**Team:** Emily Parrish, M.S. Data Analytics

**Github Profile:** https://github.com/emparrish417

**Project Choice: (***#3* *Twitter Analysis*) Leveraging one or more of the databases provided throughout the class, analyze the Elon-centric twitter dataset.

**Proposed Databases:**

Mongo: Existing feature where hashtags are separated from the original text so it eliminates several pre-processing steps.

Arango: Networking and subgrouping based on hashtag content/behavior

**Project Description/Components:**

The goal of this project is to assess the influence of Elon’s account activity on the account activity of his network. This will be looked at primarily through hashtags both in content and frequency metrics. Below lists some relevant sub-activities:

1. Hashtag statistics and EDA: For the entire dataset, look at how often Elon and prominent figures in his network (aka the accounts that interact with him the most) are using hashtags at all and understand if there is a correlation between Elon’s hashtags and different rings of his network.
   1. How often are hashtags used?
   2. Are there any prominent hashtags?
   3. On average, what hashtags are paired together?
   4. Is there an uptick in hashtag activity at a particular time?
2. Evaluate use of hashtags as a labelling technique:
   1. Perform topic modelling on text of tweets in the network
   2. Catalog hashtags into topical areas
   3. Compare relevance of hashtags to prominent topics
3. Identify subgroups based on hashtag activity and perform network analysis to potentially cluster these groups based on relationship with Elon (does he have a scientific/technology network, political network, etc.?)
4. Chronology: Were hash tagging behaviors in Elon’s network similar to Elon from a temporal standpoint? Which (if any) prominent accounts changed subgroups based on the content of their tweets (through their use of hashtags) at a particular time?

Note: The different components of the project are designed so that I can pick one or two based on project scope. They also could be aggregated into one large project.

**Additional Approaches:**

1. NLP techniques (i.e. topic modelling): use of automated unsupervised analysis techniques to compare actual content of tweets to hashtags
   1. Typically, MALLET (open-source) software package used for this.
   2. Possibly use sentiment analysis libraries like Senti-Word NET to characterize Elon’s tweets in comparison with subgroups (would deviate from the hashtag focus)