

Audio-visual Sentiment Analysis for Learning Emotional Arcs in Movies

Eric Chu, Deb Roy

MIT Media Lab, Laboratory for Social Machines

Story-Learning Machine

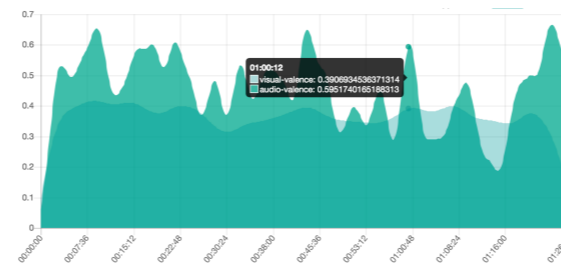
Mapping the relationship between story structure
and engagement across networks

Story-Learning Machine

Mapping the relationship between story structure and engagement across networks

1 Story anatomy

Audio-visual emotional arc

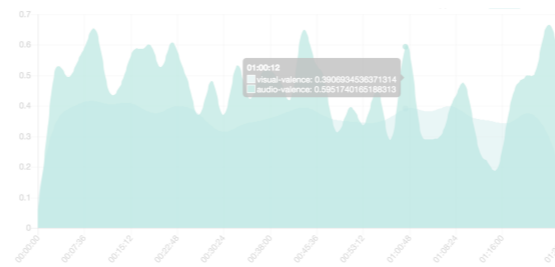


Story-Learning Machine

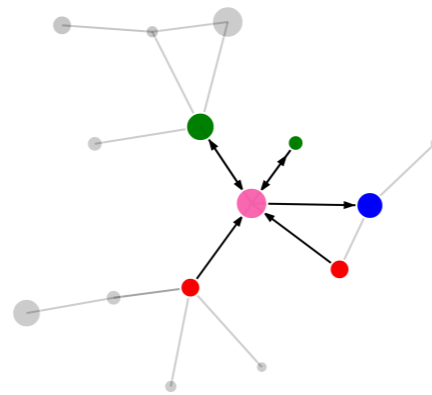
Mapping the relationship between story structure and engagement across networks

1 Story anatomy

Audio-visual emotional arc



2 Engagement analysis

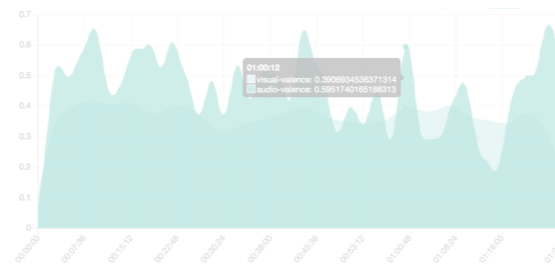


Story-Learning Machine

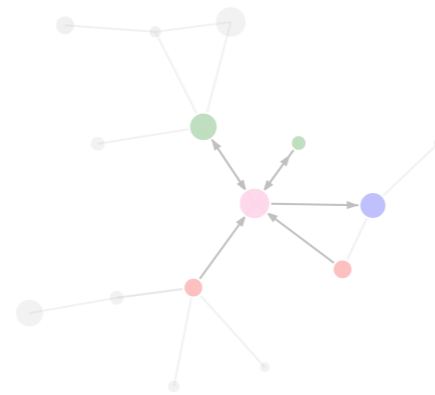
Mapping the relationship between story structure and engagement across networks

1 Story anatomy

Audio-visual emotional arc



2 Engagement analysis



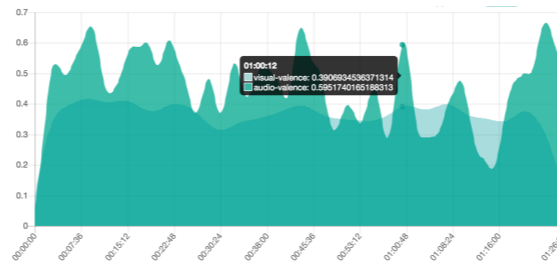
3 Intervention

Story-Learning Machine

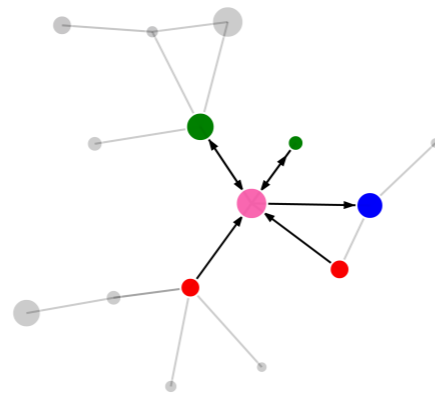
Mapping the relationship between story structure and engagement across networks

1 Story anatomy

Audio-visual emotional arc



2 Engagement analysis



3 Intervention

Background

- Why emotional arcs?
 - Narrative theory of emotional arcs (Vonnegut, Campbell, etc.)
 - Research on power of emotions
 - What Makes Online Content Viral? (Berger & Milkman, 2012)
 - Emotion and Decision Making (Lerner, 2015)

Background

- Why emotional arcs?
 - Narrative theory of emotional arcs (Vonnegut, Campbell, etc.)
 - Research on power of emotions
 - What Makes Online Content Viral? (Berger & Milkman, 2012)
 - Emotion and Decision Making (Lerner, 2015)
- Why videos?
 - Increasingly popular in social media
 - Rich medium
 - Powerful

Outline

- Part 1 — creating arcs
 - Visual
 - Smoothing arcs
 - Audio
 - Crowdsourcing ground truth
 1. Evaluation
 2. Combining audio and visual
- Part 2
 - Clustering arcs based on shape
 - Predicting engagement

Image dataset — Sentibank



famous_church $\xrightarrow{\text{sentiment lexicon}}$ 2.00



fluffy_ears $\xrightarrow{\text{sentiment lexicon}}$ 1.27

Flickr

Based on Flickr tags

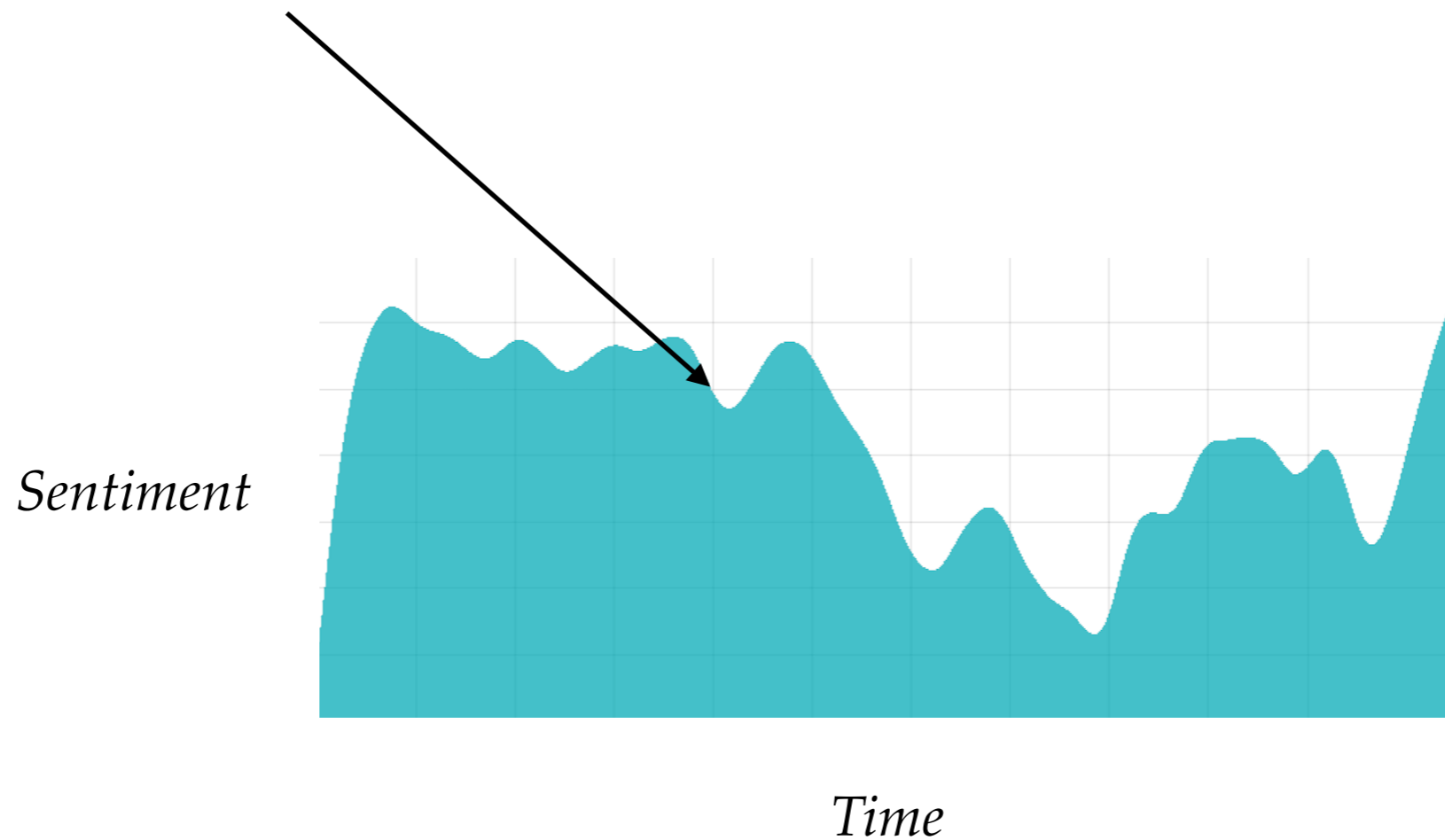
Image model

- Deep convolutional neural network to predict sentiment value

Image model

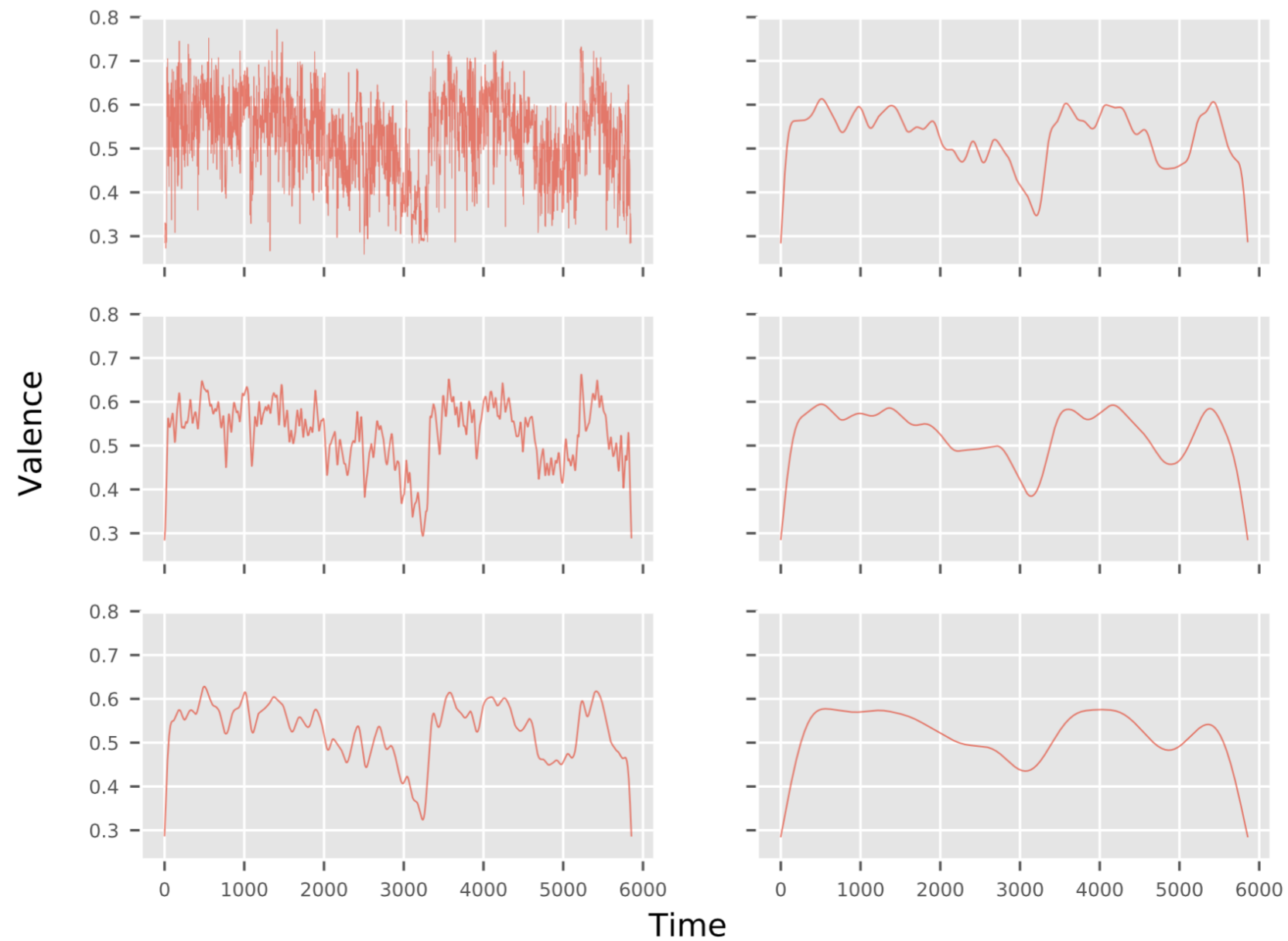
- Deep convolutional neural network to predict sentiment value

$$f_{\text{sentiment}} \left(\text{Image} \right) = 0.76$$



Constructing arcs

Smooth by convolving with Hahn window of size w



Audio

- Spotify Dataset
 - Million Song Dataset lacks sentiment / valence tag
 - Collected 600,000+ 30-second samples plus features from Spotify
 - Features: valence, energy, speechiness, etc.

Audio

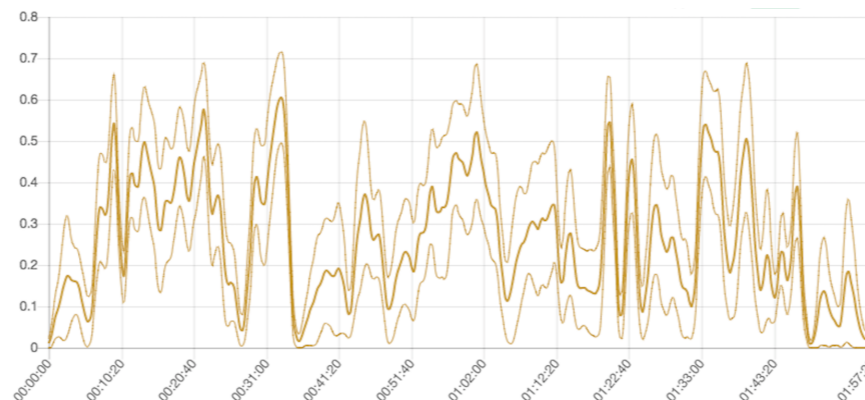
- Spotify Dataset
 - Million Song Dataset lacks sentiment / valence tag
 - Collected 600,000+ 30-second samples plus features from Spotify
 - Features: valence, energy, speechiness, etc.
- Model
 - Convert 20-second window into 96-bin mel-spectrogram
 - 5 conv layers with ELU and batch normalization, followed by fully connected layer

Audio

- Spotify Dataset
 - Million Song Dataset lacks sentiment / valence tag
 - Collected 600,000+ 30-second samples plus features from Spotify
 - Features: valence, energy, speechiness, etc.
- Model
 - Convert 20-second window into 96-bin mel-spectrogram
 - 5 conv layers with ELU and batch normalization, followed by fully connected layer
- Covariate Shift
 - Movies contain background noises, people talking, silence, etc.
 - Want to be able to weight our predictions / produce uncertainty estimates

Audio

- Spotify Dataset
 - Million Song Dataset lacks sentiment / valence tag
 - Collected 600,000+ 30-second samples plus features from Spotify
 - Features: valence, energy, speechiness, etc.
- Model
 - Convert 20-second window into 96-bin mel-spectrogram
 - 5 conv layers with ELU and batch normalization, followed by fully connected layer
- Covariate Shift
 - Movies contain background noises, people talking, silence, etc.
 - Want to be able to weight our predictions / produce uncertainty estimates
 - *Gal et. al (2015)*: at test time, set dropout prob to 0.5, pass input k times through network. Standard deviation of predictions defines confidence interval



Collecting ground truth data

Collecting ground truth data

- **Extract ~1000 30-second clips from peaks and valleys of audio and visual arcs**
 - 1-7 clips from ~100 movies
- **Each clip is annotated by 3 reviewers**


Collecting ground truth data

- **Extract ~1000 30-second clips from peaks and valleys of audio and visual arcs**
 - 1-7 clips from ~100 movies
- **Each clip is annotated by 3 reviewers**
- 4 questions:
 - 1. How positive or negative is this video clip? (1 being most negative, 7 being most positive)**
 2. How confident are you in your previous answer? (1 being least confident, 10 being most confident)
 3. Which emotion(s) does this video clip contain or convey? (check all that apply or none of the above)
 - Options: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, none of the above
 4. Which of the following contributed to your decisions? (check all that apply)
 - Options: audio, dialogue, visual (actions, scene, setting)

Evaluation

Extracted from

Mean valence rating by
annotators


$$Precision = \frac{\#(\textit{peak \& positive}) + \#(\textit{valley \& negative})}{\#clips}$$

Evaluation: precision on audio

Stddev	Audio-peak	Audio-valley
[0, 0.02)	1.0	0.921
[0.02, 0.04)	1.0	0.679
[0.04, 0.06)	0.7	0.6
[0.06, 0.08)	0.65	0.619
[0.08, 0.1)	0.632	0.615

Takeaway: confidence interval method works

Evaluation: precision on various cuts, genre

Clips extracted from	Overall
Audio-peaks	0.683
Audio-valleys	0.758
Visual-peaks	0.508
Visual-valleys	0.757

Evaluation: precision on various cuts, genre

Clips extracted from	Overall	Genre	Overall	Visual-peak
Audio-peaks	0.683	Action	0.678	0.264
Audio-valleys	0.758	Science Fiction	0.699	0.333
Visual-peaks	0.508	Thriller	0.678	0.382
Visual-valleys	0.757	Adventure	0.726	0.443
		Drama	0.660	0.520
		Fantasy	0.769	0.590
		Comedy	0.705	0.667
		Animation	0.798	0.667
		Family Film	0.760	0.722
		Romance	0.678	0.757
		Romantic Comedy	0.677	0.823

Evaluation: precision on various cuts, genre

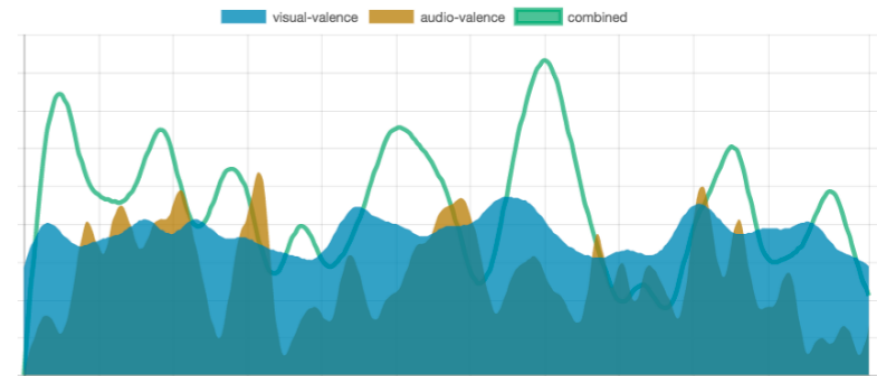
Clips extracted from	Overall	Genre	Overall	Visual-peak
Audio-peaks	0.683	Action	0.678	0.264
Audio-valleys	0.758	Science Fiction	0.699	0.333
Visual-peaks	0.508	Thriller	0.678	0.382
Visual-valleys	0.757	Adventure	0.726	0.443
		Drama	0.660	0.520
		Fantasy	0.769	0.590
		Comedy	0.705	0.667
		Animation	0.798	0.667
		Family Film	0.760	0.722
		Romance	0.678	0.757
		Romantic Comedy	0.677	0.823

Takeaways:

- Action, thriller-type movies have poor visual-peak precision
 - Flickr dataset doesn't contain images of guns, bodies, etc.
- **Need some way to globally condition on genre**

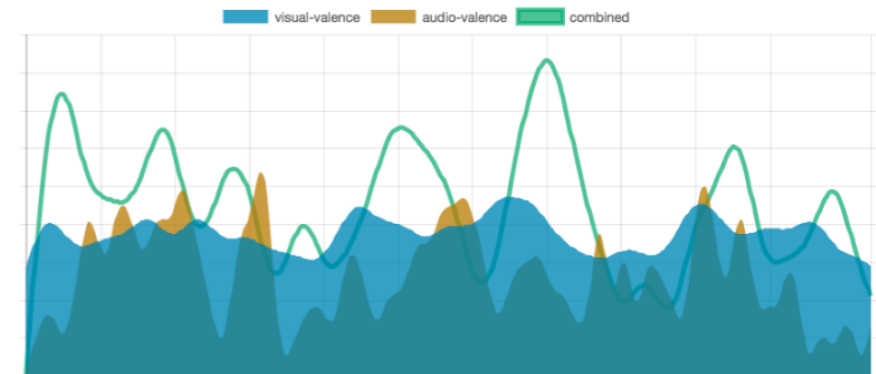
Combining audio and visual

Linear regression model to predict mean valence rating as assigned by annotators



Combining audio and visual

Linear regression model to predict mean valence rating as assigned by annotators

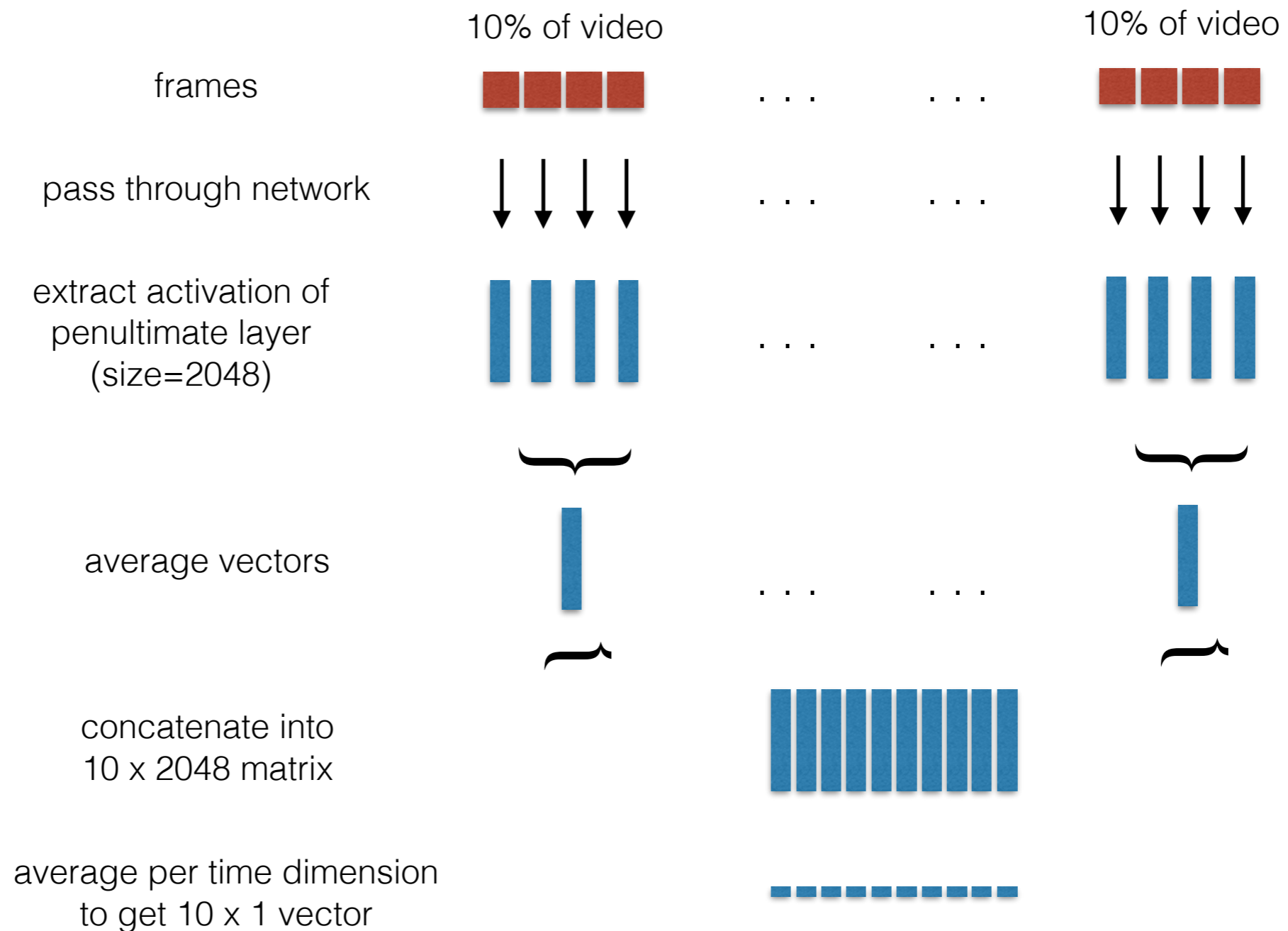


Features:

- | | |
|--------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Visual | <ol style="list-style-type: none">1. visual valence2. (visual valence) - (movie's mean visual valence)3. (max of movie's visual valence) - (visual valence)4. (visual valence) - (min of movie's visual valence)5. peakiness of visual valence |
| Audio | <ol style="list-style-type: none">6. audio valence7. (audio valence) - (movie's mean audio valence)8. (max of movie's audio valence) - (audio valence)9. (audio valence) - (min of movie's audio valence)10. peakiness of audio valence11. audio stddev |
| Other | <ol style="list-style-type: none">12. (relative) time in movie13. movie embeddings |

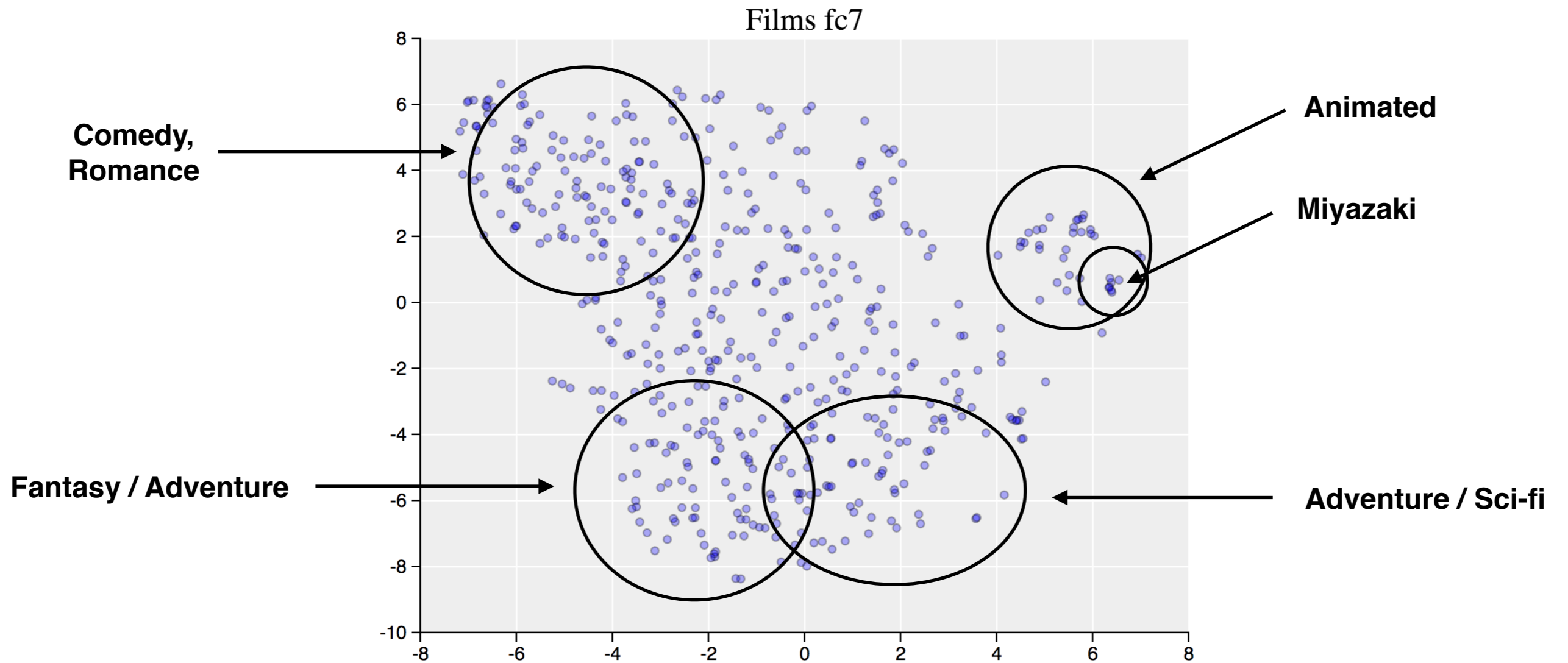
Combining audio and visual: movie embeddings

- Goal: create embeddings that capture emotional gestalt of movie, corresponds to genre
- Using model trained to predict adjective-noun label



Combining audio and visual: movie embeddings

TSNE of 10 x 2048 embeddings



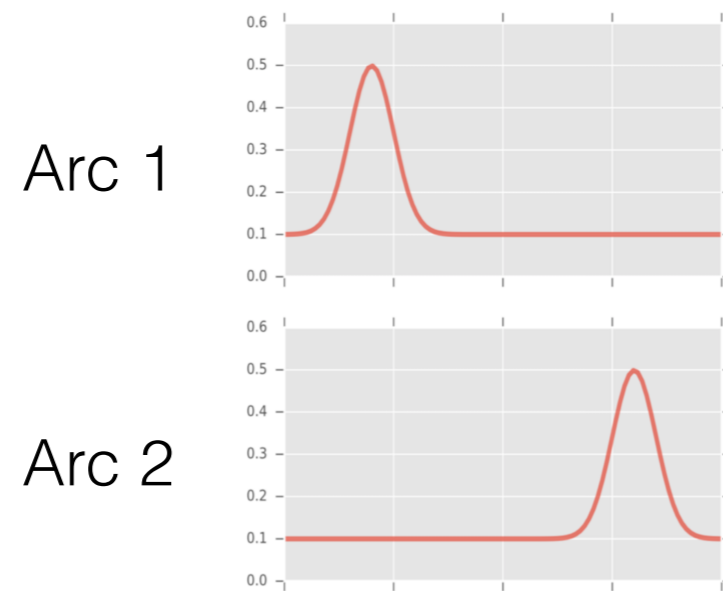
Accuracy of combined model

Accuracy measured by agreement in polarity between combined model and annotators' ratings

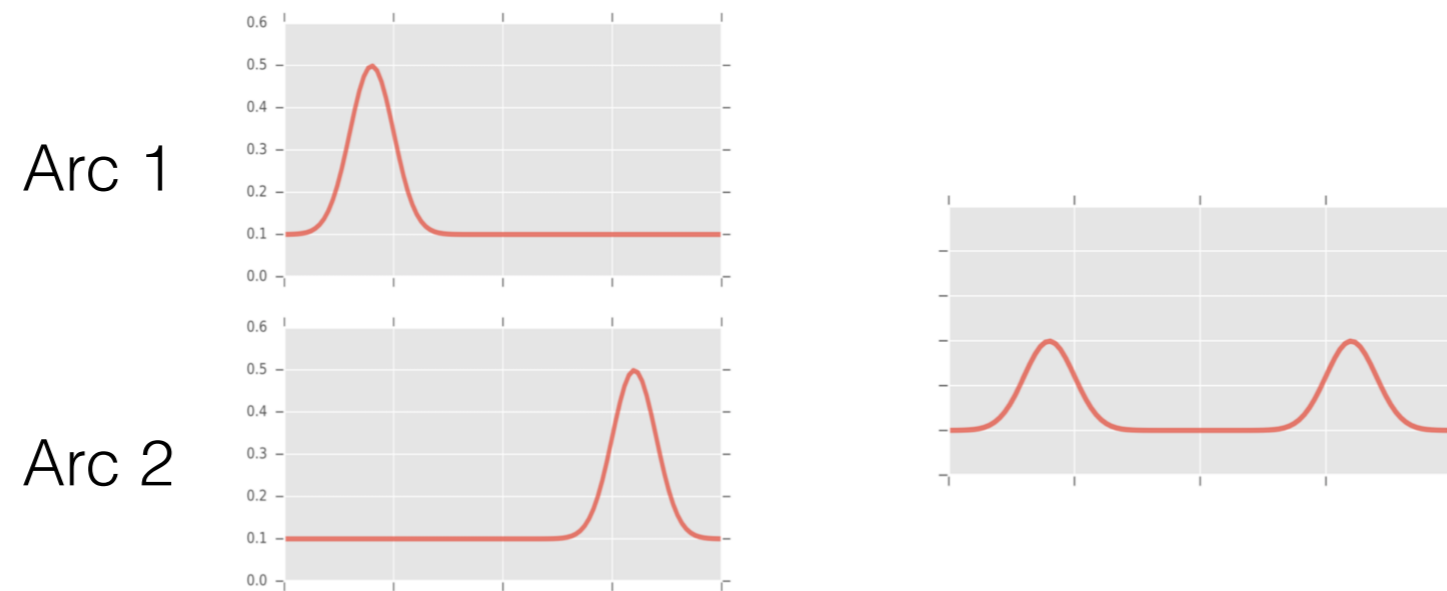
Feature Set	Accuracy
All features	0.894
No movie embedding	0.784
Audio only	0.712
Visual only	0.612

PART 2

Shape-based clustering: pathological example with k-means



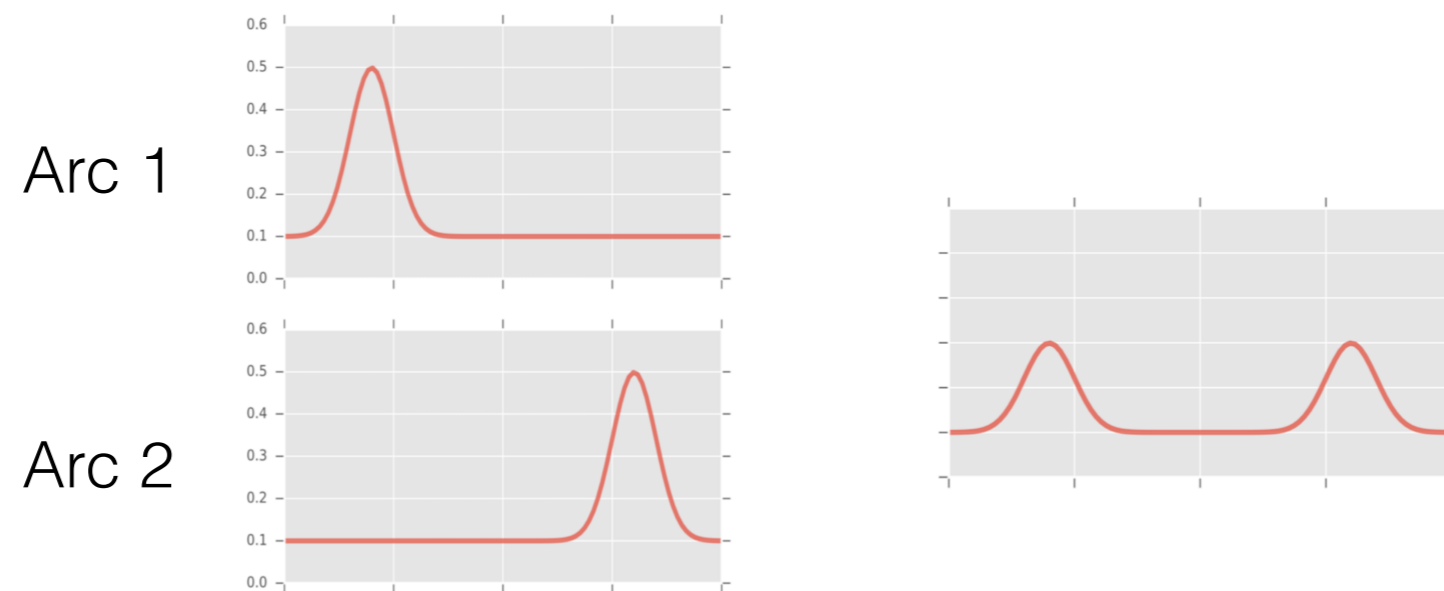
Shape-based clustering: pathological example with k-means



Problems:

1. Mean is a poor representation of cluster

Shape-based clustering: pathological example with k-means



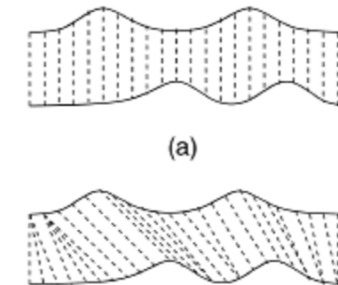
Problems:

1. Mean is a poor representation of cluster
2. Euclidean distance is a poor distance metric

Shape-based clustering: k-medoids + dynamic time warping

Fixing problems:

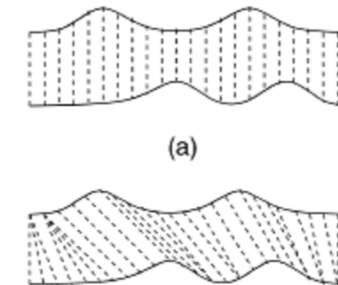
1. Mean is a poor representation of cluster
 - Use k-medoids instead of k-means
2. Euclidean distance is a poor distance metric
 - Use dynamic time warping (DTW)



Shape-based clustering: k-medoids + dynamic time warping

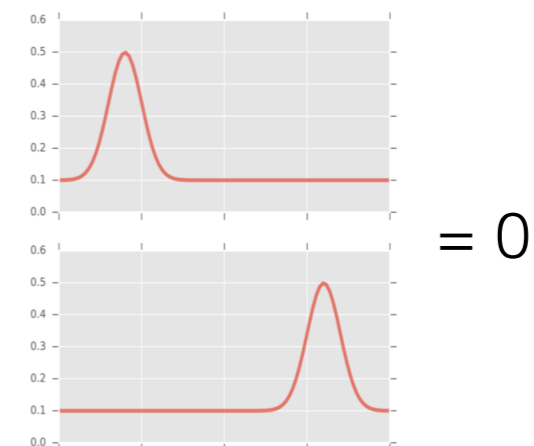
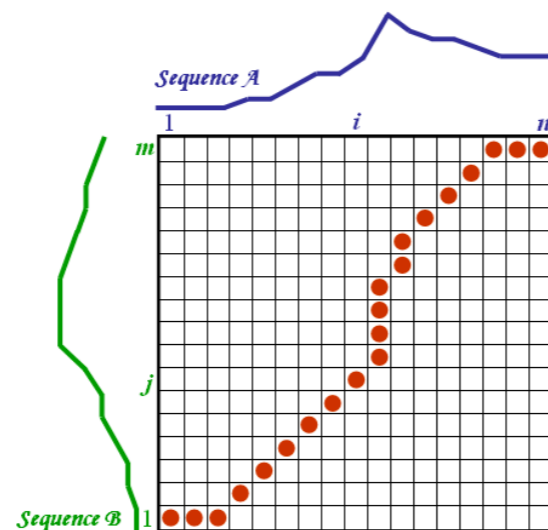
Fixing problems:

1. Mean is a poor representation of cluster
 - Use k-medoids instead of k-means
2. Euclidean distance is a poor distance metric
 - Use dynamic time warping (DTW)



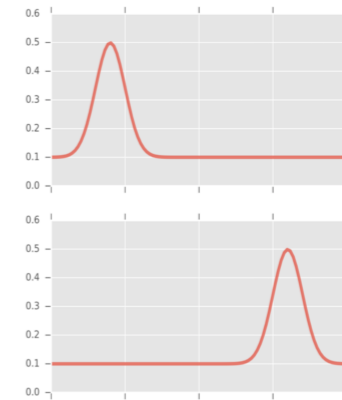
DTW:

- Given two time series A and B of length n , construct a $n \times n$ matrix M , where $M[i][j]$ contains the squared difference between A_i and B_j .
- The DTW distance between A and B is the shortest path through this matrix



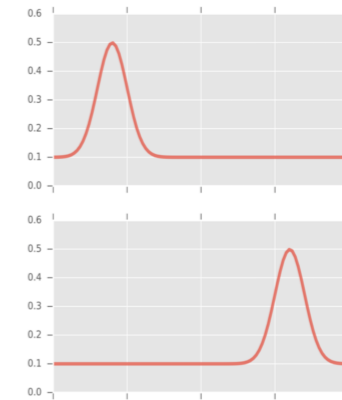
Shape-based clustering: DTW with Keogh lower bound

- But wait! These two arcs, while both characterized by a large peak, **may impact a viewer differently based on the timing of that peak.**
- So we want to **allow warping** (as provided by DTW), **but only to an extent**

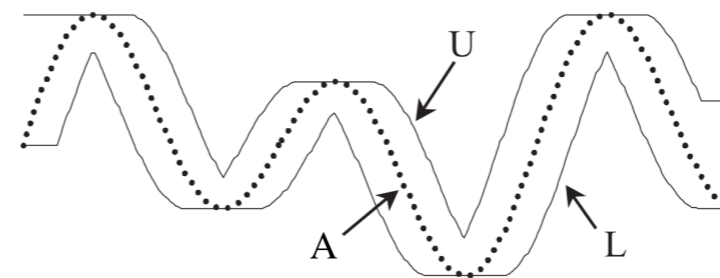
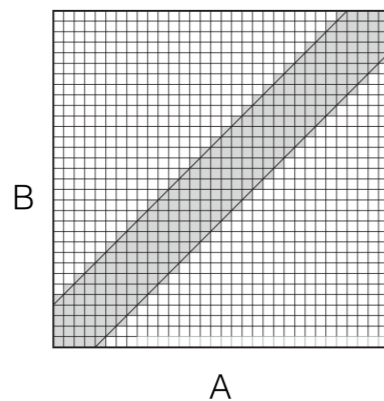


Shape-based clustering: DTW with Keogh lower bound

- But wait! These two arcs, while both characterized by a large peak, **may impact a viewer differently based on the timing of that peak.**
- So we want to **allow warping** (as provided by DTW), **but only to an extent**



- Therefore, use Keogh lower bound
 - Limit possible paths through M
 - Effectively creating 'warping window' around A defined by upper bound U and lower bound L
 - If B is within window, distance is 0



$$LB_{Keogh}(A, B) = \sqrt{\sum_{i=1}^n \begin{cases} (B_i - U_i)^2 & \text{if } B_i > U_i \\ (B_i - L_i)^2 & \text{if } B_i < L_i \\ 0 & \text{otherwise} \end{cases}}$$

Shape-based clustering results

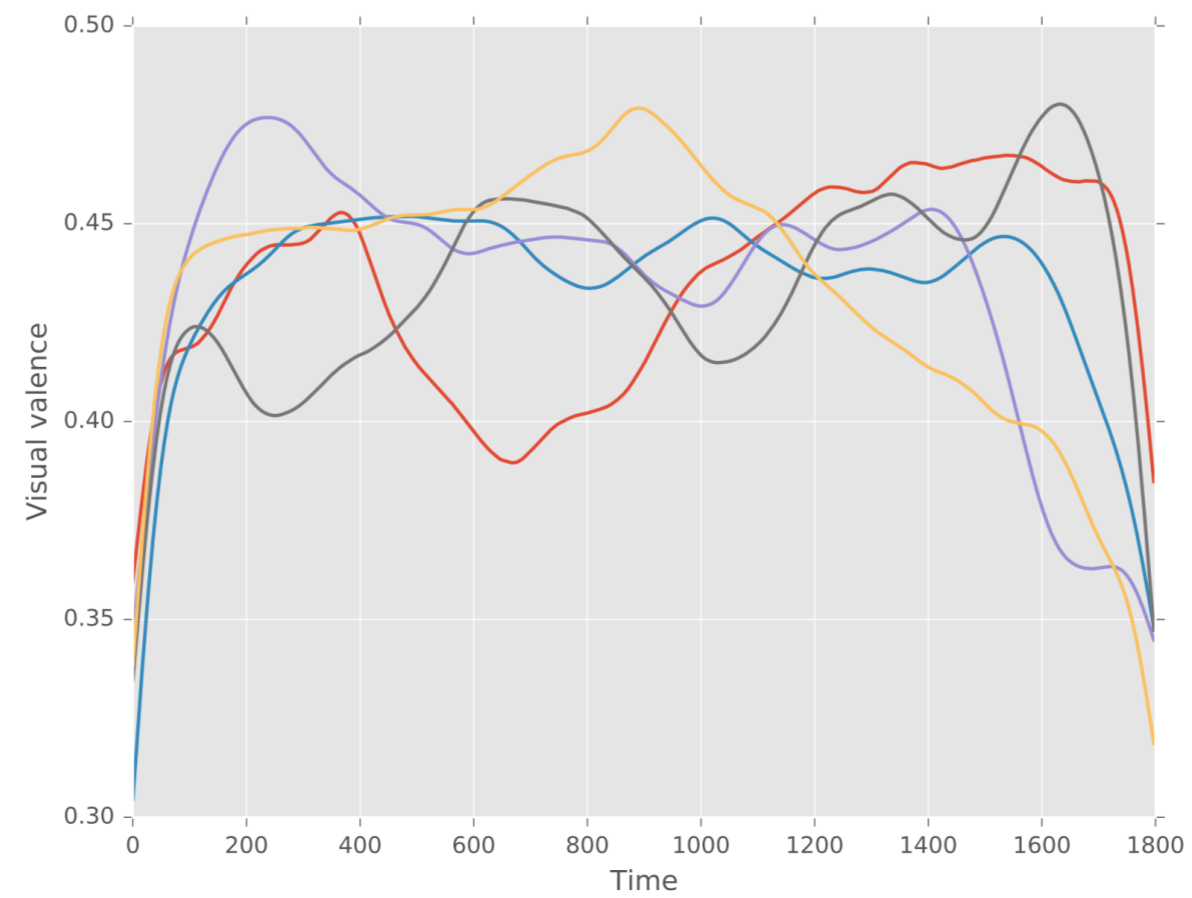
Two corpora:

1. ~500 Hollywood films
2. ~1500 Vimeo shorts from channel 'Short of the Week'

Shape-based clustering results

Two corpora:

1. ~500 Hollywood films
2. ~1500 Vimeo shorts from channel 'Short of the Week'

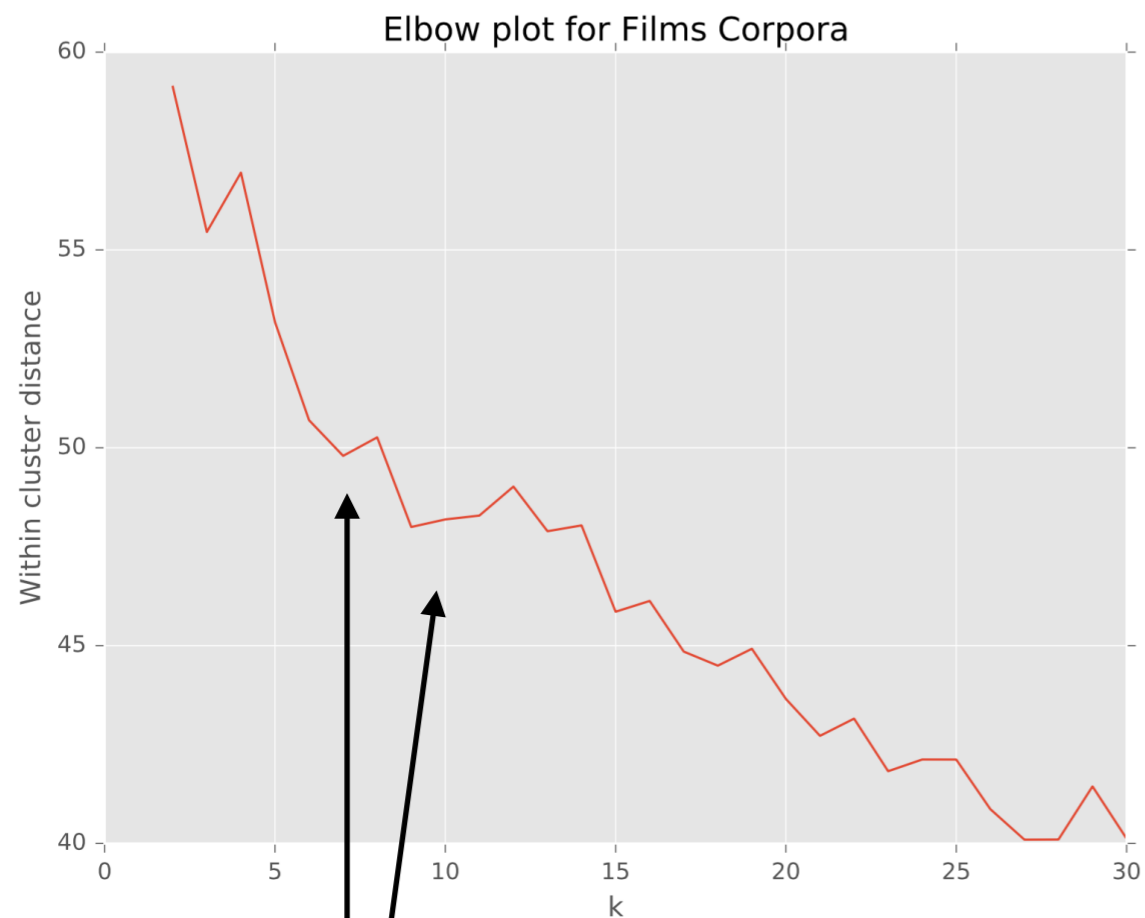


Shape-based clustering results

Two corpora:

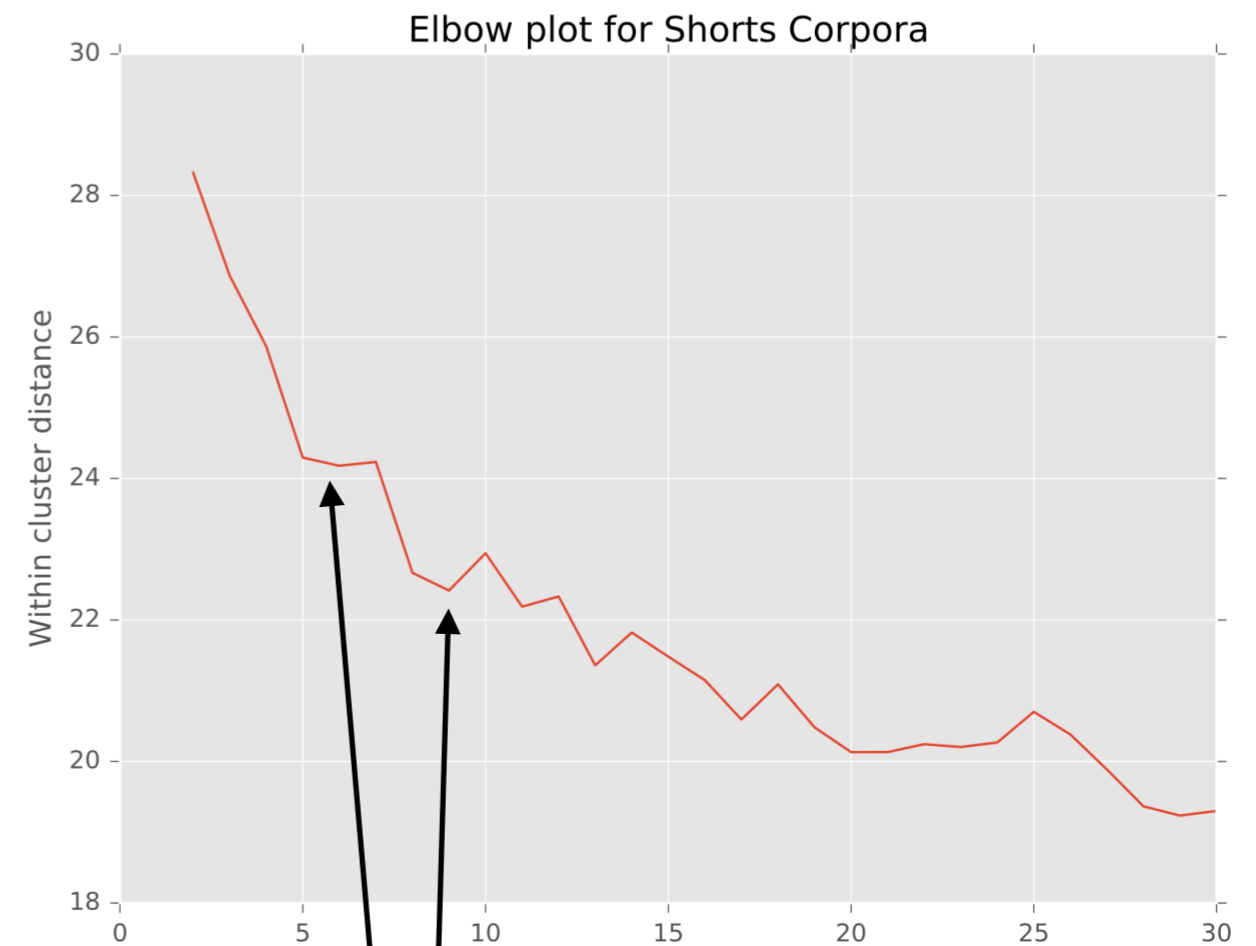
1. ~500 Hollywood films
2. ~1500 Vimeo shorts from channel 'Short of the Week'

Films



'optimal' k \approx 6, 10

Shorts



'optimal' k \approx 5, 9

Engagement analysis:
predicting the number of Vimeo comments

Engagement analysis: predicting the number of Vimeo comments

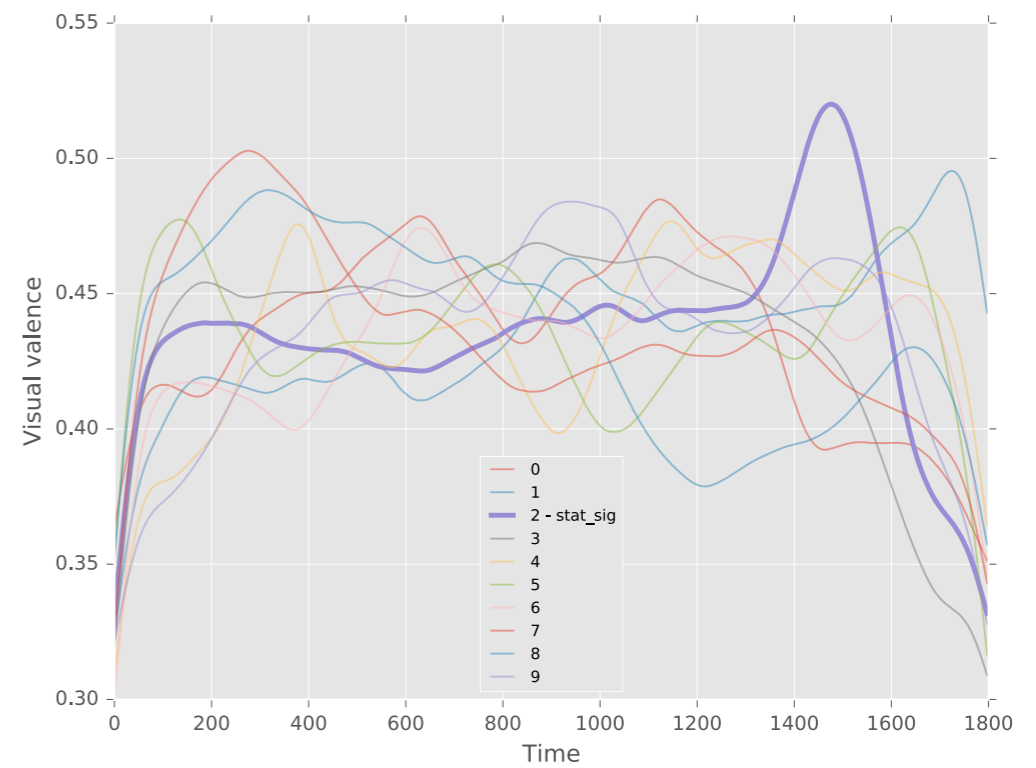
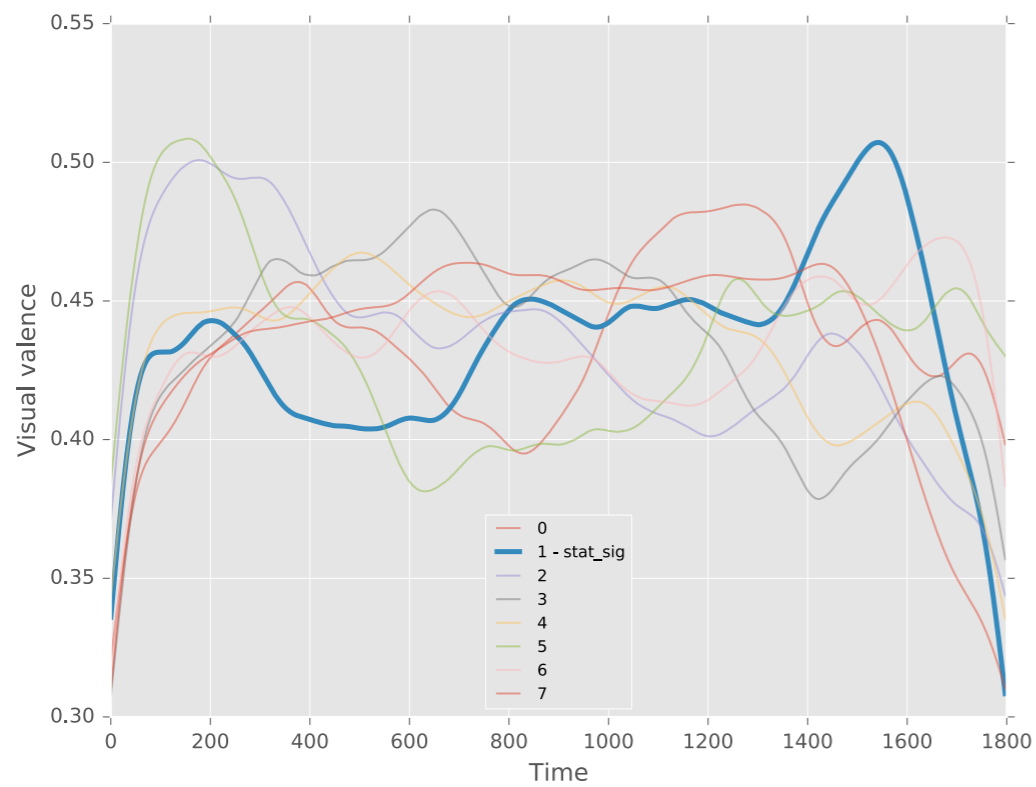
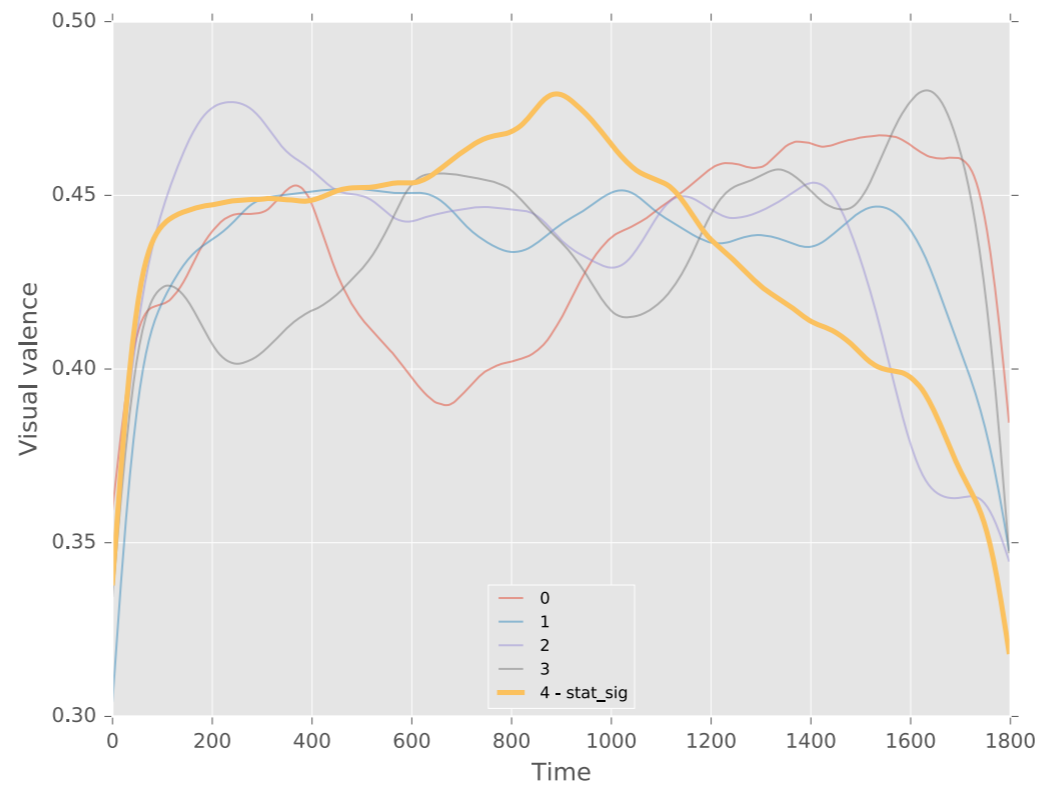
Feature	Coefficient	p-value
Intercept
Duration	-0.1180	0.001
Year	-0.1720	0.000
Month	-0.0688	0.048
Hour
Author_num_comments

Engagement analysis: predicting the number of Vimeo comments

	Feature	Coefficient	p-value
	Intercept
	Duration	-0.1180	0.001
	Year	-0.1720	0.000
	Month	-0.0688	0.048
	Hour
	Author_num_comments
	Cluster_A
	Cluster_B
	Cluster_C
	Cluster_D
	Cluster_E	0.1948	0.011

Which family of arcs does this movie belong to

Engagement analysis: stat-sig clusters — receive more comments



Conclusion

- Image and audio models to create emotional arcs
- Datasets
 - Spotify and movie clips are publicly available
- Method for clustering time series based on shape
- Showing through a (small) analysis that these families of emotional arcs can matter