**Title:** The science of suggesting a video.

**Goals of the project:**

Our goal is to investigate how various Ted Talks are related to one another, and how Ted Talks generate views for other Ted Talks. Specifically, we want to find out what factors go into play when Videos are claimed to be related, and we want to analyze how viewers feel about related videos. Our goal is to quantify and estimate the viewer opinion of related videos to see what good Ted Talks have in common, and what unsuccessful Ted Talks lack. We also want to investigate what makes a good Ted Talk title that generates views, so we can make recommendations for titling a Ted Talk.

Through this investigation we can learn more about how a network of media content such as Ted Talks are connected and how that network impacts the success of a video. Further, we can learn more about what makes for a good Ted Talk video, which will undoubtedly reveal insights on successfully creating a title for anything.

**Background and Motivation:**

**Why did we choose this topic?**

This idea appeals to us because of the breadth and detail of data available and the applicability of these insights to other market data. TED Talks are a wonderful testing ground for finding patterns because of these reasons as well as the excitement of applying our knowledge to such a well known entity like TED.

**Why is this problem interesting?**

This problem requires the use of multiple machine learning algorithms, data visualizations, and web scraping efforts to give a comprehensive view into what factors make effective cliques of videos and what constitutes a strong relationship between two videos. This can also be abstracted to business-oriented problems including how to market related products and what constitutes a group of products.

**How is this project different than prior works?**

Although this project idea is similar to prior works regarding TED Talks, it differs in scope that we seek to analyze the entire network of TED Talks to find patterns between them instead of analyzing single talks for useful information.

**Plan to Obtain Data:**

Our main dataset is located on kaggle.com and includes data scraped from the TED website. These data entries include important information such as title, description, ratings, comments, and a list of related videos.  We plan to do further scraping of the TED website for further ratings information if necessary. Some parsing may be done on the comments section to determine users’ sentiment regarding the TED talk.

**Statistical Methods or Machine Learning Algorithms we plan to use:**

* Sentiment analysis on comment sections
* K- means to group talks into different categories
* Linear regression and random forests to analyse which factors impact ratings the most.
* SVM to classify a suggestion and a video as a success and a failure (?)