Student Project Ideas

- Movie Script Analysis
    - Core: Plain text of movie scripts are all online at http://www.imsdb.com/. Popularity of quotes and quote content from all of these movies are available within each movie's page. Combine the popularity of each quote with the local features of it to determine what parts of movies are most memorable.
    - Questions to answer
        - Are memorable quotes more likely to come from the main character or from minor characters
        - Are similar scenes from movie remakes likely to be just as popular as the original?
        - How many actors are in the scene for the top quotes? The minor ones?
        - When an actor wins Best X award, such as the Oscars, what are the features of the script?
        - How do artsy movies compare to action movies? Documentaries to Romantic Comedies?
    - Drive: Social Media attention - Several of the top entries on r/dataisbeautiful are about easily accessible topics like movies. Scripts all have similar formatting, and extracting the speaker along with their quote is trivial
    - Difficulty: Both IMDB and IMSDb are easily scraped, there appears to be zero scraping protection. Matching quotes to scripts shouldn't be that hard, as strings will be identical between them, and scripts have similar formatting to distinguish between scenes
- Charity analysis
    - Core: Several databases for charities now exist, some rating the effectiveness of individual charities. What are the common features of effective charities? Scrape websites like Charity Navigator to build a profile of their top rated selections and compare to each other
    - Questions to answer:
        - How does amount of funding relate to the effectiveness rating? For entries where historical data is available, how does the effectiveness rating change with more funding?
        - Are older charities necessarily higher rated than newer ones?
        - How many board members are optimal?
        - Where available, how much overhead is optimal?
    - Drive - Analysis like this will be very quickly shared through the EA scene, and could result in several invitations to speak on the subject.
    - Difficulty: Mostly in building the scraping engine. Once data is gathered on a few hundred to a few thousand charities, just throw a grab bag of techniques at it and see what happens e.g. PCA, dimensionality reduction, etc.
- Police Stops analysis

    - Core: Look for salient features in policing using the Los Angeles Police Stops dataset.
    - Questions to answer:
        - How does policing change within a week of a major incident such as a nationally covered officer's death or protest/riot?
        - How do the profiles of arrested / stopped persons change over the course of the day? Are more nonwhites stopped in the evening? Are more vehicles stopped in the evening?
        - How do the local race distributions for each region compare against their likelihood of being stopped (if there are 50% nonwhites in Hollywood, do they make up 50% of the stopping statistics?)
        - If you gather the population of foreigners in each region through census data (Dave can show whoever picks this how to access), how does the population of foreigners affect the stopping rate?
        - How do new officers behave?
    - Drive: Social media sharing
    - Difficulty: Mostly machine learning given different constraints. The Stops dataset is mostly clean, so most time on the project is setting up learning on the large volume of data.
- College Rating Prediction
    - Core: Look through historical data on college ranking to find the main factors driving higher and lower rankings results over time. I can't seem to find the dataset, but it was about 1-4 GB of data going back 10 years, with 60+ features of over 2000 colleges
- Real Estate comparison
    - Core: Real estate companies would be interested in anything you could tell them about their market. Acquire local housing prices to make a predictive model for housing price, whether the house is available to sell soon, etc.