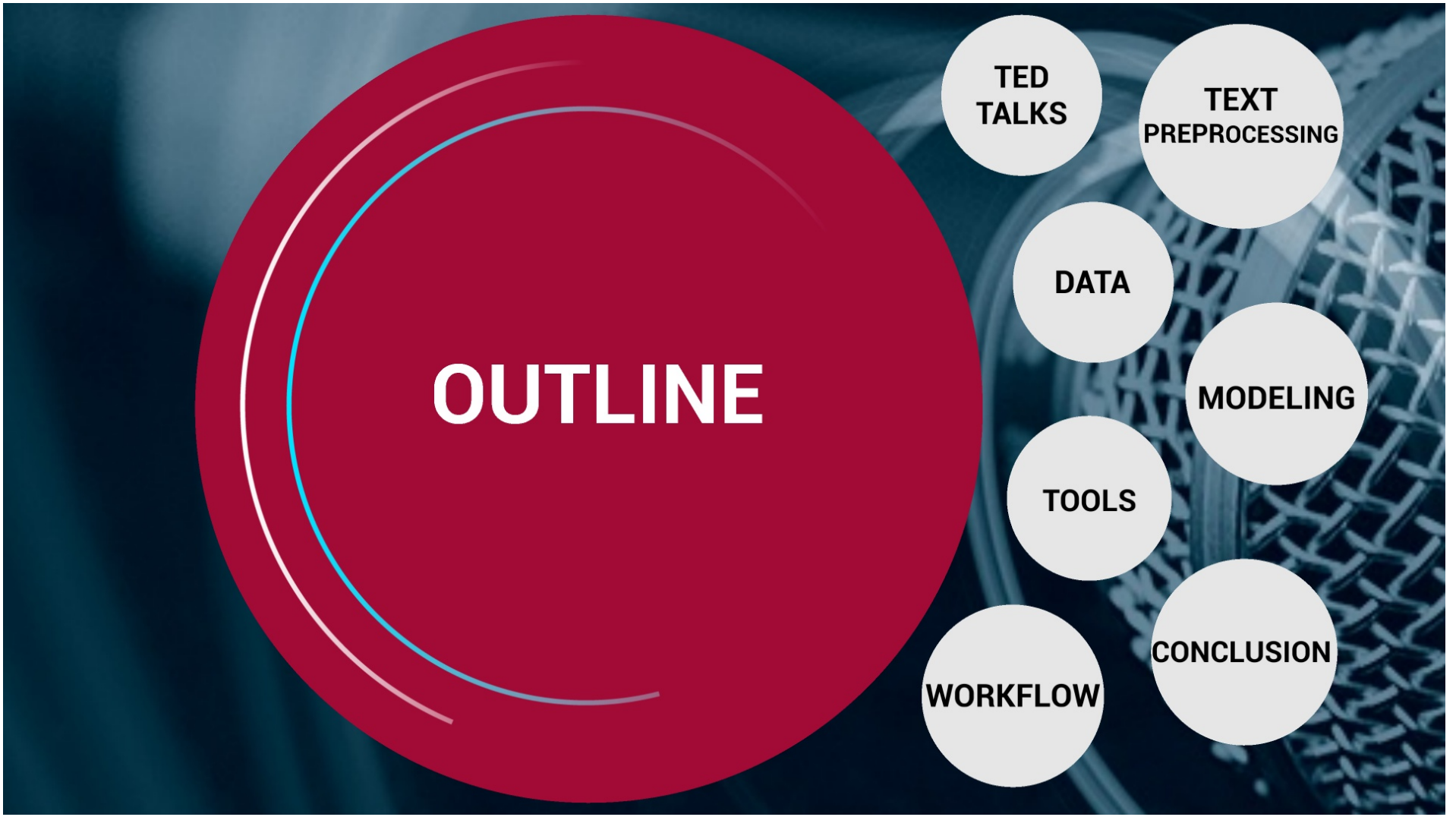
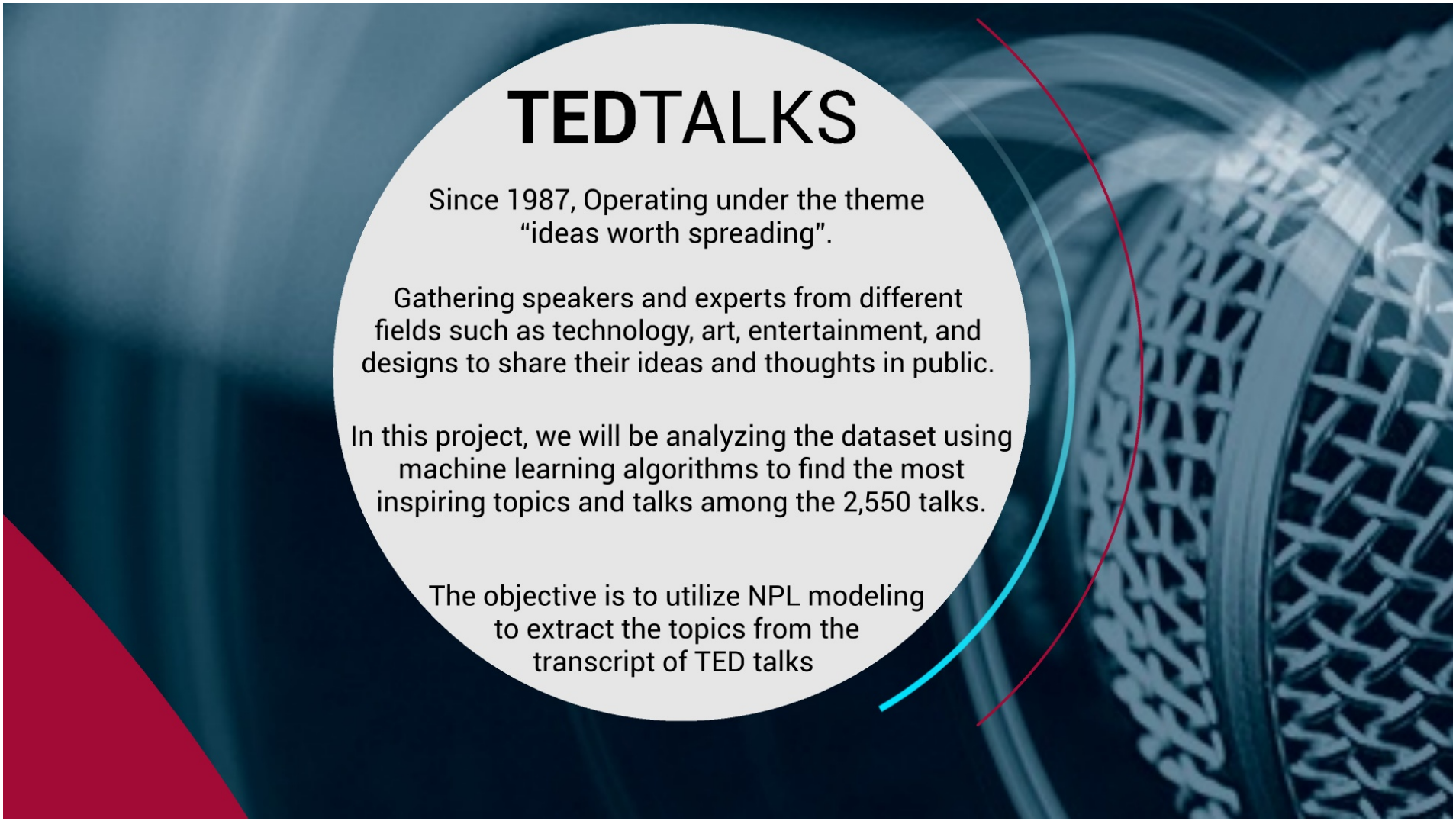
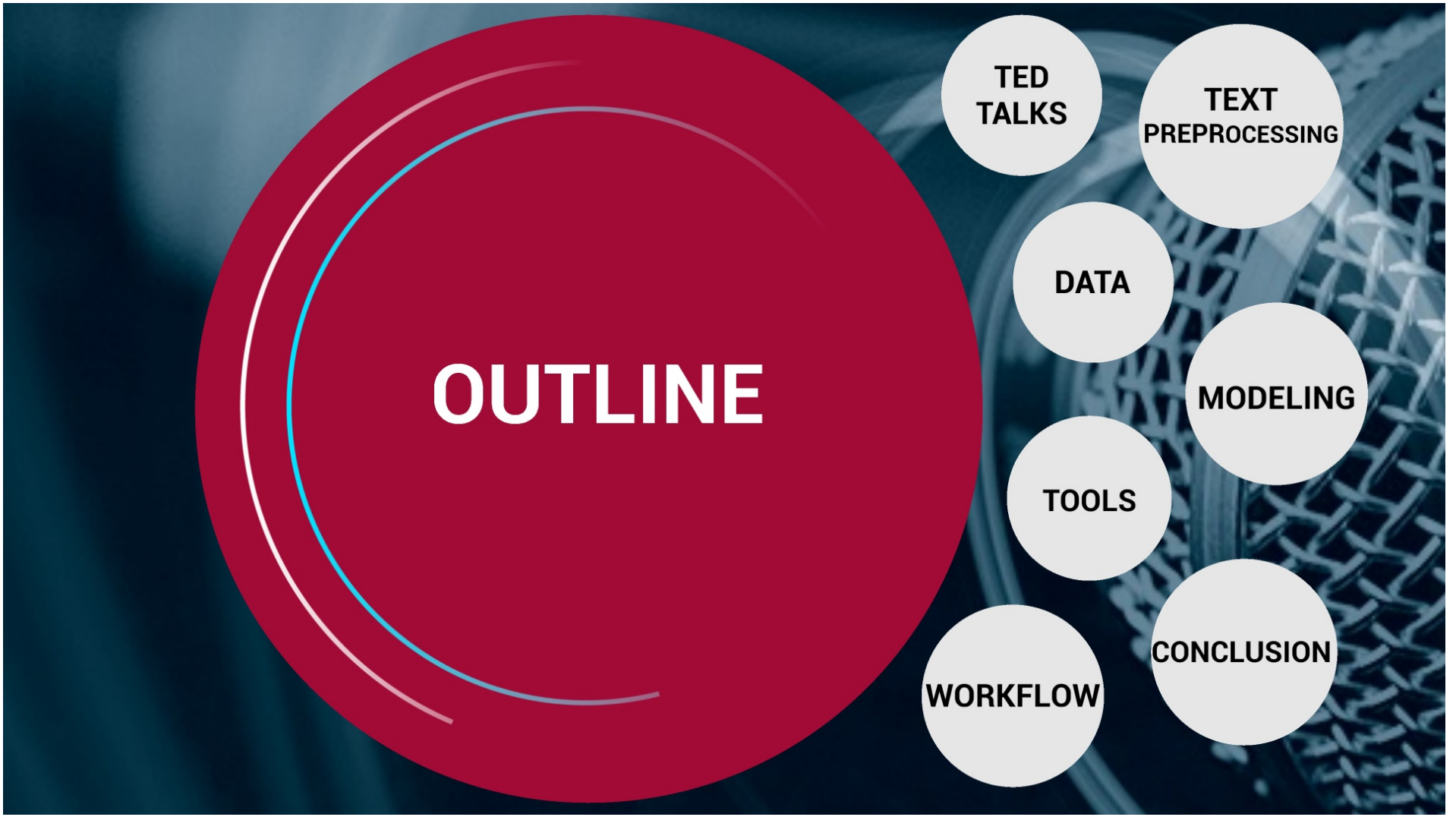
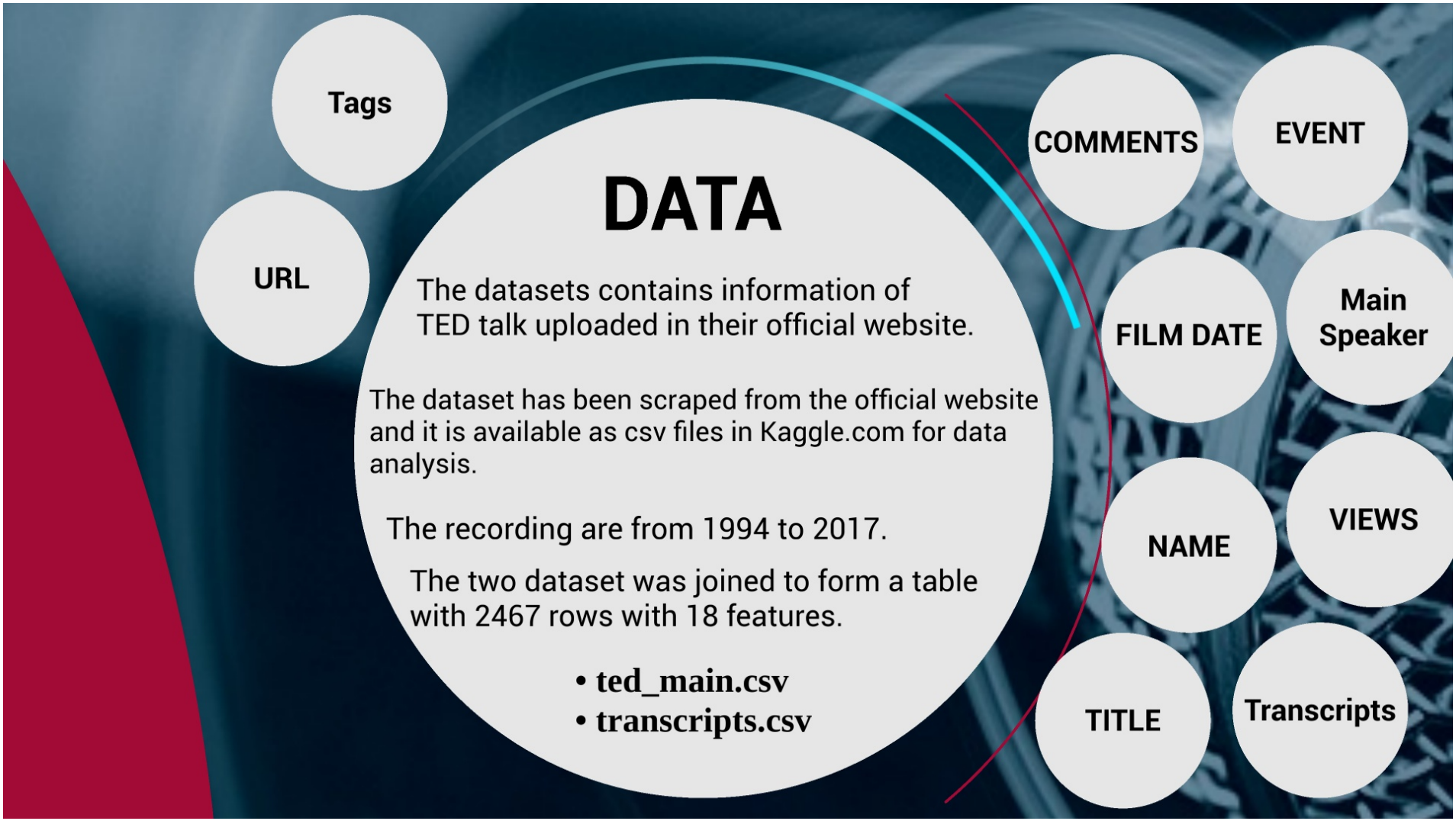
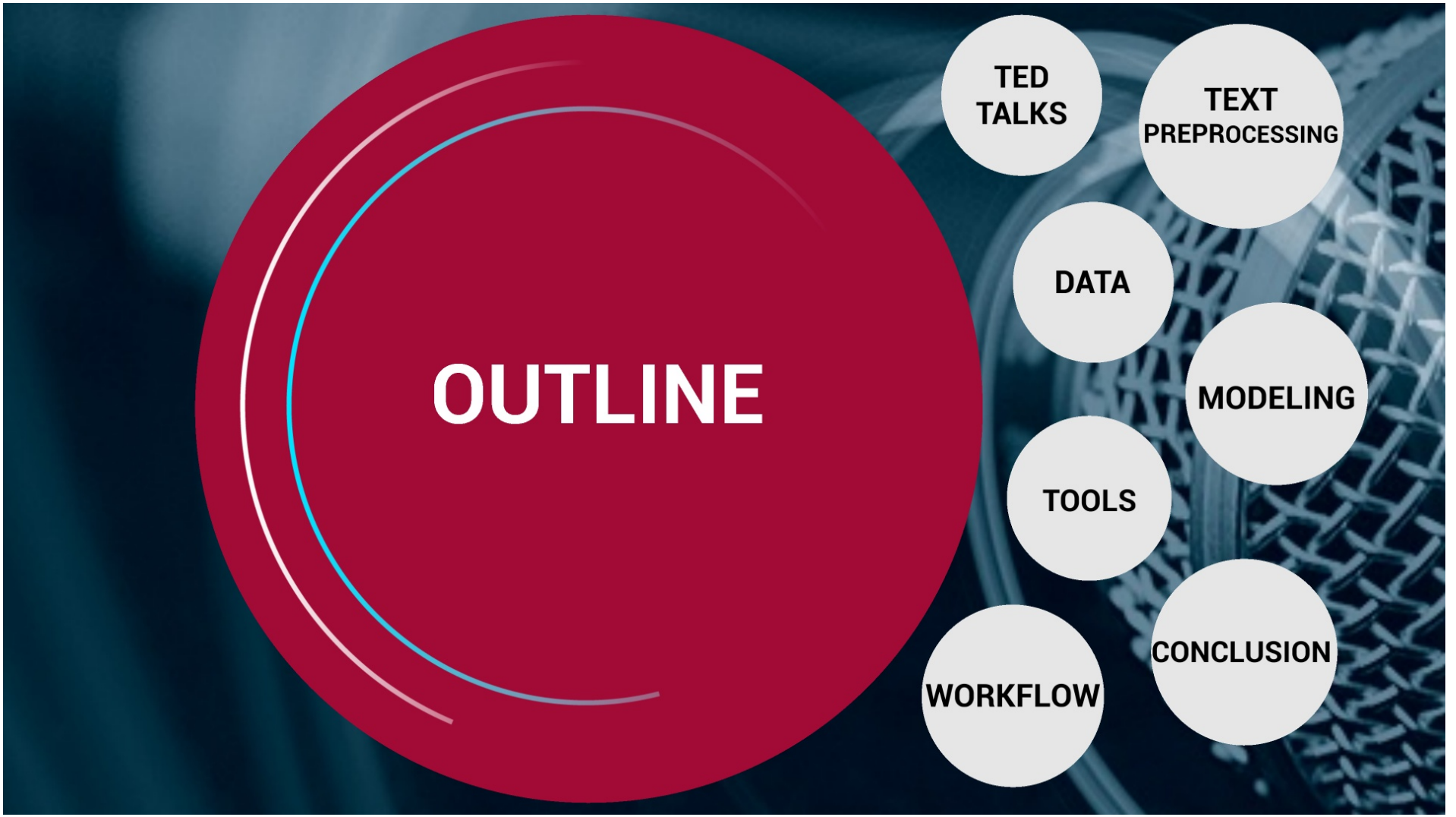

OUTLINE
WORKFLOW
VISUALIZATION
CONCLUSION
TED TALKS
Topic Modeling
Natural Language Processing
Hayat Aldhahri & Juri AlSayigh

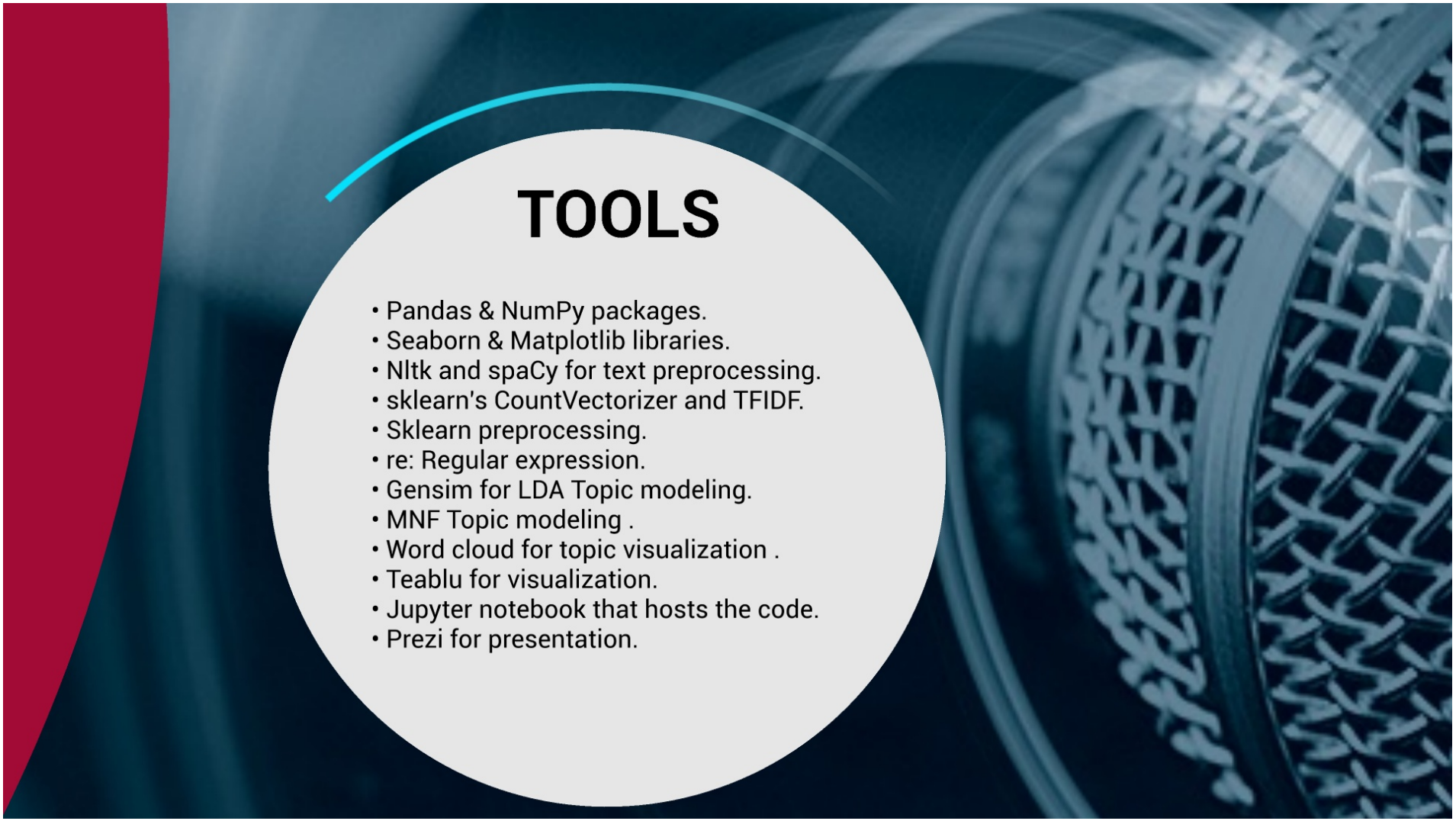

OUTLINE
TED TALKS
TEXT PREPROCESSING
DATA
MODELING
TOOLS
CONCLUSION
WORKFLOW

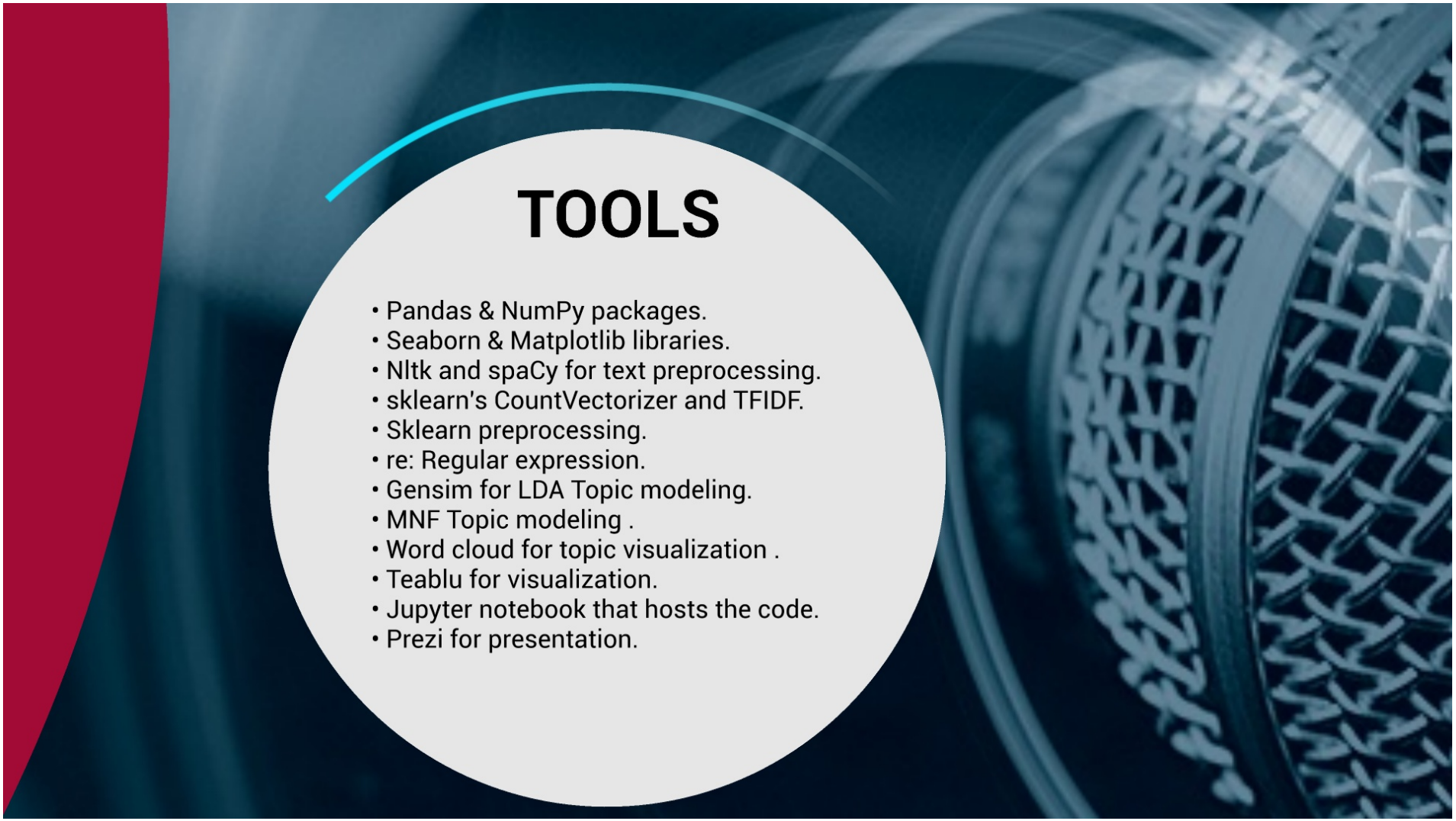

TEDTALKS
Since 1987, Operating under the theme "ideas worth spreading".
Gathering speakers and experts from different fields such as technology, art, entertainment, and designs to share their ideas and thoughts in public.
In this project, we will be analyzing the dataset using machine learning algorithms to find the most inspiring topics and talks among the 2,550 talks.
The objective is to utilize NPL modeling to extract the topics from the transcript of TED talks

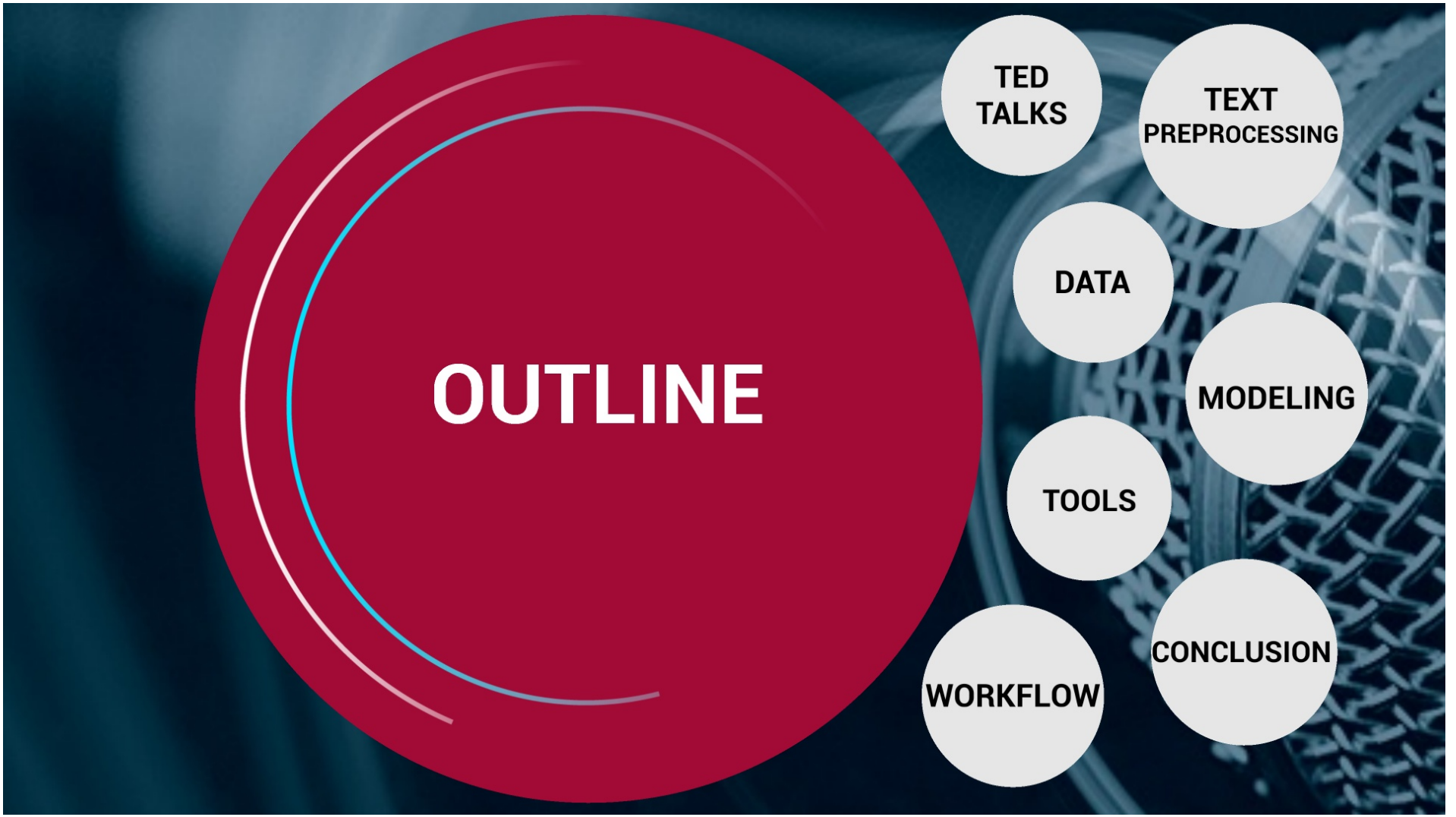

OUTLINE
TED TALKS
TEXT PREPROCESSING
DATA
MODELING
TOOLS
CONCLUSION
WORKFLOW


Tags
URL
DATA
The datasets contains information of TED talk uploaded in their official website.
The dataset has been scraped from the official website and it is available as csv files in Kaggle.com for data analysis.
The recording are from 1994 to 2017.
The two dataset was joined to form a table with 2467 rows with 18 features.
• ted_main.csv
• transcripts.csv
COMMENTS
EVENT
FILM DATE
Main Speaker
NAME
VIEWS
TITLE
Transcripts


OUTLINE
TED TALKS
TEXT PREPROCESSING
DATA
MODELING
TOOLS
CONCLUSION
WORKFLOW


TOOLS
• Pandas & NumPy packages.
• Seaborn & Matplotlib libraries.
• Nltk and spaCy for text preprocessing.
• sklearn's CountVectorizer and TFIDF.
• Sklearn preprocessing.
• re: Regular expression.
• Gensim for LDA Topic modeling.
• MNF Topic modeling .
• Word cloud for topic visualization .
• Teablu for visualization.
• Jupyter notebook that hosts the code.
• Prezi for presentation.


TOOLS
• Pandas & NumPy packages.
• Seaborn & Matplotlib libraries.
• Nltk and spaCy for text preprocessing.
• sklearn's CountVectorizer and TFIDF.
• Sklearn preprocessing.
• re: Regular expression.
• Gensim for LDA Topic modeling.
• MNF Topic modeling .
• Word cloud for topic visualization .
• Teablu for visualization.
• Jupyter notebook that hosts the code.
• Prezi for presentation.


OUTLINE
TED TALKS
TEXT PREPROCESSING
DATA
MODELING
TOOLS
CONCLUSION
WORKFLOW


OUTLINE
WORKFLOW
VISUALIZATION
CONCLUSION
TED TALKS
Topic Modeling
Natural Language Processing
Hayat Aldhahri & Juri AlSayigh

WORKFLOW

Click to edit text

WORKFLOW

Click to edit text

EDA

WORKFLOW

Click to edit text

TEXT PREPROCESSING

EDA

WORKFLOW

Click to edit text

TEXT PREPROCESSING

EDA

TOPIC MODELING

WORKFLOW

Click to edit text

TEXT PREPROCESSING

MODELING OUTPUT

EDA

TOPIC MODELING

WORKFLOW

Click to edit text

TEXT PREPROCESSING

MODELING OUTPUT

EDA

TOPIC MODELING

FINDINGS

EDA

Join the tables

ted_main.csv

2550 rows
17 columns

+

transcripts.csv

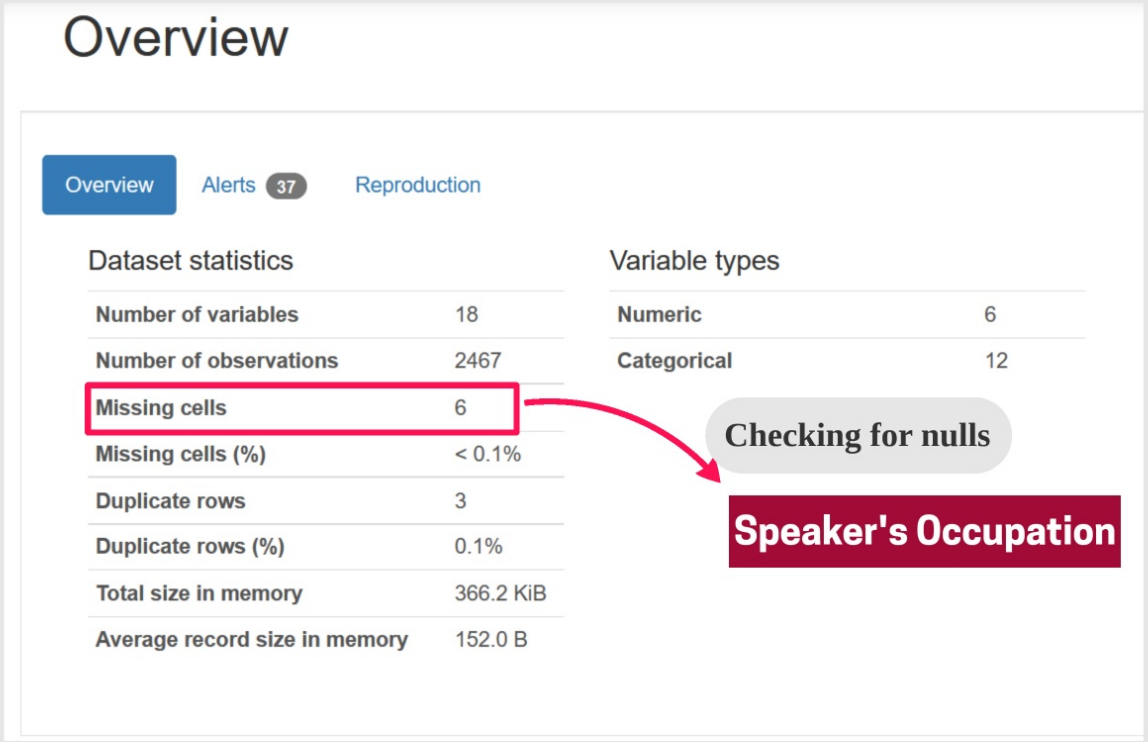
2467 rows
2 columns

=

ted_talk dataframe

2467 rows
18 columns

Pandas Profiling



WORKFLOW

Click to edit text

TEXT PREPROCESSING

MODELING OUTPUT

EDA

TOPIC MODELING

FINDINGS

Text Preprocessing on transcript column

1. Change the text to lower case
2. Remove everything between bracket
3. Remove stop words
4. Remove frequent words
5. Remove rare words
6. Apply lemmatization
7. Attempt to correct spelling but it was already correct
8. Return the transcript column to the Data frame

index 1 before text preprocessing

Good morning. How are you?(Laughter)It's been great, hasn't it? I've been blown away by the whole thing. In fact, I'm leaving. (Laughter)There have been three themes running through the conference which are relevant to what I want to talk about. One is the extraordinary evidence of human creativity in all of the presentations that we've had and in all of the people here.

index 1 after text preprocessing

good morning great hasnt ive blown away whole thing fact im leaving three theme running conference relevant want talk extraordinary evidence human creativity presentation weve variety range second put u place idea whats happen term future idea may play outi interest education

WORKFLOW

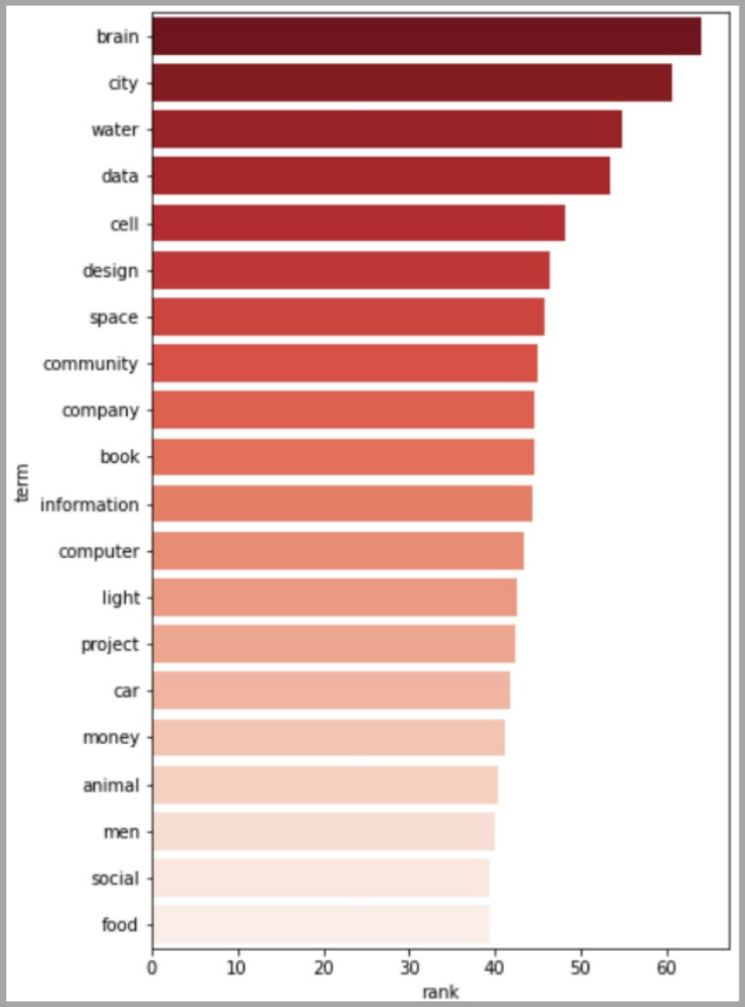
Click to edit text

- EDA
- TEXT PREPROCESSING
- TOPIC MODELING
- MODELING OUTPUT
- FINDINGS

Before Topic Modeling

The figure shown gives a list of words rank

we used TFIDF as our count vectorizer considering only 1-grams.



Topic Modeling

LDA	NMF
0: ['economist', 'sustainable', 'survey', 'financial', 'investment']	0: ['girl', 'men', 'mother', 'father', 'god']
1: ['brain', 'music', 'robot', 'sound', 'play']	1: ['cell', 'dna', 'gene', 'molecule', 'cancer']
2: ['la', 'sorry', 'oh', 'welcome', 'chose']	2: ['water', 'ocean', 'animal', 'fish', 'plant']
3: ['economist', 'sustainable', 'survey', 'financial', 'investment']	3: ['patient', 'cancer', 'disease', 'health', 'doctor']
4: ['economist', 'sustainable', 'survey', 'financial', 'investment']	4: ['computer', 'data', 'design', 'information', 'machine']
5: ['city', 'water', 'data', 'community', 'company']	5: ['city', 'car', 'design', 'space', 'street']
6: ['economist', 'sustainable', 'survey', 'financial', 'investment']	6: ['brain', 'memory', 'animal', 'behavior', 'activity']
7: ['economist', 'sustainable', 'survey', 'financial', 'investment']	7: ['robot', 'machine', 'leg', 'animal', 'intelligence']
8: ['economist', 'sustainable', 'survey', 'financial', 'investment']	8: ['universe', 'planet', 'earth', 'star', 'space']
9: ['web', 'internet', 'link', 'page', 'government']	9: ['government', 'company', 'money', 'dollar', 'business']

in the **LDA** model results: words are not related to each other and repeated.
NMF optimizes the topic molding where the topic modeling results are much better with specific topics.

Topic Modeling

LDA	NMF
0: ['economist', 'sustainable', 'survey', 'financial', 'investment']	0: ['girl', 'men', 'mother', 'father', 'god'] Family
1: ['brain', 'music', 'robot', 'sound', 'play']	1: ['cell', 'dna', 'gene', 'molecule', 'cancer'] Genetics
2: ['la', 'sorry', 'oh', 'welcome', 'chose']	2: ['water', 'ocean', 'animal', 'fish', 'plant'] Nature
3: ['economist', 'sustainable', 'survey', 'financial', 'investment']	3: ['patient', 'cancer', 'disease', 'health', 'doctor'] Health
4: ['economist', 'sustainable', 'survey', 'financial', 'investment']	4: ['computer', 'data', 'design', 'information', 'machine'] Technology
5: ['city', 'water', 'data', 'community', 'company']	5: ['city', 'car', 'design', 'space', 'street'] Civil
6: ['economist', 'sustainable', 'survey', 'financial', 'investment']	6: ['brain', 'memory', 'animal', 'behavior', 'activity'] Biology
7: ['economist', 'sustainable', 'survey', 'financial', 'investment']	7: ['robot', 'machine', 'leg', 'animal', 'intelligence'] Robotics
8: ['economist', 'sustainable', 'survey', 'financial', 'investment']	8: ['universe', 'planet', 'earth', 'star', 'space'] Space
9: ['web', 'internet', 'link', 'page', 'government']	9: ['government', 'company', 'money', 'dollar', 'business'] Governme

in the **LDA** model results: words are not related to each other and repeated.
NMF optimizes the topic molding where the topic modeling results are much better with specific topics.

WORKFLOW

Click to edit text

TEXT PREPROCESSING

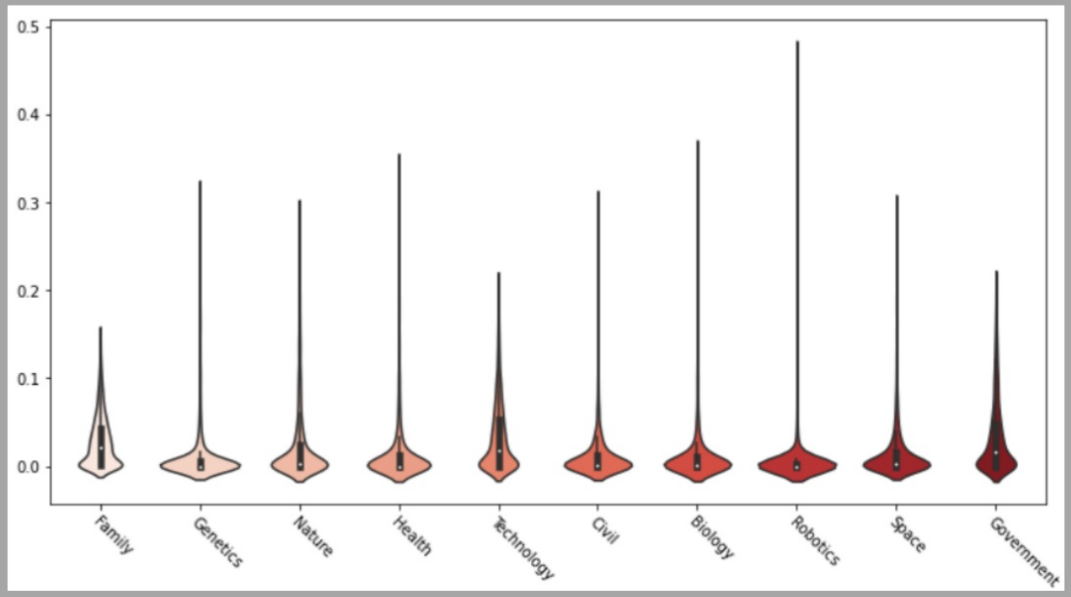
MODELING OUTPUT

EDA

TOPIC MODELING

FINDINGS

Topic Modeling Output



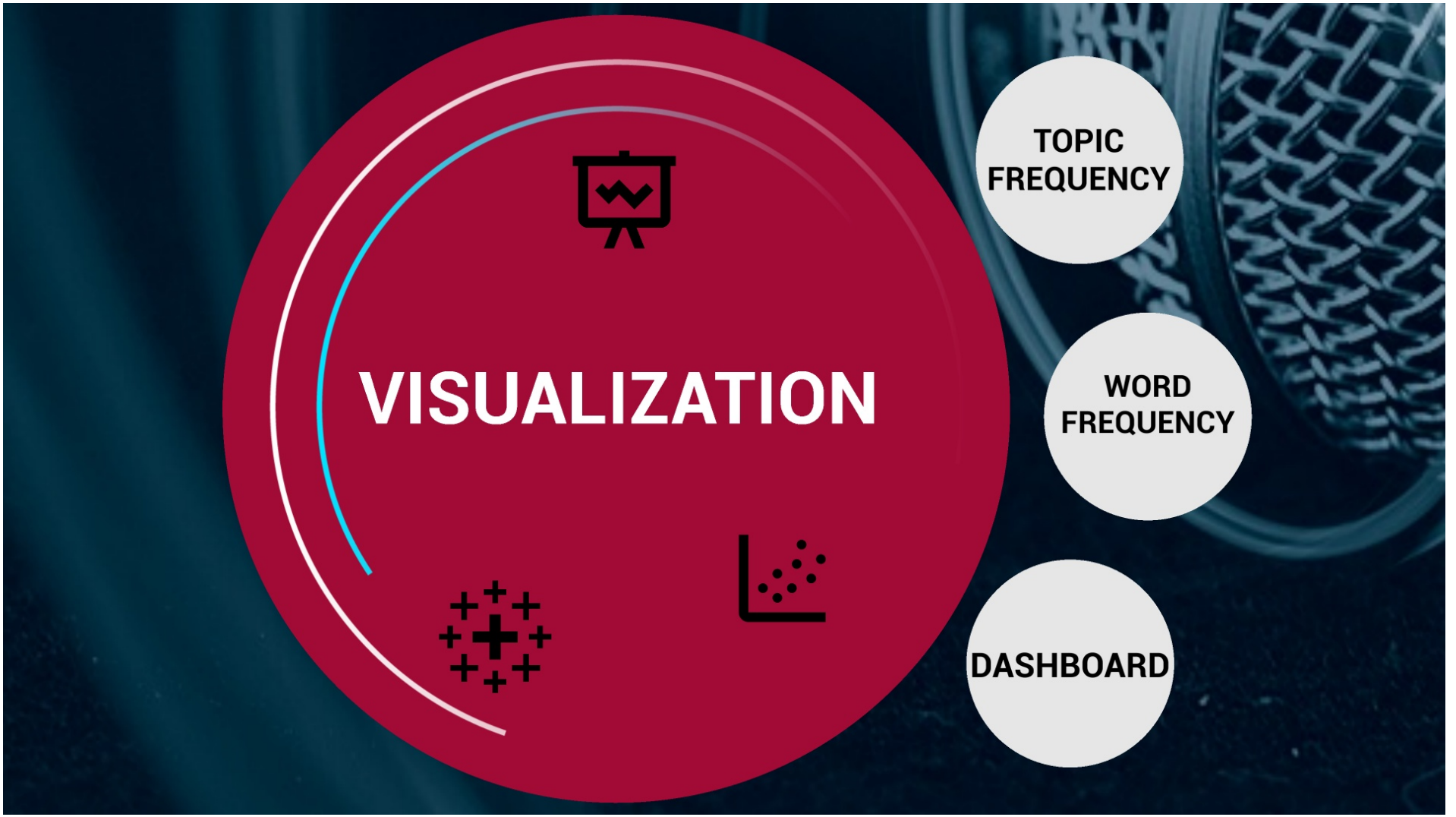
The above figure illustrates the distribution for each topic.
Topic “Robotics” appears to have the strongest outliers among the others.
Topics “Family, Technology and Government” appear to be most frequent and well distributed.

Whats done after Topic Modeling?

A new column was created in the dataframe with the topics for further Analysis.

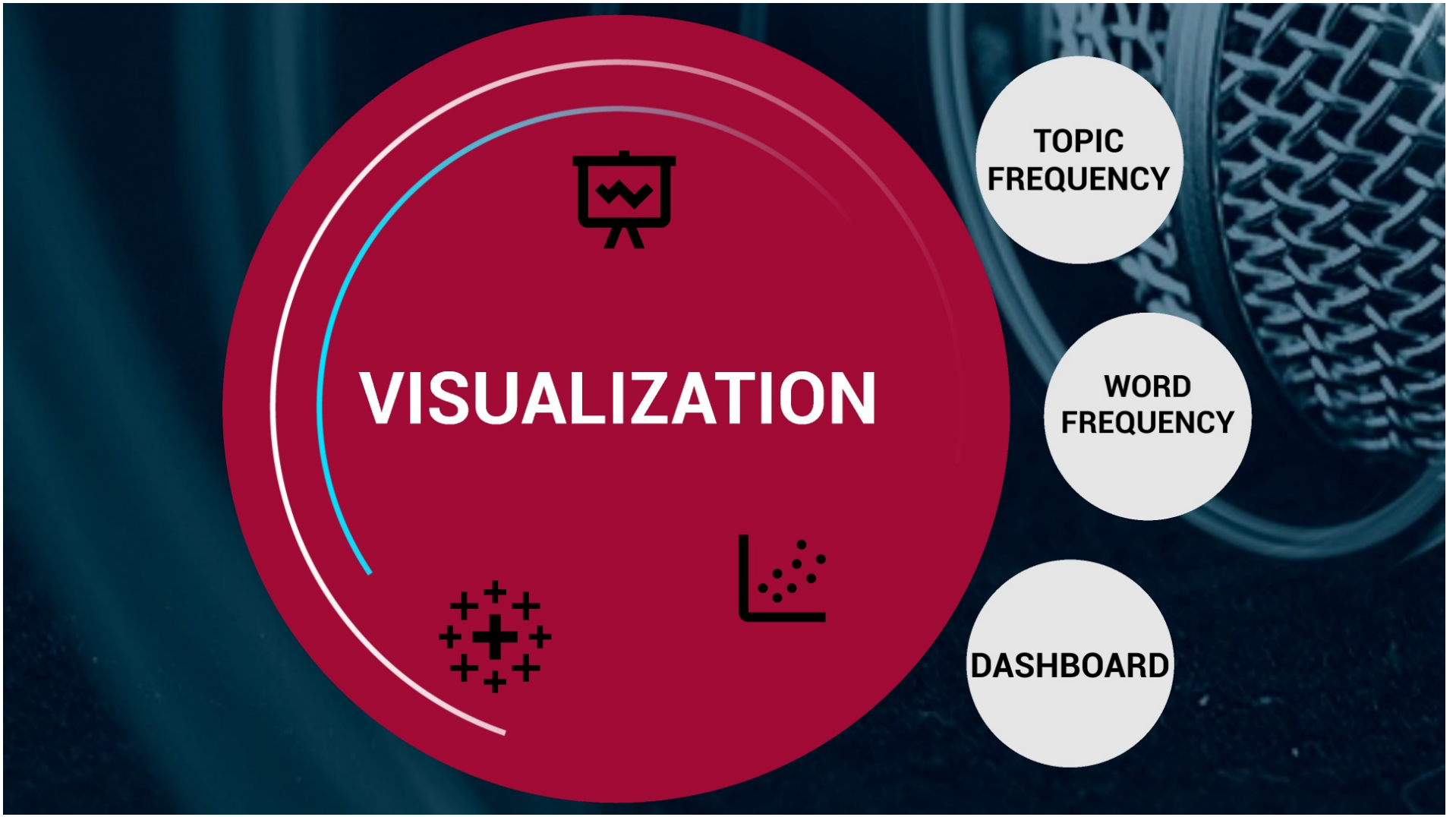
ments	description	duration	event	film_date	languages	main_speaker	name	num_speaker	published_date	ratings	related_talks	speaker_occupation	tags	title	url	views	transcript	Topic
4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	1140825600	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 19645}, {...	[{'id': 865, 'hero': 'https://pe.tedcdn.com/im...	Author/educator	['children', 'creativity', 'culture', 'dance',...	Do schools kill creativity?	https://www.ted.com/talks/ken_robinson_says_sc...	47227110	good morning great hasnt ive blown away whole ...	Family
265	With the same humor and humanity he exuded in ...	977	TED2006	1140825600	43	Al Gore	Al Gore: Averting the climate crisis	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 544}, {'i...	[{'id': 243, 'hero': 'https://pe.tedcdn.com/im...	Climate advocate	['alternative energy', 'cars', 'climate change...	Averting the climate crisis	https://www.ted.com/talks/al_gore_on_averting_...	3200520	thank much chris truly great honor opportunity...	Government
124	New York Times columnist David Pogue takes aim...	1286	TED2006	1140739200	26	David Pogue	David Pogue: Simplicity sells	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 964}, {'i...	[{'id': 1725, 'hero': 'https://pe.tedcdn.com/i...	Technology columnist	['computers', 'entertainment', 'interface desi...	Simplicity sells	https://www.ted.com/talks/david_pogue_says_sim...	1636292	hello voice mail old friend ive called tech su...	Technology
200	In an emotionally charged talk, MacArthur-winn...	1116	TED2006	1140912000	35	Majora Carter	Majora Carter: Greening the ghetto	1	1151367060	[{'id': 3, 'name': 'Courageous', 'count': 760}...	[{'id': 1041, 'hero': 'https://pe.tedcdn.com/i...	Activist for environmental justice	['MacArthur grant', 'activism', 'business', 'c...	Greening the ghetto	https://www.ted.com/talks/majora_carter_s_tale...	1697550	youre today im happy youve heard sustainable d...	Civil
593	You've never seen data presented like this. Wi...	1190	TED2006	1140566400	48	Hans Rosling	Hans Rosling: The best stats you've ever seen	1	1151440680	[{'id': 9, 'name': 'Ingenious', 'count': 3202}...	[{'id': 2056, 'hero': 'https://pe.tedcdn.com/i...	Global health expert; data visionary	['Africa', 'Asia', 'Google', 'demo', 'economic...	The best stats you've ever seen	https://www.ted.com/talks/hans_rosling_shows_t...	12005869	10 year ago took task teach global development...	Government


OUTLINE
WORKFLOW
VISUALIZATION
CONCLUSION
TED TALKS
Topic Modeling
Natural Language Processing
Hayat Aldhahri & Juri AlSayigh

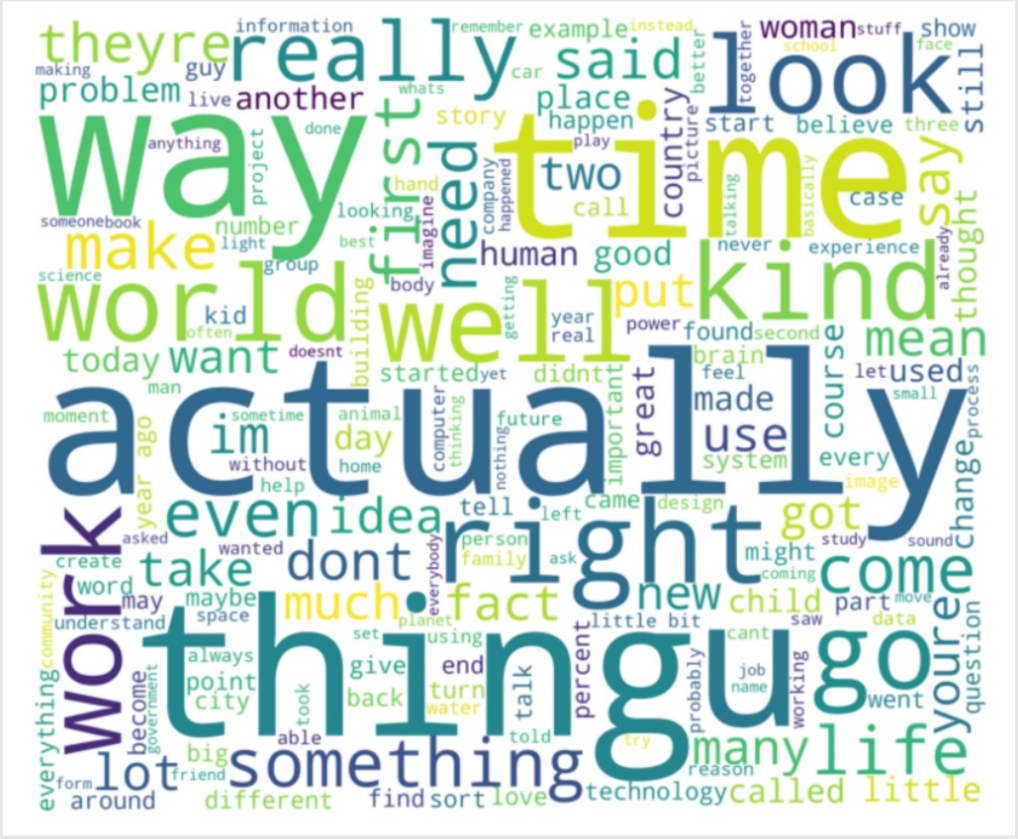

VISUALIZATION
TOPIC FREQUENCY
WORD FREQUENCY
DASHBOARD

TED Talks Topic Frequency

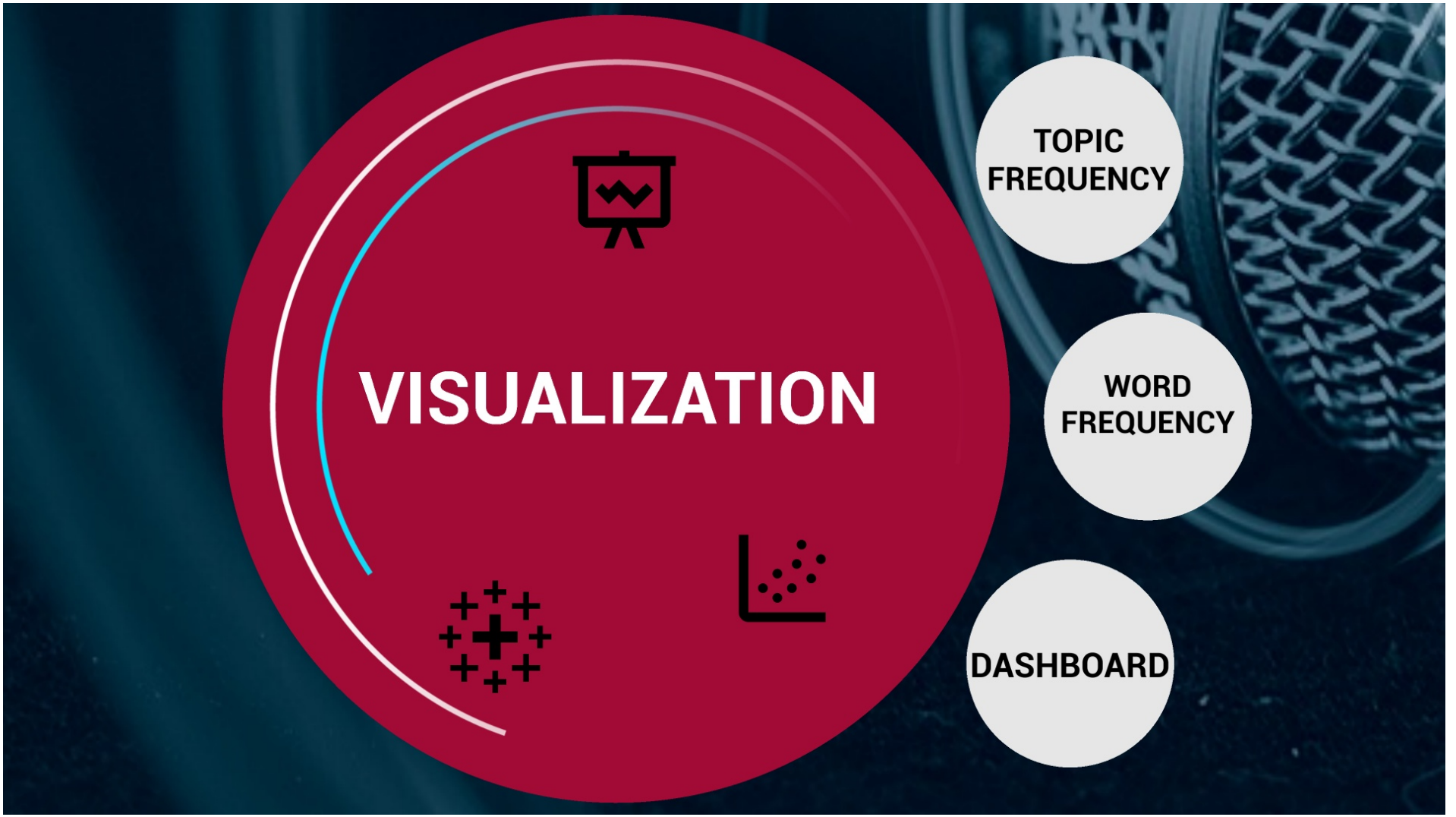
Topics frequency appearance in Ted Talks in a word cloud visualization


VISUALIZATION
TOPIC FREQUENCY
WORD FREQUENCY
DASHBOARD

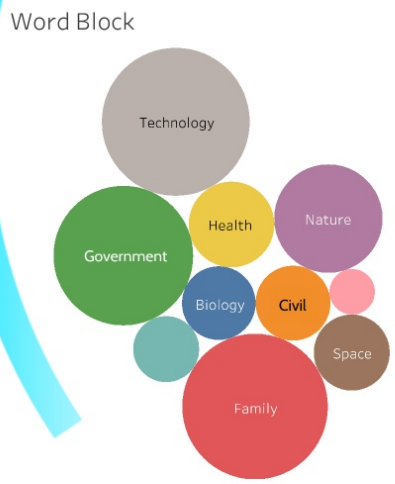
Word Frequency per TedTalk

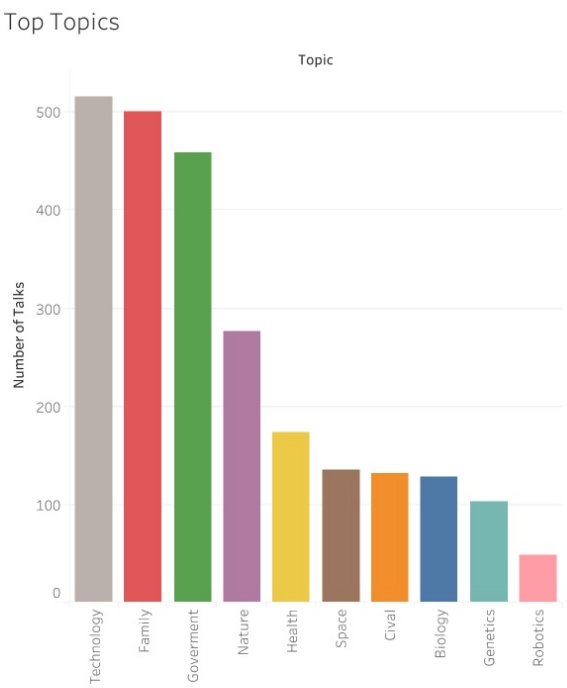


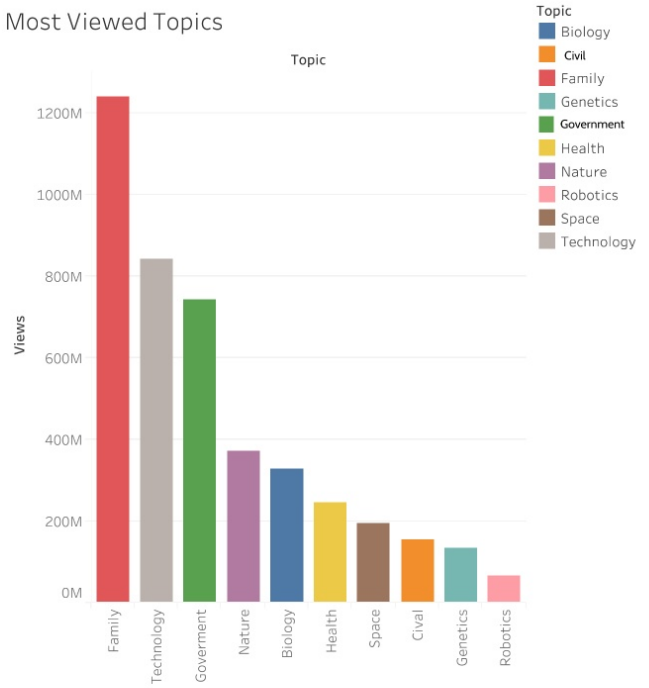
a word cloud created to illustrate the most frequent words mentioned in Ted Talks since 1994


VISUALIZATION
TOPIC FREQUENCY
WORD FREQUENCY
DASHBOARD

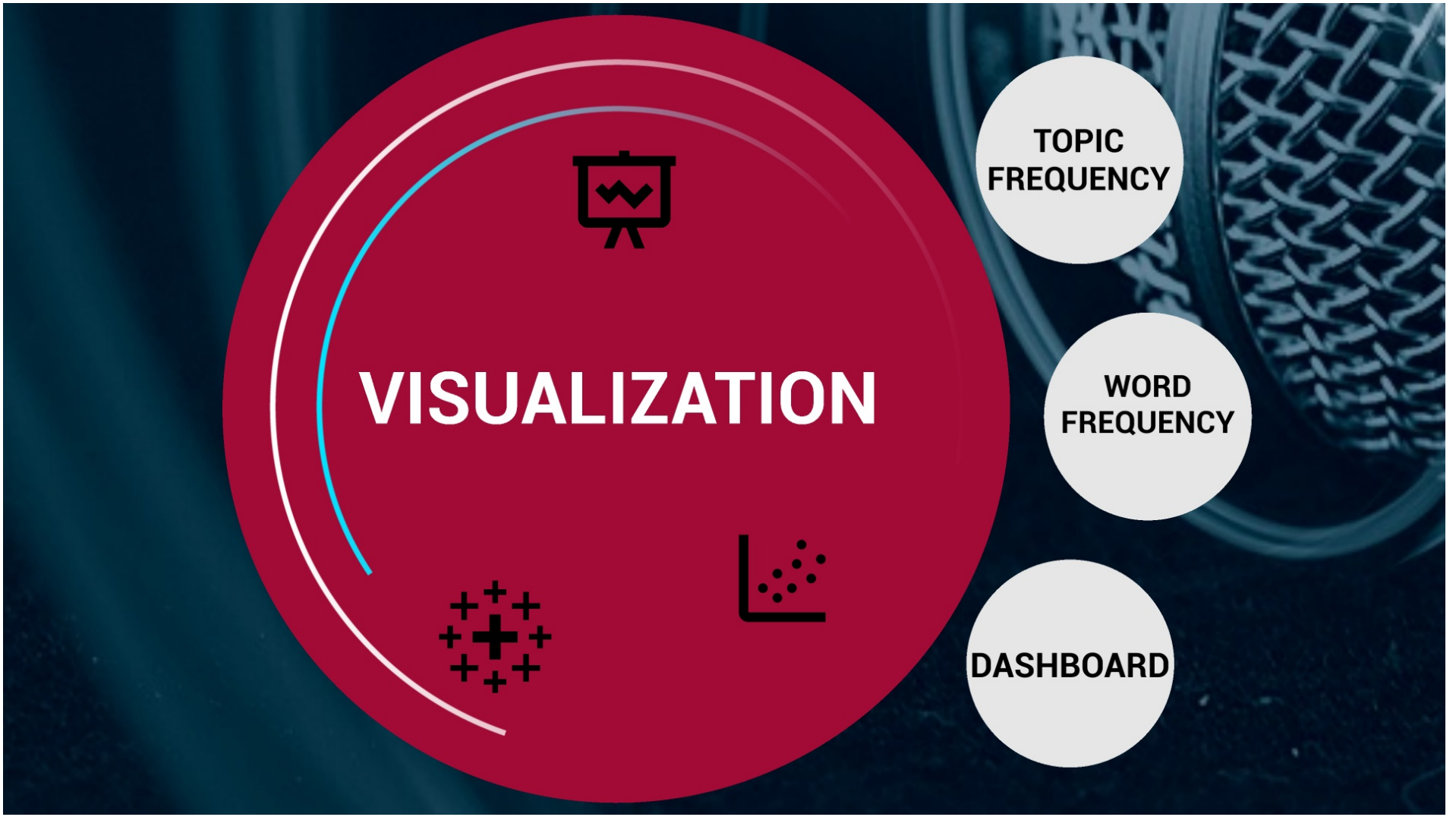
DASHBOARD





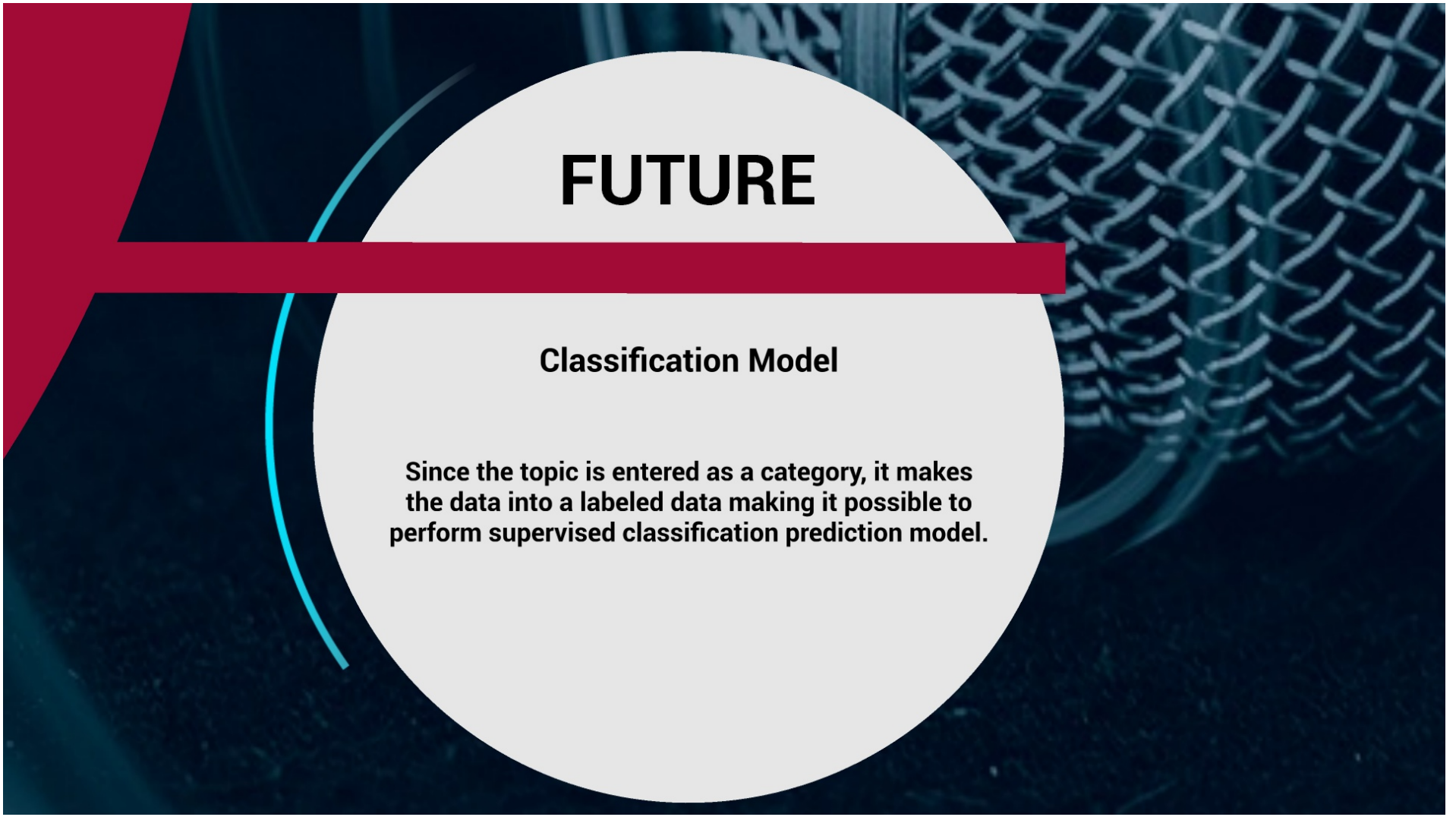


https://public.tableau.com/app/profile/hayat4538/viz/NLP_TedTalk/Dashboard1?publish=yes


VISUALIZATION
TOPIC FREQUENCY
WORD FREQUENCY
DASHBOARD


OUTLINE
WORKFLOW
VISUALIZATION
CONCLUSION
TED TALKS
Topic Modeling
Natural Language Processing
Hayat Aldhahri & Juri AlSayigh


Thank You
Hayat & Juri
FUTURE


FUTURE
Classification Model
Since the topic is entered as a category, it makes the data into a labeled data making it possible to perform supervised classification prediction model.


Thank You
Hayat & Juri
FUTURE


OUTLINE
WORKFLOW
VISUALIZATION
CONCLUSION
TED TALKS
Topic Modeling
Natural Language Processing
Hayat Aldhahri & Juri AlSayigh