**Capstone 2 project ideas:**

1. [Predicting outcome for H-1B eligibility in the US](https://www.kaggle.com/elraphabr/predicting-outcome-for-h-1b-eligibility-in-the-us)**:** below is the link from kaggle.

https://www.kaggle.com/elraphabr/predicting-outcome-for-h-1b-eligibility-in-the-us/data

This dataset contains five year's worth of H-1B petition data, with approximately 3 million records overall. The columns in the dataset include case status, employer name, worksite coordinates, job title, prevailing wage, occupation code, and year filed.

**Inspiration:**

Is the number of petitions with Data Engineer job title increasing over time?

Which part of the US has the most Hardware Engineer jobs?

Which industry has the most number of Data Scientist positions?

Which employers file the most petitions each year?

**2) Trending YouTube Video Statistics:** below is the link from kaggle.

https://www.kaggle.com/datasnaek/youtube-new

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day. Each region’s data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. The data also includes a category\_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

**Inspiration:**

Sentiment analysis in a variety of forms?

Categorising YouTube videos based on their comments and statistics?

Training ML algorithms like RNNs to generate their own YouTube comments?

Analysing what factors affect how popular a YouTube video will be?

Statistical analysis over time?

**3) MovieLens 20M Dataset:** below is the link from kaggle.

https://www.kaggle.com/grouplens/movielens-20m-dataset

The datasets describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These data were created by 138493 users between January 09, 1995 and March 31, 2015. This dataset was generated on October 17, 2016. Users were selected at random for inclusion. All selected users had rated at least 20 movies.

**Inspiration:**

Which genres receive the highest ratings? How does this change over time?

Determine the temporal trends in the genres/tagging activity of the movies released?

4**) MovieLens 20M Dataset:** below is the link from kaggle.

<https://www.kaggle.com/asindico/reviews-analysis-and-prediction/data>

This is a list of over 1,500 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more provided by [Datafiniti's Product Database](https://datafiniti.co/products/product-data/). The dataset includes basic product information, rating, review text, and more for each product.

**Inspiration:**

What are the most reviewed Amazon products?

What are the initial and current number of customer reviews for each product?

How do the reviews in the first 90 days after a product launch compare to the price of the product?

How do the reviews in the first 90 days after a product launch compare to the days available for sale?

Map the keywords in the review text against the review ratings to help train sentiment models.

5**) News Aggregator Dataset:** below is the link from kaggle.

https://www.kaggle.com/uciml/news-aggregator-dataset

This dataset contains headlines, URLs, and categories for 422,937 news stories collected by a web aggregator between March 10th, 2014 and August 10th, 2014.

News categories included in this dataset include business; science and technology; entertainment; and health. Different news articles that refer to the same news item (e.g., several articles about recently released employment statistics) are also categorized together.

**Inspiration:**

Can we predict the category (business, entertainment, etc.) of a news article given only its headline?

Can we predict the specific story that a news article refers to, given only its headline?

Note: I also wanted to do project on deep learning, but wanted to discuss it first with you as this may take longer and I am not sure about if my laptop will have sufficient amount of resources. Will discuss it with my Mentor in next meeting.