chr)    (chr)  (dbl)
## 1       16     Cash      NA 1.890 IF NWPL Rocky Mountains Fixed Price
## 2       16     Cash      NA 1.910 IF NWPL Rocky Mountains Fixed Price
## 3       16     Cash      NA     NA IF NWPL Rocky Mountains             Basis
## 4       16     Cash      NA     NA IF NWPL Rocky Mountains             Basis
## 5       17     ROM       NA 2.060 IF NWPL Rocky Mountains Fixed Price
## 6       17     ROM       NA 2.080 IF NWPL Rocky Mountains Fixed Price
## 7       17     ROM       NA     NA IF NWPL Rocky Mountains             Basis
## 8       17     ROM       NA     NA IF NWPL Rocky Mountains             Basis
## 9       18 37226      NA 2.395 IF NWPL Rocky Mountains Fixed Price
## 10      18 37226      NA 2.415 IF NWPL Rocky Mountains Fixed Price
## ...     ...     ...     ...     ...
## Variables not shown: header3 (chr)
```

# Data Visualization



# Programming and Abstraction

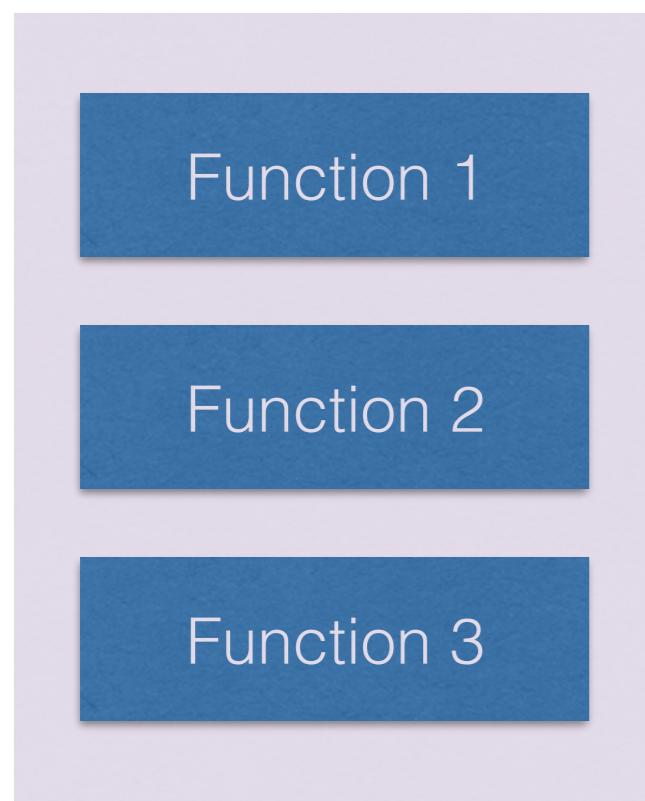
```
library(readr)
library(dplyr)
## Data were obtained from http://cran-logs.rstudio.com
cran <- read_csv("data/2016-07-20.csv.gz", col_types = "ccicccccci")
cran %>% filter(package == "filehash") %>% nrow
...
```

```
## pkgname: package name (character)
## date: YYYY-MM-DD format (character)
num.download <- function(pkgname, date) {
    ## Construct web URL
    src <- sprintf("http://cran-logs.rstudio.com/%s/%s.csv.gz",
                   substr(date, 1, 4), date)

    ## Construct path for storing local file
    dest <- file.path("data", basename(src))

    ## Don't download if the file is already there!
    if(!file.exists(dest))
        download.file(src, dest, quiet = TRUE)
    cran <- read_csv(dest, col_types = "ccicccccci", progress = FALSE)
    cran %>% filter(package == pkgname) %>% nrow
}
```

# Products and Tooling



R package

**library(mypackage)**

Shiny app

Movie explorer

**Filter**

Minimum number of reviews on Rotten Tomatoes  
10 80 300

Year released  
1940 1970 2014

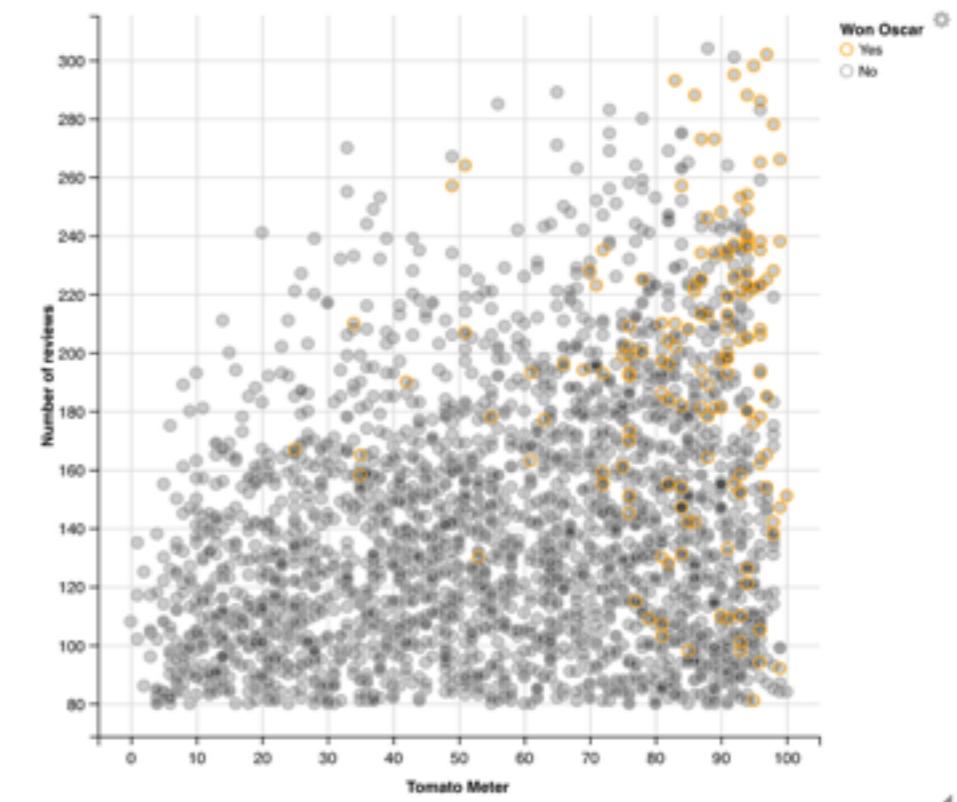
Minimum number of Oscar wins (all categories)  
0 4

Dollars at Box Office (millions)  
0 800

Genre (a movie can have multiple genres)  
All

Director name contains (e.g., Miyazaki)  
Tom

Cast names contains (e.g. Tom)



# Major Themes

- knitr, markdown, R markdown
- Tidy data, dplyr, tidyr, lubridate, regular expressions
- Principles of data graphics, ggplot2, mapping
- Functions, functional programming, object oriented programming
- R packages, Shiny apps