CSE 519 -- Data Science (Fall 2021)
Prof. Steven Skiena
Semester Projects

Proposal Due: Thursday, October 21, 2021
Progress Report Due: Thursday, November 11, 2021
Final Report Due: Thursday, December 2, 2021

The project will involve concentrated work in one research project related to data science. Certain aspects of the project will have to change to support the size of the class, but we are structuring the projects and grading in such a way as to provide meaningful feedback for large numbers of students.

The possible choices of projects will be constrained to those listed below, so there will be several groups working independently on each project. You will not be allowed to propose your own project topic, but each topic leaves enough room to pursue distinct directions so that I expect to see variety among the submissions. Each project will have a TA or other graduate student serving as "captain" for the project. Your captain is the first line of defense for individual discussions about your team's particular project. I will also be encouraging the use of Piazza for questions/discussions about your project, as well as occasional "town halls" about each project in or immediately after class.

While I anticipate that much of the work will be done as the deadline approaches, it is important to get started early enough to discover insurmountable roadblocks in data acquisition or problem definition before it is too late. The project proposal and progress reports have been instituted to ensure people get serious well in advance of the final deadline.

Each group will be responsible to turn in 3-5 page project proposals/literature search and progress reports as of the dates above. I will award almost half of the total grade for each project on the strength of the preliminary reports. This is to encourage starting early, and to make sure that you and I both know what you are to do before it is too late to avoid trouble.

My hope is that the best of the submissions for one or more of these topics will lead to published work. This has been the case several times when I have taught such a course.

Rules of the Game

- From past experience, I have found that students tend to herd on a small subset of possible projects. Further, they are irresistibly drawn to poor choices: that offer the least available data, that have limited ways to evaluate success, or little freedom for creativity in modeling. I have colored the topics I think best in red and marked (***) Please consider projects which seem less popular -- likely they will prove better to work on.

- Take the time to investigate multiple options during the proposal phase. You should have gotten some preliminary results by then, enough to provide confidence you will be able to successfully complete the project.
- I reserve the right to reassign the groups with the weakest proposals to other topics for the second phase of the project, both to rescue them from a bad situation and rebalance the class.
- A substantial part of the grading of the projects will be done by peers: each student will be charged with grading three other projects on the same topic. Part of your grade for the course will be a function of how seriously you take this, e.g the quality of your feedback. Peer review is how research papers are selected for publication, and I hope this will be revealing. I am hoping that this review will provide insights to improve your own projects as well.
- To preserve anonymity in submissions, it is important that the papers you submit online for grading **not** contain your names or ID numbers. **Indeed, the peer grading form/rubric will include a question if the grader can figure out whose paper they have, and if so we will take off points.**
- Expect to be asked to do a trial peer grading exercise to test our system soon. Please respond promptly.
- Group sizes should usually range from two to three students. The amount of work per group is expected to scale linearly with the effort involved. The projects are large enough that I do not think students working by themselves can achieve the critical mass to get a substantial project done, and are strongly discouraged.
- Each student will be asked to proportion their group's effort among all the team members they are working with. So make sure you work with people who you like, trust, and respect.

I will possibly add another few projects later, so check back occasionally.

Twitter and Circadian Rhythms
Captain: **(Adi) Adithya V Ganesan** <adithya.virinchipuramganesa@stonybrook.edu>

Social media analysis sheds considerable light on human behavior, gaining statistical strength from the scale of such interactions. In this project, we will analyze Twitter data to get insight into factors affecting how much sleep different populations receive.

Prof. Jason Jones of Sociology has used Twitter for studying sleep patterns, most notably his study of the performance of NBA basketball players after unusually late nights, as measured by late night tweets. A good way to start this project is to read papers on this work, particularly:

Jones, J. J., Kirschen, G. W., Kancharla, S., & Hale, L. (2019). Association between late-night tweeting and next-day game performance among professional basketball players. Sleep Health, 5(1), 68-71.

Kryger, M. (2017). What can tweets tell us about a person's sleep?. Journal of Clinical Sleep Medicine, 13(10), 1219-1221.

Prof Jones has been collecting a 1% sample of all tweets from February 2015 through to the present. You will be provided with a dataset that is a long list of timestamps extracted from Twitter data with locations and characteristics. Timestamps are server-based UTC timestamps of an activity (posting a tweet, in this case). Locations are user-supplied, free response text strings that may or may not indicate a geographical place. Characteristics are mutually-exclusive category labels for the user (e.g. the user described themself as a doctor, a nurse, etc.).

Demonstrate that this data can be used to plot circadian rhythms -- the cycles when people are active. For example, we presume that users are more active during local daylight hours. You should be able to show how this activity shifts seasonally as local daylight hours shift.

We are particularly interested in how the circadian rhythm of different groups differ. We might expect that musicians have a different sleep pattern than dentists, and truckers from teachers. Your task is to characterize Twitter bedtimes per characteristic group. Perhaps "Twitter bedtime" can be defined as the onset of the 8-hour block of time with least activity within each 24-hour noon-to-noon local day. Contrast Twitter bedtimes across user characteristics. How stable is this bedtime? For whom is it most variable?

There are many interesting ways to break down this data. Possibilities include:

- *Sunlight and seasonal effects* -- The time of sunrise varies with latitude and season. A reasonable hypothesis is that sunrise affects the time people awaken in the morning. Can you produce a satisfying sleep map video showing wakeup times by place for each (say) week of the year? Are bedtimes in sync with this, or do people get less sleep in summer and more sleep in winter? Are their periodic changes in the temperate regions, which have less fluctuation in sunrise/set times than more polar regions?
- *Weekend and holiday effects* -- How do holidays (where people are less likely to travel to work) change people's sleep schedules, and how long does it take for them to recover?
- *Time zone effects* -- Because time zones are broad, there is more light at 7AM on the eastern side of a time zone than the west. All presumably work the same conventional hours (traditionally 9AM-5PM). A reasonable hypothesis is that people on the east end of a time zone wake up earlier than the west -- is this true?

- *Daylight savings time effects* -- How do sleep times adjust to annual one-hour time changes in fall and spring? Do people gain/lose an hour of sleep, or does it come from other activities? How long does the change in sleep times last?
- *Job title effects* (*****) -- The self-reported description string for each Twitter account gives insight into how people see themselves. Often it contains information as to career and interests. Are there differences in the sleep patterns of truckers and teachers? Prof. Jones has done some work on these strings, but the first task is to classify them into groups.
- *Age, gender and nationality effects* -- Do sleep schedules vary significantly among people of different demographic groups? Be careful: these effects must be separated from sampling and geographic biases.
- *Time course effects* -- As the years go by, are people getting more or less sleep? Be careful: these effects must be decoupled from changes in the popularity of Twitter and all the other factors discussed above.
- *Travel effects* -- Geolocation tags and self-reported tweets should identify people who had cross-country travels. How did this affect their sleep patterns, and for how long?

I am much more interested in projects which do a sophisticated and careful (statistically defensible) analysis of one of these questions than a hack programming job that does a lousy job tabulating all of them.

There are many effective ways to mess up your analysis. Some of these include:

- *Bots and corporate sources* -- Mechanical sources do not sleep according to a human schedule. Companies are most active during the business day regardless of location or season. One must try to identify and eliminate such sources.
- *Sampling effects* -- You are only seeing 1% of all posted tweets, which means you are unlikely to see the first and last tweet of the day for people. An important orthogonal analysis might be to analyze the full profile of a smaller number of people, with limitations imposed by Twitter APIs.
- *Statistical sharpness* - You need the right statistic to pick up the subtle phenomenon of the first and last tweet of the day. Perhaps it is over or under representation in each time window over a baseline, but you need to do some clear thinking to get something meaningful.

Fantasy Premier League (FPL)

Captain: **Tanzir Islam Pial** <tanzirislam.pial@stonybrook.edu>

The English Premier League runs each year from August to May. Parallel to it, runs the fantasy premier league (FPL). In FPL, a user has to make a team under certain constraints. When players in his team do something good/bad on the pitch in real life, his team earns/loses points. Every year more than eight million people around the world play this game.

Official website of FPL: https://fantasy.premierleague.com
Rules of playing FPL: https://fantasy.premierleague.com/help/rules

Vast amount of data related to FPL is publicly available on the web. FPL has an official API through which you can collect data. People have created datasets using FPL data and python wrappers for easily accessing the FPL API too. These are few of them:

1. https://buildmedia.readthedocs.org/media/pdf/fpl/latest/fpl.pdf
2. https://github.com/vaastav/Fantasy-Premier-League
3. https://www.kaggle.com/ollywelch/fpl-dataset

There are a number of interesting directions this project can take.

- *Player prediction model* -- You can create a model that can predict the scores of each footballer in a given gameweek(GW) and then create the most optimal team for that gameweek under the budget and other constraints. For example if your model is predicting the team for a GW that starts on November 13, 2020, it can use any data available on November 12, 2020 and create the team based on that. You should be evaluating the performance of your model for at least 38 GWs and show where your model would have ranked in that GW in the global leaderboard. You should create baseline models (That do not use any training) who pick players based on very simple algorithms (Like picking the highest scoring players from the previous gw) and compare your final model's performance against that of the baseline model.
- *Team construction model* -- Instead of predicting for a single GW, you can take one step further and create a model that plays an entire season! You should be evaluating your model's performance similarly as above. It will be easy to obtain great performance by overfitting on data that existed only after the date your construction was supposedly performed, so you must be careful to validate this correctly.
- *Optimal transfer strategy* -- Given a team and its budget for a GW, what is the most optimal transfer strategy if you just wanted to maximize your points for the following GW? You will evaluate your performance for all real users (who made a transfer in that GW) for at least 10 gameweeks and show how many more points users could have scored if they had taken your model's strategy in that GW.
- *How good are top fantasy players?* -- Another direction this project can take is analyzing the top fantasy players. You will show analytically how their strategies differ from lower ranking players. There are players who consistently place among the top 1% every year, including, surprisingly, the Chess world champion Magnus Carlsen (https://fantasy.premierleague.com/entry/132645/history). You must come up with claims regarding their strategy backed by data.

- *Real vs. Fantasy Performance* -- You can also analyze how much the FPL potential of players of a club translates into their real life performance on the pitch. Find out some scenarios where a club's footballers have strong performance in FPL scoreboard but the club is struggling on the league table.
- *Analysis of real life transfer activities* -- You will gather all transfer data where a footballer has moved from a Premier League (PL) club to another PL club. You have to predict how well he will perform in his new club with respect to FPL after the transfer. For this task you can use any data till the transfer date and then predict that player's score for a full/half season.

You are free to explore your own ideas related to FPL too.

The analysis tasks are more open-ended. So you must come up with strong benchmarks / evaluation criteria if you are doing those tasks.

Value or Significance Scores of Places

Captain: Zhan Shi <shi.zhan@stonybrook.edu>

The late, great store Book Review in Huntington NY just closed, presumably a victim of a landlord seeking more rent. It was the best bookstore on Long Island, and a real community treasure. I met my wife there for our first date in the travel section -- but now it is gone and eventually destined to be replaced by just another chain store. Is there a way to quantify the community value of a business in an objective way, so the proprietors are given any help they need to ensure survival?

Certain businesses like stores and restaurants are irreplaceable to a community, being of particular meaning, value, or significance as a function of popularity, distinctiveness, and/or age/history. A beloved local independent bookstore is generally more important to preserve than any single McDonalds restaurant. Important local businesses are often under threat because there is no way to recognize their importance to the community.

This project asks you to build a dataset regarding businesses in a given town or region, quantifying variables relevant to their importance or significance. You need a large enough, rich enough data set to make this meaningful. It may be comprehensive by type (e.g restaurants, churches, book stores, or clothing stores) or by location (a big city, state, or ideally the entire country), but it needs enough heft (both number of rows and number of columns) to be interesting.

Then produce a score to measure the merit of each business. Ideally you may produce multiple scores and see where they agree and disagree. If you can figure out independent measures or proxies for what we want to assess the score by, this would be great.

Finally, explore the consequences of your measure. What do visualizations tell you about how merit is distributed geographically and culturally? Use it for studies of other phenomena. Do beloved places go out of business more or less often than other places? What are the most valuable endangered businesses?

Retrieval Patterns of Physical Objects from Retail Sales Data (***)
Captain: tanzirislam.pial@stonybrook.edu

A data set concerning sales at a hardware store can be used to address an interesting old-style psychology problem. Given a bunch of items to grab, where all items must be grabbed, do people grab the small/light ones first or the big/heavy ones? The most fragile or squishy or what when? This data set has millions of baskets of items, ordered by when they were given to a cashier.

There are two laws of psychophysics that come to mind. Fitts' Law would predict the largest packages would be the easiest to grab, but not necessarily the first. Maybe one could look at inter-item scan times as a function of scanned item size. Hick's Law would predict the rate of grabbing would increase from first item to last as the number of choices decreases. Again inter-item scan times would be the first place to look. What does the data suggest?

A small, local chain of retail stores has graciously made its market basket data available to us for analysis. This data consists of records of several million sales, over several years, of tens of thousands of different products. This is a rich dataset, which offers many directions for possible investigation. A tiny example of the file is available here. Fields include:

Date,Transaction Time,Customer Number,Sales Header,STR,Item Number,Item Description,Net Sales Units,Net Sales,Cost,Gross Margin,Gross Margin %,Class Code,Class Name,Department Code,Department Name

The order of the transactions in the database presumably reflect the order in which the cashier scanned them (*first thing -- check out whether this is true*). One possible formulation of our task is -- given the set of items in a transaction, predict the order in which they will be scanned.

Teams are encouraged to think creatively about what directions to pursue, although some ideas are given below:

- *Product embeddings* -- Build distributed representations of products (embeddings), perhaps using category data or the network of co-purchased items (or both). Do these embeddings provide interesting insights into which products are similar or related?

- *Size and weight analysis* -- Can you map each object to a likely size, weight, texture, bulkiness measurement to inform your analysis?
- *Validate Fitt's and Hick's Law here* -- Look at the times between items to see if we can detect evidence of these phenomena in the data sets.
- *Time series modeling of scanning order* -- Is there evidence that scanning order differs with different cashiers or time of day? Is there temporal locality in the data (likely reflecting the same cashier) or do all people behave the same.

This data is not for public distribution. Any student who takes on this project will have to sign an NDA, and pledge to keep the data private. All student reports will be made available to the company for review, which may suggest specific directions for future work.

Understanding Flight Delays

Captain: (Adi) Adithya V Ganesan <adithya.virinchipuramganesa@stonybrook.edu>

Flight delays are a bane of the modern world. About 31.1% of the flights were delayed by at least 15 minutes in 2013. These delays cause inconvenience and economic issues for passengers, airlines and the airports.

This project asks you to make a study of flight delay data to better understand how commercial aviation works, and predict when delays are likely. We list some possible questions below. You should do a few in depth rather than many shallowly, and *we encourage you to come up with your own questions/modeling challenges related to flight delays.*

Questions:

Weather, Climate and Geography:

1. Are airlines better prepared for extreme climates (time of the year) than weather changes?
2. What weather factors dictate the flight delay?
3. Can we predict the delay by knowing the weather conditions ahead of time?
4. What can you inform of the airports situated in atypical geographic locations in regards to flight delay?

Boarding delays:

1. What are the issues that cause boarding delays? What is the distribution and its effect on delay?
2. Does the cabin crew size have a correlation with the delay?

Flight Maintenance:

1. Does flight maintenance play an important role in decreasing the delay?

2. Do larger commercial flights delay more than smaller and medium sized flights?
3. Typically what kind of issues lead to delay in flights under this category?

Airports and Airlines:

1. How frequent are delays in busiest airports compared to moderately busy and least busiest airports?
2. Do the airlines that run a tight ship (smaller market size, customer capacity, target routes etc) do a better job at keeping time than the big players?
3. What is the delay rate of newer airlines?

Other Analysis:

1. Do flight delays have a cascading effect? What is the extent of the effect a delayed arrival has on departures scheduled afterwards?

Useful Resources:

1. https://www.bts.dot.gov/explore-topics-and-geography/topics/understanding-reporting-causes-flight-delays-and-cancellations-0
2. A Review on Flight Delay Prediction
3. Flight data for United States aviation can be found here.
4. Weather API

U.S. Covid-19 Vaccinations Analysis at County Level

Captain: Zhan Shi <shi.zhan@stonybrook.edu>

Up to 09/20, there are about 55.4% people in the U.S. fully vaccinated and nearly 65% people have at least one dose, which is a significant achievement. But recently the vaccination rate seems to grow slowly, and the Covid-19 infectious number continues to increase. We are in a new challenging situation. And the Covid-19 Vaccinations data could be found here.

This project asks you to study available vaccination data to inform us of interesting things about discrepancy in rates among communities (say county-level data in the U.S.) and ways that we can increase the vaccination rate. The issues here are complicated -- some economic, some cultural, some scientific. Some ideas are proposed below, but you should propose other illuminating questions and use data to answer them.

Questions:

Vaccinations Analysis

1. Which factor of the county may affect the number of COVID-19 vaccinations, like economic, education, state policy, location and so on.
2. Do personal factors like age affect the vaccination rate?

3. Are there some exceptional counties that behave extremely differently from the surrounding counties? And what makes them special？

Prediction and Outlook

1. Looking at the current vaccination trend, when over 80% people could be vaccinated for all regions to reach a sort of herd immunity?
2. According to Covid-19 infectious numbers for each county, if higher vaccinated counties have a relatively lower infectious rate? Are there any violations? And why?
3. By analyzing the different aspects of Covid-19 related data and county data, could you give some insights or suggestions for the county to improve vaccinations and reduce daily infectious numbers?

Job Title Dataset Development and Analysis -- Prestige and Mobility (***)
Captain: Lu Chen luchen2@cs.stonybrook.edu

The prestige, compensation, requirements of mobility of jobs are largely encapsulated by their title. Professors probably rank higher in prestige and education than salary. The wait staff in the front of a restaurant probably has both higher prestige and compensation than those cleaning the dishes in back, although possibly similar education. This project asks you to build a comprehensive data set concerning jobs and title and then analyze it to say interesting things about how the world works.

This project starts by developing a data set with one row per job title, and columns capturing attributes of these jobs -- ideally quantities like average salary, number of people with this title, average education level, measures of prestige, happiness/satisfaction, industrial sector, average age, etc. There are presumably large government data sets about jobs and employment and titles -- do a serious search for these and clean/integrate them to form the foundations of your dataset. Retain code numbers and other important information. Maybe the career center on campus also has interesting data, or standard business data sets held by our campus library.

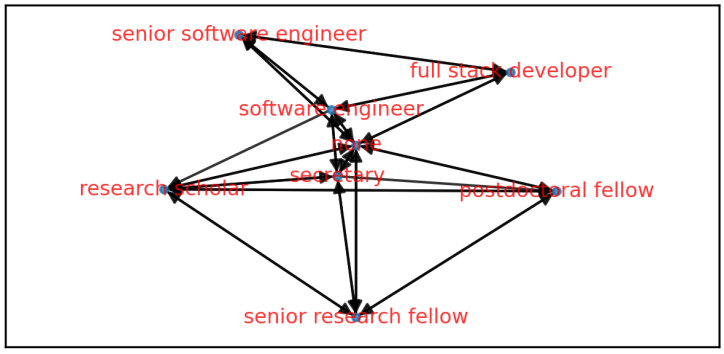
There is considerable room for innovation in building this dataset after you establish the foundations through government statistics:

- Different datasets will use different sets of job titles. You should be careful to unify titles into subsets that mean similar things.
- Certain columns will have large numbers of missing values. Set up an imputed value column which is complete next to the original raw data for comparison.
- There are presumably clearly defined hierarchies within any industry. Assistant Professors are outranked by Associate and full professors. Organizing professions into natural hierarchies will be a useful thing to do for imputation, etc.

- There are many other types of data which can be used to deepen this analysis. Analysis of resumes found on the web may be informative. LinkedIn has a tremendous amount of information about job offerings and titles. On social media like Facebook and Twitter and Google people talk a lot about jobs (can you quantify how much chatter there is?)
- Is there relevant data here from other countries? My guess is that relative salaries and prestige differ substantially across different countries.

Our particular interest is analyzing job prestige and transitions through social media analysis. But we need you to build a comprehensive dataset of job properties before playing with this. We collected 25,033 self-reported job titles from millions of public twitter biographies from 2015-2021. With this longitudinal data of six-year time period, we could observe the occupational changes of people, reflecting graduation (e.g.,: from *Student* to *SDE*) and career advancement (e.g.,: from *SDE* to *Senior SDE*). This motivates questions like:

1. Job Popularity: Which jobs are mentioned most ?
    a. Most popular jobs reported on Twitter ***V.S.*** Most popular jobs in the real world.
    b. Do high-salary jobs have more exposure on Twitter?
2. Popular Transitions: Which transitions are more frequent ? and what do they mean?
    a. It could mean job title rephrase (VP -> vice president)
    b. ... career promotion (SDE ->Senior SDE) ...
3. Transition Directionality: Which transitions are one-way? And what do they mean?
    a. SDE->senior SDE is supposed to be more common than Senior SDE->SDE.
4. Auto-Job Categorization: Which jobs should fall into the same bucket?
    a. Can you tell the different jobs apart from unnormalized titles? (using graph partitioning, node embeddings, char-embeddings, etc)



    b. Standard Occupational Classification (SOC): https://www.bls.gov/soc/
    c. Job salaries estimation: https://www.bls.gov/oes/2020/may/oes_nat.htm
5. Topmost Master: Which are the most prestigious jobs and how do they correlate to wage?

GitHub Demo with data : https://github.com/zjlxgxz/SBU-CSE-519-2021-Job-Graph-Demo

Segmenting Books by Themes and Topics
Captain: tanzirislam.pial@stonybrook.edu

The StonyBook project (www.stonybook.org) seeks to do natural language processing (NLP) on a book or even library scale, working on our corpus of 50,000 novels. We have developed a pipeline for book processing, and present a variety of types of analysis on our website for each of these novels.

One type of analysis we do not really do at this time is chunking or segmenting books into "incidents" or "actions". In particular, we would like to be able to take a given book and identify which parts are romantic, where there are action scenes, and whether something is intended to be funny.

Perhaps the simplest version of this problem is genre identification -- given a book, label it as a romance or mystery or adventure… It might be good to start here to get a handle on the problem. Identify a set of possible genres, build a training set and develop an appropriate classifier. This is a problem which has previous work, so look for it. To get you started, here is a dataset containing books and their genres: https://github.com/uchidalab/book-dataset
And the related paper: https://arxiv.org/abs/1610.09204

More complicated is the question of getting the appropriate segments and labeling them in a single book. For segmenting a book into multiple meaningful segments denoting a switch of context, this paper may be helpful: https://arxiv.org/abs/2011.04163 The authors are from Stony Brook. Contact them to get started on segmenting!

Perhaps start with labeling individual chapters, but it is necessary to either find appropriate training data or figure out how to do this in an unsupervised manner. Getting annotated data for books is hard because books are long and thus expensive for humans to read. But perhaps you can leverage existing resources in a clever way. Note that episodes are often just a few paragraphs within a chapter, so recognizing the boundaries is hard.

Approaches to segmentation in an unsupervised way might involve topic modeling techniques like latent Dirichlet allocation (LDA), or identifying language from different lexicons like Wordnet, Roget's thesaurus, or another resource. Properly evaluating the quality of such a segmentation requires some thinking

Finally, there is an interesting visualization aspect to this -- how can you represent your analysis of a book graphically so we can see what its structure is.