Introduction to the STAT207 Course

Case Study:

What datasets do you find interesting?

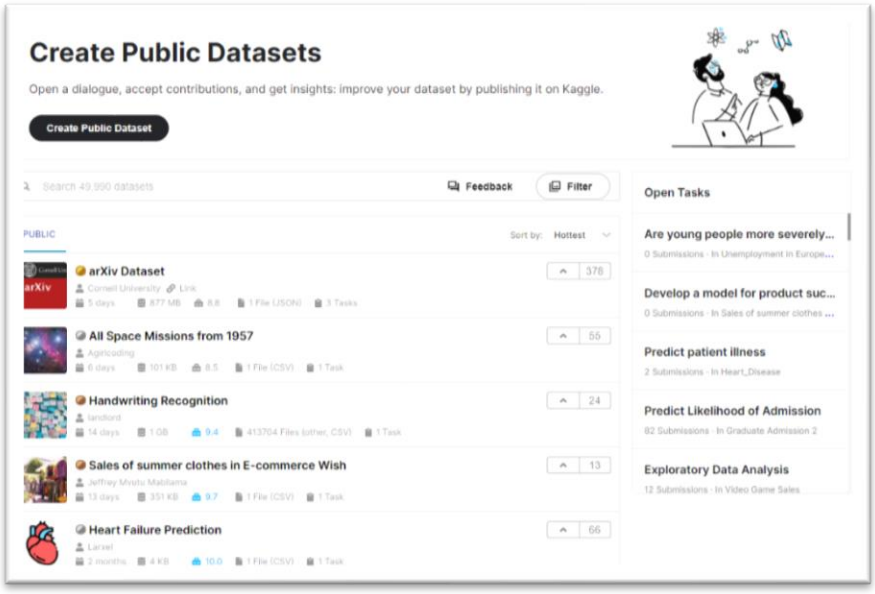
Purpose of this Lecture:

In this lecture we will cover the following topics.

- About you
- About me
- What is data science?
- Data science vs. statistics
- Course Goals
- Why use Python for data science?
- Why study data science?
- Skills needed by a data scientist
- Course website and syllabus
- Course Github enterprise organization
- Lecture format
- Lab format

ABOUT YOU!

What types of data sets would you like to **gain insights from, make predictions with,** and/or **use to help make better decisions?**

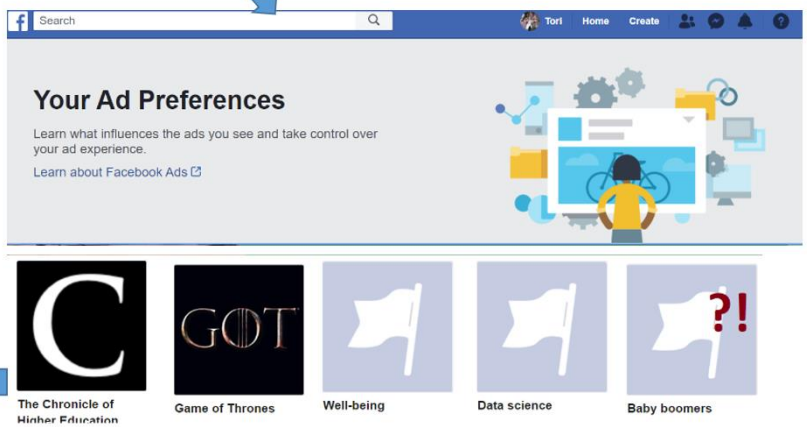


https://www.kaggle.com/datasets

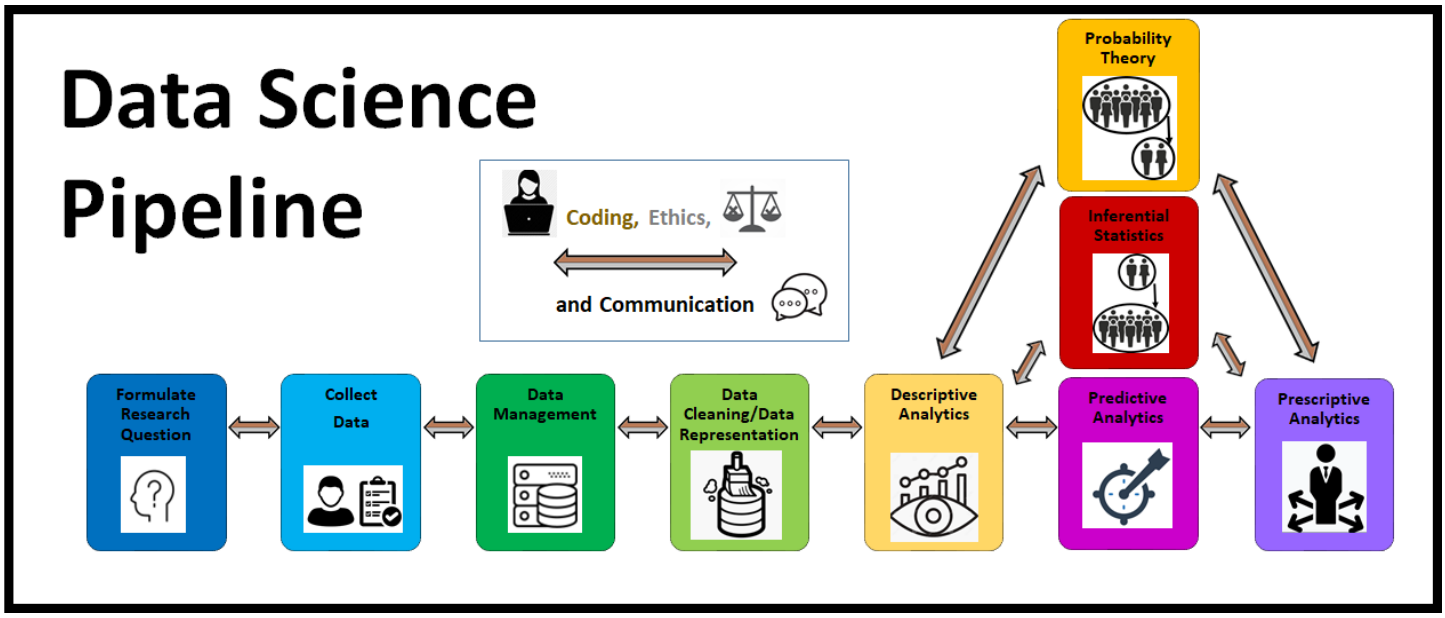
What places have you been able to find fun and interesting datasets from in the past?

About me

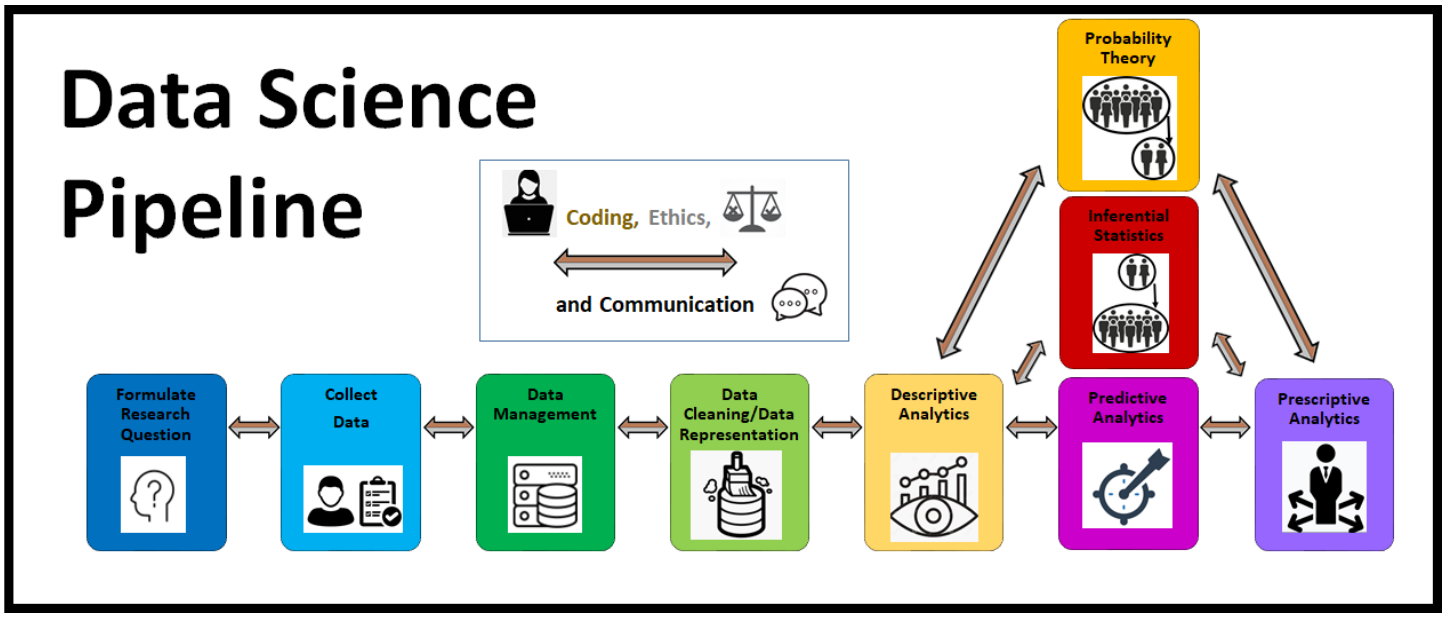
- Online Advertising
- TV Advertising
- Narcotics Detection
- Gene Expression Analysis
- Get Out the Vote Initiatives



WHAT IS DATA SCIENCE?

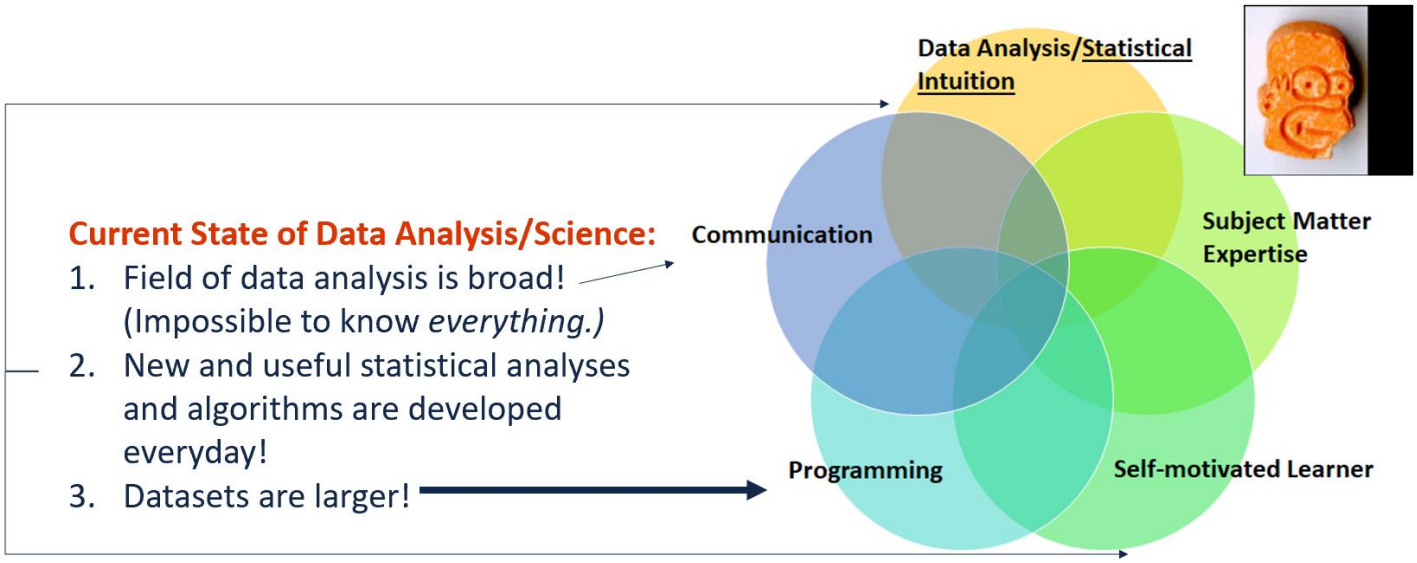


STATISTICS VS. DATA SCIENCE



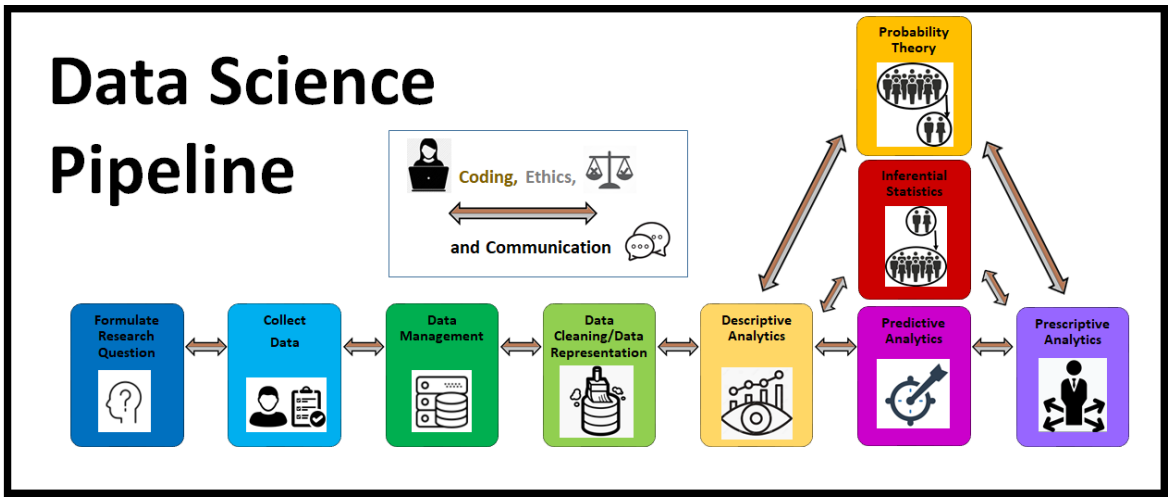
Current State of Data Analysis/Science:

1. Field of data analysis is broad! (Impossible to know *everything.)*
2. New and useful statistical analyses and algorithms are developed everyday!
3. Datasets are larger!



https://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

COURSE GOALS



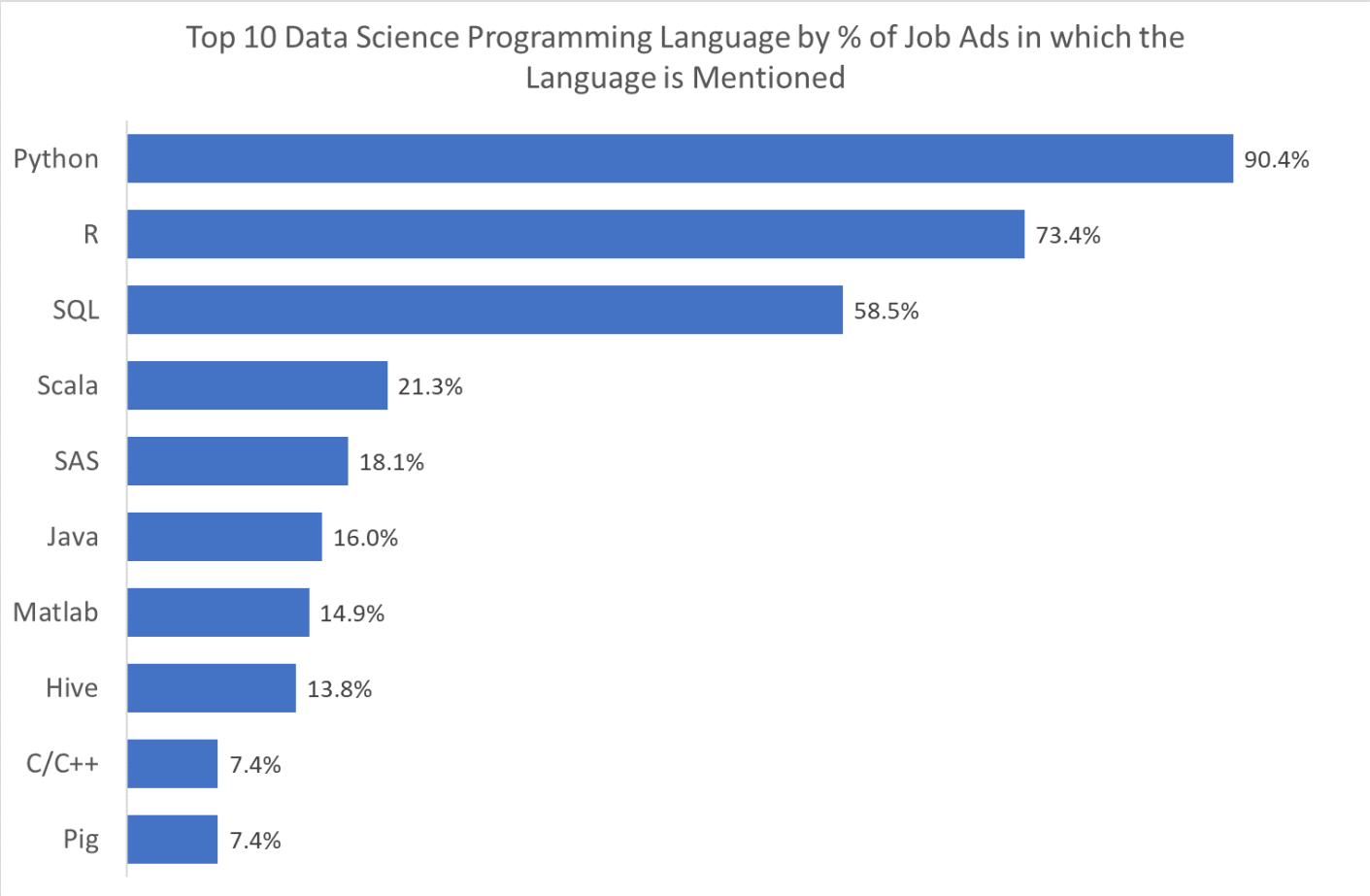
1. **Survey** of the data science pipeline

2. Using **Python,** complete a **beginning-to-end data science project.**

3. When conducting a more advanced data science project, **develop an intuition** for:
    a. what **questions to ask**
    b. how to *efficiently* learn **new algorithms**, **models**, **functions** etc
    c. What **search terms** to look up
    d. **what to research**

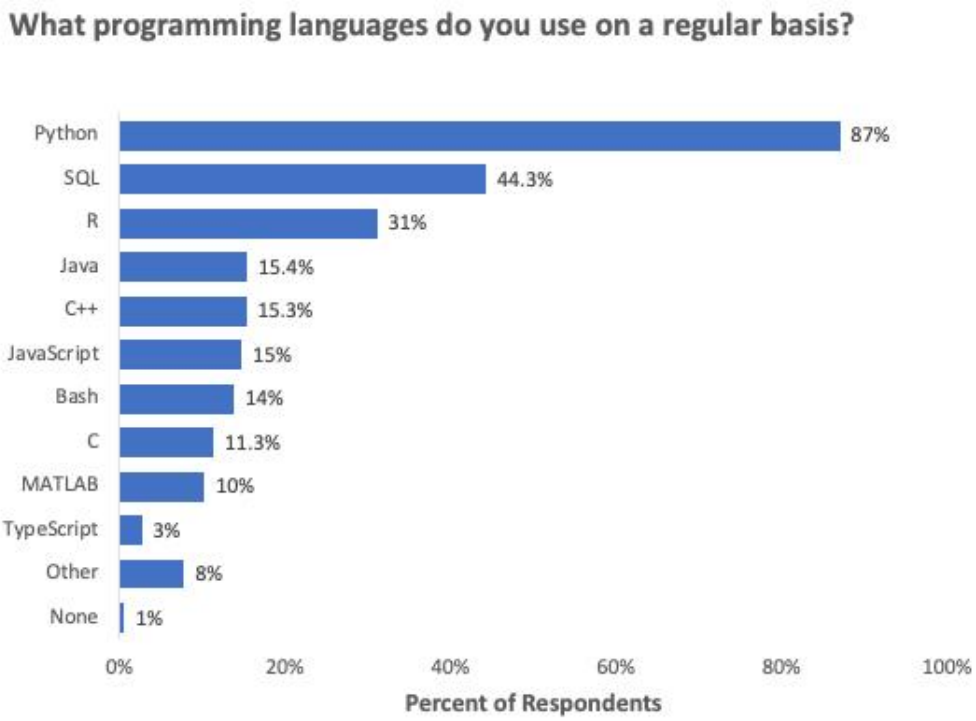
4. **Topics covered:**
    a. http://courses.las.illinois.edu/spring2022/stat207/course_topics.html

WHY USE PYTHON FOR DATA SCIENCE?



https://towardsdatascience.com/which-programming-language-should-data-scientists-learn-first-aac4d3fd3038

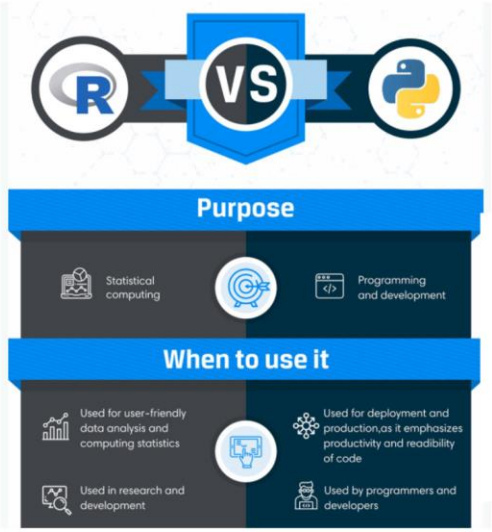
What are some ways we could **have collected this data?**

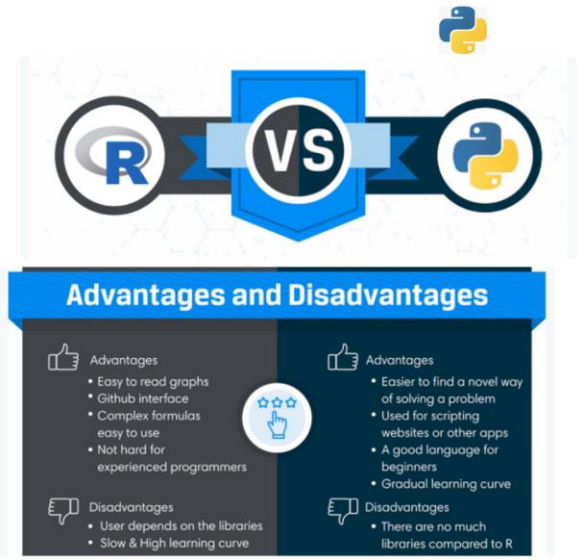


Note: Data are from the 2019 Kaggle ML and Data Science Survey. You can learn more about the study here: https://www.kaggle.com/c/kaggle-survey-2019.
A total of 19717 respondents completed the survey; the percentages in the graph are based on a total of 14762 respondents who provided an answer to this question.

BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING
Copyright 2020 Business Over Broadway

What if we wanted **make an inference** about whether Python is the most used programming language of **ALL DATA SCIENTISTS** using this **sample of data scientists**? What might we be interested to know about how the data was collected?



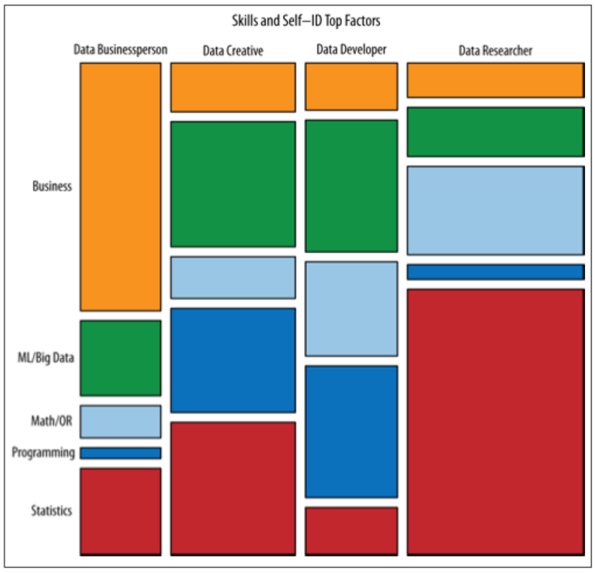


- www.stackoverflow.com has great answers to many of the questions you could ask for Python!
- Working in a big team to automate something? Python is great!

https://www.superdatascience.com/blogs/learn-all-the-pros-and-cons-of-python-vs-r-programming

WHY STUDY DATA SCIENCE?

https://www.superdatascience.com/blogs/learn-all-the-pros-and-cons-of-python-vs-r-programming



http://radar.oreilly.com/2013/06/theres-more-than-one-kind-of-data-scientist.html

•**Data Businesspeople** are the product and profit-focused data scientists. They're leaders, managers, and entrepreneurs, but with a technical bent. A common educational path is an engineering degree paired with an MBA.

•**Data Creatives** are eclectic jacks-of-all-trades, able to work with a broad range of data and tools. They may think of themselves as artists or hackers, and excel at visualization and open source technologies.

•**Data Developers** are focused on writing software to do analytic, statistical, and machine learning tasks, often in production environments. They often have computer science degrees, and often work with so-called "big data".

•**Data Researchers** apply their scientific training, and the tools and techniques they learned in academia, to organizational data. They may have PhDs, and their creative applications of mathematical tools yields valuable insights and products.

Course Website and Syllabus

Canvas Page: https://compass2g.illinois.edu/

- Your grades
- Lecture markups
- Lecture videos
- Discussion
- Zoom Meeting Links

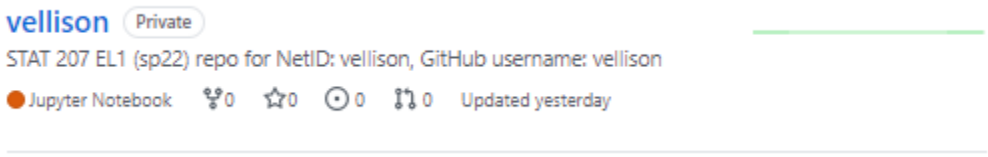
Course Website: **http://courses.las.illinois.edu/spring2022/stat207/**

- Course schedule and *incomplete* lecture notes (to be filled out in the lecture).
- Syllabus
- Assignment and Project Information
- Tech Guides
- Course Content List
- Course Staff Info

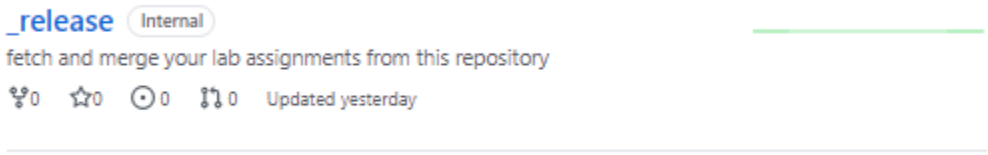
Course Github Enterprise Organization

https://github-dev.cs.illinois.edu/stat207-sp22-el1

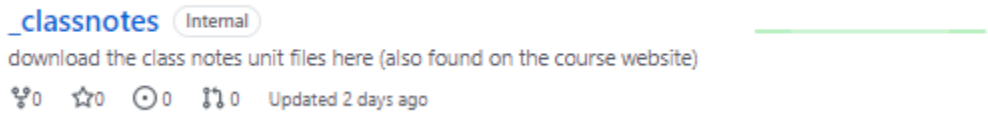
1. **Your netid repository**
    - **push** your completed lab assignments here for grading.



2. **_release repository**
    - **fetch** and **merge** (ie. download) your weekly lab assignments from here.



3. **_classnotes repository**
    - **pull** (ie. download) the lecture note materials here.



Lecture Format

During Lecture

- **Lectures are Synchronous and In-Person:** attendance strongly encouraged if you are able to, but not required!

- **"Skeleton" Lecture Unit Materials Posted Before Class**
  - http://courses.las.illinois.edu/spring2022/stat207
  - https://github-dev.cs.illinois.edu/stat207-sp22-el1/_classnotes

- **Lecture Unit Folder Includes:**
  - Slides pdf *(conceptual)*
  - Jupyter Notebook *(applications)*
  - Jupyter Notebook pdf copy
  - csv files (sometimes)

After Lecture

- Lecture Markups Posted on Canvas

- Lecture Video Posted on Canvas

Lab Format

During Lab

Labs are Synchronous and In-Person:

- 5 points for attendance at each lab
- 50 total points for lab attendance
- 4 lab misses penalty free

Lab Purpose

Work on lab assignments and ask the TA and CAs questions

- Individual lab assignment [25 points]
- Group lab assignment [5 points]
  - Groups of 2-3
  - Contribution report
  - Only one team member needs to submit

After Lab

Submit your lab assignment materials to Github by the following **Wednesday night 11:59pm CST** at the latest.