CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

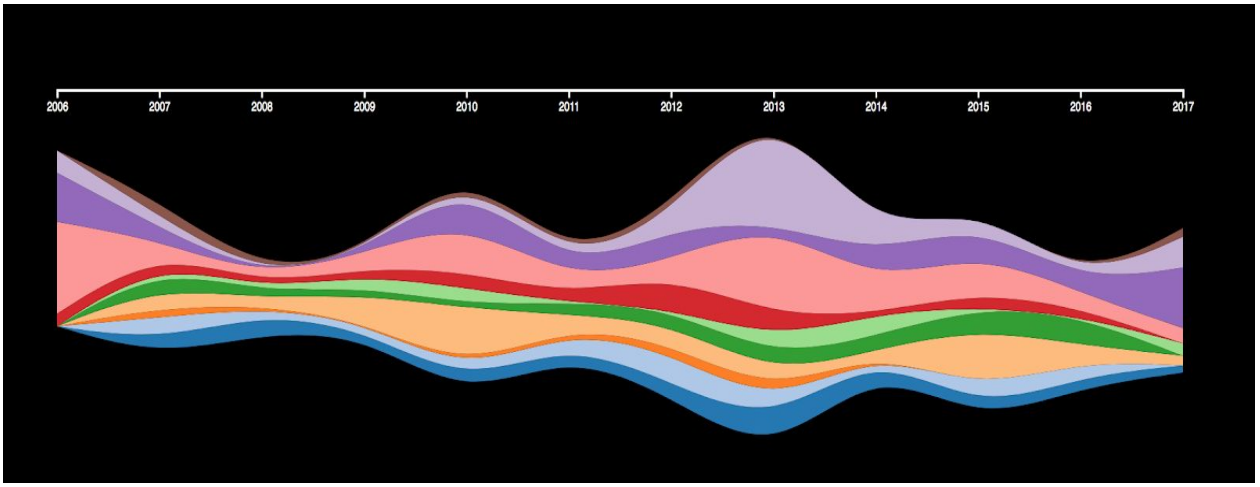
Process Book

May 16, 2018

CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL



Project Name: tedtalktimeline

TED started as an annual, invite-only conference in 1984, but has since morphed into a global brand. Since 2006, TED.com has become a hub for thought-provoking ideas delivered by some of the world's brightest minds. With more than 2,500 talks recorded since its inception, the TED library continues to grow. Topics range from the future of the internet, to living stress-free, to even a love poem for lonely prime numbers.

Github Link:

https://github.com/tedtalktimeline/tedtalktimeline.github.io.git

CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

Table of contents

tedtalktimeline.github.io GAURI JOSHI, NEHA BANDAL

Initial Project Proposal

Summary

TED is a media organization which posts talks online for free distribution, under the slogan "ideas worth spreading".TED was founded in February 1984 as a conference, which has been held annually since 1990.There are hundreds of different topics included in TED.

Aim of this project is to present different ted topics in user friendly manner, so that it will be easy for users to select topic and speaker quickly and easily and then watch video for that topic. For that major task was to organize ted talk data in efficient and effective manner. So the first step was data modeling, and linking different visualizations together.

Main objective of this project is show the timeline of different Ted Talks based on different categories and number of views.

Questions that can be answered from this project are:

1. Which are the most viewed talks in which year of all time? What does this tell us?
2. What kind of topics attract the maximum discussion and debate (in the form of comments)?
3. Diversity of speakers(Different category of individuals sharing their related or unrelated experience)?
4. Show related videos for each tedtalk
5. Top 10 Ted Talks

Problems

We ran into fairly significant problems in finalizing visualizations. There were few different combinations and ideas but due to some data restrictions we finalized the one which we thought more user friendly and efficient. The databases we were relying on proved to be sparsely populated, and reliable data was not available in any of the sources we turned to.After much discussion and collecting data for this initial proposal, we decided that the difficulties in locating data were large enough to merit a switch in the direction we were taking our project in.

CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

Updated Project Proposal

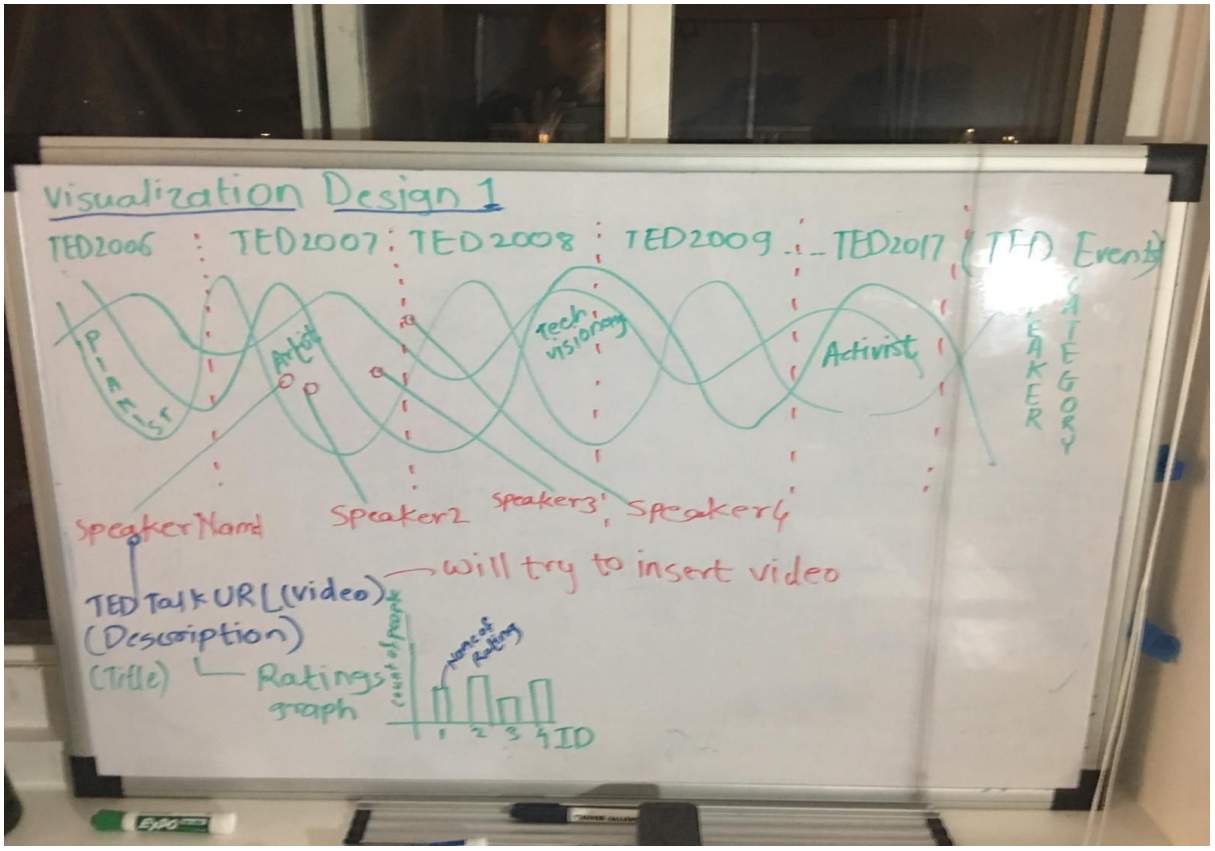
Design process

Finalized visualization:

After thinking more and more and trying different combinations of visualization we finalized below combination of visualizations.
Specifically, we plan on looking at metrics such as Ted topics, Ted Speakers, and Rating data for that ted talk, and analyzing this data for things such as related ted talks. We will obtain the data by running various Processing and Python libraries.
We have been able to acquire the necessary data and have started the coding process.

Data representation sketches:

Design 1:

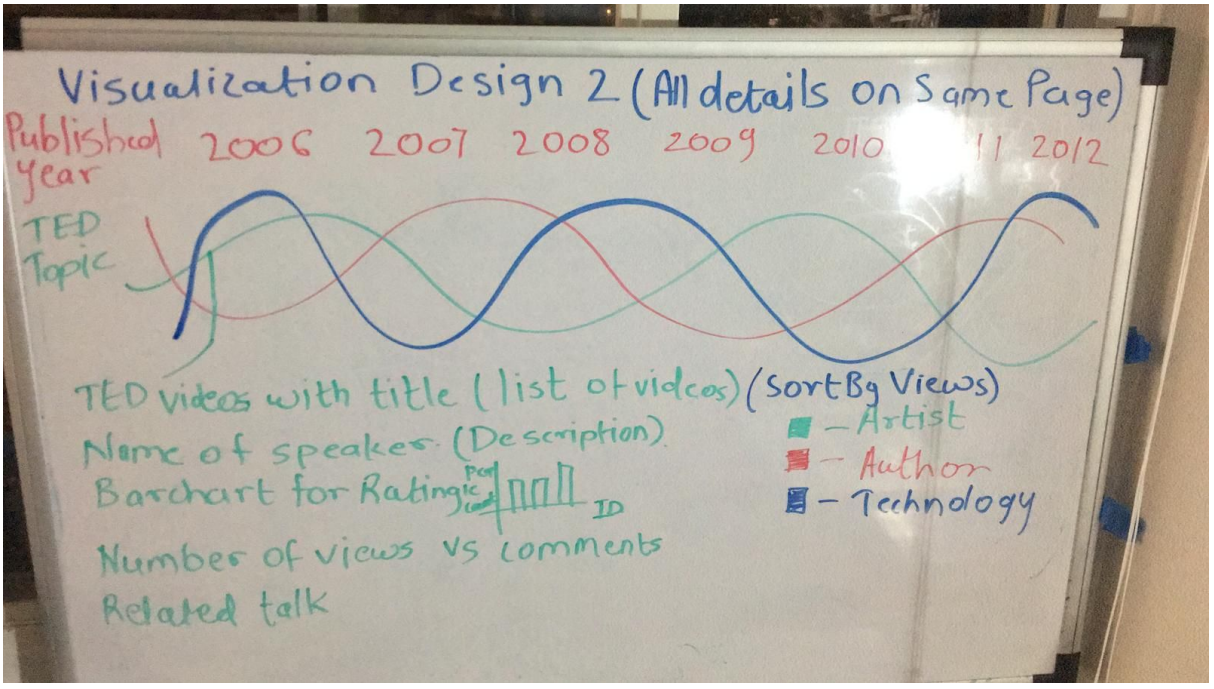


CS 560: FINAL PROJECT

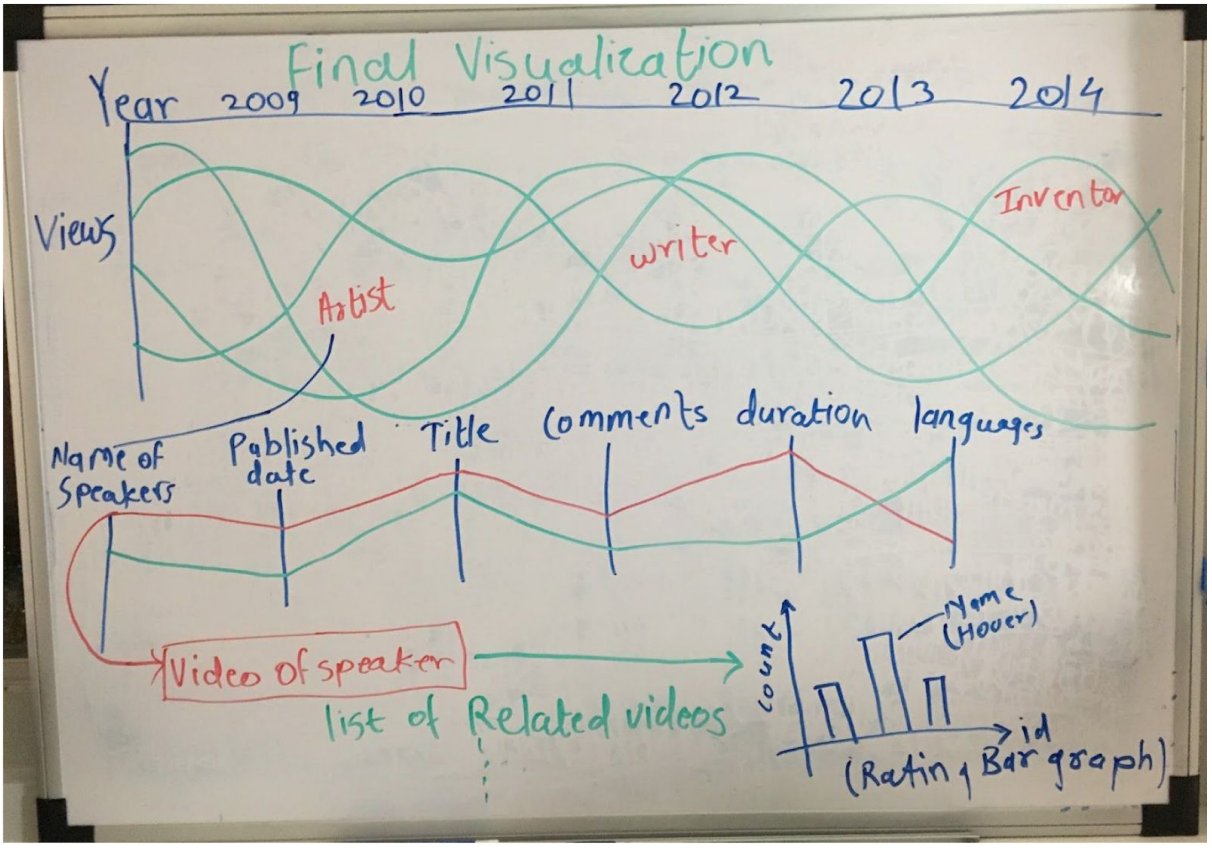
tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

Design 2:



Design 3:



CS 560: FINAL PROJECT

tedtalktimeline.github.io GAURI JOSHI, NEHA BANDAL

Process Book

Introduction

Features:

1. Visualization with timeline to show different categories of Ted Talks or occupations of authors. This is primary objective of our project. We are planning to implement Sinusoidal wave to show Ted Talks Timeline (X-Axis - Year/Event and Y-Axis - Number of Views for different categories or occupations of authors). Most of the objectives mentioned above will be covered in this visualization like most viewed talks of all time, diversity of speakers.
2. When a user clicks/hover on any particular category, the parallel coordinate will be shown with different attributes like published date, the name of the speaker/s, title, duration, comments, etc. revealing information about the particular category of Ted Talk. We are unsure about how things will work out for few attributes (Title, Speaker) on parallel coordinates. We may refer this link for parallel coordinate graph. http://bl.ocks.org/syntagmatic/4020926
3. Clicking/hovering on particular line from parallel plot will give details about that particular Speaker along with link to his/her Ted Talk videos. This feature can act as a filter for speakers.
4. We are also plot visualization to show related talks to particular title in parallel plot. This feature will cover the 'How is each TED Talk related to every other TED Talk?' objective along with filter for Ted Talk.

Data:

The Ted Talks Timeline is based on talks subject and speaker statistics aggregated from
https://www.kaggle.com/rounakbanik/ted-talks.
These datasets contain information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. It contains information about all talks including number of views, number of comments, descriptions, speakers and titles. If require we will find more relevant data.

Data Processing:

Based on our initial analysis there are different columns which requires cleanup and processing. For e.g. Film Date and Published Date are UNIX timestamps we will need to convert it into Javascript/D3 Date format. Also we will also require to normalize data

to some extent as few columns consist of JSON objects. For e.g. ratings, related talks etc. Also duration column is in seconds and we converted it in minutes.

Tasks:

Before we jump into the tasks, we should really preface them. With all the data that we have, we chose to break the analysis down into four main components: Ted topics analysis, Speakers analysis, Rating data analysis/comparison, and related videos analysis.

Analysis task:

With this visualization, we wanted to show different ted topics *(Architect, Artist, Designer, Engineer, Entrepreneur, Inventor, Musician, Photographer, Writer, Actor, Activist, Physicist)* with increase and decrease in views over years (2006-2017). While selecting the ted topic, the user can see how the data including Speakers,videos and ratings change with time, see a breakdown of the related videos, as well as see the end result of the speakers and rating data analysis.

After selecting topic we added parallel coordinate to visualize all the data for selected ted topic, it includes Year, Speakers, Title, Views, Duration, Comments, Languages. The user can therefore see, compare, and contrast the variations in duration of Ted talks, languages ted talk is published across 12 different ted talk topics.

Related Work:

We referred below visualizations while working on our project.

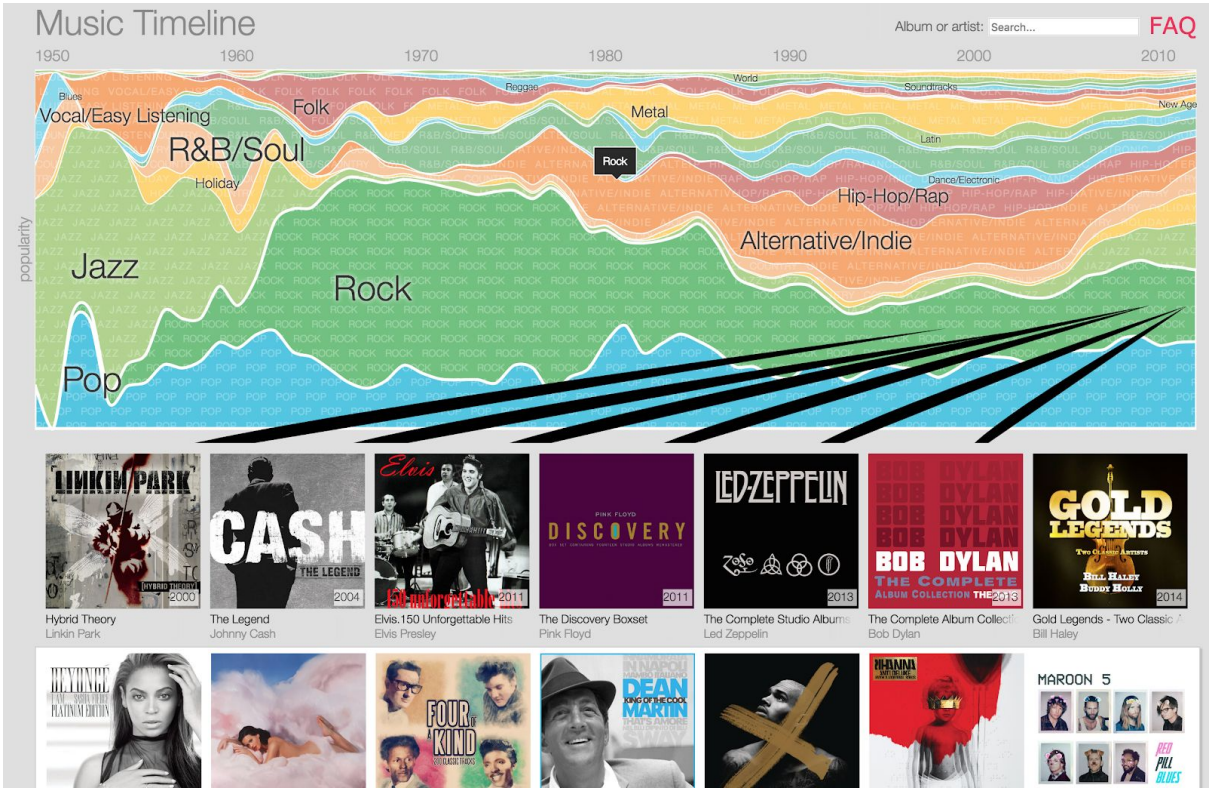
1. https://www.kaggle.com/rounakbanik/ted-data-analysis
2. https://research.google.com/bigpicture/music/
3. https://mef-bda503.github.io/pj-sevgilit/files/TED_Talks.html
4. https://www.kaggle.com/lpang36/analysis-of-ted-talk-ratings

CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

Reference Design:



Process

Data collection:

Choosing the Ted topic, video information: We selected 12 different ted topics (Architect, Artist, Designer, Engineer, Entrepreneur, Inventor, Musician, Photographer, Writer, Actor, Activist, Physicist) and included all speakers who gave ted talks on these topics. Major task is to manually select most viewed topics as streamgraph won't look good with more than certain numbers.

In dataset we got many speakers associated with one topic and different speakers have different ted videos which gave us idea of parallel coordinate. While creating parallel coordinate major task was selecting quantitative data as text won't fit in your parallel coordinate, at the same time we need to show title and speaker of selected ted talk so we created textbox which includes all qualitative as well as quantitative data.

Another task was filtering rating data. Rating data column was in the form of json data so it was required to flat that column to get different key(rating ID) and value(Rating name) and count associated with that.

Data analysis:

In this project we are focusing on Overview+Detail, Zooming, and Focus+Context Interfaces concepts. Along with these features we will have Interaction, Great storytelling(including effective use of colors, animation, annotation/ labels, depth in layers, simplicity and consistency).

In the process of data analysis major task was to transform ted talk video data which is url into embedded video.

The goal was to discover useful information just by visualizing short but important amount of data and then details on demand in next visualizations which will suggest conclusion ultimately.

Data modeling:

1. Modeled data for different visualization for streamgraph modeled ted talk categories data by selecting few categories.
2. For parallel coordinates modeled data with multiple speakers and all quantitative data columns. Like views, comments, duration, languages, speaker name, title of ted talk etc,
3. Finally we are showing video and rating graph for selected ted talk.

CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

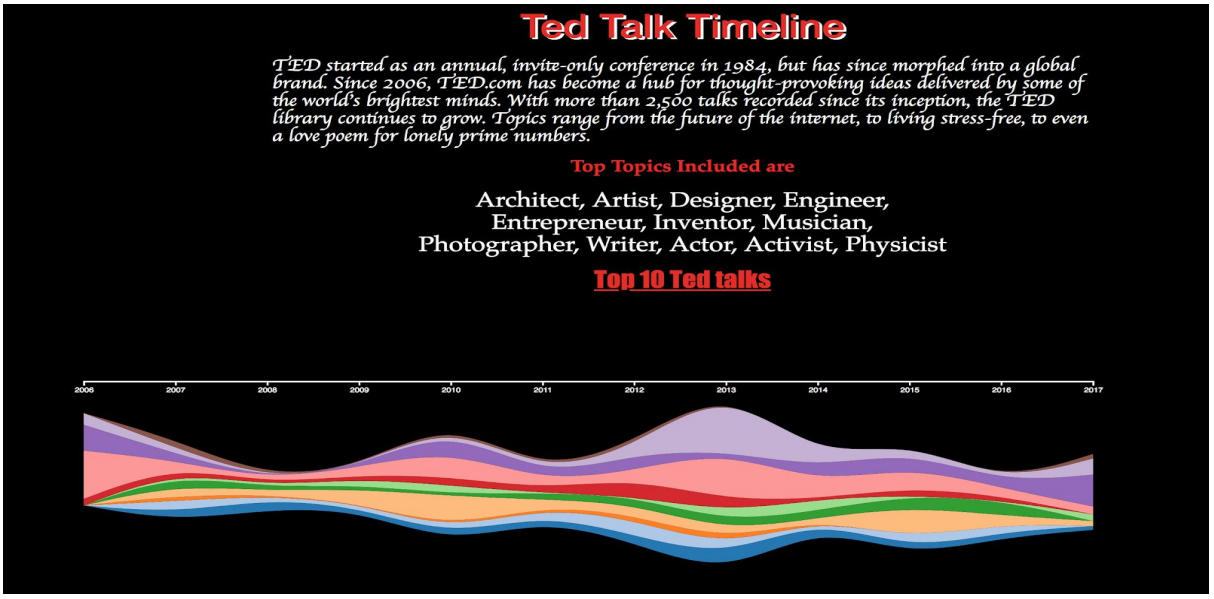
Timeline building

Visualization patterns

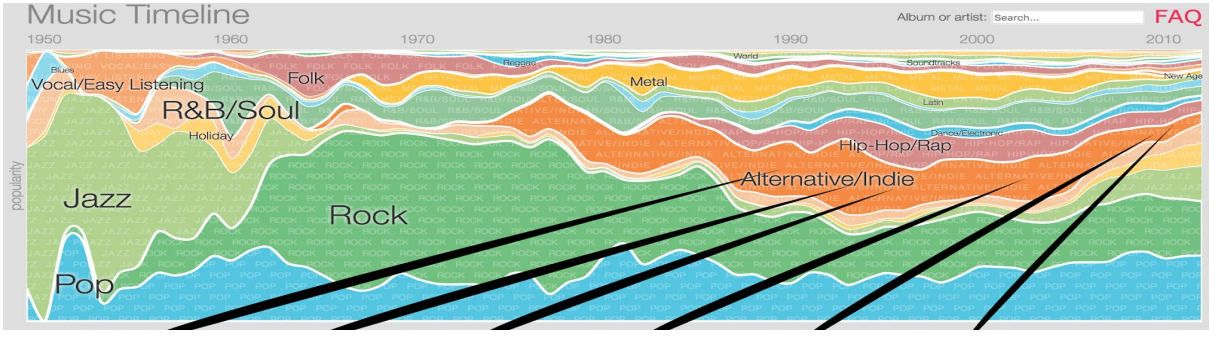
Streamgraph:

After discussing on different visualizations we finalized to start design with streamgraph to show timeline. Streamgraph is used to show increase and decrease in number of views for different topics over years.
We got data from 2006 till 2017, so we decided to visualize timeline from 2006 to 2017 which is a major part of website.

Different topics:
Architect, Artist, Designer, Engineer, Entrepreneur, Inventor, Musician, Photographer, Writer, Actor, Activist, Physicist



Reference: https://research.google.com/bigpicture/music/



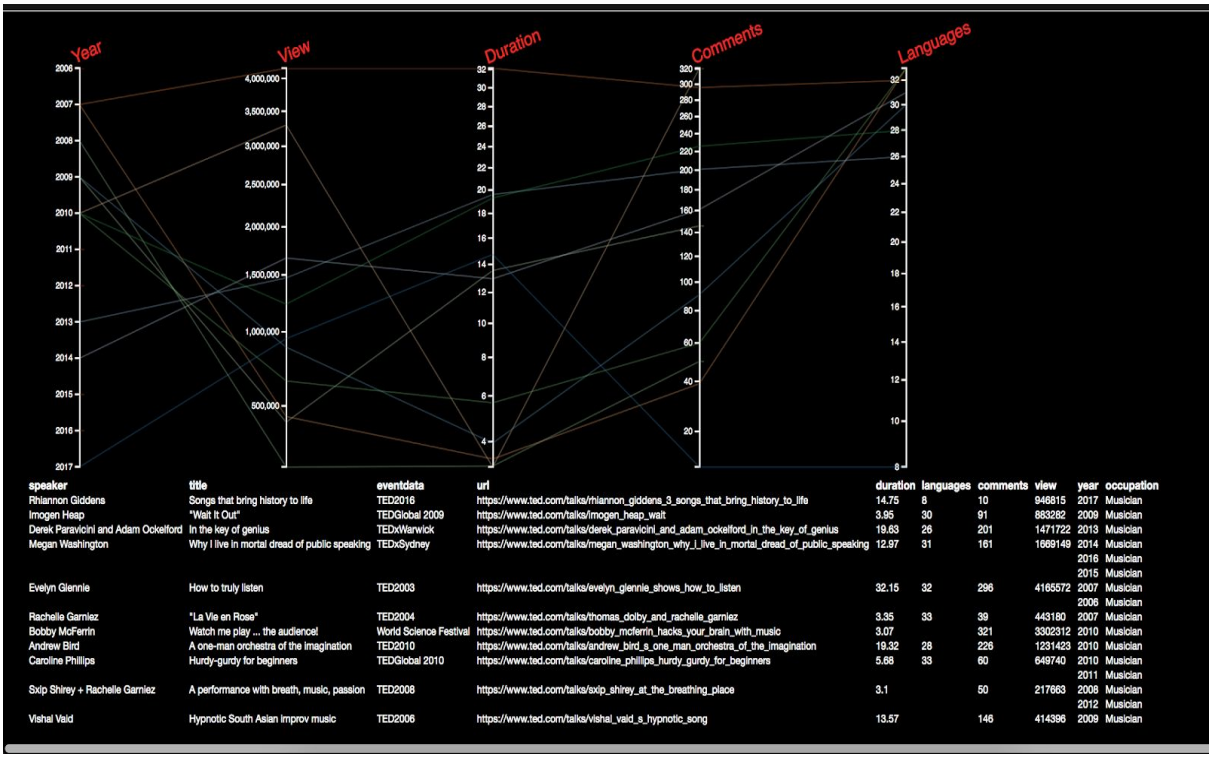
CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

Parallel Coordinates

After selecting ted topic we collect data for that ted topic and generate parallel coordinate for the same with both quantitative and qualitative values.
It includes year, comments, languages, views and duration and then in description we are adding speaker name, title to see details about ted talk.
Here we have added brushing feature in parallel coordinates to select as per requirements. So you can brush over any y axis and select data, this will update data in description.



speaker	title	eventdata	url	duration	languages	comments	view	year	occupation
Rhiannon Giddens	Songs that bring history to life	TED2016	https://www.ted.com/talks/rhiannon_giddens_3_songs_that_bring_history_to_life	14.75	8	10	946815	2017	Musician
Imogen Heap	"Wait It Out"	TEDGlobal 2009	https://www.ted.com/talks/imogen_heap_wait	3.95	30	91	883282	2009	Musician
Derek Paravicini and Adam Ockelford	In the key of genius	TEDxWarwick	https://www.ted.com/talks/derek_paravicini_and_adam_ockelford_in_the_key_of_genius	19.63	26	201	1471722	2013	Musician
Megan Washington	Why I live in mortal dread of public speaking	TEDxSydney	https://www.ted.com/talks/megan_washington_why_i_live_in_mortal_dread_of_public_speaking	12.97	31	161	1669149	2014	Musician
								2016	Musician
								2015	Musician
Evelyn Glennie	How to truly listen	TED2003	https://www.ted.com/talks/evelyn_glennie_shows_how_to_listen	32.15	32	296	4165572	2007	Musician
								2006	Musician
Rachelle Garniez	"La Vie en Rose"	TED2004	https://www.ted.com/talks/thomas_dolby_and_rachelle_garniez	3.35	33	39	443180	2007	Musician
Bobby McFerrin	Watch me play ... the audience!	World Science Festival	https://www.ted.com/talks/bobby_mcferrin_hacks_your_brain_with_music	3.07		321	3302312	2010	Musician
Andrew Bird	A one-man orchestra of the imagination	TED2010	https://www.ted.com/talks/andrew_bird_s_one_man_orchestra_of_the_imagination	19.32	28	226	1231423	2010	Musician
Caroline Phillips	Hurdy-gurdy for beginners	TEDGlobal 2010	https://www.ted.com/talks/caroline_phillips_hurdy_gurdy_for_beginners	5.68	33	60	649740	2010	Musician
								2011	Musician
Sxip Shirey + Rachelle Garniez	A performance with breath, music, passion	TED2008	https://www.ted.com/talks/sxip_shirey_at_the_breathing_place	3.1		50	217663	2008	Musician
								2012	Musician
Vishal Vaid	Hypnotic South Asian improv music	TED2006	https://www.ted.com/talks/vishal_vaid_s_hypnotic_song	13.57		146	414396	2009	Musician

CS 560: FINAL PROJECT

tedtalktimeline.github.io

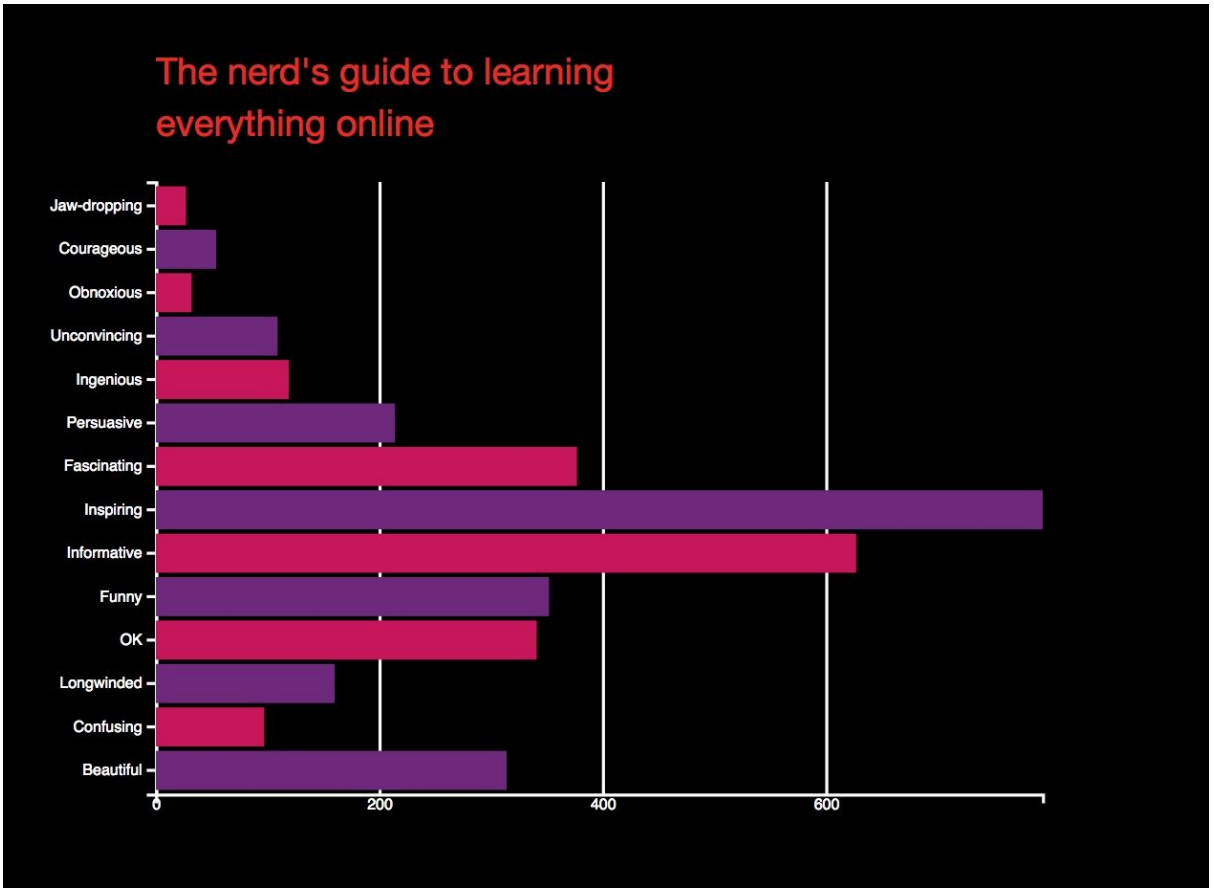
GAURI JOSHI, NEHA BANDAL

Bar chart:

This will give you rating data bar chart for ted video which will help user to understand what other users are thinking about selected ted talk like is it funny or serious etc., data we got has rating data column which is in json format.so we first flattened that column and then used each value to generate bar chart.

Example:

[{'id': 10, 'name': 'Inspiring', 'count': 957}, {'id': 22, 'name': 'Fascinating', 'count': 686}, {'id': 9, 'name': 'Ingenious', 'count': 1443}, {'id': 7, 'name': 'Funny', 'count': 325}, {'id': 23, 'name': 'Jaw-dropping', 'count': 420}, {'id': 8, 'name': 'Informative', 'count': 470}, {'id': 25, 'name': 'OK', 'count': 31}, {'id': 3, 'name': 'Courageous', 'count': 40}, {'id': 1, 'name': 'Beautiful', 'count': 53}, {'id': 24, 'name': 'Persuasive', 'count': 161}, {'id': 21, 'name': 'Unconvincing', 'count': 8}, {'id': 2, 'name': 'Confusing', 'count': 2}, {'id': 26, 'name': 'Obnoxious', 'count': 8}, {'id': 11, 'name': 'Longwinded', 'count': 4}]



CS 560: FINAL PROJECT

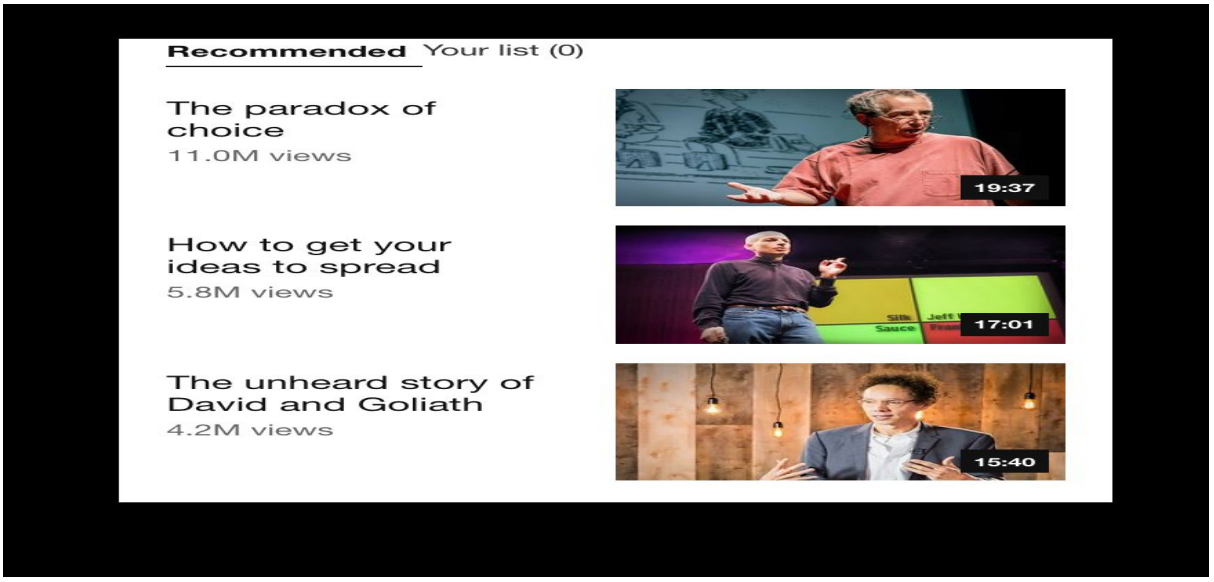
tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

Embedding Video:

Major and more attractive part of project is embedding ted video on website, it was the most complex task as data we got for video is tedtalk url not youtube video url so we first needed to embed it and then link it on website. For this d3 has a attribute called **Sandbox** which we used to embed video and adding source url into it. Because of this video got linked into website so instead of clicking on link you can directly watch video here.

Below selected video user can see related videos of the running videos which was one of our objective question,



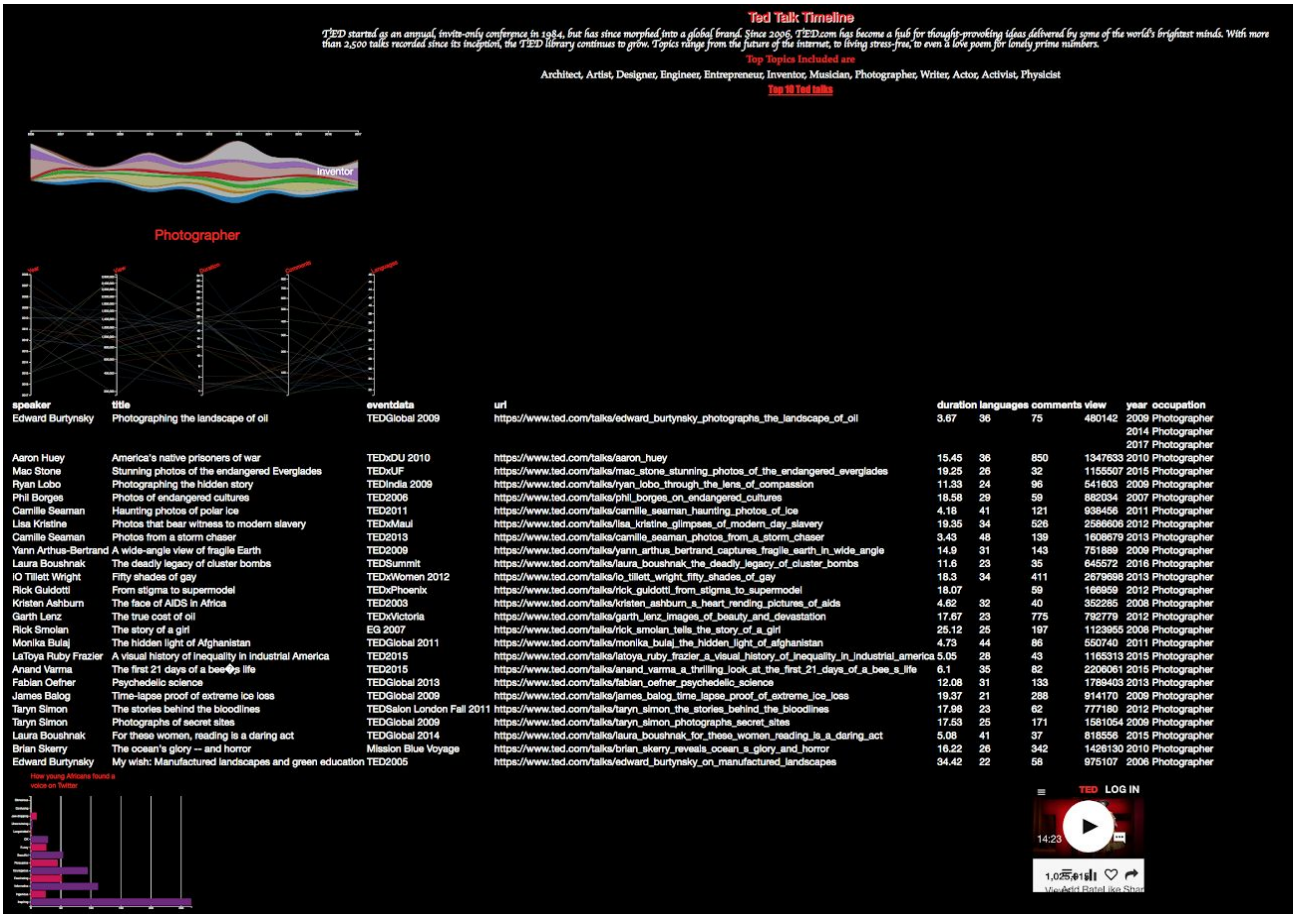
CS 560: FINAL PROJECT

tedtalktimeline.github.io

GAURI JOSHI, NEHA BANDAL

Dashboard:

If we see dashboard, it works on onclick events i.e selecting ted topic from **streamgraph** takes you to **parallel coordinate and description with rows** and clicking on each row of description will take you to **rating bar graph for ted video** and and **actual ted video**.



So on website you can see:

1. Streamgraph
2. Parallel Coordinate
3. Description with each row containing Speaker name, title, ted event, URL, duration, languages,comments, view, year and occupation.
4. Rating data bar chart
5. Video for ted talk

Challenges:

- Data cleaning and filtering and flat the rating json data.
- Trying to figure out way of embedding video into video as URL is tedtalk URL.
- Getting all visualizations on one page
- Onclick events

Logistics

Visualization patterns implementation were split evenly among team members and task of coding as well as documentation involved equal amount of efforts by all team members. In this project we followed agile programming model as all team members found it most efficient way to code. Constant discussions were involved whenever problems arose. Similarly all team members were working independently on their own tasks without any dependency on other team members.

Conclusion

The most obvious conclusion is user can find and watch ted videos quickly and efficiently with detailed and effective data.
We hope that everyone will enjoy this visualization as much as we had fun making it. Overall process of creating and linking all visualizations was difficult than we thought. Many things came during implementation and took us to new direction of implementation, particularly in terms of feasibly and design. When there were many interactive components like linking multiple visualization together, efficiency and design were trade-offs. It was the first time dealing with complex and huge data, especially video ones. Overall it was a great learning and work experience.