Statistical Computing: The Course

Biostatistics 140.776
Roger D. Peng

rdpeng.github.io/Biostat776

About Me

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

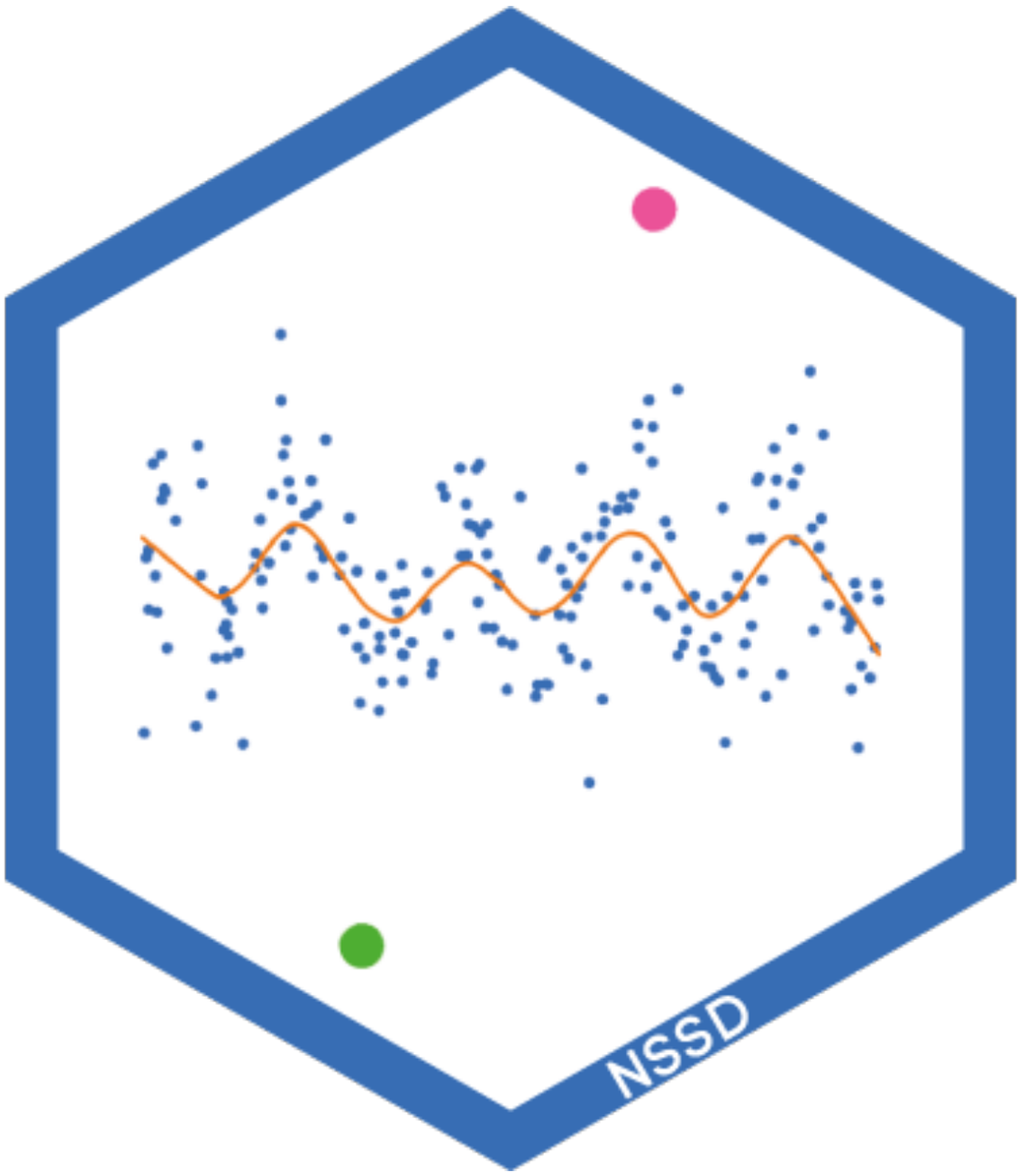
About Me

- 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

@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

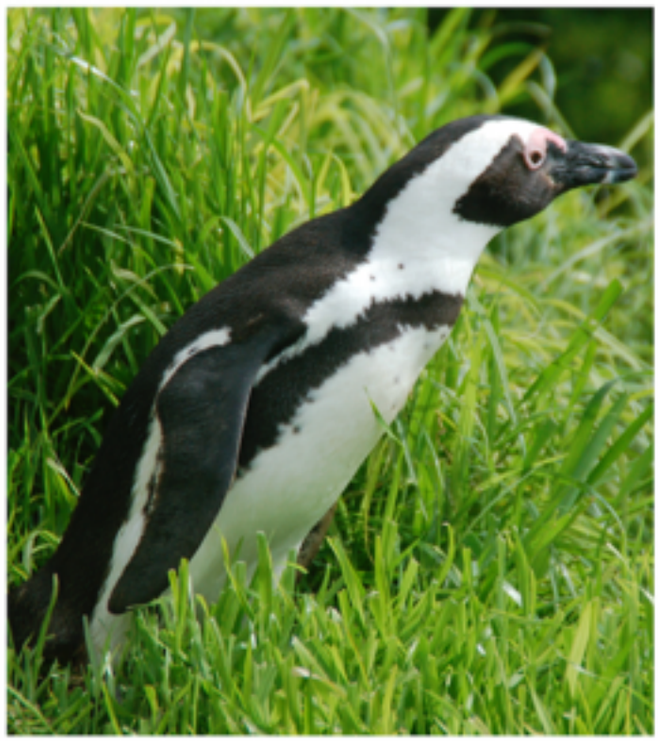
Subscribe in iTunes: http://goo.gl/hz3AHL

Course Logistics

- **Instructor**: Roger D. Peng (Dept. of Biostatistics)
- **Web Site**: rdpeng.github.io/Biostat776
- **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

Exploratory Data Analysis with R

Roger D. Peng

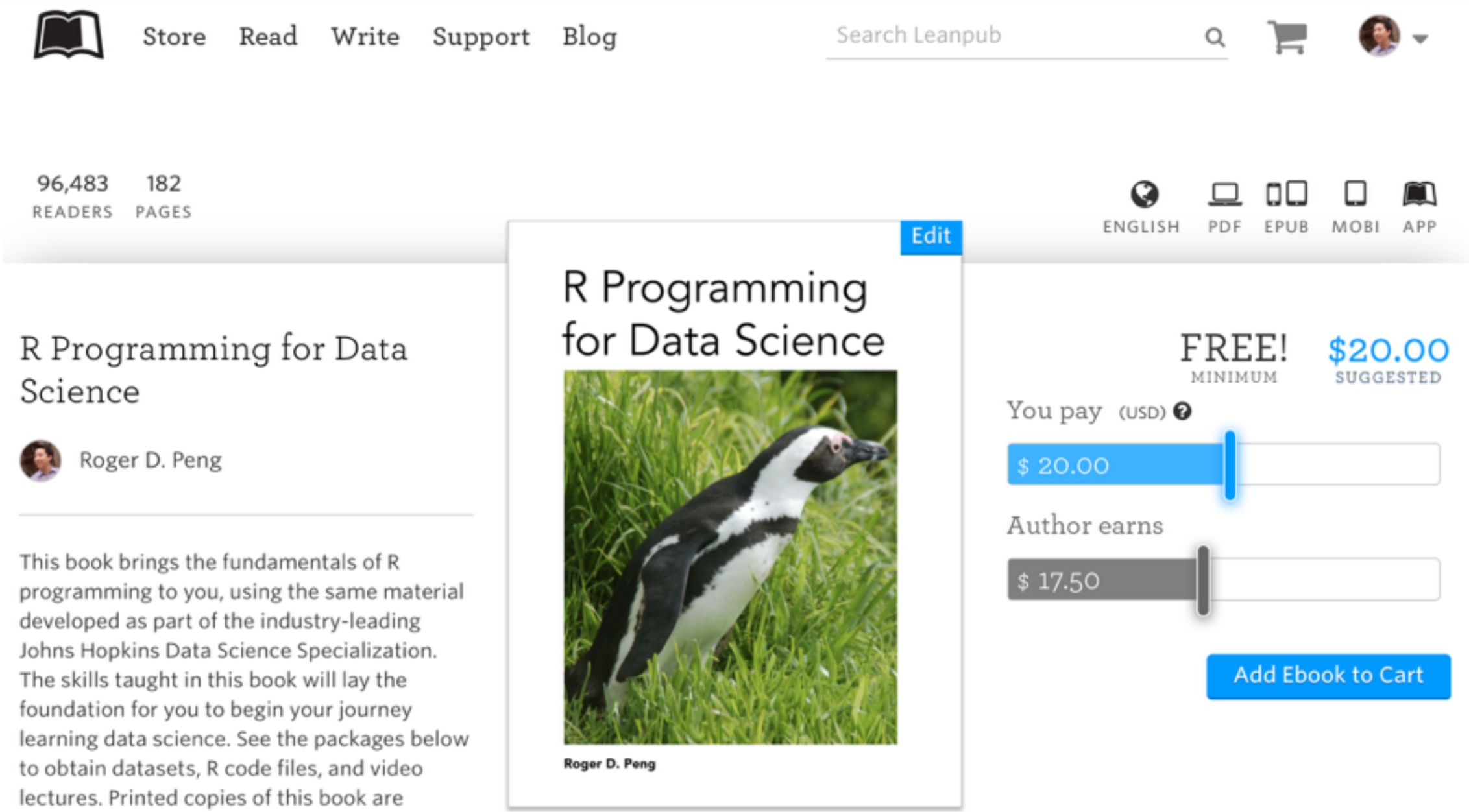
leanpub.com/exdata

Report Writing for Data Science in R

Roger D. Peng

leanpub.com/reportwriting

Leanpub



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

UPDATED 28 DAYS AGO

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

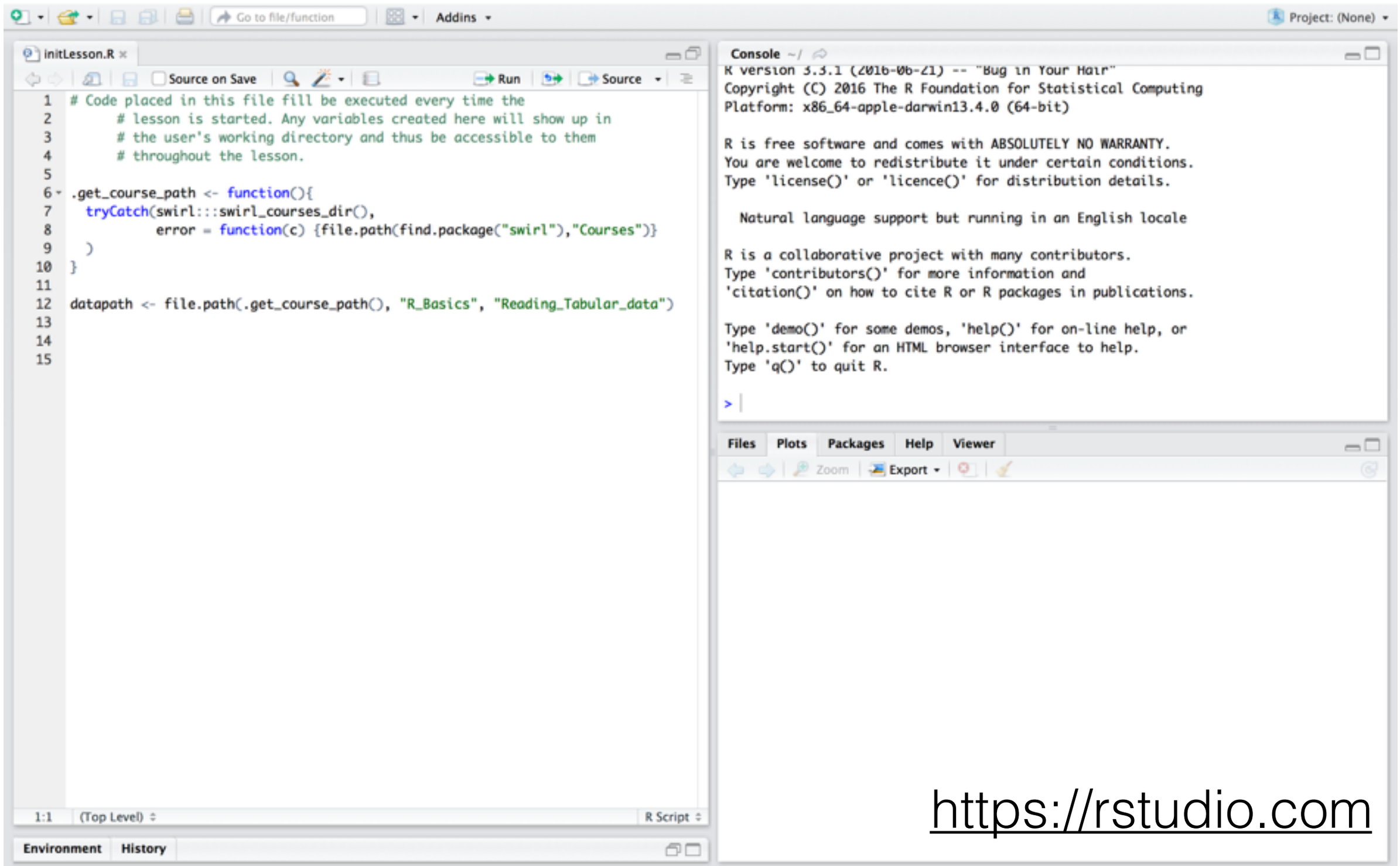
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

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!

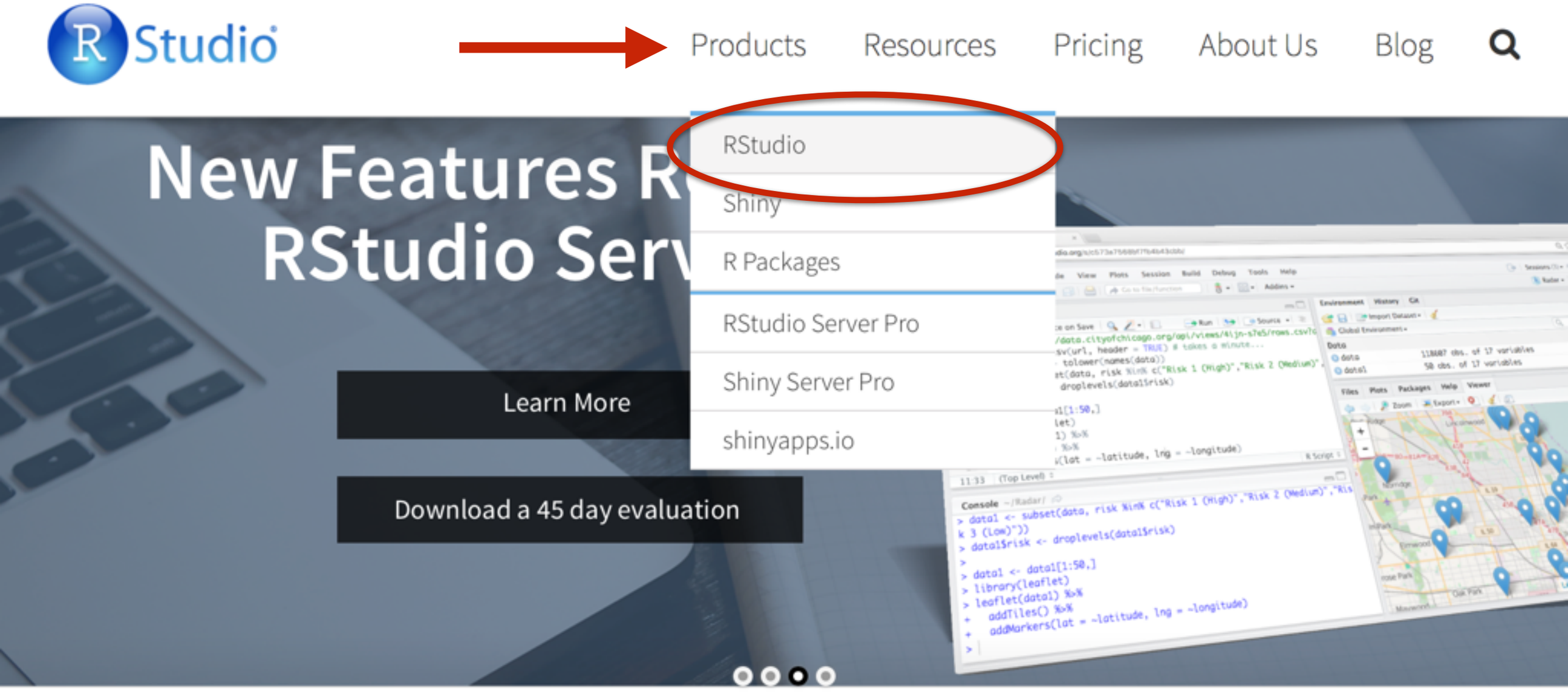
- The latest release (Tuesday 2016-06-21, Bug in Your Hair) R-3.3.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available 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


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
CODE
WORKSPACE

ZIP explorer


rmarkdown
Shiny
tidyr

RStudio

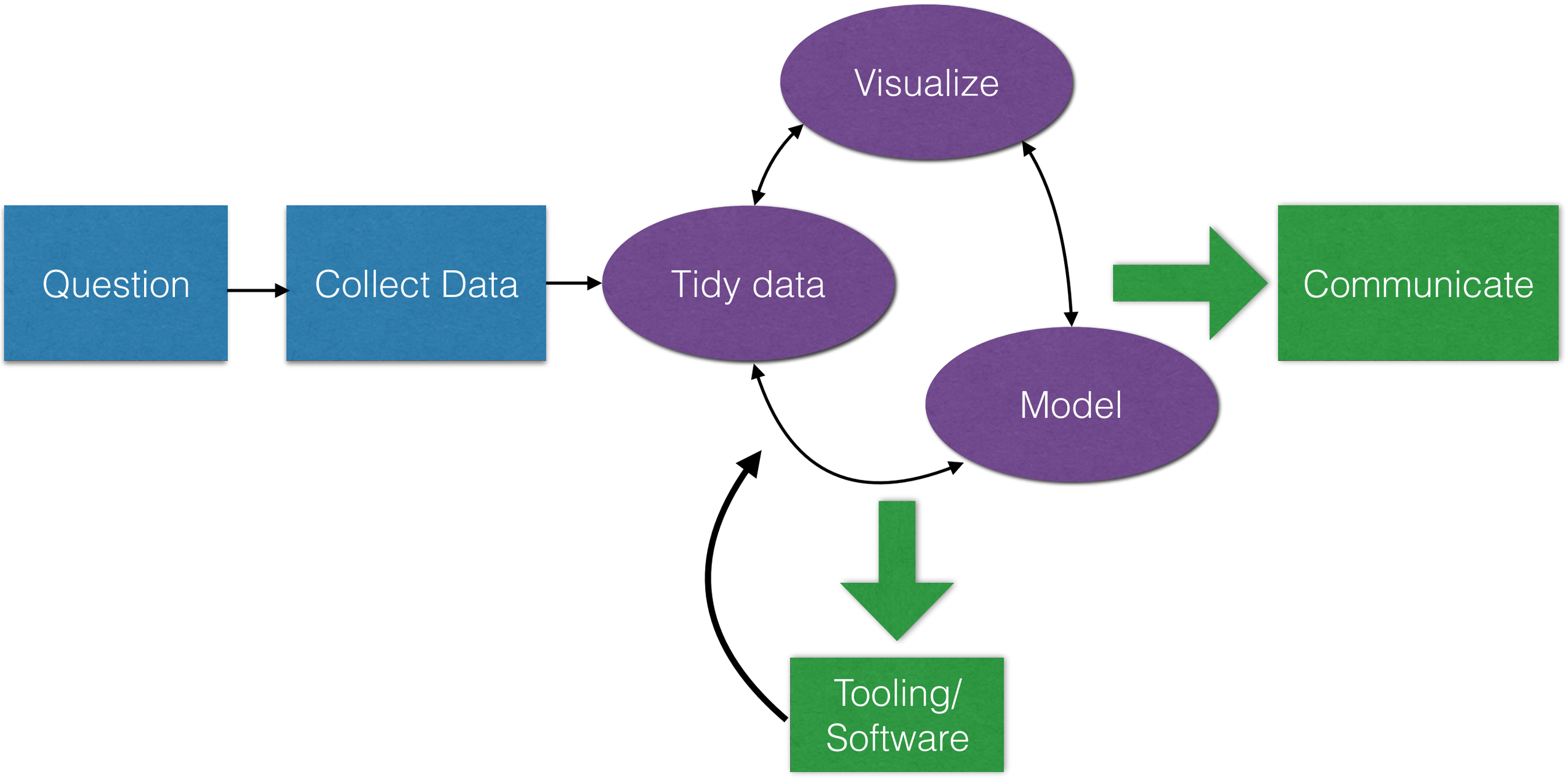
Choose Your Version of RStudio

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

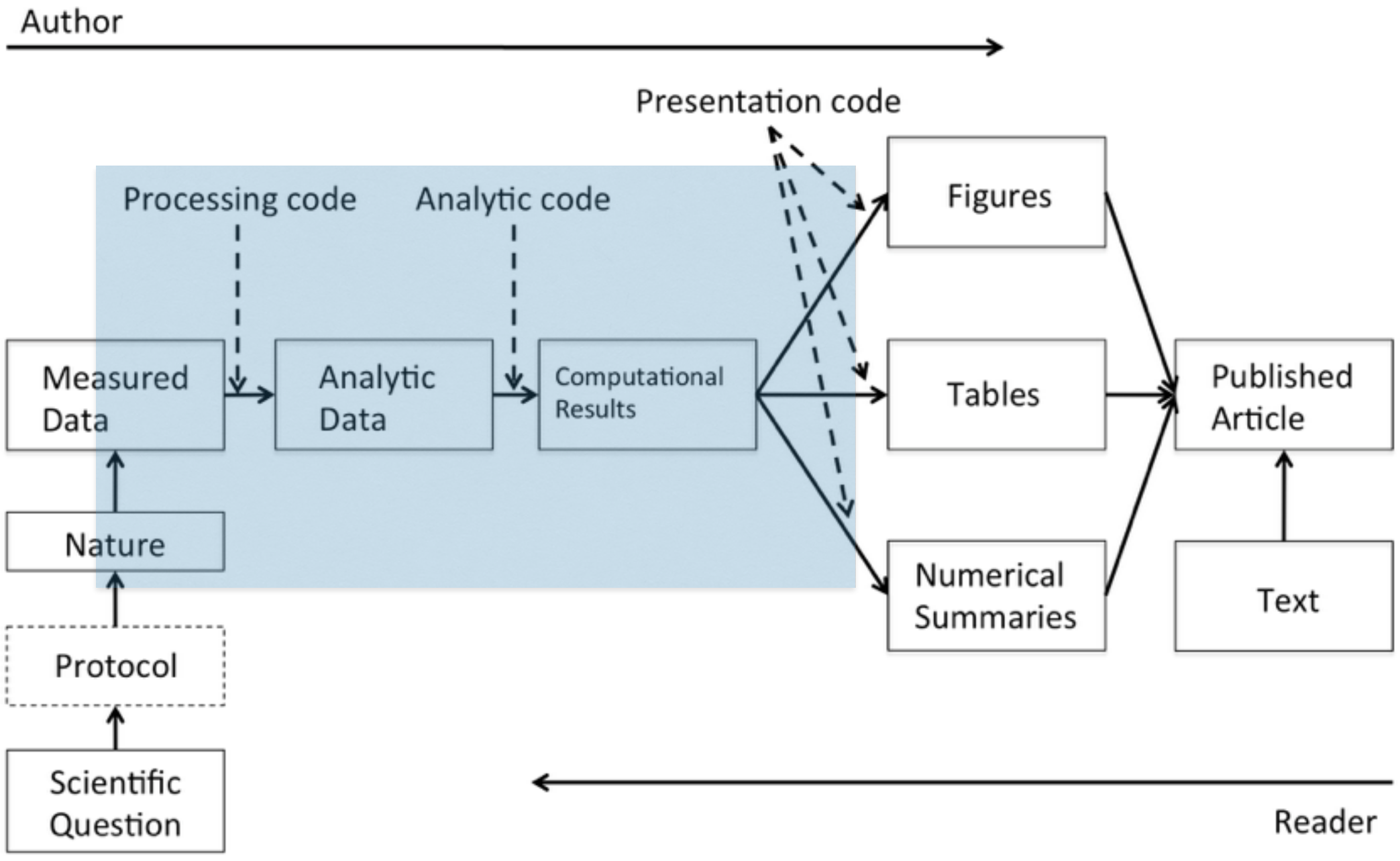
Data Science Workflow



Major Themes

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

Reproducible Research?



Tidying Data

Enron North America - West Gas

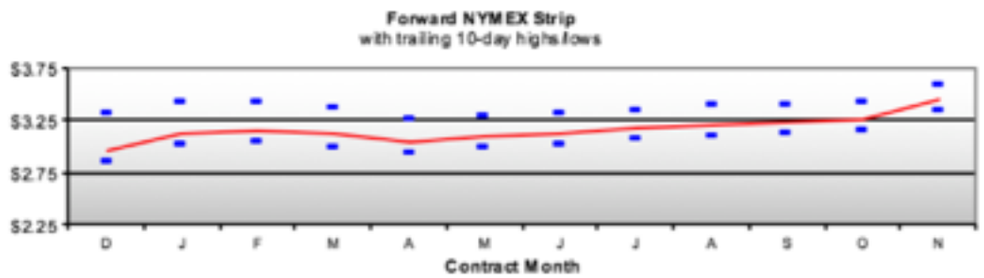
November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX		IF NWPL Rocky Mountains			
			Fixed Price		Basis	
	SETTLE	Δ	BID	OFFER	BID	OFFER
Cash			1.890	1.910		
ROM			2.060	2.080		
Dec-01	2.960	0.090	2.395	2.415	(0.565)	(0.545)
Dec-01 to Mar-02	3.088	0.083	2.594	2.614	(0.494)	(0.474)
Apr-02 to Oct-02	3.166	0.084	2.581	2.601	(0.585)	(0.565)
Nov-02 to Mar-03	3.651	0.090	3.356	3.376	(0.295)	(0.275)
One Year Strip*	3.165	0.084	2.634	2.654	(0.530)	(0.510)



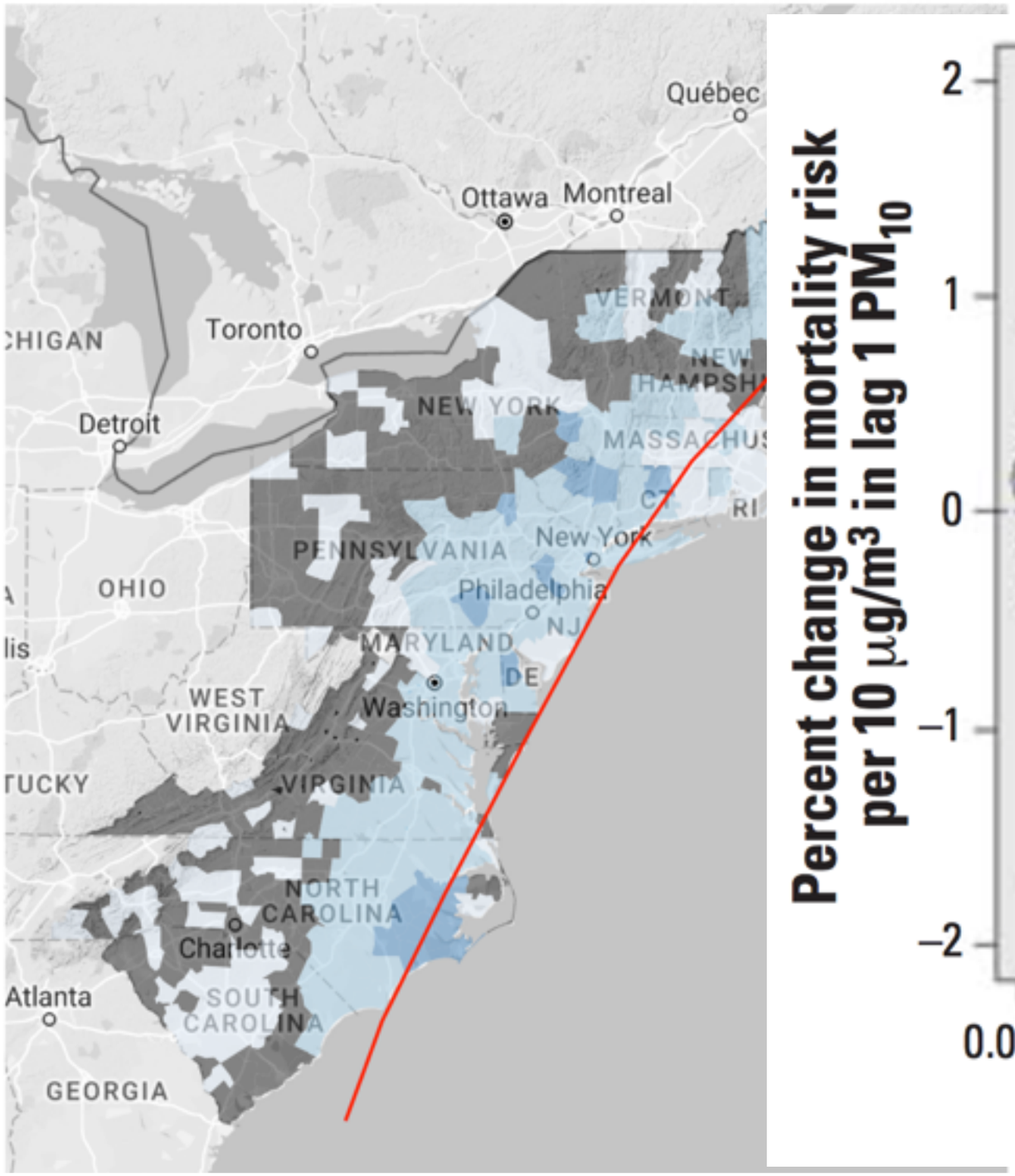
	IF CIG Rocky Mountains				IF EL Paso Permian				IF EL Paso San Juan			
	Fixed Price		Basis		Fixed Price		Basis		Fixed Price		Basis	
	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER
Cash	1.940	1.960			2.375	2.395			2.450	2.470		
ROM	1.960	1.980			2.420	2.440			2.350	2.370		
Dec-01	2.345	2.365	(0.615)	(0.595)	2.700	2.720	(0.260)	(0.240)	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)	2.855	2.875	(0.233)	(0.213)	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)	3.009	3.029	(0.158)	(0.138)	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)	3.499	3.519	(0.153)	(0.133)	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.551	2.571	(0.614)	(0.594)	2.982	3.002	(0.182)	(0.162)	2.817	2.837	(0.347)	(0.327)

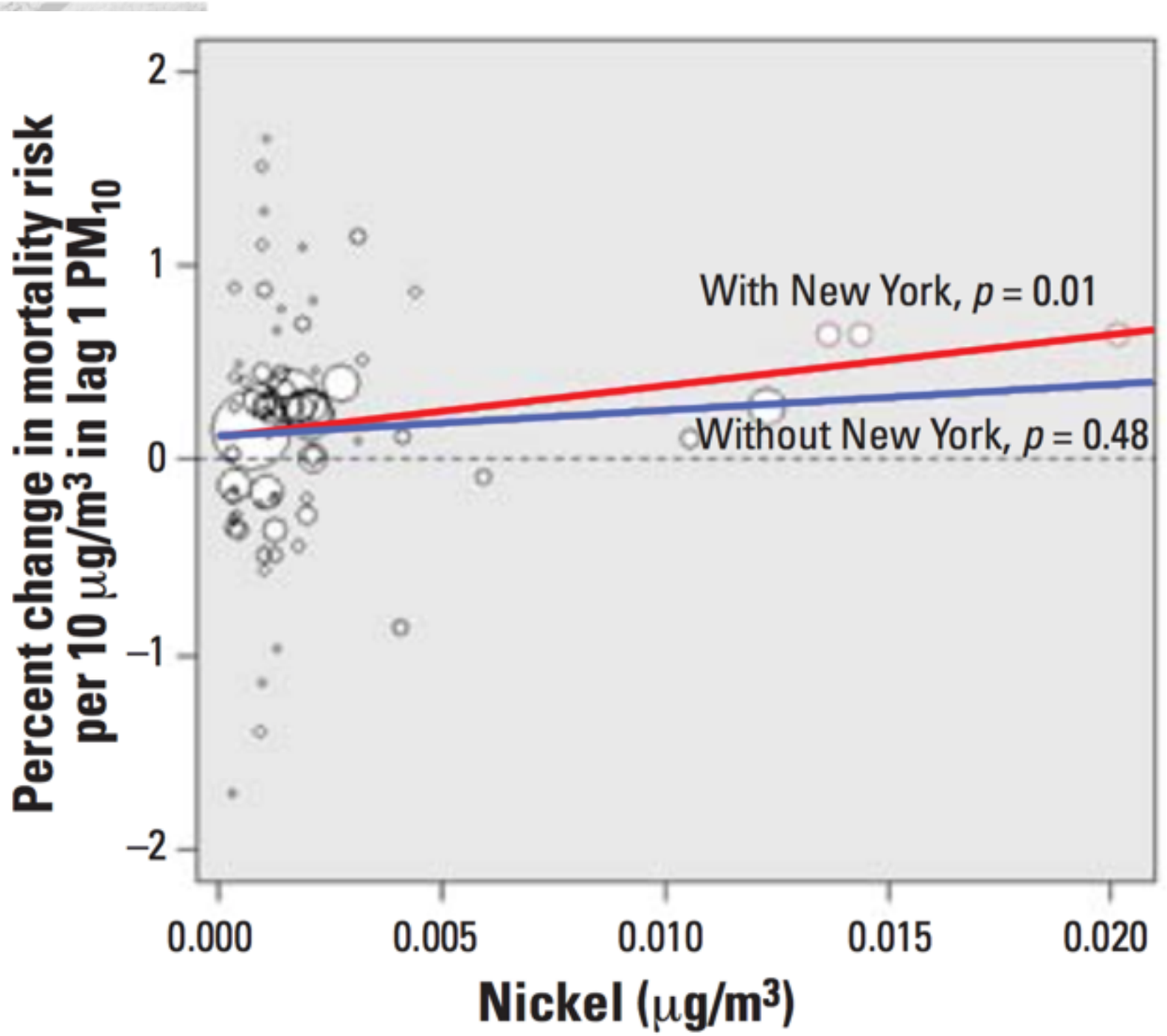
	AECO / NIT				IF NWPL Canadian Border (Sumas)				IF PEPL TX-OK			
	Fixed Price		Basis		Fixed Price		Basis		Fixed Price		Basis	
	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER
Cash	2.376	2.396			2.480	2.500			2.530	2.550		
ROM	2.398	2.418			2.460	2.480			2.530	2.550		
Dec-01	2.552	2.572	(0.408)	(0.388)	2.800	2.820	(0.160)	(0.140)	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)	2.892	2.912	(0.196)	(0.176)	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)	2.796	2.816	(0.370)	(0.350)	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)	3.706	3.726	0.055	0.075	3.531	3.551	(0.120)	(0.100)
One Year Strip*	2.676	2.696	(0.488)	(0.468)	2.880	2.900	(0.285)	(0.265)	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

```r
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
```

```r
## 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)

Function 1

Function 2

Function 3

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