# Research Question & Literature Review

# Description of data

This data set (insert the citation for the dataset) was found on the data sharing website [Kaggle.com](http://Kaggle.com). It consists of data scraped from the [TED.com](http://TED.com) website and includes information about the videos including number of views and comments, speaker names, and topics. It also includes a full transcript of each talk that is provided along with the videos by [TED.com](http://TED.com). The data was scraped on 25th September 2017 and includes videos that were published up until 21st September 2017.

As typical for social media data, the views and comments on videos are heavily skewed, with more recent videos showing many more views per day than videos that have been online for a longer period of time. For this reason views and comments have been logged to account for this heavy skewness. Previous research (Boppolige & Gurtoo, 2017) has suggested that after a certain period of days the views and comments videos on [TED.com](http://TED.com) receive become representative of the total amount of views they receive overall. This research cited 46 days for comments and 134 days for views. As only one paper has cited this time line we used this as a guideline and set a conservative limit of 150 days that a video must have been online to be included in our analysis. This will ensure that videos have had adequate time to reach their representative views and comments.

A subset of the videos included in this data set (give the N) were musical performances with limited speaking. As research questions are interested in predicting popularity (as measured by views and comments) using textual analysis of the speech, music videos were removed. Any videos that had less than xxxx words were excluded from the analysis, this includes both musical performances and videos where the transcript was not included in the data file.

# Graphs

Add the following graphs and have a paragraph for each graph that details how it answers the research question – back these up with stats

* Views and comments by:
  + Occupation
  + Theme of talk
  + Gender
  + (Sentiment Analysis)
  + Pos vs Neg sentiment
  + Emotions
  + Laughter ? If we have time??

# Conclusion and Future Directions

One question that arises in the literature is whether views and comments are adequate measures of online content’s performance (Nelson-Field, Riebe, & Newstead, 2013), especially if we are asking questions about virality. Nelson-Field, Riebe, and Newstead (2013) suggest that sharing behaviour may be a better predictor of a video’s success. The dataset we analysed did not have a measure of how many times a video was shared thus preventing us from considering this outcome variable as a measure of the video’s success. Scraping the [TED.com](http://TED.com) website to gather this information is a potential future direction for this project allowing us to have a more accurate outcome measure of a video’s performance.

Another question that arose while analysing this data set was whether the advice that TED gives its speakers is useful in creating highly engaging talks. The research questions focused on in our initial data analysis were based on the extant literature on viral marketing and social sharing. However, TED provides clear directions to their speakers on how to make their talk a hit. In the next stage of this project we intend to test these instructions to see if they pay off.