Introduction to the STAT207 Course

***Case Study*:**

What datasets do you find interesting?

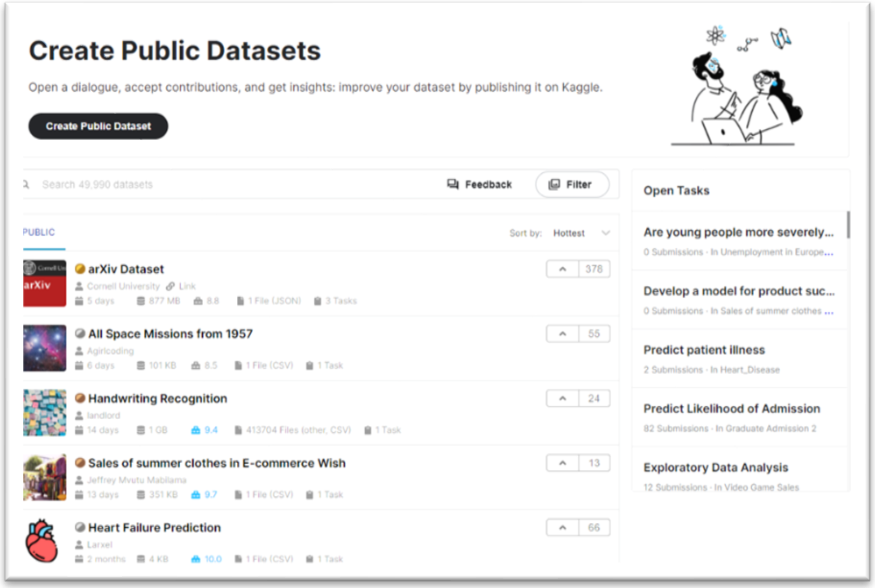
Purpose of this Lecture:

In this lecture we will cover the following topics.

1. About you
2. About me
3. What is data science?
4. Data science vs. statistics
5. Course Goals
6. Why use Python for data science?
7. Why study data science?
8. Skills needed by a data scientist
9. Course website and syllabus
10. Course Github enterprise organization
11. Lecture format
12. Lab format

1. 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?**



Create Public Datasets

Open a dialogue, accept contributions, and get insights: improve your dataset by publishing it on Kaggle.

Dataset	Owner	Updated	Size	Usability	Files	Tasks	Votes
arXiv Dataset	Cornell University · Link	5 days	877 MB	8.8	1 File (JSON)	3 Tasks	378
All Space Missions from 1957	Agirlcoding	6 days	101 KB	8.5	1 File (CSV)	1 Task	55
Handwriting Recognition	landlord	14 days	1 GB	9.4	413704 Files (other, CSV)	1 Task	24
Sales of summer clothes in E-commerce Wish	Jeffrey Mvutu Mabilama	13 days	351 KB	9.7	1 File (CSV)	1 Task	13
Heart Failure Prediction	Larxel	2 months	4 KB	10.0	1 File (CSV)	1 Task	66

Open Tasks

Are young people more severely...
0 Submissions · In Unemployment in Europe...

Develop a model for product suc...
0 Submissions · In Sales of summer clothes ...

Predict patient illness
2 Submissions · In Heart_Disease

Predict Likelihood of Admission
82 Submissions · In Graduate Admission 2

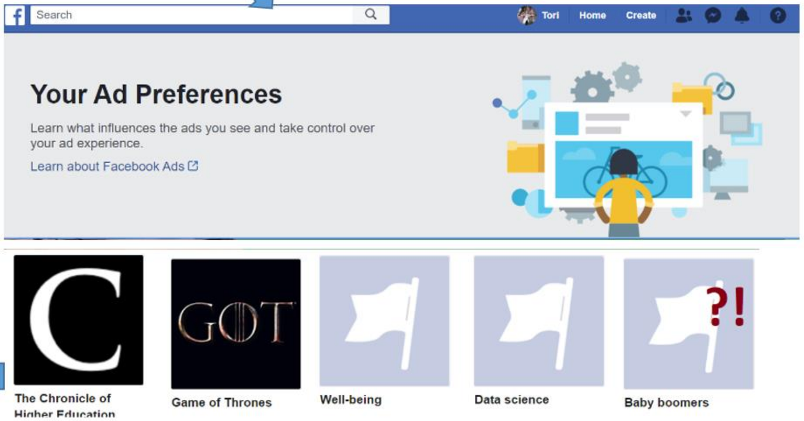
Exploratory Data Analysis
12 Submissions · In Video Game Sales

https://www.kaggle.com/datasets

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

2. About me

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



Your Ad Preferences

Learn what influences the ads you see and take control over your ad experience.

Learn about Facebook Ads

The Chronicle of Higher Education

Game of Thrones

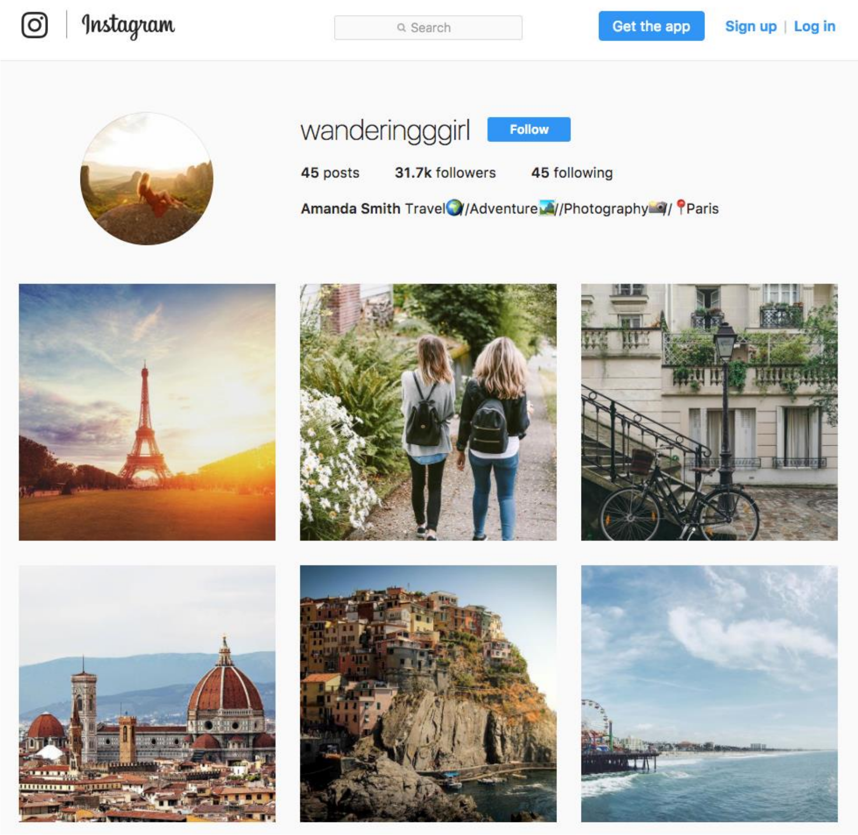
Well-being

Data science

Baby boomers

3. What is data science?

Do you think this is a real or fake Instagram account?



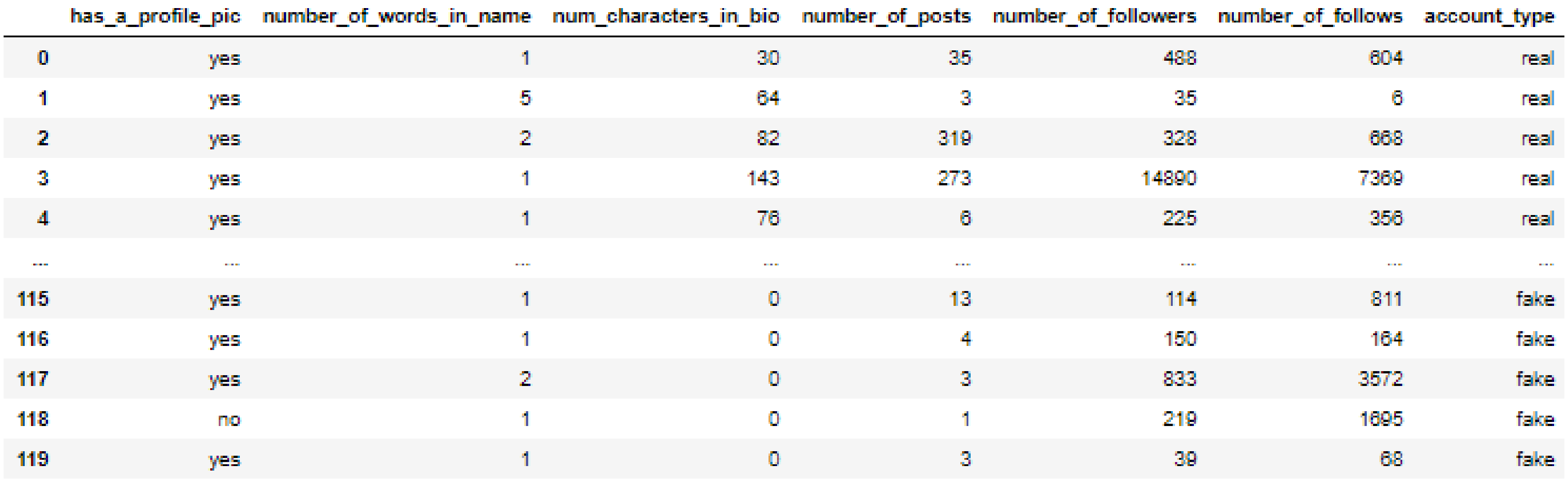
wanderingggirl

45 posts 31.7k followers 45 following

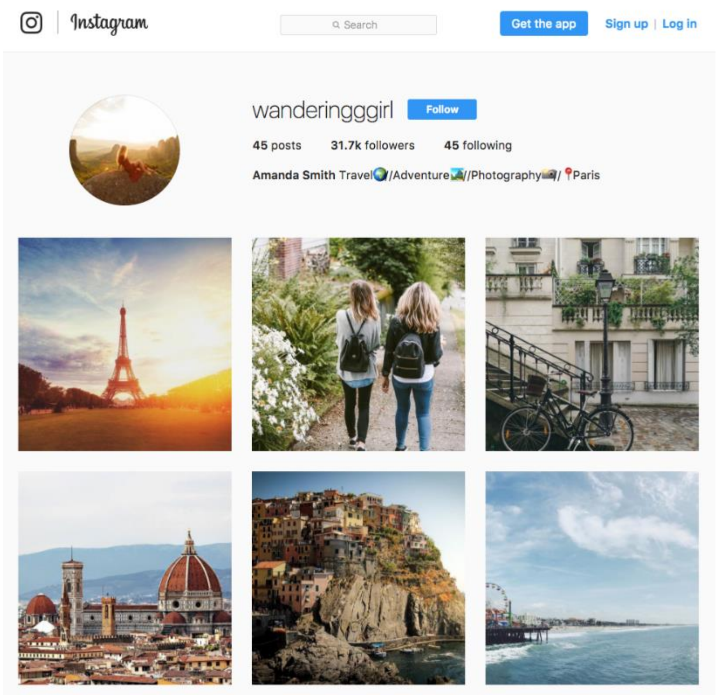
Amanda Smith Travel🌍/Adventure🏞️/Photography📸/ 📍Paris

Dataset

Suppose that we have at our disposal the dataset below that is comprised of 60 fake Instagram accounts and 60 real Instagram accounts.



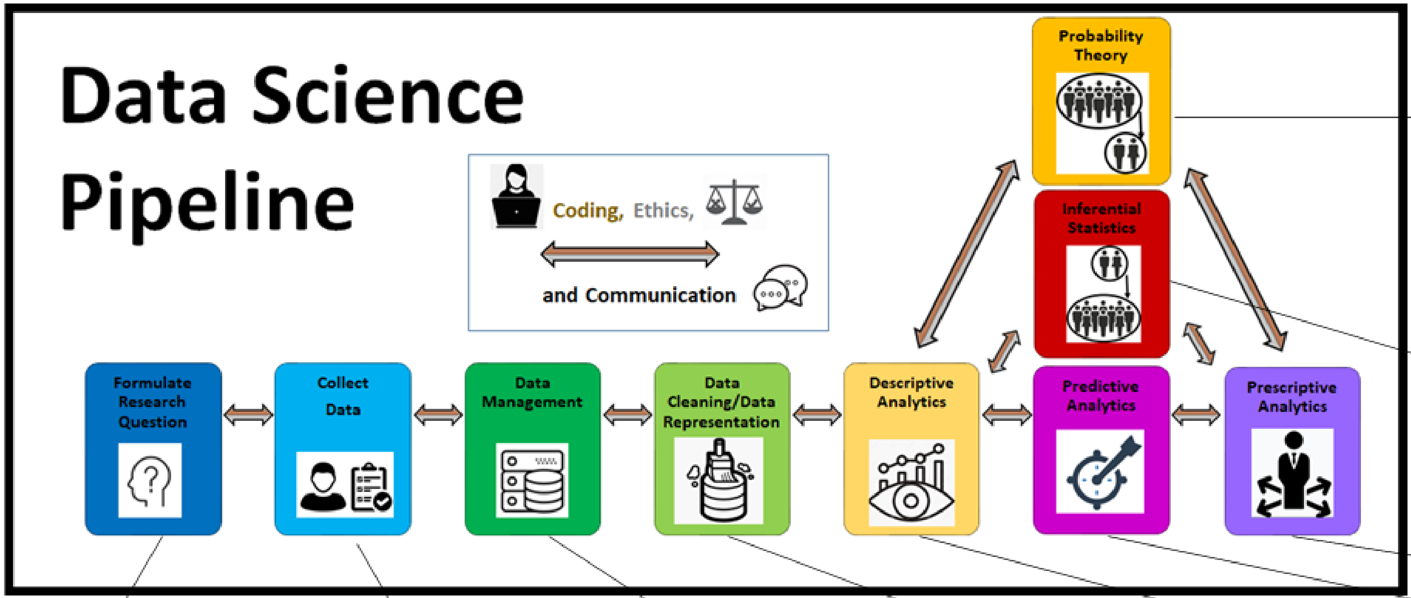
	has_a_profile_pic	number_of_words_in_name	num_characters_in_bio	number_of_posts	number_of_followers	number_of_follows	account_type
0	yes	1	30	35	488	604	real
1	yes	5	64	3	35	6	real
2	yes	2	82	319	328	668	real
3	yes	1	143	273	14890	7369	real
4	yes	1	76	6	225	356	real
...	...	...	...	...	...	...	...
115	yes	1	0	13	114	811	fake
116	yes	1	0	4	150	164	fake
117	yes	2	0	3	833	3572	fake
118	no	1	0	1	219	1695	fake
119	yes	1	0	3	39	68	fake



wanderingggirl

45 posts 31.7k followers 45 following

Amanda Smith Travel🌍/Adventure🏞/Photography📷/ 📍Paris



Data Science Pipeline

Coding, Ethics, and Communication

Formulate Research Question

Collect Data

Data Management

Data Cleaning/Data Representation

Descriptive Analytics

Probability Theory

Inferential Statistics

Predictive Analytics

Prescriptive Analytics

Is it uncommon for a fake account to have more than 1000 followers? How might you use this dataset to figure this out?

What larger population of Instagram users could we use this dataset to learn more about?

What's a research question you might want to answer with this dataset?

What might you be interested to know about how this dataset was collected? Why?

What are some common types of files that datasets can be stored in? What kind of issues might we encounter if our dataset was comprised of 1,000,000,000,000 Instagram accounts?

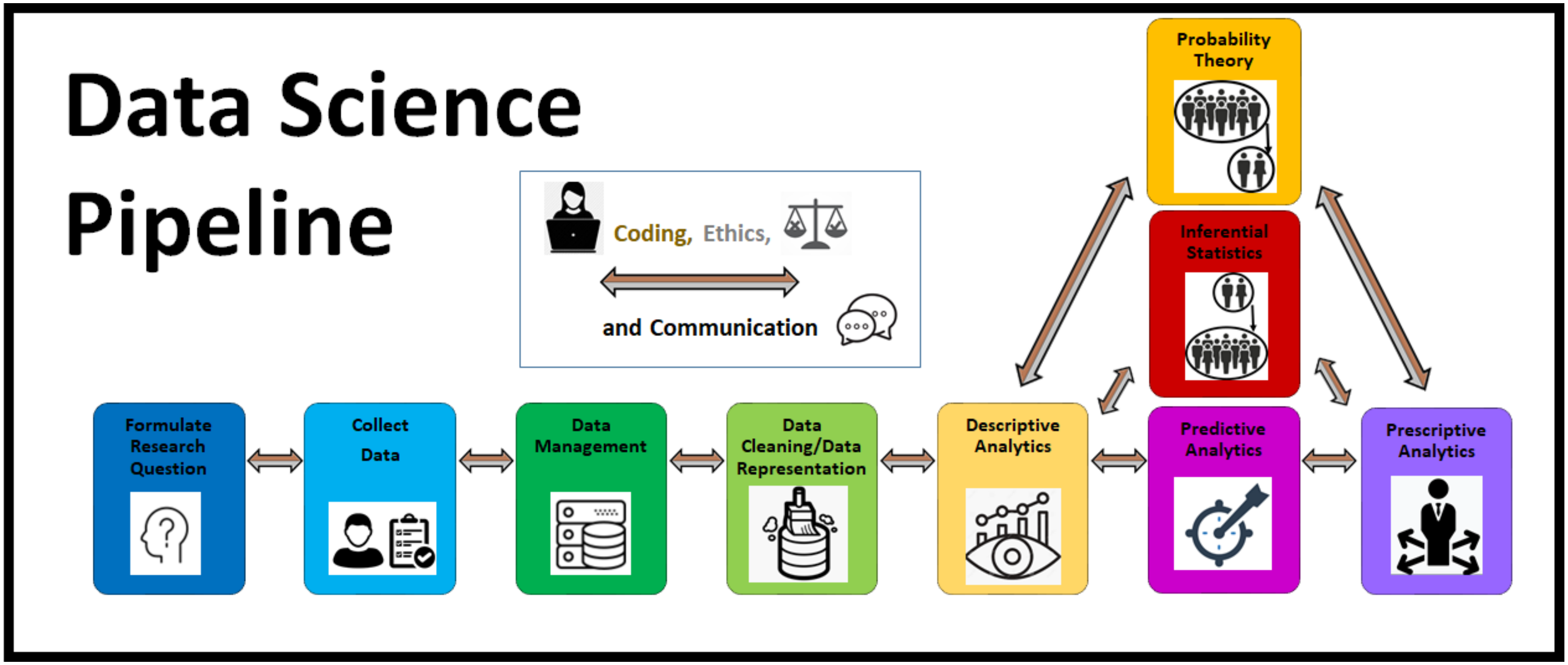
What are some ways that our dataset might not be "clean"?

What is one type of plot we could use to learn more about this dataset?

How might we want to use this dataset to make predictions?

How might a person use this this data to make "good decisions"?

Statistics vs. Data Science



Data Science

Data science places more focus on the ___________________ nature of this pipeline and how decisions or insights discovered

made _______________ in the pipeline can influences insights or decisions made __________________ in the pipeline (or vice

versa).

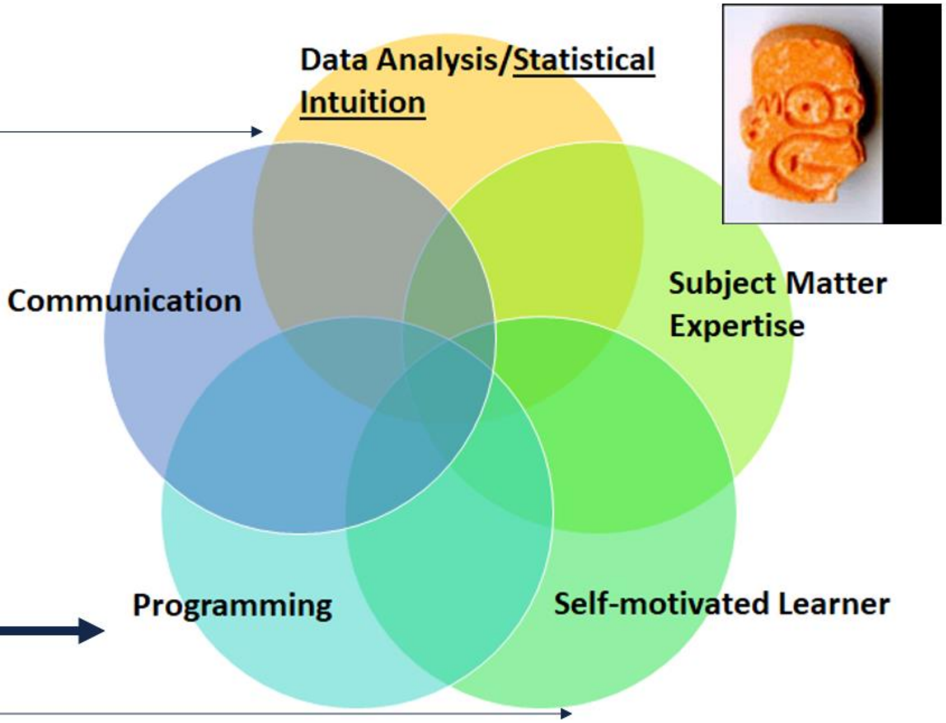
Statistics

Mostly focuses on just: 1.) probability theory, 2.) inferential statistics, and 3.) ___________ predictive analytics techniques in a way that focuses less on the other parts of the data science pipeline.

SKILLS NEEDED BY A DATA SCIENTIST *AND WHY*

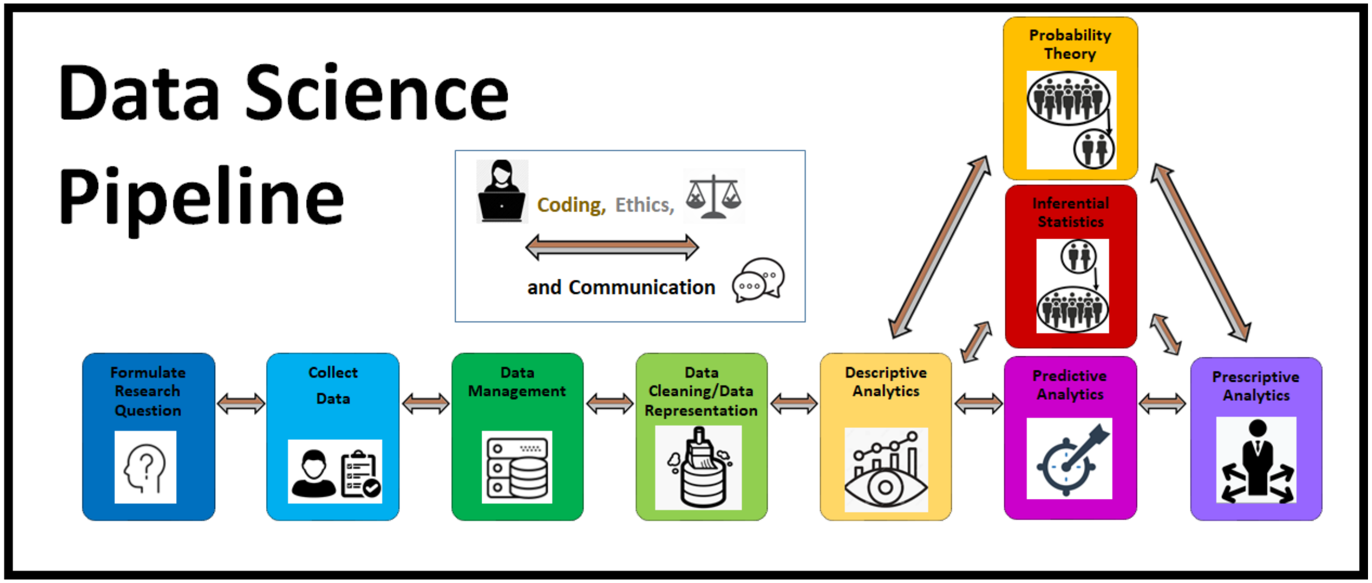
Current State of Data Analysis/Science:

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

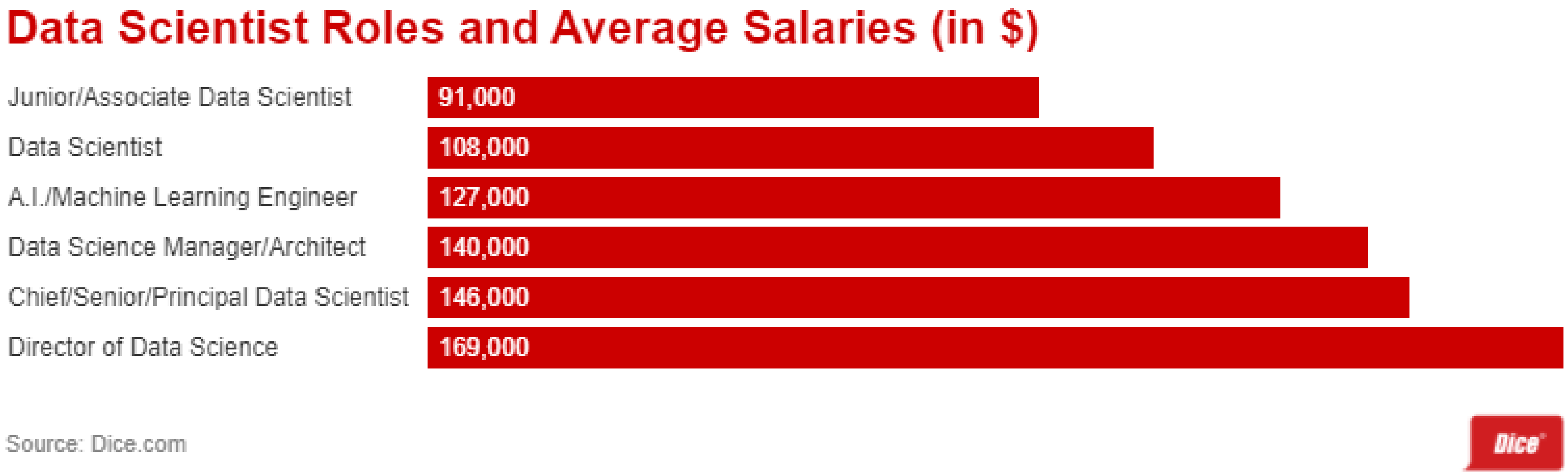
COURSE GOALS



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/fall2022/stat207/course_topics.html

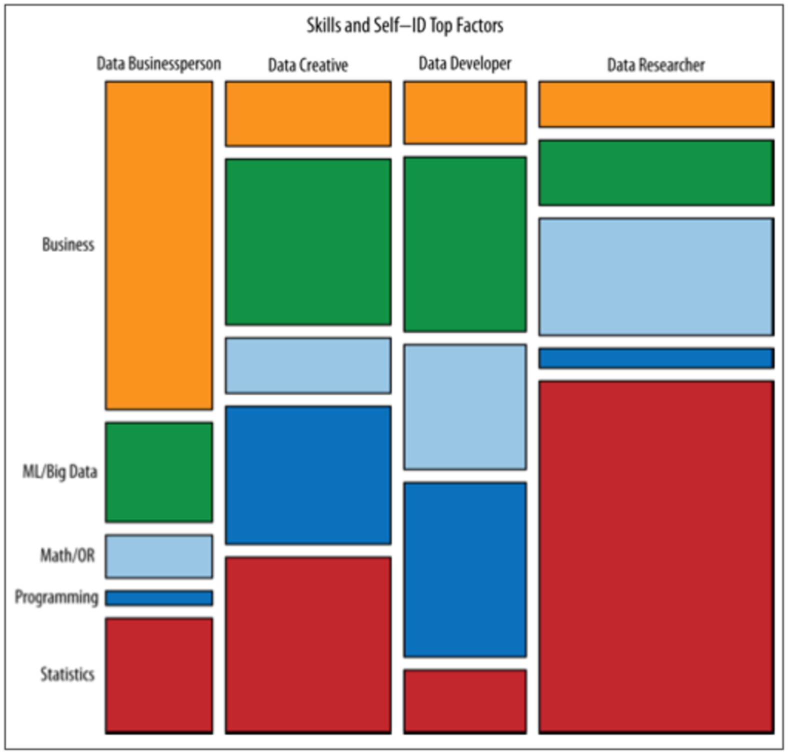
WHY STUDY DATA SCIENCE?



Data Scientist Roles and Average Salaries (in $)

Source: Dice.com

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://canvas.illinois.edu/courses/30296

1. Your grades
2. Lecture markups (in Files tab)
3. Post-Lecture videos
   - *Note: There was a **9% difference** between:*
     - *the median final grade of students **who regularly attended lecture** last semester and*
     - *the median final grade students **who did not regularly attend lecture** last semester.*

Course Website: http://courses.las.illinois.edu/fall2022/stat207/

1. Course schedule and *incomplete* lecture notes (to be filled out in the lecture).
2. Syllabus
3. Assignment and Project Information
4. Tech Guides
5. Course Content
6. Course Staff Info

UIUC Coursework Github Enterprise Organization

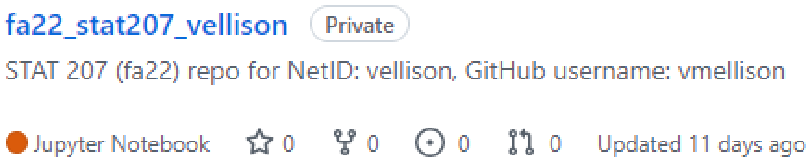
https://github.com/illinois-cs-coursework

1. **Your netid repository**

   URL Format: https://github.com/illinois-cs-coursework/fa22_stat207_YOUR_NETID

   - **push** your completed lab assignments here for grading

     Ex:



fa22_stat207_vellison Private

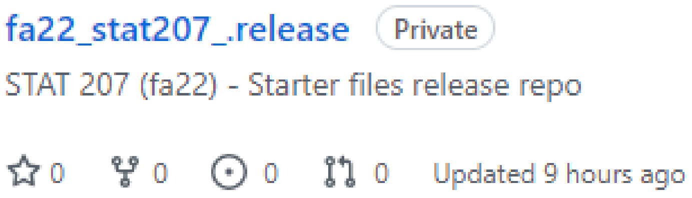
STAT 207 (fa22) repo for NetID: vellison, GitHub username: vmellison

● Jupyter Notebook ☆ 0 0 0 0 Updated 11 days ago

2. **fa22_stat207_.release repository**

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

     https://github.com/illinois-cs-coursework/fa22_stat207_.release



fa22_stat207_.release Private

STAT 207 (fa22) - Starter files release repo

☆ 0 0 0 0 Updated 9 hours ago

Lecture Format

During Lecture

- **Lectures are Synchronous and In-Person:** attendance strongly encouraged if you are able to, but not required! (+1 Bonus Point for Attendance and Completing the Summary)

- **"Skeleton" Lecture Unit Materials Posted Before Class**
  - http://courses.las.illinois.edu/fall2022/stat207

- **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 **Tuesday night 11:59pm CST** at the latest.