

Abstract geometric lines in black on a white background, forming various overlapping polygons and shapes.

# CLUSTERING OF YOUTUBE NEWS CHANNEL VIDEOS

BASED ON TITLE & DESCRIPTION

Jan Paul Kieffer  
Kong Hyeon Kim  
Chi Ying Zoe Lai  
Max Wurthmann



# AGENDA

---

Motivation & Goals

---

Exploratory Data Analysis

---

Models & Methods

---

Tasks & Plan

---

Potential Problems

# MOTIVATION

- What contents do English news channels post on YouTube?
- Is there any focus on what topics they post?



# GOALS

## Clustering by Topics

- Training-based Models
- Non-training-based Models

## Comparing Clusters of each Channel

- Distance
- Size, Popularity
- Development over time
- Relation between Title and Views

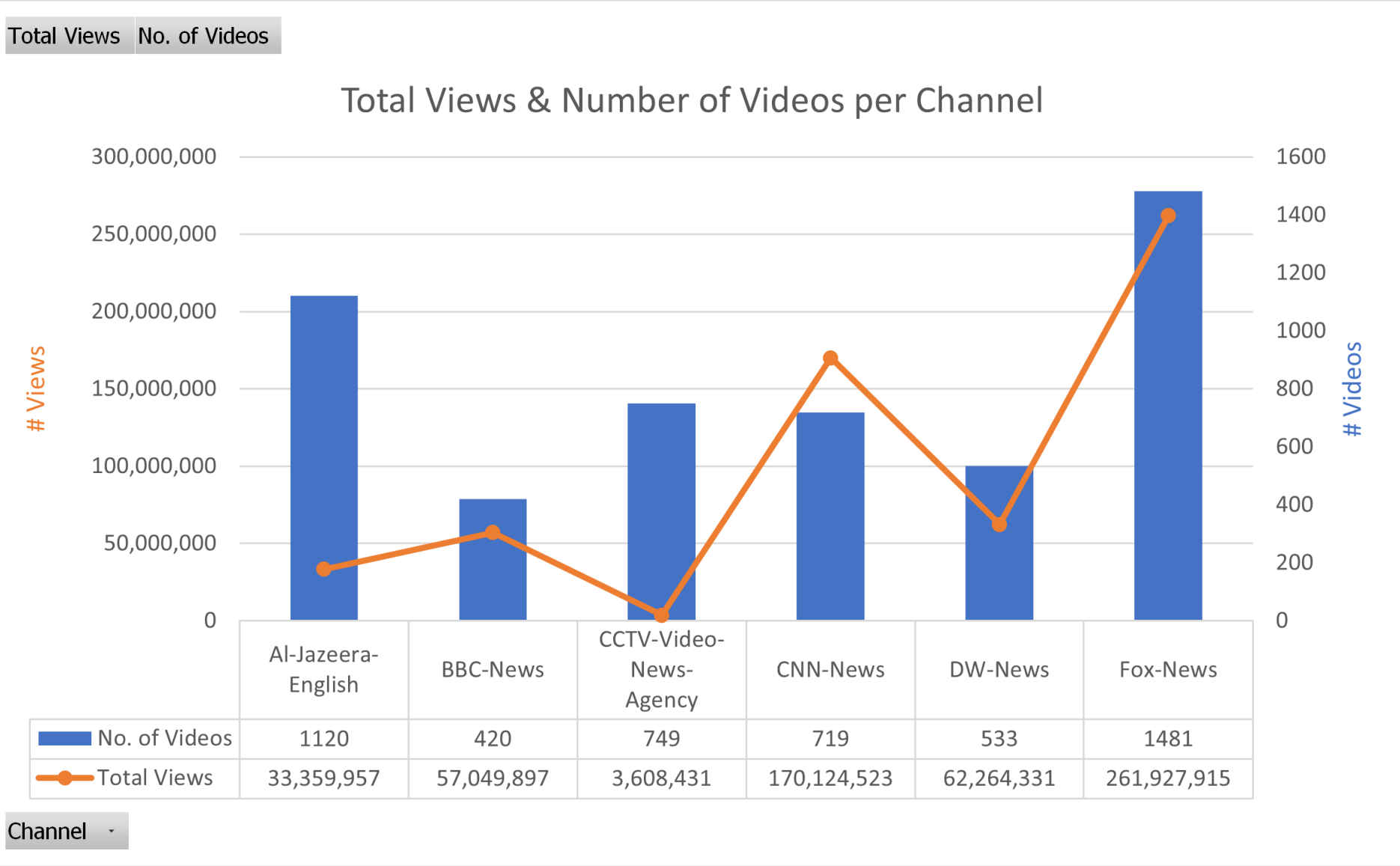
## Comparing Clusters across Channels

- Differences
- Similarities

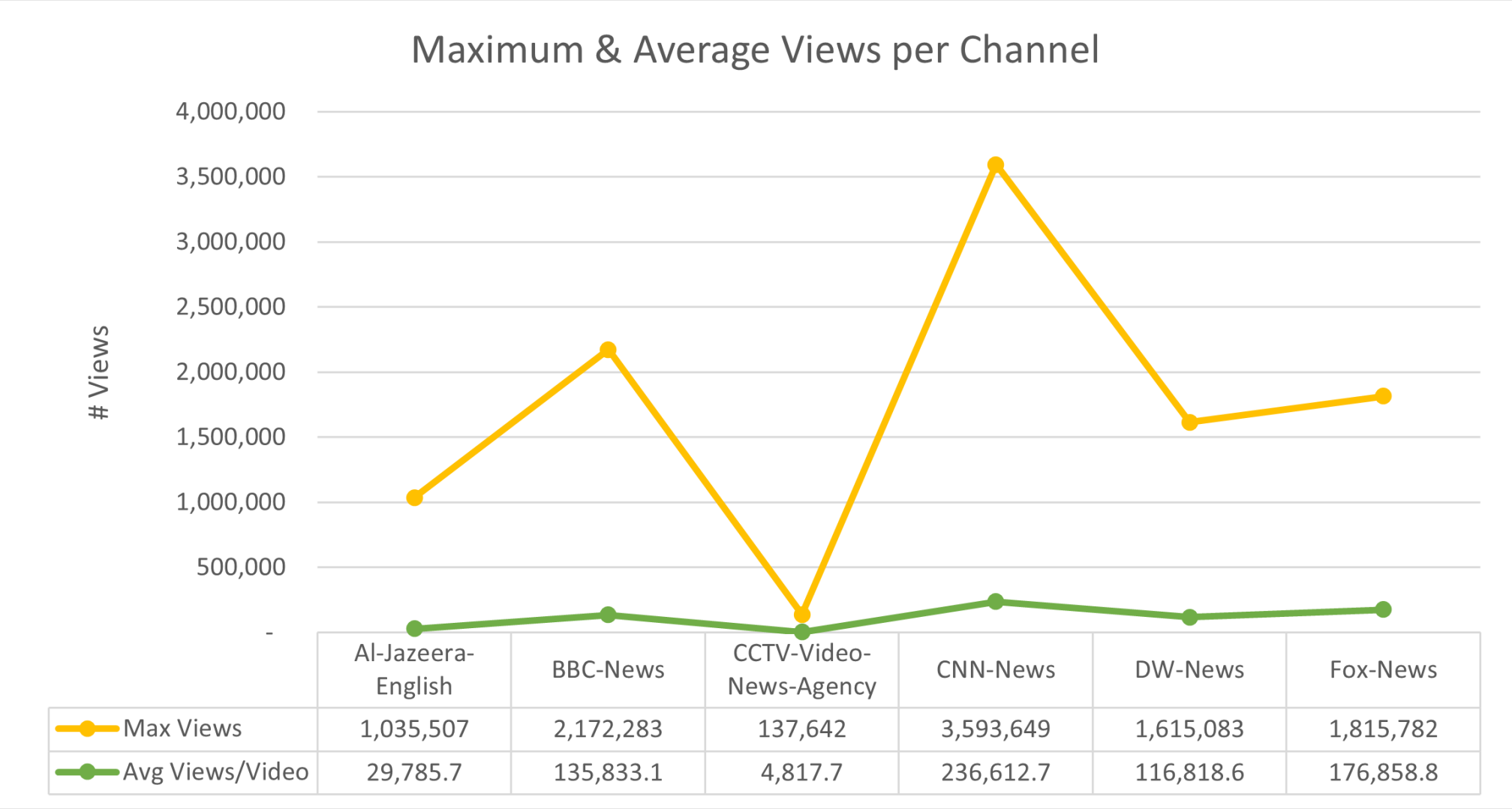
# DATA COLLECTION

- YouTube Channels covered:
  1. DW-News
  2. CNN-News
  3. BBC-News
  4. Al-Jazeera-English
  5. Fox-News
  6. CCTV-Videos-News-Agency
- Dated:
  - Since 25.03.2023
  - Current dataset up to 17.05.2023
  - To be updated until modelling starts
- Content includes:  
Title, Description, Views , Video Length,  
Upload Time, Data Retrieval Time

# EXPLORATORY DATA ANALYSIS



# EXPLORATORY DATA ANALYSIS





# MODELS & METHODS

## Embeddings

- Doc2Vec
- Contextual sentence embeddings using a pretrained Model
  - BERT, T5
  - Concatenate title and description, or
  - Input separately and then combine results

## Visualization

- Wordcloud
- Self-organizing Maps
- Reduction and Plot
  - PCA
  - t-SNE



# MODELS & METHODS

## Clustering

- Whole dataset:  
Representation of overall news
- Separate dataset:  
Insights about each news channel
- Unsupervised methods
  - K-Means Clustering
  - Self-organizing Maps

## Topic Inference

- Path 1 (unsupervised):
  1. Perform clustering
  2. Extract an n-gram or tokens that are especially frequent
- Path 2 (supervised):
  1. Annotate parts of the data
  2. Finetune a retrained model to predict the topic

Paul — Problem Formulation, Model Overview

Zoe — Data Preprocessing, Slides Preparation

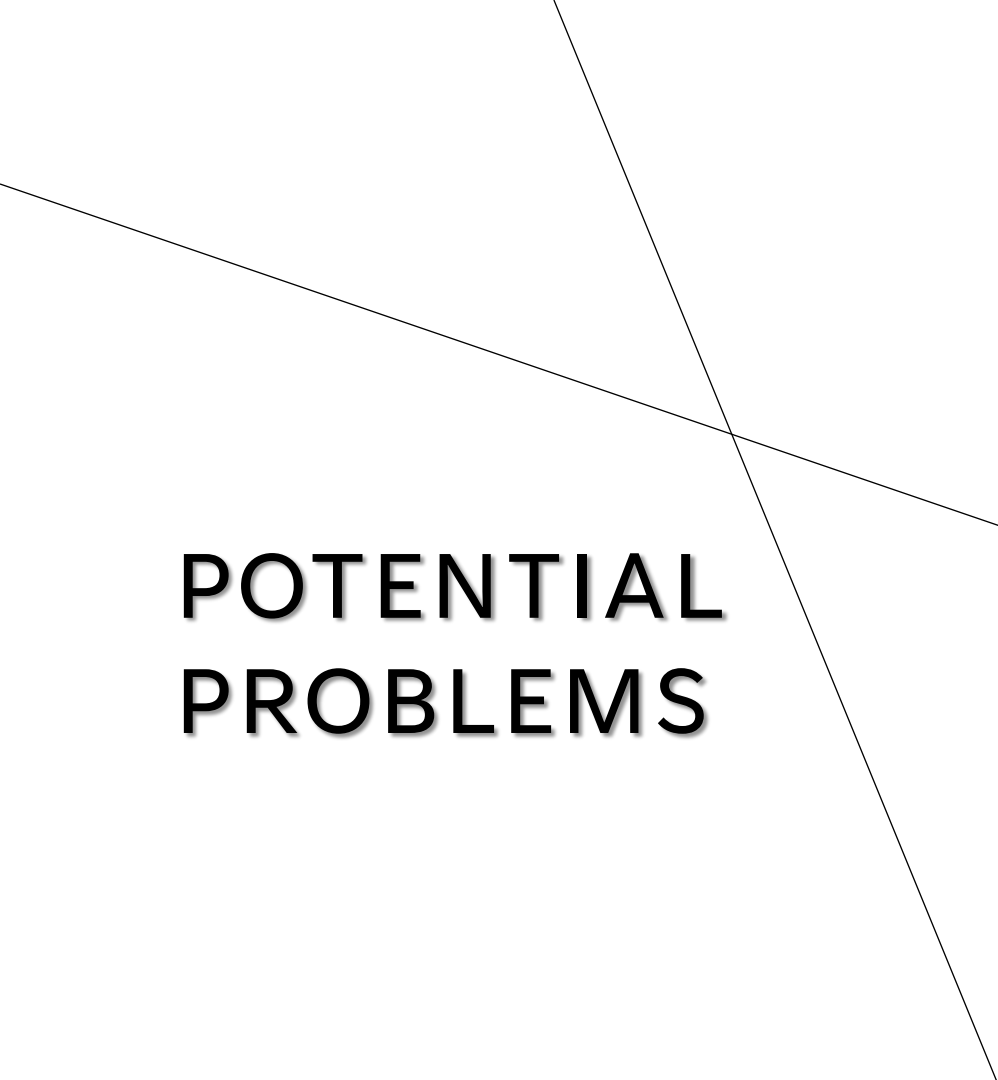
Max — Models and Methods

All — Model Application

**TASKS**

# SCHEDULE

Timeline	Progress
End of May	Data preprocessing
Mid-June	Model selection
End of June	Model application
Mid-July	Evaluation



# POTENTIAL PROBLEMS

- Insufficient dataset
- Incomplete video descriptions
- During scraping deleted commas and semicolons.

A series of white, thin, overlapping geometric lines on a black background, forming an abstract, angular shape on the left side of the slide.

THANK YOU