Lessons from teaching data science to over a million people

Sean Kross

CRUNCH Conference
Budapest
2017-10-20

A little about me

- Formerly: The Johns Hopkins Data Science Lab
- Currently: The University of California San Diego
- Main interests:
  - Data Science
  - Online Education
  - Open Science

NYU

UNIVERSITY OF MARYLAND 18 56

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Penn UNIVERSITY of PENNSYLVANIA

UC San Diego

The Johns Hopkins Data Science Lab

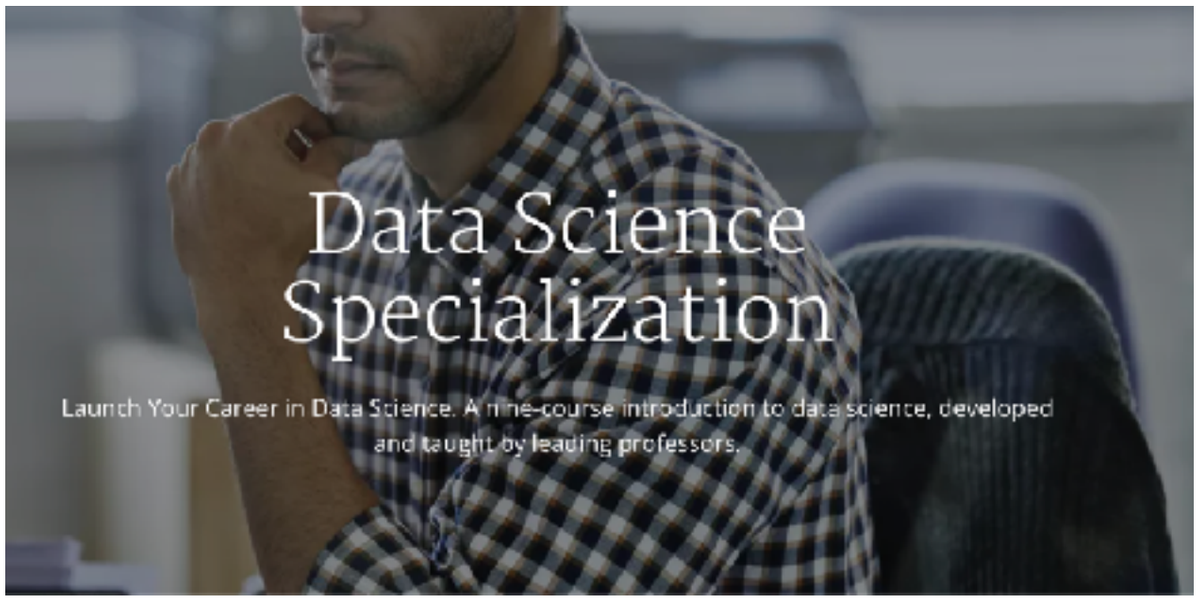
Jeff Leek
@jtleek

Roger Peng
@rdpeng

Brian Caffo
@bcaffo

jhudatascience.org

coursera


Data Science Specialization
Launch Your Career in Data Science. A nine-course introduction to data science, developed and taught by leading professors.


Executive Data Science Specialization
Be The Leader Your Data Team Needs. Learn to lead a data science team that generates first-rate analyses in four courses.


Mastering Software Development in R Specialization
Build the Tools for Better Data Science. Learn to design software for data tooling, distribute R packages, and build custom visualizations

?

Part 1: The Data Science Specialization

‹ All Collections

Collection idea for us?

Practical Data Science for Stats - a PeerJ Collection

Data Science | Statistics | Scientific Computing and Simulation | Computer Education | Computational Science | Social Computing | Software Engineering | Science and Medical Education | Computational Biology | Human-Computer Interaction | Anthropology | Programming Languages | Visual Analytics | Graphics | Data Mining and Machine Learning

September 27, 2017 preprint

Forecasting at scale

1,145 downloads 3,615 views

Sean J Taylor, Benjamin Letham

https://doi.org/10.7287/peerj.preprints.3190v2

September 1, 2017 preprint

How to share data for collaboration

768 downloads 3,430 views

Shannon E Ellis, Jeffrey T Leek

https://doi.org/10.7287/peerj.preprints.3139v5

Practical Data Science for Stats

The "Practical Data Science for Stats" Collection contains preprints focusing on the practical side of data science workflows and statistical analysis. Curated by Jennifer Bryan and Hadley Wickham.

There are many aspects of day-to-day analytical work that are almost absent from the conventional statistics literature and curriculum. And yet these activities account for a considerable share of the time and effort of data analysts and applied statisticians.

The goal of this collection is to

https://peerj.com/collections/50-practicaldatascistats/

- **Forecasting at Scale** by Sean Taylor and Benjamin Letham (Facebook)

- **How to Share Data for Collaboration** by Shannon Ellis and Jeffrey Leek (Johns Hopkins Data Science Lab)

- **Opinionated Analysis Development** by Hilary Parker (Stitch Fix)

- **Data Organization in Spreadsheets** by Karl Broman and Kara Woo (The University of Wisconsin & DataCamp)

Rationale: “Let’s put in-person courses online to augment in-person teaching.”

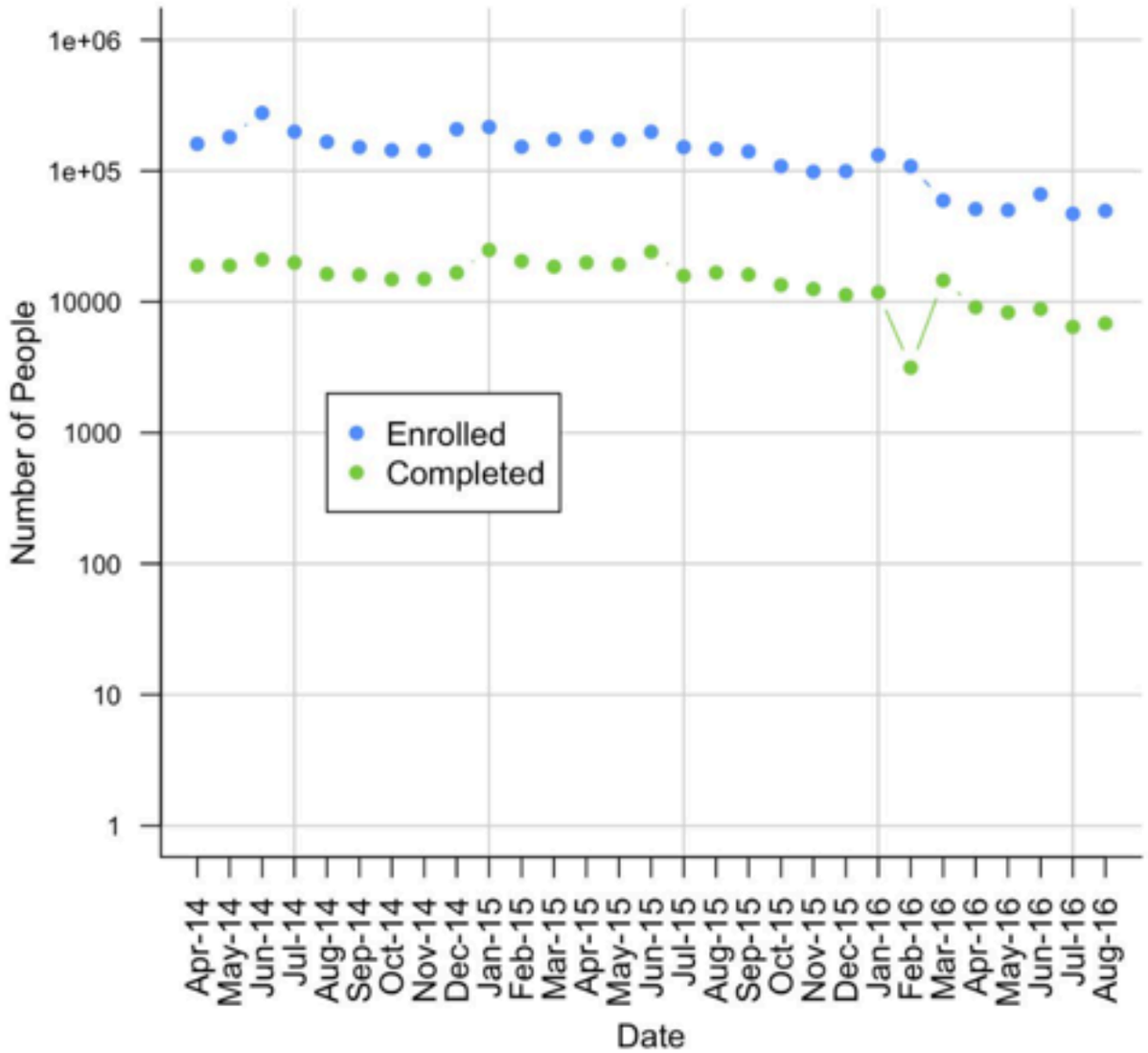
Course	Enrollment	Completions
Mathematical Biostatistics Boot Camp 1	109,789	4,150
Mathematical Biostatistics Boot Camp 2	23,842	944
Computing for Data Analysis	243,987	21,069
Data Analysis	193,126	6,500
Biostatistics Case Study	39,140	3,322

We were on to something.

Nine Courses

1. The Data Scientist's Toolbox
2. R Programming
3. Getting and Cleaning Data
4. Exploratory Data Analysis
5. Reproducible Research
6. Statistical Inference
7. Regression Models
8. Practical Machine Learning
9. Developing Data Products

Enrollment and Completions of the Data Science Specialization



Key innovations

Give everything away for free

Leanpub

YouTube

GitHub

Regression Models for Data Science In R

Brian Caffo

The Unix Workbench

Sean Kross

Mastering Software Development in R

Roger D Peng Sean Kross Brooke Anderson

Developing Data Products In R

DDP

Brian Caffo

R Programming for Data Science

Roger D. Peng

The Elements of Data Analytic Style

Jeff Leek

Report Writing for Data Science in R

Roger D. Peng

EXECUTIVE DATA SCIENCE

A GUIDE TO TRAINING AND MANAGING THE BEST DATA SCIENTISTS

Statistical inference for data science

Brian Caffo

How to be a modern scientist

@jtleek

Exploratory Data Analysis with R

Roger D. Peng

Capstones -> Portfolios -> Jobs

Run *every* course *every* month

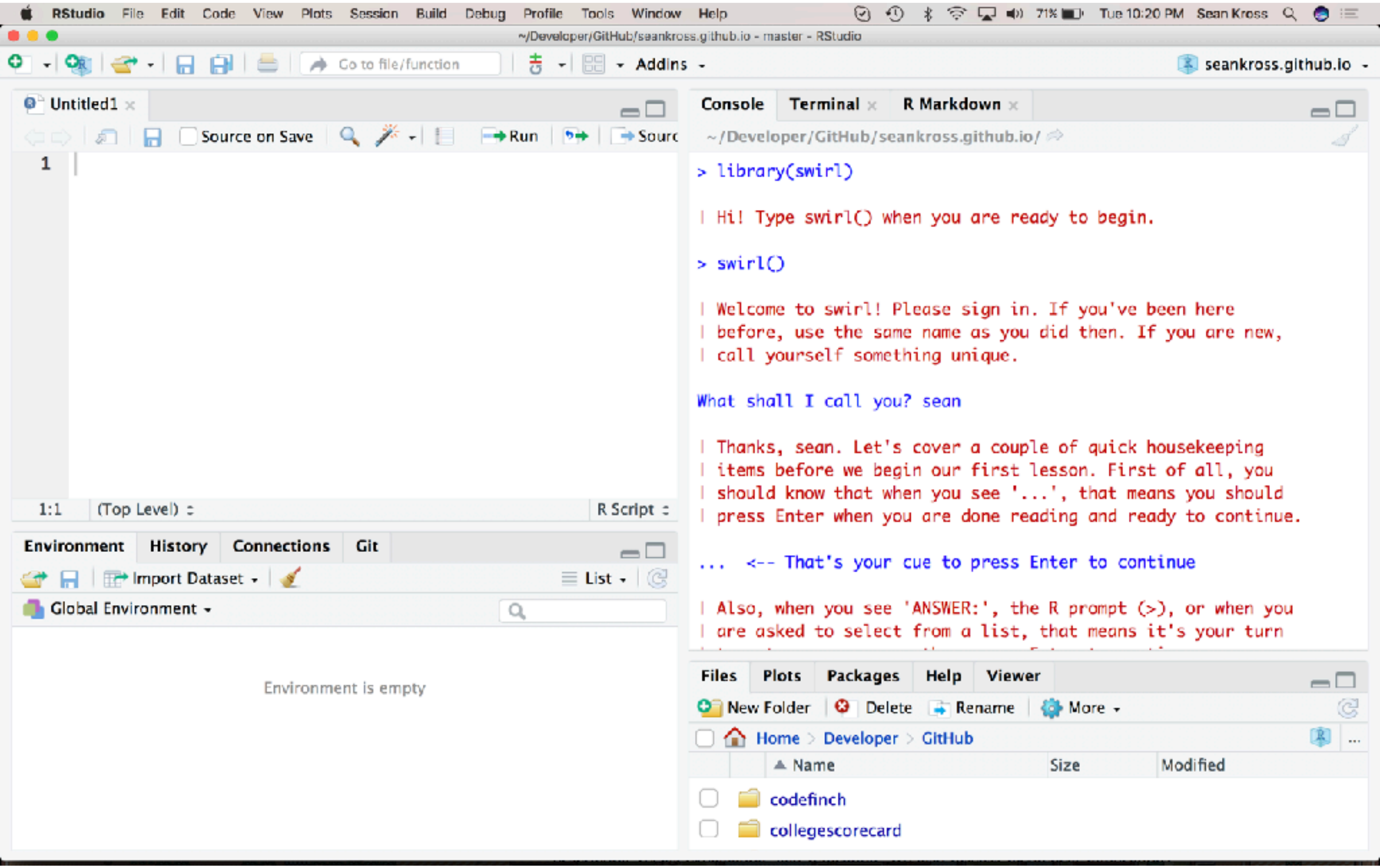
Integrate content

https://ubc-mds.github.io/

{swirl}

Learn R, in R.

swirl teaches you R programming and data science
interactively, at your own pace, and right in the R console!


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
71% Tue 10:20 PM Sean Kross
~/Developer/GitHub/seankross.github.io - master - RStudio
Go to file/function
Addins
seankross.github.io
Untitled1
Source on Save
Run
Sourc
1
1:1
(Top Level)
R Script
Environment
History
Connections
Git
Import Dataset
List
Global Environment
Environment is empty
Console
Terminal
R Markdown
~/Developer/GitHub/seankross.github.io/
> library(swirl)
| Hi! Type swirl() when you are ready to begin.
> swirl()
| Welcome to swirl! Please sign in. If you've been here
| before, use the same name as you did then. If you are new,
| call yourself something unique.
What shall I call you? sean
| Thanks, sean. Let's cover a couple of quick housekeeping
| items before we begin our first lesson. First of all, you
| should know that when you see '...', that means you should
| press Enter when you are done reading and ready to continue.
... <-- That's your cue to press Enter to continue
| Also, when you see 'ANSWER:', the R prompt (>), or when you
| are asked to select from a list, that means it's your turn
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Developer
GitHub
Name
Size
Modified
codefinch
collegescorecard

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help 71% Tue 10:21 PM Sean Kross

~/Developer/GitHub/seankross.github.io - master - RStudio

Go to file/function | Addins | seankross.github.io

Untitled1

Source on Save | Run | Sourc

```
1
```

1:1 (Top Level) | R Script

Environment | History | Connections | Git

Import Dataset | List

Global Environment

Values

num_vect	num [1:4] 0.5 55 -10 6
tf	logi [1:4] TRUE FALSE TRUE FALSE

Console | Terminal | R Markdown

~/Developer/GitHub/seankross.github.io/

```
| values 0.5, 55, -10, and 6.

> num_vect <- c(0.5, 55, -10, 6)

| That's the answer I was looking for.

  |=======                                                |  13%
| Now, create a variable called tf that gets the result of
| num_vect < 1, which is read as 'num_vect is less than 1'.

> tf <- num_vect < 1

| That's correct!

  |========                                               |  16%
| What do you think tf will look like?

1: a single logical value
2: a vector of 4 logical values

Selection:
```

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > Developer > GitHub

	Name	Size	Modified
	codefinch		
	collegescorecard		

Lessons from Alumni: The Data Science Specialization

- Data scientists want to create online artifacts. See Stitch Fix and Stack Overflow's technical blogs. Also see open source software projects like Prophet, a forecasting library from Facebook.

- Data Scientists want to be able to do in-house data science training.

- Domain expertise is important but so is buy-in from management.

Part 2: Executive Data Science

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

TechRepublic

BIG DATA

How companies can develop internal data science expertise instead of hiring more Ph.D.s

TC News Video Events Crunchbase Search

Developer Airbnb artificial intelligence technology vacation rental

Airbnb is running its own internal university to teach data science

Zillow®

DataCamp

Imputation Update

```
To: sprmngr@zillow.com
From: blucjay@zillow.com
Date: 04/04/16 12:56:33
Subject: Imputation Update

I imputed the missing values in the properties dataset and
the results look reasonable. I set up a meeting for you and
Chelsea to discuss the details of the prediction model. It's
in the Sherman conference room at 2pm.

- Jay
```

Possible Answers

1. Sounds good.
2. Great work Jay, thanks.
3. Thanks. If anything comes up let me know.

Submit Answer

Introducing the Variables

Got it!

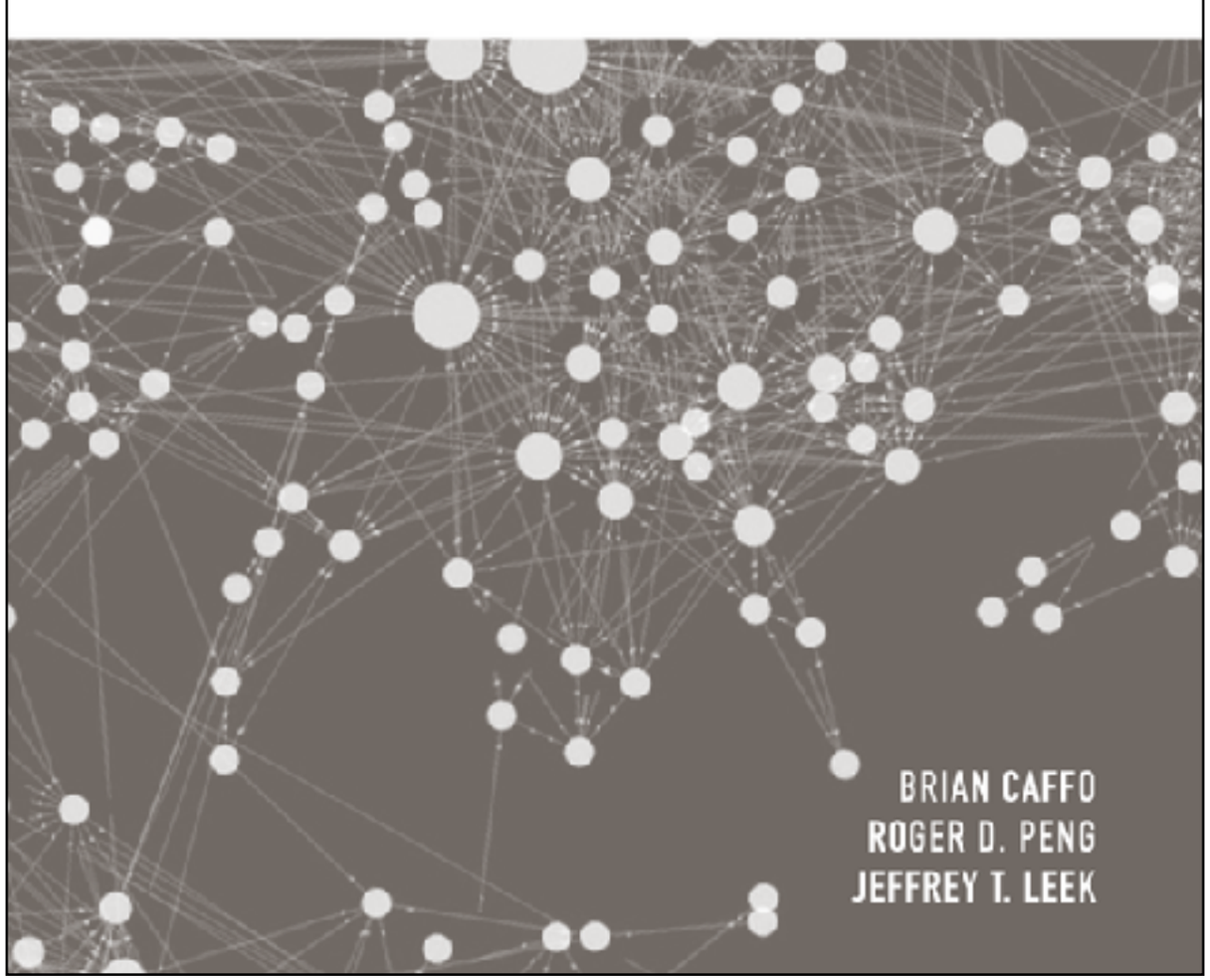
Considering Time

What are the implications of using time as a variable in our price model?

- I: The value of a home could change over time even if nothing else about the home changed.
- II: The value of a home shouldn't change over time especially when nothing else about the home changed.
- III: Time is likely a important predictor and therefore a model that accounts for time is likely more accurate.
- IV: If time is an important predictor a model that includes time will likely be less accurate.

EXECUTIVE DATA SCIENCE

A GUIDE TO TRAINING AND MANAGING THE BEST DATA SCIENTISTS



BRIAN CAFFO
ROGER D. PENG
JEFFREY T. LEEK

Lessons from Alumni: Executive Data Science

- Data science technologies tend to “trickle up.” (Especially good to know if you develop data science technologies.)

- Invest your precious time into basic statistics over basic programming.

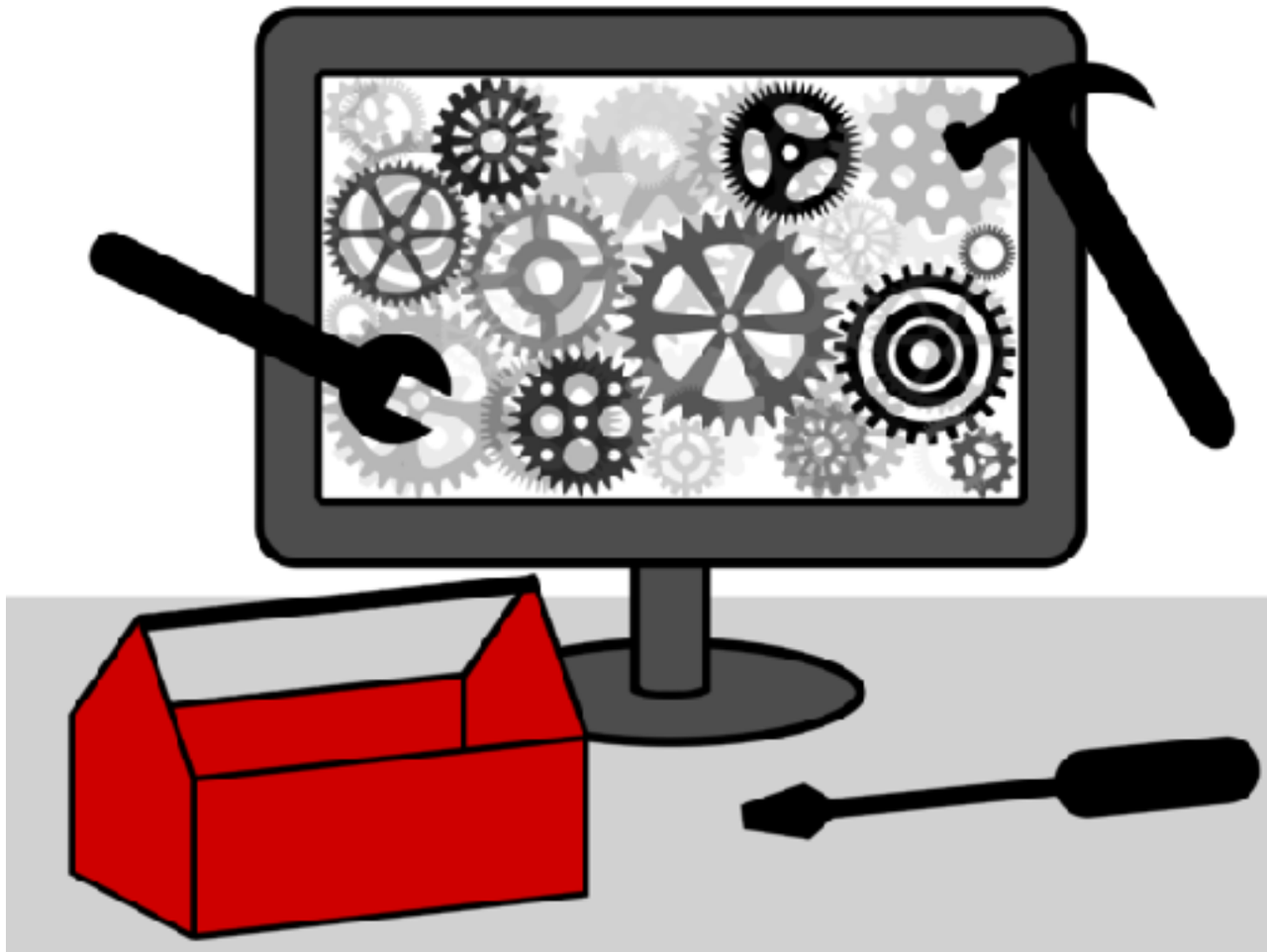
- Your expectations for developing an analysis should resemble your expectations for developing software.

Part 3: Mastering Software Development in R

Data Scientist

Data Engineer

Mastering Software Development in R



Roger D Peng Sean Kross Brooke Anderson

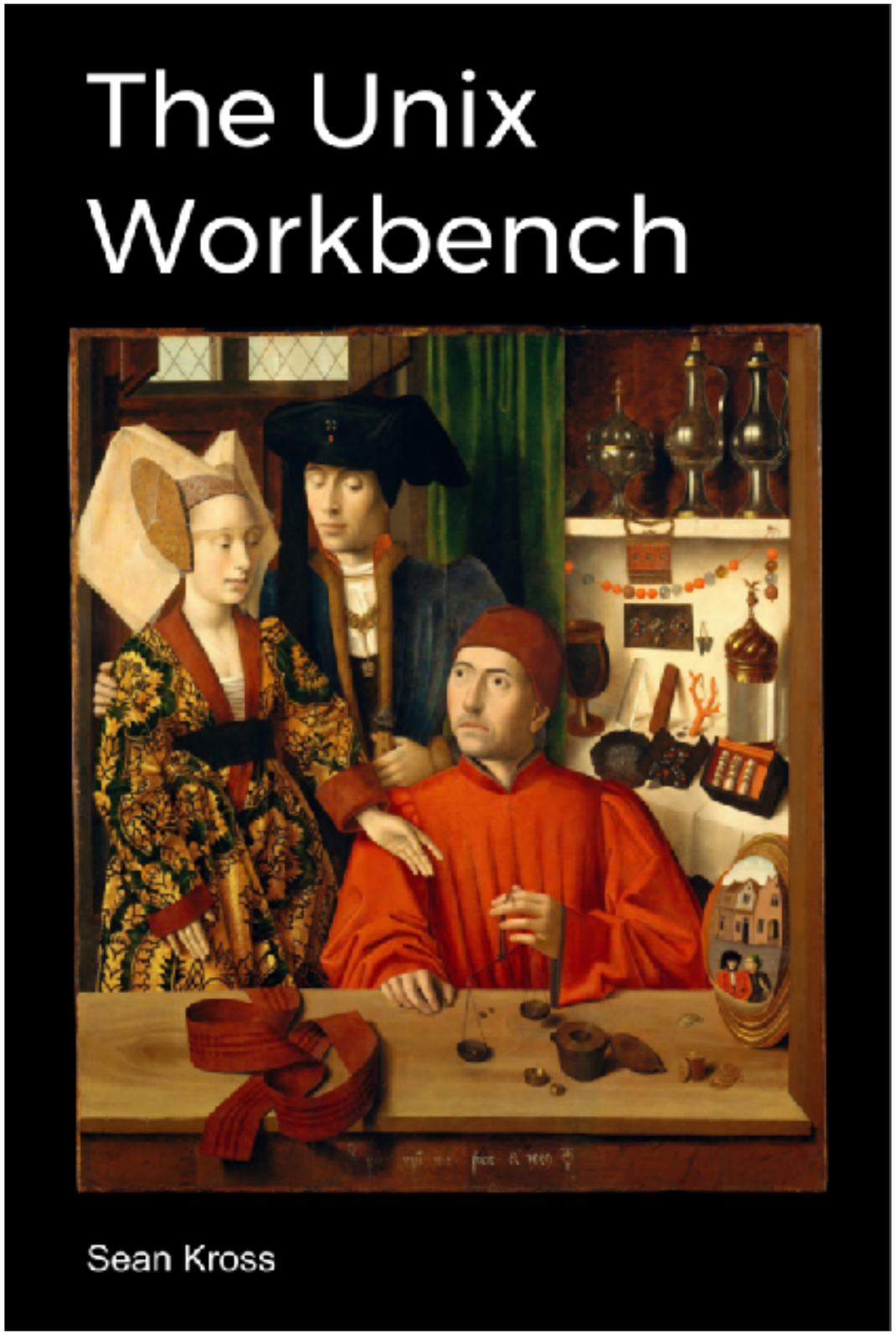
Lessons from Alumni:
Mastering Software Development in R

- The field of data science is still taking form.

- Experimentation with roles can give you a competitive advantage.

- Taking risks is easier of you're part of a community.

?



- **Learn how to use the command line from the ground up.**
- **No previous experience expected.**
- **The gateway into computationally intensive tasks.**
- **Includes an introduction to cloud computing.**
- **Free!**

leanpub.com/unix

Thank you!

Questions?

Link to these slides: seankross.com/crunch-talk/

Let’s talk: seankross@ucsd.edu

Find me on Twitter: @seankross