Jarret Perkins
Evan Segaul
David Taylor

# Tracking TED Talk Trends

https://github.com/esegaul/tedtalkvis

## Background and Motivation:

TED Talks are lectures given by researchers, scientists, and cultural commentators about popular ideas in technology, sociology, politics, and everything in between. We are motivated to examine TED Talk data because we are all avid watchers of the videos that they publish of the lectures, and we believe that the topics covered in TED Talks are good indicators of the ideas that are most popular in science and technology at the time of the talks. We are all interested in new developments in the realm of technology, and we're interested in looking at how certain innovations and ideas change in popularity over time. Moreover, we have an interest in natural language processing (NLP) technologies, and this TED Talk dataset will allow us to gain experience in extracting generalized topics from text data. We have an understanding of machine learning algorithms and wish to expand our knowledge base into NLP with this project. The dataset that we found on Kaggle is ideal for our project because it is mostly complete and has many interesting features that can be visualized.

## Objectives:

In this project, we want to generate insights about the popularity of TED Talks over the years. There are many different factors that go into the success of a TED Talk, so we want to explore what features make a video popular so that a speaker can effectively spread their knowledge to as many people possible. Additionally, our dataset is filled with information about comments and tags that are on the talks, so we can explore the relationship between Talks and its implication on their success. Something that we may end up considering is finding a pattern in the popularity of topics over time, and using that data to create predictions about the popularity of emerging topics in the future.

## Data and Processing:

Our project will not involve many of the typical data processing tasks like wrangling, cleanup, or aggregation since our data is coming from an existing Kaggle dataset (https://kaggle.com/rounakbanik/ted-talks) that is extremely clean and well structured.

The main operation that we will be performing with our data before creating our visualization is feature engineering. When performing feature engineering, we will be combining different categories that exist in our dataset to create new data points that might be useful in our analysis. While we don't know what features we will create before we start analyzing our dataset, some examples of this could be the number of words spoken per time in the video or topic of the video from topic modeling analysis.

## Must-Have Features:

List the features that are absolutely necessary for the project to be successful.

- Visualize the frequency of topics over time (some form of line chart)
    - Popularity on y-axis and time on the x-axis
    - Use this information to see bubbles that appear as topics gain popularity
    - Ability to highlight certain topics to better visualize their trend
    - Encode ratings with color or size of data point
- Encode similarity between video topics using tags
    - The x-axis and the y-axis will contain top tag values across videos and the color of the heat map will encode how frequently those tags appear together
- Encode similarity between the impressions of videos
    - Group by top impression and display number of videos there

    - Allow for zoom feature to go down another level to see videos groups by similar videos
    - Allows us to see how videos are related by impression. Successive levels reveal more a more similar impression left on viewers

## Optional Features:

- Seeing what “clickbait” titles have in common
- Inspecting elements on the individual visualizations to queue more information about specific datatypes
- For heat map, encode number of views for videos with tag pairing with size

## Project Schedule:

- 11/2-11/3: Perform initial data analysis and feature engineering with Python
- 11/8: Give project update -- describe challenges with data and basic insights from preliminary analysis
- 11/9-11/18: Develop prototypes together, meeting several times over the course of a couple of weeks. Work together on all visualizations.
- 11/19-12/2: Finalize visualizations and document processes in process book. Prepare for presentation

## Visualization Design:

### Potential Designs:

- A line graph with a line for each topic showing the number of videos released in each time period.
- A heatmap that uses its axes as two selected categories (tags, speaker occupation, comments) and the size/color is encoded by the number of views.

- A graph connecting related talks where node color describes the primary tag of the video.
- A graph connecting talks by a normalized the number of tags in common.
- Group by most popular tag, sub-group by next most popular tag, sub-sub-group by next most popular tag
- Graph encoding relatedness by a normalized score of ratings. Basically, take the count of the ratings, divide by total views, and then consider a video to be similarly rated for a given quality if their two normalized scores are within a given threshold of each other. Apply this across all of the ratings and that encodes how related two talks are based on the impression they left on the audience
- Heatmap that uses tag values as the axes and the quantitative value is the number of videos in which those tags appear together.

## Final Design:

A dual-view visualization consisting of:

- A heatmap with tag values as the axes and views as the quantitative value to find what two fields are linked together. The tag values are categorical and their linking is quantitative, so a heatmap is appropriate.
- A visualization where each video is grouped (encoded by a bubble with size as the number of videos) by its top rating value. In that subgroup, one can see each of those videos grouped by its next highest rating value. In terms of encoding, we will use space to show related groups and the size of the bubbles to show the number of videos in each group.
- A line graph showing the number of videos for each time period. A line graph is appropriate due to the quantitative venture of the number of videos among a time series