

Introduction

JSC 370: Data Science II

January 8, 2024

Instructors

- Meredith Franklin: meredith.franklin@utoronto.ca
- Jun Ni (Jenny) Du: junni.du@mail.utoronto.ca

I will do the lectures on Mondays and Jenny will run the labs on Wednesdays. If you have questions you can email either of us.

My Background

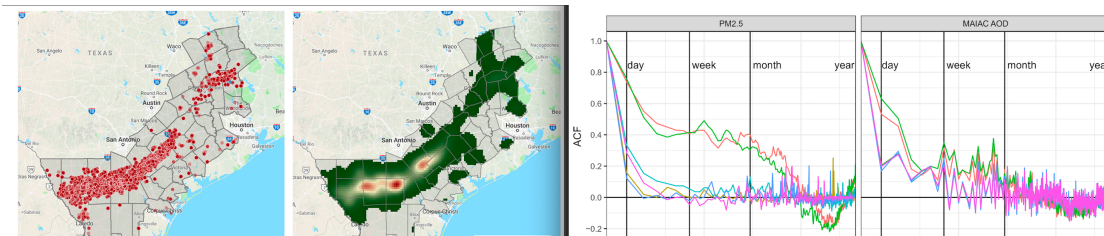
- In 2021 moved from Los Angeles where I was an Assistant/Associate Professor of Biostatistics at University of Southern California
- From Canada originally, McGill math for BSc, Ottawa/Carleton Institute of Math for MSc, Harvard for PhD, UChicago for postdoc
- At U of T I'm an Associate Professor with tenure in the Department of Statistical Science (51%) and the School of the Environment (49%)

My Teaching

- Founded a Master's of Health Data Science program at USC that launched in 2020
- Co-taught the introduction data science course
- Taught graduate-level spatial statistics, inference, linear models
- This semester I am also teaching STA255

My Research

- Spatial statistical methods for environmental data
- Data science techniques for remote sensing data/imagery
- Focus on pollution (air, noise) and climate (ghg, land cover change)
- Machine learning becoming a big part of environmental research





Course Goals

Through this course, you will hone the techniques used in Data Science. You will learn:

- Programming in R (Python for ML), and tools Markdown, Git
- Exploratory data analysis – generating hypotheses and building intuition
- Data visualization – showing data through interpretable summaries
- Data collection – data scraping, wrangling, cleaning
- Statistical (machine learning) algorithms

- Building a github.io website

Quercus + Git + Piazza

Course website - lecture slides, labs, data

<https://jsc370.github.io/JSC370-2024/>

Quercus - announcements, homework solutions, lab solutions, guest speaker reflections, grading

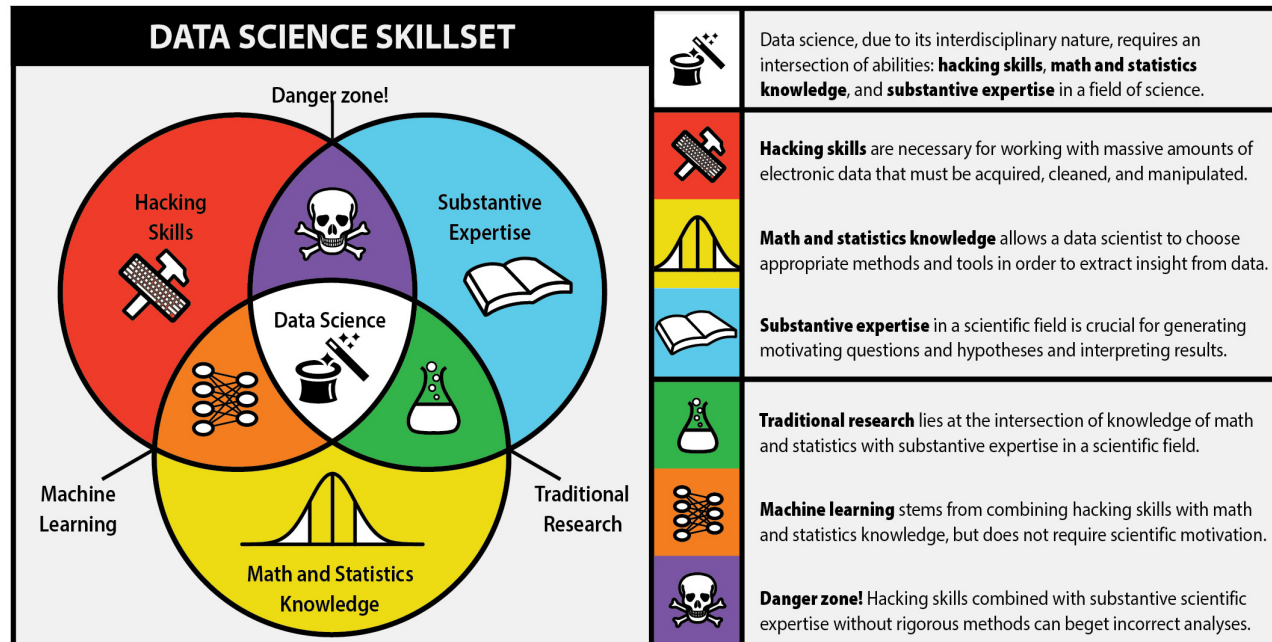
<https://q.utoronto.ca/courses/340934>

Piazza - questions and discussion

<https://piazza.com/class/lr58osfpxu12ms/>

What is data science?

- Data science is an exciting discipline that allows you to turn raw data



Data science can be really cool

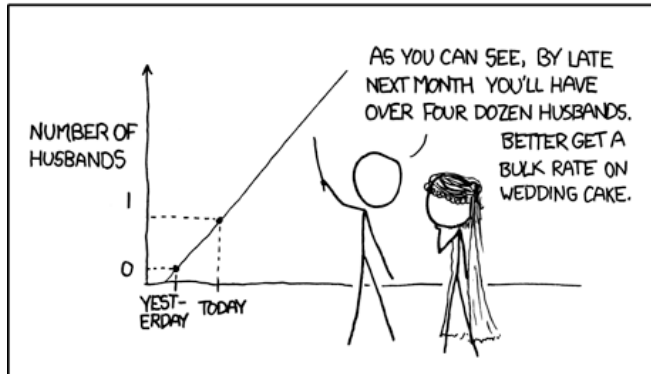
WHENEVER I LEARN A
NEW SKILL I CONCOCT

OH NO! THE KILLER
MUST HAVE FOLLOWED
HER ON VACATION!

BUT TO FIND THEM WE'D HAVE TO SEARCH
THROUGH 200 MB OF EMAILS LOOKING FOR
SOMETHING FORMATTED LIKE AN ADDRESS!

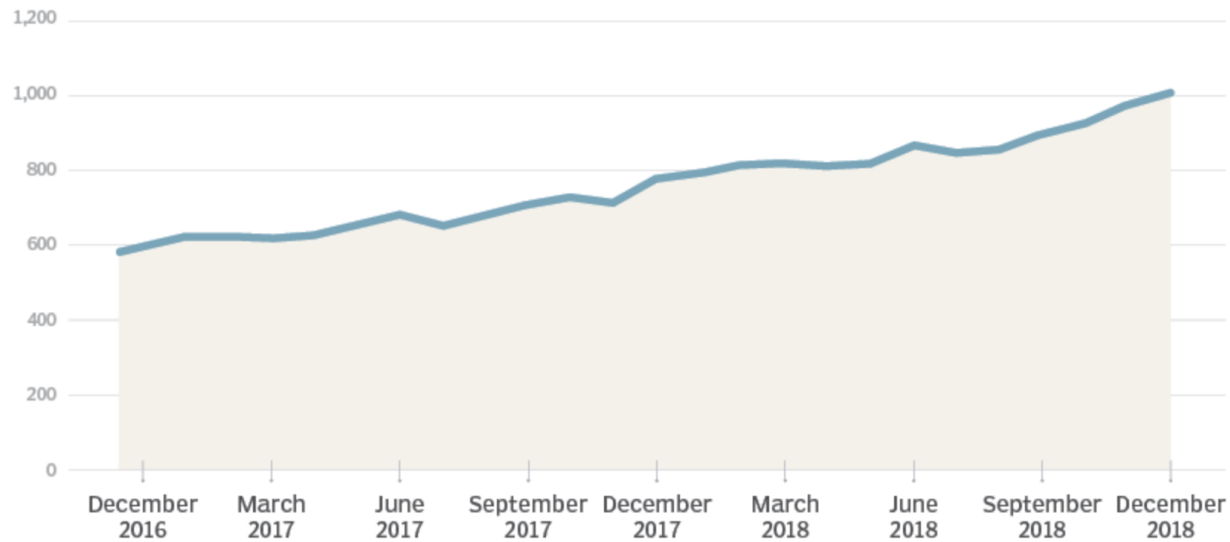
With great power comes great responsibility

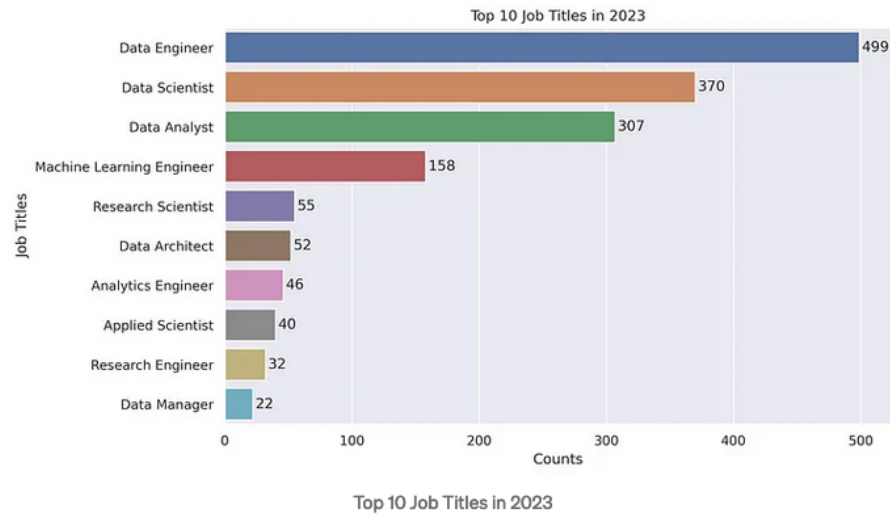
MY HOBBY: EXTRAPOLATING



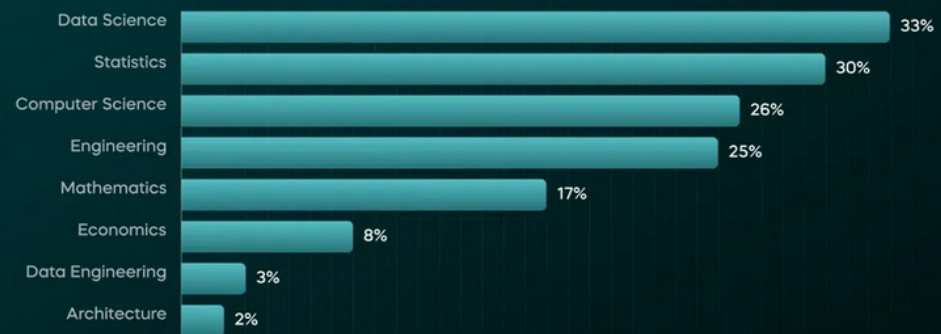
Data scientists are in high demand

Data scientist job postings, per 1 million postings on Indeed





Required Degrees



Data Scientists in Demand

Also see [here](#), and [here](#), and [here](#)

A good [data science subreddit](#) to follow - it provides insights on jobs, academic programs, and there are AMAs from industry leaders.

Another good resource is [Towards Data Science](#)

What is this course?

This course is a introduction to the world of data science following on from where JSC270 left off.

The course will teach language agnostic skills that are easily transferable, with examples done in R.

You can use any language/tool you prefer. But I can only guarantee help if you are using R and RStudio.

What is R?

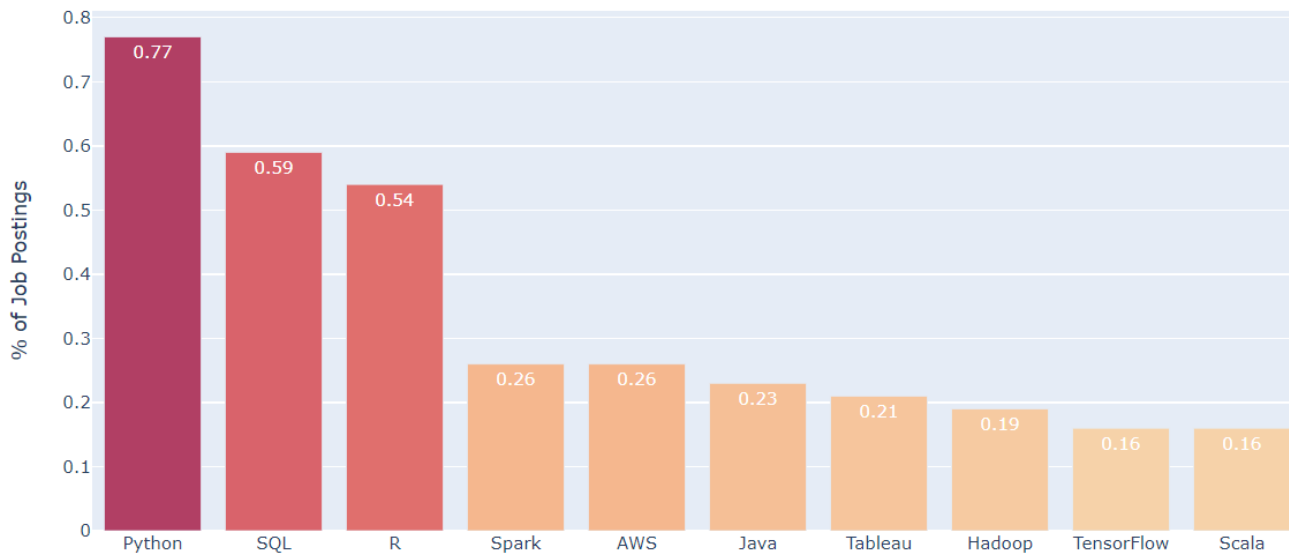


R is a language and environment for statistical computing and graphics. -- <https://r-project.org>

Created by statisticians for statisticians.

Over 20,000 packages added to CRAN.

10 Most In-Demand Data Science Skills in 2021



History of R

Originates from S, which was developed by Bell Labs in the 1970s

First versions of R were developed by Robert Gentleman and Ross Ihaka of U Aukland in mid-1990s

R is intended for statisticians but used by many (>2M users!)

R is open source, has nice graphics and visualizations

A lot of help is available online (Stack Overflow, R package vignettes, Journal of Statistical Software)

R Data Science Resources

1) R Programming for Data Science, 2022. Roger Peng. <https://bookdown.org/rdpeng/rprogdatascience/>

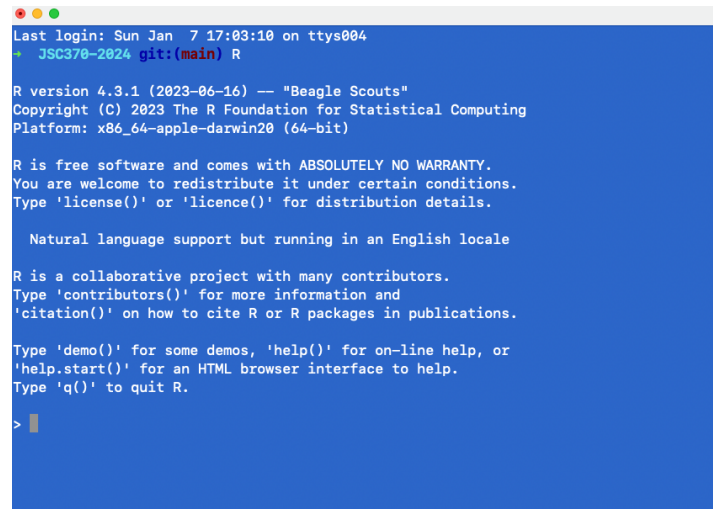
Supplementary References

2) R for Data Science, 2023 Garrett Golemund and Hadley Wickham. <http://r4ds.hadley.nz/>

3) Exploratory Data Analysis with R, 2020 Roger Peng <https://bookdown.org/rdpeng/exdata/>

4) Mastering Software Development in R, 2020 Roger Peng, Sean Kross,

R in the terminal



```
Last login: Sun Jan  7 17:03:10 on ttys004
+ JSC370-2024 git:(main) R

R version 4.3.1 (2023-06-16) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

What is RStudio?



RStudio is an integrated development environment (IDE) for R.

<https://posit.co/download/rstudio-desktop/>



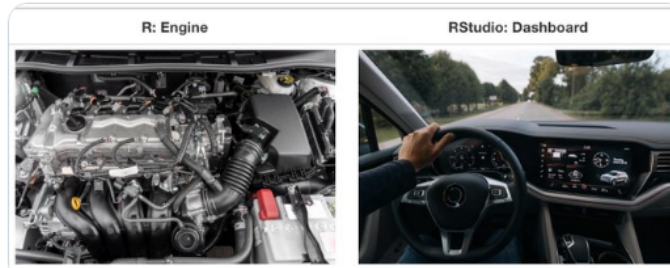
ModernDive
@ModernDive



Now that the school year has begun, let's start at the beginning! Two common questions

First, what's the difference between R & RStudio? Let's use a 🚗 analogy!

- R is like a car's engine. It does the work!
- RStudio is like a car's dashboard. You interact with R using RStudio!

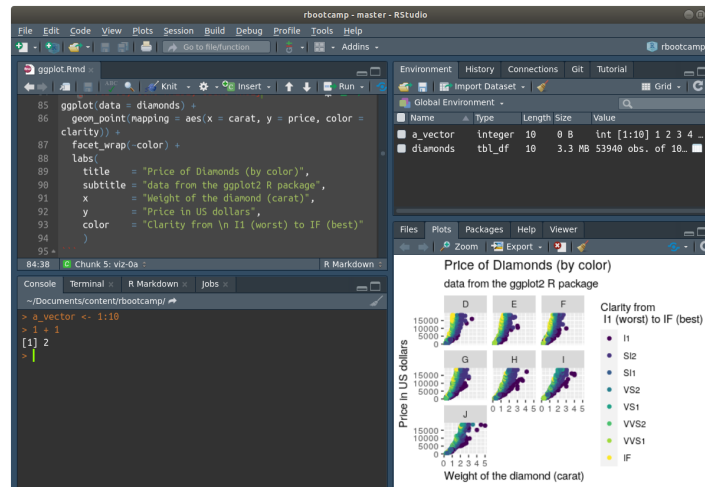


9:11 AM · Sep 10, 2019



👍 35 👤 See ModernDive's other Tweets

R + RStudio



GitHub

- Version control is necessary in the trade of data science and is used in industry and academia
- Building up a solid GitHub profile will put you in a good position for job hunting
- You will build a github.io website as part of this course



First Week

The lab exercises can be found on the course website in the schedule

<https://jsc370.github.io/jsc370-2024/>

Download the Rmd files

Submit individually completed lab at the end of day Wednesday

Next Week

Lecture 1-3 pm Monday Jan 15 (Version control)

Lab 1-3 pm Wednesday Jan 17 (Version control)

