

Statistical Computing

Biostatistics 140.776

Kasper Daniel Hansen

< khansen@jhsph.edu | www.hansenlab.org >

Department of Biostatistics

McKusick-Nathans Institute of Genetic Medicine

Johns Hopkins University

About me

Statistical Genomics

Epigenomics / Transcriptomics / Metabolomics

Exploratory Data Analysis

Small n / Large p

Software / Bioconductor



Course designed by Roger Peng



Current location: Australia (sabbatical)



Roger D. Peng
@rdpeng

Following



I'm trying to learn the rules of cricket. Send help.

3:40 PM - 27 Aug 2017

Logistics

Instructor: Kasper D. Hansen

TA: Weixiang Fang

Meets: TuTh 1:30-2:50 in W2008

Office Hour: Tu 3-4 (Welch 111) or email.

Note: Lecture Oct 10 is cancelled.

Important dates:

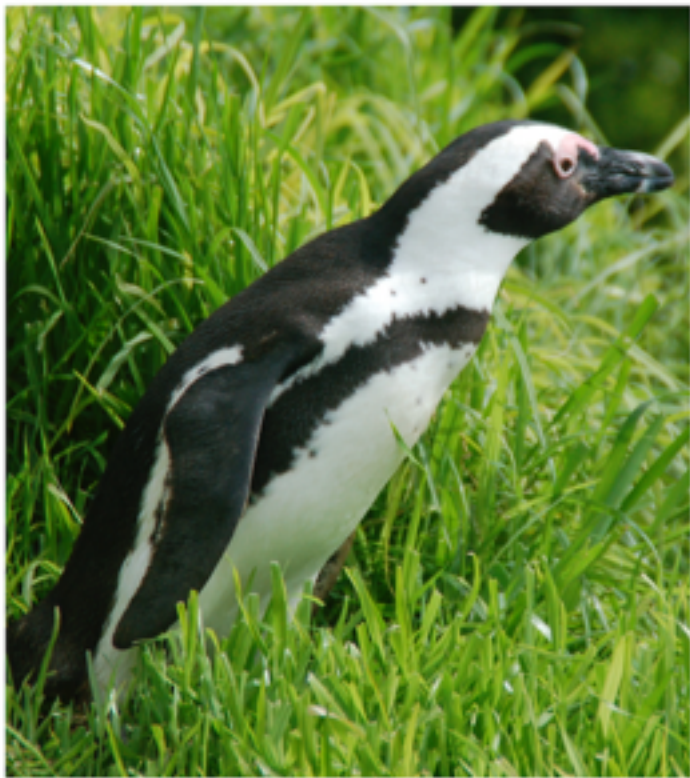
Home Work 1 - Sept 15.

Home Work 2 - Oct 6.

Home Work 3 - Oct 20.

Textbooks

R Programming
for Data Science



Roger D. Peng

leanpub.com/rprogramming

Exploratory Data
Analysis with R



Roger D. Peng

leanpub.com/exdata

Report Writing for
Data Science in R



Roger D. Peng

leanpub.com/reportwriting

Leanpub



Store Read Write Support Blog

Search Leanpub



96,483
READERS

182
PAGES



ENGLISH



PDF



EPUB



MOBI



APP

Edit

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](#).

FREE SAMPLE

R Programming for Data Science



Roger D. Peng

UPDATED 28 DAYS AGO

FREE!
MINIMUM

\$20.00
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/

Mastering Software Development in R

Roger D. Peng, Sean Kross, and Brooke Anderson

2016-08-31

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](#)

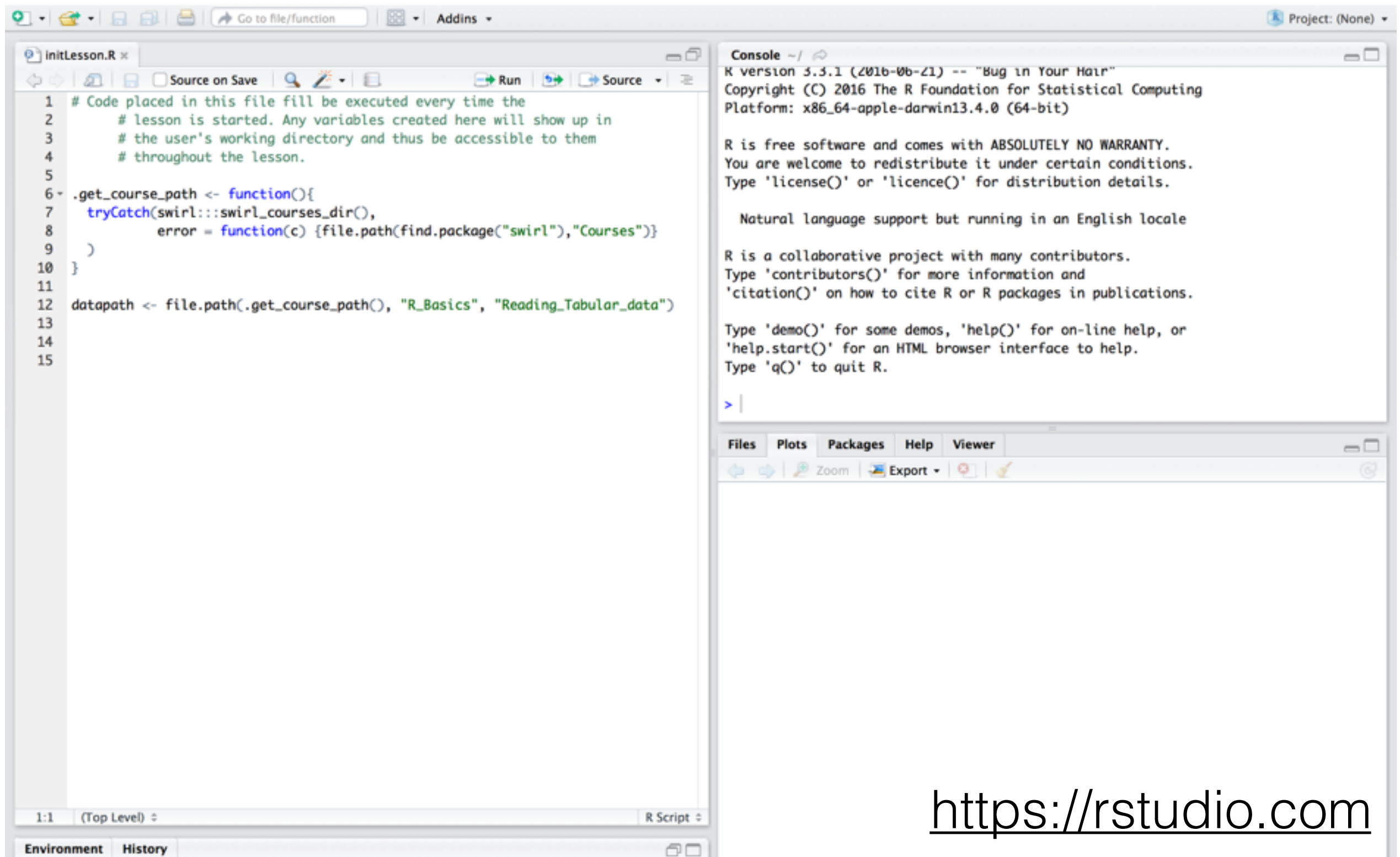
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

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



<https://rstudio.com>

RStudio



Products

Resources

Pricing

About Us

Blog



New Features R RStudio Serv

Learn More

Download a 45 day evaluation

RStudio

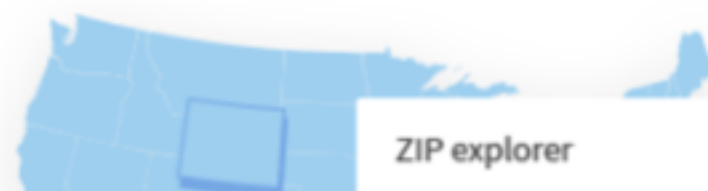
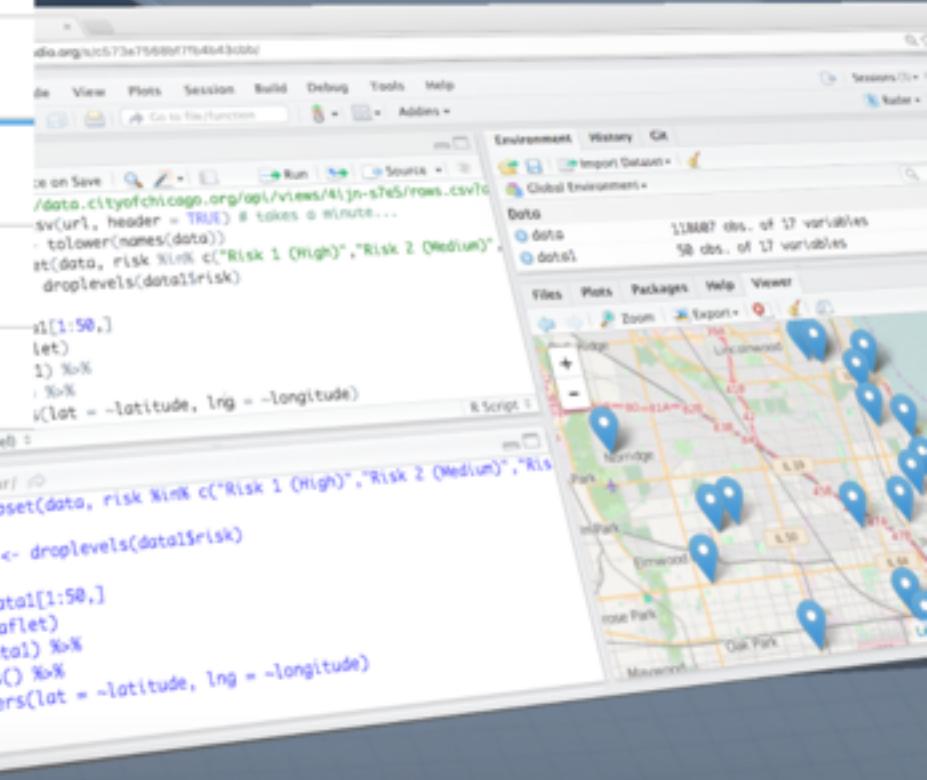
Shiny

R Packages

RStudio Server Pro

Shiny Server Pro

shinyapps.io



RStudio

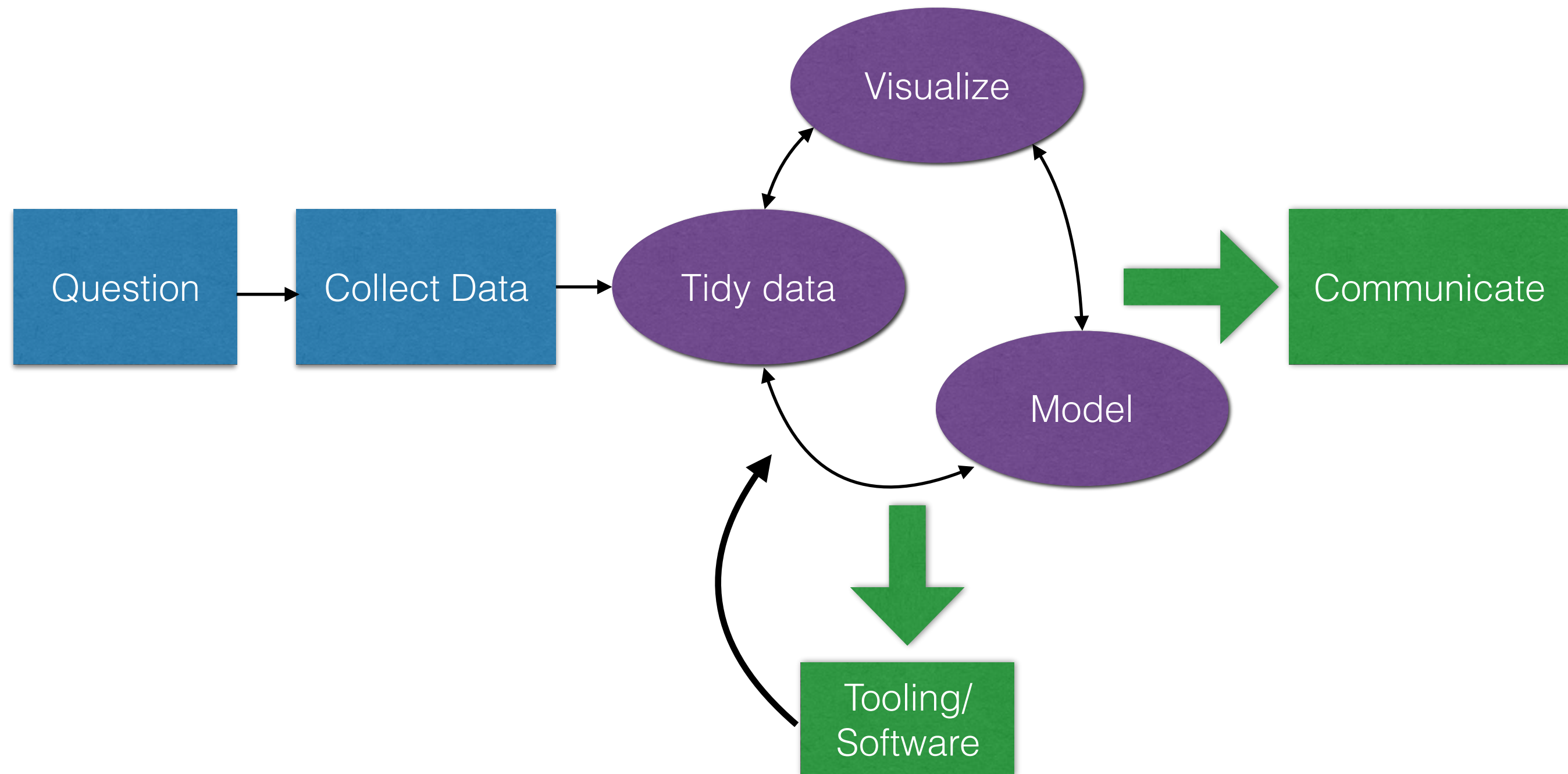
Choose Your Version of RStudio

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

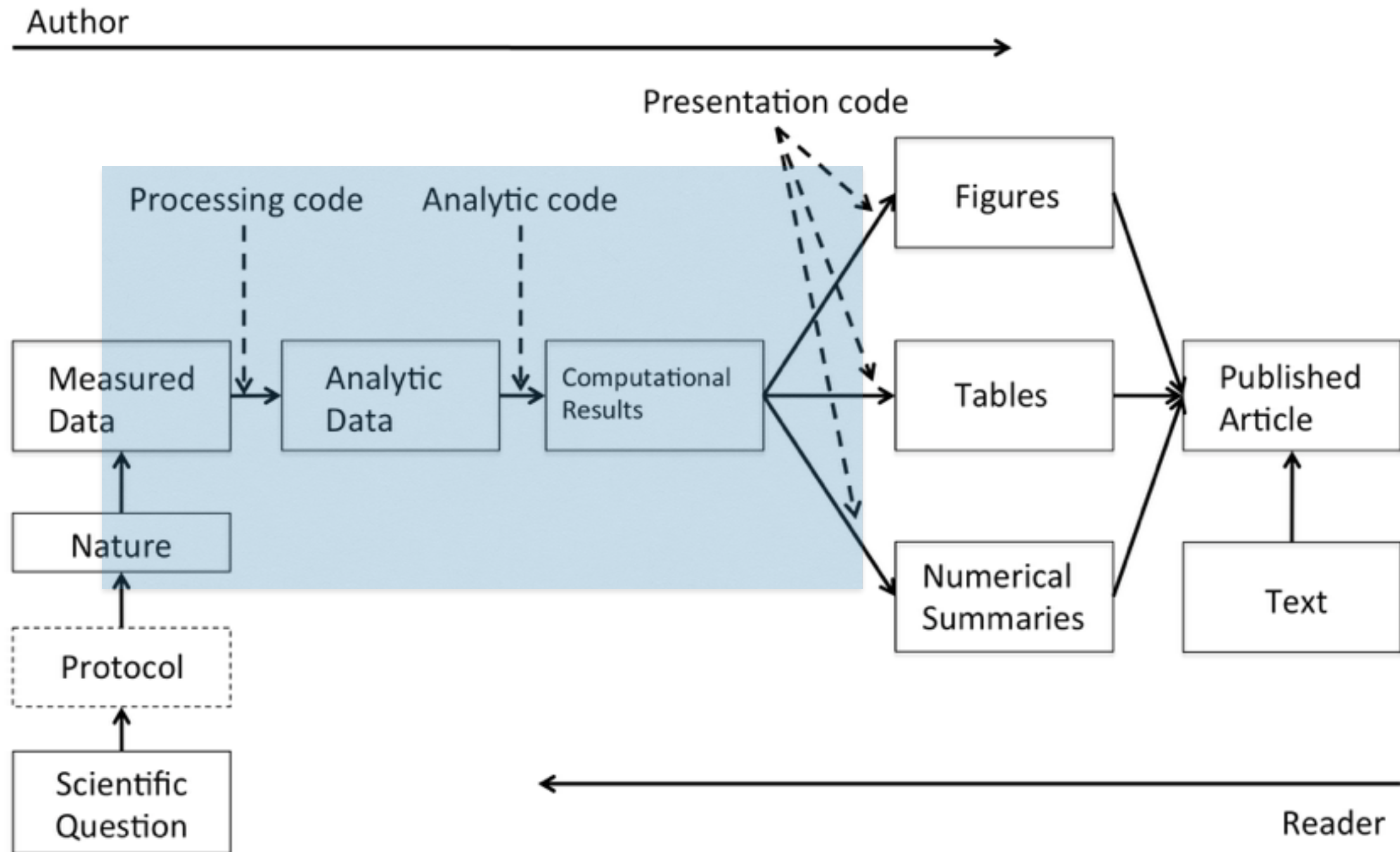
Data Science Workflow



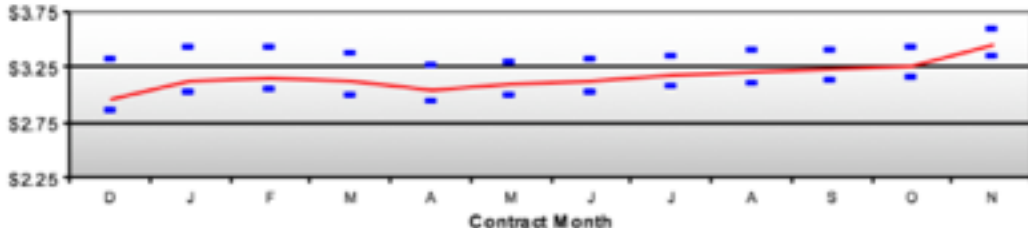
Major Themes

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

Reproducible Research?



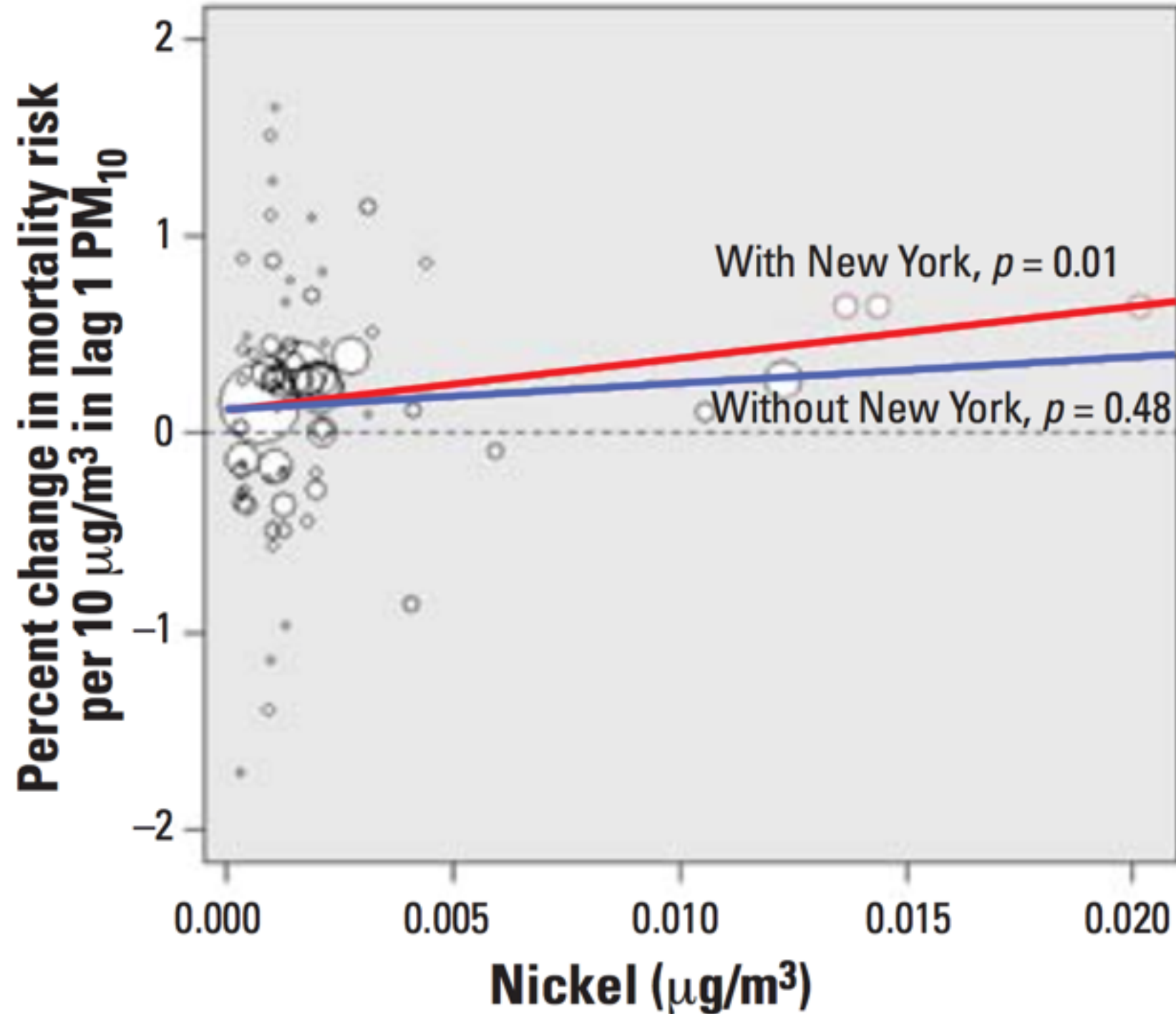
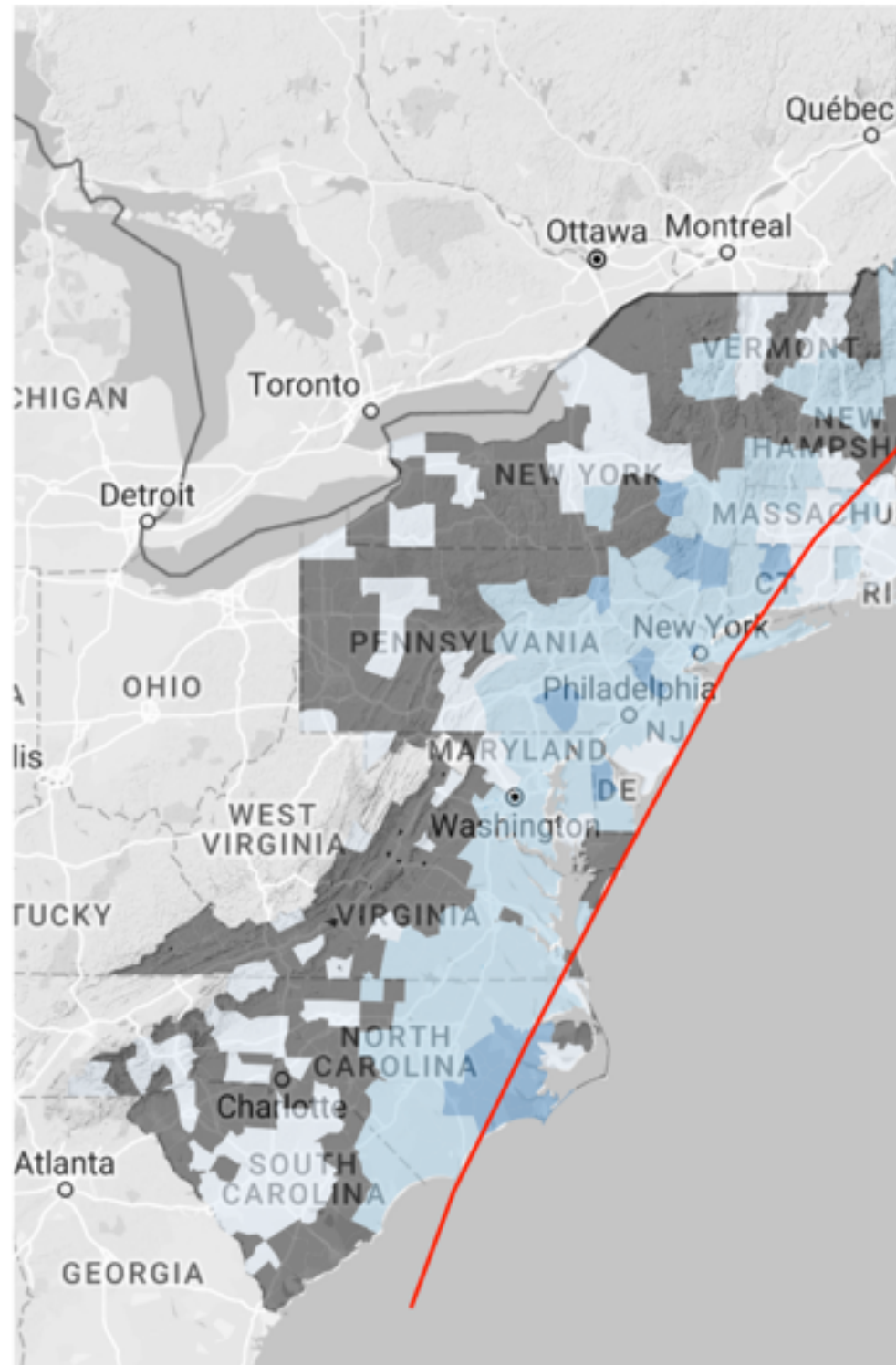
Tidying Data

	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																				
6		Houston Office										Regional Offices										
7		Barry Tycholiz (713) 853-1587										Mark Whitt (303) 575-6473										Denver
8		Kim Ward (713) 853-0685										Paul Lucci (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
12		Forward Prices (US\$/MMBtu)																				
14		Cash ROM		NYMEX		Forward NYMEX Strip with trailing 10-day highs/lows										IF NWPL Rocky Mountains						
16	SETTLE			Δ											Fixed Price Basis							
17																BID OFFER BID OFFER						
18				Dec-01	2.960	0.090											1.890 1.910					
19				Dec-01 to Mar-02	3.088	0.083											2.060 2.080					
20				Apr-02 to Oct-02	3.166	0.084											2.395 2.415 (0.565) (0.545)					
21				Nov-02 to Mar-03	3.651	0.090											2.594 2.614 (0.494) (0.474)					
22				One Year Strip*	3.165	0.084											2.581 2.601 (0.585) (0.565)					
23																	3.356 3.376 (0.295) (0.275)					
24																	2.634 2.654 (0.530) (0.510)					
25		Cash ROM				IF CIG Rocky Mountains										IF EL Paso Permian						
26						Fixed Price Basis										Fixed Price Basis						
27						BID OFFER BID OFFER										BID OFFER BID OFFER						
28				Dec-01			1.940 1.960										2.375 2.395					
29				Dec-01 to Mar-02			1.960 1.980										2.420 2.440					
30				Apr-02 to Oct-02			2.345 2.365 (0.615) (0.595)										2.700 2.720 (0.260) (0.240)					
31				Nov-02 to Mar-03			2.548 2.568 (0.540) (0.520)										2.855 2.875 (0.233) (0.213)					
32				One Year Strip*			2.471 2.491 (0.695) (0.675)										3.009 3.029 (0.158) (0.138)					
33							3.311 3.331 (0.340) (0.320)										3.499 3.519 (0.153) (0.133)					
34							2.551 2.571 (0.614) (0.594)										2.982 3.002 (0.182) (0.162)					
35		Cash ROM				AECO / NIT										IF NWPL Canadian Border (Sumas)						
36						Fixed Price Basis										Fixed Price Basis						
37						BID OFFER BID OFFER										BID OFFER BID OFFER						
38				Dec-01			2.376 2.396										2.480 2.500					
39				Dec-01 to Mar-02			2.398 2.418										2.460 2.480					
40				Apr-02 to Oct-02			2.552 2.572 (0.408) (0.388)										2.800 2.820 (0.160) (0.140)					
41				Nov-02 to Mar-03			2.616 2.636 (0.472) (0.452)										2.892 2.912 (0.196) (0.176)					
42				One Year Strip*			2.661 2.681 (0.505) (0.485)										2.796 2.816 (0.370) (0.350)					
43							3.216 3.236 (0.435) (0.415)										3.706 3.726 0.055 0.075					
44							2.676 2.696 (0.488) (0.468)										2.880 2.900 (0.285) (0.265)					
45																	IF PEPL TX-OK					
																	Fixed Price Basis					
																	BID OFFER BID OFFER					
																	2.530 2.550					
																	2.530 2.550					
																	2.828 2.848 (0.133) (0.113)					
																	2.958 2.978 (0.130) (0.110)					
																	3.046 3.066 (0.120) (0.100)					
																	3.531 3.551 (0.120) (0.100)					
																	3.041 3.061 (0.123) (0.103)					

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



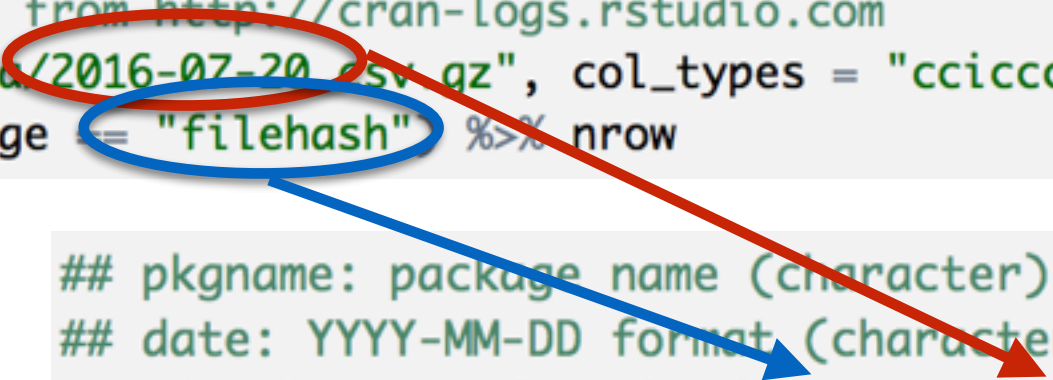
Programming and Abstraction

```
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

Movie explorer

Filter

Minimum number of reviews on Rotten Tomatoes

10 80 300

Year released

1,940 1,970 2,014

Minimum number of Oscar wins (all categories)

0 1 2 3 4

Dollars at Box Office (millions)

0 800

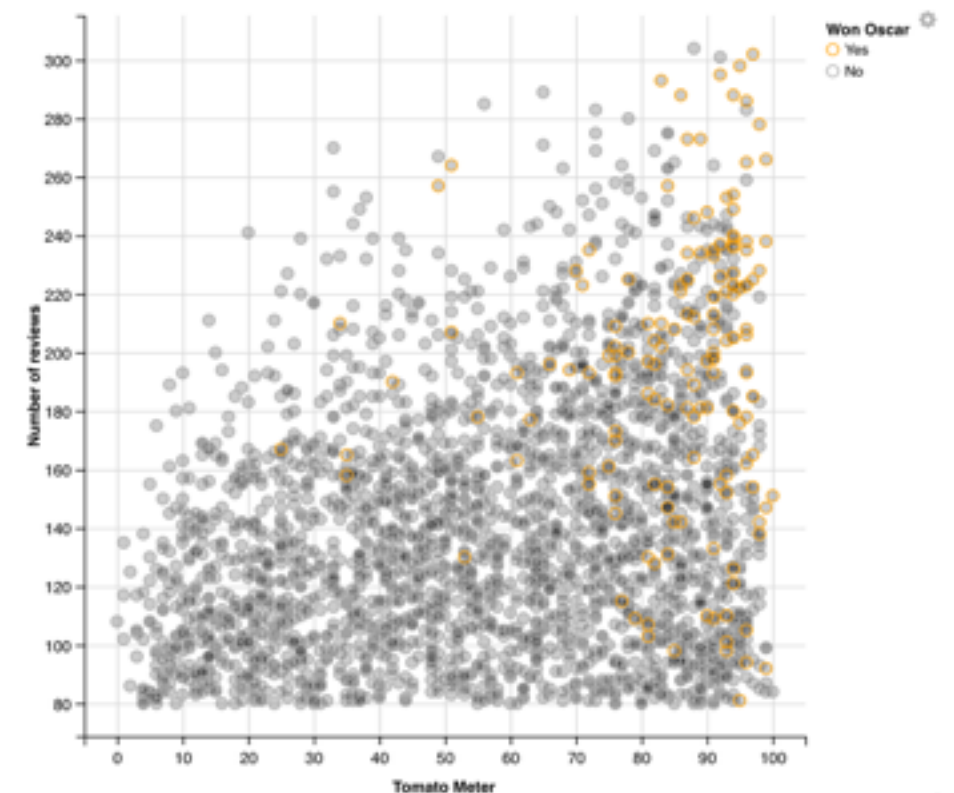
Genre (a movie can have multiple genres)

All

Director name contains (e.g., Miyazaki)

Cast names contains (e.g. Tom

Shiny app



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