Language Technologies Institute

Carnegie Mellon University

Multimodal AI: Understanding Human Behaviors

Louis-Philippe (LP) Morency

MultiComp Lab

PhD students: **Chaitanya Ahuja, Volkan Cirik, Paul Liang, Hubert Tsai, Alexandria Vail, Torsten Wörtwein and Amir Zadeh**

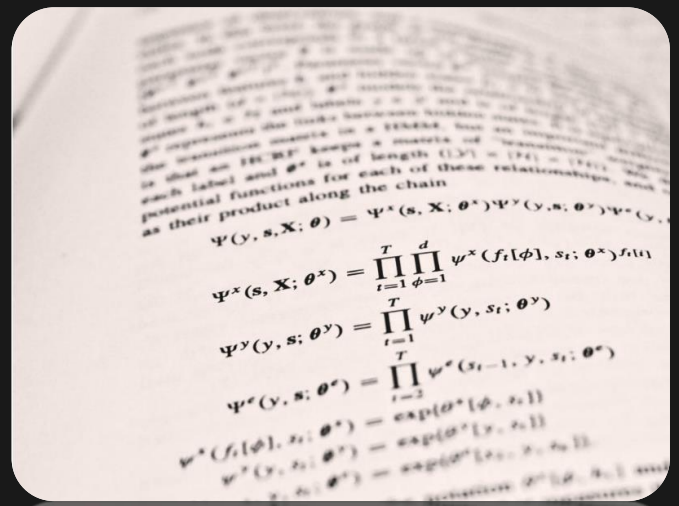
Postdoctoral researcher: **Jeffrey Girard**

Lab coordinator: **Nicole Siverling**

Project assistant: **John Friday**

MultiComp Lab

Communication Dynamics



$$\Psi(\mathbf{y}, \mathbf{s}, \mathbf{X}; \boldsymbol{\theta}) = \Psi^x(\mathbf{s}, \mathbf{X}; \boldsymbol{\theta}^x)\Psi^y(\mathbf{y}, \mathbf{s}; \boldsymbol{\theta}^y)\Psi^c(\mathbf{y}, \mathbf{s}; \boldsymbol{\theta}^c)$$

$$\Psi^x(\mathbf{s}, \mathbf{X}; \boldsymbol{\theta}^x) = \prod_{t=1}^{T}\prod_{\phi=1}^{d}\psi^x(f_t[\phi], s_t; \boldsymbol{\theta}^x)$$

$$\Psi^y(\mathbf{y}, \mathbf{s}; \boldsymbol{\theta}^y) = \prod_{t=1}^{T}\psi^y(\mathbf{y}, s_t; \boldsymbol{\theta}^y)$$

$$\Psi^c(\mathbf{y}, \mathbf{s}; \boldsymbol{\theta}^c) = \prod_{t=2}^{T}\psi^c(s_{t-1}, \mathbf{y}, s_t; \boldsymbol{\theta}^c)$$

Multimodal AI

Mental Health

Multimodal AI Technologies

Robots

Virtual Humans

Ubiquitous

Mobile

Wearable

Online

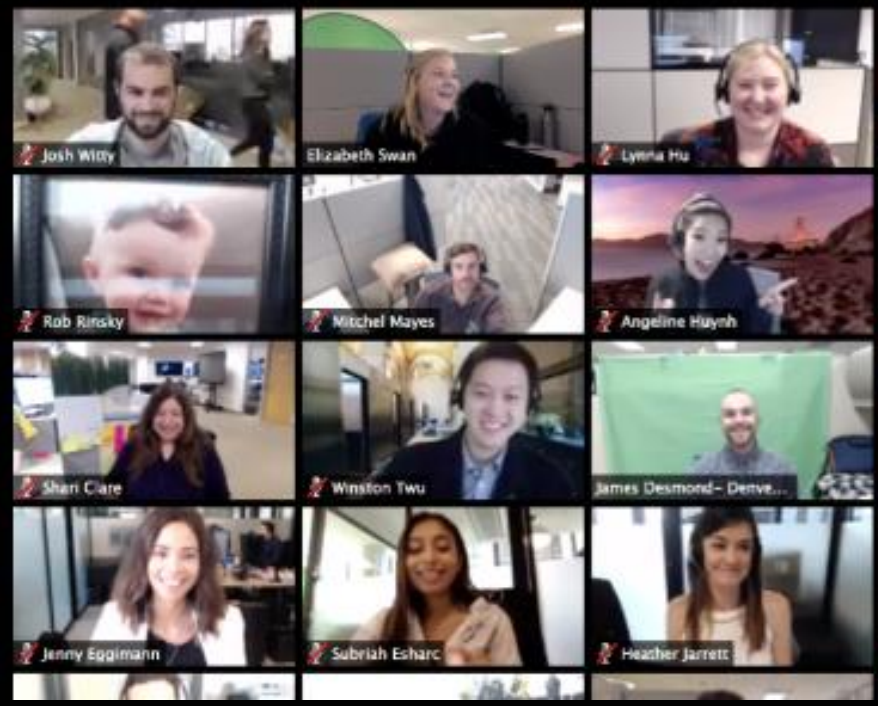
Multimodal AI Technologies

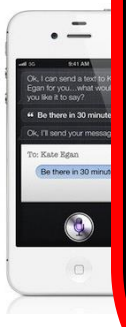
Robots

Virtual Humans

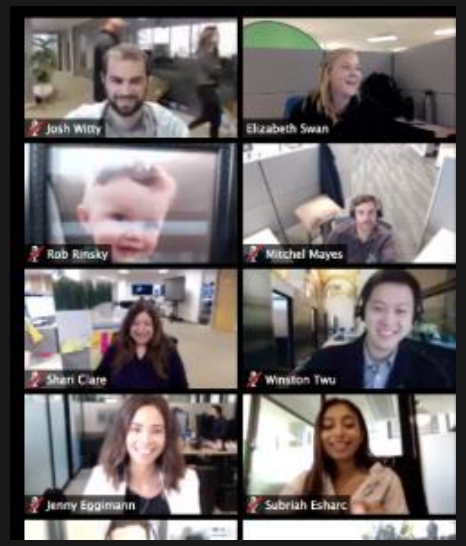
Ubiquitous

Video Conferencing





Multimodal Communicative Behaviors



Verbal

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Vocal

- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual

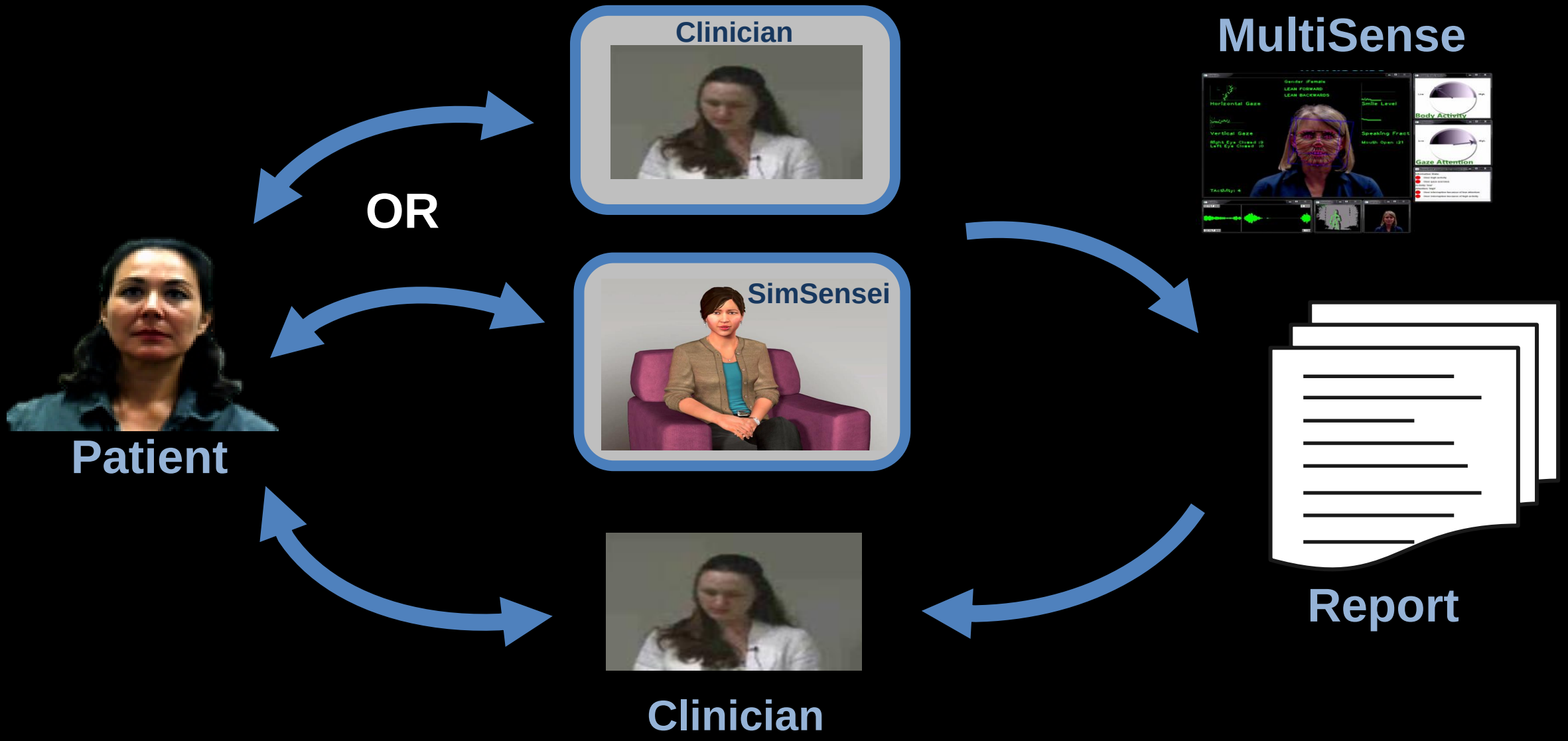
- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Mental Health by the Numbers

1 in 4 adults

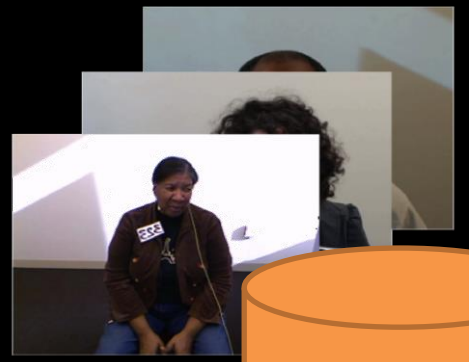
#1 cause of disability

Multimodal AI for Mental Health Assessment



Behavioral Indicators of Psychological Distress

Distress Assessment Interview Corpus



Data

Dictionary of Multimodal Behavior Markers

[IVA 2011, FG 2013, ICASSP 2013, 2015, ACII 2013, 2017, Interspeech 2013, 2017, ICMI 2013, 2014, 2018]

Psychosis

- Speech disfluency
- Language structure
- Gaze patterns

…

Depression

- Smile dynamics
- Gaze aversion
- Vowel space

…

Suicidal Ideation

- Lexical markers
- Voice quality
- Prosodic cues

…

HARVARD

UNIVERSITY OF PITTSBURGH

UPMC

COLUMBIA UNIVERSITY

Scientific Discoveries

① Smile Dynamics - Behavior Indicators

Depressed vs Non-depressed

Number of smiles	=	
Smile duration	−	+
Smile intensity	−	+

Surprising!

S. Scherer, G. Stratou, J. Boberg, J. Gratch, A. Rizzo and L.-P. Morency. Automatic Behavior Descriptors for Psychological Disorder Analysis. IEEE Conference on Automatic Face and Gesture Recognition, 2013

Scientific Discoveries

② Negative Expressions - Behavior Indicators

PTSD vs Non-PTSD

Overall population =

Men only + −

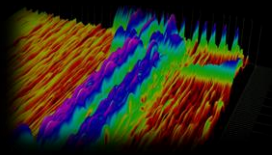
Women only − +

Opposite!

G. Stratou, S. Scherer, J. Gratch and L.-P. Morency. Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences. International Conference on Affective Computing and Intelligent Interaction, 2013

Scientific Discoveries

③ Speech Patterns - Behavior Indicators



	Suicidal	vs	Non-suicidal
First person pronouns (e.g., me, my, mine, I)	+		−

	Repeater	vs	Non-repeater
Voice tenseness	−		+

Important!

V. Venek, S. Scherer, L.-P. Morency, A. Rizzo and J. Pestian, Adolescent Suicidal Risk Assessment in Clinician-Patient Interaction, IEEE Transactions on Affective Computing, January 2016

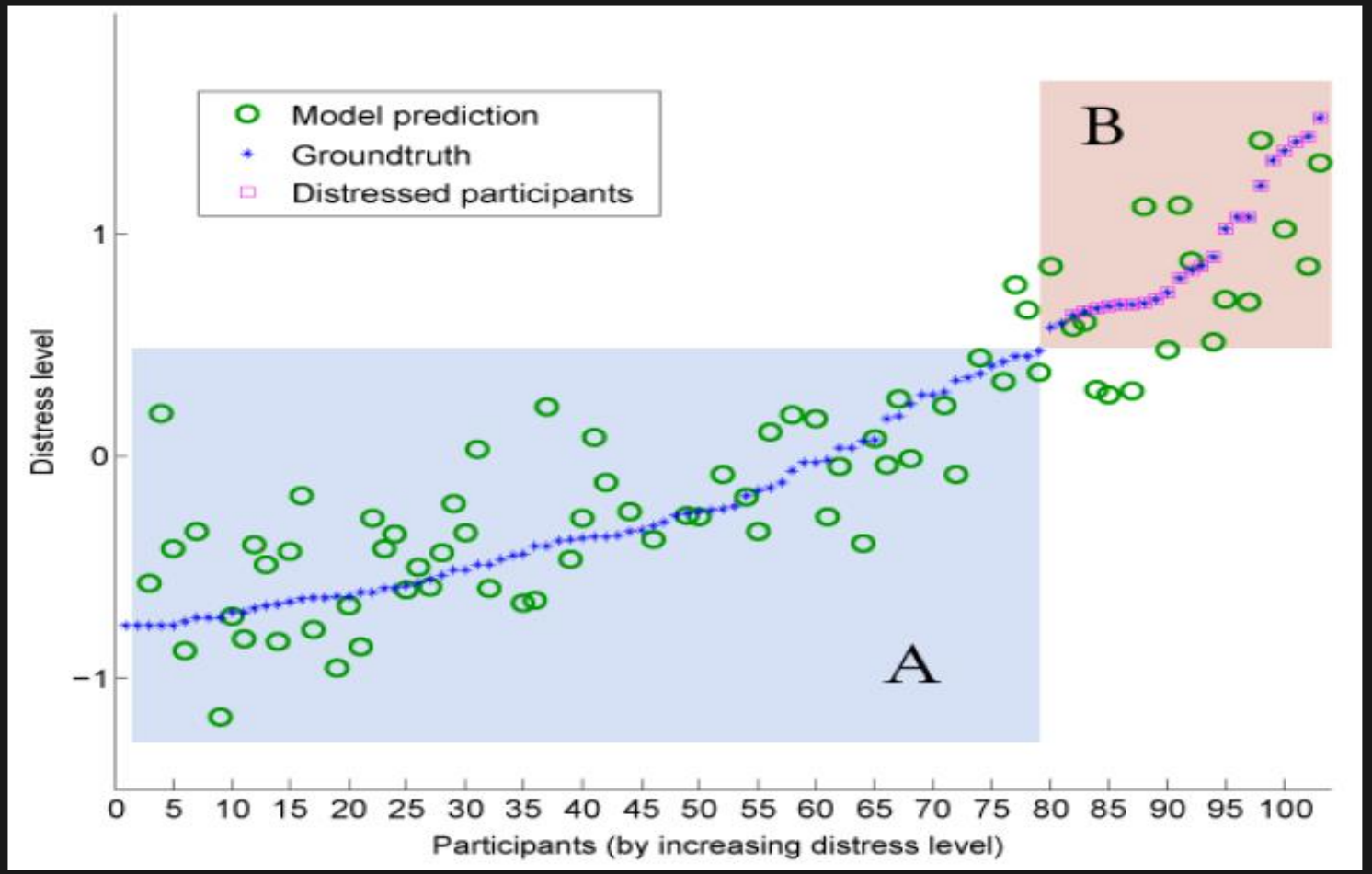
Multimodal AI: Automatic Distress Level Prediction

Verbal

We saw the yellow dog

Vocal

Visual



Multimodal Machine Learning

Verbal

We saw the yellow dog

Visual

Vocal

Multimodal Machine Learning

Disorders
- Depression
- Distress
- Autism

Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration

http://multicomp.cs.cmu.edu/social-iq

Social-IQ: A QA Benchmark for Artificial Social Intelligence

[CVPR 2019]

Social Intelligence

1250 Videos

7500 Questions

30k Correct Answers

22.5k Incorrect Answers

Social Gathering

Intimate Moments

Debates

Discussion

Opinion Sharing

TV Shows

Social Phenomena

*a. Are people **getting along**?*
*b. How is the **atmosphere in the room**?*

Mental State and Attitude

*a. Was the man **hurt when insulted**?*
*b. Was the woman **brave**?*

Multimodal Behavior

*a. How did the man **show his discontent**?*
*b. How did the woman **respond to the rude person**?*

Core Challenges in Multimodal Machine Learning

- Representation
- Alignment
- Fusion
- Translation
- Co-learning

Multimodal Machine Learning: A Survey and Taxonomy

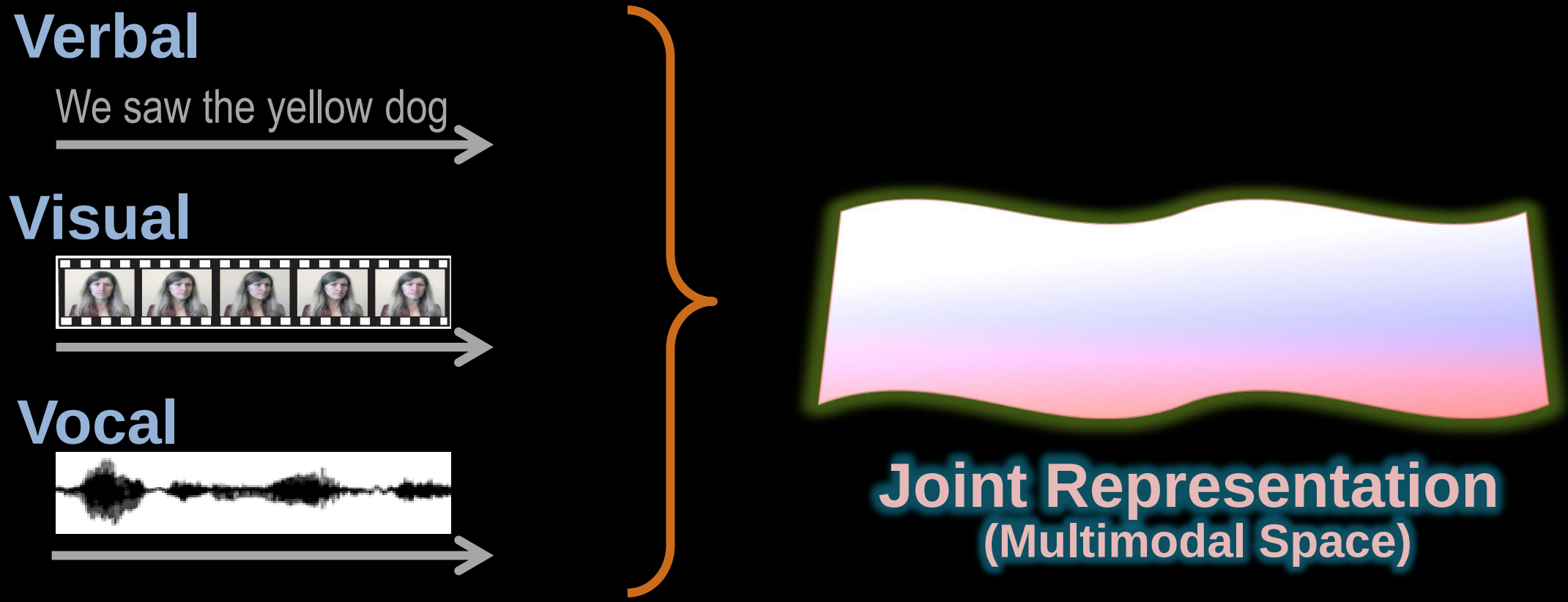
By Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

https://arxiv.org/abs/1705.09406

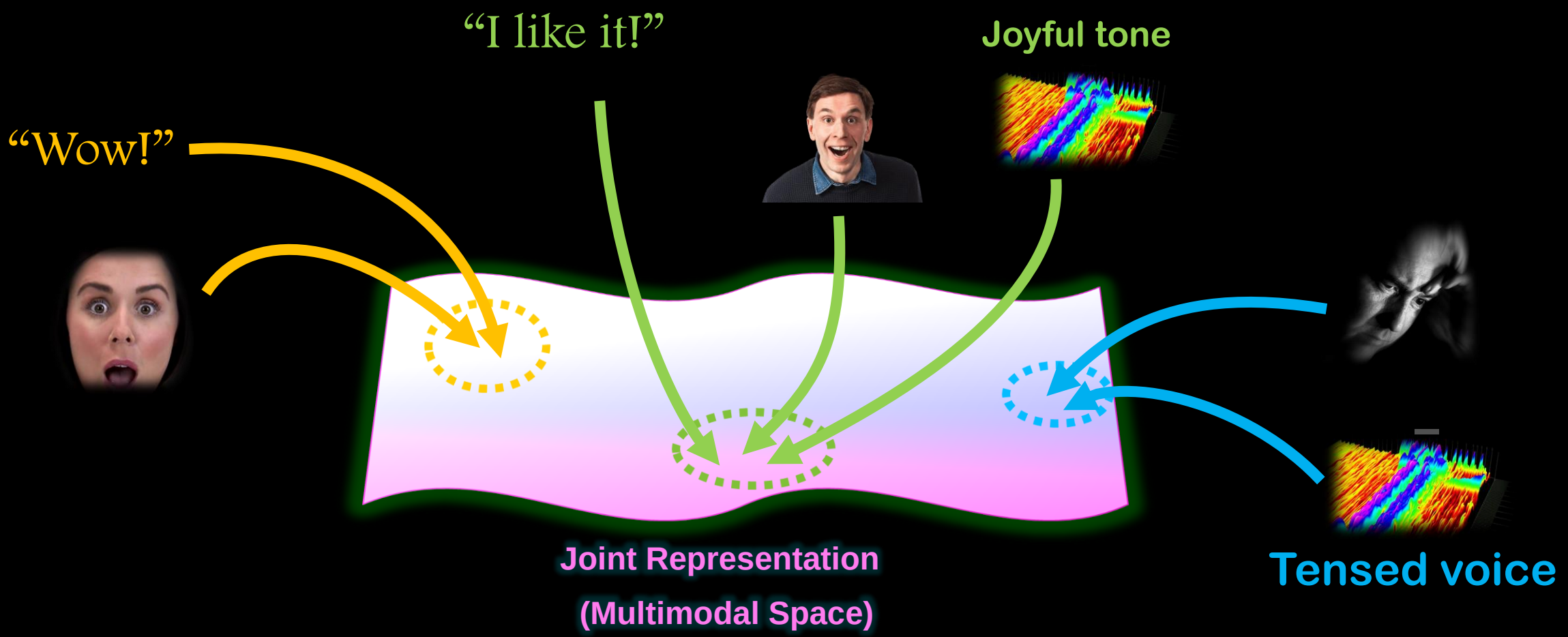
- ☑ **5 core challenges**
- ☑ **37 taxonomic classes**
- ☑ **253 referenced citations**

Core Challenge 1: Multimodal Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the **complementarity** and **redundancy**.



Joint Multimodal Representation



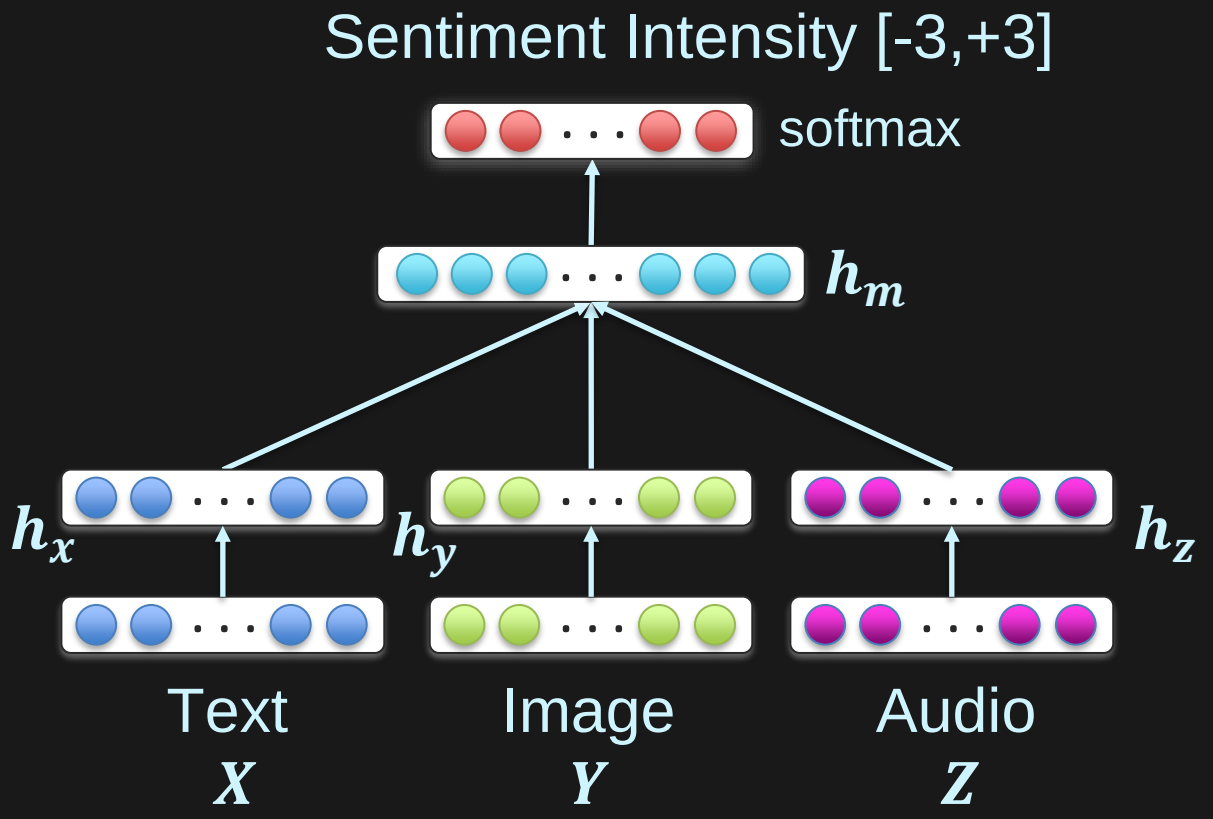
Multimodal Sentiment Analysis

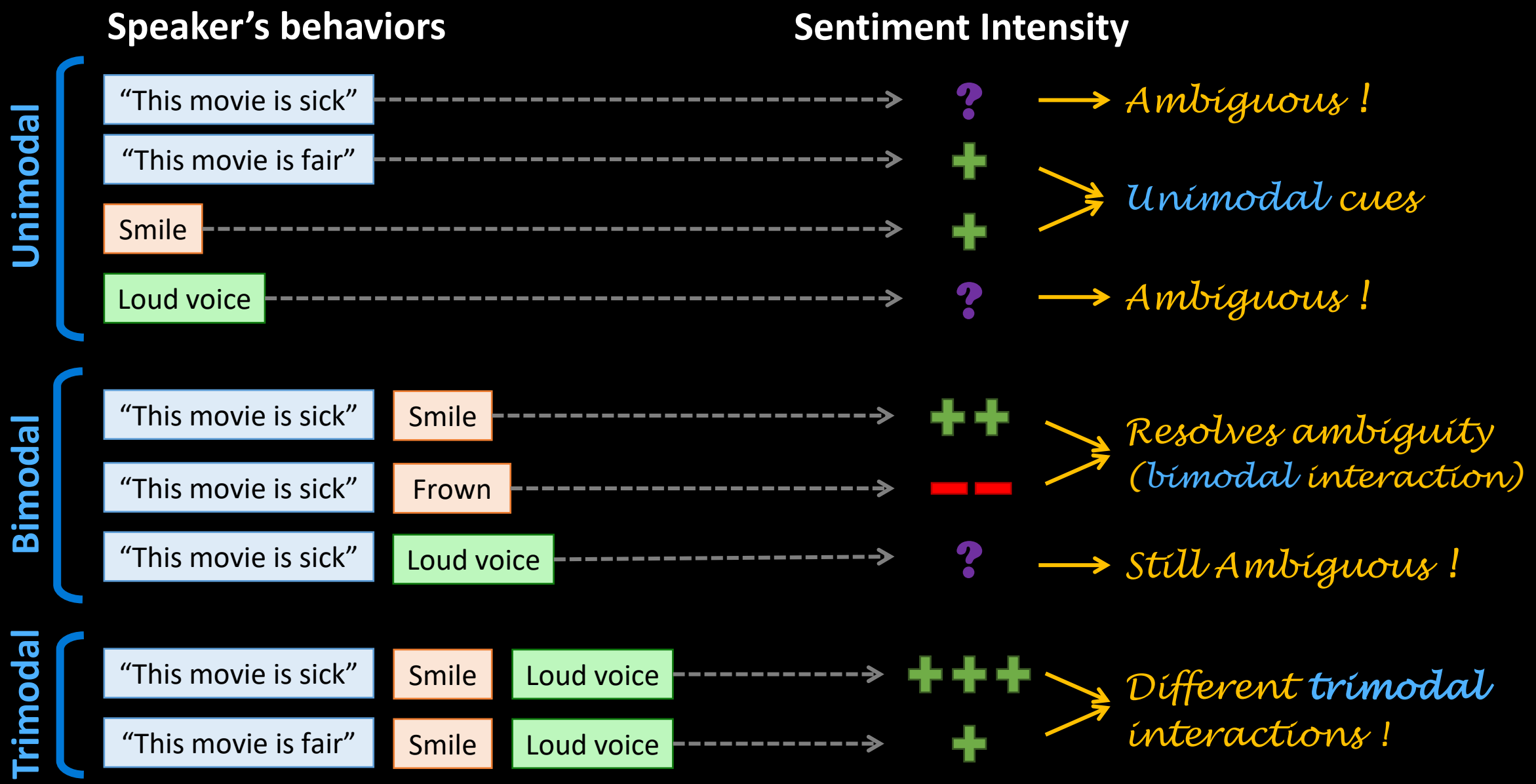
MOSI dataset (Zadeh et al, 2016)

- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

Multimodal joint representation:

$$h_m = f\left(W \cdot [h_x, h_y, h_z]\right)$$




Speaker's behaviors
Sentiment Intensity
Unimodal
"This movie is sick"
"This movie is fair"
Smile
Loud voice
?
+
+
?
Ambiguous !
Unimodal cues
Ambiguous !
Bimodal
"This movie is sick"
Smile
"This movie is sick"
Frown
"This movie is sick"
Loud voice
++
−−
?
Resolves ambiguity
(bimodal interaction)
Still Ambiguous !
Trimodal
"This movie is sick"
Smile
Loud voice
"This movie is fair"
Smile
Loud voice
+++
+
Different trimodal
interactions !

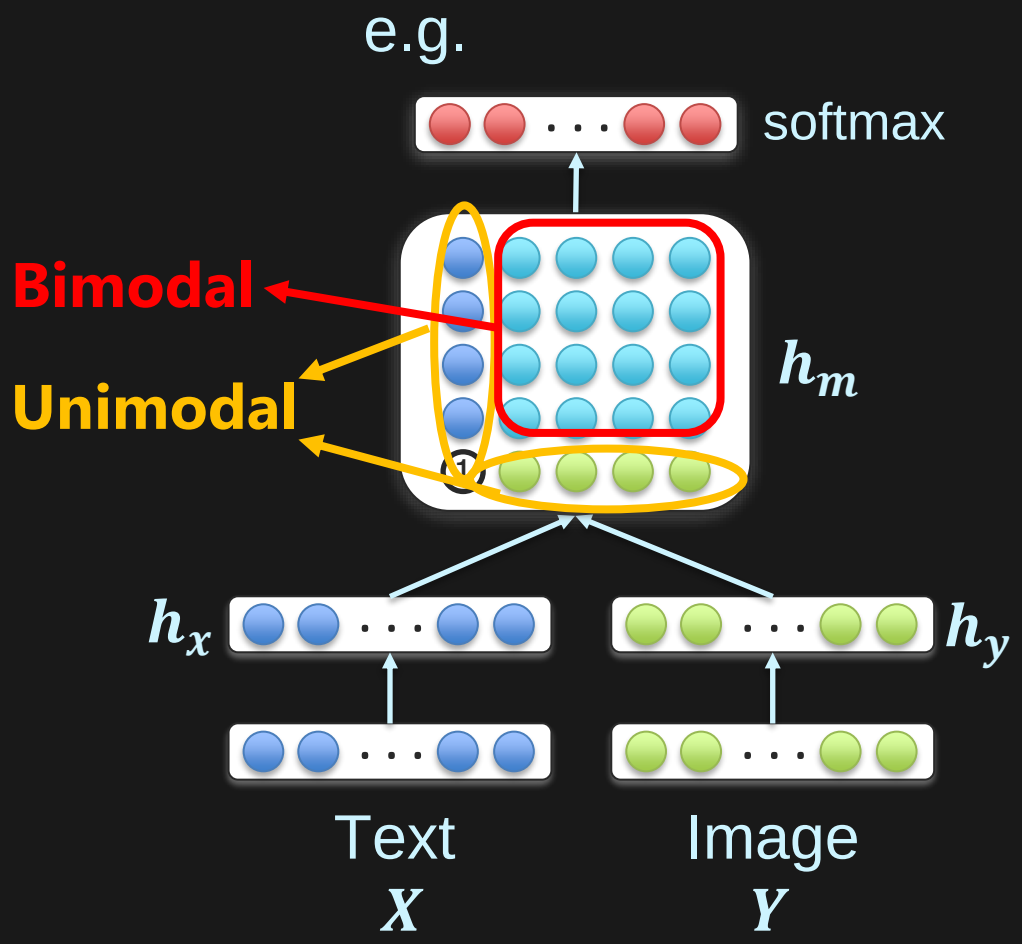
Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} = \begin{bmatrix} h_x & h_x \otimes h_y \\ 1 & h_y \end{bmatrix}$$

Important !

[Zadeh, Jones and Morency, **EMNLP 2017**]



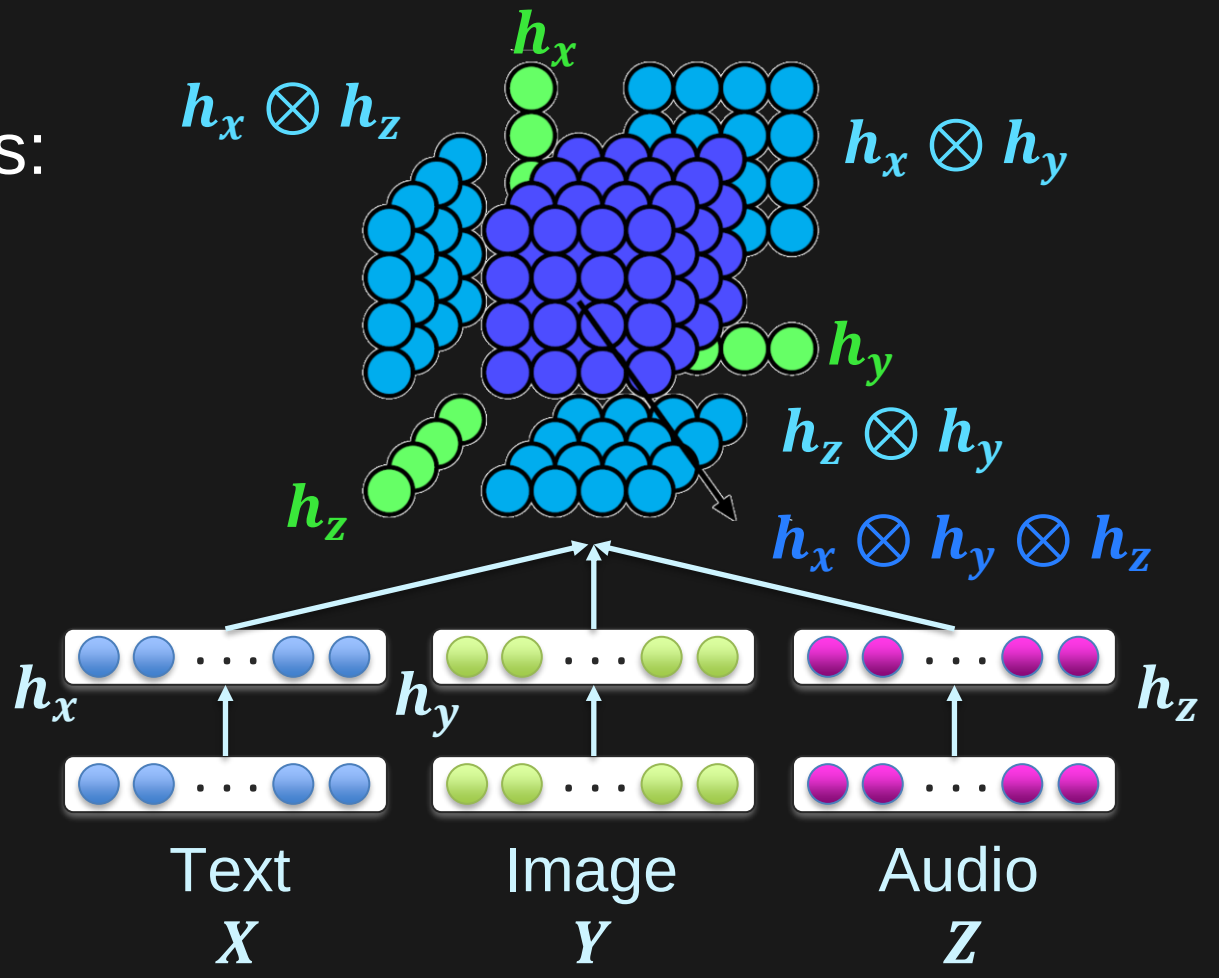
Multimodal Tensor Fusion Network (TFN)

Can be extended to three modalities:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_z \\ 1 \end{bmatrix}$$

Explicitly models unimodal, bimodal and trimodal interactions !

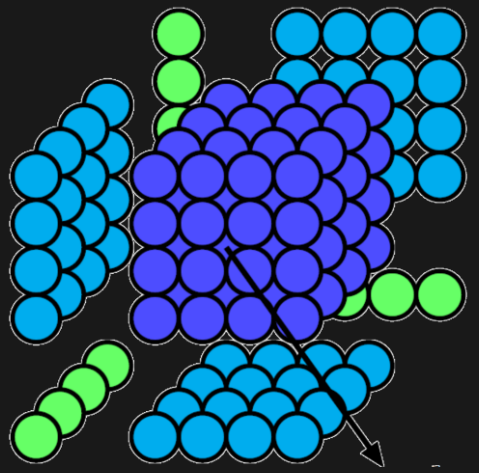
[Zadeh, Jones and Morency, **EMNLP 2017**]



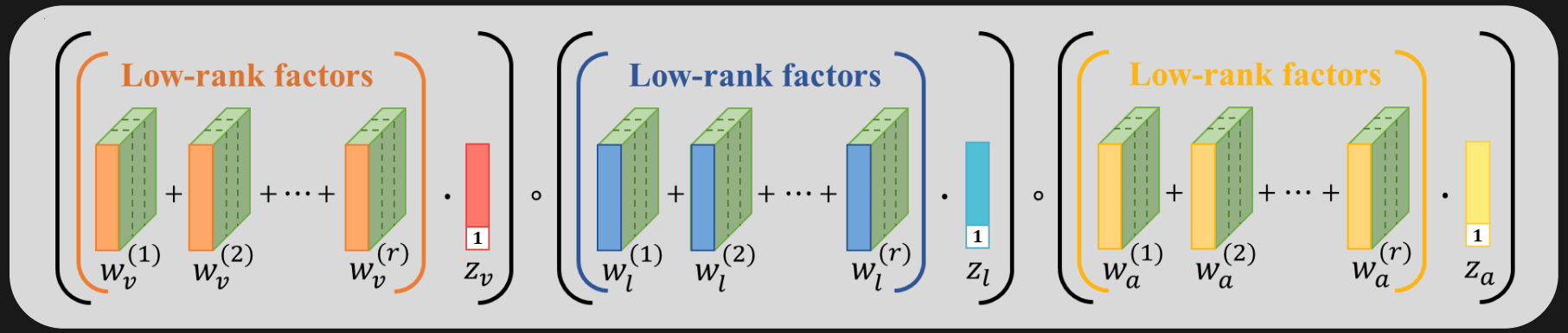
Improving Efficiency of Multimodal Representations [ACL 2018]

Tensor Fusion Network: Explicitly models unimodal, bimodal and trimodal interactions

[Zadeh, Jones and Morency, **EMNLP 2017**]



Efficient Low-rank Multimodal Fusion



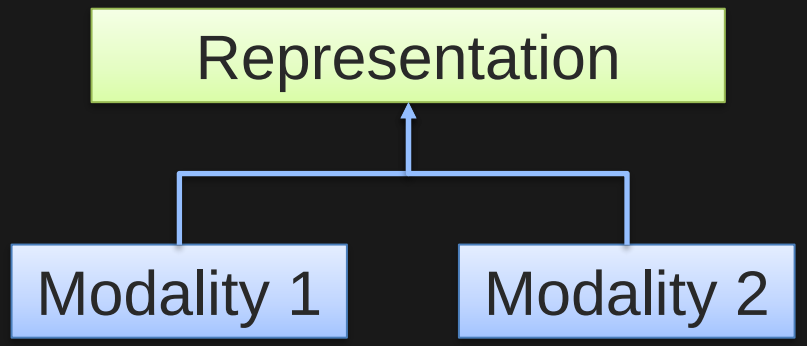
Canonical Polyadic Decomposition

[Liu, Shen, Bharadwaj, Liang, Zadeh and Morency, **ACL 2018**]

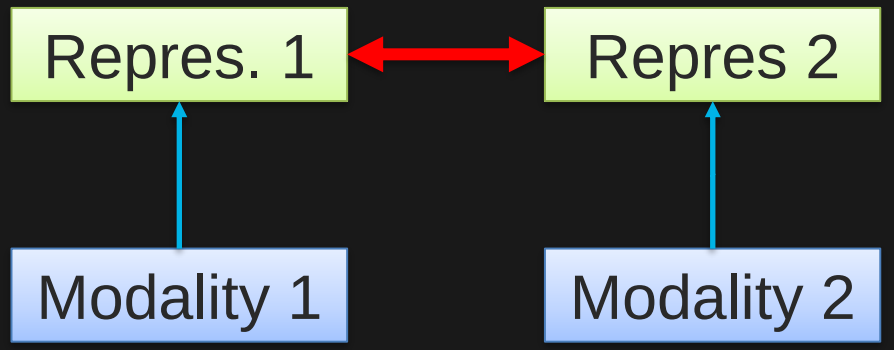
Core Challenge 1: Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

Ⓐ **Joint representations:**



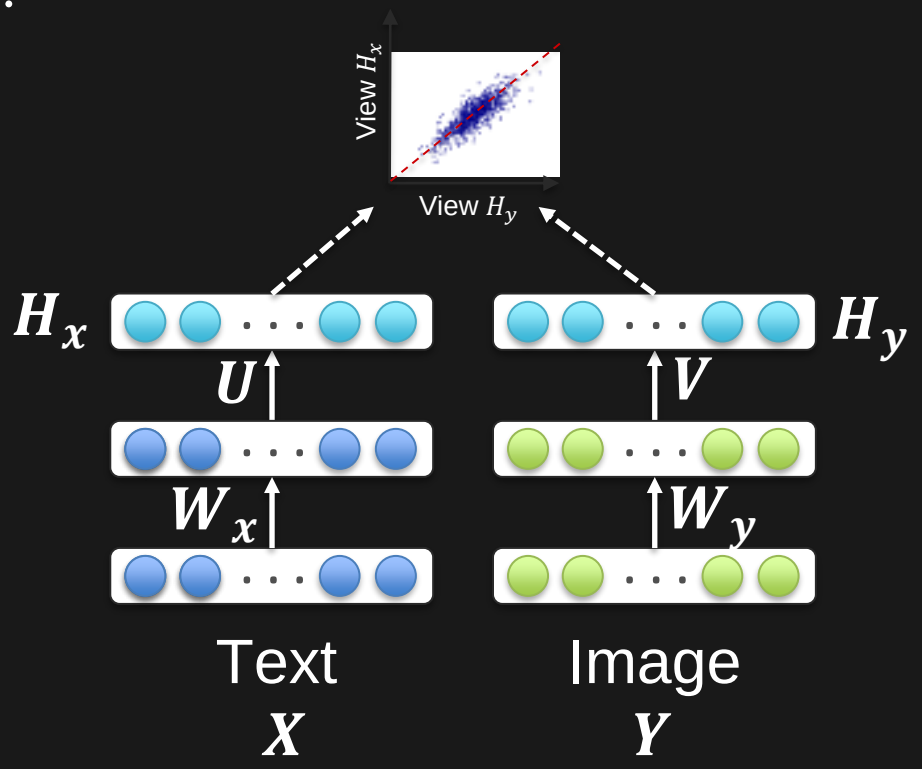
Ⓑ **Coordinated representations:**



Coordinated Representation: Deep CCA

Learn linear projections that are maximally correlated:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr(\boldsymbol{u}^T\boldsymbol{X}, \boldsymbol{v}^T\boldsymbol{Y})$$



Andrew et al., ICML 2013

Toward Debiasing Sentence Representations [ACL 2020]

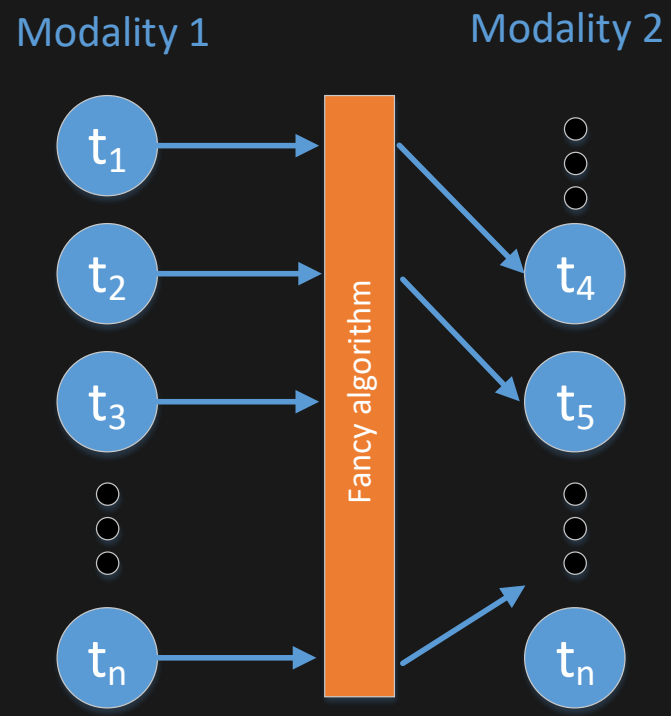
“The boy is coding.” OR “The girl is coding.”

“The boys at the playground.”
OR
“The girls at the playground.”

RESEARCH QUESTION: How to debias multimodal representations?

Core Challenge 2: Alignment

Definition: Identify the direct relations between (sub)elements from two or more different modalities.



A Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

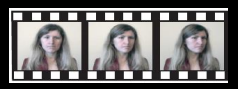
Uses internally latent alignment of modalities in order to better solve a different problem

Explicit (Multimodal) Alignment

Alignment and Representation: Multimodal Transformer

[ACL 2019]

Visual

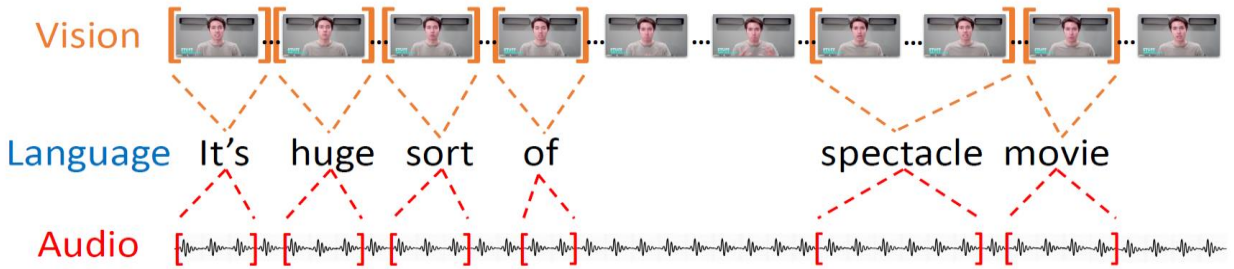


Vocal

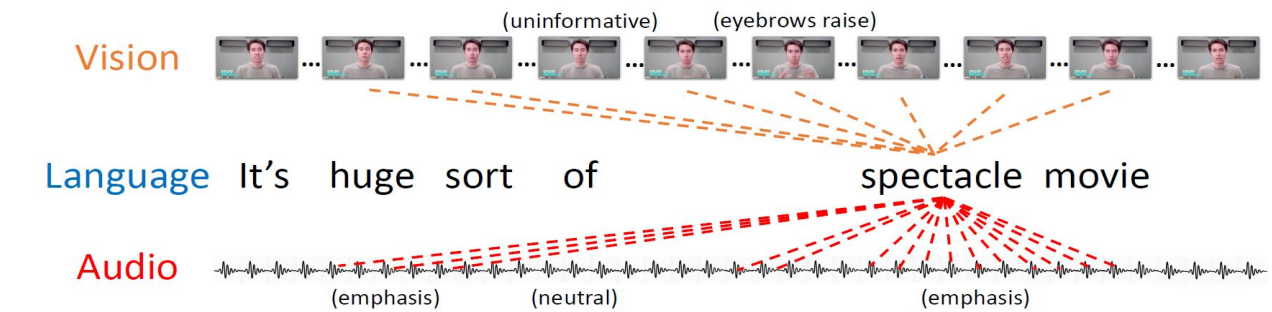
Verbal

"I like…"

Predefined Word-level alignment



Automatic Cross-Modal alignment



time

Multimodal representation

Representation

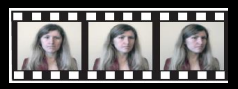
Alignment

Multimodal Transformer [ACL 2019]

Alignment

Representation

Visual



Visually contextualizing the verbal modality

Verbal

"I like…"

"spectacle"

Similarities:
(attentions)

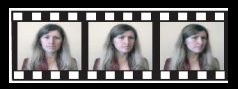
time

Multimodal Transformer [ACL 2019]

Alignment

Representation

Visual



Visually contextualizing the verbal modality

Verbal

"I like…"

"spectacle"

Residual connection

" New visually-contextualized representation of language"

Similarities:
(attentions)

Visual embeddings:

Correlated visual information

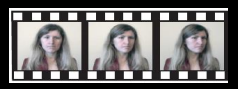
time

Multimodal Transformer [ACL 2019]

Alignment

Representation

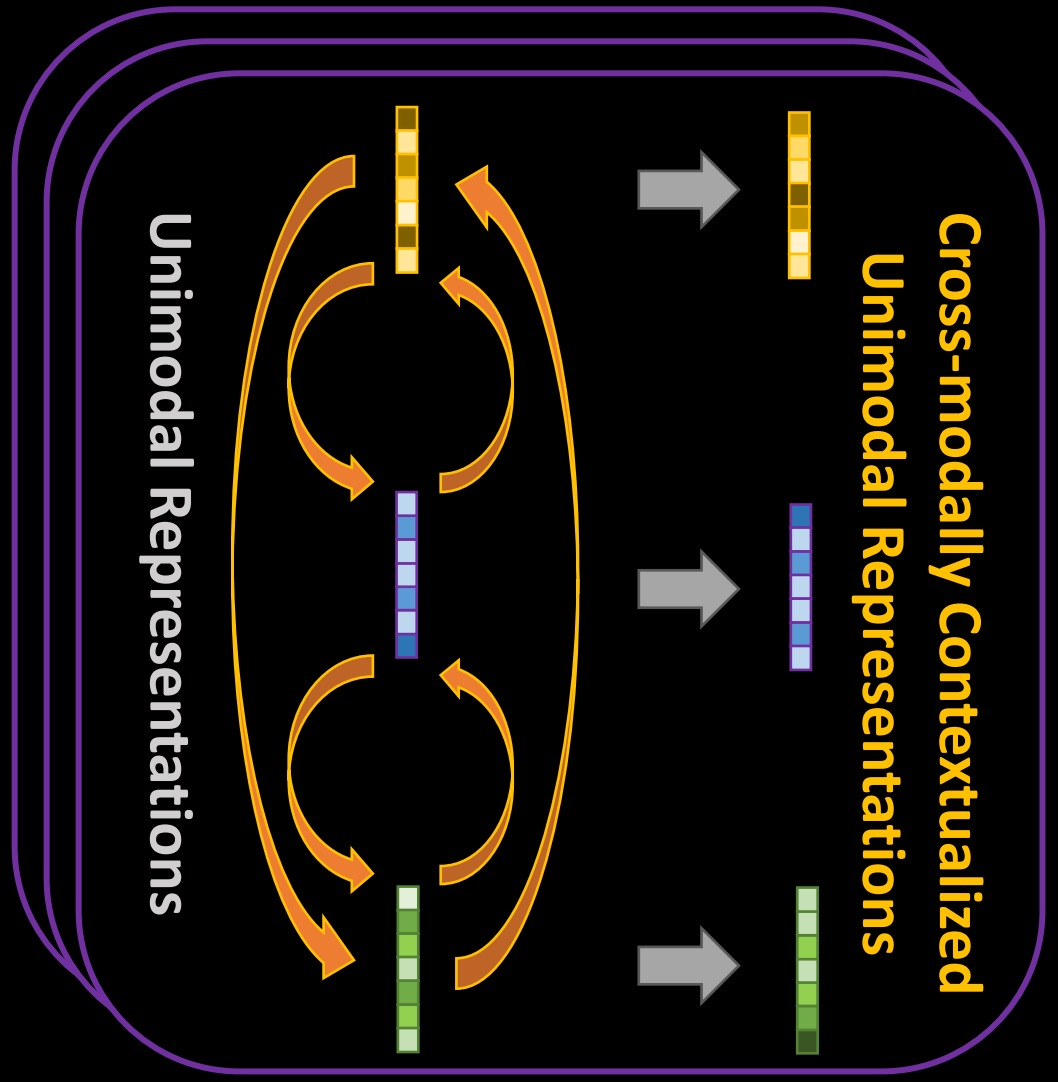
Visual



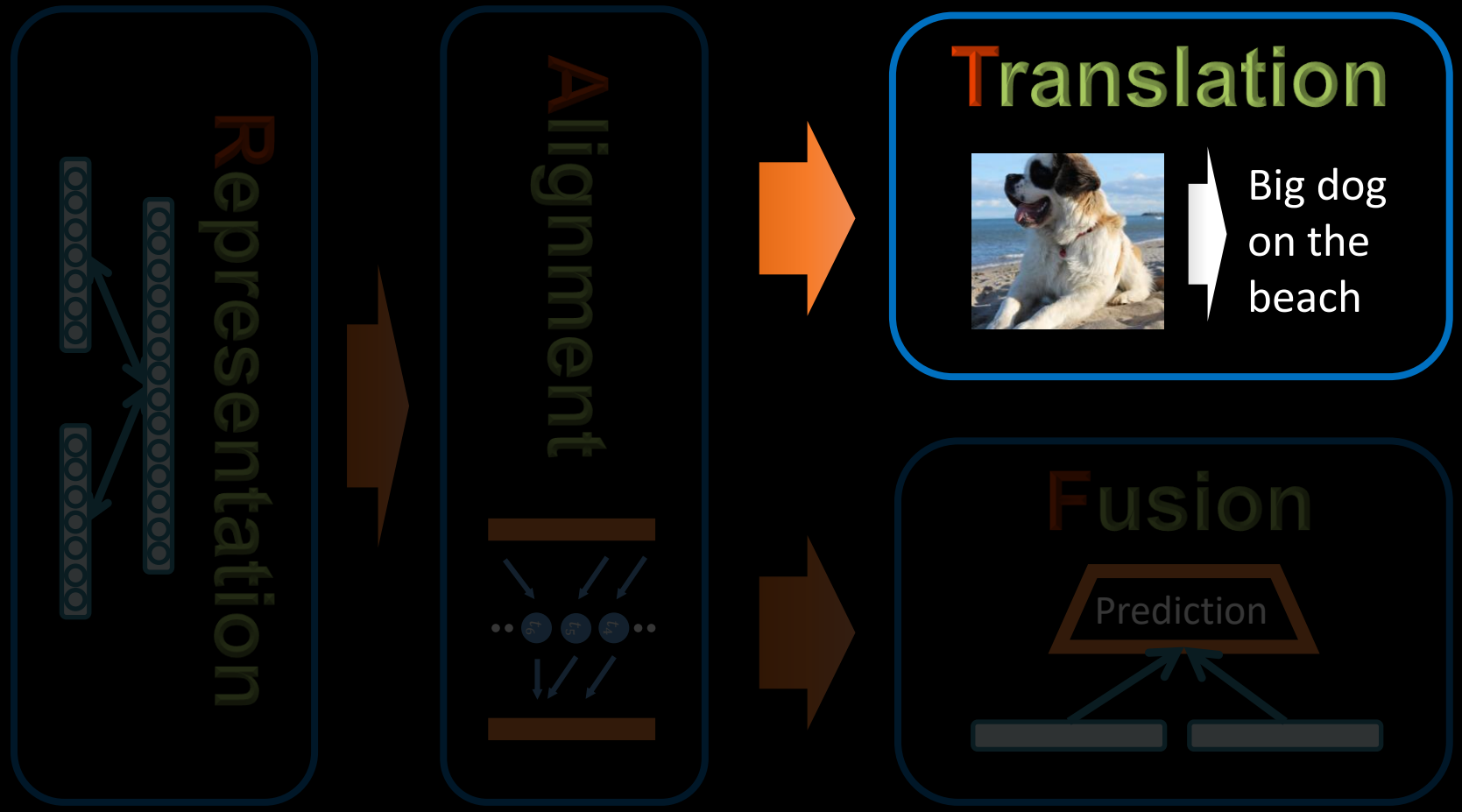
Vocal

Verbal

"I like…"



Multimodal AI – Core Challenges [Survey: TPAMI 2019]



Language-to-Pose [3DV 2019]

Translation

Story Narrative

"Characters walk hands-in-hands slowly while talking and then decide to leave the road and explore the forest…"

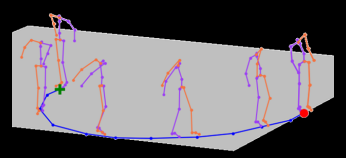
Movie Animation

More than a week for professional animators to create 15 seconds !

Language-to-Pose [3DV 2019]

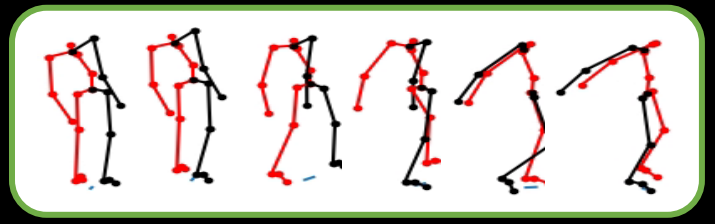
Translation

Visual



Verbal

"A person walks in a circle"



y_1 y_2 y_3 y_4 y_5 y_6

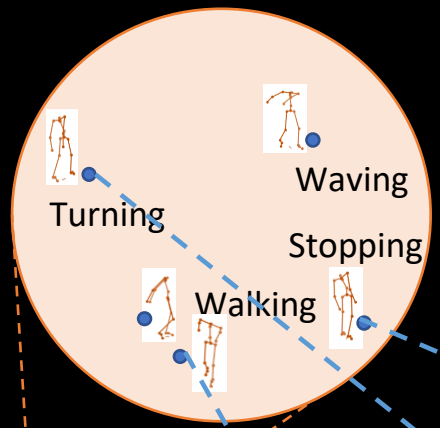
Pose Encoder $q_e(y; \Psi_e)$

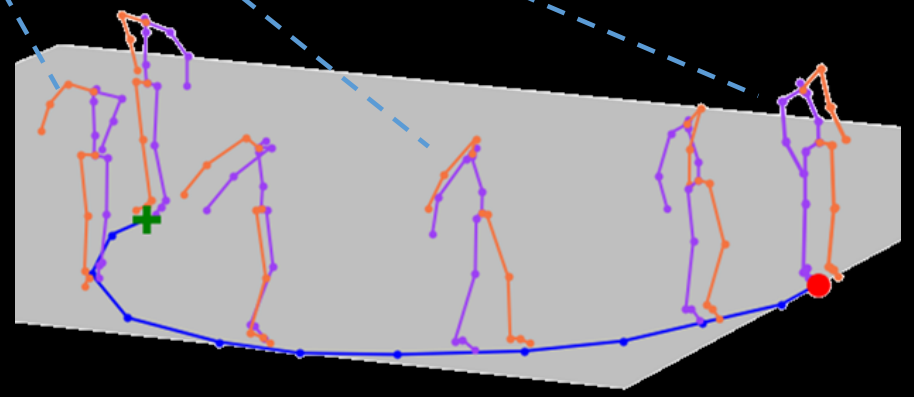
Language Encoder $p_e(x; \Phi_e)$

x_1 x_2 x_3 x_4 x_5 x_6

"A person walks in a circle"

Joint Embedding Space $\mathbb{Z}$

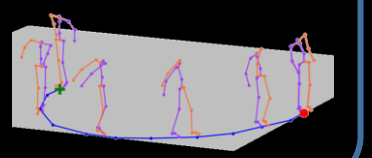




Language-to-Pose [3DV 2019]

Translation

Visual

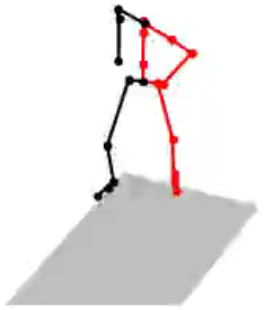


Verbal

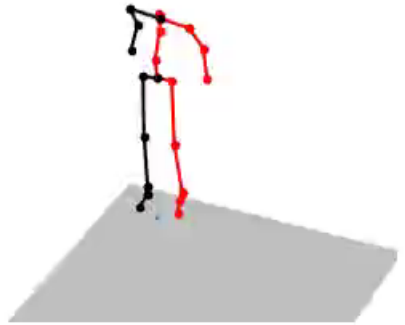
“A person walks in a circle”

Results from our JL2P model

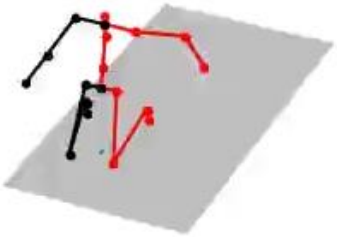
(Joint Language-to-Pose)



a person jogs a few steps



A person steps forward then turns around and steps forwards again.

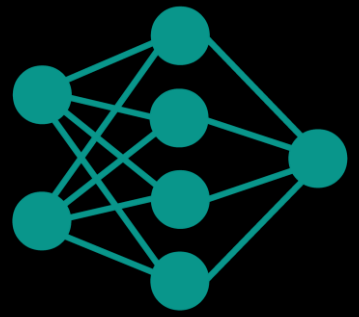


A kneeling person raises their arms to the sides and stand up.

Language-to-Network [ACL 2020]

Rhinoceros Auklet: Breeding adults are cloudy gray, ... a thick orange-yellow bill ...

Acadian Flycatcher: olive-green above with a whitish eye-ring ... two distinct white wing-bars.



✓ Rhinoceros Auklet

?

✗ Acadian Flycatcher

Language-to-Action [ACL 2020]

Multi-step instruction:

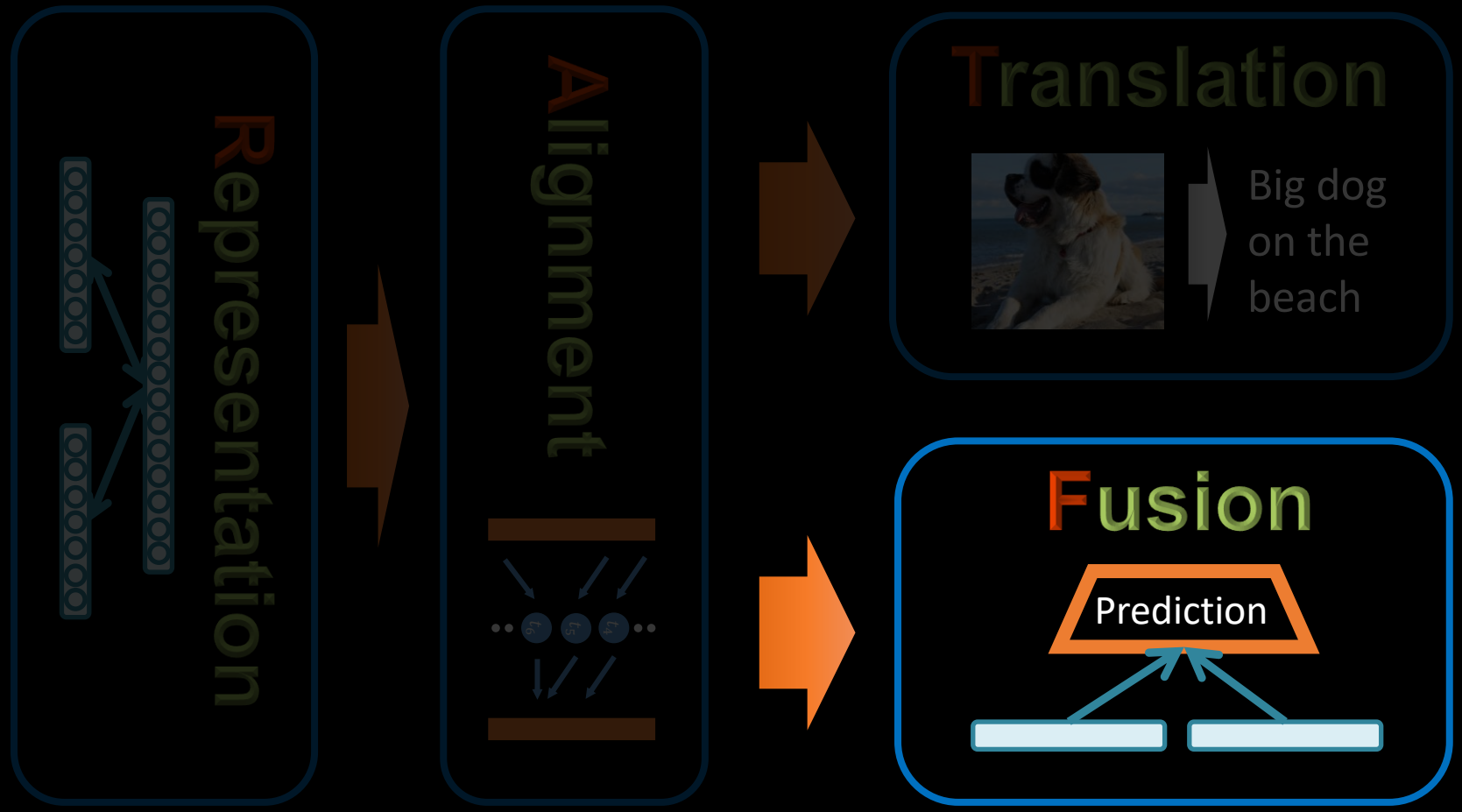
1. **Go** to the **entrance** of the **lounge area**.
2. **On your right** there will be **a bar**.
3. **On top** of the **counter**, you will see **a box**. **Bring** me **that**.

Refer360 dataset

- 17,135 annotated instances
- 2,000 panaromic 360 degrees scenes
- 43.8 average number of words per instructions

https://github.com/volkancirik/refer360

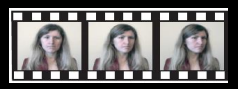
Multimodal AI – Core Challenges [Survey: TPAMI 2019]



Fusion

Multimodal Sentiment Analysis [ACL 2018]

Visual



+

Vocal

+

Verbal

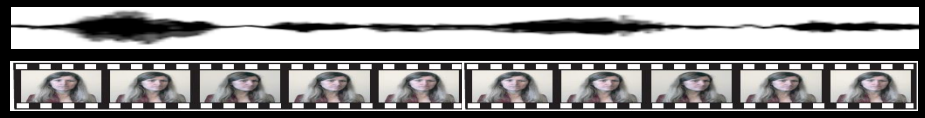
"I like…"

CMU- MOSEI dataset

- 23,000 video segments
- 1,000 speakers (from vlogs)
- 3 modalities (verbal, vocal,visual)

https://github.com/A2Zadeh/CMU-MultimodalSDK

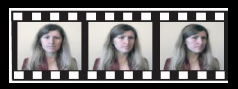
"He's average presenter when…"



Fusion

Multimodal Sentiment Analysis [AAAI 2018]

Visual

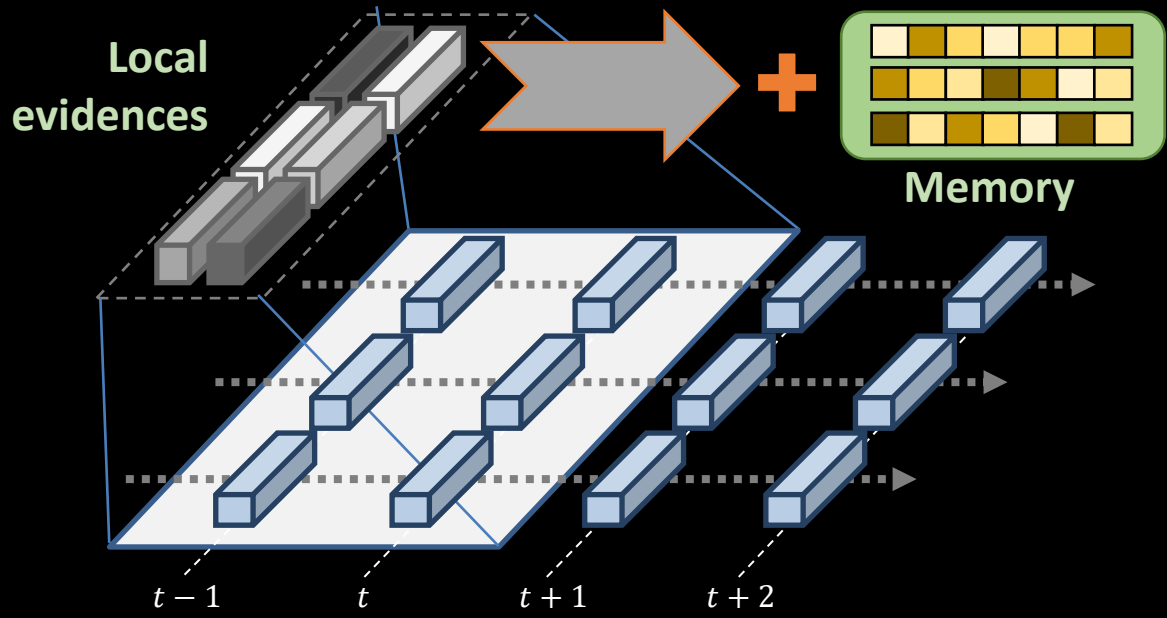


Vocal

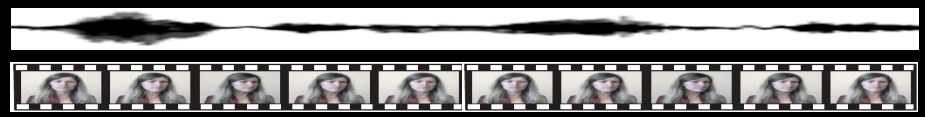
Verbal

"I like…"

Memory Fusion Network

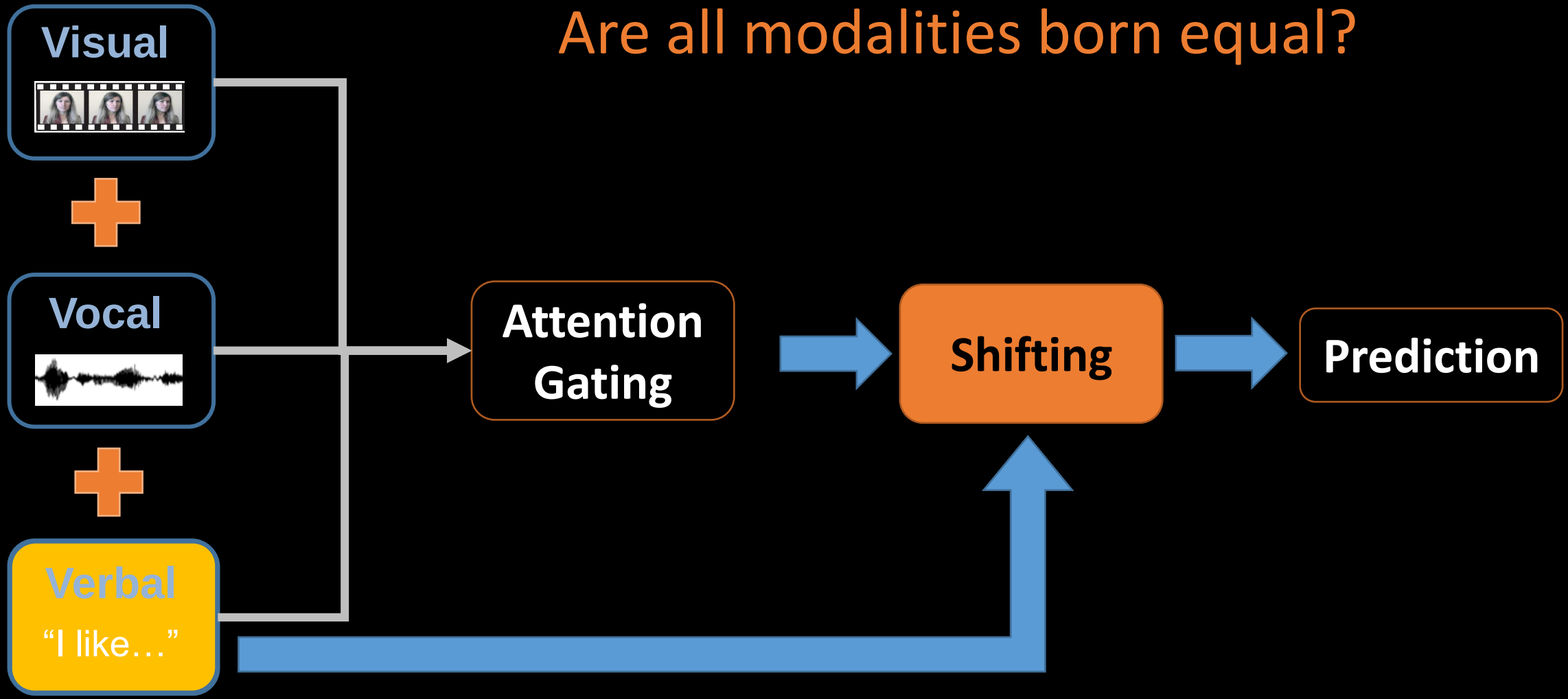


"He's average presenter when…"

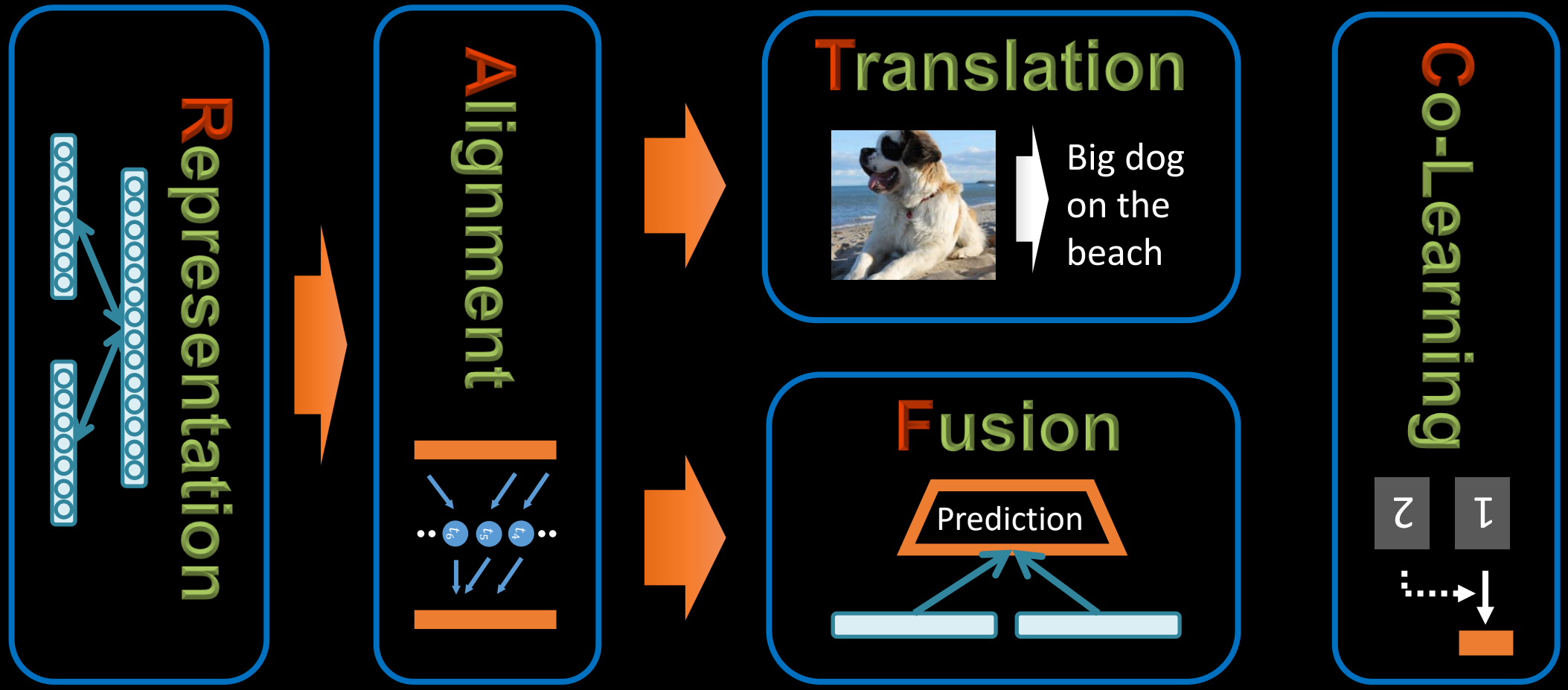


Multimodal Adaptation Gate [ACL 2020]

Are all modalities born equal?



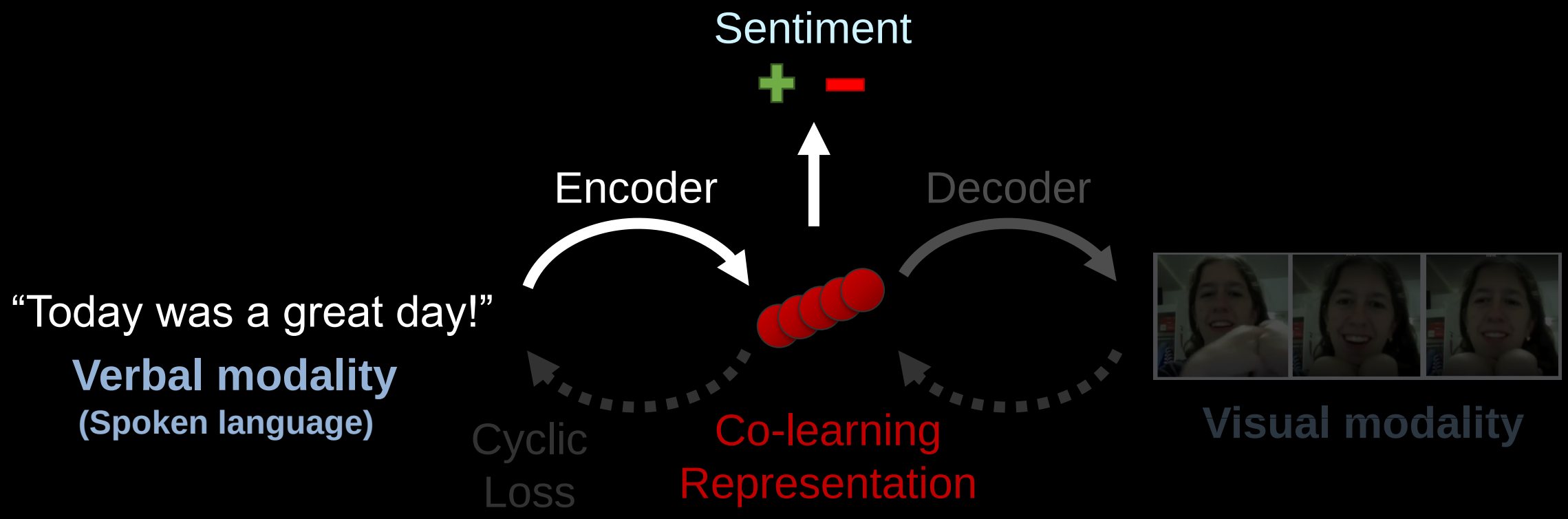
Multimodal AI – Core Challenges [Survey: TPAMI 2019]



Co-Learning

Representations from Cross-modal Translation

[AAAI 2019]

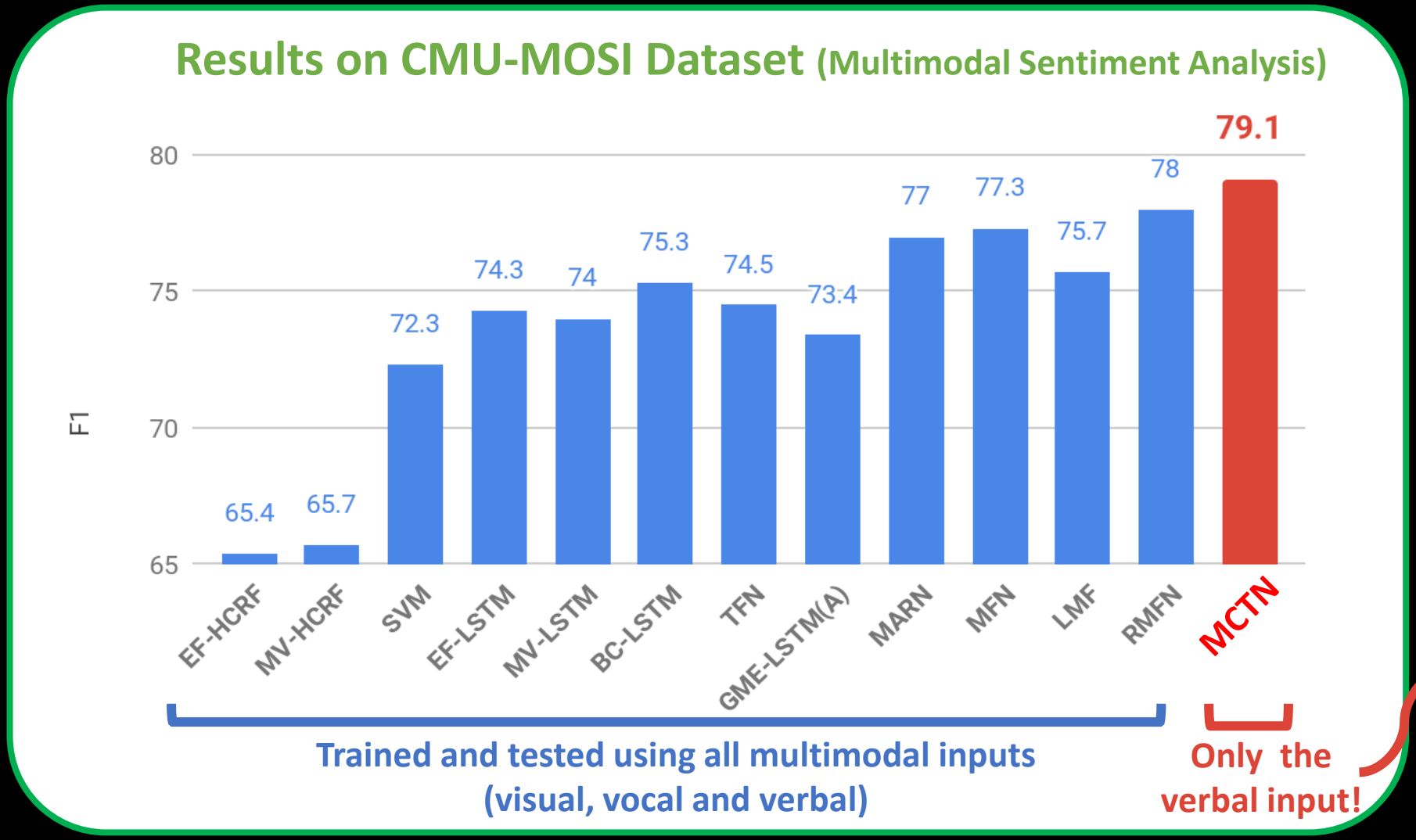


Multimodal Cyclic Translation Network (MCTN)

Representations from Cross-modal Translation

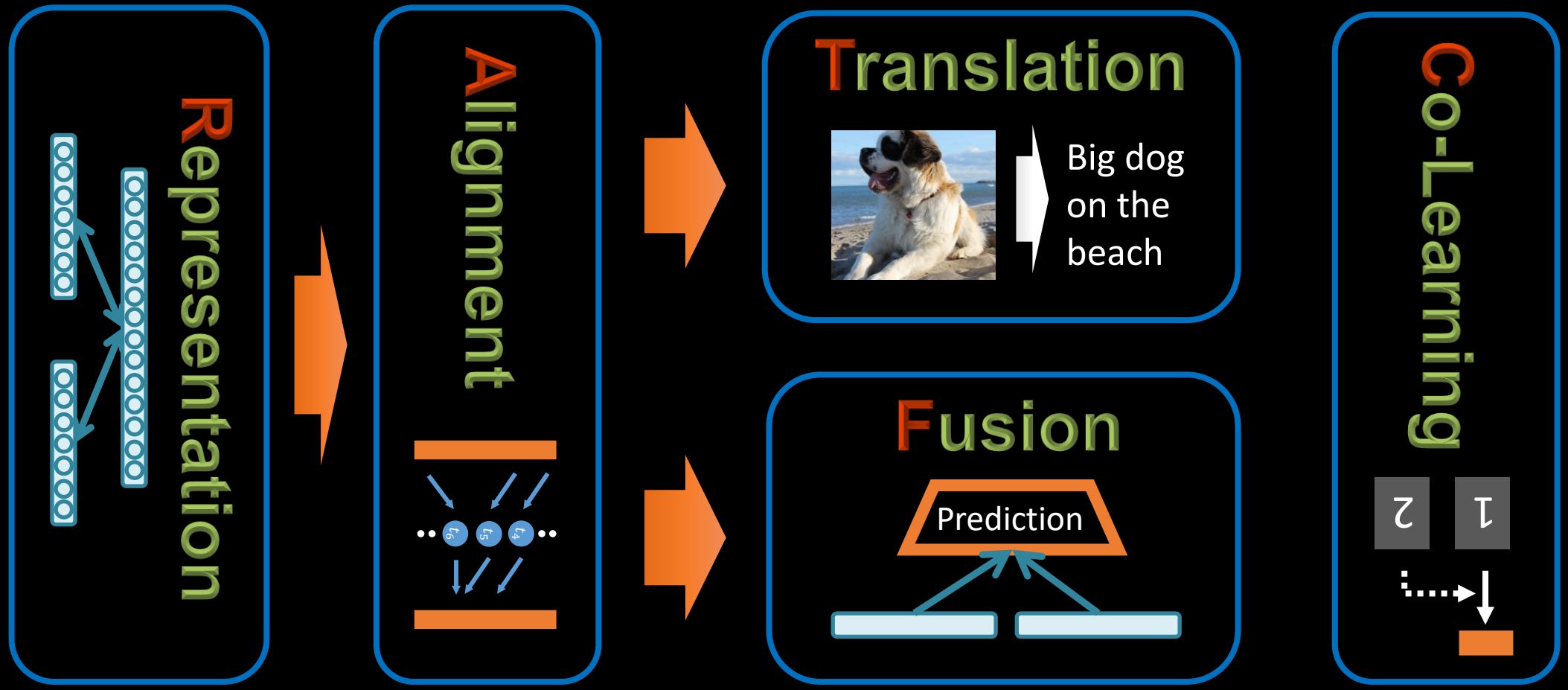
Co-Learning

[AAAI 2019]



Trained using multimodal co-learning but tested with only verbal input!

Multimodal AI – Core Challenges [Survey: TPAMI 2019]



Multimodal Machine Learning – Taxonomy

[Survey: TPAMI 2019]

Representation

- Joint
  - *Neural networks*
  - *Graphical models*
  - *Sequential*
- Coordinated
  - *Similarity*
  - *Structured*

Alignment

- Explicit
  - *Unsupervised*
  - *Supervised*
- Implicit
  - *Graphical models*
  - *Neural networks*

Fusion

- Model agnostic
  - *Early fusion*
  - *Late fusion*
  - *Hybrid fusion*
- Model-based
  - *Kernel-based*
  - *Graphical models*
  - *Neural networks*

Translation

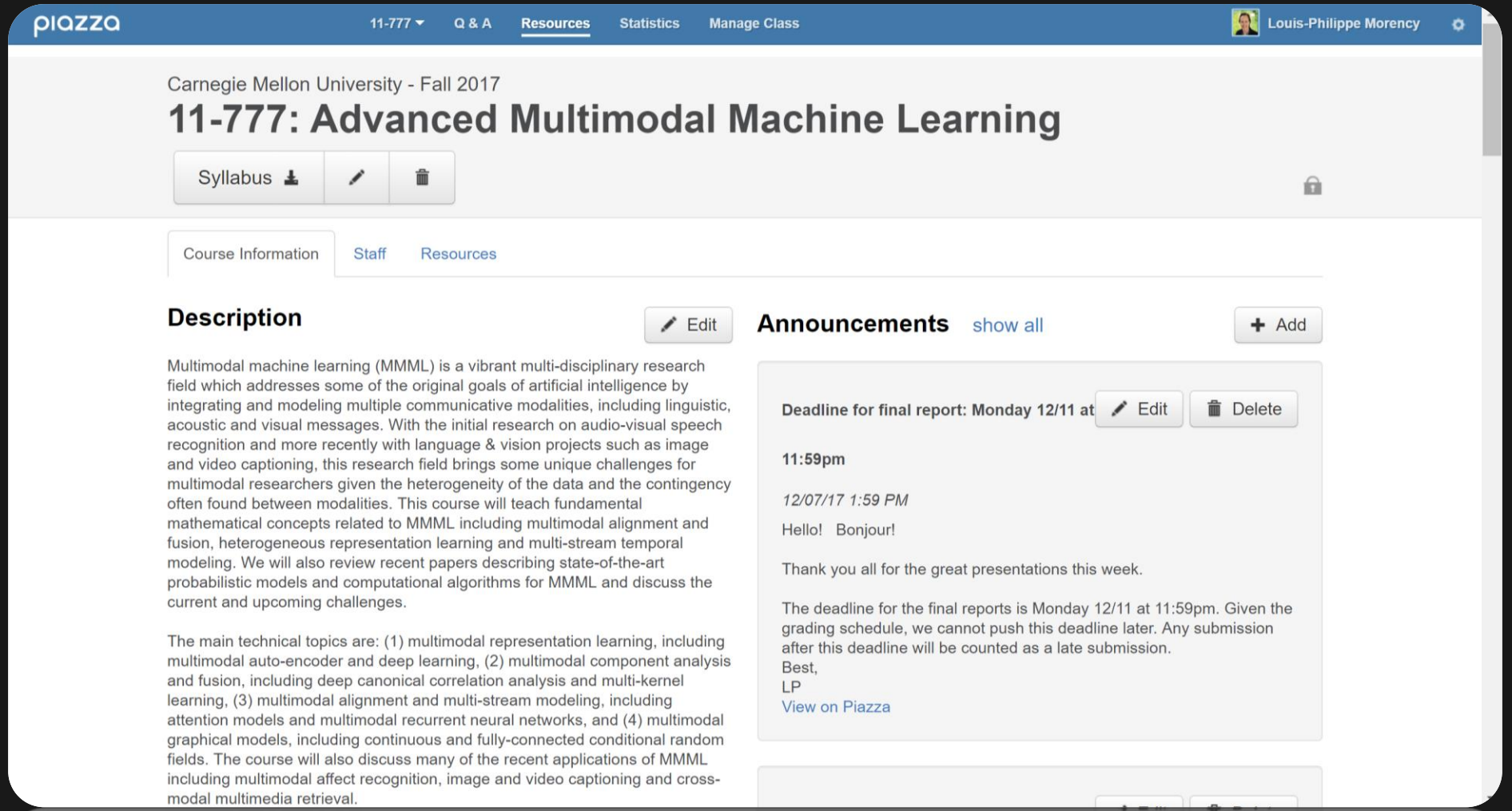
- Example-based
  - *Retrieval*
  - *Combination*
- Model-based
  - *Grammar-based*
  - *Encoder-decoder*
  - *Online prediction*

Co-learning

- Parallel data
  - *Co-training*
  - *Transfer learning*
- Non-parallel data
  - *Zero-shot learning*
  - *Concept grounding*
  - *Transfer learning*
- *Hybrid data*
  - *Bridging*

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, **Multimodal Machine Learning: A Survey and Taxonomy**, https://arxiv.org/abs/1705.09406

CMU Course on Multimodal Machine Learning



https://piazza.com/cmu/fall2019/11777/home

MERCI !

http://multicomp.cs.cmu.edu/