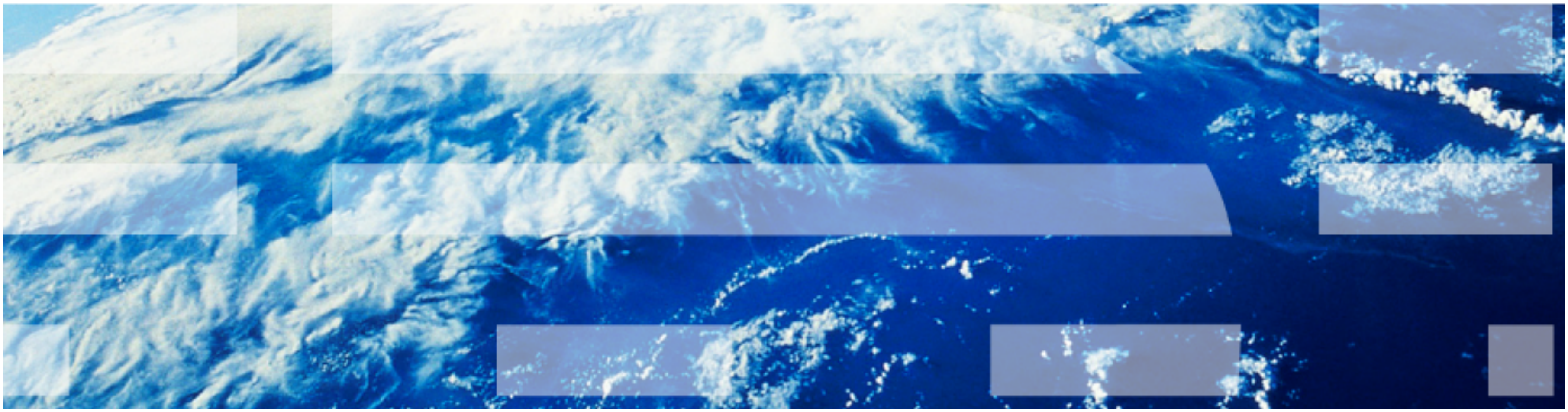
E6895 Advanced Big Data Analytics:

Data Analytics for Video Popularity

Ziyu He (zh2255)
Haoxiang Gao (hg2412)



April 7, 2016

Current Work

- Data Collection and Preprocessing
  - Pipeline to crawl data from different sources efficiently
  - Different approaches to handle unstructured and multi-typed data
  - Summarized as features

- Model Training and Interpretation
  - Machine learning model for prediction
  - Validation and inference

- Item Based Recommendation and Visualization
  - Item based recommender
  - Data Storage and visualization

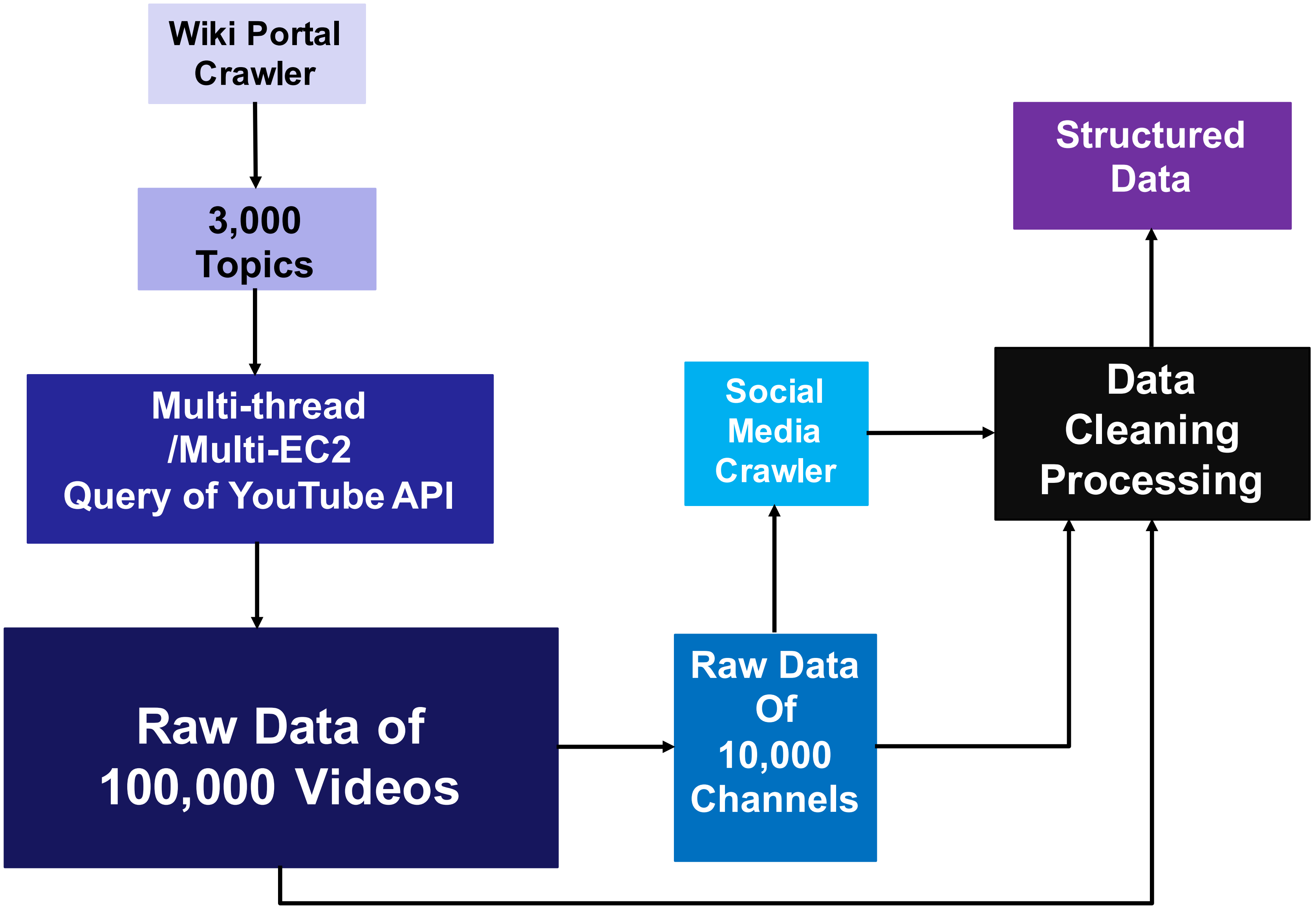
Core Features

- Direct Quantitation of Popularity of a Channel
  - Number of viewers/subscribers/comments

- Viewers Opinion (aggregation of videos stats)
  - Sentiment analysis score of video comments
  - Ratio of likes/dislike, comment/viewer, fav/viewer
  - Social media reference score

- Quantitative Description of Channel and Its Content
  - Frequency of publish
  - Duration
  - Content category
  - Topics obtained from topic modeling on description
  - Named entity recognition score
  - Characteristics of content creator: social media behavior

Data Collection Pipeline



Wiki Portal Crawler

3,000 Topics

Multi-thread /Multi-EC2 Query of YouTube API

Raw Data of 100,000 Videos

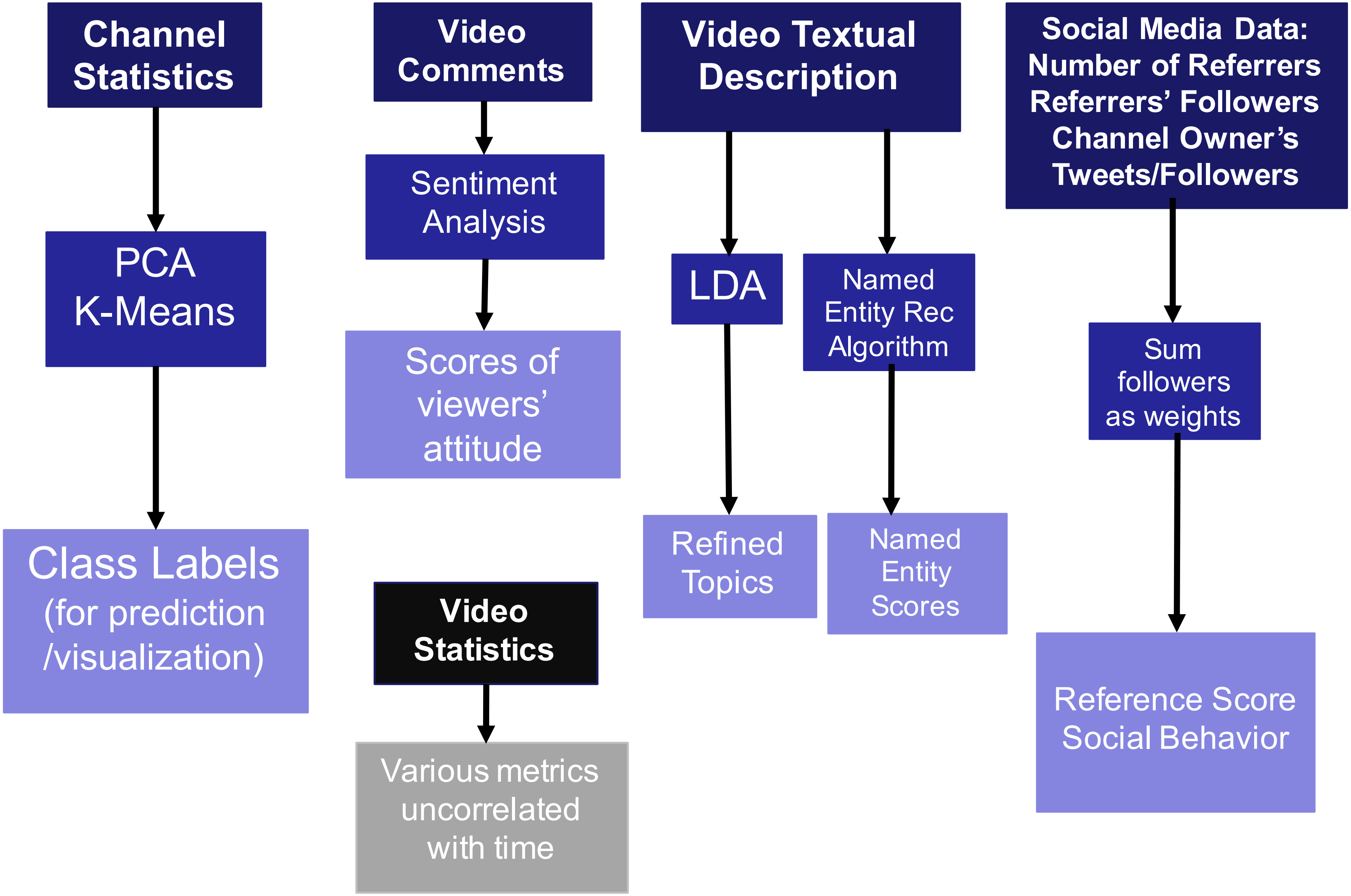
Raw Data Of 10,000 Channels

Social Media Crawler

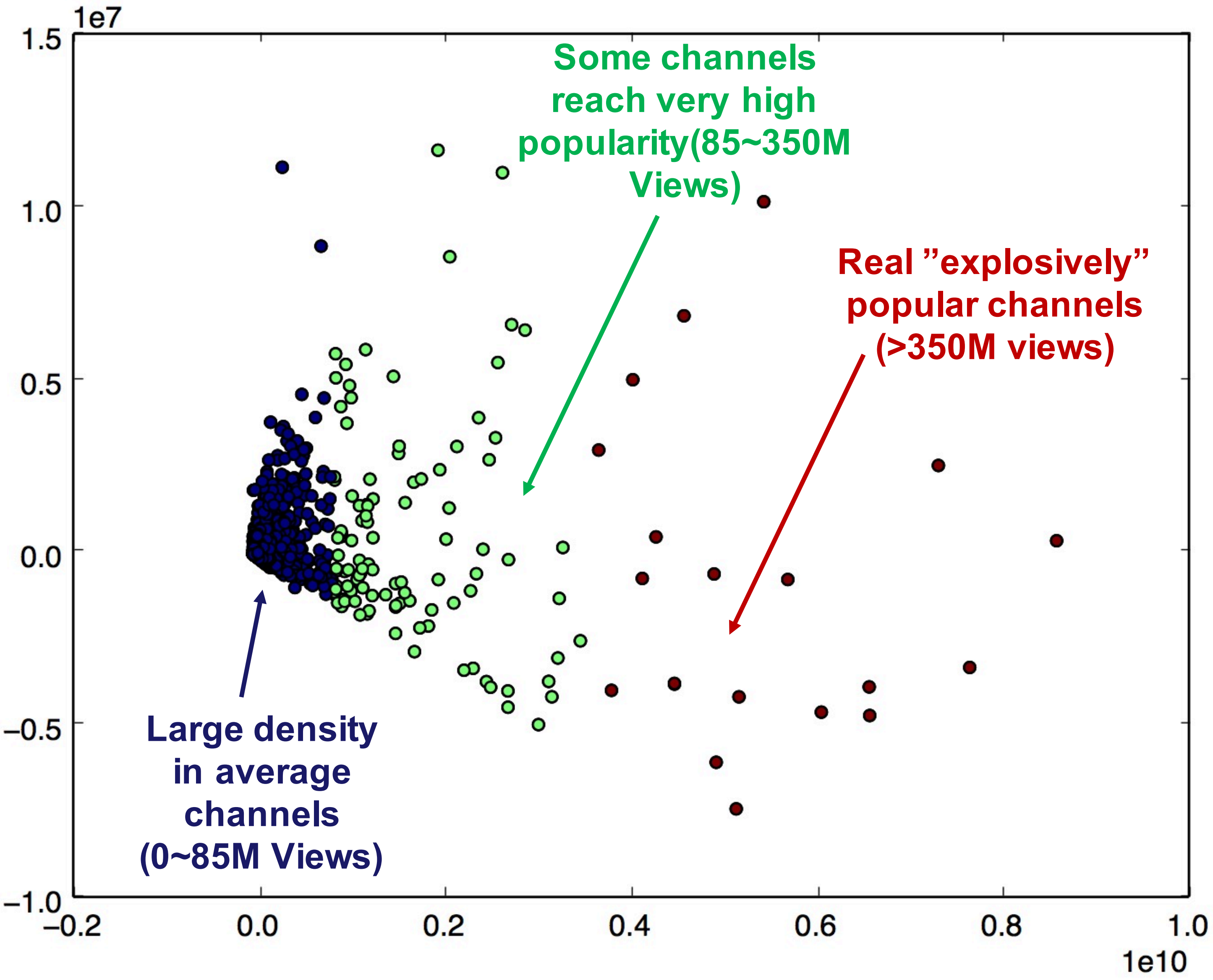
Data Cleaning Processing

Structured Data

Data Preprocessing



Results: PCA of Popularity Clusters



Some channels reach very high popularity(85~350M Views)

Real ”explosively” popular channels (>350M views)

Large density in average channels (0~85M Views)

Results: Model Validation

Consider all of the features obtained from data processing and predict on class labels

5-Fold Cross-Validation:
•1.3% general error rate
•0.5% error rate for label “1” prediction
•5% error rate for label “2” prediction
•5% error rate for label “3” prediction

Achieved low error rate for label “1” and general case, while worse error rate for label “2” and “3”. Probably due to unbalanced portion of different classes, which is similar to facial detection problem and can be potentially improved with cascaded classifiers

Results: Inference of Features

1. Frequency of Publishing Videos
2. Reference on social media
3. Neutral/Compound comment sentiment
4. Activity of channel owner on social media
5. Rate of subscription
6. Rate of “Likes”
7. Rate of comments
8. Positive comment sentiment
9. Occurrence of named entities
10. Duration
11. Negative sentiment

Results: Categories

1. "Comedy”
2. "Drama”
3. “Horror”
4. “Documentary”
5. “Education”
6. “People and Blogs”
7. “Anime/Animation”
8. “Foreign”
9. “Nonprofits & Activism”
10. “Family”
11. “Action/Adventure”
12. “Sci-fi”
13. “Thriller”

Future Work

(1) Visualization of Video Recommender System

(2) A robust solution for topic modelling other than LDA