# E-Musical

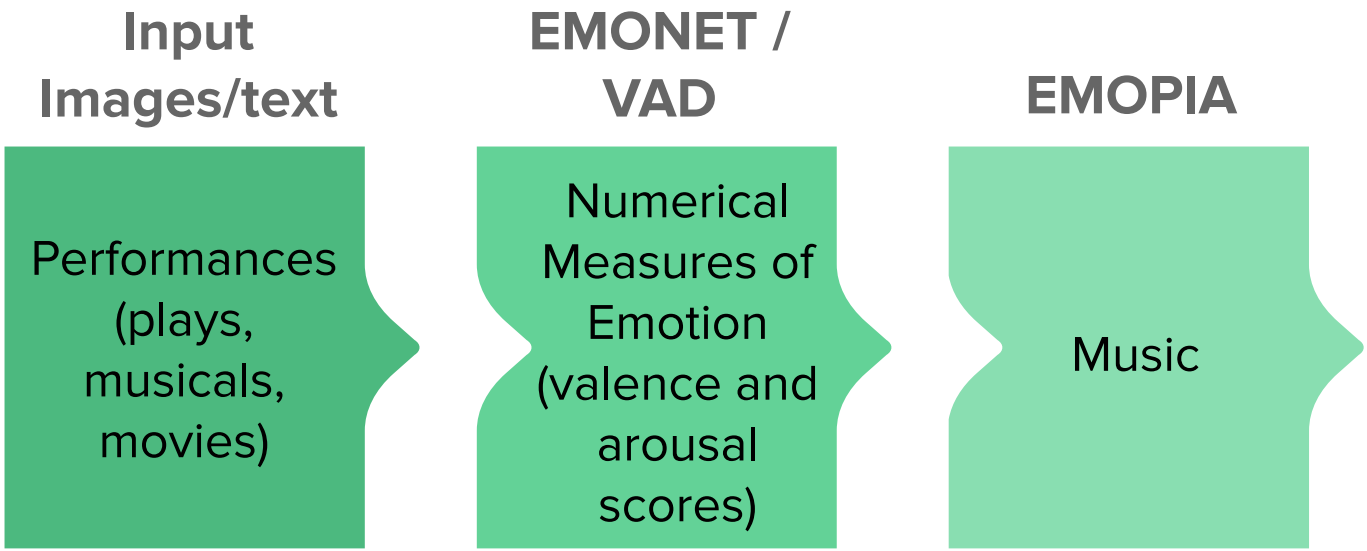
Art and Machine Learning: Project 3
Group 16
Nikitha Murikinati, Aarthi Ramsundar, Qi Xuan Teo, Audrey Zhang
March, 2022

# Project concept

"Where words fail, music speaks" -- *Hans Christian Andersen*

- Music is one of the most universal ways through which human emotion is captured and expressed
- Our project seeks to leverage the emotional landscape of performative pieces to generate music
- Our inspiration is to generate soundtracks for videos or theme music for fictional characters, using video/text inputs

# Pipeline


Input
Images/text
EMONET /
VAD
EMOPIA
Performances (plays, musicals, movies)
Numerical Measures of Emotion (valence and arousal scores)
Music

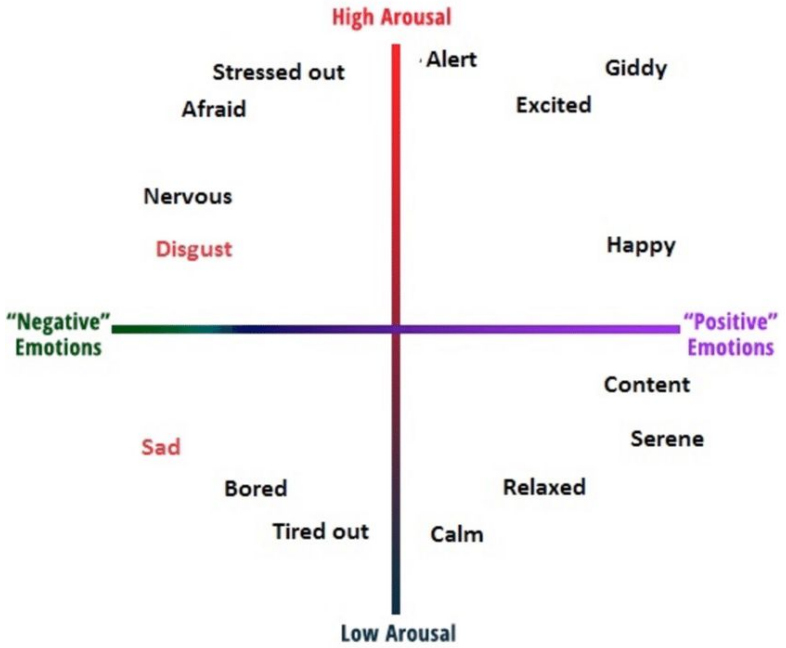
# Two-Dimensional Representation of Emotion

Valence:

*The level of pleasantness in the emotion, ranging from unpleasant (negative) to pleasant (positive)*

Arousal:

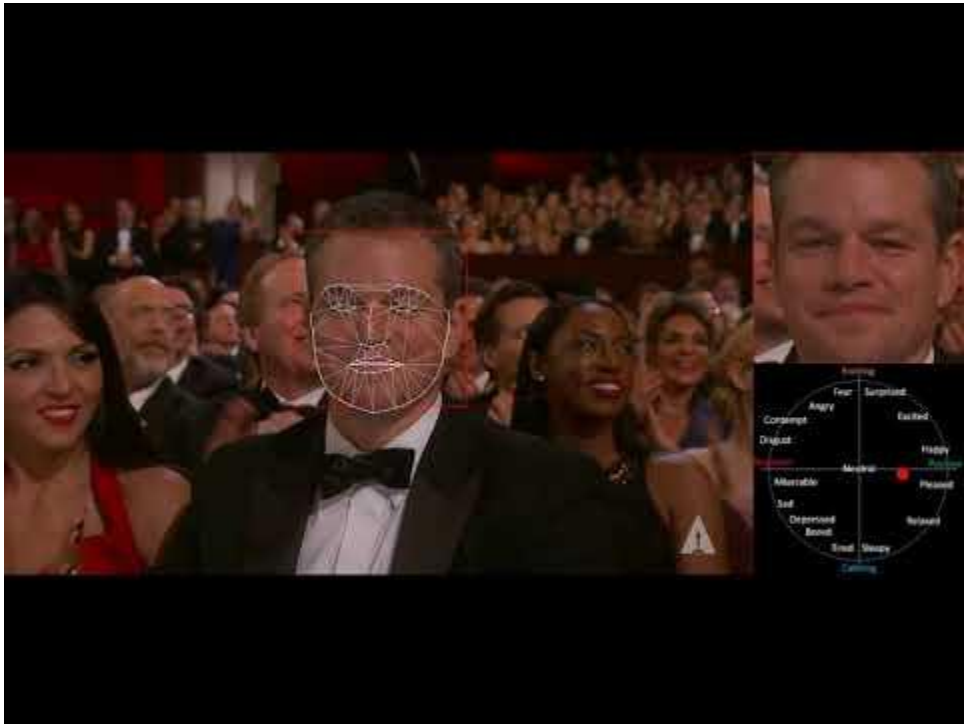
*The level of intensity of emotion, ranging from calm (low) to excited (high)*




Online Measuring of Available Resources - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-two-dimensions-of-emotions-Valence-negative-positive-and-arousal-low-high-Every_fig1_318031044 [accessed 29 Mar, 2022]

# EMONET

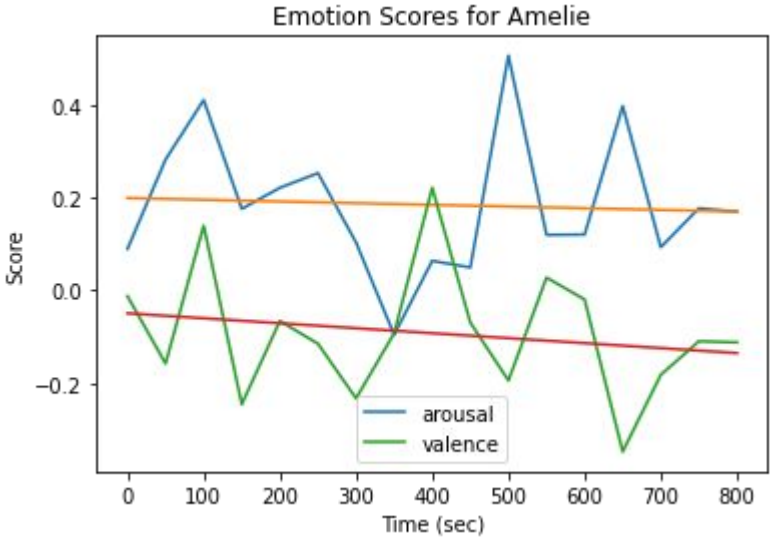
“Estimation of continuous valence and arousal levels from faces in naturalistic conditions” -- *Toisoul et al., 2021*

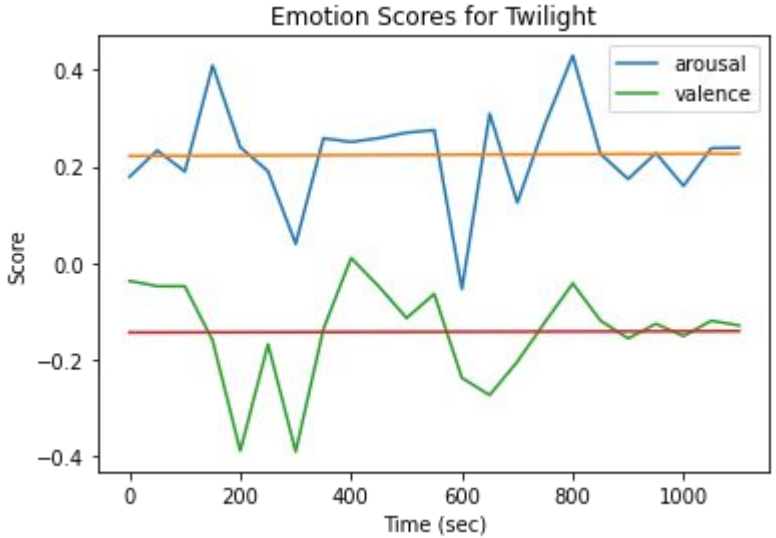



Code adaptations:

- Output predicted valence and arousal scores instead of an emotion category
- Intake a sequence of images and output a corresponding sequence of valence and arousal scores, in order
- Downsample the image sequence as needed (instead of frame-by-frame predictions) to limit output size

# Arousal and Valence Scores



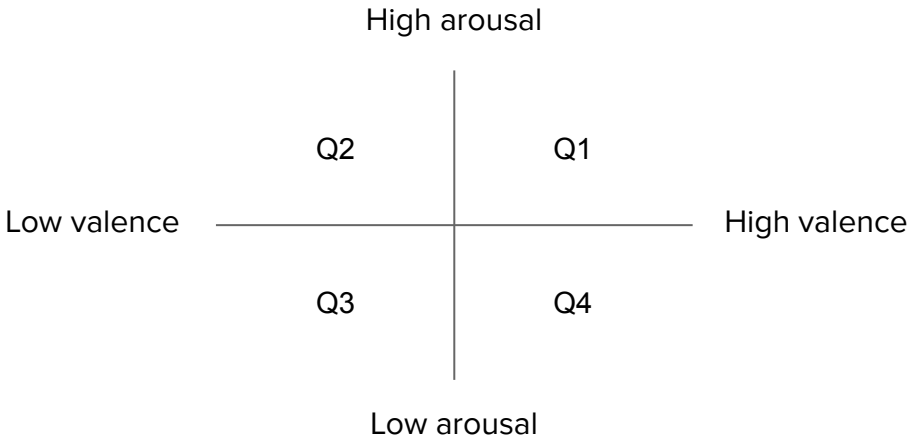
# Text-Based VAD Analysis

- Modified from Atmaja et. al. (2019)
- Tokenizes sentences and words -
  - Compares words with ANEW database to obtain Valence, Arousal, Dominance values
  - Looks ahead for negation (e.g. ‘not’, ‘no’), and multiplies VAD values by -1 if found
- VAD values are averaged across subsections (song/scene) to obtain EMOPIA input
- How to obtain text?
  - Youtube transcripts - provides both text, as well as timestamps to aid in alignment

# EMOPIA

"EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation" -- *Hung et al., 2021*

*Allows for emotion-conditioned symbolic music generation using a transformer trained on the EMOPIA dataset*




Code adaptations:

- Combine a sequence of valence and arousal values to generate an emotion classification tag for a performance rather than looking at a single valence/arousal value pair
- Combined values using different methods (average, max, min, median) for experimentation

# Sample results: Twilight (movie scene)


Image generated emotions:

- High arousal, low valence (Q2)

Text generated emotions:

- Low arousal, high valence (Q4)

# Sample results: Amelie (movie scene)


Image generated emotions:

- High arousal, high valence (Q1)

Text generated emotions:

- Low arousal, high valence (Q4)

# Reflections

- A lot of image-generated emotion values landed in quartile 2 (low valence, high arousal ➜ exciting and unpleasant scenes):
  - Potentially an artifact of using average scores across a scene
  - Could also be a result of using centercrop to get images to the correct size
  - There are multiple actors in a scene, who may have different expressions that ‘cancel out’ when the algorithm is predicting overall emotion
  - Emotions in performances are also expressed in movement/scenery/backdrop/color, which get lost when using image stills and looking at human facial expressions only
- Image-generated scores and text-generated scores sometimes differ
  - Different average valence and arousal scores for text and image
  - Full emotional weight of the scene can not be captured from one single modality
  - Multimodal methods (text, audio, images) may be required to combine input from different modalities into a single emotion score.

# Thank you!

# References

B.T. Atmaja, K. Shirai, M. Akagi, Deep Learning-based Categorical and
Dimensional Emotion Recognition for Written
and Spoken Text, International Seminar on Science and Technology, Surabaya - Indonesia,
2019.

Hung, Hsiao-Tzu, et al. "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation." *arXiv preprint arXiv:2108.01374* (2021).

Toisoul, Antoine, et al. "Estimation of continuous valence and arousal levels from faces in naturalistic conditions." *Nature Machine Intelligence* 3.1 (2021): 42-50.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior research methods, 45(4), 1191-1207.