Multimodal AI

Lecture 1 - Introduction

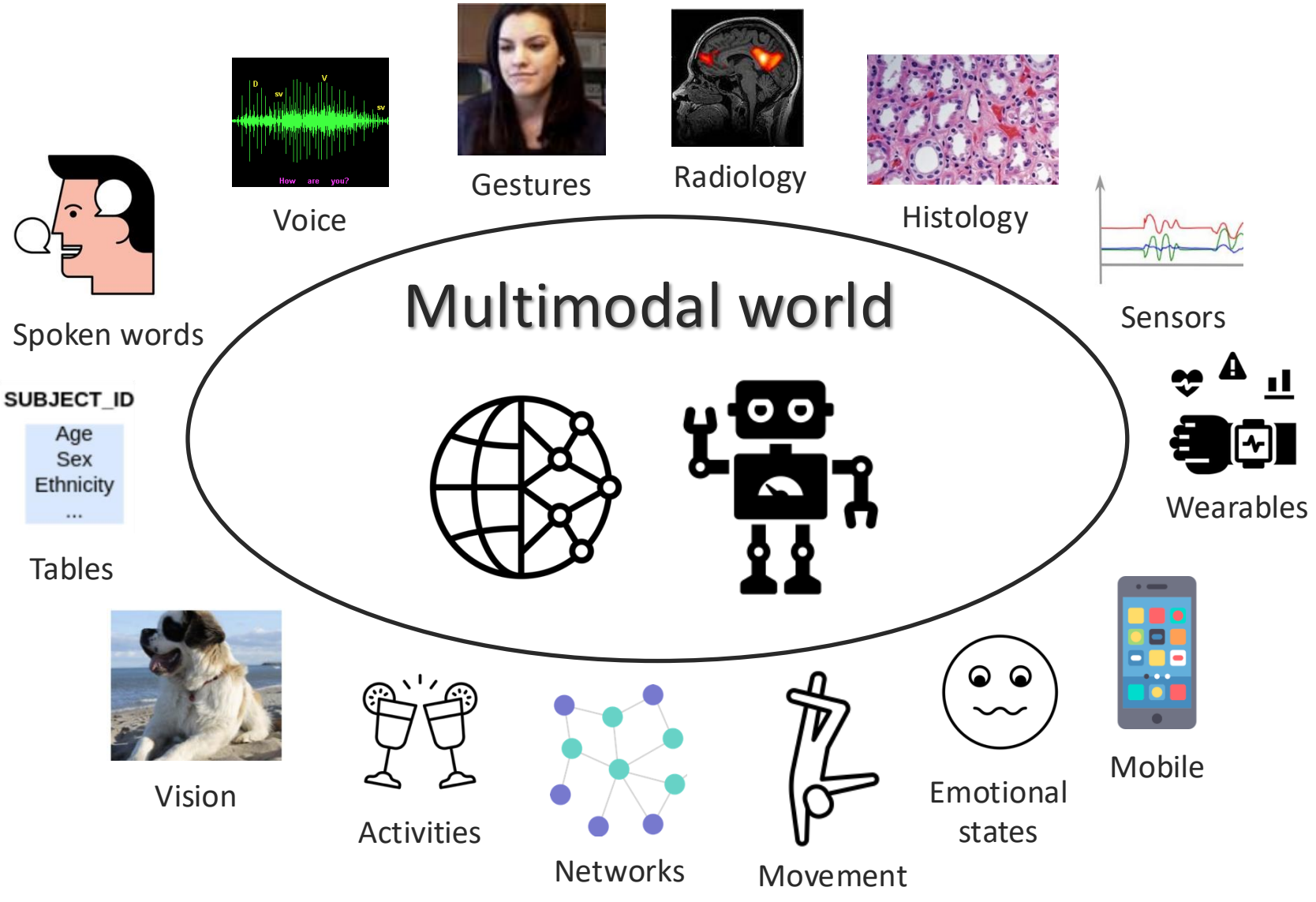
Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

mit media lab

MIT EECS

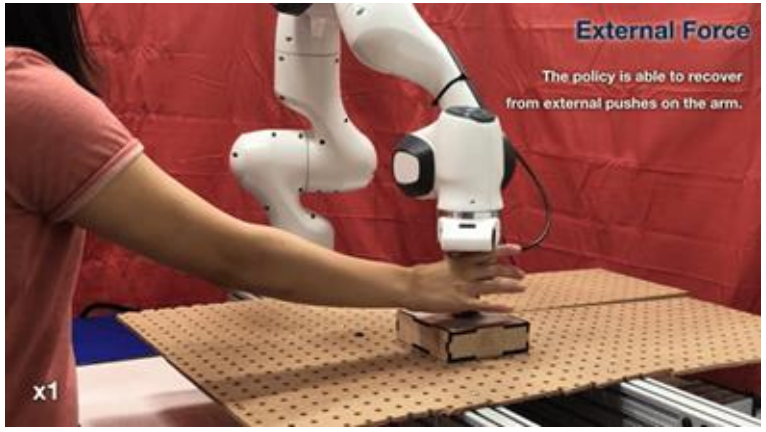
Multimodal AI

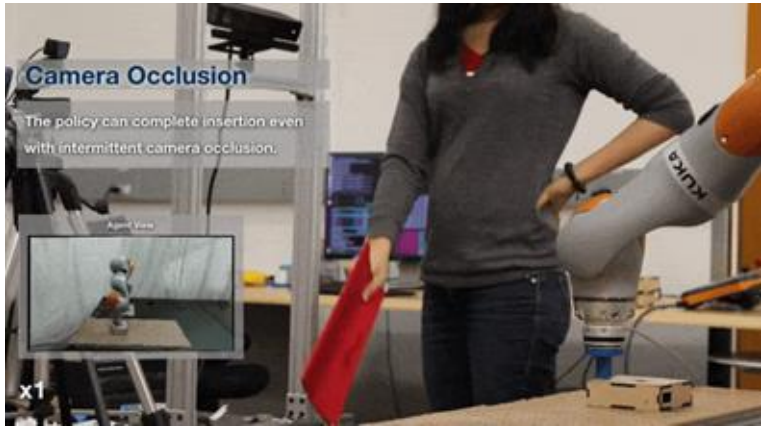


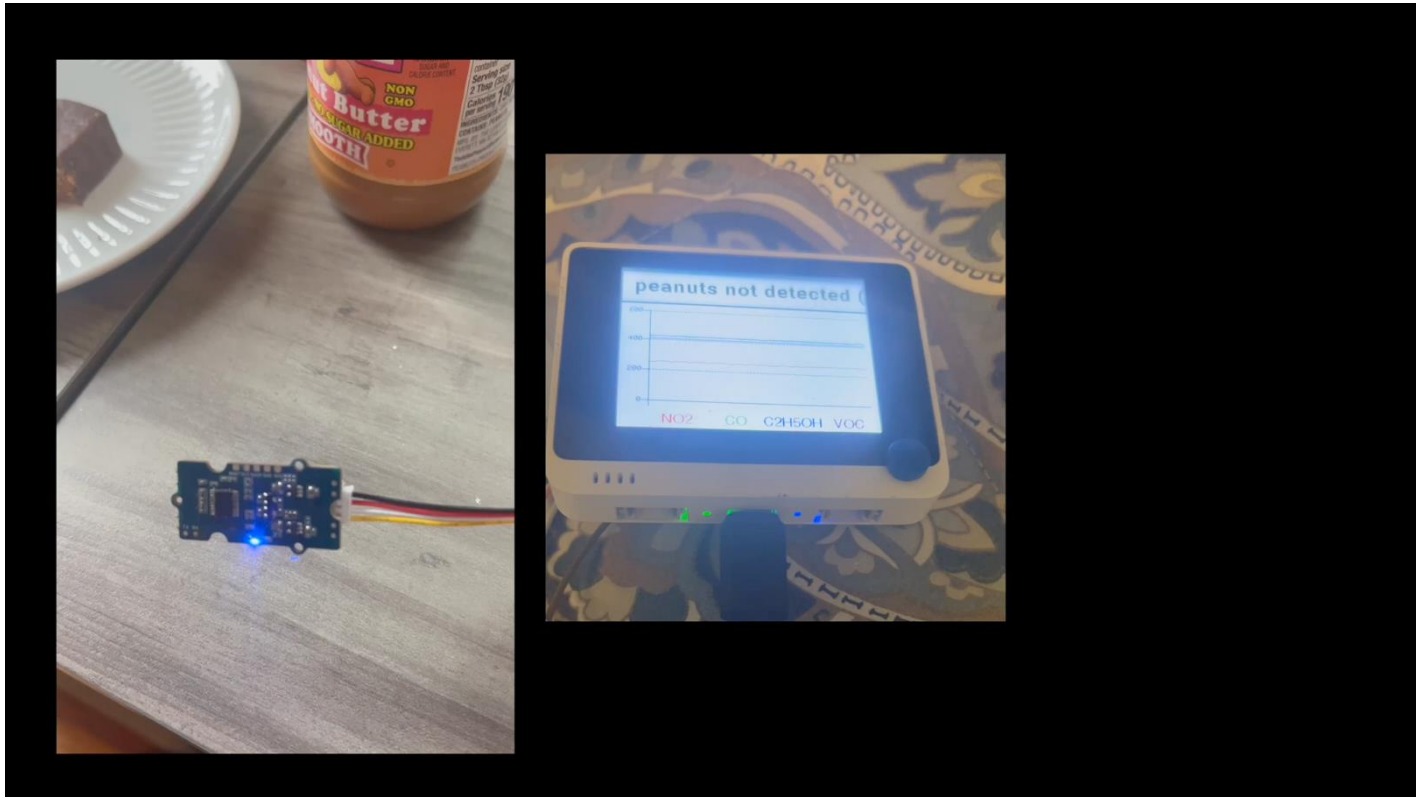
multisensory intelligence

AI for Physical Sensing

Sensing in physical systems, manufacturing, smart cities, IoT, robotics





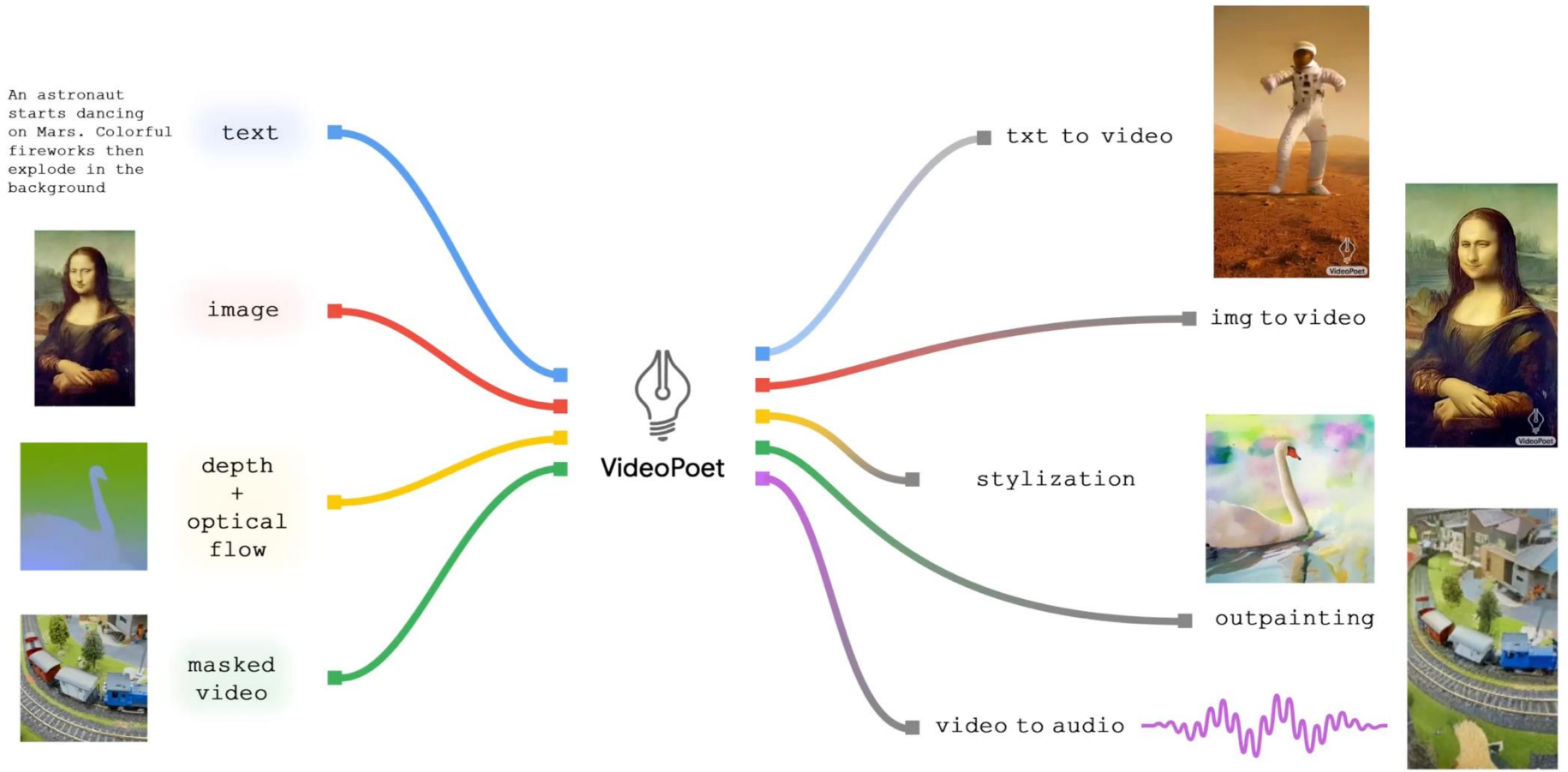


[Lee et al., Making Sense of Vision and Touch: Learning Multimodal Representations for Contact Tasks. ICRA 2019]
[Feng et al., SmellNet: A Large-scale Hierarchical Database for Real-world Smell Recognition. ICLR 2026]

multisensory intelligence

Multimodal Generative AI

Multimedia, content creation, creativity and the arts

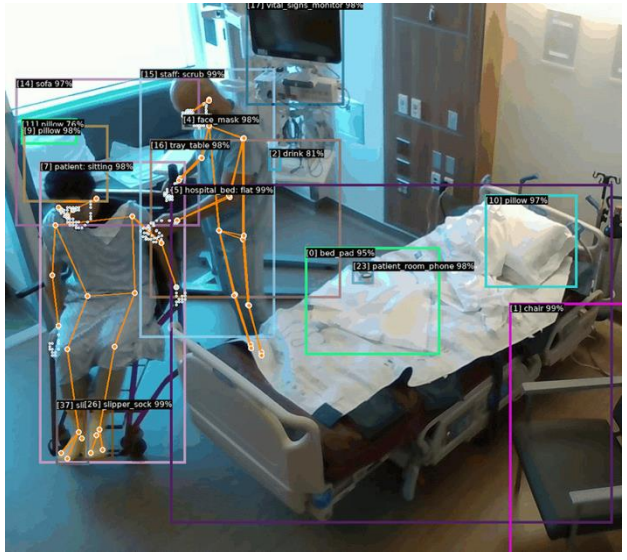


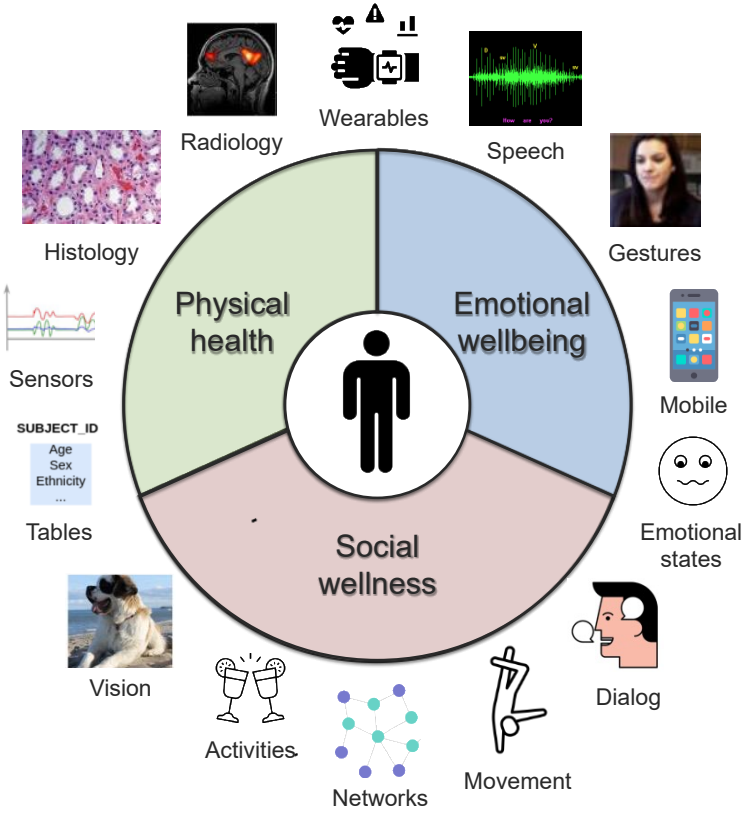
[Kondratyuk et al., VideoPoet: A Large Language Model for Zero-shot Video Generation. ICML 2024]

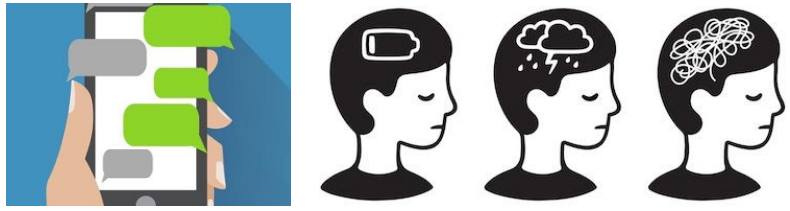
multisensory intelligence

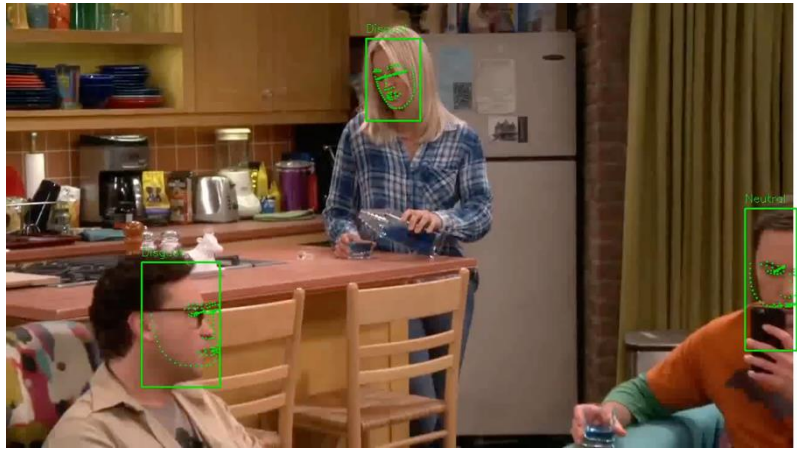
Holistic Health: Physical, Social, and Emotional

Majority of medical indicators will not be taken in the doctor's office









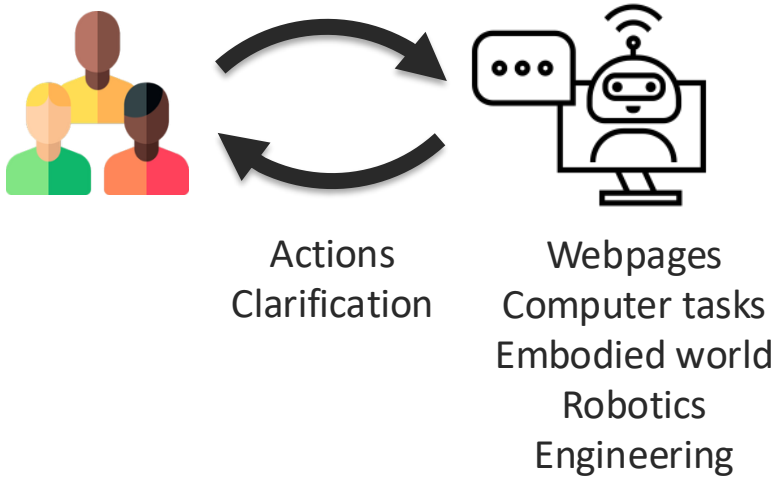
[Dai et al., Clinical Behavioral Atlas. NEJM AI 2025]
[Hu et al., OpenFace 3.0: An Open-source Foundation Model for Facial Behavior Analysis. FG 2025]
[Mathur et al., Advancing Social Intelligence In AI: Technical Challenges and Open Questions. EMNLP 2024]

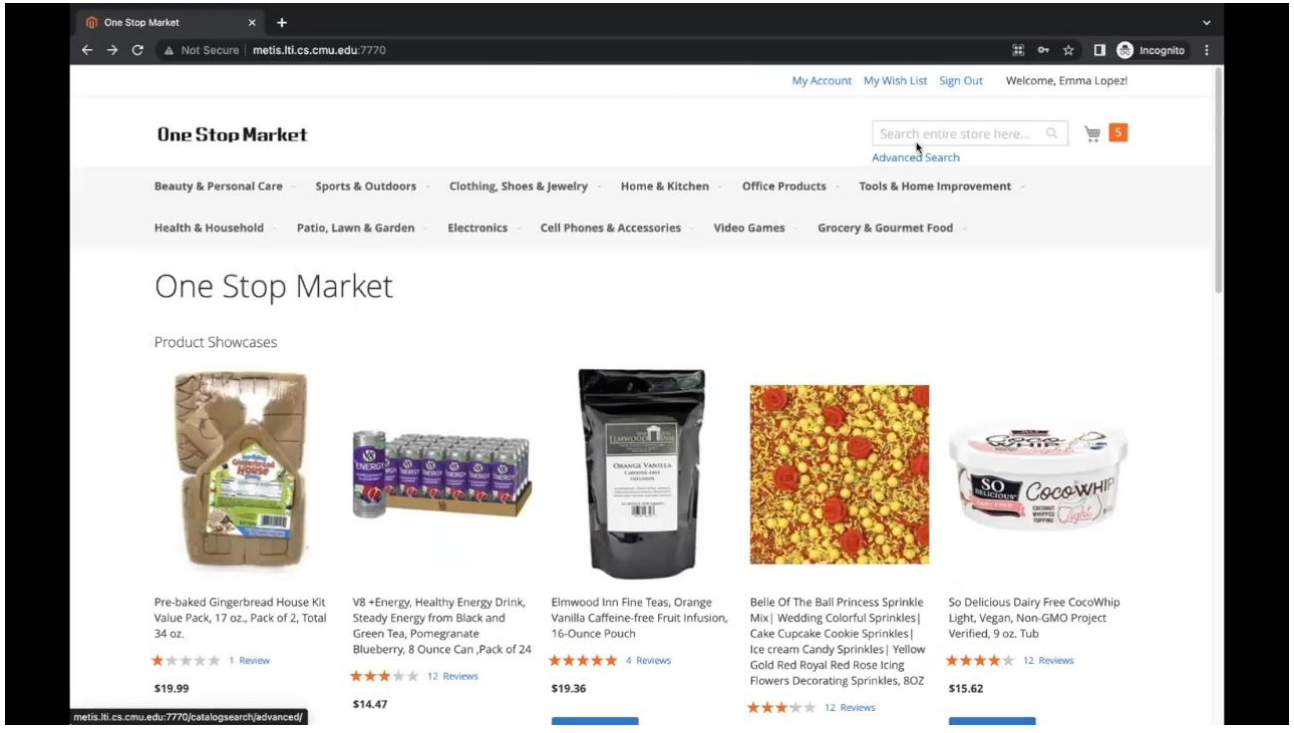
multisensory intelligence

Interactive Multimodal Agents

AI agents for the web and digital automation

Example task: Purchase a set of earphones with at least 4.5 stars in rating and ship it to me.

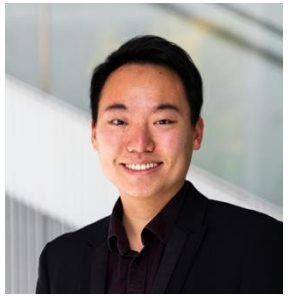




[Zhou et al., WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR 2024]
[Jang et al., VideoWebArena: Evaluating Multimodal Agents on Video Understanding Web Tasks. ICLR 2025]

multisensory intelligence

Your Teaching Team, Spring 2026



Paul Liang
ppliang@mit.edu
Course instructor

Dimitris Bertsimas
dbertsim@mit.edu
Course instructor

Jinhua Zhao
jinhua@mit.edu
Course instructor

Sang-Gook Kim
sangkim@mit.edu
Course instructor

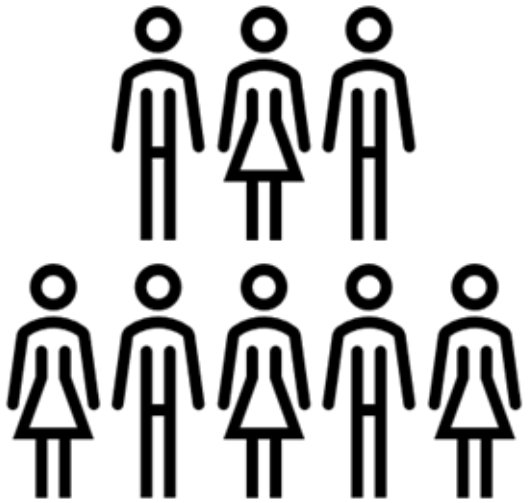
Edgar Morfin
emorfin@mit.edu
Teaching Assistant

Valdemar Danry
vdanry@mit.edu
Teaching Assistant

David Dai
dvdai@mit.edu
Teaching Assistant

multisensory intelligence

Time for Introductions!



Your name, department and programs

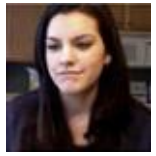
Your favorite modality(ies)!

Previous research experience in AI

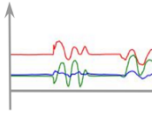
Why are you interested in this course?

Course Overview

1. AI for new modalities: data, modeling, evaluation, deployment


Gestures


Radiology

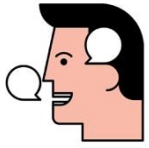

Sensors

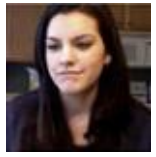

Wearables

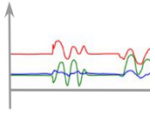

Mobile


Networks

2. Multimodal AI: connecting multiple different data sources


Language


Gestures


Sensing


Actuation

multisensory intelligence

Learning Objectives

1. Study recent technical achievements in AI research
2. Improve critical and creative thinking skills
3. Understand future research challenges in AI
4. Explore and implement new research ideas in AI

Preferred Pre-requisites

1. Some knowledge of programming (ideally in Python)
2. Some basic understanding of modern AI capabilities & limitations
3. Bring external (non-AI) domain knowledge about your problem
4. Bonus: worked on AI for some modality

Course delivery format

- 1.5-hour class every Tuesday and Thursday, 230pm to 4pm.
- Homework and reading assignments outside of class, 2-week duration.
- In-class midterm exam, with short and long answer questions.
- Significant research project outside of class, with reports and presentations.

Lecture Topics

(subject to change, based on student interests and course discussions)

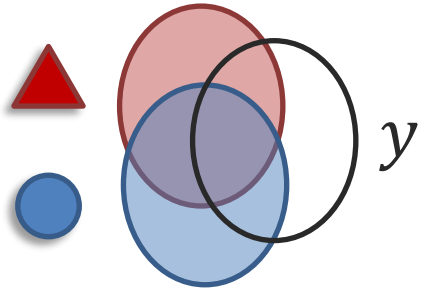
Module 1: Introduction to multimodal AI

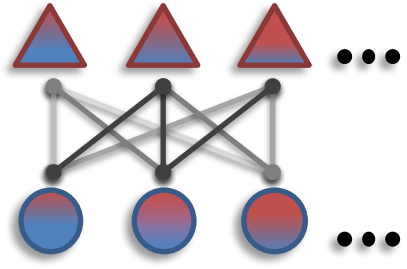
Week 1 (2/3 & 2/5): Introduction to multimodal AI and research tasks

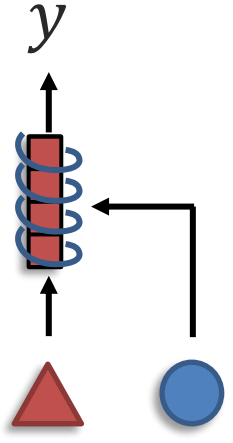
Week 2 (2/10 & 2/12): Unimodal data, structure, and information

Week 3 (2/17 & 2/19): Multimodal fusion

Week 4 (2/24 & 2/26): Multimodal representations







multisensory intelligence

Lecture Topics *(subject to change, based on student interests and course discussions)*

Module 2: Multimodal foundation models

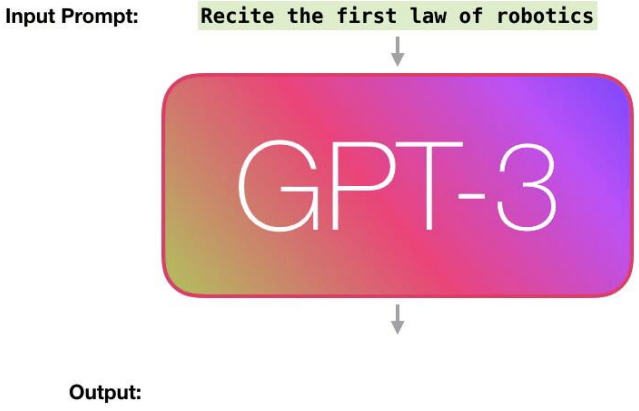
Week 5 (3/3 & 3/5): Multimodal foundation models

Week 6 (3/10 & 3/12): Multimodal alignment & transfer

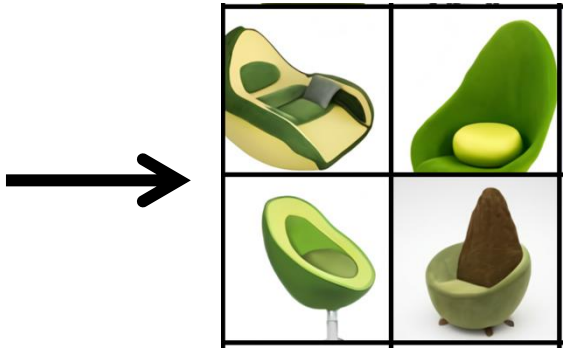
Week 7 (3/17 & 3/19): Multimodal generation

Week 9 (3/31 & 4/2): Multimodal reasoning

Week 10 (4/7 & 4/9): Multimodal interaction & agents



An armchair in the shape of an avocado



multisensory intelligence

Lecture Topics *(subject to change, based on student interests and course discussions)*

Module 3: Multimodal application domains

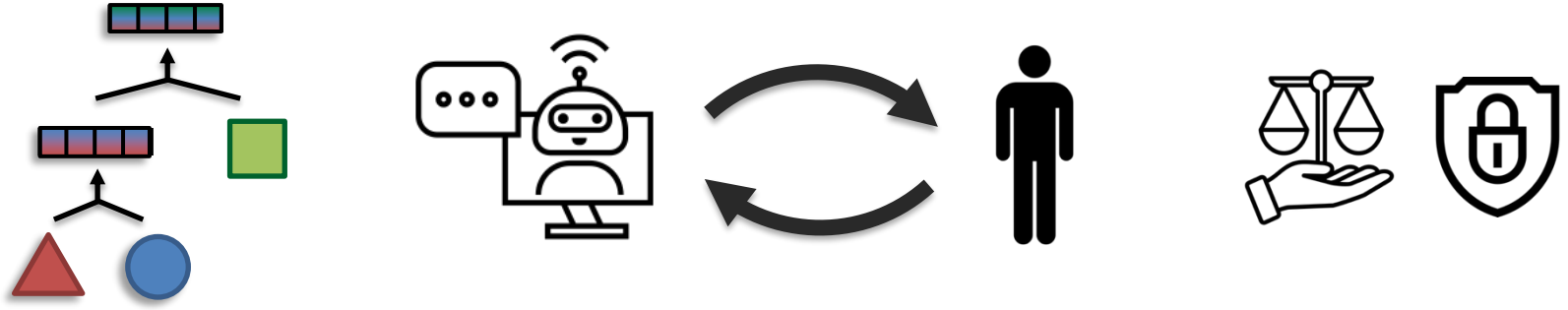
Week 11 (4/14 & 4/16): Multimodal in design

Week 12 (4/21 & 4/23): Multimodal in manufacturing

Week 13 (4/28 & 4/30): Multimodal in cities & transportation

Week 14 (5/5 & 5/7): New research directions

Week 15 (5/12): Project final presentations



multisensory intelligence

Grading Overview

- 35% for homework assignments:
  - 5 homework assignments, roughly 2 weeks for each.
- 15% for reading assignments:
  - 5 reading assignments covering state-of-the-art research in multimodal AI, released with the homework, roughly 2 weeks for each.
- 15% for a in-class midterm exam.
- 35% for a high-quality research project:
  - 5% for proposal with literature review
  - 10% for midterm progress report
  - 20% for final report and presentation

Homework Assignments

HW 1: Multimodal data processing and visualization.
- Find, process, visualize, and label your unique multimodal dataset of interest.

HW 2: Multimodal fusion and supervised learning.
- Train unimodal and multimodal models from scratch to get good predictive performance.

HW 3: Adapting and fine-tuning multimodal LLMs.
- Use LLMs to ask and answer open-ended questions about your data modality.

HW 4: Multimodal reasoning with reinforcement learning.
- Solving harder tasks requiring interpretable and multi-step thinking.

HW 5: Multimodal interactive agents.
- Multi-turn systems that act on the environment grounded in your data modality.

Reading Assignments

- 5 readings assignments, with usually 2 required papers and some suggested (but optional) papers, and 5-6 discussion probes.
- Three main assignment parts:
  - **Reading notes:** Read the assigned papers and summarize the main take-away points
    - *Optional: if you have clarification questions about the papers*
  - **Paper scouting:** Scout for extra papers, blog posts or other resources related to these question probes
  - **Discussion points:** Reflect on the question probes related to the reading papers and prepare discussion points.

Course Project

- Meant to introduce students to the research process:
  - Reading literature
  - Becoming familiar with application domains
  - Trying state-of-the-art models
  - Performing error analysis
  - Developing and iterating on new methods
  - Evaluating and analyzing results
- Done in groups of 2-3 students.
- Group will be assigned a mentor.
- Equal contribution tracked via regular updates, presentations, and GitHub.

Course Project Timeline

- Week 2: **Project preferences:** you should have selected your teammates, have ideas about your dataset and task
- Week 4: **Proposal report:** literature review and research ideas
- Week 6: Setting up datasets and baselines
- Week 9: **Midterm report:** baselines and initial results for new ideas
- Week 11: Updated results for research idea
- Week 13: Error analysis and ablations
- Week 15: **Final presentations:** describing explored research ideas with results, analysis, and discussion
- Week 16: **Final report**

Absences and Late Submissions

- Lectures are not recorded, students expected to attend live
  - If you plan to miss more than one lecture this semester, let us know as soon as possible.
- Homework and reading assignment wildcards (2 per student)
  - 24-hours extension, max 1 per assignment
- Project report wildcards (2 per team)
  - Either for proposal, midterm, or final report
  - 24-hours extension, can be used together

Course Websites

- Course website
  - A public version of the course information
  - https://mit-mi.github.io/mmai-course/spring2026/
- We will setup canvas for submissions

Course Websites

Classes	Tuesday Lectures	Thursday Lectures	HWs
Week 1 2/3 & 2/5	**Course introduction (All)** • Multimodal core challenges • Course syllabus	**Multimodal datasets (Paul)** • Research tasks and datasets • Intro to AI research	**HW1 out**
Week 2 2/10 & 2/12	**Datasets tutorial** • Data processing and visualization • Pytorch and modeling	**Unimodal representations (Paul)** • Dimensions of heterogeneity • Common model architectures	
Week 3 2/17 & 2/19	**Multimodal fusion (Dimitris)** • Cross-modal interactions • Early and late fusion	**Explainable fusion (Dimitris)** • Dynamic and explainable fusion • Attention-based fusion	**HW1 due** **HW2 out**
Week 4 2/24 & 2/26	**Fusion tutorial** • Higher-order interactions • Multimodal fusion models	**Multimodal transformers (Paul)** • Self-attention & transformers • Multimodal transformers	
Week 5 3/3 & 3/5	**Multimodal foundation models (Paul)** • Multimodal pre-training • Multimodal fine-tuning	**Multimodal LLMs tutorial** • Instruction tuning • Fine-tuning approaches	**HW2 due** **HW3 out**
Week 6 3/10 & 3/12	**Multimodal alignment (Paul)** • Multimodal grounding • Aligned representations	**Cross-modal transfer (Paul)** • Modality transfer and co-learning • Self-training and multitask learning	
Week 7 3/17 & 3/19	**Multimodal generation (Paul)** • Translation, summarization, creation • Model evaluation and ethics	***In-class midterm***	**HW3 due**
Week 8 3/24 & 3/26	***Spring Break – No lectures***		

multisensory intelligence

Behavioral Study of Multimodal

Language
and gestures

David McNeill

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

McGurk effect

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Behavioral Study of Multimodal

Language
and gestures

David McNeill

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

McGurk effect

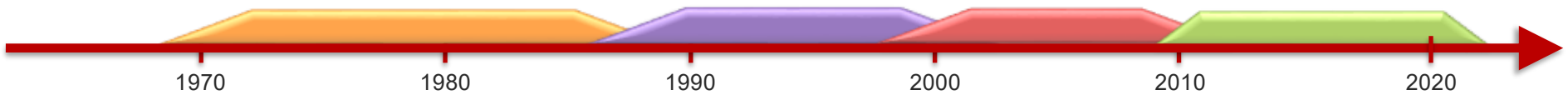
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Prior Research in 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 …)
  - The "foundation model" era (2020s until …)



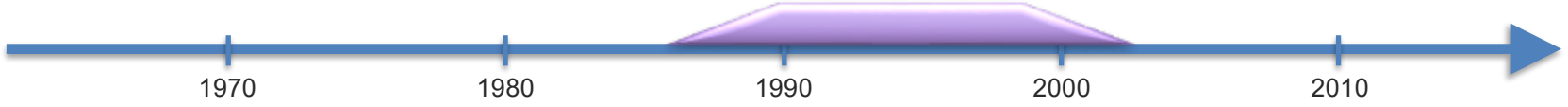
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

The "Computational" era (Late 1980s until 2000)

Multimodal interfaces (HCI)

Put that there by Architecture Machine Group, MIT [1982]



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

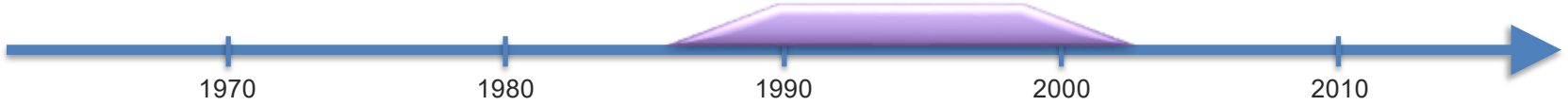
multisensory intelligence

The "Computational" era (Late 1980s until 2000)

Multimodal interfaces (HCI)

pFinder: Real-time Tracking of human body

by C. Wren, A. Azarbayejani, T. Darrell and A. Pentland [1995]

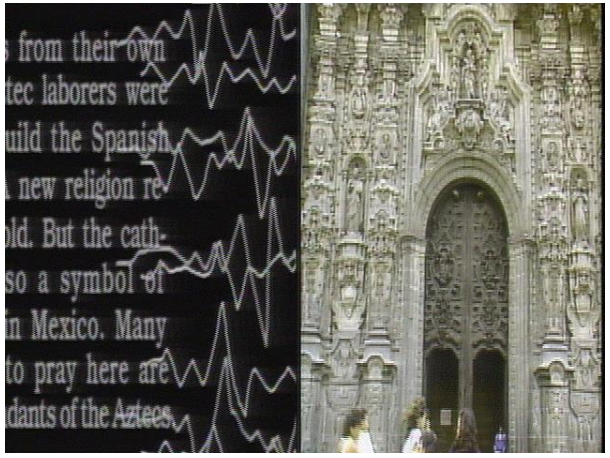


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

The "Computational" era (Late 1980s until 2000)

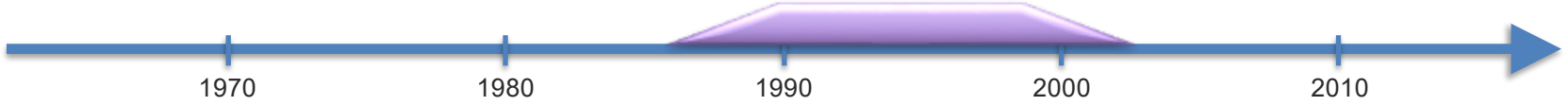
Multimedia Computing



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."



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

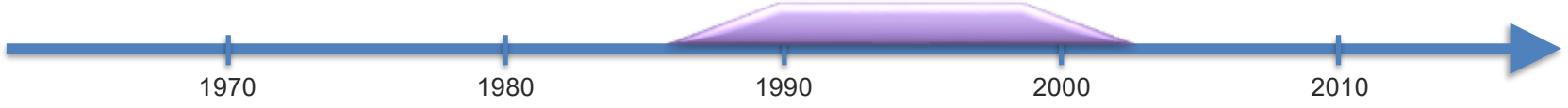
multisensory intelligence

The "Computational" era (Late 1980s until 2000)

Multimedia Computing

Multimedia content analysis

- **Shot-boundary detection (1991 -)**
  - Parsing a video into continuous camera shots
- **Still and dynamic video abstracts (1992 -)**
  - Making video browsable via representative frames (keyframes)
  - Generating short clips carrying the essence of the video content
- **High-level parsing (1997 -)**
  - Parsing a video into semantically meaningful segments
- **Automatic annotation (indexing) (1999 -)**
  - Detecting pre-specified events/scenes/objects in video



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

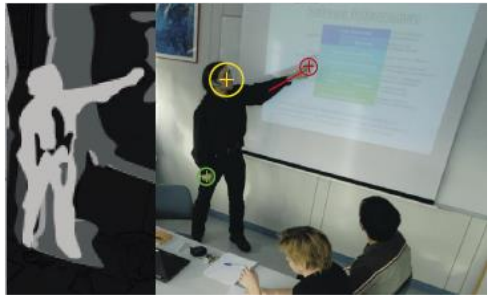
multisensory intelligence

The "Interaction" era (2000s)

Modeling Human Multimodal Interaction

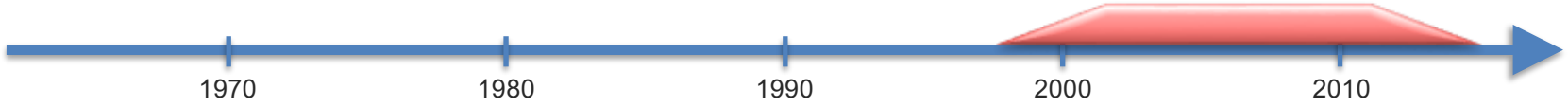
AMI Project [2001-2006, IDIAP]

- 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



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

The "Interaction" era (2000s)

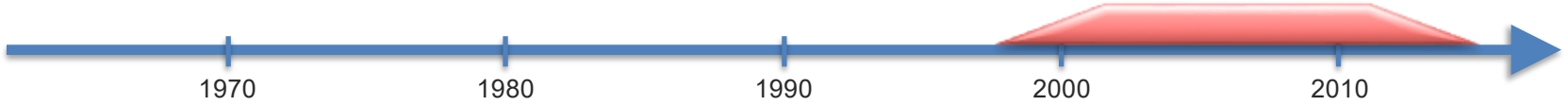
Modeling Human Multimodal Interaction

CALO Project [2003-2008, SRI]

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

SSP Project [2008-2011, IDIAP]

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



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

The "Interaction" era (2000s)

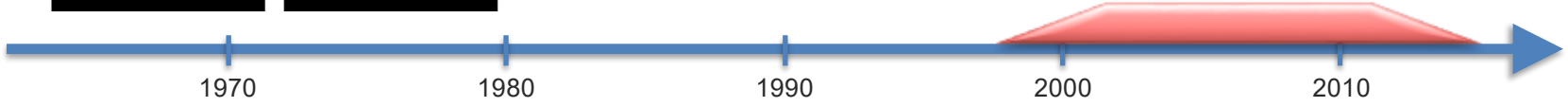
Many new challenges and multimodal corpora !!

Audio-Visual Emotion Challenge (AVEC, 2011-)

- Emotional dimension estimation
- Standardized training and test sets
- Based on the SEMAINE dataset

Emotion Recognition in the Wild Challenge (EmotiW 2013-)

- Standardized training and test sets
- Challenging in-the-wild test sets



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

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]

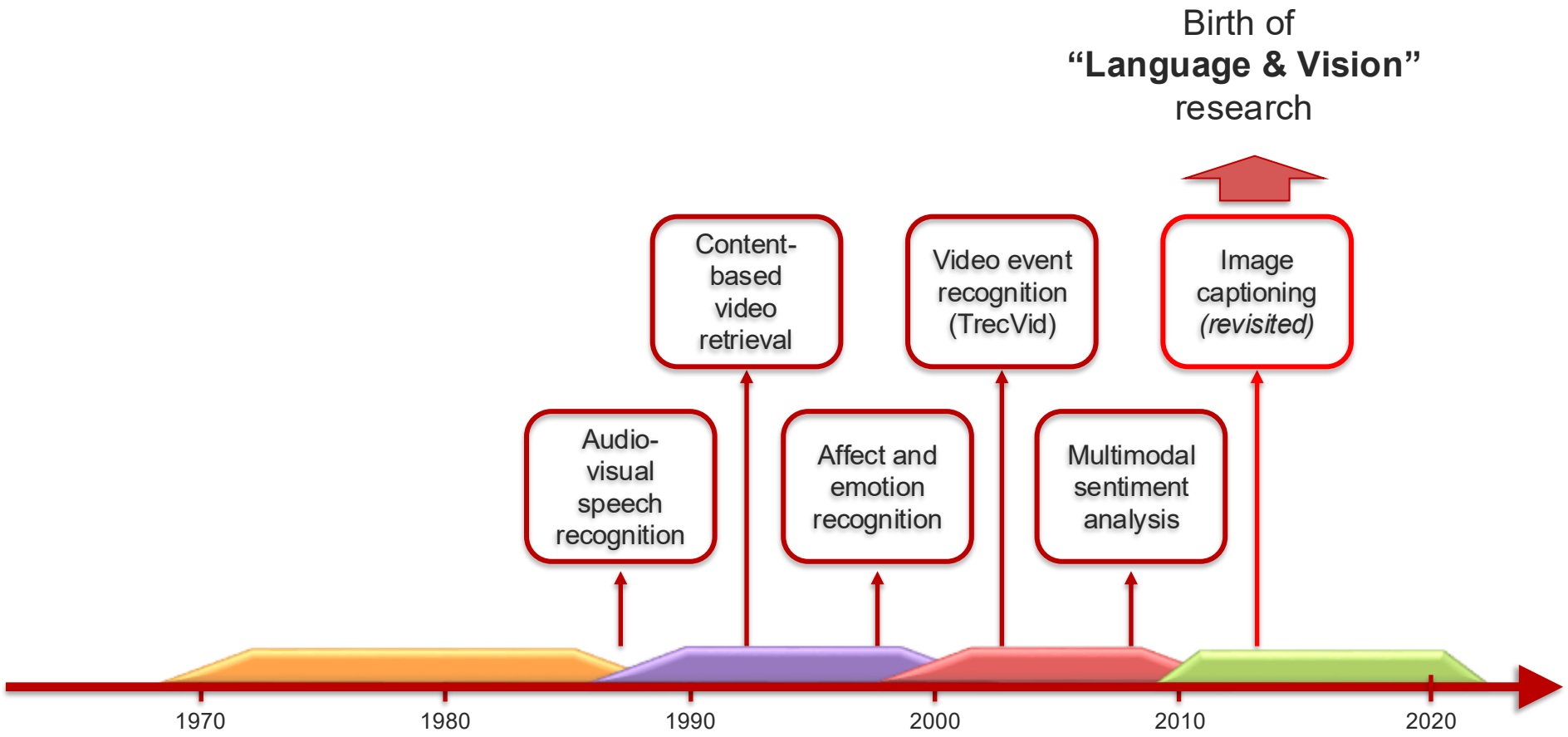
Key enablers for multimodal deep learning research:

- New large-scale multimodal datasets
- Faster compute and GPUs
- High-level visual features
- “Dimensional” linguistic features

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

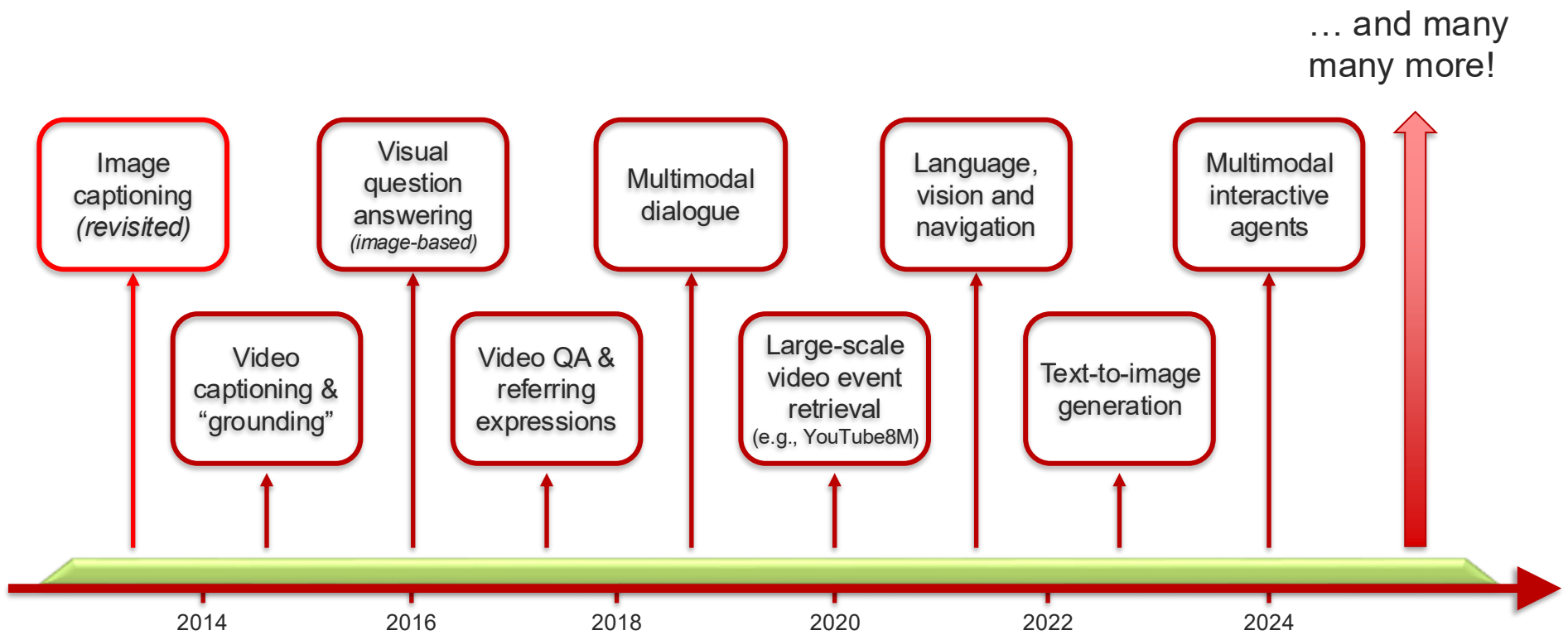
multisensory intelligence

Multimodal Research Tasks



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Multimodal Research Tasks

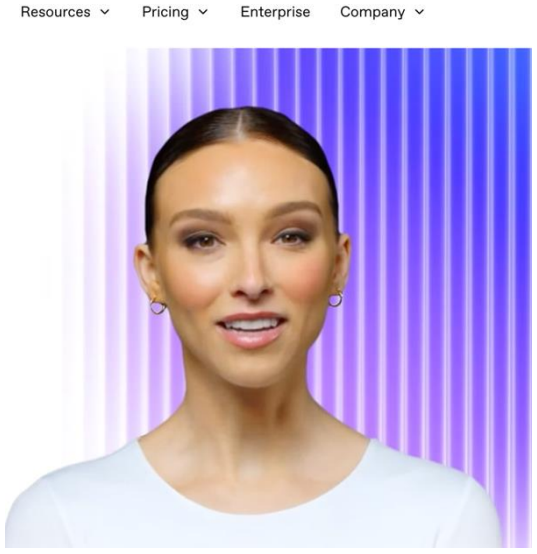


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

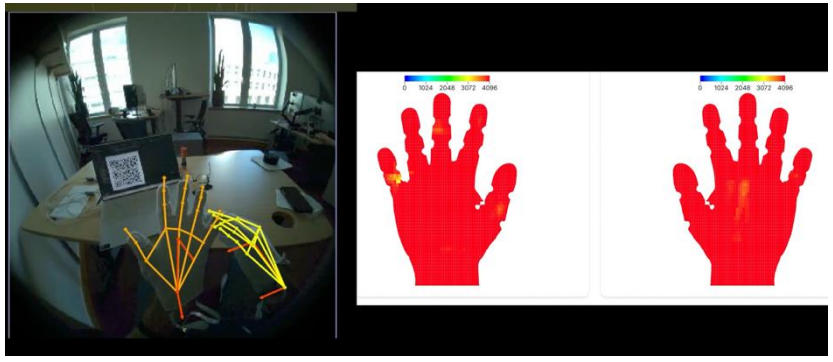
In 2025:

Video generation and world models

Virtual agents for health and wellbeing



Intelligent tactile + olfactory interfaces



Multimodal AI – Surveys, Tutorials, Courses

Foundations and Recent Trends in Multimodal Machine Learning

Paul Liang, Amir Zadeh and Louis-Philippe Morency

☑ 6 core challenges
☑ 50+ taxonomic classes
☑ 700+ referenced papers

https://arxiv.org/abs/2209.03430

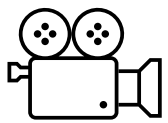
Tutorials: ICML 2023, CVPR 2022, NAACL 2022

multisensory intelligence

What is a Modality?

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:

Raw Modalities		Abstract Modalities
Speech signal	Language	Sentiment intensity
Image	Detected objects	Object categories

multisensory intelligence

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

A research-oriented definition…

***Multimodal* is the science of**

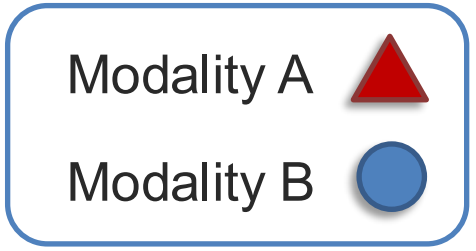
heterogeneous and interconnected data

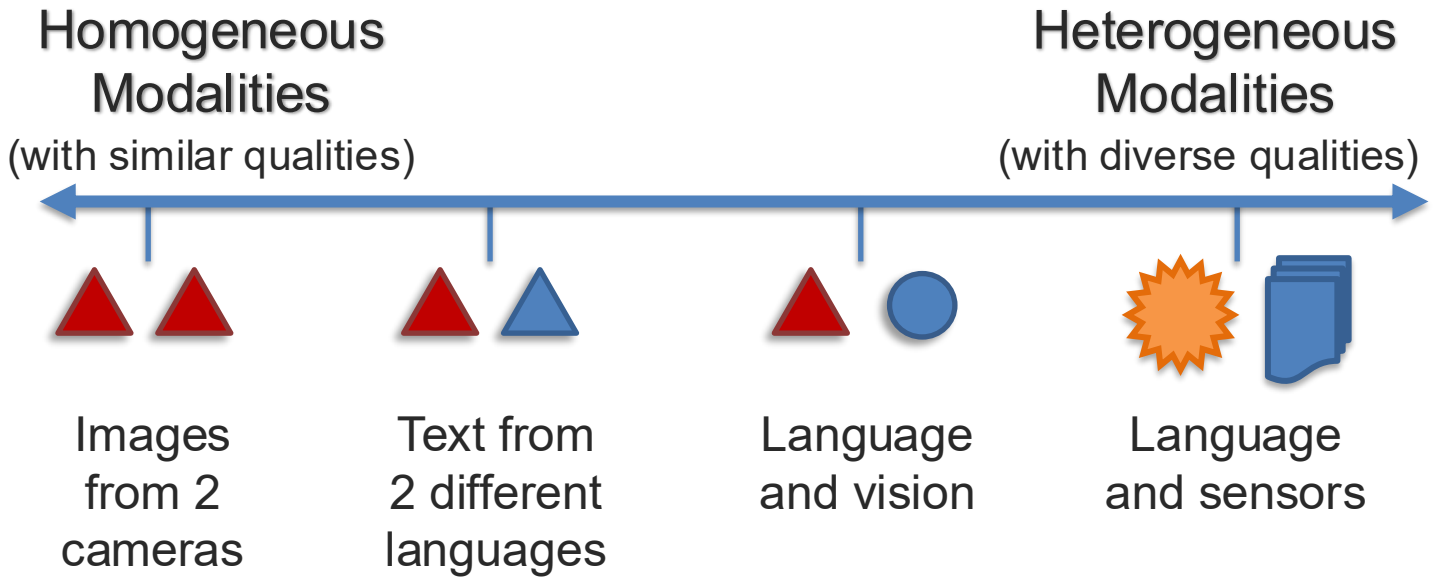
Connected + Interacting

multisensory intelligence

Heterogeneous Modalities

Information in different modalities shows diverse qualities, structures, & representations.



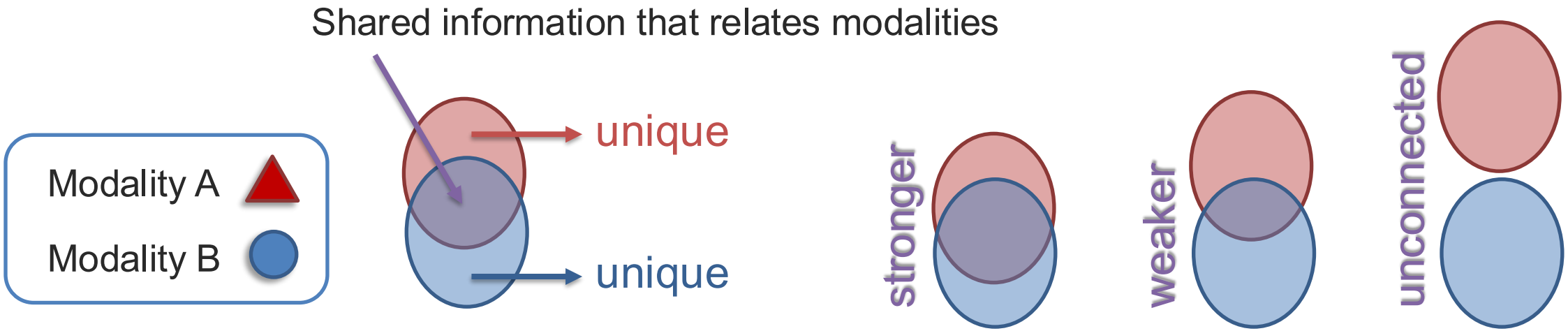


Examples:

Abstract modalities are more likely to be homogeneous

multisensory intelligence

Connected Modalities

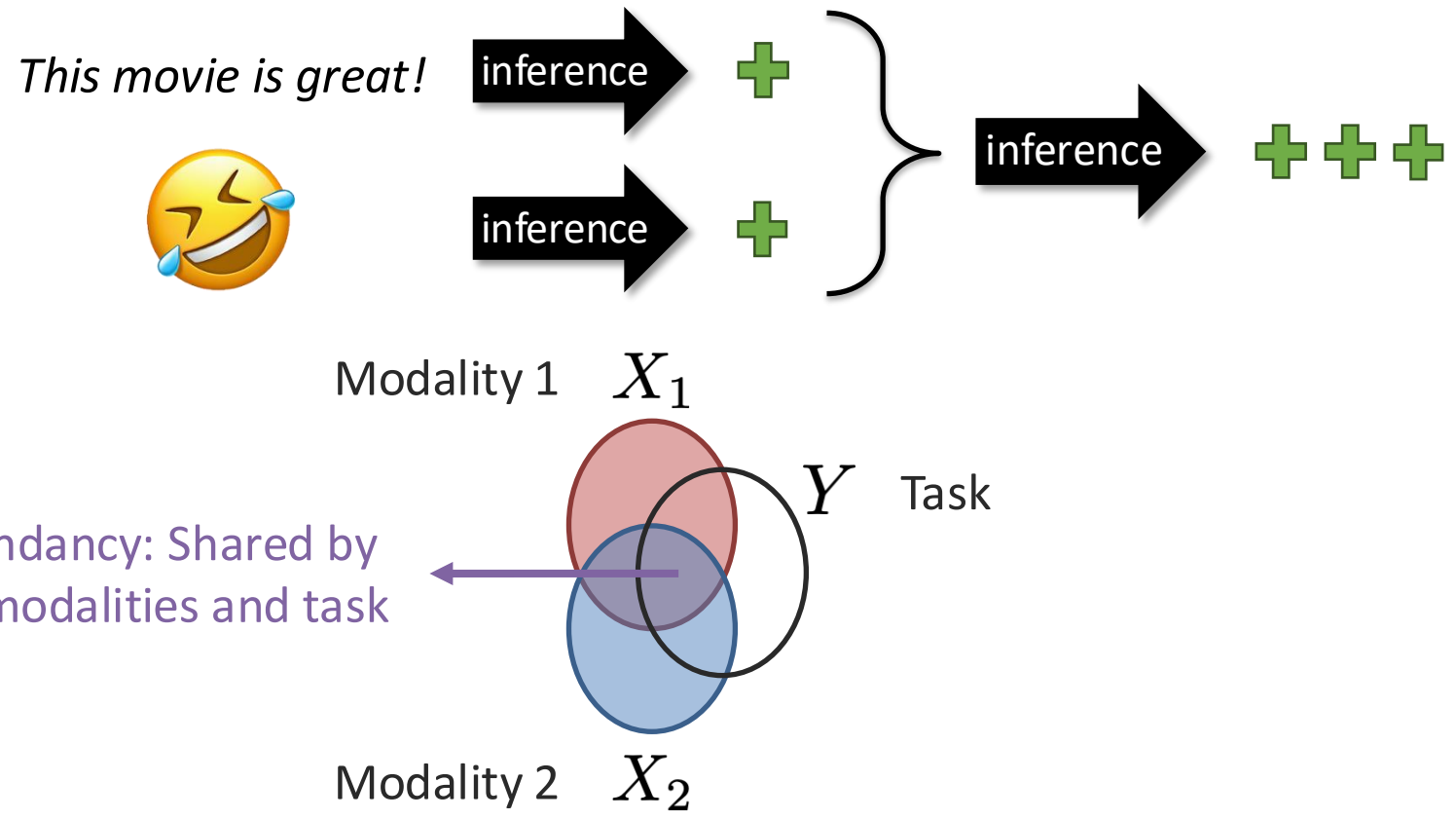


A teacup on the right of a laptop in a clean room.

multisensory intelligence

Interacting Modalities

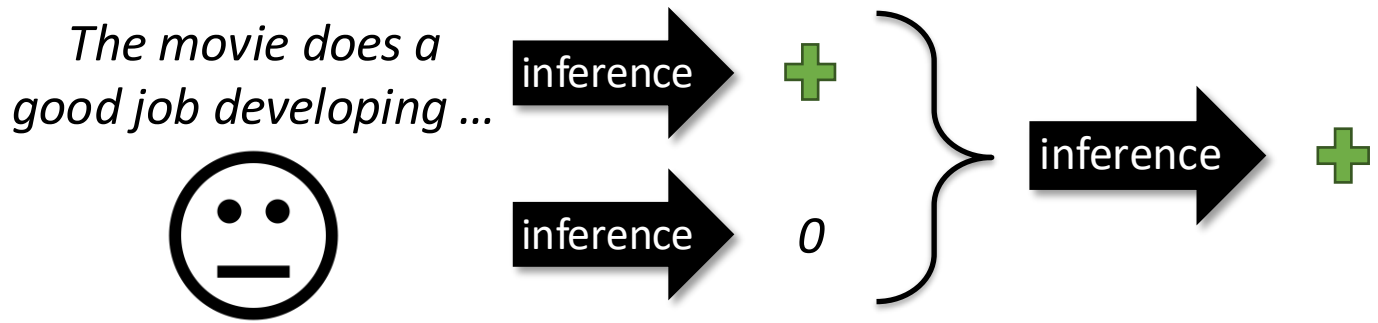
Interactions: How modalities *combine* to provide information for a task.

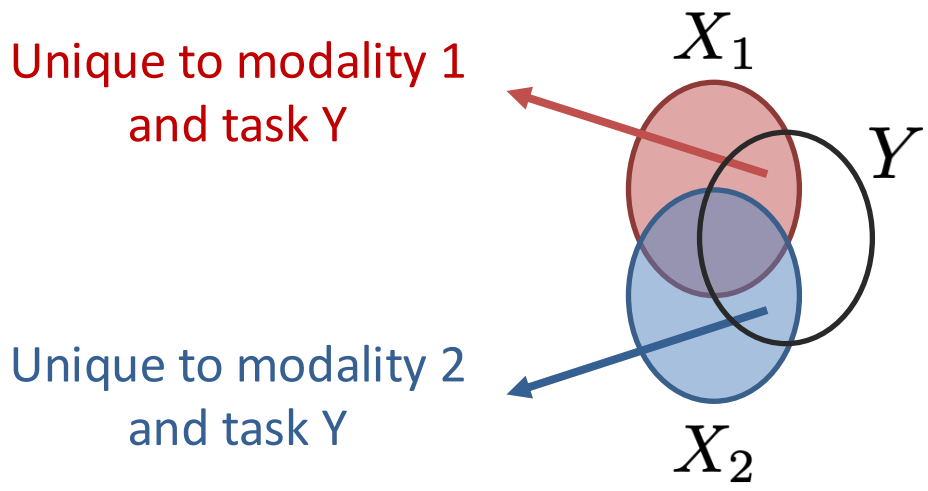


multisensory intelligence

Interacting Modalities

Interactions: How modalities *combine* to provide information for a task.

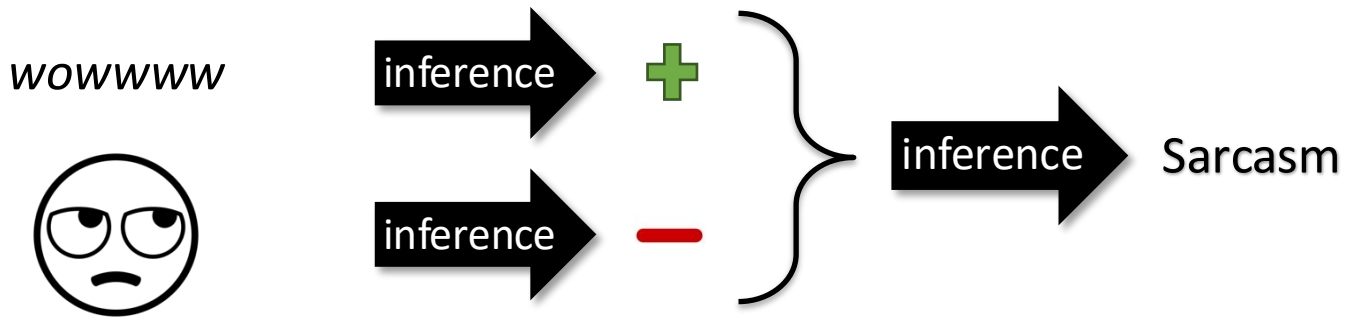


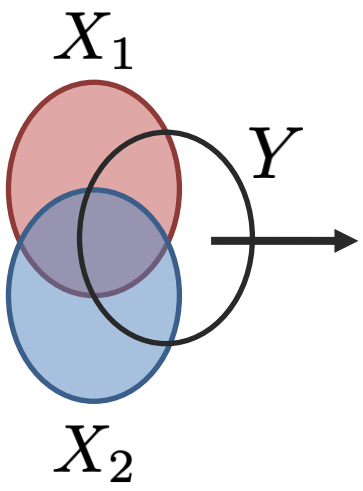


multisensory intelligence

Interacting Modalities

Interactions: How modalities *combine* to provide information for a task.





Synergy: Emerging information from multimodal interaction

What is Multimodal? → *Why is it hard?* → *What is next?*

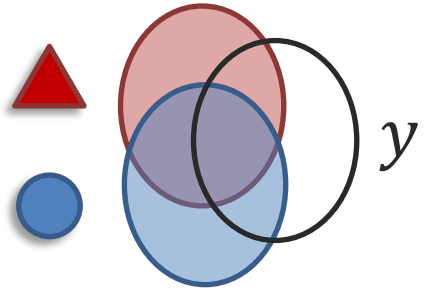
Heterogeneous

+

Connected

+

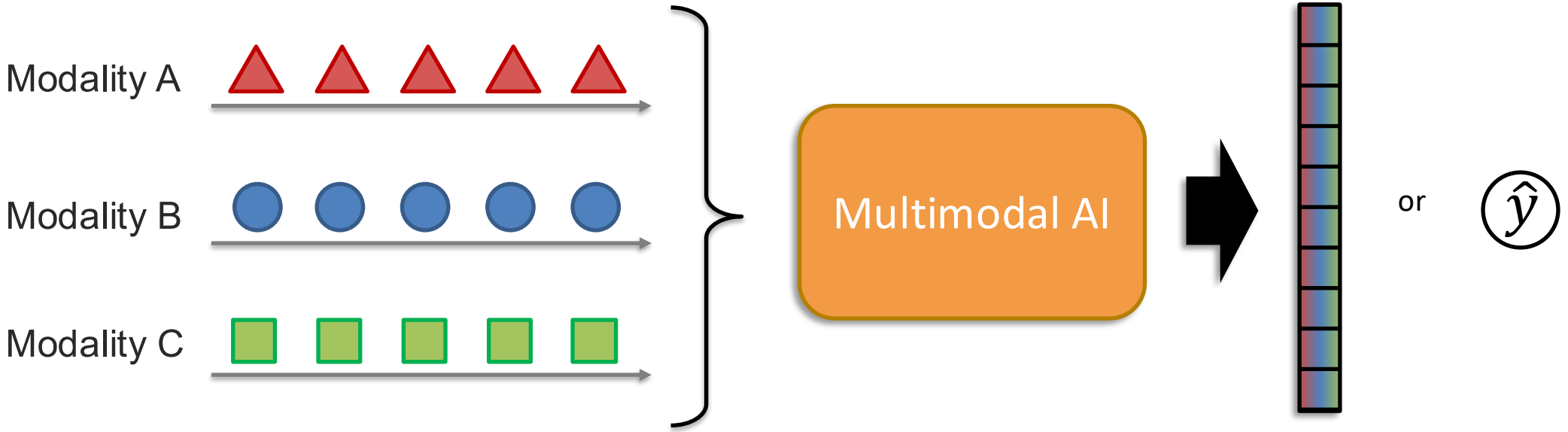
Interacting



Multimodal is the scientific study of heterogeneous and interconnected data ☺

multisensory intelligence

Multimodal AI Challenges

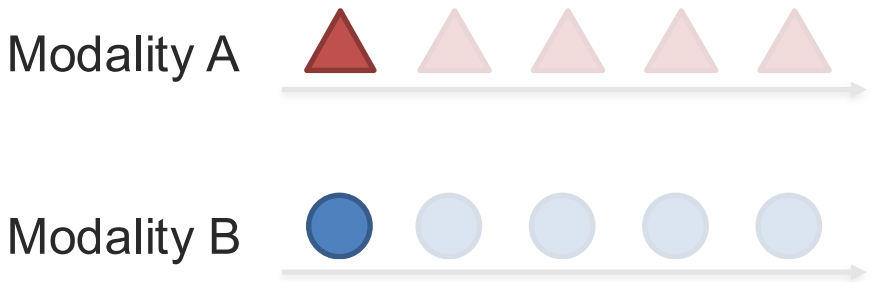


Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

➡ This is a core building block for most multimodal modeling problems!

Individual elements:

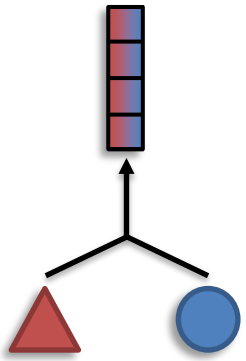


Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities.

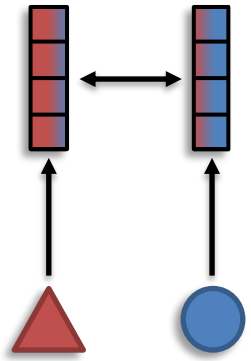
Sub-challenges:

Fusion



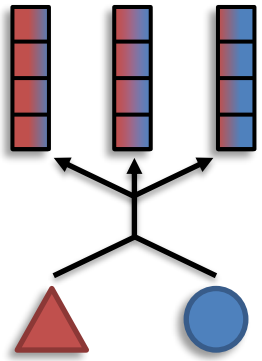
modalities > # representations

Coordination



modalities = # representations

Fission



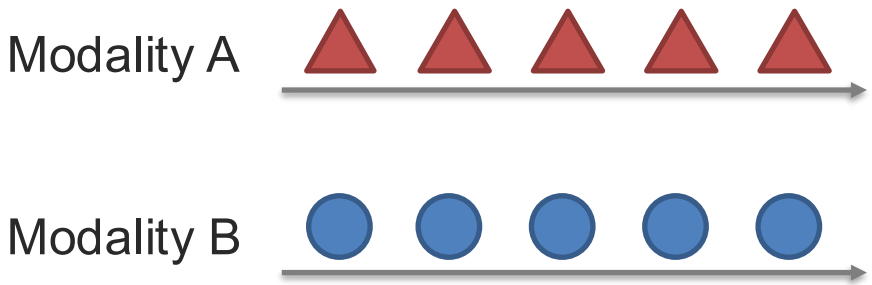
modalities < # representations

Challenge 2: Alignment

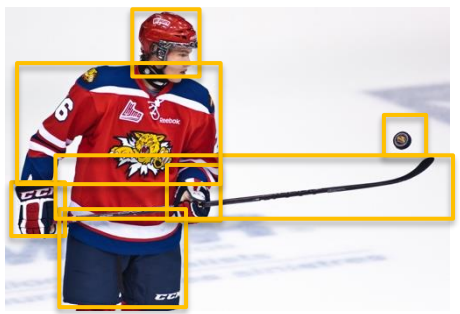
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure.

➡ Most modalities have internal structure with multiple elements

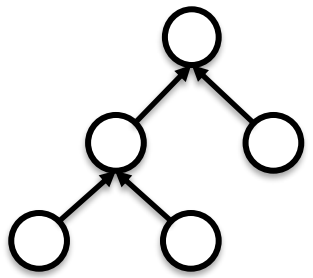
Elements with temporal structure:



Other structured examples:



Spatial



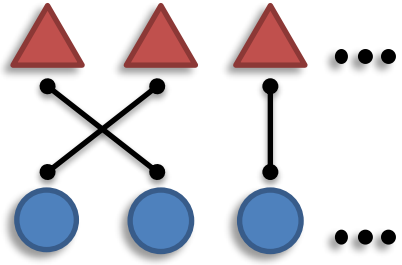
Hierarchical

Challenge 2: Alignment

Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure.

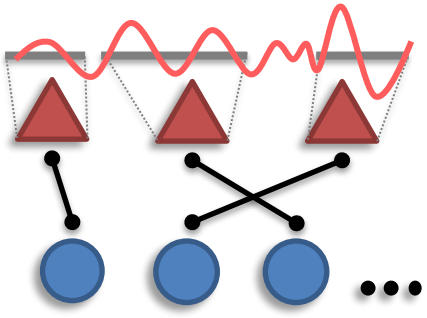
Sub-challenges:

Discrete connections



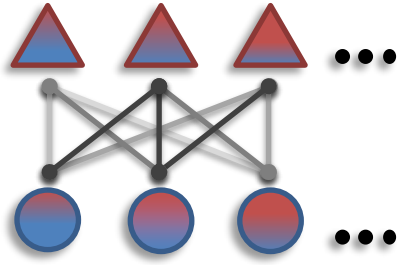
Explicit alignment
(e.g., grounding)

Continuous alignment



Granularity of individual elements

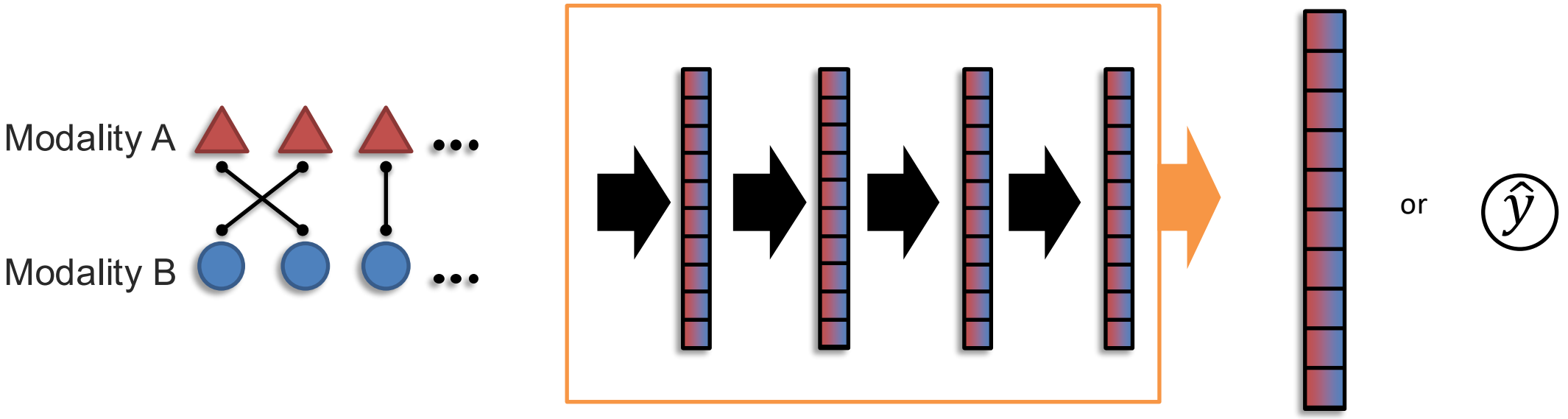
Contextualized representation



Implicit alignment + representation

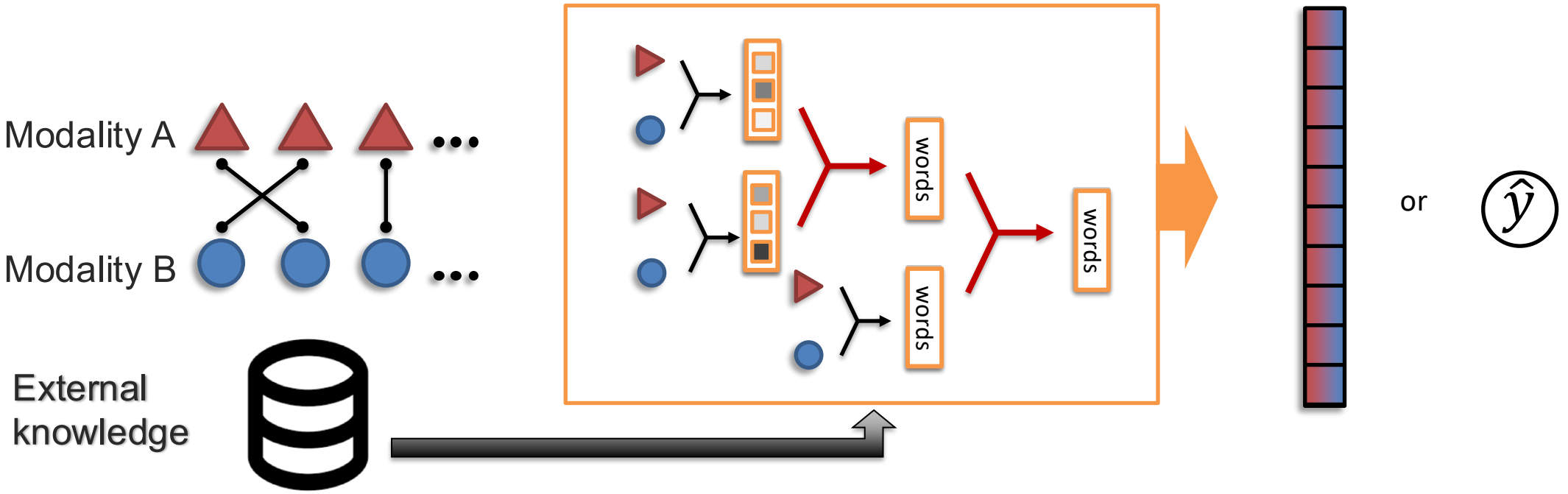
Challenge 3: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



Challenge 3: Reasoning

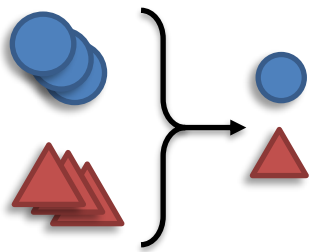
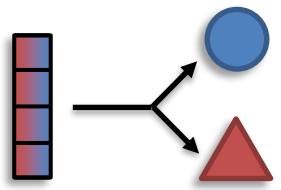
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



Challenge 4: Generation

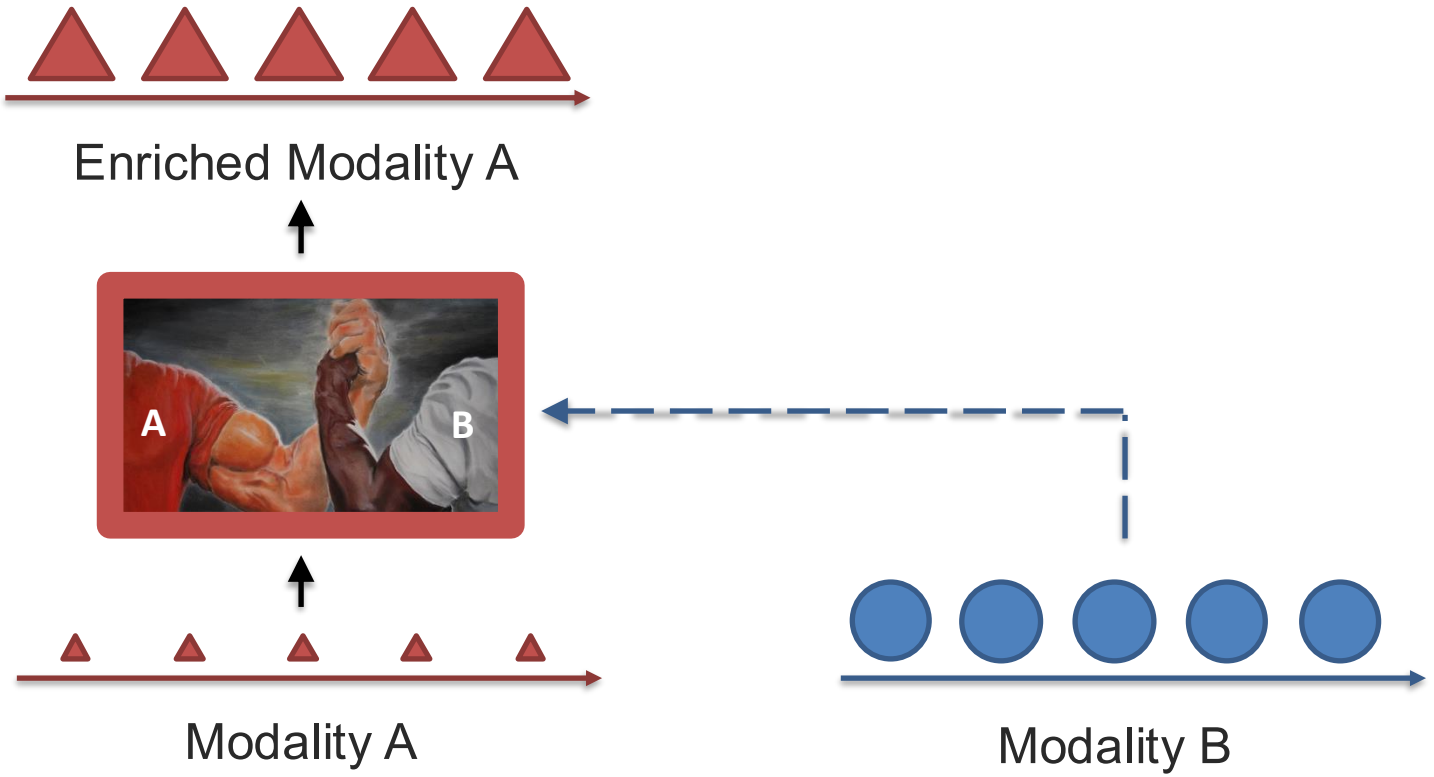
Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.

Sub-challenges:

	Summarization	Translation	Creation
			
Information: (content)	Reduction	Maintenance	Expansion
			

Challenge 5: Transference

Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources.

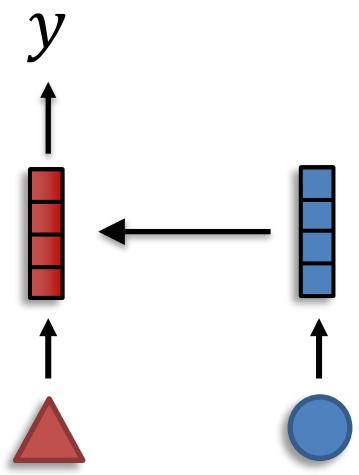


Challenge 5: Transference

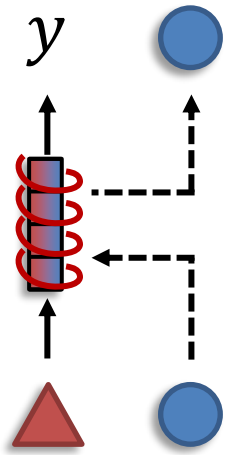
Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources.

Sub-challenges:

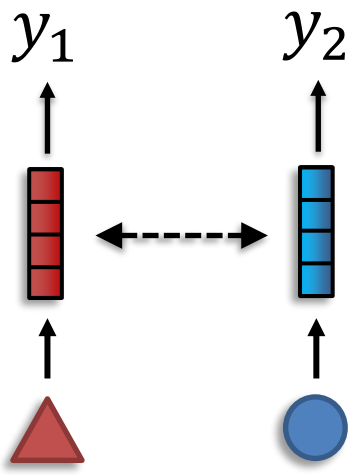
Transfer



Co-learning



Model Induction

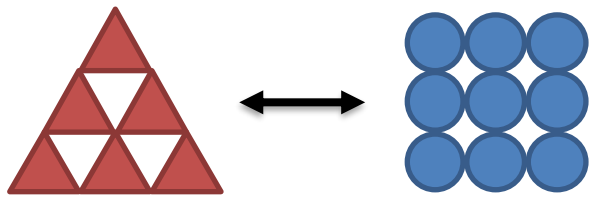


Challenge 6: Quantification

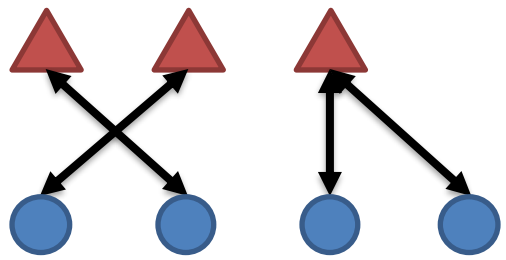
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions, and the multimodal learning process.

Sub-challenges:

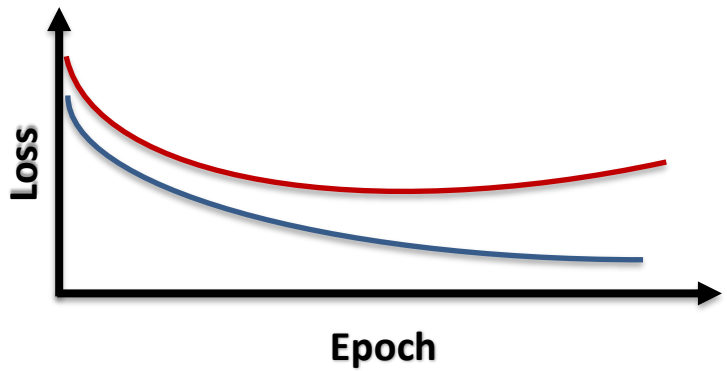
Heterogeneity



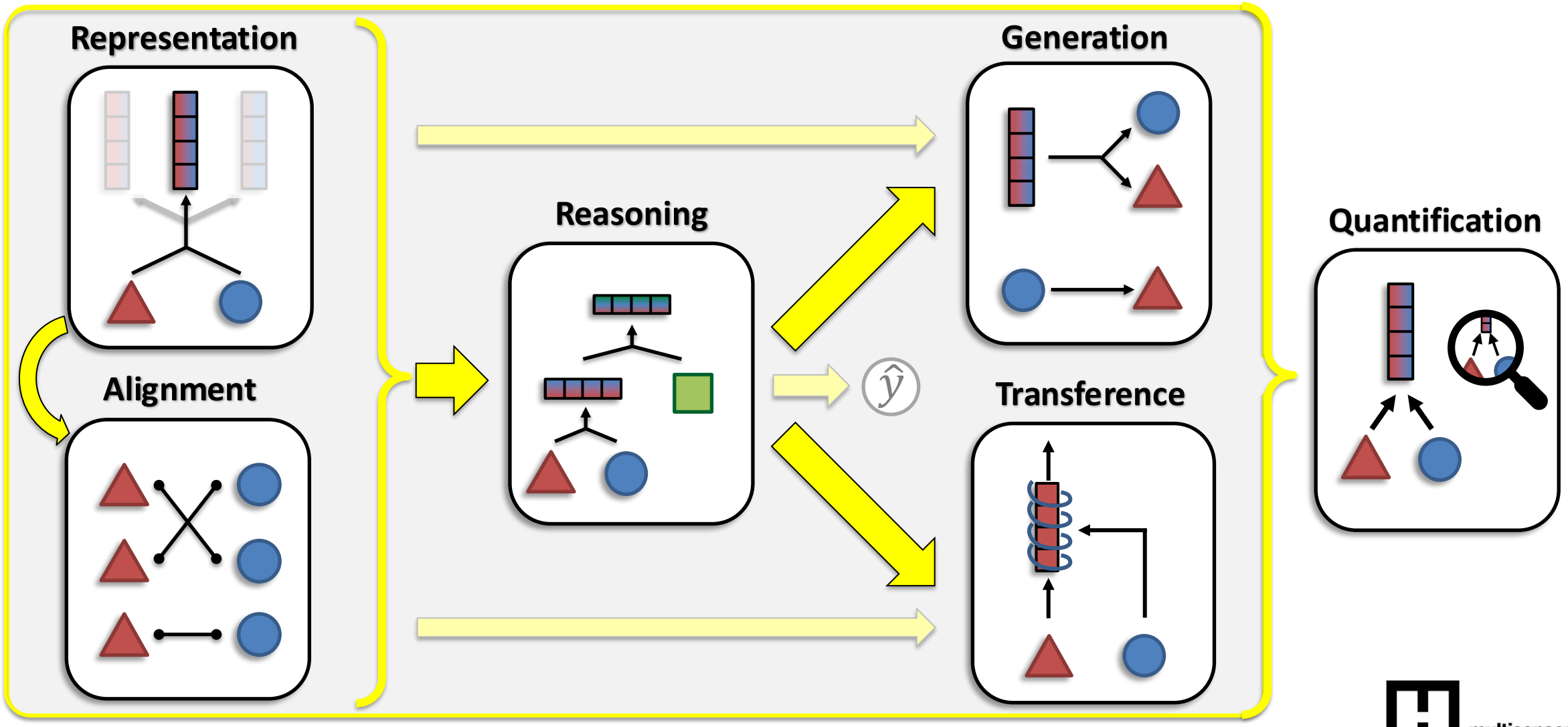
Interactions



Learning



Summary of Core Multimodal Challenges



What is Multimodal? → *Why is it hard?* → *What is next?*

Heterogeneous

+

Connected

+

Interacting

Representation

Alignment

Reasoning

Generation

Transference

Quantification

Multimodal LLMs

Generative AI

Physical sensing

Holistic health

Agents

Assignments for This Coming Week

This Thursday: lecture on **multimodal datasets and tasks.**

Mingle with classmates and try to form teams. Start thinking about what project you want to work on!

We will release a more formal project preference form this week to help with team matching.