The democratization of data science education

Sean Kross
UC San Diego Design Lab
2017-10-11

A little about me

- New! Improved!
- Advised by Philip Guo
- Main interests:
  - Data Science
  - Online Education
  - Open Science

NYU

UNIVERSITY OF MARYLAND 18 56

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

PennDesign

The Johns Hopkins Data Science Lab

Jeff Leek

Roger Peng

Brian Caffo

‹ 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 3,490 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.

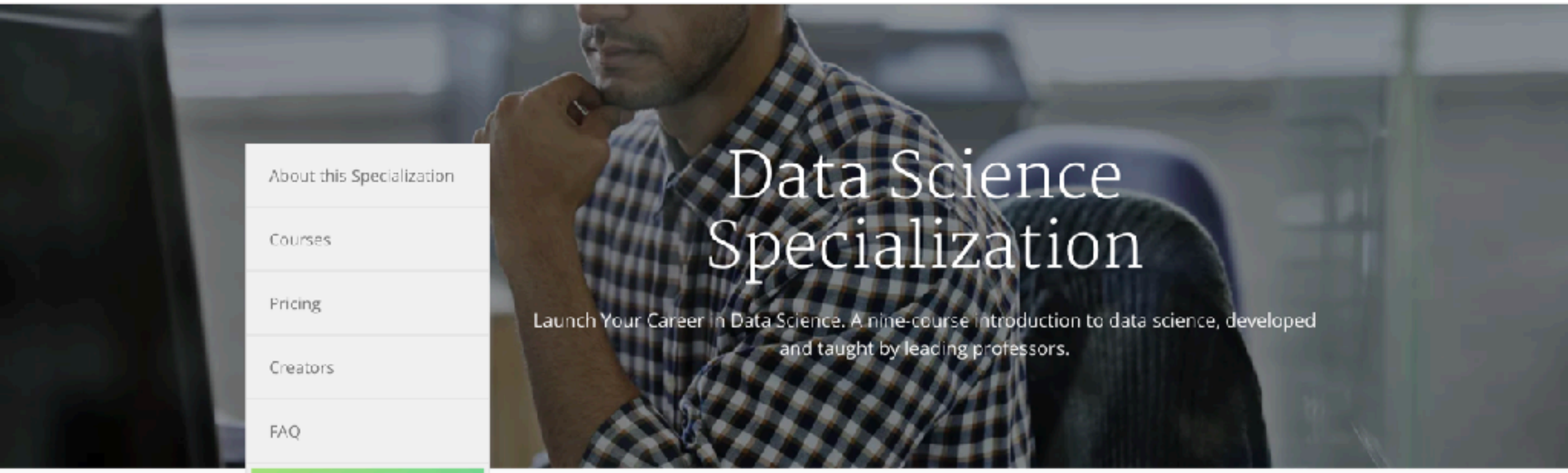
The goal of this collection is to

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

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.



Data Science Specialization

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

About This Specialization

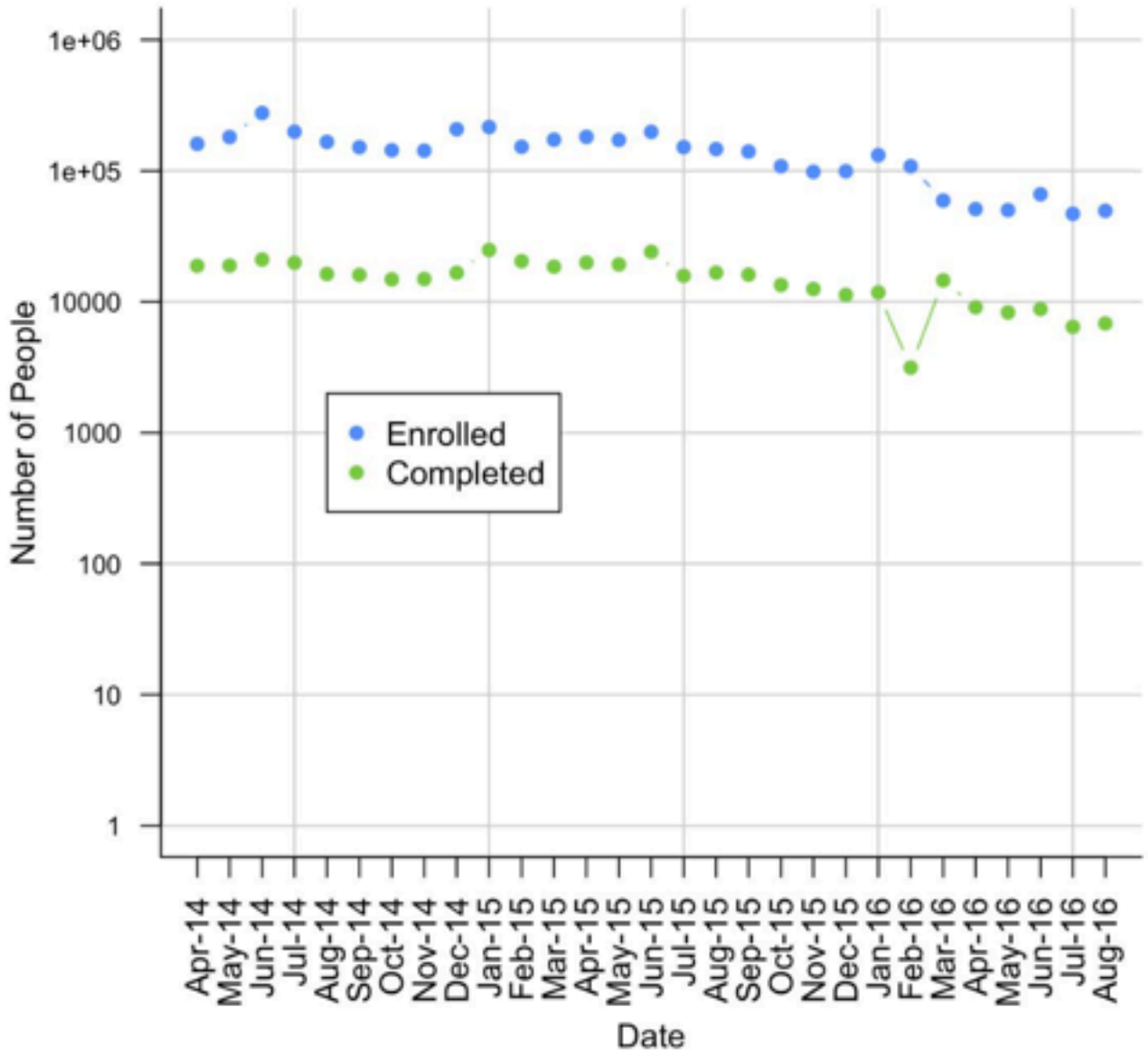
Ask the right questions, manipulate data sets, and create visualizations to communicate results.

This Specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. In the final Capstone Project, you'll apply the skills learned by building a data product using real-world data. At completion, students will have a portfolio demonstrating their mastery of the material.

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



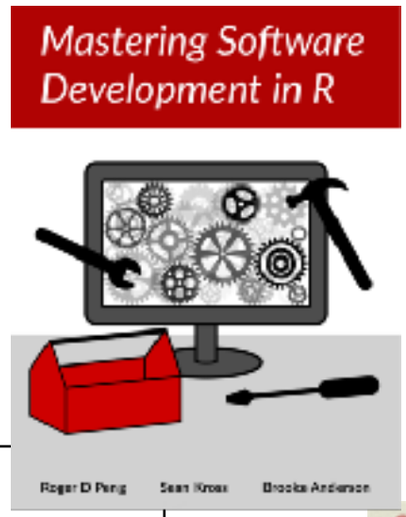
Key innovations

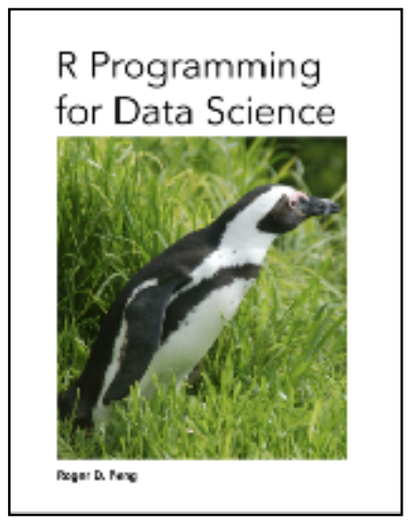
Give everything away for free

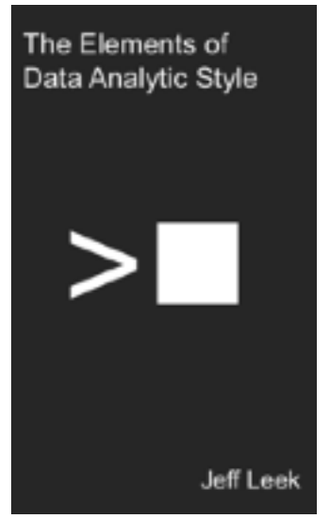
Leanpub

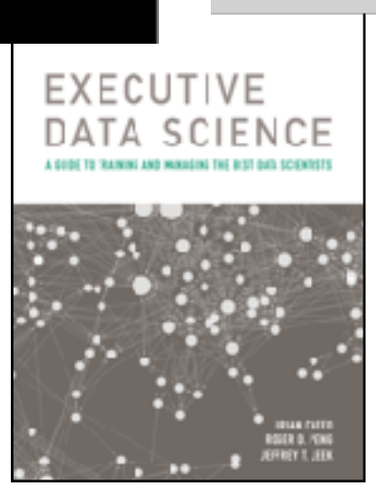
YouTube

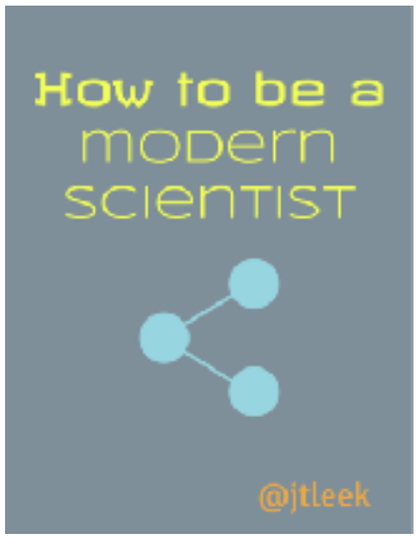
GitHub

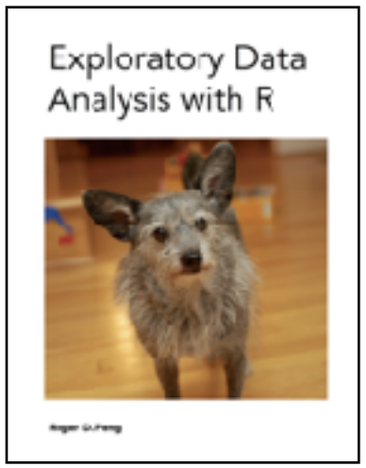












Capstones -> Portfolios -> Jobs

Run *every* course *every* month

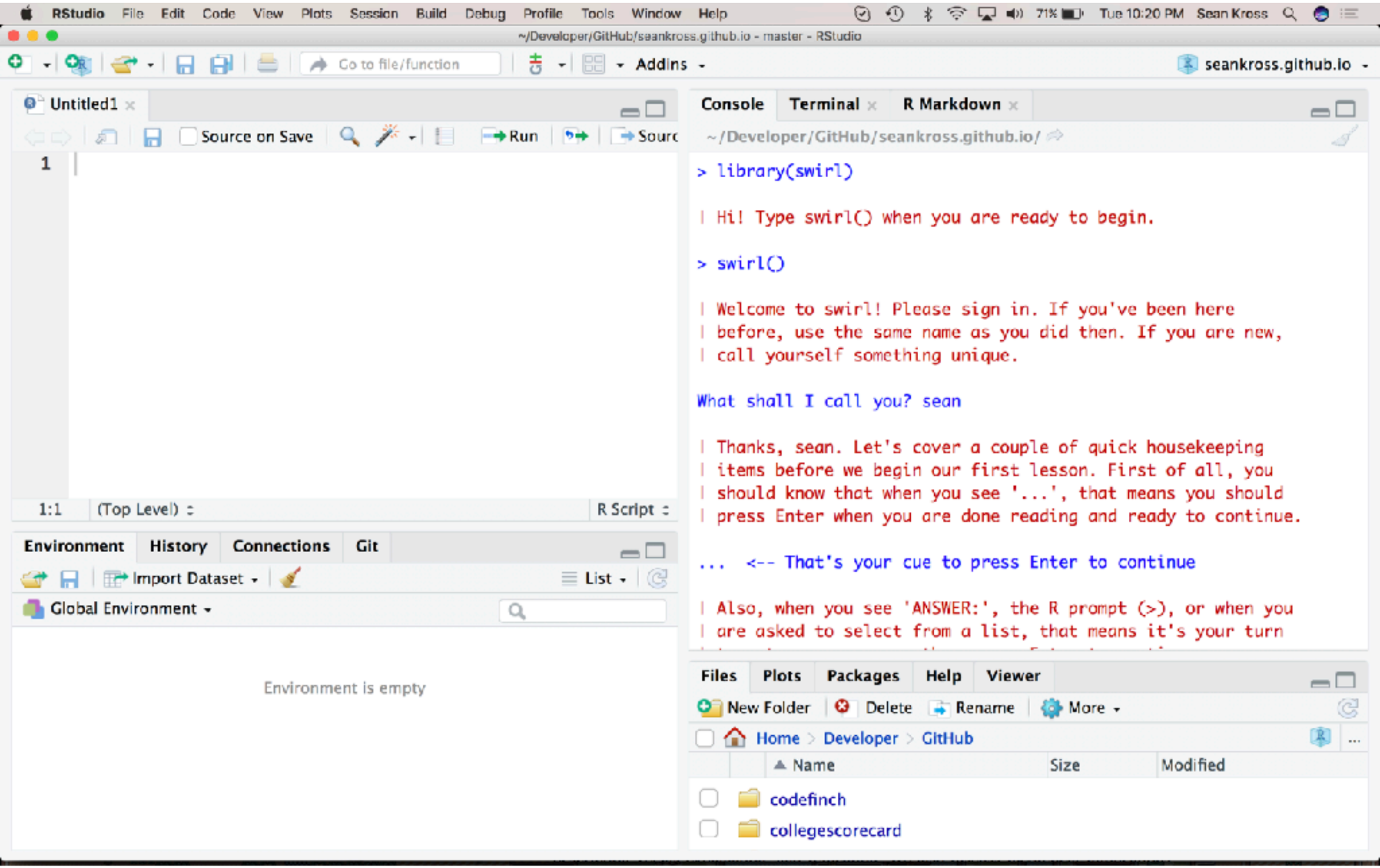
Integrate content

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		

Three areas I would like to concentrate on:

1. How can we develop new interactive learning systems for data science?

2. The great courses of the world are sitting in professor's file cabinets. How can we make online course content creation easier?

3. How do folks do data analysis? Why do they make certain choices during an analysis? How do upstream decisions made during an analysis affect downstream results?

Thank you!

Questions?

Link to these slides: seankross.com/dlab-talk-dss/

Let’s talk: seankross@ucsd.edu

Find me on Twitter: @seankross