Jessica Wulzen - Capstone Project Ideas
Project idea 1: Poker
http://poker.cs.ualberta.ca/irc_poker_database.html

Data: Database of 10 million + hands of No Limit Texas Holdem
Topic: Preflop poker strategy. Determine how successful each 2 card hand is based on number of players, table position, betting dynamic, and hand outcome. Use this to determine what the playable range of cards is by position, and examine how different types of strong hands (pocket pairs, suited connectors, Ace+X) differ.

Project pros:
-Enormous amount of data
-Topic with a lot of different directions to take analysis
-Opportunity to create interesting visualizations, summarizing complicated poker strategy concepts visually

Project cons:
-Not clear how ML algorithms could be used, seems more focused on analyzing data vs making predictions

Project idea 2: Online toxicity
-Kaggle competition
https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification
Data: Kaggle dataset of 1.8 million online comments, with
Topic: Can you help detect toxic comments — *and* minimize unintended model bias? That's your challenge in this competition.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both part of Alphabet), builds technology to protect voices in conversation. A main area of focus is machine learning models that can identify toxicity in online conversations, where toxicity is defined as anything *rude, disrespectful or otherwise likely to make someone leave a discussion*.

Project pros:
-Topic I find very interesting/relevant
-Opportunities to expand beyond the scope of this project into moderation of online communities (youtube comments, reddit posts, twitch chat, etc)
-Part of a popular Kaggle competition, allowing for visibility + the ability to learn from how others are tackling similar questions

Project cons:
-Data comes as a single download, lowering the opportunities for data wrangling (this may be offset

by opportunities to work with the text of comments)
-As a popular + well-funded competition, there are many people working on this making it a less 'unique' project

Project idea 3: Music reviews
https://developer.spotify.com/documentation/web-api/
https://www.kaggle.com/nolanbconaway/pitchfork-data

Data: Kaggle dataset which has scraped 15+ years of music reviews from Pitchfork.com, combined with Spotify API to capture data on musical attributes
Topic: Analyze data of reviews to look at connections between the language of the review and the score assigned. Combine this with data from Spotify about each song's individual characteristics (such as energy, key, danceability, etc) to see how critics feel about musical trends.

Pros:
-Many directions to take analysis
-Uses data from multiple sources + involves taking data from an API
-Interesting and relatable idea

Cons:
-Similarly to the poker question - not clear where ML algorithms would be used
-Process of linking Pitchfork reviews with the Spotify data may be difficult to implement/out of scope