

# Statistical Computing: The Course

Biostatistics 140.776

Roger D. Peng

[rdpeng.github.io/Biostat776](https://rdpeng.github.io/Biostat776)

# 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



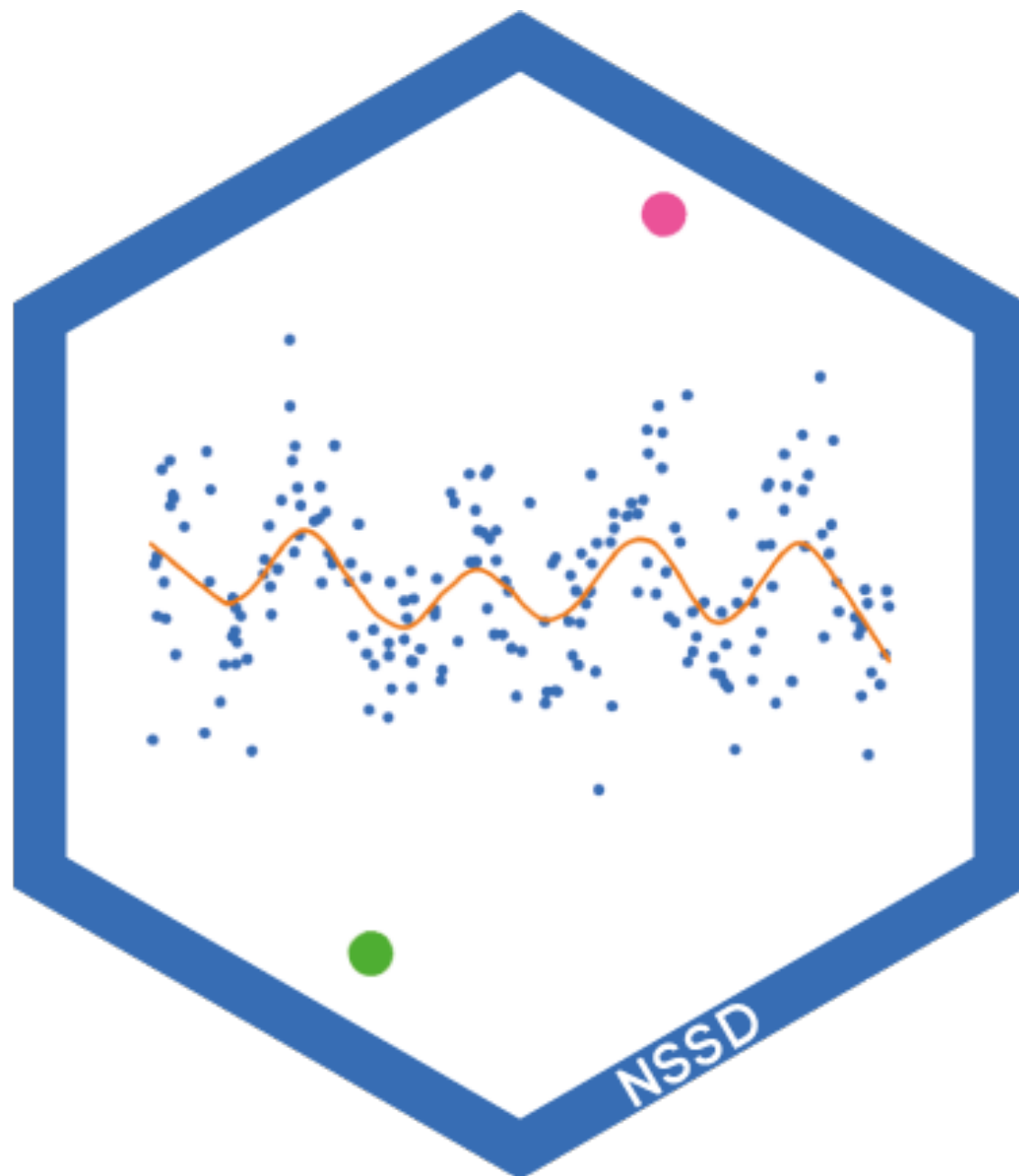




# simplystats

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

@simplystats

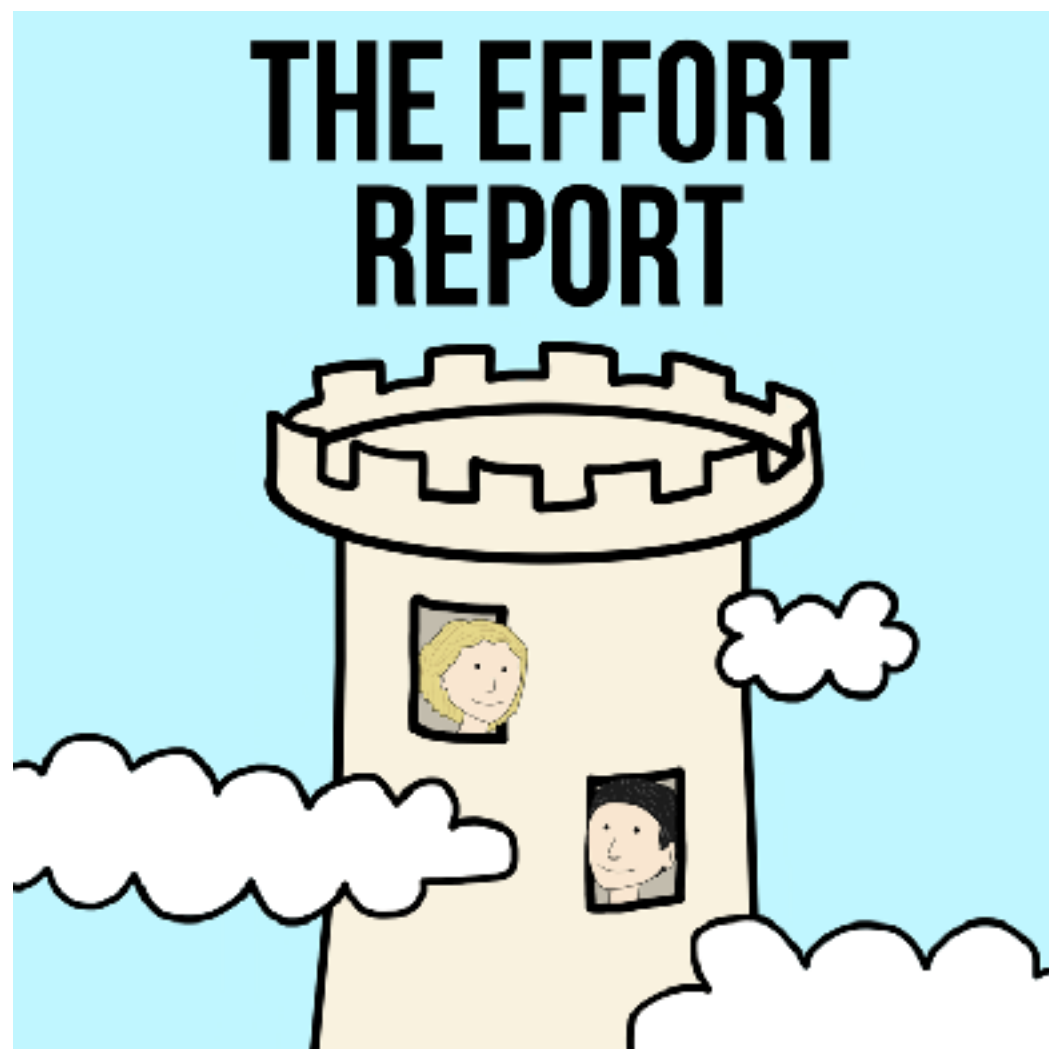


# Not So Standard Deviations

(with Hilary Parker of Stitch Fix)



<http://nssdeviations.com>



## **The Effort Report**

(with Elizabeth Matsui of UT Austin)



<http://effortreport.libsyn.com>

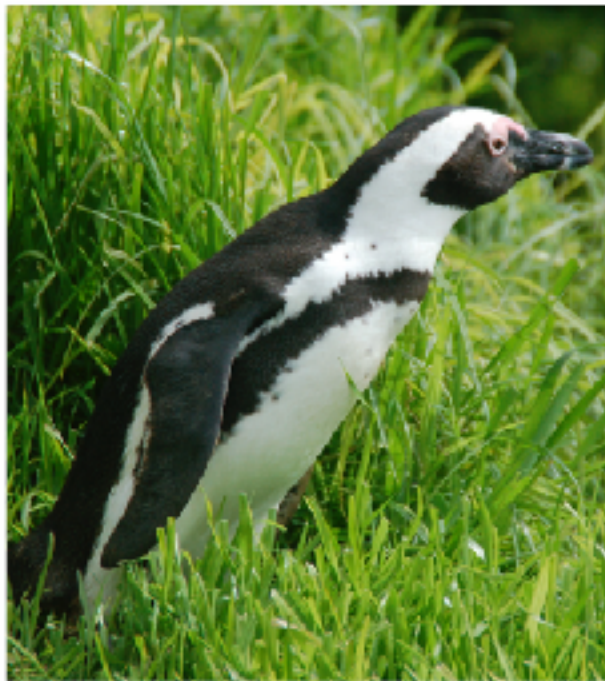
# Course Logistics

- **Instructor:** Roger D. Peng (Dept. of Biostatistics)
- **Web Site:** [rdpeng.github.io/Biostat776](https://rdpeng.github.io/Biostat776)
- **Meets:** T/Th 1:30—2:50pm in W2008
- **Office Hour:** Wednesday 11:30am—1:30pm (E3535), or email me, or if my door is open, come in!
- **TA Office Hour:** Thursday 3:30-4:30pm



# Textbooks

## R Programming for Data Science



Roger D. Peng

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

## Exploratory Data Analysis with R



Roger D. Peng

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

## Mastering Software Development in R



Roger D Peng

Sean Kross

Brocke Anderson

[leanpub.com/msdr](https://leanpub.com/msdr)




# Leanpub

182  
PAGES

ENGLISH PDF EPUB MOBI APP

## 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. Printed copies of this book are [available through Lulu](#).

Table Of Contents 


## R Programming for Data Science



Roger D. Peng

LAST UPDATED ON 2018-03-02

Edit

 You own this book! [View it in your Library](#)

Free! \$20.00  
MINIMUM SUGGESTED 

YOU PAY

\$20.00

AUTHOR EARN\$

\$16.00

Packages Details

- ☒ The Book
- ☐ The Book + Datasets + R Code Files
- ☐ The Book + Lecture Videos (HD) + Datasets + R Code Files

YOU PAY (US\$)

\$20.00

EU customers: Price excludes VAT.  
VAT is added during checkout.

Add Ebook to Cart

# Textbooks

## **R Programming for Research**

***Colorado State University, ERHS 535***

***Brooke Anderson and Rachel Severson***

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

Report Writing for  
Data Science in R



**Roger D. Peng**

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

# Grading

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



# Software

- R (of course)
- Make sure you have the **latest version** installed (version 3.5.1)
- Obtain R from **<https://cran.r-project.org>**
- 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](#)

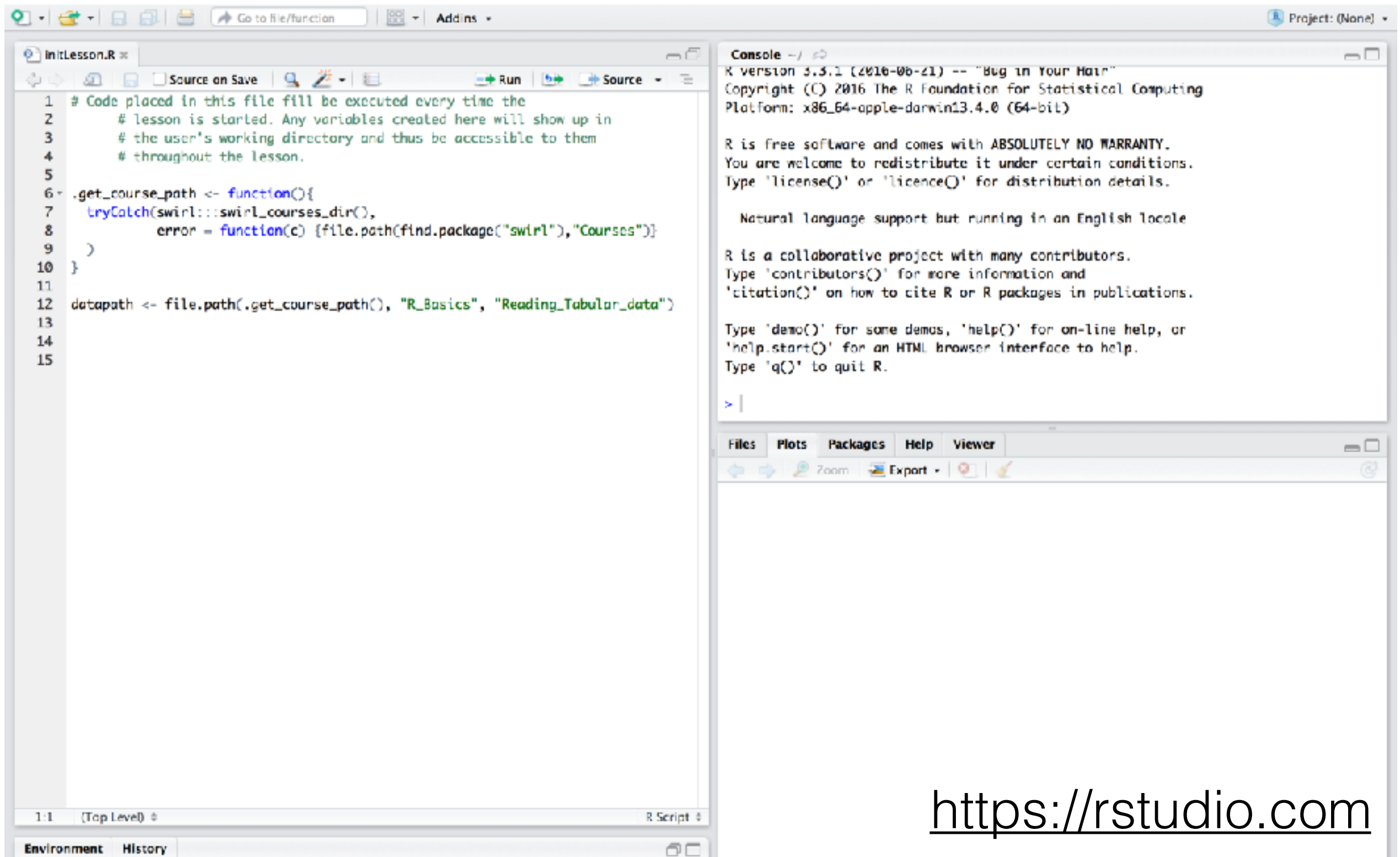
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

### 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](#)

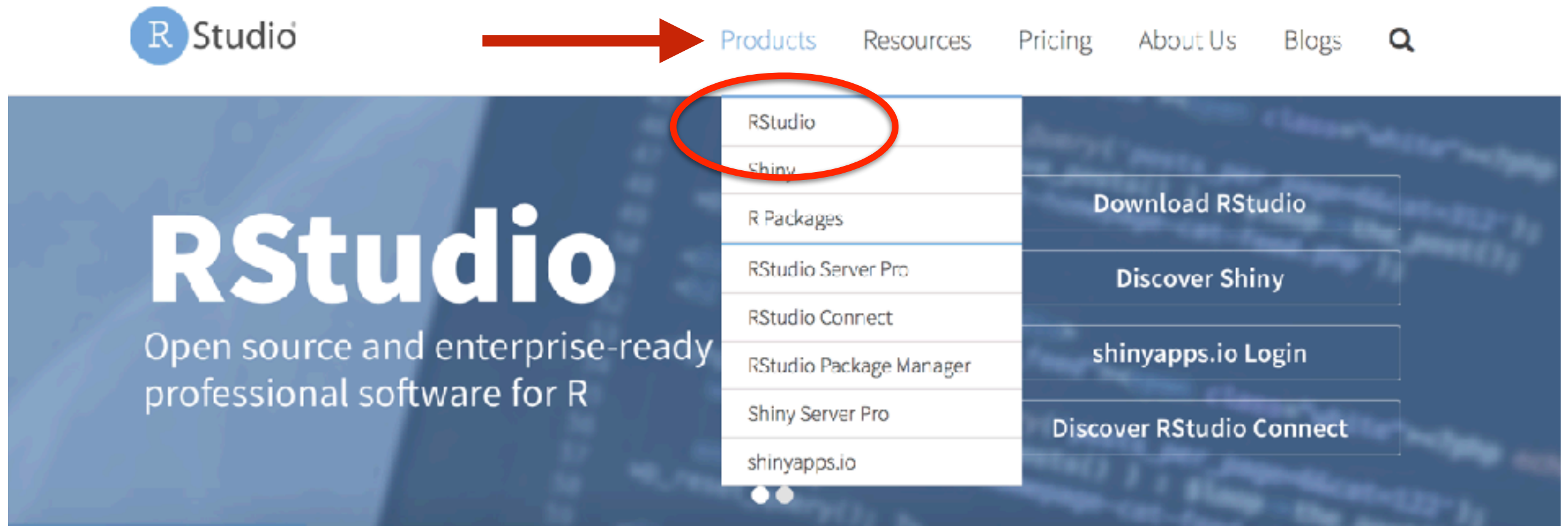
# RStudio IDE



<https://rstudio.com>



# RStudio



# RStudio

## Choose Your Version of RStudio

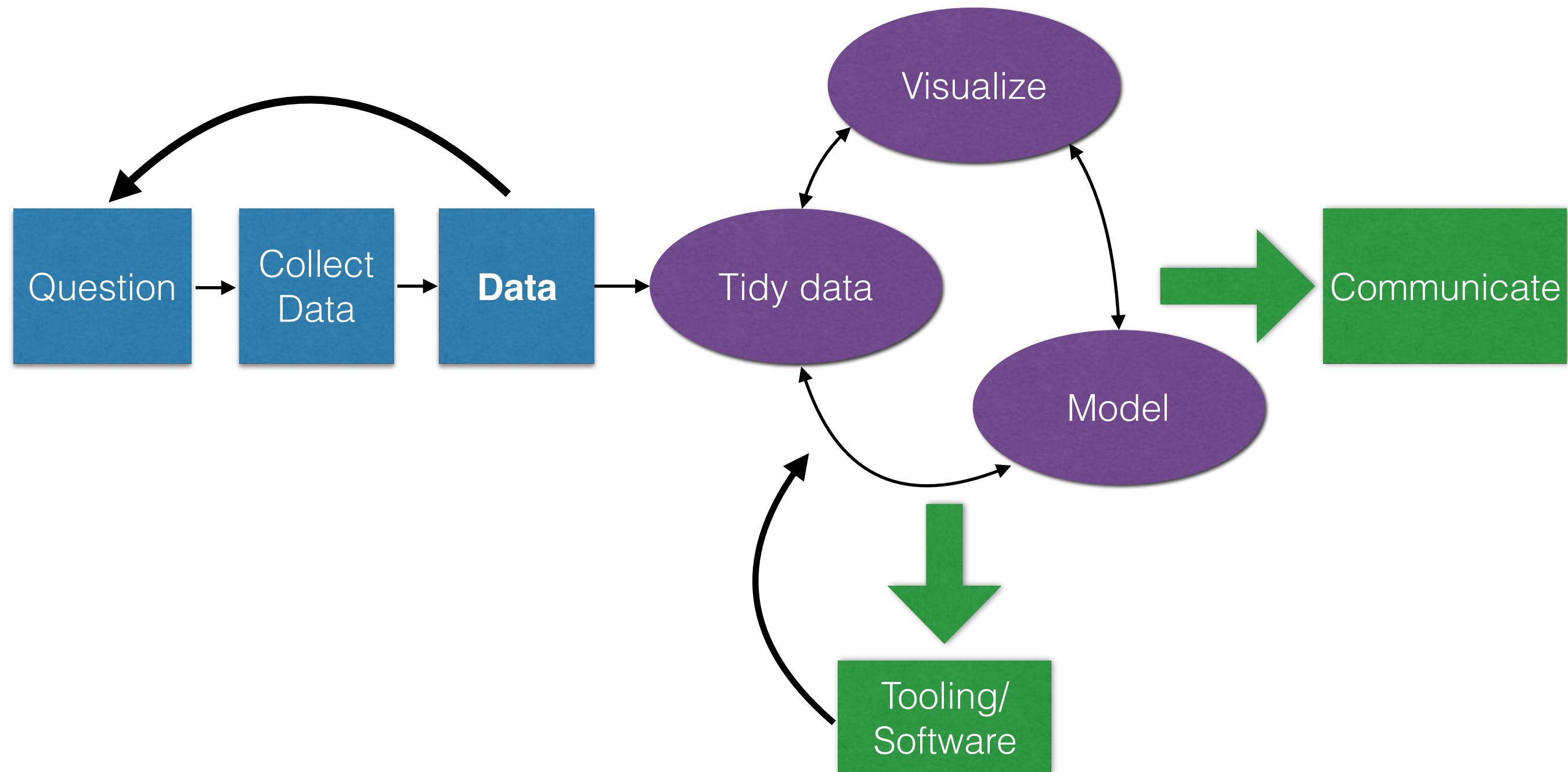
RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. [Learn More](#)

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

Intermission



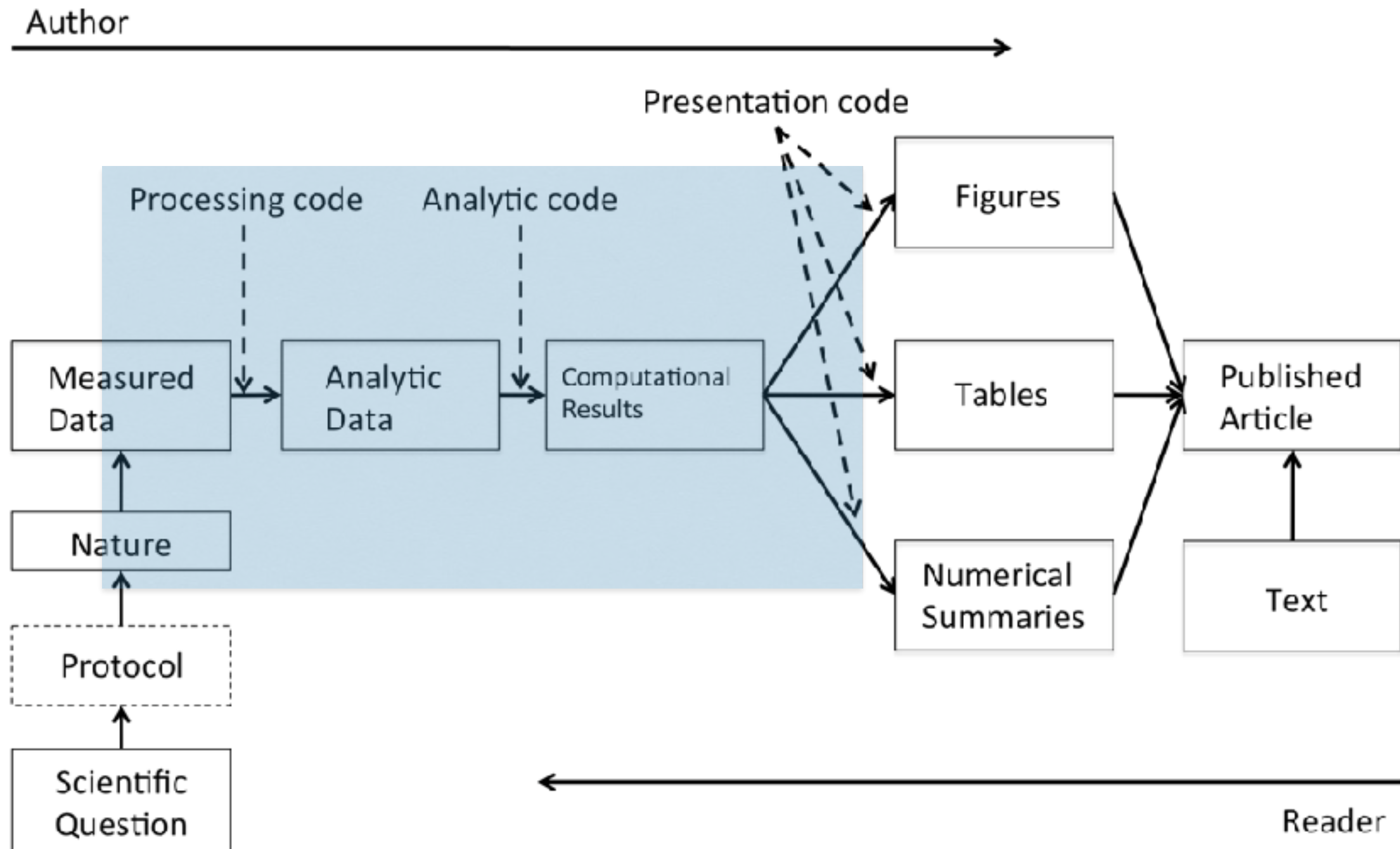
# Data Science Workflow



# Major Themes

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

# Reproducible Research?





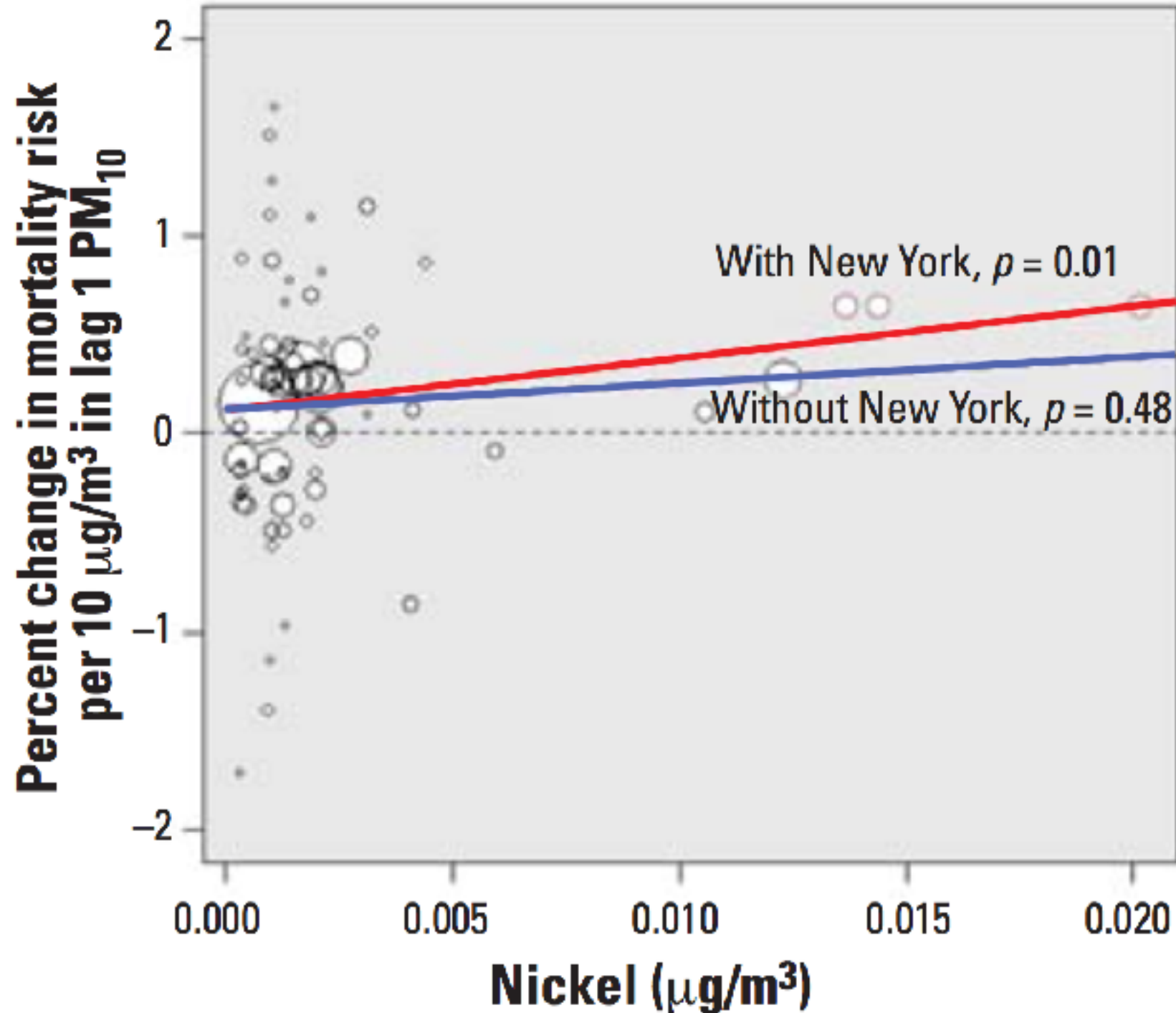
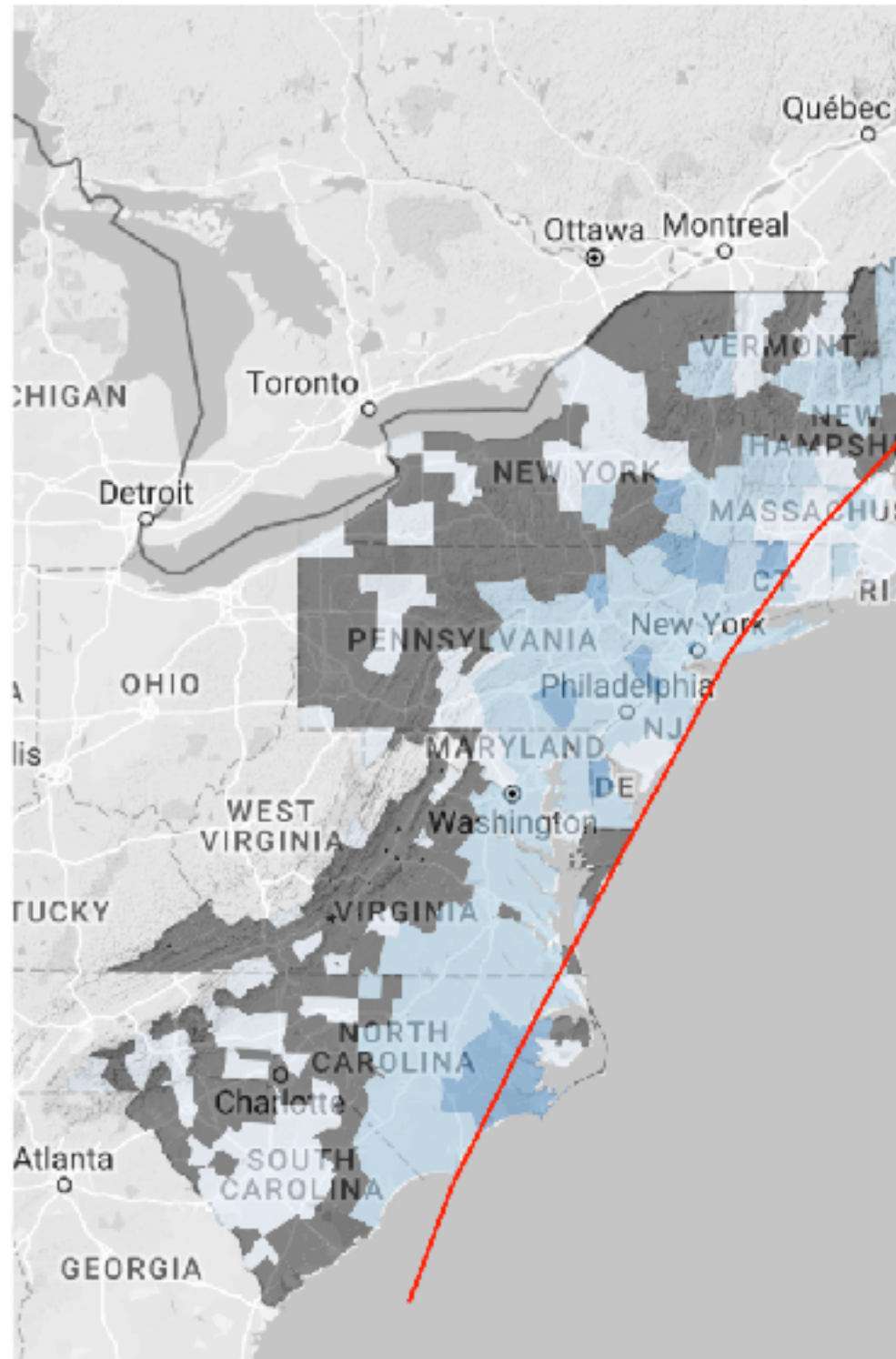
# Tidying Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Enron North America - West Gas																				
2	November 9, 2001																				
3																					
4	ENA - West Gas Contacts																				
5																					
6	Houston Office										Regional Offices										
7	Barry Tycholiz (713) 853-1587										Mark Whitt (303) 575-6473 Denver										
8	Kim Ward (713) 853-0685										Paul Luccl (303) 575-6474 Denver										
9	Stephanie Miller (713) 853-1688										Tyrell Harrison (303) 575-6478 Denver										
10	Philip Polsky (713) 853-5181										Dave Fuller (503) 464-3732 Portland										
11																					
12	Forward Prices (US\$/MMBtu)																				
13																					
14																					
15																					
16																					
17																					
18																					
19																					
20																					
21																					
22																					
23																					
24																					
25																					
26																					
27																					
28																					
29																					
30																					
31																					
32																					
33																					
34																					
35																					
36																					
37																					
38																					
39																					
40																					
41																					
42																					
43																					
44																					
45																					

# Tidying Data

```
## Source: local data frame [280 x 7]
##
##      row row_info to_date value             header1      header2
##      (int)   (chr)   (chr) (dbl)             (chr)        (chr)
## 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





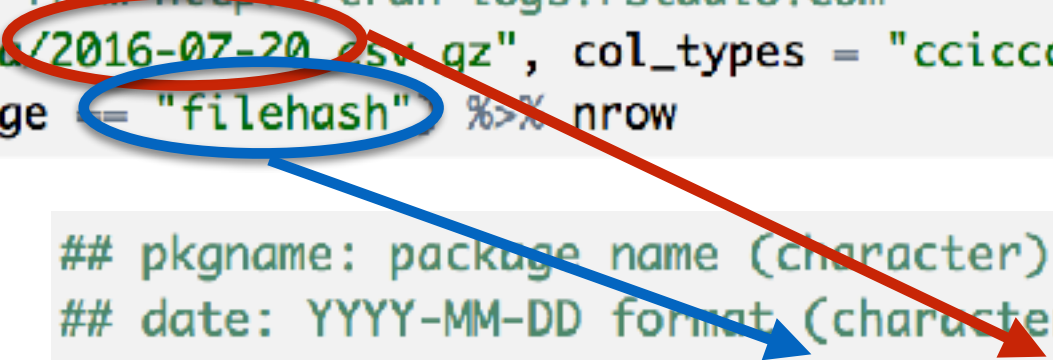
# Programming and Abstraction

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





# Products and Tooling

R package

**library(mypackage)**

Function 1

Function 2

Function 3

Shiny app

Movie explorer

Filter

Minimum number of reviews on Rotten Tomatoes

10 80 100

Year released

1910 1970 2014

Minimum number of Oscar wins (all categories)

0 4

Dollars at Box Office (millions)

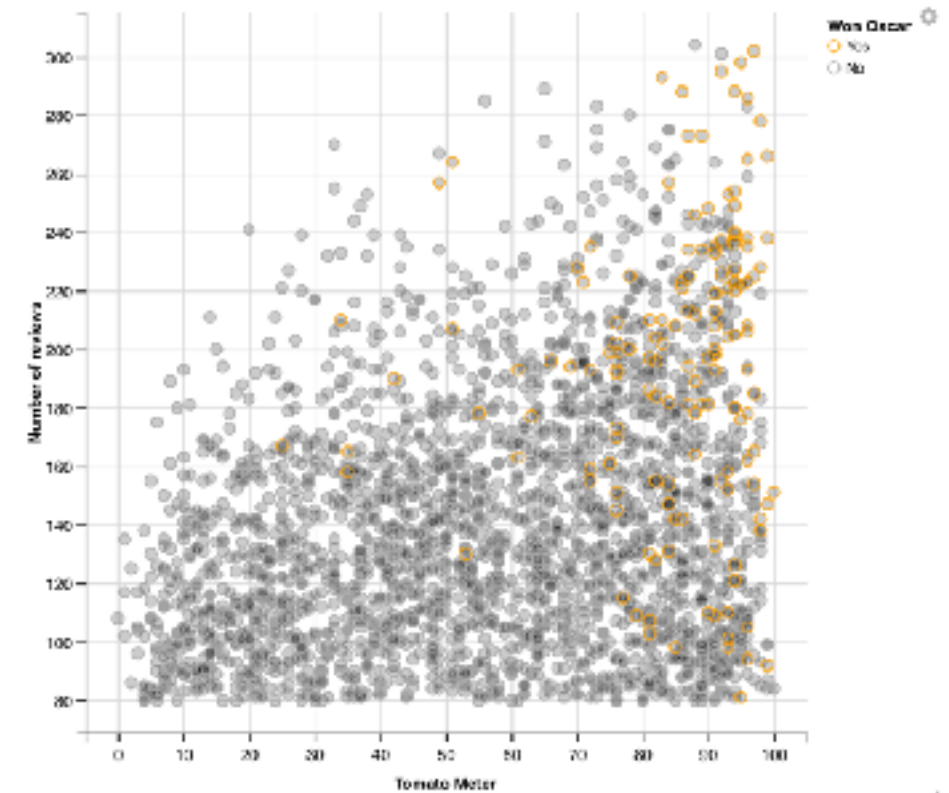
0 100

Genre (a movie can have multiple genres)

All

Director name contains (e.g., Miyazaki)

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