**Data Engineering | Project Proposal**

**Question/Need:**

1. Promoting positive climate change conversations via Twitter
2. Can we identify “communities” of users that have the potential to form bridges between groups with differing opinions on climate change?

**Data Description:**

1. Tweets about climate change (via Twitter API)
2. Goal: 100,000+ unique tweets

**Tools (Algorithms):**

1. Twitter API and Tweepy - data acquisition tools/libraries
2. AWS and/or Docker - cloud processing and containers (MongoDB & Spark)
3. MongoDB and PyMongo - data storage and processing
4. Spark and PySpark - large-scale data handling and processing
5. NLTK and spaCy - text/language pre-processing
6. Sklearn (PCA, LSA, NFM) - topic modeling
7. Streamlit and/or Flask - web application and model deployment
8. Additional algorithms, tools, and visualization libraries as needed

**MVP Goal:**

1. Check the limitations of data acquisition, if any
2. Finalize data acquisition
3. Initial text preprocessing
4. Basic bar chat plot of various topic associations/terms, word cloud