Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 1.1: Introduction

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.*
Spring 2021 edition taught by Yonatan Bisk

Your Instructor and TAs This Semester (11-777)

Louis-Philippe Morency
morency@cs.cmu.edu
Course lecturer

Ta-Chung Chi
tachungc@andrew.cmu.edu
TA

Xuandi Fu
xuandifu@cmu.edu
TA

Martin Q. Ma
qianlim@cmu.edu
TA

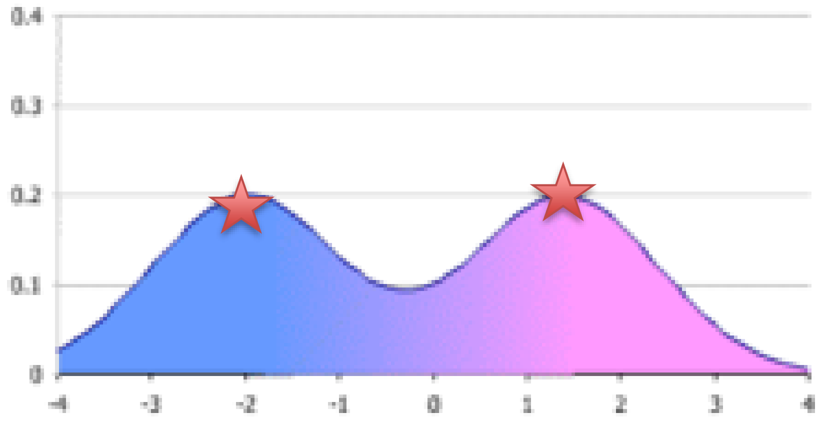
Tianqin Li
tianqinl@cs.cmu.edu
TA

Lecture Objectives

- Introductions
- What is Multimodal?
  - Multimodal communicative behaviors
- A historical view of multimodal research
- Core technical and conceptual challenges
  - Representation, alignment, translation, fusion and co-learning
- Course syllabus and project assignments
  - Grades and course structure

What is Multimodal?

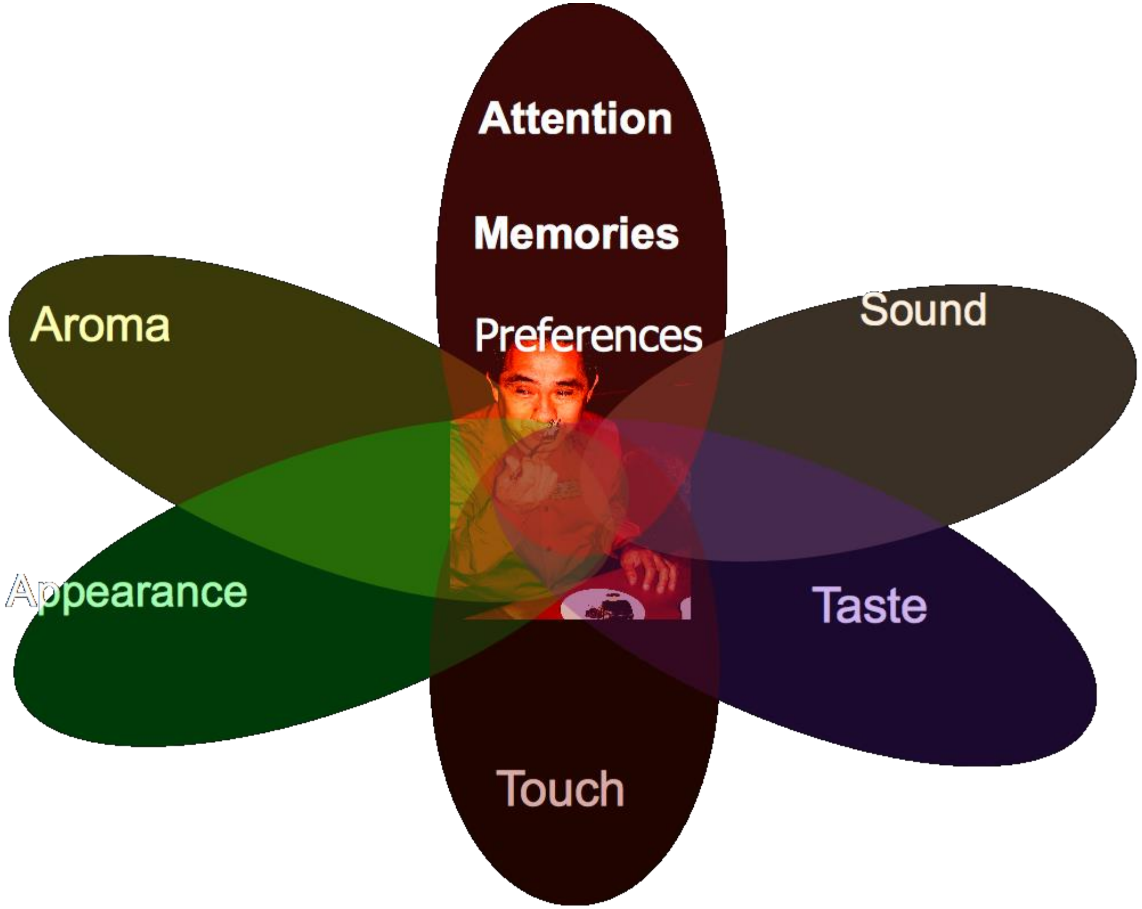
What is Multimodal?



Multimodal distribution

- Multiple modes, i.e., distinct "peaks" (local maxima) in the probability density function

What is Multimodal?



Sensory Modalities

Multimodal Communicative Behaviors

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

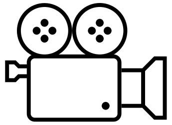
Vocal
- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

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

What is Multimodal?

Modality

Modality refers to the way in which something expressed or perceived.



from a sensor

Raw Modalities (closest from sensor) → Abstract Modalities (farthest from sensor)

Examples:

Speech signal	Language	Sentiment intensity
Image	Detected objects	Object categories

Multimodal: from multiple modalities

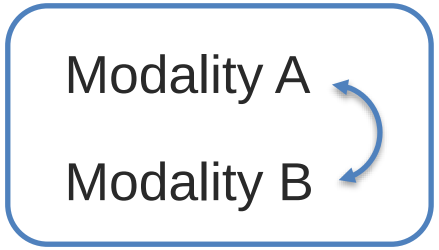
Examples of Modalities

- Natural language (both spoken or written)
- Visual (from images or videos)
- Auditory (including voice, sounds and music)
- Haptics / touch
- Smell, taste and self-motion
- Physiological signals
  - Electrocardiogram (ECG), skin conductance
- Other modalities
  - Infrared images, depth images, fMRI

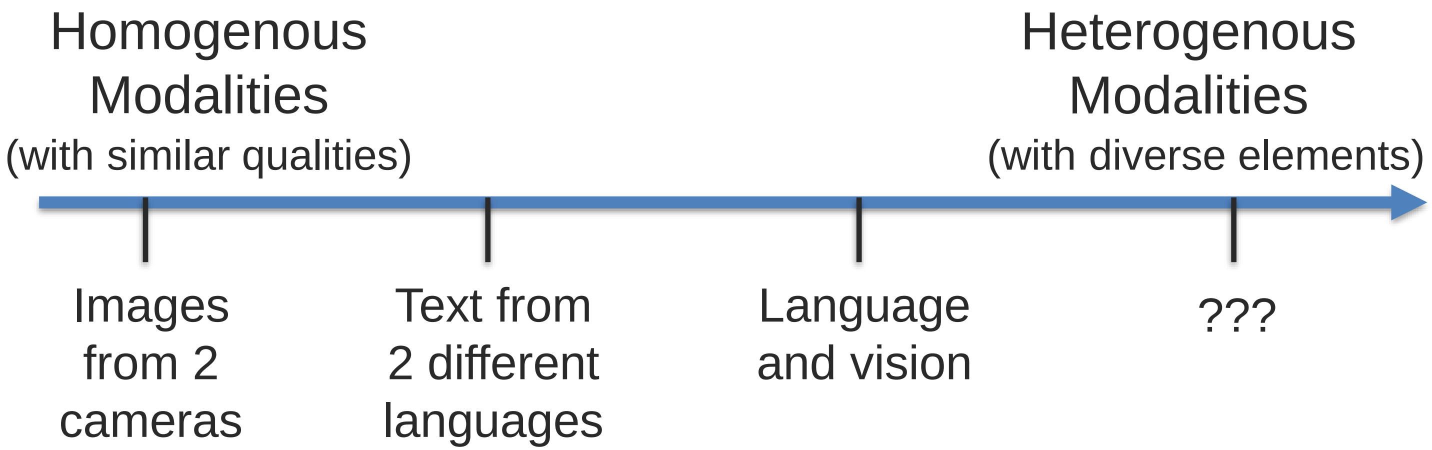
What is Multimodal?

Heterogeneity

Information present in the different modalities will often show diverse qualities and elements.



Examples:



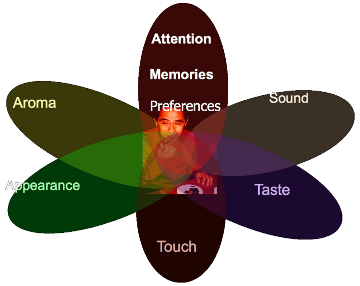
What is Multimodal?

Multimodal Machine Learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data

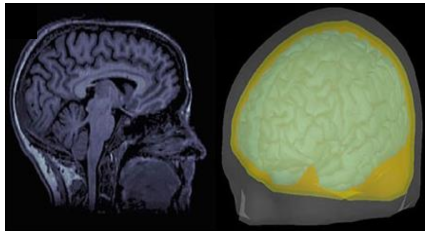
Multimodal Artificial Intelligence studies computer agents able to demonstrate intelligence capabilities such as understanding, reasoning and planning, through multimodal experiences, and data

***Multimodal* is the science of heterogenous data ☺**

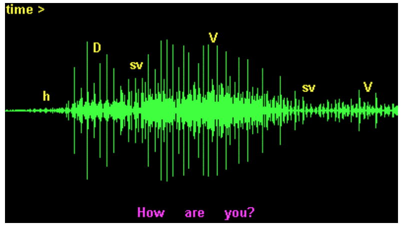
Multiple Communities and Modalities



Psychology



Medical

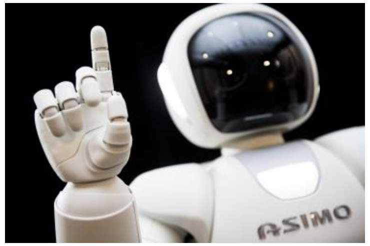


Speech

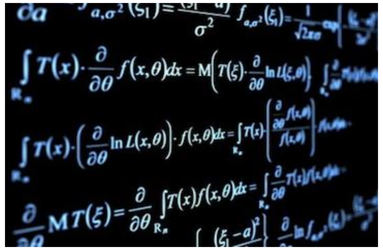
Vision

Language

Multimedia



Robotics



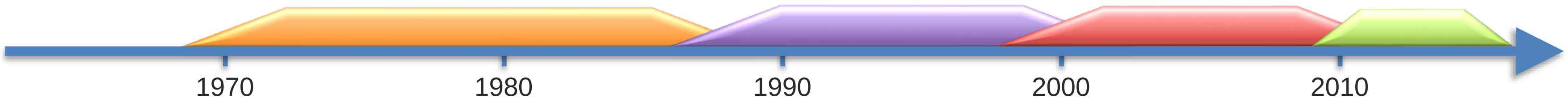
Learning

A Historical View

Prior Research on "Multimodal"

Four eras of multimodal research

- The "behavioral" era (1970s until late 1980s)
- The "computational" era (late 1980s until 2000)
- The "interaction" era (2000 - 2010)
- The "deep learning" era (2010s until …)
  - Main focus of this course

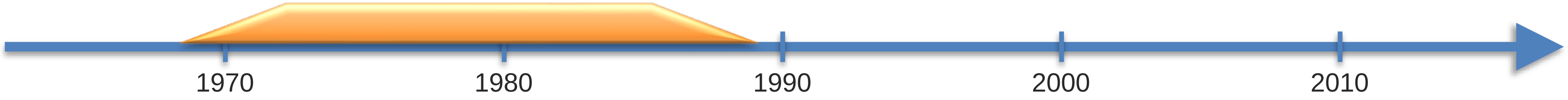


Language and Gestures

David McNeill
University of Chicago
Center for Gesture and Speech Research

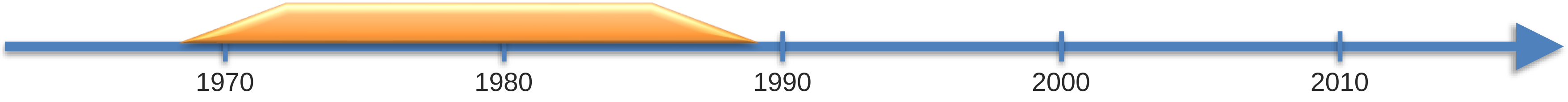
"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

- TRIVIA: Justine Cassell was a student of David McNeill



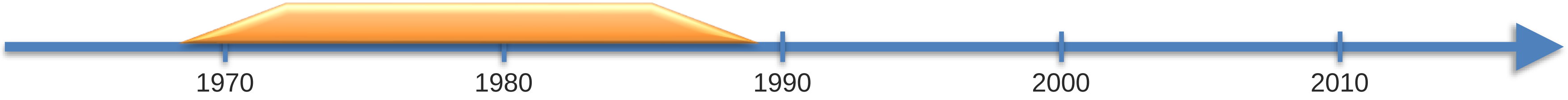
The McGurk Effect (1976)

[Hearing lips and seeing voices – Nature]



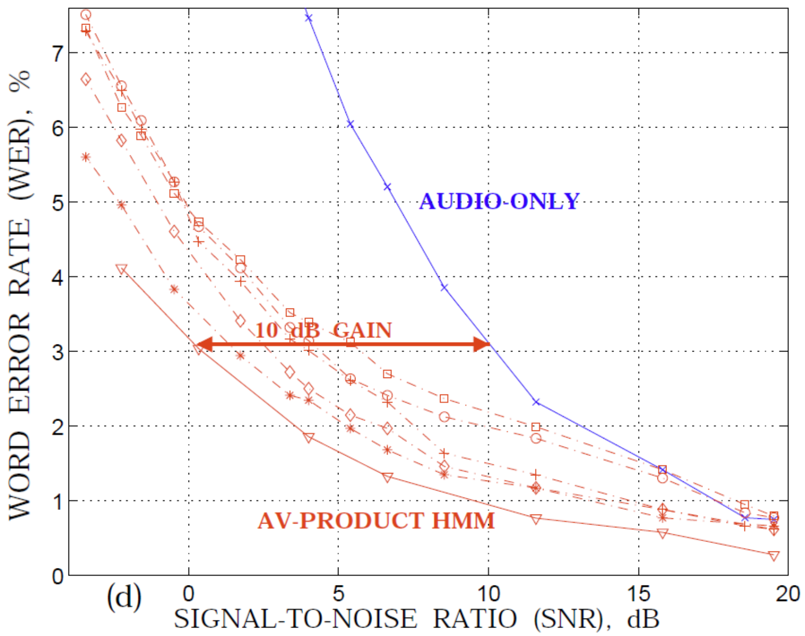
The McGurk Effect (1976)

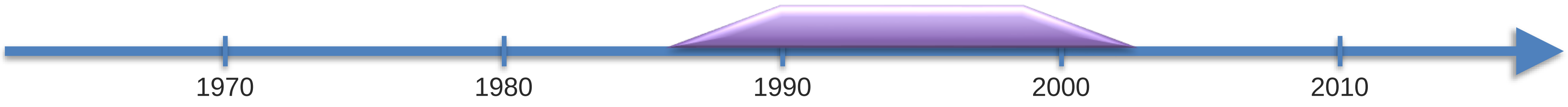
[Hearing lips and seeing voices – Nature]



➢ The "Computational" Era(Late 1980s until 2000)

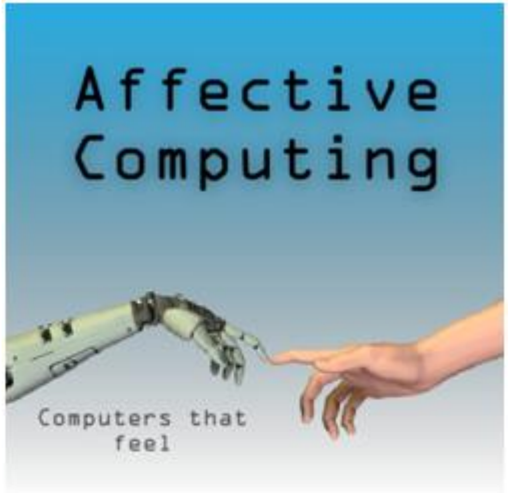
1) Audio-Visual Speech Recognition (AVSR)





➢ The "Computational" Era (Late 1980s until 2000)

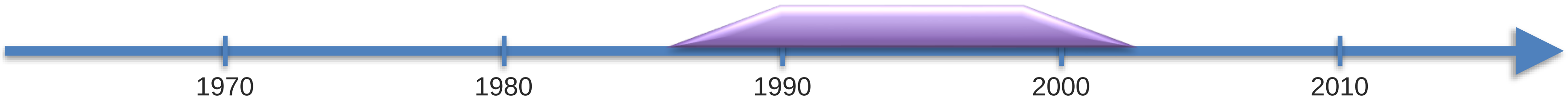
2) Multimodal/multisensory interfaces



Rosalind Picard

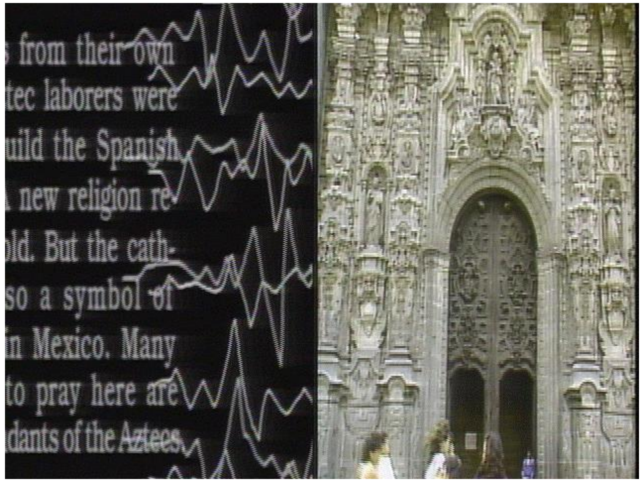
***Affective Computing** is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena.*

- TRIVIA: Rosalind Picard came from the same group (MIT, Sandy Pentland)



➢ The “Computational” Era (Late 1980s until 2000)

3) Multimedia Computing

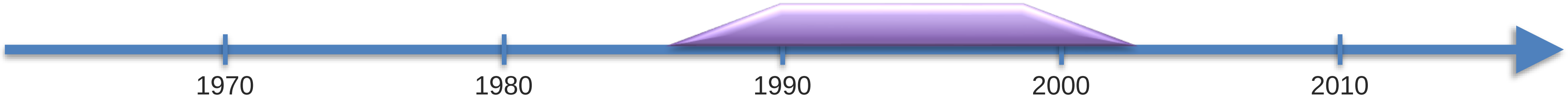


Carnegie Mellon University

informedia digital video understanding

[1994-2010]

“The Informedia Digital Video Library Project automatically combines speech, image and natural language understanding to create a full-content searchable digital video library.”

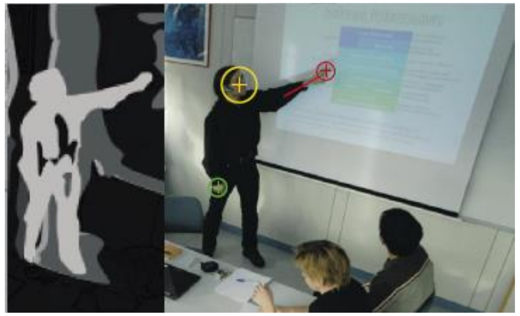


➢ The “Interaction” Era (2000s)

1) Modeling Human Multimodal Interaction

AMI Project [2001-2006, IDIAP]

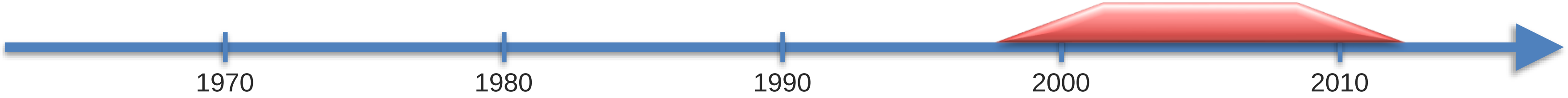
- 100+ hours of meeting recordings
- Fully synchronized audio-video
- Transcribed and annotated



CHIL Project [Alex Waibel]

- Computers in the Human Interaction Loop
- Multi-sensor multimodal processing
- Face-to-face interactions

❑ TRIVIA: Samy Bengio started at IDIAP working on AMI project



➢ The "Interaction" Era (2000s)

1) Modeling Human Multimodal Interaction

PAL
Personalized Assistant that Learns

CALO Project [2003-2008, SRI]

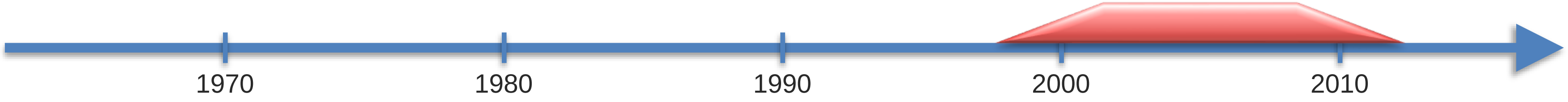
- Cognitive Assistant that Learns and Organizes
- Personalized Assistant that Learns (PAL)
- Siri was a spinoff from this project

Social Signal Processing Network

SSP Project [2008-2011, IDIAP]

- Social Signal Processing
- First coined by Sandy Pentland in 2007
- Great dataset repository: http://sspnet.eu/

❑ TRIVIA: LP's PhD research was partially funded by CALO ☺



➢ The "deep learning" era (2010s until …)

Representation learning (a.k.a. deep learning)

- Multimodal deep learning [ICML 2011]
- Multimodal Learning with Deep Boltzmann Machines [NIPS 2012]
- Visual attention: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [ICML 2015]

Key enablers for multimodal research:

- New large-scale multimodal datasets
- Faster computer and GPUS
- High-level visual features
- "Dimensional" linguistic features

Our course focuses on this era!

Core Technical Challenges

Core Challenges in “Deep” Multimodal ML

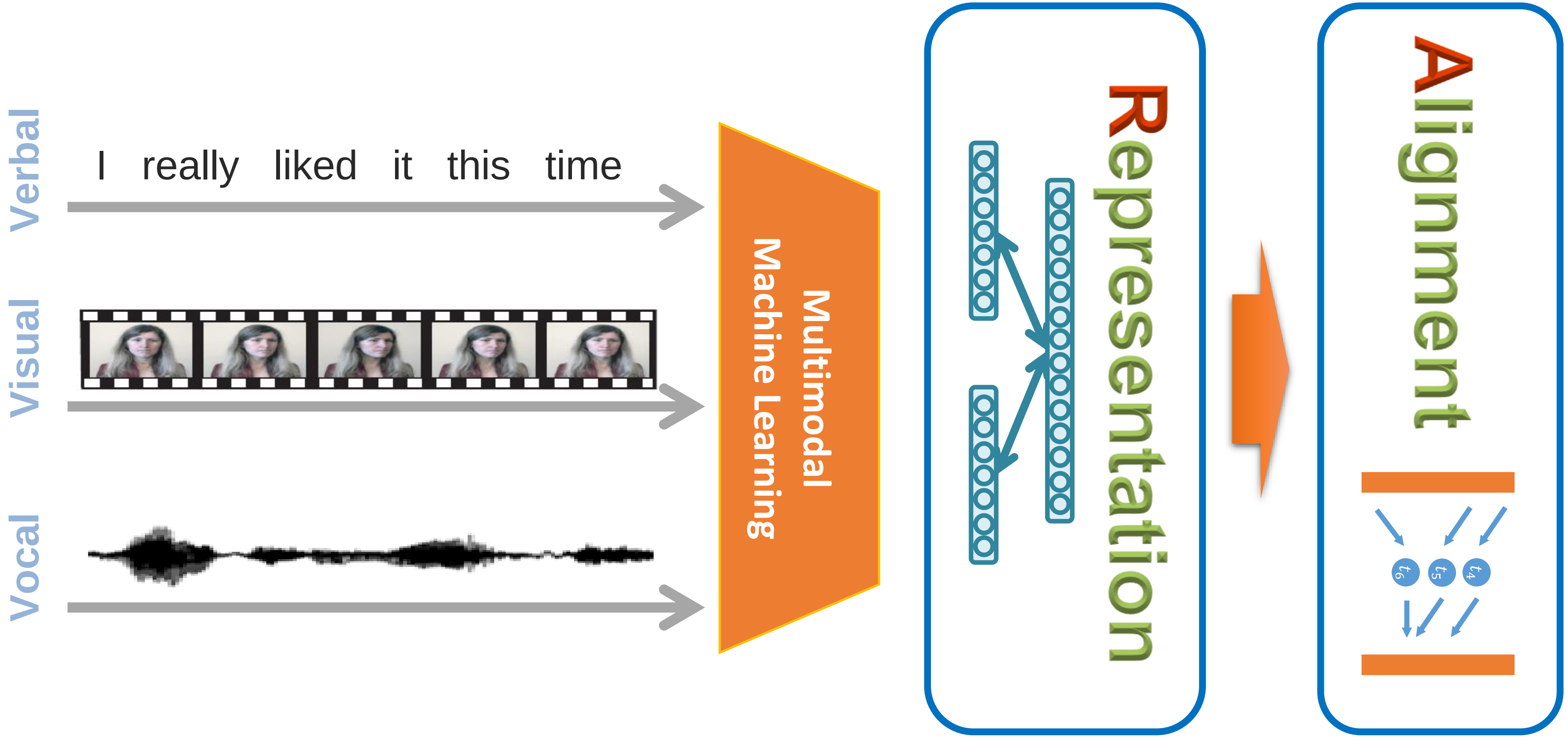
Multimodal Machine Learning: A Survey and Taxonomy

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

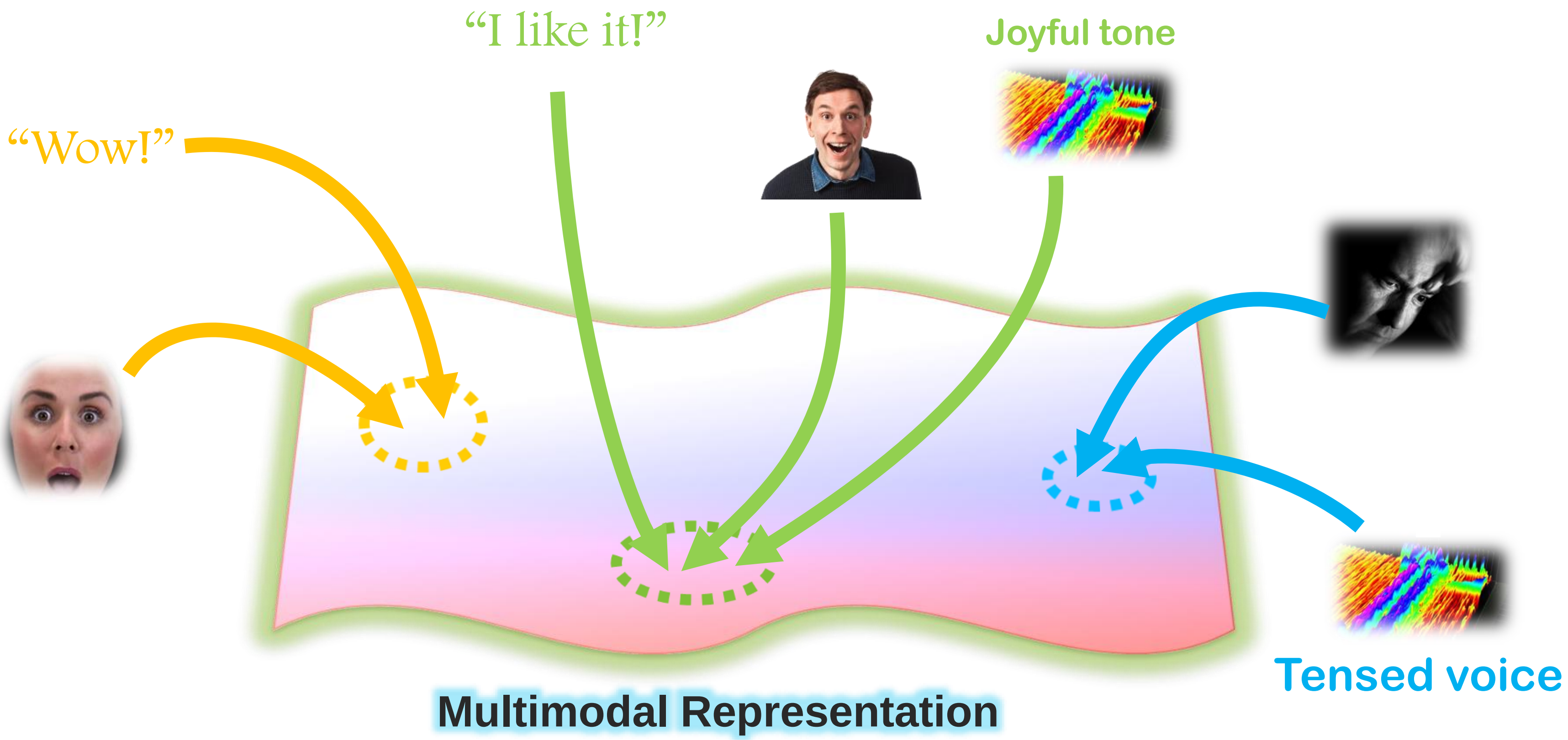
https://arxiv.org/abs/1705.09406

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

First Two Core Challenges



Core Challenge 1: Representation



Core Challenge 1: Early Examples

Audio-visual speech recognition
[Ngiam et al., ICML 2011]

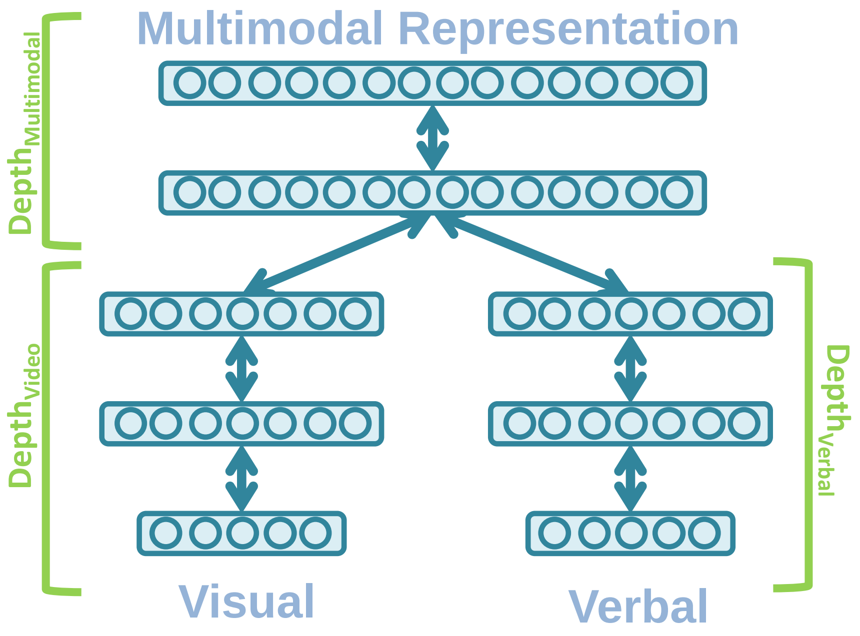
- Bimodal Deep Belief Network

Image captioning
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

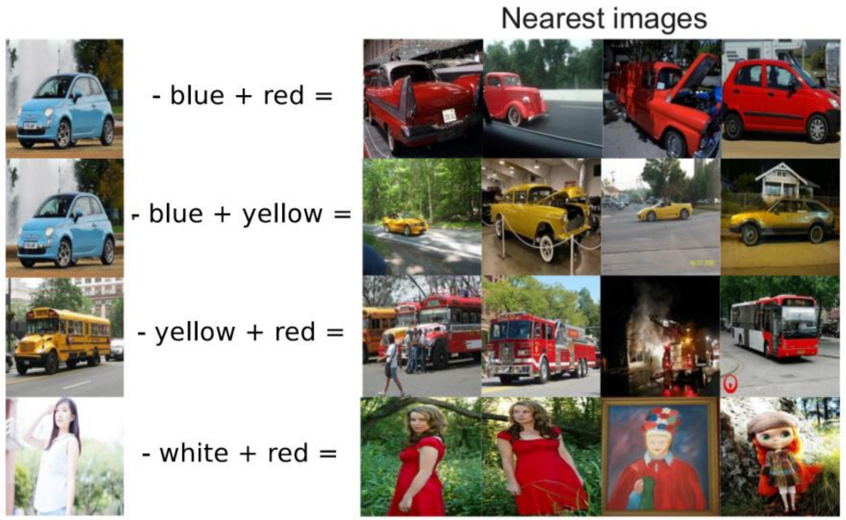
Audio-visual emotion recognition
[Kim et al., ICASSP 2013]

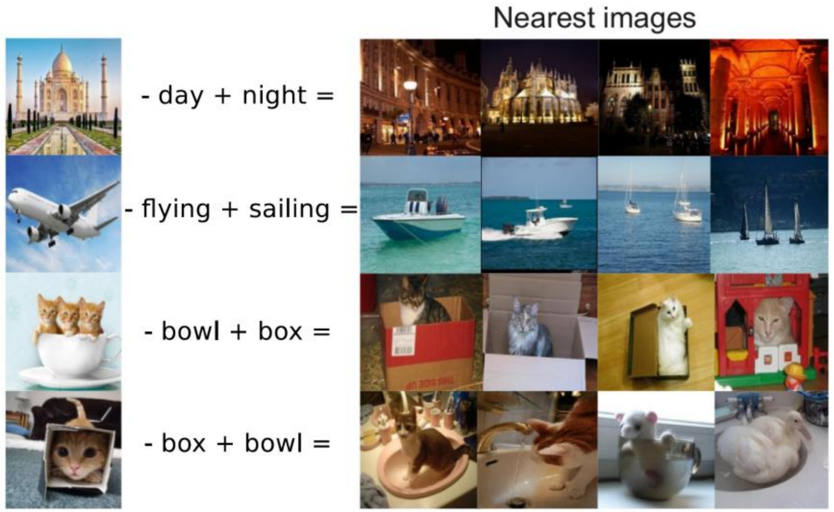
- Deep Boltzmann Machine



Core Challenge 1: Early Examples

Multimodal Vector Space Arithmetic



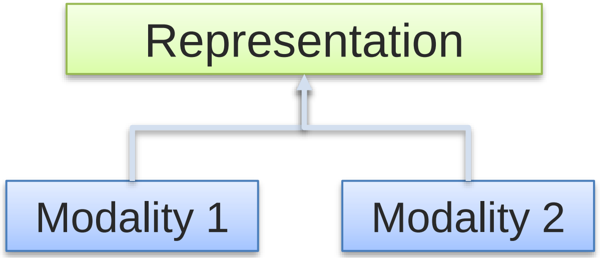


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

Core Challenge 1: Representation

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

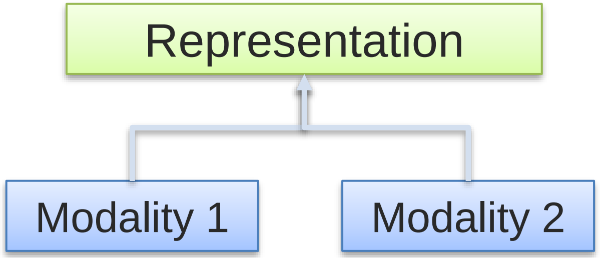
Ⓐ Joint representations:



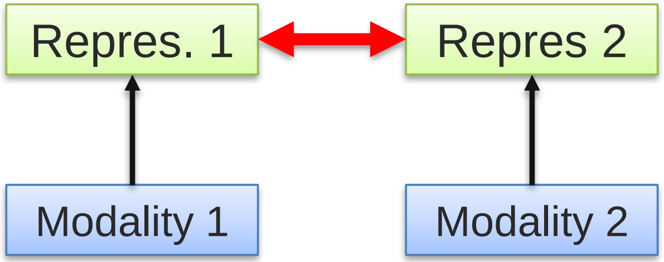
Core Challenge 1: Representation

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

Ⓐ **Joint representations:**

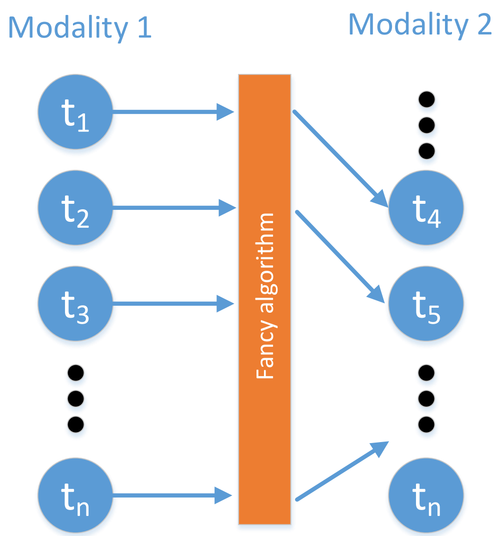


Ⓑ **Coordinated representations:**



Core Challenge 2: Alignment

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



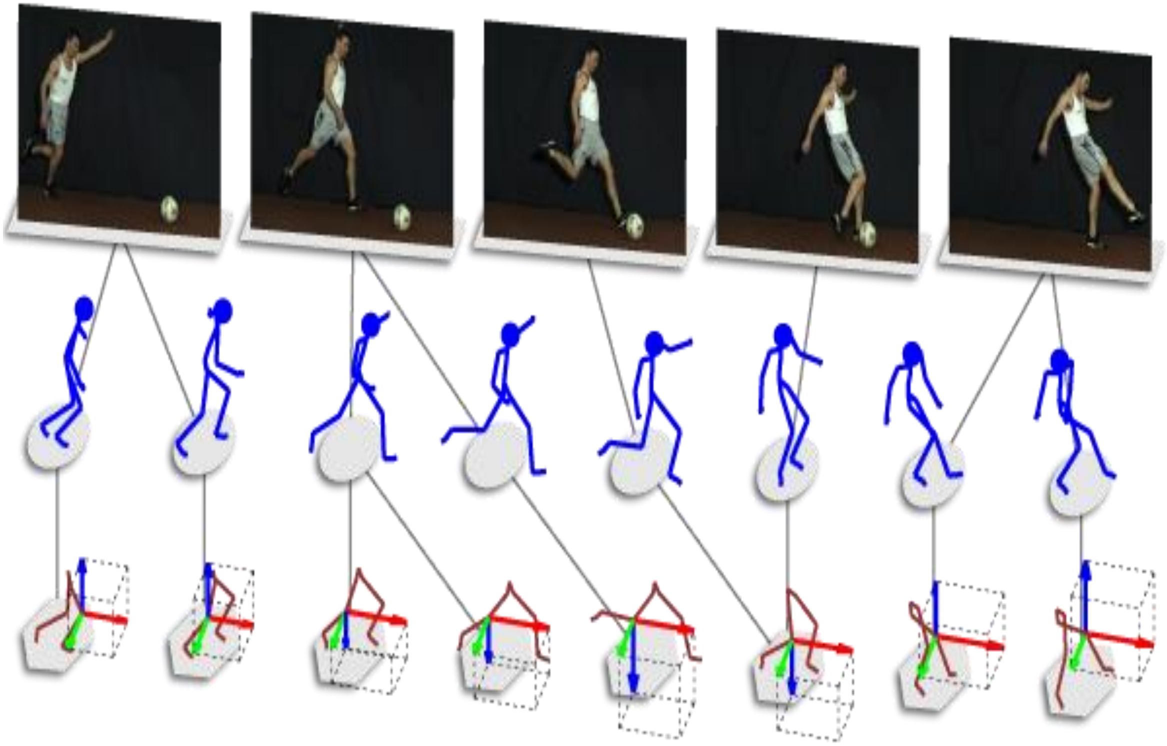
A Explicit Alignment

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

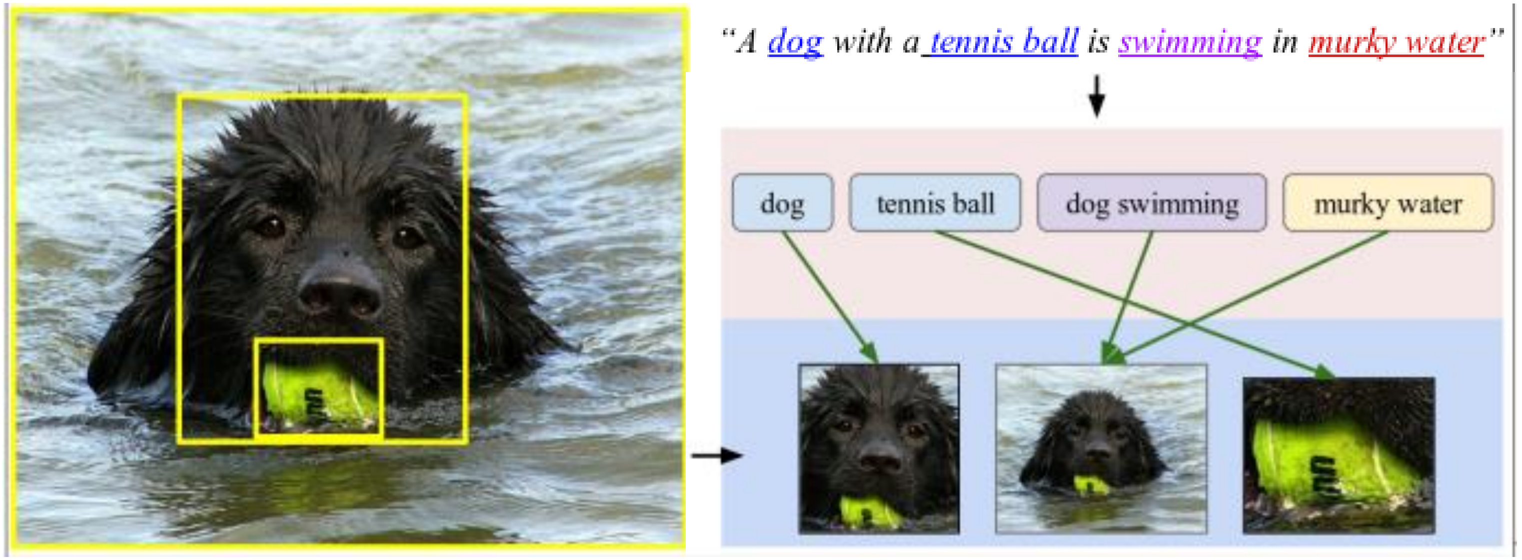
Core Challenge 2: Explicit Alignment



Applications:
- Re-aligning asynchronous data
- Finding similar data across modalities (we can estimate the aligned cost)
- Event reconstruction from multiple sources

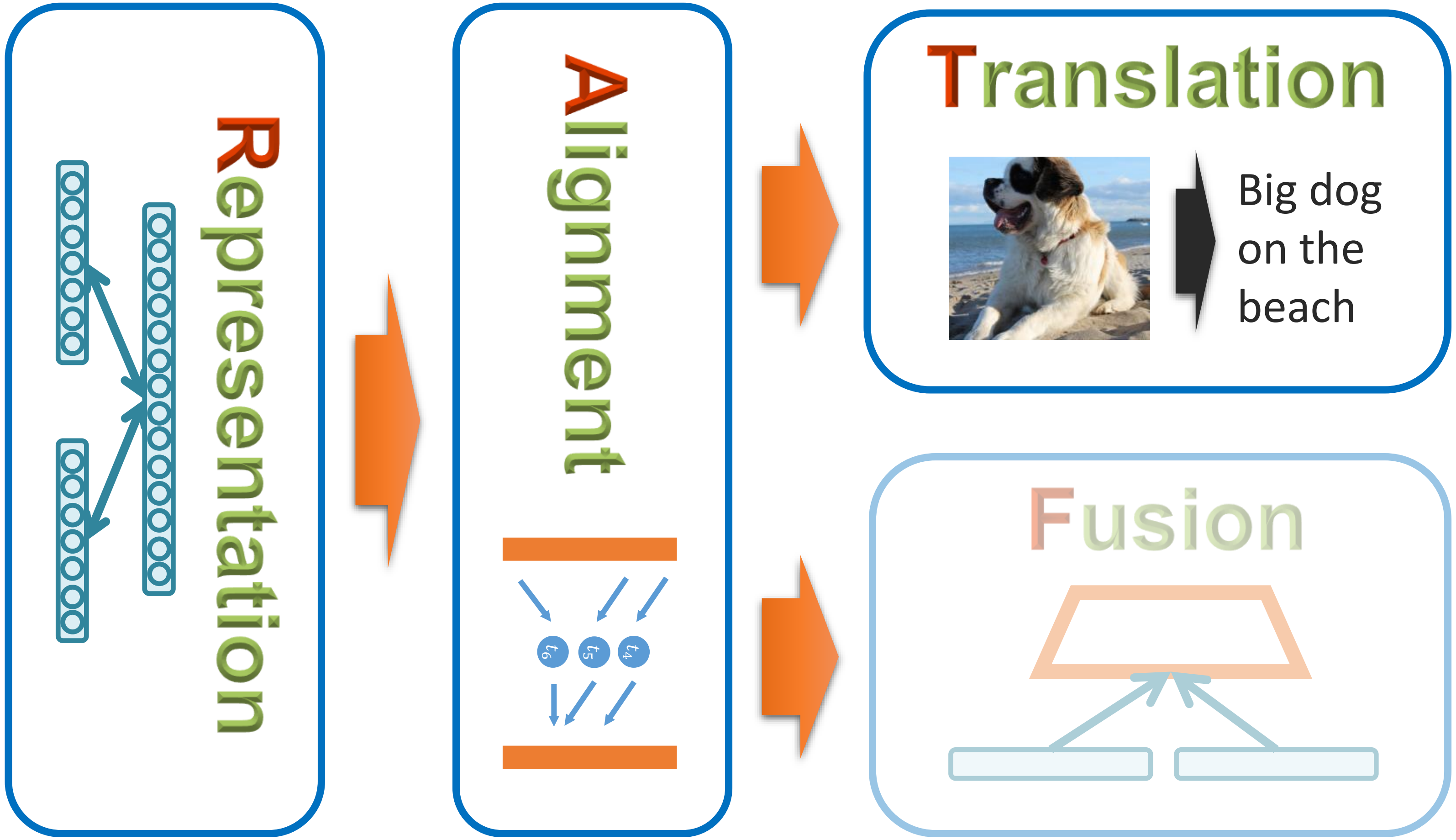
Core Challenge 2: Explicit Alignment

Core Challenge 2: Implicit Alignment



Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

Two More Core Challenges – Conceptual-level Challenges



Core Challenge 3 – Translation

Visual gestures
(both speaker and listener gestures)

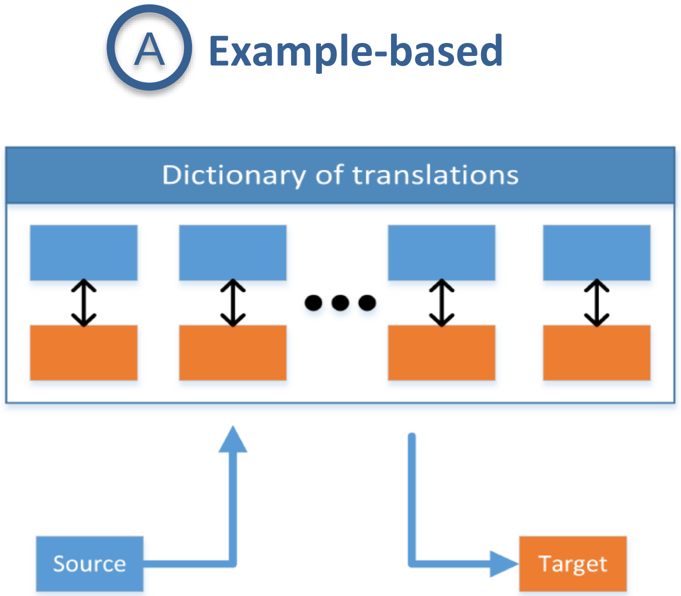
←

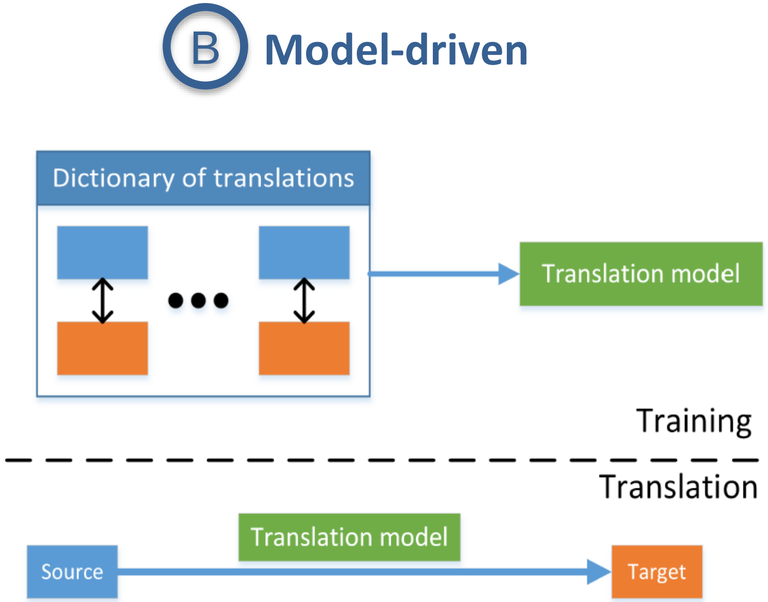
Transcriptions
+
Audio streams

Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013

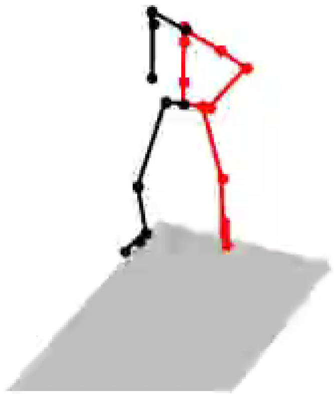
Core Challenge 3: Translation

Definition: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.

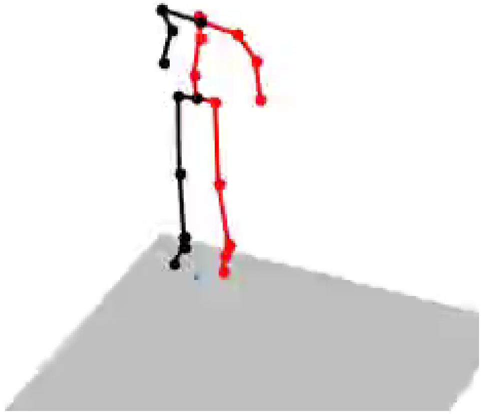




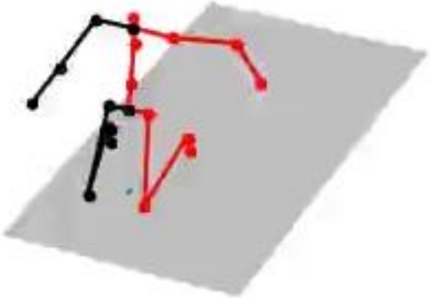
Core Challenge 3: Translation - Example



a person jogs a few steps



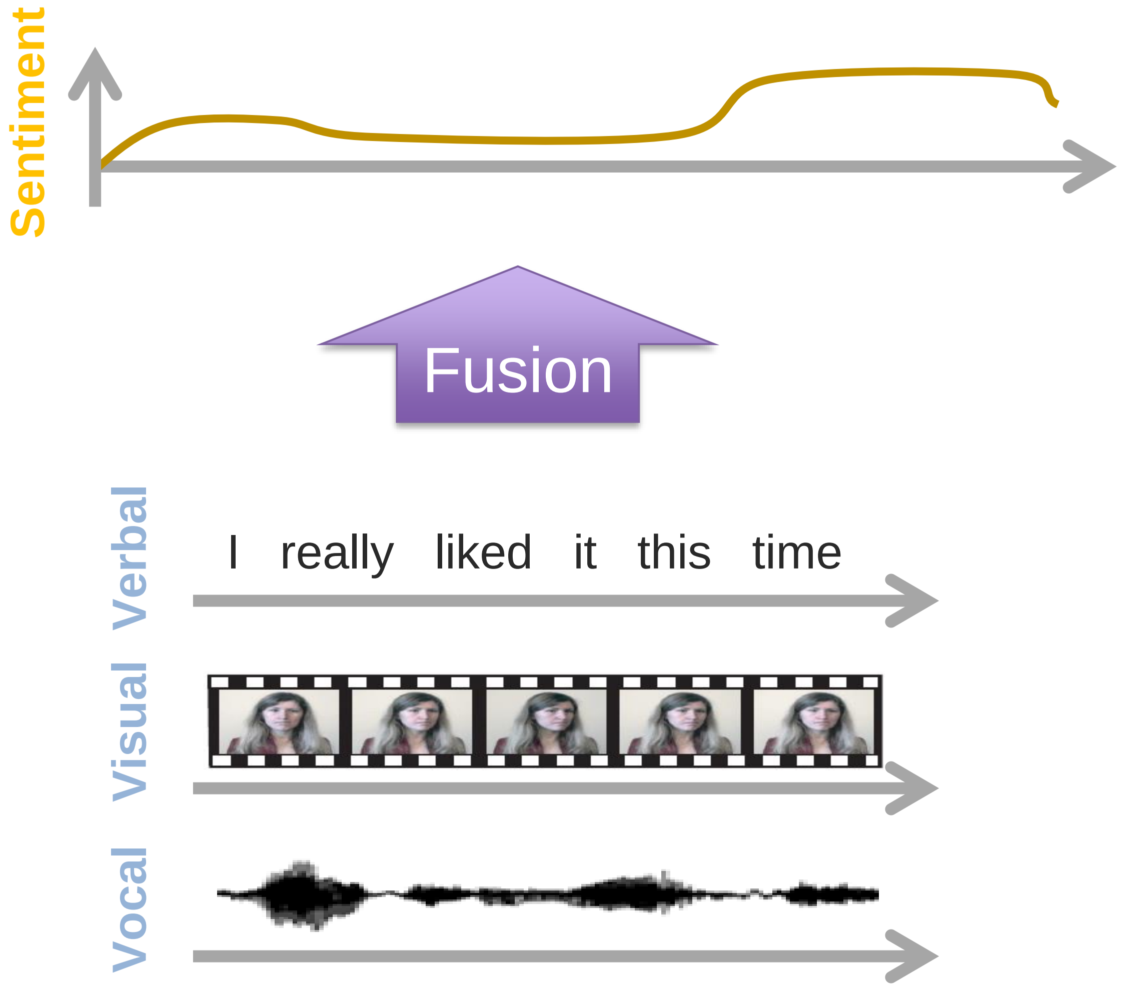
A person steps forward then turns around and steps forwards again.



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

Ahuja, C., & Morency, L. P. (2019). Language2Pose: Natural Language Grounded Pose Forecasting. *Proceedings of 3DV Conference*

Core Challenge 4: Fusion

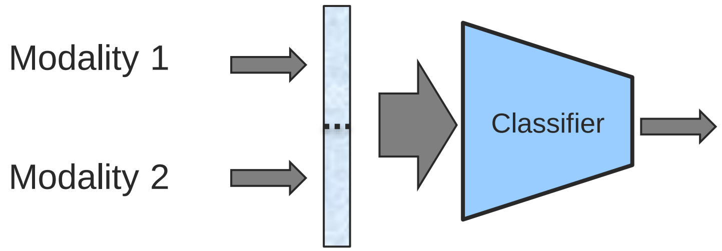


Core Challenge 4: Fusion

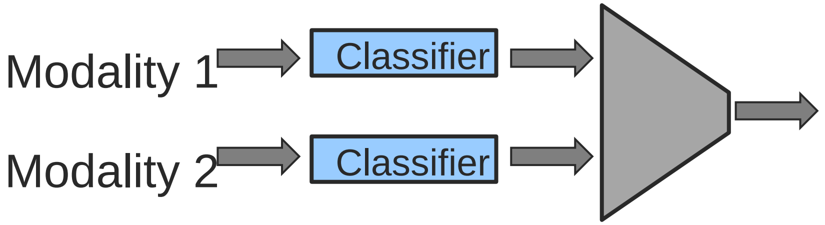
Definition: To join information from two or more modalities to perform a prediction task.

(A) Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion

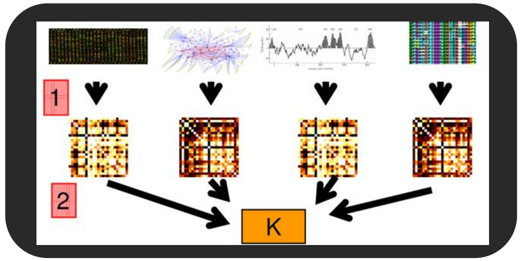


Core Challenge 4: Fusion

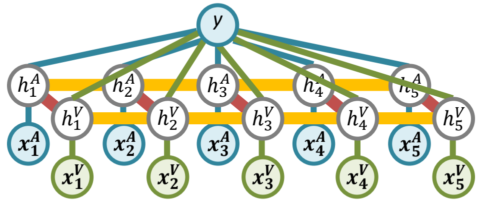
Definition: To join information from two or more modalities to perform a prediction task.

Ⓑ **Model-Based (Intermediate) Approaches**

1) **Deep neural networks**
2) **Kernel-based methods**
3) **Graphical models**

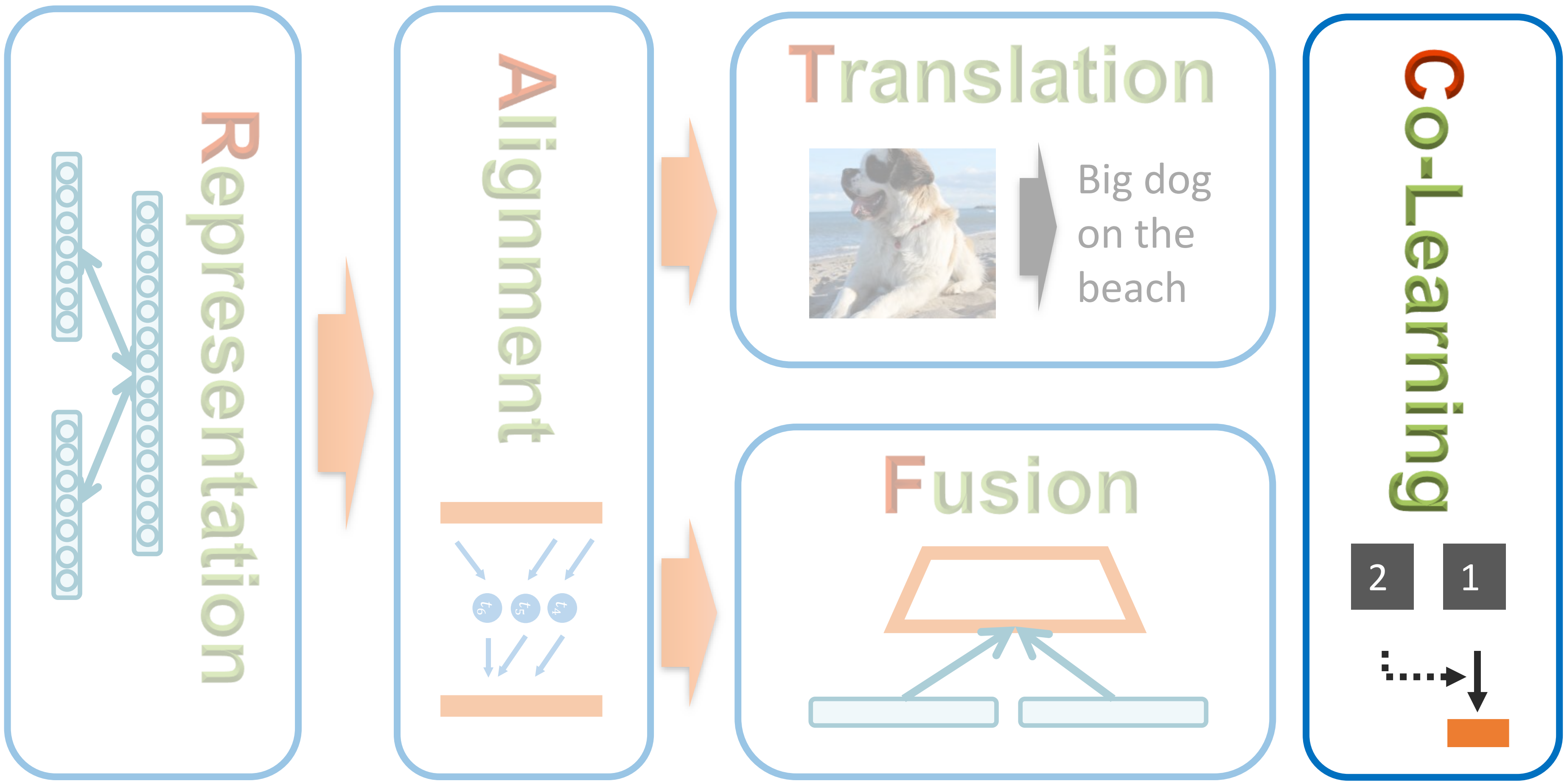


Multiple kernel learning



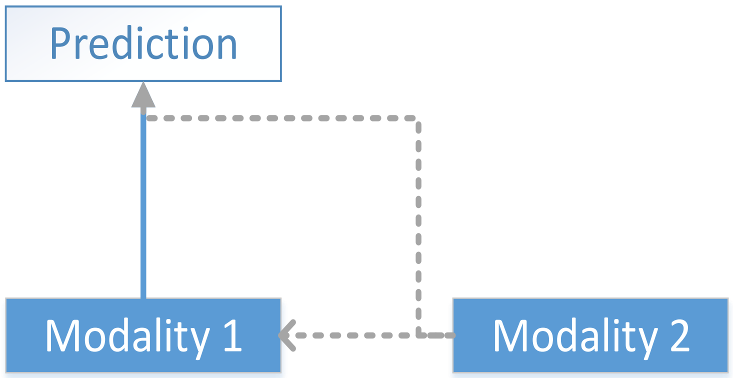
Multi-View Hidden CRF

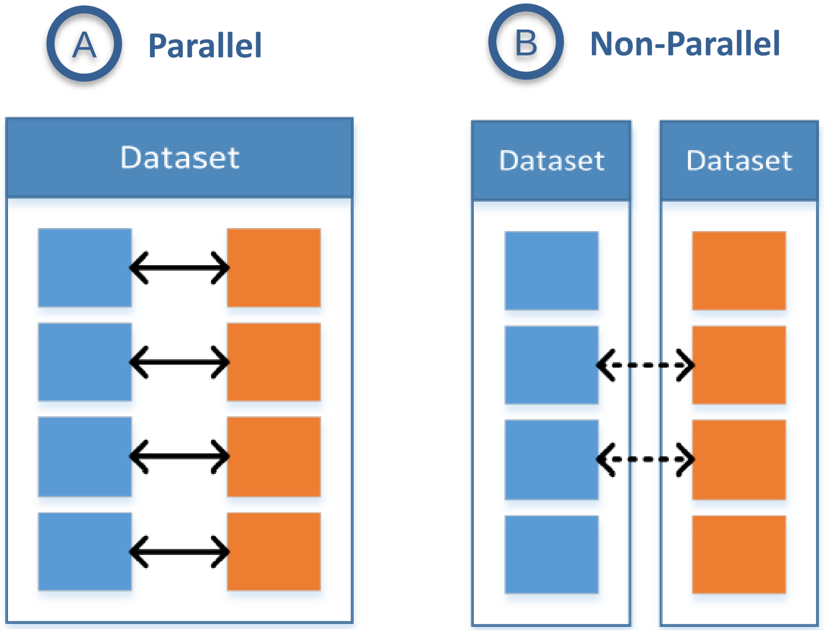
One Last Core Challenge



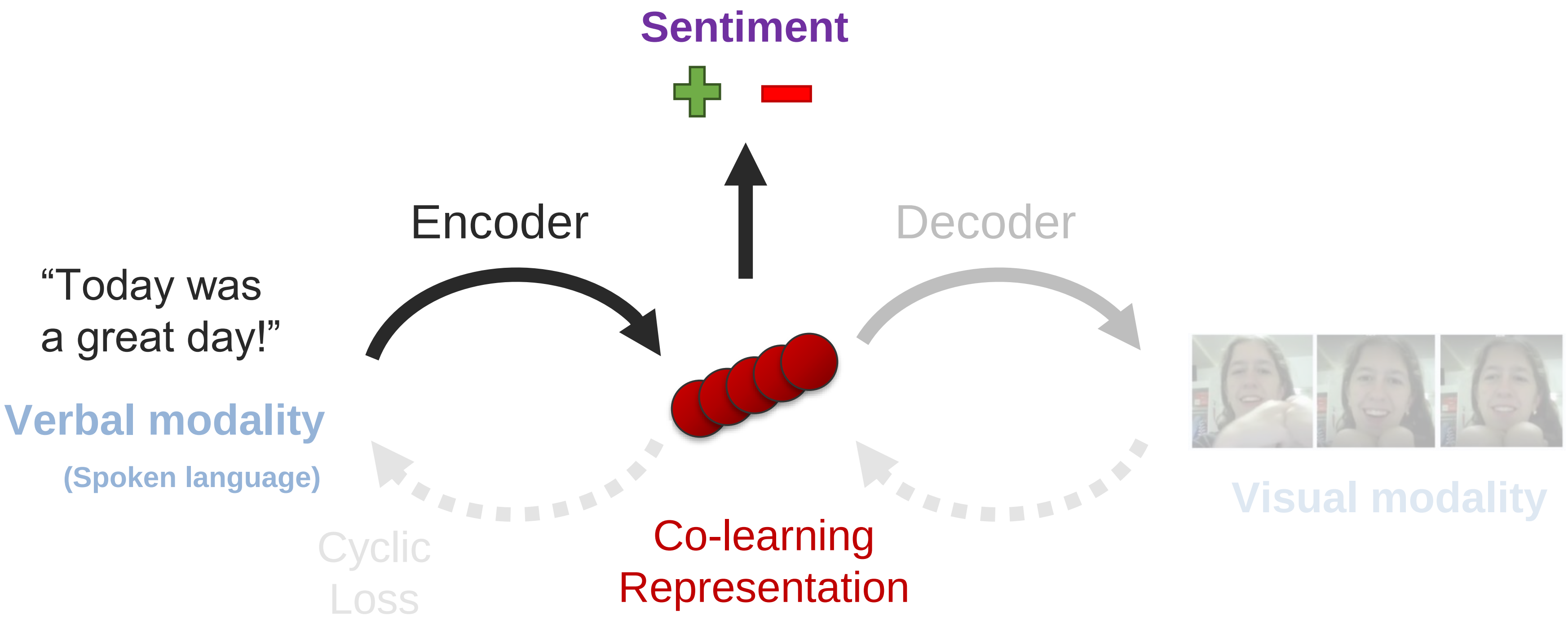
Core Challenge 5: Co-Learning

Definition: Transfer knowledge between modalities, including their representations and predictive models.



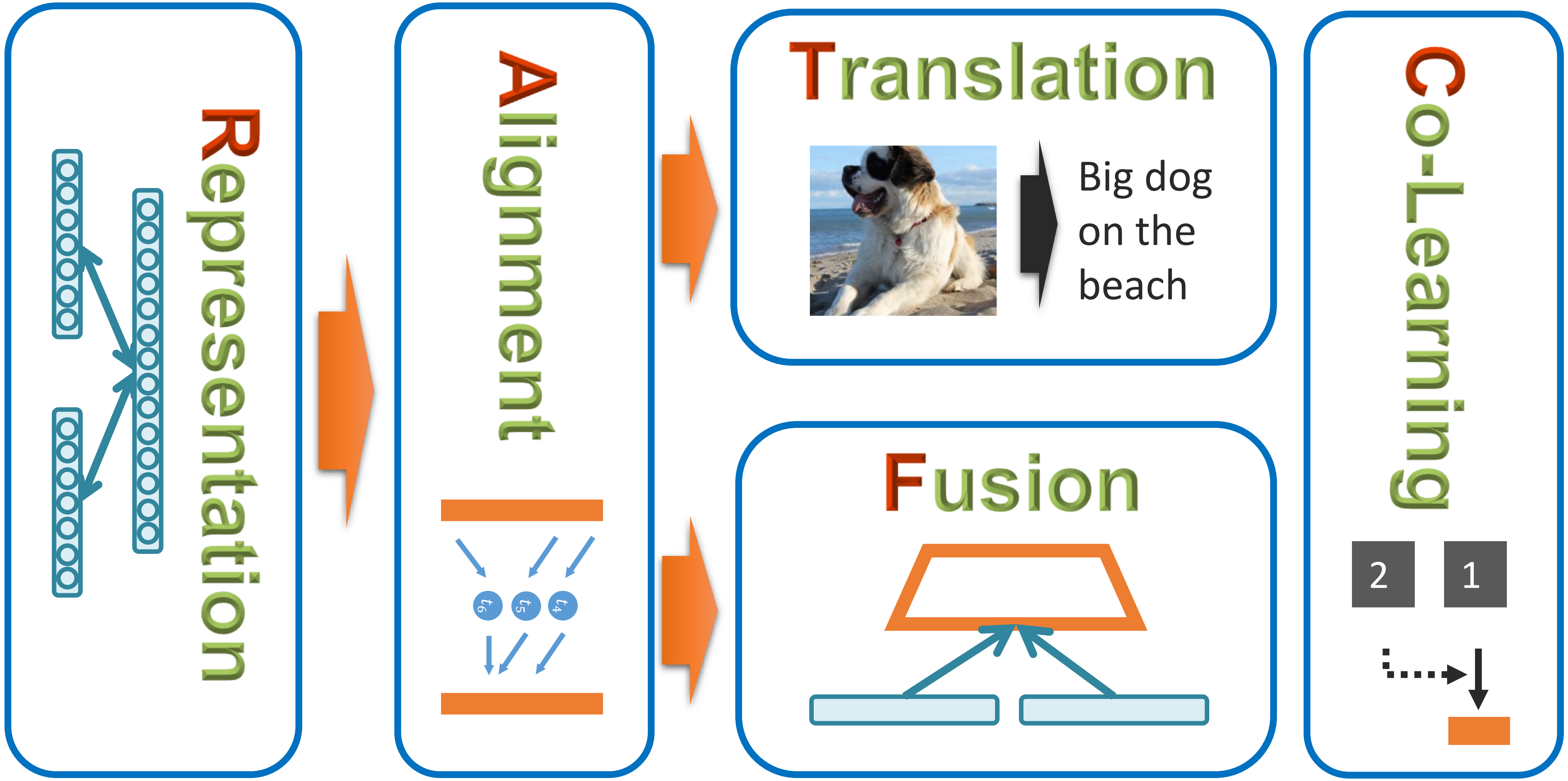


Core Challenge 5: Co-Learning



Pham et al., Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities, https://arxiv.org/abs/1812.07809

Five Multimodal Core Challenges



Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Taxonomy of Multimodal Research

[https://arxiv.org/abs/1705.09406]

Representation

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

Translation

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

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*

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

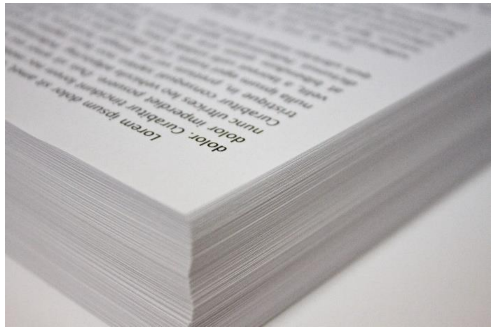
Real world tasks tackled by MMML

- Affect recognition
  - Emotion
  - Persuasion
  - Personality traits
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

Course Syllabus

Three Course Learning Paradigms

Course lecture participation
(15% of your grade)



Reading assignments
(12% of your grade)

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)\\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)\\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

Course project assignments
(73% of your grade)

Course Recommendations and Requirements

1. Ready to read about 6 papers this semester !
   - Curated list of research papers for the 6 reading assignments
   - Summarize one paper and contrast it with other papers
2. Already taken a machine learning course
   - Strongly recommended for students to have taken an introduction machine learning course
   - 10-401, 10-601, 10-701, 11-663, 11-441, 11-641 or 11-741
3. Motivated to produce a high-quality course project
   - Projects are designed to enhance state-of-the-art algorithms
   - Four project assignments, to help scaffold the project tasks

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)\\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)\\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

Course Project Timeline

Pre-proposal *(due Wednesday Sept. 15)*

- Define your dataset, research task and teammates

First project assignment *(due Sunday Sept. 26)*

- Study related work to your selected research topic

Second project assignment *(due Sunday Oct 10)*

- Experiment with unimodal representations

Midterm project assignment (due Monday Nov. 1)

- Implement and evaluate state-of-the-art model(s)

Final project assignment (due Sunday Dec. 5)

- Implement and evaluate new research ideas

Course Project Guidelines

- Dataset should have at least two modalities:
  - Natural language and visual/images
- Teams of 3, 4 or 5 students
- The project should explore algorithmic novelty
- Possible venues for your final report:
  - NAACL 2022, ACL , IJCAI 2022, ICML 2022, ICMI 2022
- We will discuss on Thursday about project ideas
- GPU resources available:
  - Amazon AWS and Google Cloud Platform

Process for Selecting your Course Project

- **Thursday 9/2:** Lecture describing available multimodal datasets and research topics
- **Tuesday 9/7:** Let us know your dataset preferences for the course project
- **Thursday 9/19:** During the later part of the lecture, we will have an interactive period to help with team formation. More details to come
- **Wednesday 9/15:** Pre-proposals are due. You should have selected your teammates, dataset and task

Equal Contribution by All Teammates!

- Each team will be required to create a GitHub repository which will be accessible by TAs
- Each report should include a description of the task from each teammate
- Please let us know soon if you have concerns about the participation levels of your teammates

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 8/31 & 9/2	**Course introduction** • Research and technical challenges • Course syllabus and requirements	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 *(read)* 9/7 & 9/9 Due: 9/10, 9/13	**Basic concepts: neural networks** • Language, visual and acoustic • Loss functions and neural networks	**Basic concepts: network optimizat** • Gradients and backpropagatio • Practical deep model optimizat
Week 3 *(read)* 9/14 & 9/16 Due: 9/17, 9/20	**Visual unimodal representations** • Convolutional kernels and CNNs • Residual network and skip connection	**Language unimodal representatio** • Language models • Gated recurrent networks
Week 4 *(proj)* 9/21 & 9/23 *Assign. due: 9/26*	***Project hours (first assignment)***	**Multimodal representation learni** • Multimodal auto-encoders • Multiview clustering
Week 5 *(read)* 9/28 & 9/30 Due: 10/1, 10/4	**Multimodal alignment** • Explicit - dynamic time warping • Implicit - attention models	**Alignment and representation** • Self-attention models • Pretrained models
Week 6 *(proj)* 10/5 & 10/7 *Assign. due: 10/10*	***Project hours (second assignment)***	**Alignment and representation** • Multimodal transformers • Video-based alignment

Project preferences due on Tuesday 9/8

Pre-proposals due on Wednesday 9/16

First assignment due on Sunday 9/26

Second assignment due on Sunday 10/10

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 *(read)* 10/12 & 10/14 Due: 10/15, 10/18	**Alignment and translation** • Module networks • Tree-based and stack models	***Mid semester Break – No Class –***
Week 8 *(read)* 10/19 & 10/21 Due: 10/22, 10/25	**Graphical and Generative Models** • Probabilistic graphical models • Generative adversarial networks	***Project hours (midterm report)***
Week 9 *(proj)* 10/26 & 10/28 *Assign. due: 10/31*	**Language, Vision and Actions** • Action as a modality • Embodied language grounding	**Fusion and co-learning** • Multi-kernel learning and fusio • Few shots learning and co-lear
Week 10 11/2 & 11/4	***Project presentations (midterm)***	***Project presentations (midterm)***
Week 11 *(read)* 11/9 & 11/11 Due: 11/12, 11/15	**Reinforcement learning** • Markov decision process • Q learning and policy gradients	**Multimodal RL** • Deep Q learning • Multimodal applications
Week 12 11/16 & 11/18	**New research directions** • Recent approaches in multimodal ML	***Project Hours (final report)***

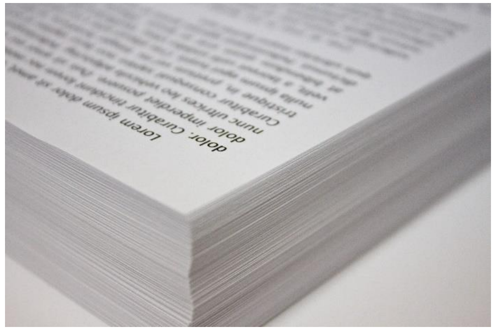
Midterm assignment due on Sunday 10/31

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 13 11/23 & 11/25	***Thanksgiving Week – No Class –***	
Week 14 *(proj)* 11/30 & 12/2 *Assign. due: 12/5*	***Project presentations (final)***	***Project presentations (final)***

Final assignment due on Sunday 12/5

Course Grades



$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$
$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$
$$h_t = o_t \tanh(c_t)$$

- Lecture highlights 15%
- Reading assignments 12%

- Project preferences/pre-proposal 3%
- First project assignment 10%
- Second project assignment 10%
- Mid-term project assignment
  - Report and presentation 20%
- Final project assignment
  - Report and presentation 30%

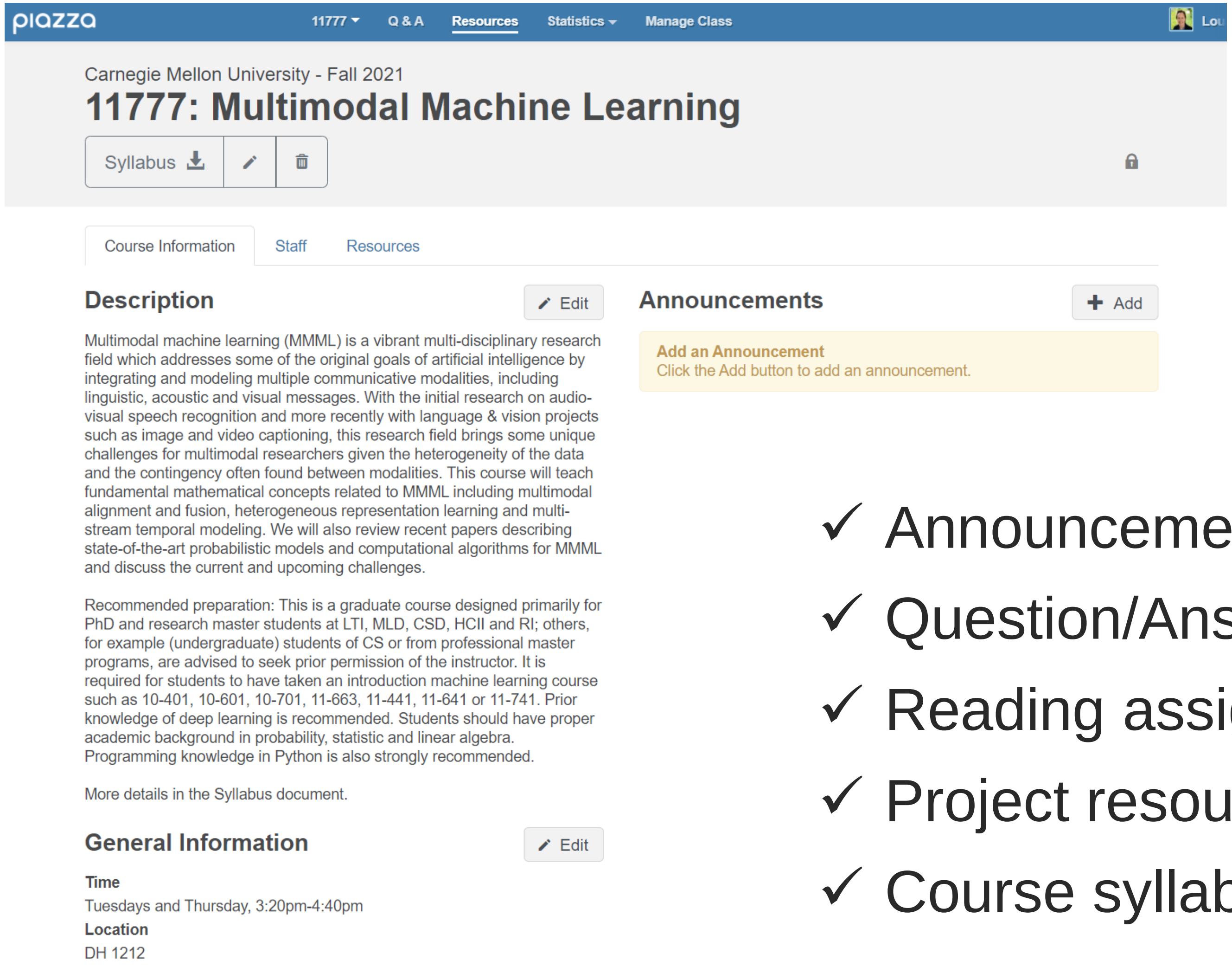
Lecture Participation – Highlight Forms

- Students should summarize lecture highlights
  - Each lecture is split in 3 segments (~30mins each)
  - One highlight statement for each segment
    - This is the main takeaway from this segment
  - Optionally, students can include related question
- Highlights submitted the same day as the lecture
  - Lectures are expected to be in-person
- Questions will be summarized by TAs
  - Answers posted on Piazza

Reading Assignments

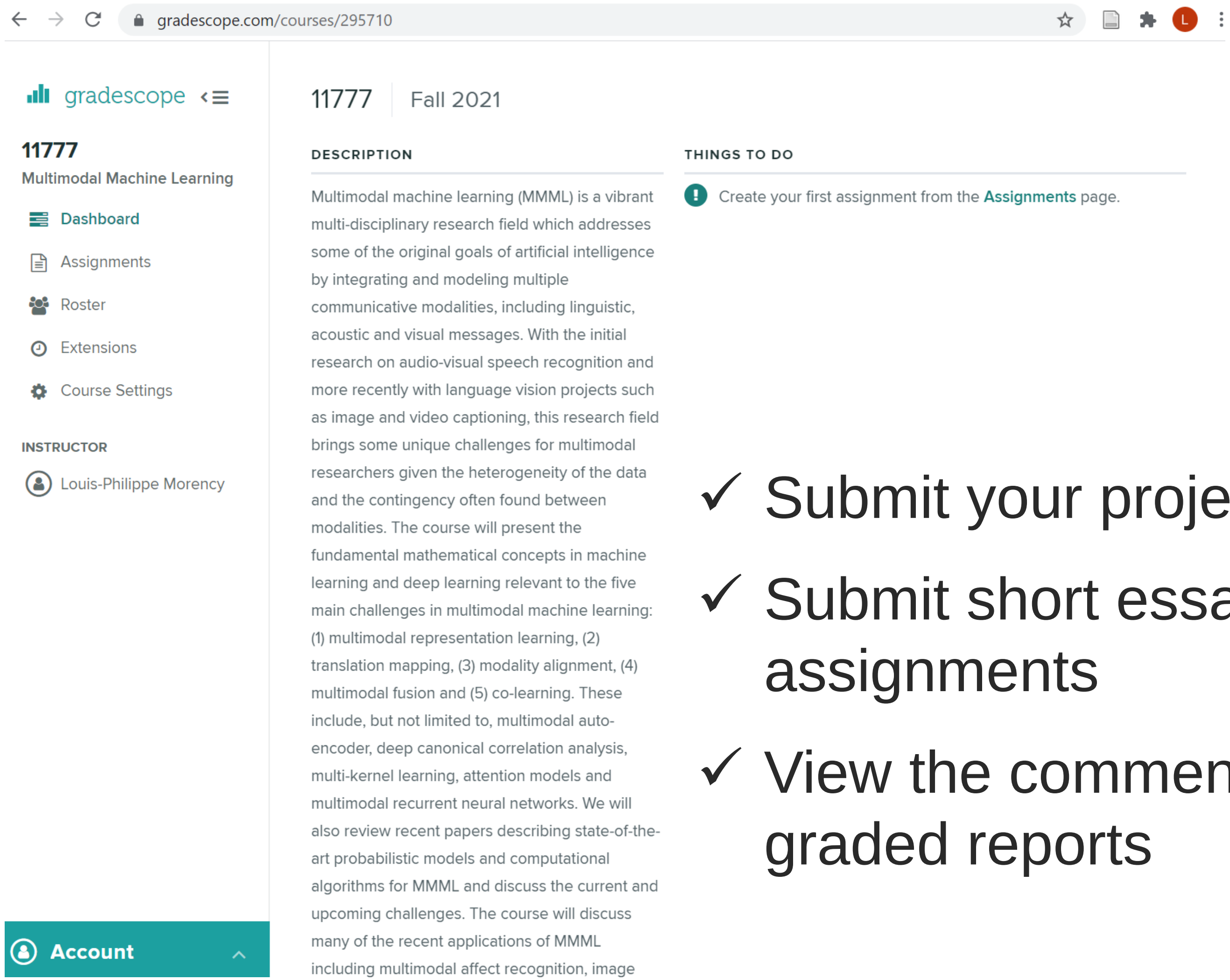
- 3-4 papers for each reading assignment
  - **Each student will read only one paper!**
  - Then you will create a short summary to help others
- Discussions with your study group
  - 9-10 students in each study group
  - Read other's summaries. Ask questions!
  - Write a short essay to compare papers and suggest ideas
- Graded based on summary and discussion
  - 1 point for the summary and 1 point for the short essay

Piazza https://piazza.com/cmu/fall2021/11777/info



- ✓ Announcements
- ✓ Question/Answers
- ✓ Reading assignments
- ✓ Project resources
- ✓ Course syllabus

Gradescope



- ✓ Submit your project assignments
- ✓ Submit short essays from reading assignments
- ✓ View the comments from your graded reports

Spring 2022 Edition of the MMML Course !

Yonatan Bisk

ybisk@cs.cmu.edu

https://yonatanbisk.com/

More details about the Spring edition to come later!

Project Preferences – Due Tuesday 9/7

- Post your project preferences:
  - List of your ranked preferred projects
    - Use alphanumeric code of each dataset
    - Detailed dataset list in the "Lecture1.2-datasets“ slides
  - Previous unimodal/multimodal experience
  - Available CPU / GPU resources
- For topics or datasets not in the list:
  - Include a description with links (for other students)

https://piazza.com/cmu/fall2021/11777/info