Introduction to Data Science

Stat 220 Bastola

2022-01-05

Something about me

- First year at Carleton
- Originally from Nepal
- PhD in Applied Statistics from UC-Riverside
- Diverse education background
- Avid learner and traveler

Figure 1: Me without mask

COVID-19 related policies

- Stay home when sick. (Even if you don't have COVID-19, you should stay home if you aren't feeling well.)
- Follow CDC on testing, quarantine, and isolation.
- Follow the College mask-wearing policy

What is data science?

Data Science:

- the science of extracting meaningful information from data

> *Computer science is more than just programming; it is the creation of appropriate abstractions to express computational structures and the development of algorithms that operate on those abstractions. Similarly, statistics is more than just collections of estimators and tests; it is the interplay of general notions of sampling, models, distributions, and decision-making. [Data science] is based on the idea that these styles of thinking support each other. - Michael Jordan, UC Berkeley*

Data Scientist

The “data scientist” mashup:

- “The definitions of data science are converging around the intersection of mathematics, statistics, and computer science—with some area of application (e.g., finance, biology, political science).”
- “I have heard data scientists referred to equally as
  - ‘the computer scientist who was the best of his peers in his statistics courses’ and
  - ‘the statistician who was the best of his peers in his computer science courses.”’
- Jennifer Lewis Priestley Data Science: The Evolution or the Extinction of Statistics?

Data Science Education

- Many schools now offer degrees in some form of data science (data analytics)
- A B.S. (or Masters) in Data Science includes courses like:
    - Intro Stats, Intro Programming, Intro Data Science
    - Regression (modeling)
    - Machine Learning, Data mining
    - Database management
    - Data visualization
    - Big Data
    - Applications (econ, poli sci, bio)
    - Ethics

Stat 220: Data Science

Focus on the “soup to nuts” approach to problem solving

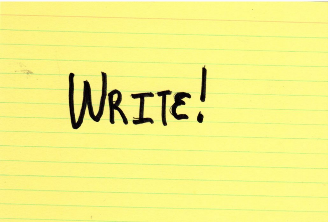
- data wrangling
    - reshaping, cleaning, gathering
- learning from data
    - EDA tools
    - statistical learning methods
    - network data, spatial data
- communication
    - reproducibility
    - effective visualization

How to make friends and succeed in Data Science?

1. *Actively follow along!* RMarkdown documents will be provided for you each week - use these to take notes and run code “live” in class.
2. *Ask questions!* This is new for everyone, no question is a bad question.
3. *When you don’t know if something will work, try it!* Experimentation is key in this class.
4. *BRING YOUR LAPTOP.* Charged, everyday.

Tell me something about yourself

- Your name?
- Gender Pronouns?
- Why are you interested in data science?
- Rate your R Skills from 1 to 10.



Class Instruction

```
# Main Course Webpage for Instruction
```

https://deepbas.io/courses/stat220/

- Bookmark this page
- Should be checked multiple times a day

Collaborative notes

- Each day, two of you will collaborate on notes to share with the class
- Creates a crowd-sourced version of what we do in class
- Helps anyone who needs to miss class
- You'll do this 3x throughout the course
- Sign up here
- Notes are due 24 hours after class, count as a HW assignment

Necessary skills to be mastered

- programming with data
- statistical modeling
- domain knowledge
- communication

Why aren't we learning Python?

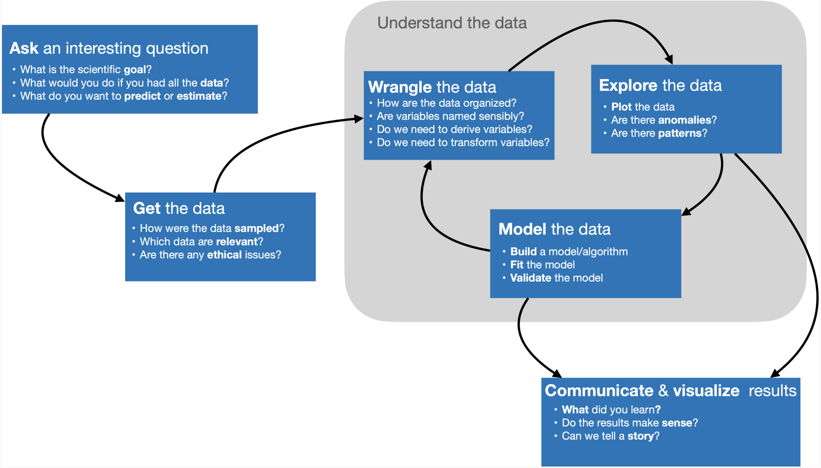
When Hadley Wickham was asked "Why R?"

> *And the second reason, which is both a huge strength of R and a bit of a weakness, is that R is **not just a programming language**. It was designed from day 1 to be an **environment that can do data analysis**. So, compared to the other options like Python, you can get up and running in R doing data science, learning much, much less about programming to get started. And that generally makes it like **easier to get up and running if you don't have formal training in computer science or software engineering**.*
> *-Hadley Wickham, Advice to Young (and Old) Programmers: A Conversation with Hadley Wickham*

Using R Markdown for data science

- You will use R Markdown for all work in this class
- A Markdown (`.Rmd`) file contains
  - R code
  - written answers, description of results, report, etc.
- The Markdown file is `knit` to generate an output document
  - pdf, html, word
  - presentations (html, beamer pdf)
  - dashboards, interactive graphics (html)
- Markdown is designed for **reproducibility**!
- The slides I produce for this class are R Markdown's `beamer`

Data Science in a nutshell



Ask an interesting question

- What is the scientific **goal**?
- What would you do if you had all the **data**?
- What do you want to **predict** or **estimate**?

Get the data

- How were the data **sampled**?
- Which data are **relevant**?
- Are there any **ethical** issues?

Understand the data

Wrangle the data

- How are the data organized?
- Are variables named sensibly?
- Do we need to derive variables?
- Do we need to transform variables?

Explore the data

- **Plot** the data
- Are there **anomalies**?
- Are there **patterns**?

Model the data

- **Build** a model/algorithm
- **Fit** the model
- **Validate** the model

Communicate & **visualize** results

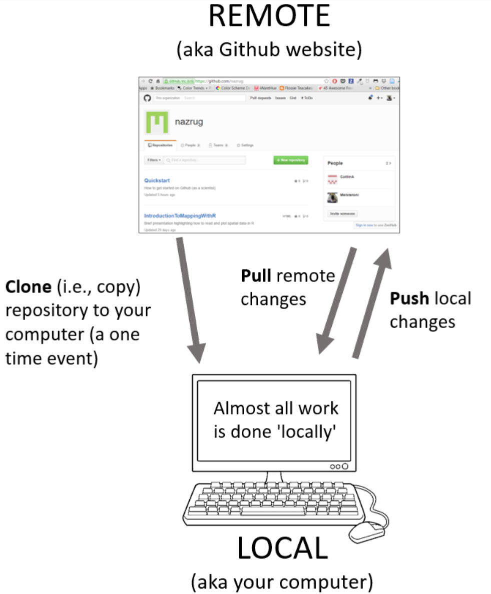
- **What** did you learn**?**
- Do the results make **sense**?
- Can we tell a **story**?

Image adapted from work of Joe Blitzstein, Hanspeter Pfister, and Hadley Wickham

Version Control using Git and Github

- Git doesn't automatically track changes the way a tool like Google Docs does
- Starting with github and going to Rstudio is easier for beginner
- Do commit, push, pull until you get used to it!
- Commit: Telling Git that you made changes
  - can be done from within RStudio

The git cycle



REMOTE

(aka Github website)

Clone (i.e., copy) repository to your computer (a one time event)

Pull remote changes

Push local changes

Almost all work is done 'locally'

LOCAL

(aka your computer)

Source: http://ohi-science.org/data-science-training/

Using GitHub and Rstudio for data science

- Rstudio lets you create git controlled projects
  - create a GitHub repo
  - make a Rstudio project using your cloned repo
  - edit/create files (.rmd, .r, .csv, . . .)
  - **commit** changes to your local computer using git
  - **push** changes to the GitHub repo (online)
  - **pull** changes made by others to your computer

Setup Guides

- What you need to do
    - read the Rstudio for Stat220 page
    - read the GitHub for Stat220 page
    - read the Software for Stat220 page

What will a typical day/week look like?

Before class:

- Some reading/video to introduce some topics
- Work on homework/projects, come with questions

During class:

- Mini lectures
- Hands-on programming

For rest of the class

- In Maize or on your laptop: make sure you have a `test-assignment` R project and a `course-content` R project.
- Work on the `test-assignment.Rmd` file in the `test-assignment` repo
    - Ask me questions
    - By class time Wednesday, push your completed `test-assignment.Rmd` and `test-assignment.md` files to GitHub.
    - Worth 10 points toward homework score. (5/10 for successful push to GitHub!)