**Proposal - 04/11/2018 - Brendan Sanderson & Ruben Abbou - CMSC 23900**

* **Dataset:**

We plan to create a visualization of Kanye West’s song lyrics by album. We will use a dataset that consists of all all of the song lyrics for each of Kanye’s songs in each of his albums. The dataset was created by pulling all of the song lyrics from Genius. We have added the dataset to the repository. It is named “Kanye Lyrics.xlsx”.

* **Preprocessing:**

We will read in all of the song lyrics from the dataset file. For each song, we will count the occurrences of each word and then we will then total the occurences of each word from each of the albums and calculate the occurrence frequency. We will do this either using Python or R and create a new data file that consists of the top 50 word occurrences in each album. We will also not include articles or short words such as [“intro”, "a", "an", "the", "this", "that", "of", "for", "or", "and", "on", "to", "be", "if", "we", "you", "in", "is", "at", "it", "rt", "mt", "with"].

.

* **Visualization:**



*Sample visualization taken from an Apple keynote*

We will make a visualization similar to the visualization shown above for each of Kanye’s albums in the dataset, by release date. The words shown will be the most occuring words in each album with a size and color that is related to each word’s number of occurrences. More common words will be larger and be in a brighter color. In the middle, we will display the album cover image. Also, we will add a final visualization of all the most occuring words in all of the albums. We will display the words in the order that they first occur.

The goal is to study the relation between each album and the words used the most by the author of the lyrics and create a visualization that is pretty and would nice on a poster.