# *Analysis of Trending Youtube Videos*

Daniel Leung, Princeton University

## Introduction

- Youtube is a lucrative business, being able to make a video trending would be very useful
- Kaggle Dataset: Likes, views, trending date, categories, thumbnails for 40,950 videos in 2017 and 2018

## Related Work

- **Thumbnails influence clickthrough rate (views)**
  - Could we generate or figure out which thumbnails could increase our views?
- **Trending words change over time**
  - Could there be a correlation between video titles and descriptions and the date they become trending?
- **Subscriber count affects if a video becomes trending**
  - Do less popular Youtubers use different words to make their videos trending?

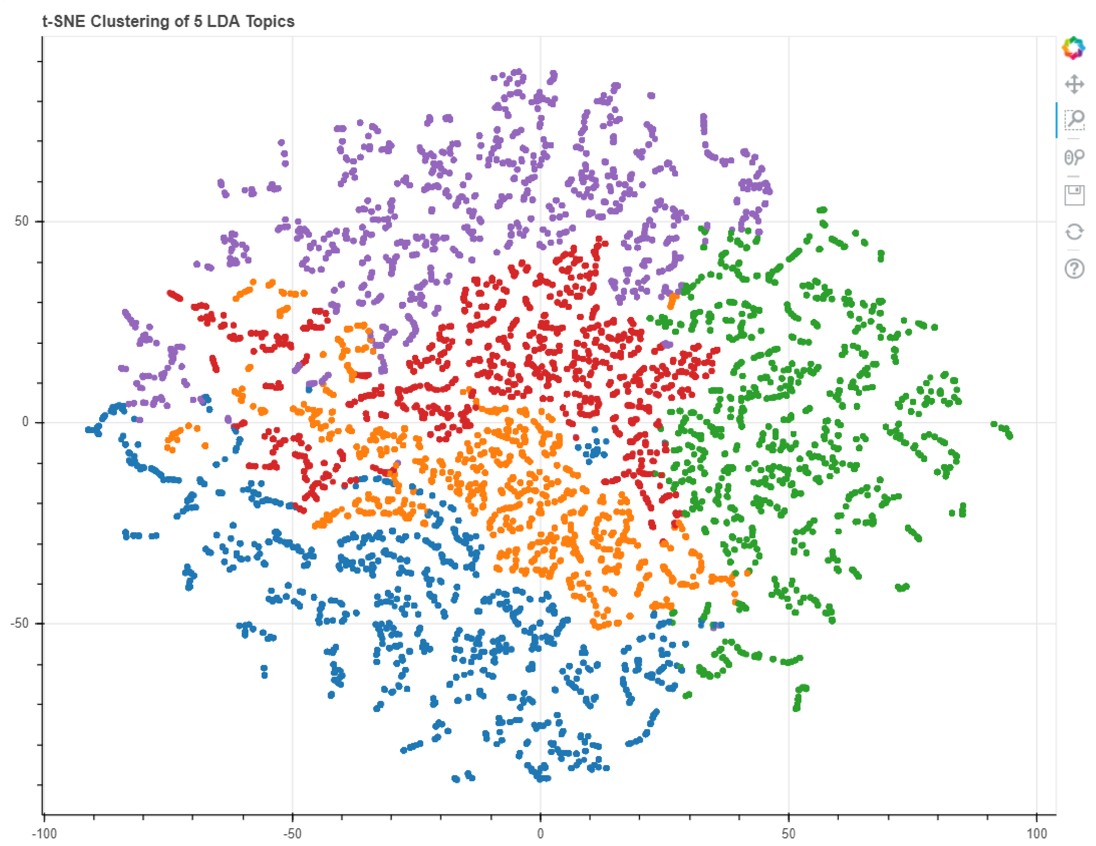
## Methods

- **Latent Variable Modeling**
  - Latent Dirichlet Allocation (5 and 10 components)
  - Principal Component Analysis (5 and 10 components)
  - T-distributed Stochastic Neighbor Embedding
- **Prediction**
  - Linear Regression (L2 penalty)
  - Ridge Regression
  - Ada Boost (50 estimators)
- **Generation**
  - Generative Adversarial Network (100 noise, 64 batch)

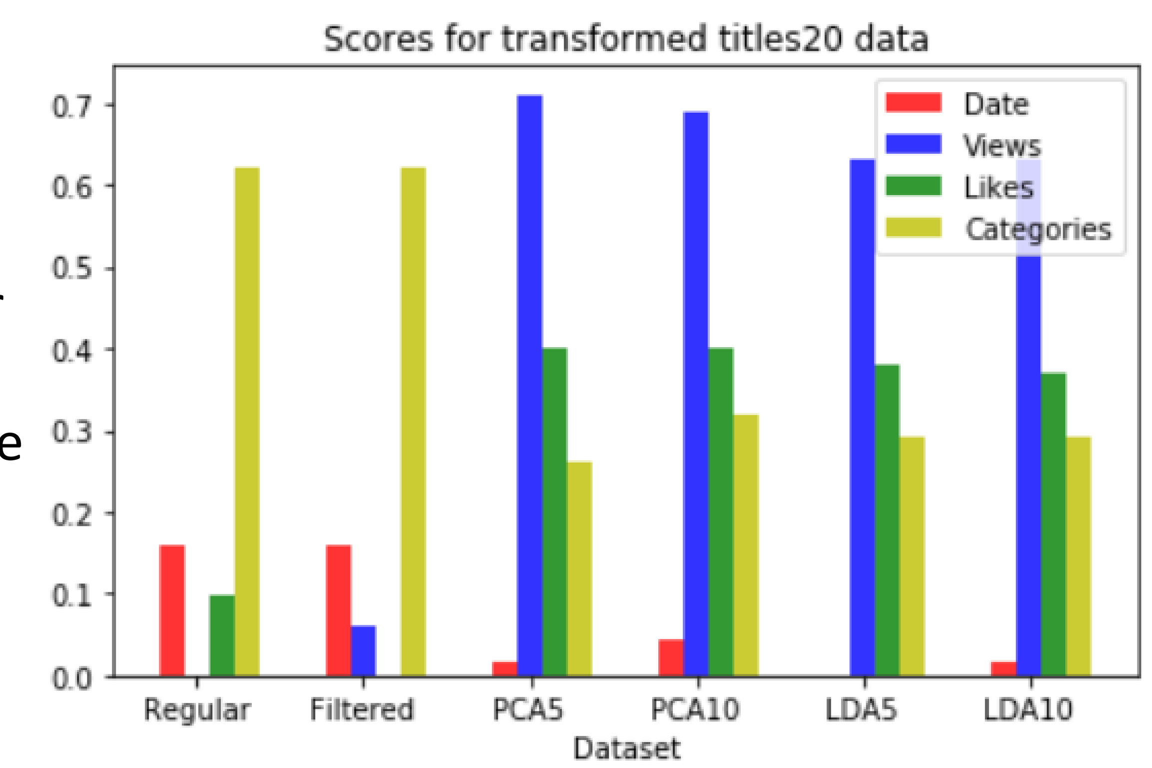
## Latent Structure

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Latent Topic 1 | live | tri | cake | one | ' | make | " | v | " | ft |
| Latent Topic 2 | offici | video | trailer | hd | audio | music | lyric | ft | 2 | first |
| Latent Topic 3 | 2018 | makeup | super | ' | bowl | v | life | full | challeng | commerci |
| Latent Topic 4 | 2017 | 2018 | new | star | last | ' | best | award | show | war |
| Latent Topic 5 | make | day | $ | test | food | new | 5 | break | time | 2018 |

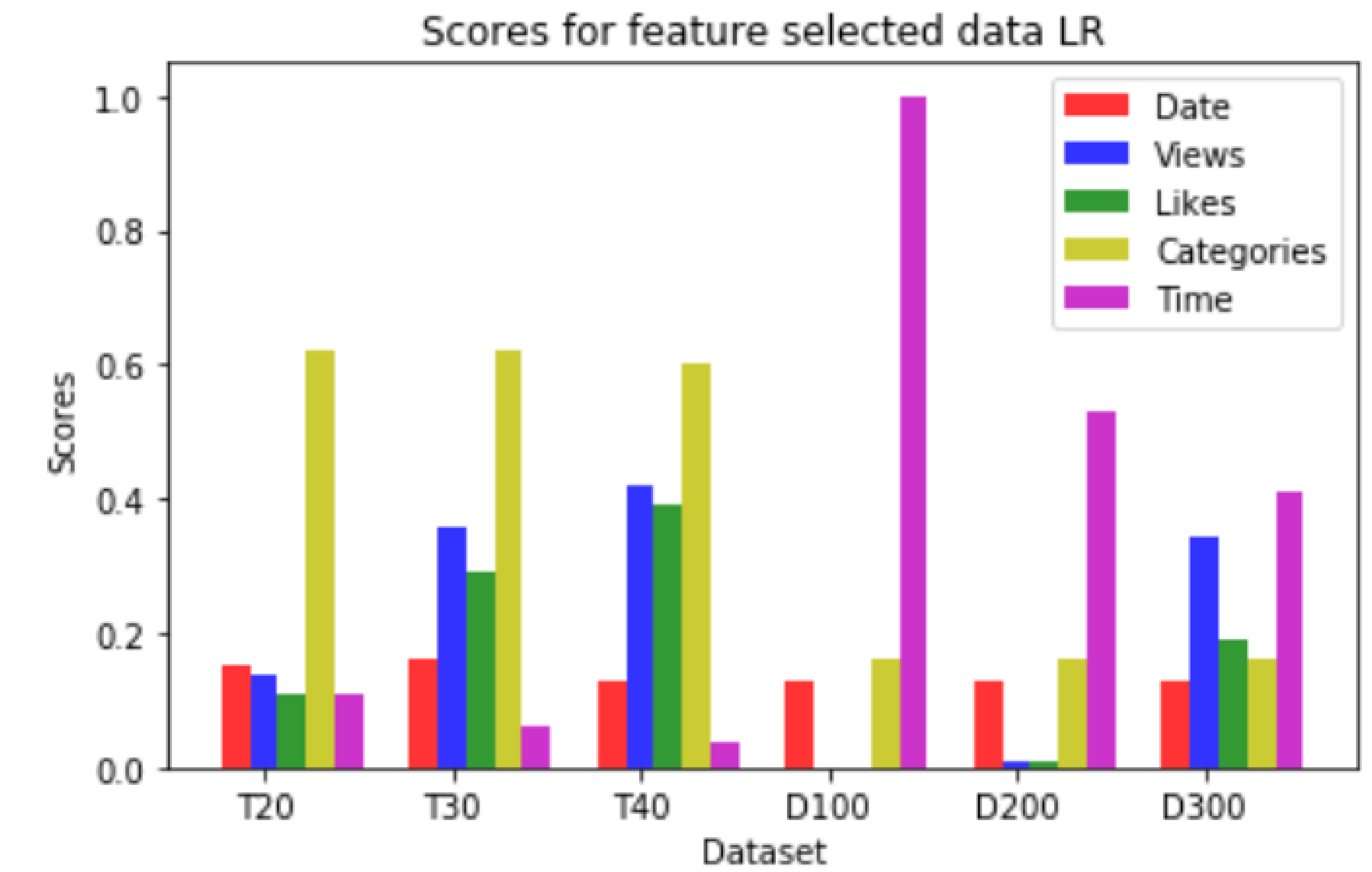
- In the table, we can see the top words for each latent topic
- The plot on the right colors each video with its corresponding latent topic, and then applies TSNE on the data for 2 dimensions to see how well LDA did




- The bar graph shows how performance for views and likes increases with feature selected data, however prediction for dates and categories drop



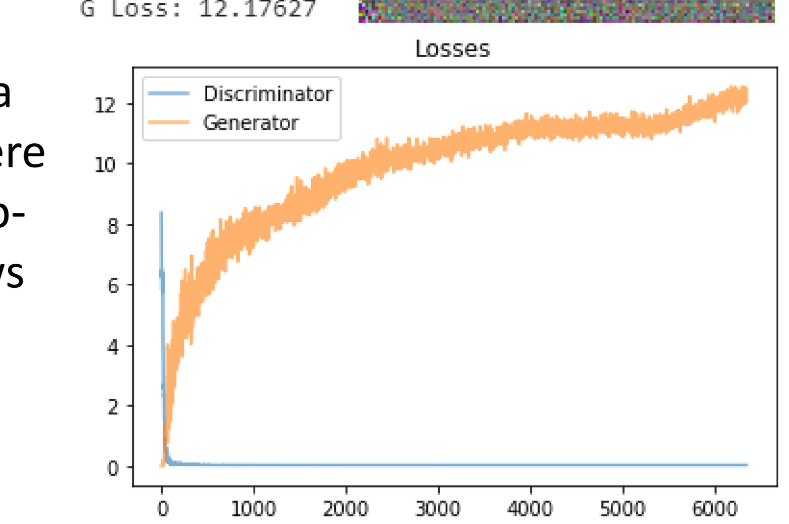

## Feature Selection




- We performed feature selection by increasing the count threshold for our bag-of-worlds model. We see that increasing the threshold improves results for titles and descriptions

## Thumbnails

- We used a Generative Adversarial Network on the thumbnail images of videos in the entertainment category
- We see that the generated image does not look very promising, the chart also shows a high generation loss and low discriminator loss
- This means our thumbnail data is still too diverse such that there is not a common format thumbnail that guarantees more views


## Prediction

- Top weighted words for select topics for each dataset

| | | | | | |
|---|---|---|---|---|---|
| | titles20 | | | | |
| Music | mv | billboard | bjork | sampl | chainsmok |
| Sports | espn | gopro | wwe | nba | candid |
| Entertainment | wwhl | ellen | choreographi | versu | bachelor |
| Science & Tech | mission | numberphil | tech | smartphon | smarter |
| | titles20 (filtered) | | | | |
| Music | mv | billboard | bjork | chainsmok | audio |
| Sports | espn | gopro | wwe | candid | nba |
| Entertainment | wwhl | babish | choreographi | versu | snl |
| Science & Tech | mission | numberphil | tech | smartphon | smarter |
| | descriptions300 | | | | |
| Music | coconut | festiv | station | hulkbust | luci |
| Sports | keep | asmr | health | easi | foot |
| Entertainment | half | kimmel | blind | fluffi | jona |
| Science & Tech | roy | celeb | jenner | comput | hope |

- We see that prediction results on titles data performs much better and different models work better for each metric

| | Titles | | | | Description | | | |
|---|---|---|---|---|---|---|---|---|
| | Date | Views | Likes | Category | Date | Views | Likes | Category |
| LR | **0.16** | 2,928,258 | 0.070 | **0.63** | 0.13 | 12,262,134 | 0.38 | 0.16 |
| ADA | 0.050 | **1,377,108** | 0.19 | 0.35 | 0.077 | 1,853,795 | 0.13 | 0.26 |
| Ridge | 0.160 | 2,087,068 | **0.052** | 0.60 | 0.13 | 3,394,866 | 0.091 | 0.16 |

- Other metrics for predicting select categories of each video

| | Titles | | | Descriptions | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Music | 0.84 | 0.79 | 0.81 | 0.20 | 0.19 | 0.19 |
| Sports | 0.79 | 0.84 | 0.81 | 0.043 | 0.024 | 0.031 |
| Entertainment | 0.60 | 0.64 | 0.62 | 0.26 | 0.37 | 0.31 |
| Science & Tech | 0.60 | 0.63 | 0.61 | 0.045 | 0.020 | 0.028 |

- Top weighted words for predicting views, likes, and dislikes

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Views | rewind | offici | infin | lovato | maluma | delic | shape | trap | la | twice |
| Likes | sidemen | closer | app | glynn | hustl | speechless | keynot | stirl | span | minnesota |
| Dislikes | domest | utah | c | nra | bbq | zombi | fergi | access | net | jess |

## Previous Channel Success

- We filtered our dataset to include only the top 5000 youtubers and appended subscriber count and total channel video view count to each bag of words data point. This would predict views and likes considering the text and the channel's subscriber and view count
- We see that this improves our prediction on likes while views prediction does not improve

| | Views | Likes |
|---|---|---|
| LR | 4023199 | 0.062 |
| ADA | 18826038 | 0.19 |
| Ridge | 2390284 | 0.039 |