How to AI (Almost) Anything

Lecture 4 – Multimodal AI & Alignment

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

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

Assignments for This Coming Week

For project:

- I gave feedback and assigned primary TA.
- Meet with me and primary TA every other week.

Reading assignment due tomorrow Wednesday (3/5).

This Thursday (3/6): second reading discussion on **modern AI architectures**.

Scaling laws for multimodal models

Not all tokens are all you need?

Lecture Topics

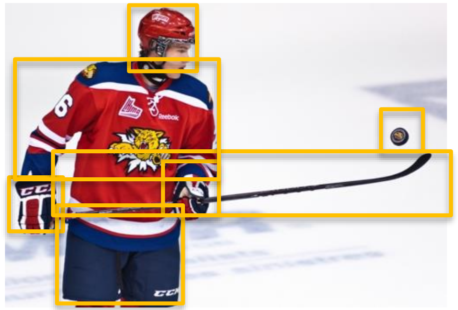
(subject to change, based on student interests and course discussions)

Module 1: Foundations of AI

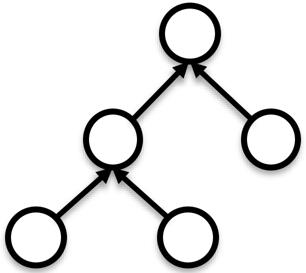
Week 1 (2/4): Introduction to AI and AI research

Week 2 (2/11): Data, structure, and information

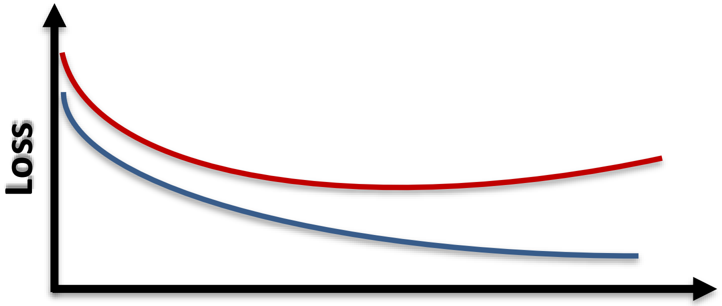
Week 4 (2/25): Common model architectures



Spatial



Hierarchical



multisensory intelligence

Lecture Topics

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

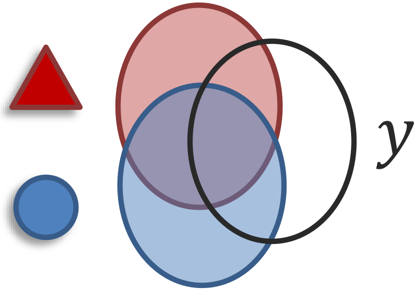
Module 2: Foundations of multimodal AI

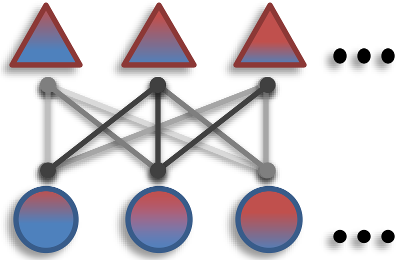
Week 5 (3/4): Multimodal connections and alignment

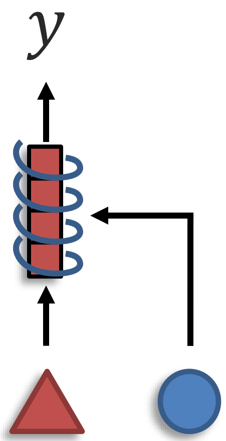
Week 6 (3/11): Multimodal interactions and fusion

Week 7 (3/18): Cross-modal transfer

Week 8 – No class, spring break







multisensory intelligence

Today’s lecture

1. Introduction to multimodal AI
2. Principles of heterogeneity, connections, interactions
3. Core multimodal challenges
4. Multimodal alignment

[credit: some slides in this lecture were 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

Behavioral Study of Multimodal

Language
and gestures

David McNeill

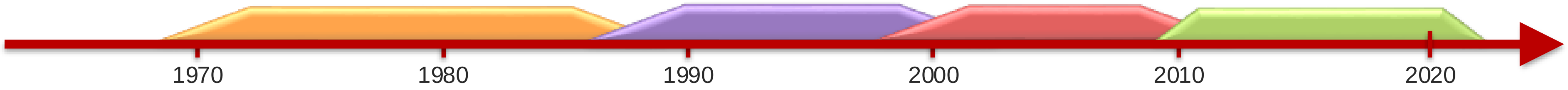
"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

McGurk effect

Prior Research in Multimodal

Four eras of multimodal research

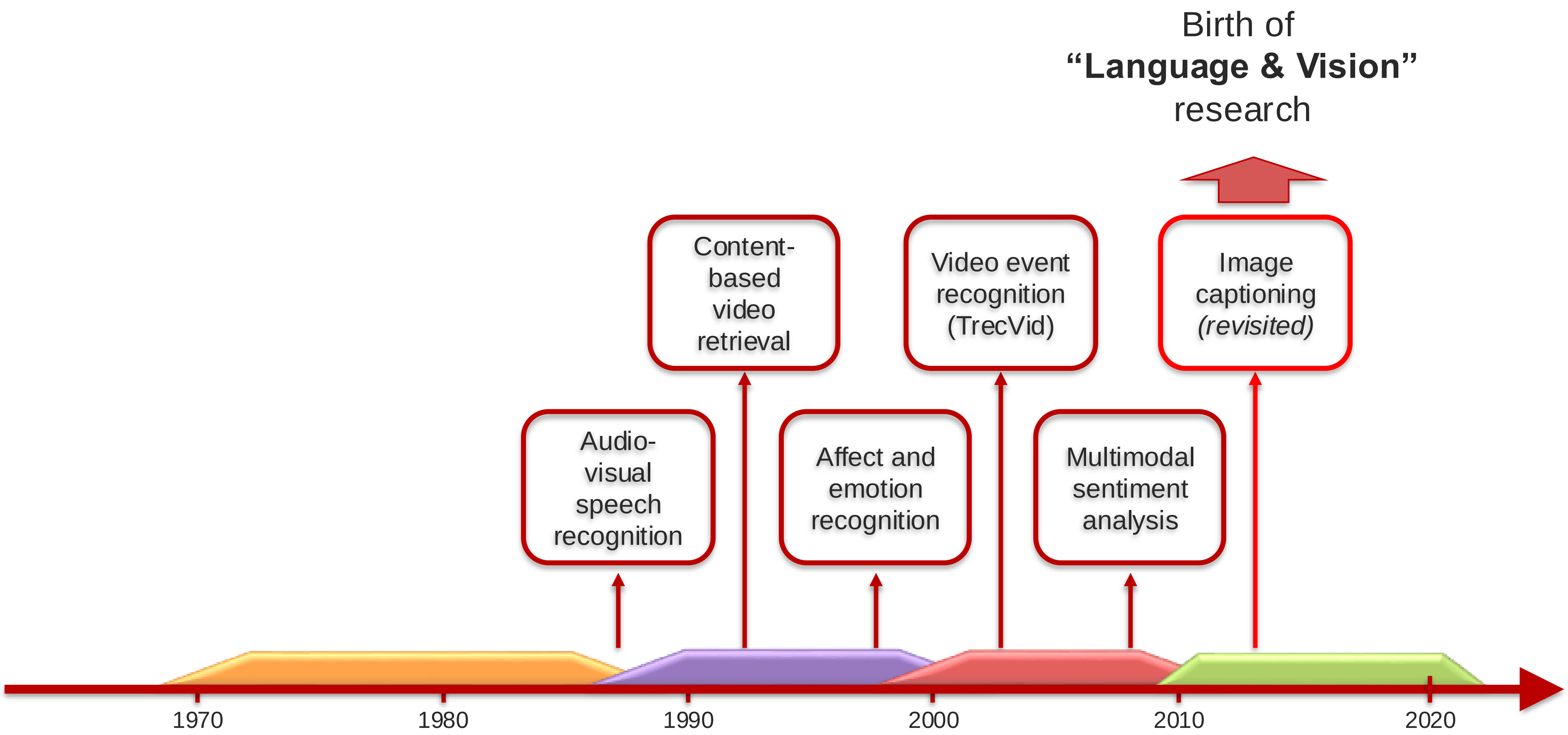
- The “behavioral” era (1970s until late 1980s)
- The “computational” era (late 1980s until 2000)
- The “interaction” era (2000 - 2010)
- The “deep learning” era (2010s until …)
  - The “foundation model” era (2020s until …)



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

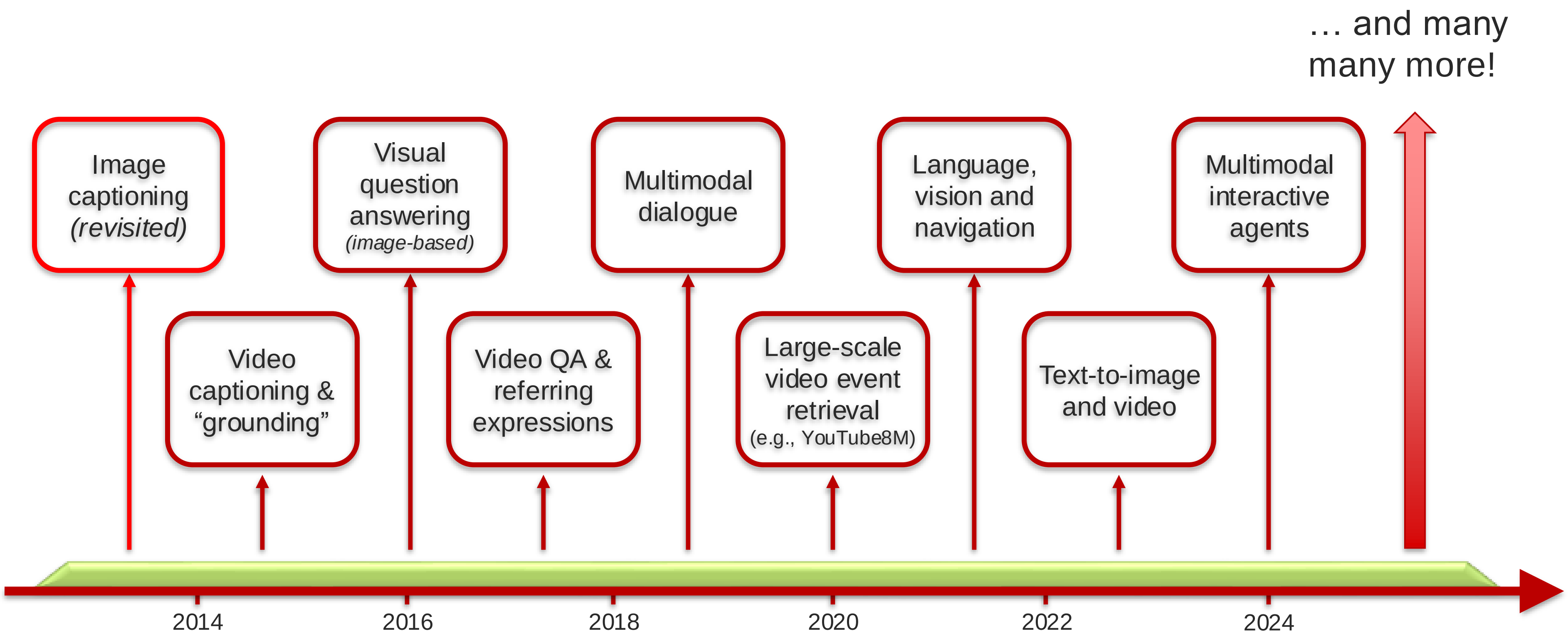
multisensory intelligence

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]

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

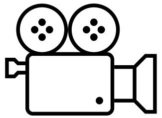
Tutorials: ICML 2023, CVPR 2022, NAACL 2022

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:

Speech signal

Image

Language

Detected objects

Sentiment intensity

Object categories

multisensory intelligence

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

A research-oriented definition…

***Multimodal* is the science of**

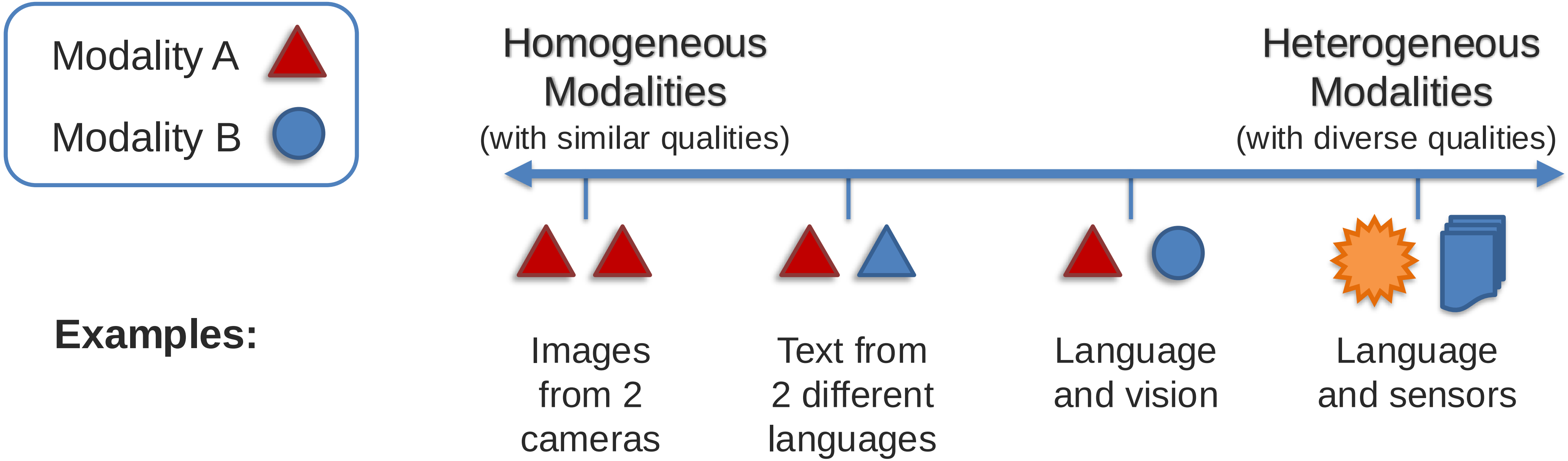
heterogeneous and interconnected data

Connected + Interacting

multisensory intelligence

Heterogeneous Modalities

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



Abstract modalities are more likely to be homogeneous

multisensory intelligence

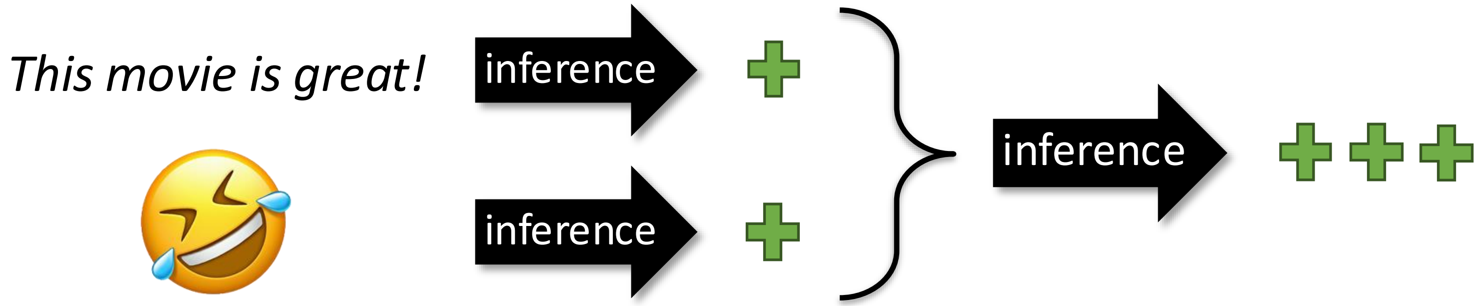
Connected Modalities

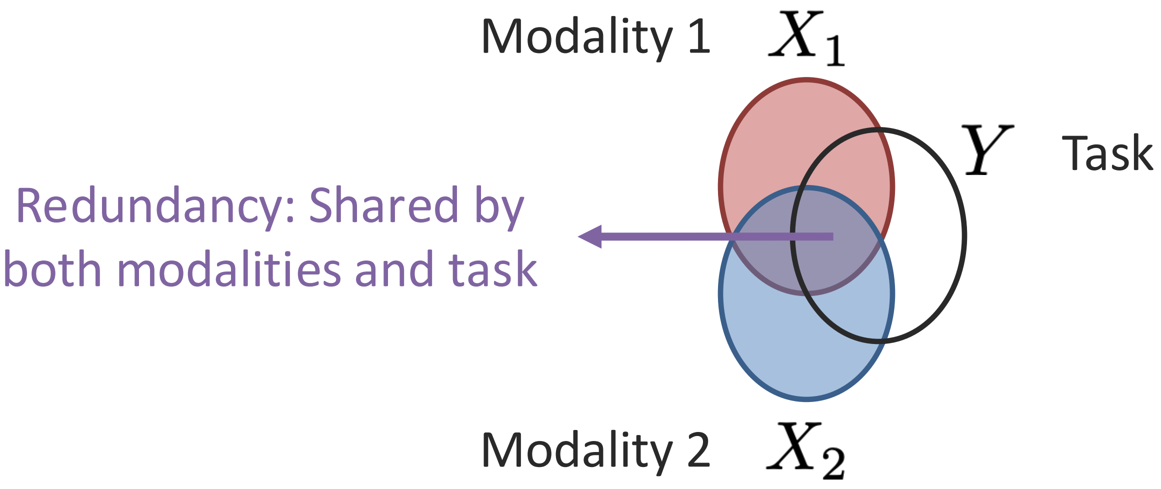
Shared information that relates modalities

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

Interacting Modalities

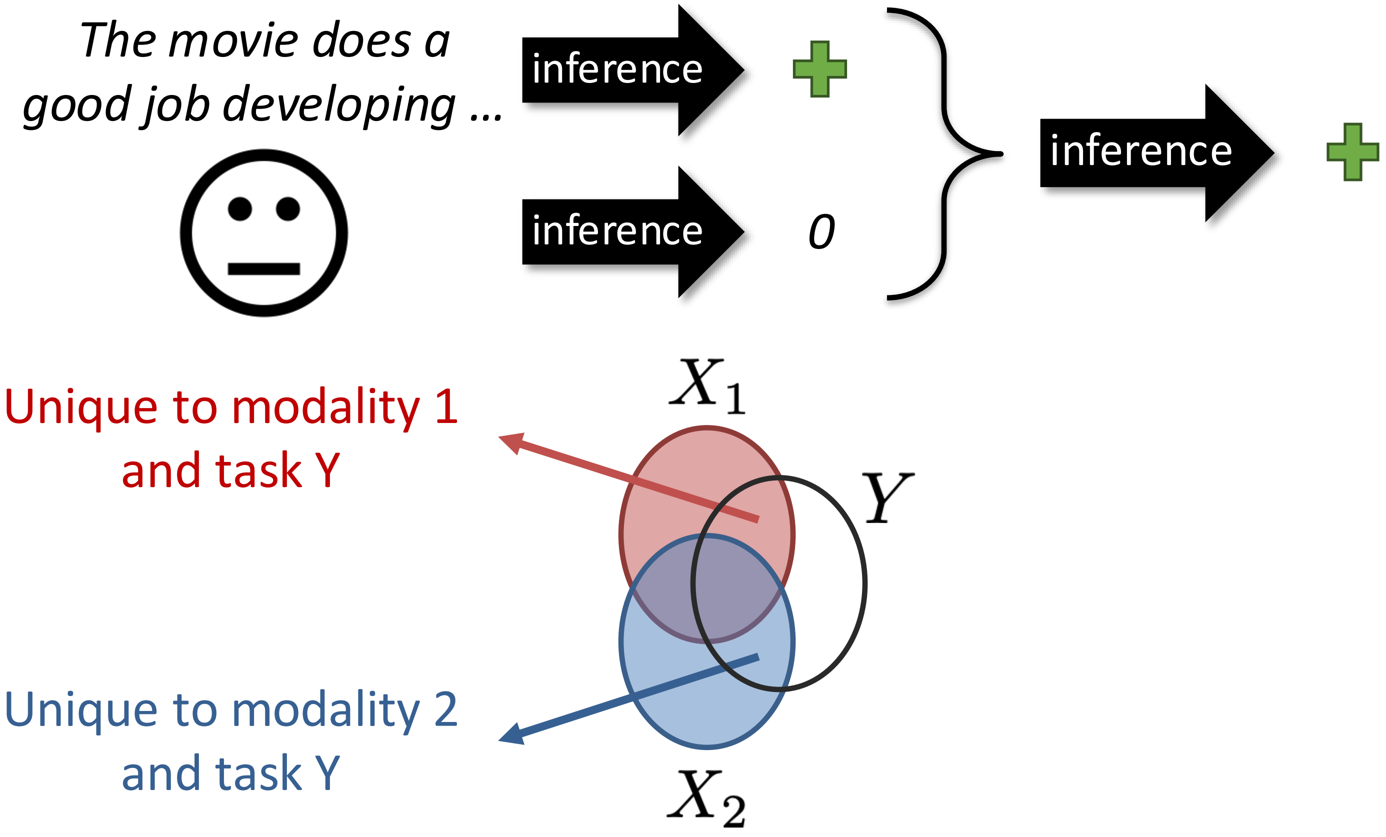
Interactions: How modalities *combine* to provide information for a task.





Interacting Modalities

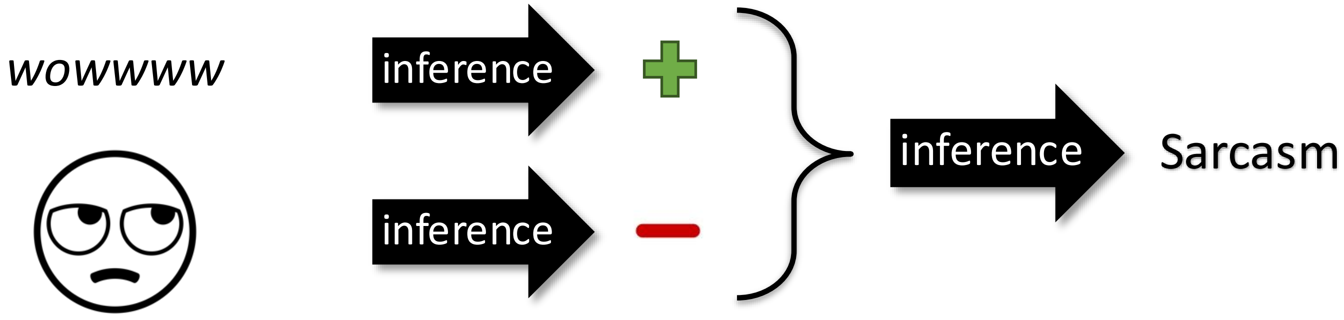
Interactions: How modalities *combine* to provide information for a task.

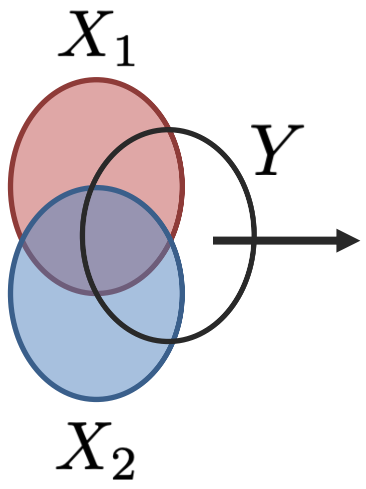


multisensory intelligence

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?*

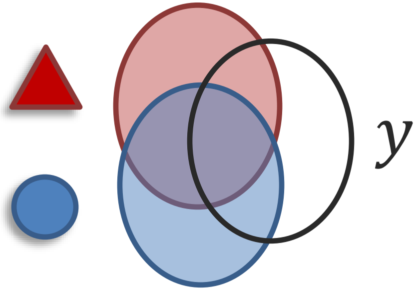
Heterogeneous

➕

Connected

➕

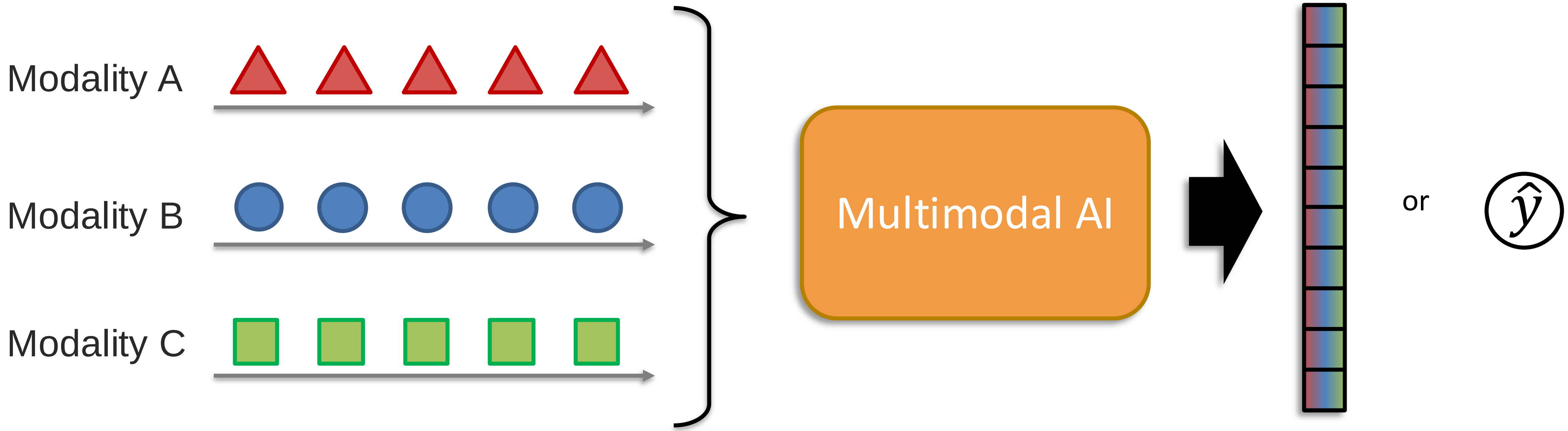
Interacting



Multimodal is the scientific study of heterogeneous and interconnected data ☺

multisensory intelligence

Multimodal AI Challenges



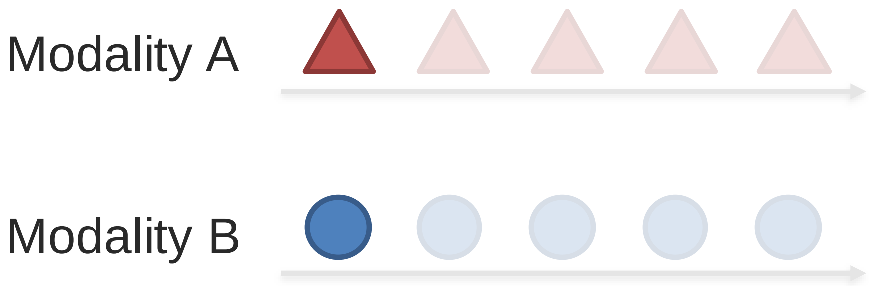
multisensory intelligence

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:

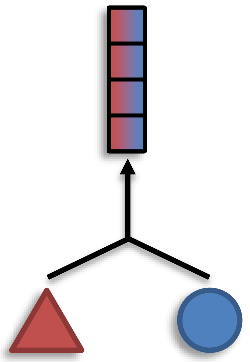


Challenge 1: Representation

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

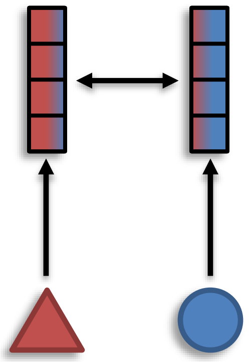
Sub-challenges:

Fusion



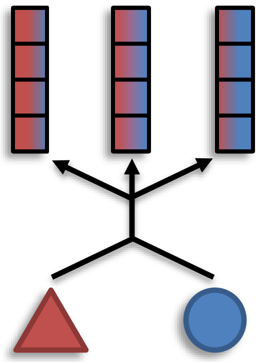
modalities > # representations

Coordination



modalities = # representations

Fission



modalities < # representations

Challenge 2: Alignment

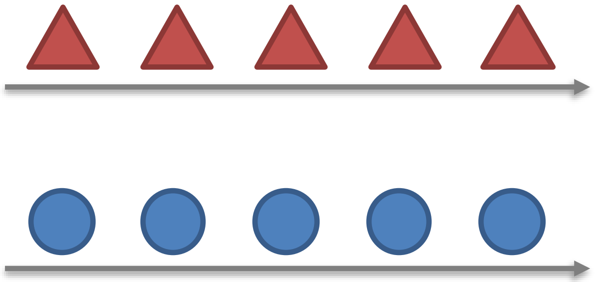
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure.

Most modalities have internal structure with multiple elements

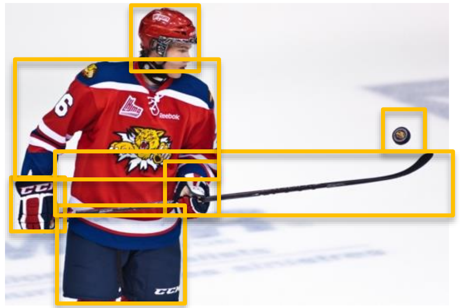
Elements with temporal structure:

Modality A

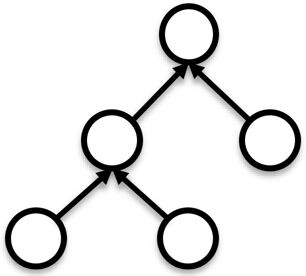
Modality B



Other structured examples:



Spatial



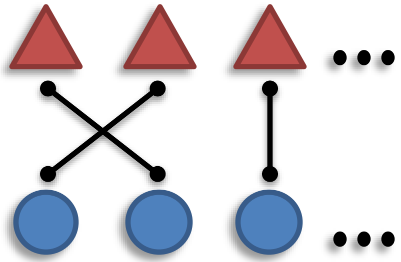
Hierarchical

Challenge 2: Alignment

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

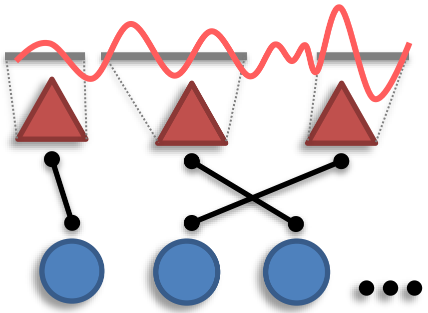
Sub-challenges:

Discrete connections



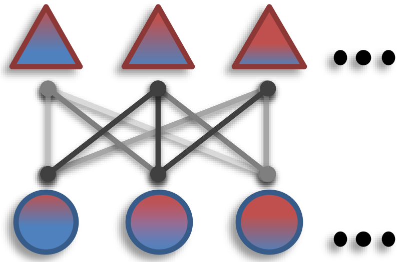
Explicit alignment
(e.g., grounding)

Continuous alignment



Granularity of
individual elements

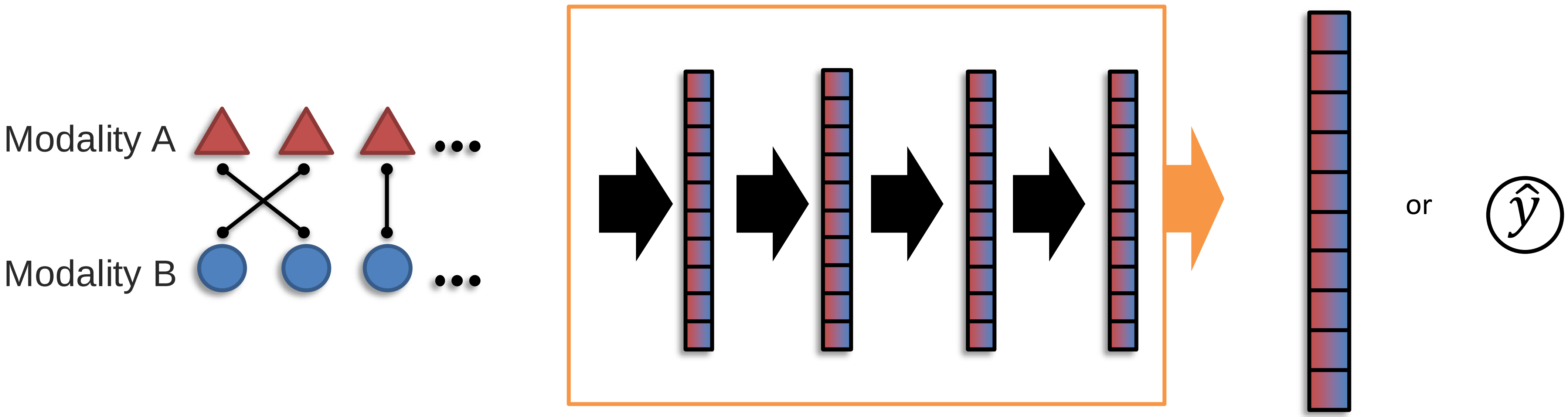
Contextualized representation



Implicit alignment
+ representation

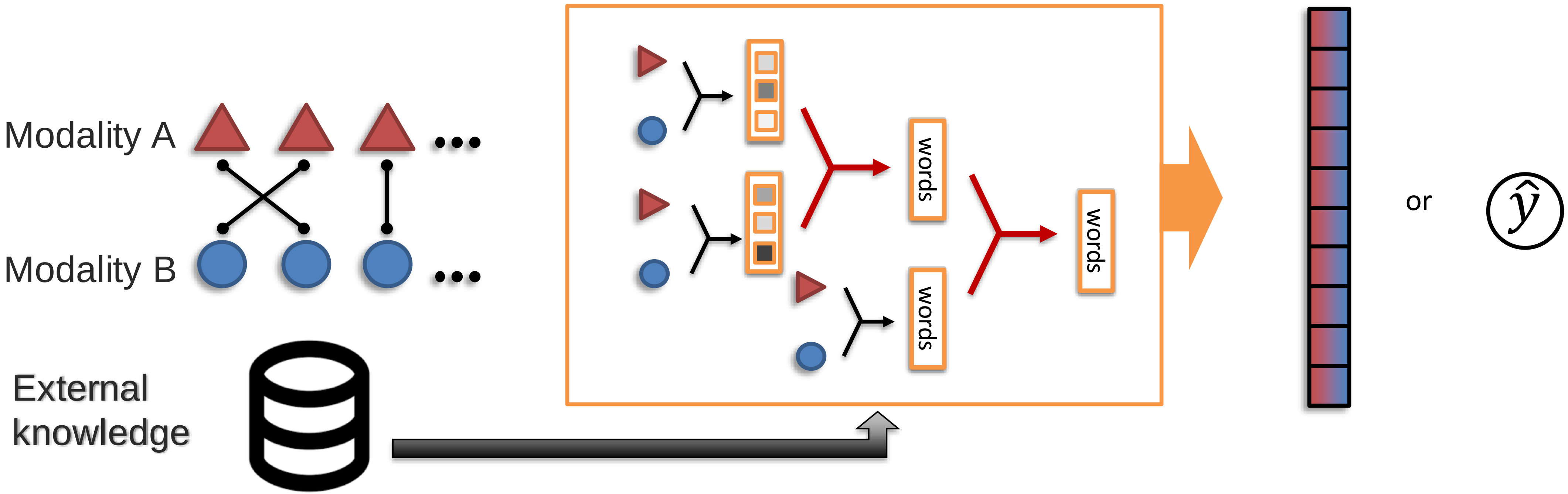
Challenge 3: Reasoning

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.

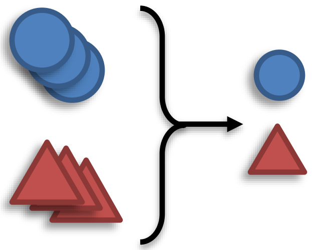


Challenge 4: Generation

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

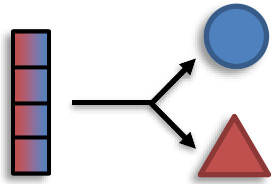
Sub-challenges:

Summarization



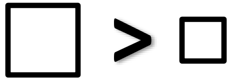
Translation

Creation

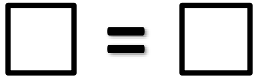


Information: (content)

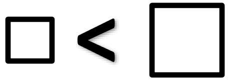
Reduction



Maintenance

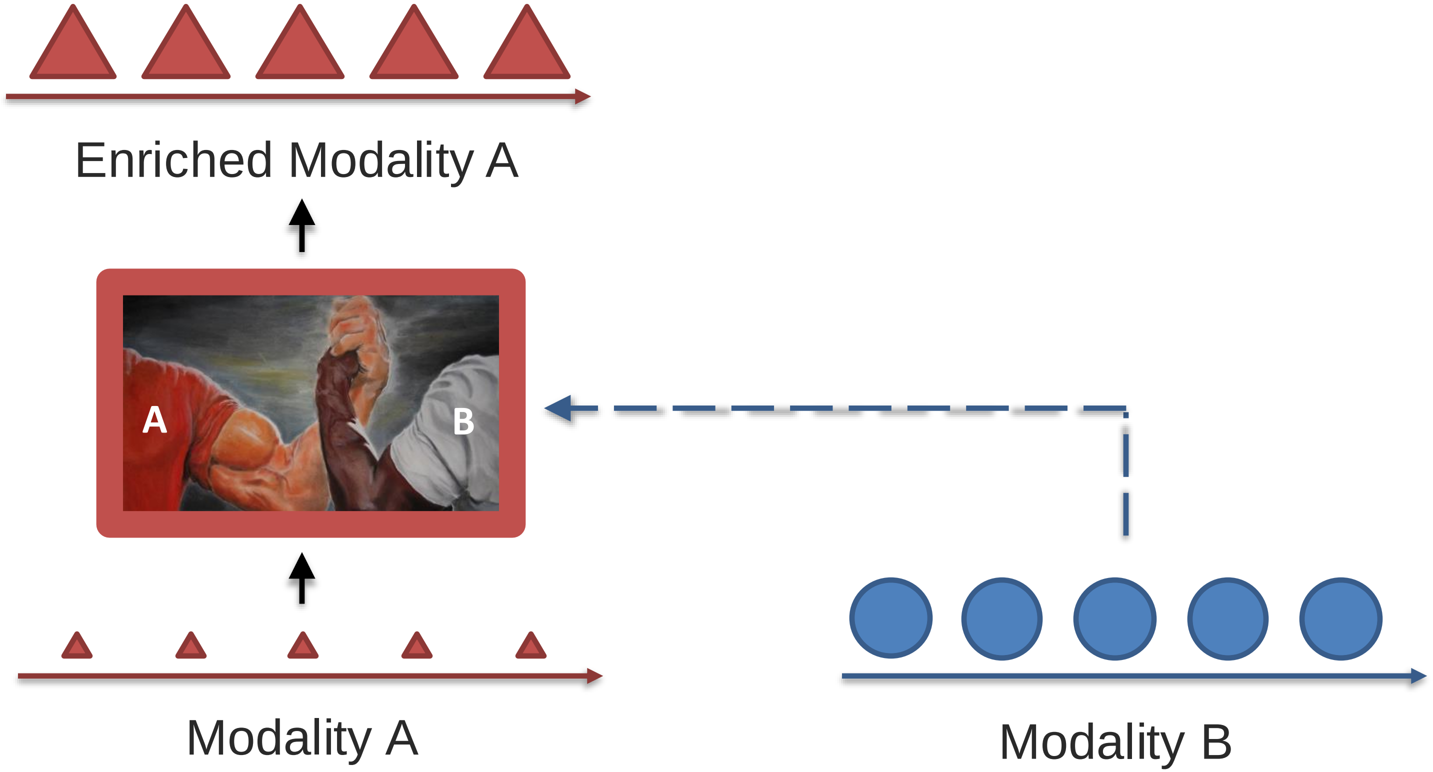


Expansion



Challenge 5: Transference

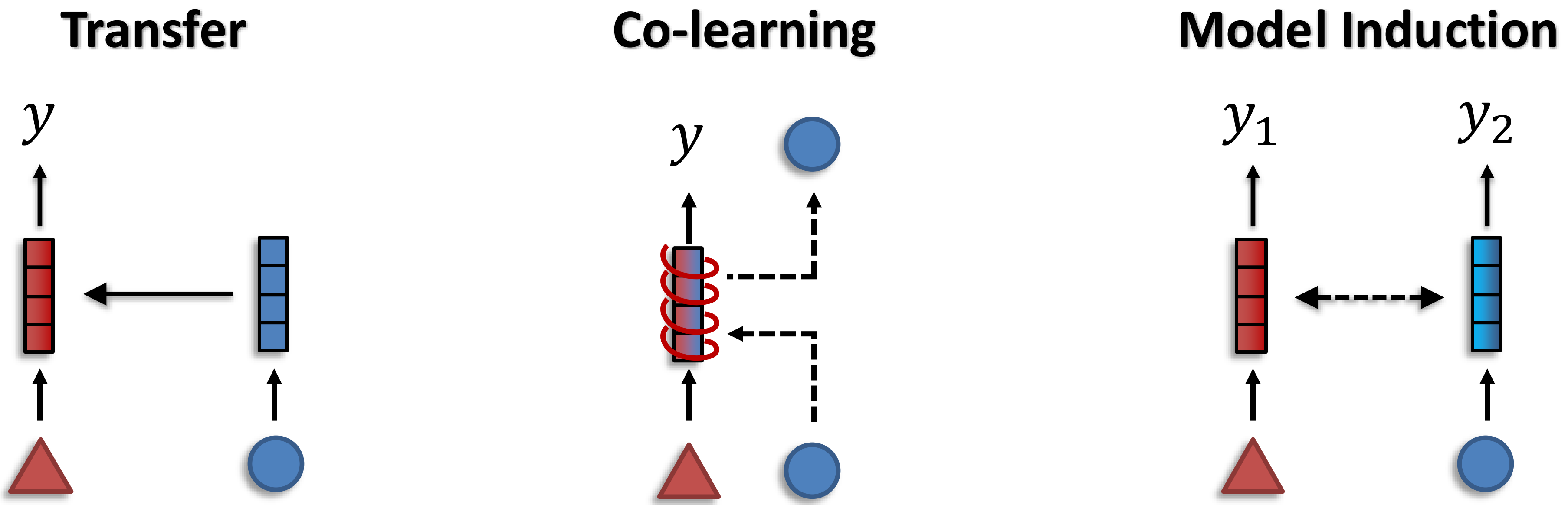
Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources.



Challenge 5: Transference

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

Sub-challenges:

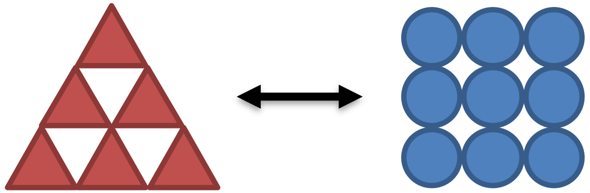


Challenge 6: Quantification

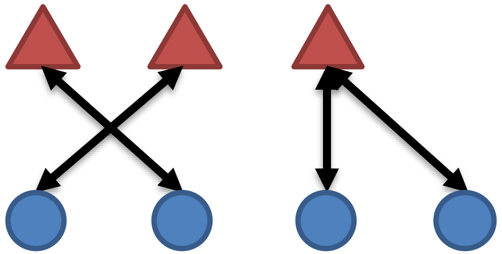
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions, and the multimodal learning process.

Sub-challenges:

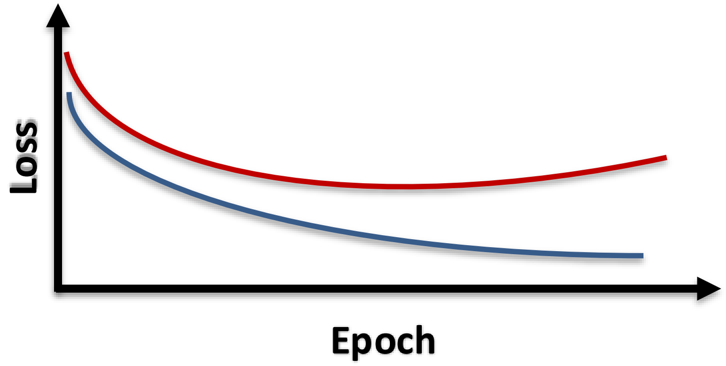
Heterogeneity



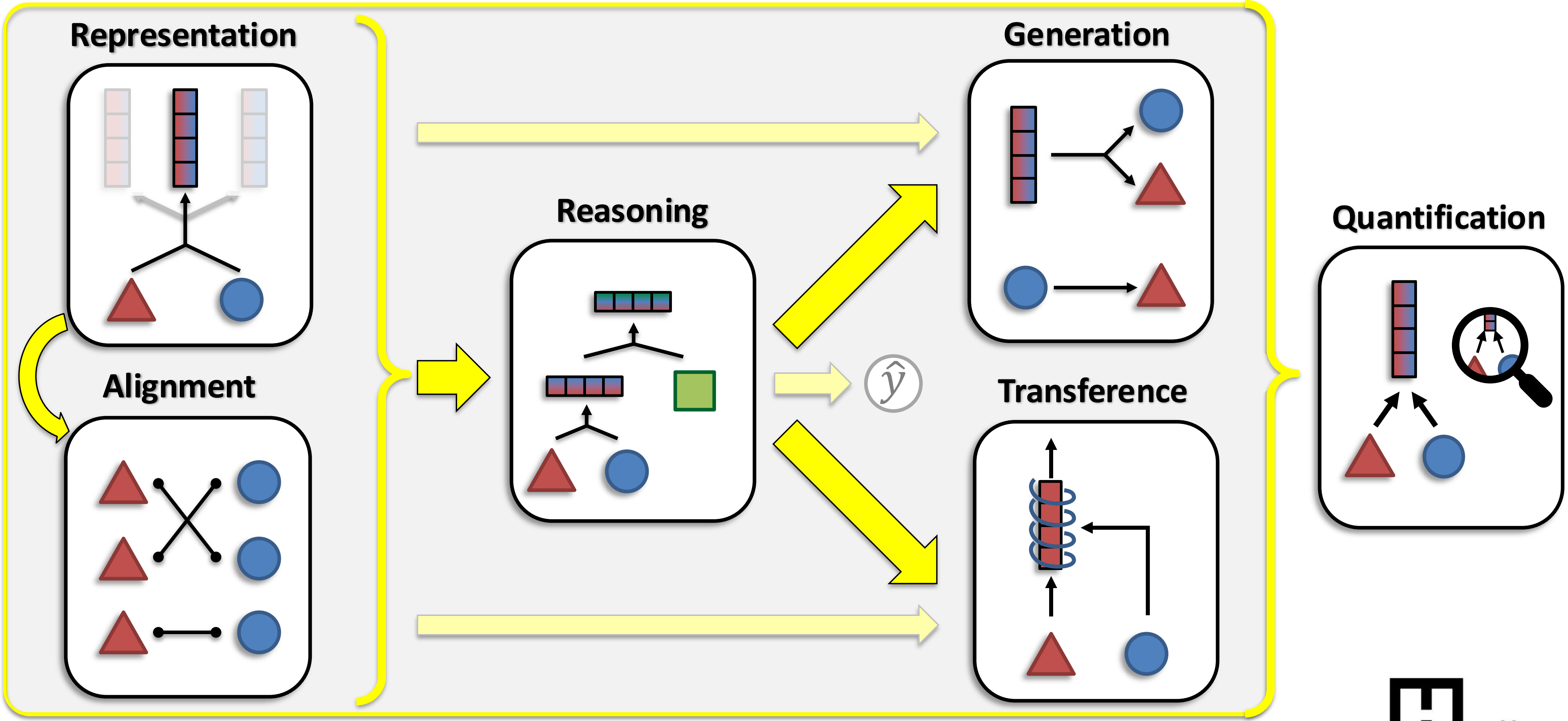
Interactions



Learning



Summary of Core Multimodal Challenges



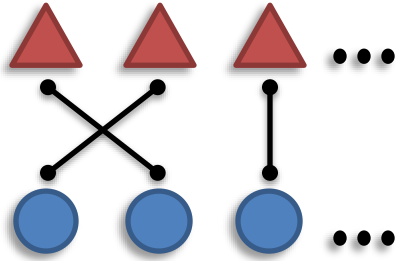
multisensory intelligence

Challenge 2: Alignment

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

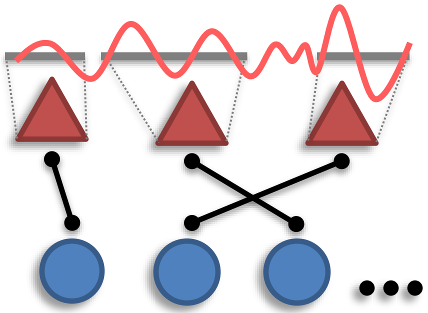
Sub-challenges:

Discrete connections



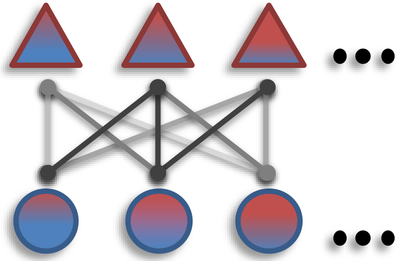
Explicit alignment
(e.g., grounding)

Continuous alignment



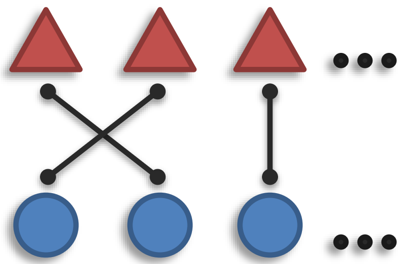
Granularity of individual elements

Contextualized representation

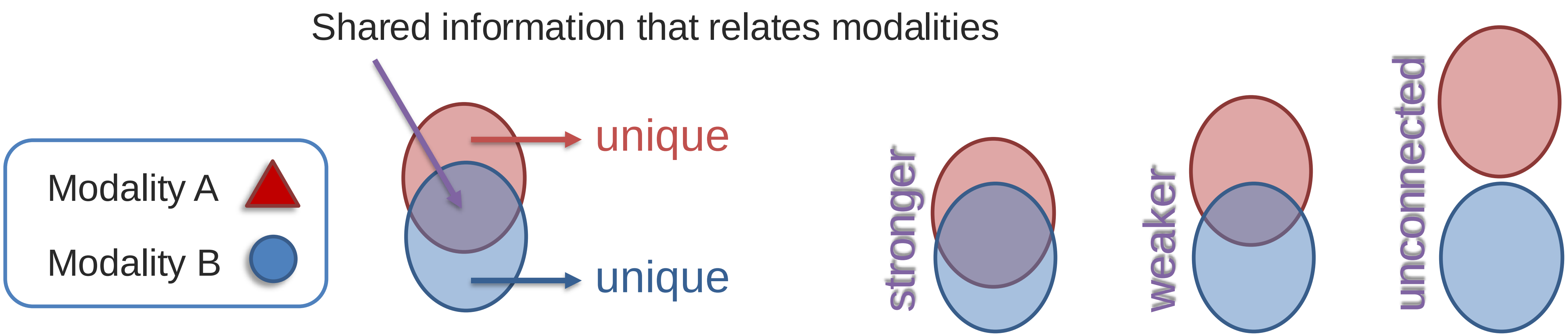


Implicit alignment + representation

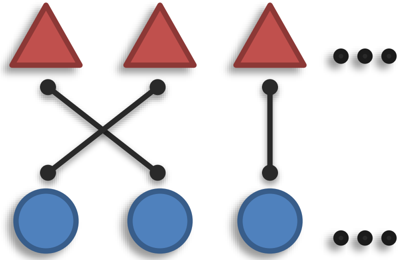
Challenge 2a: Discrete Alignment



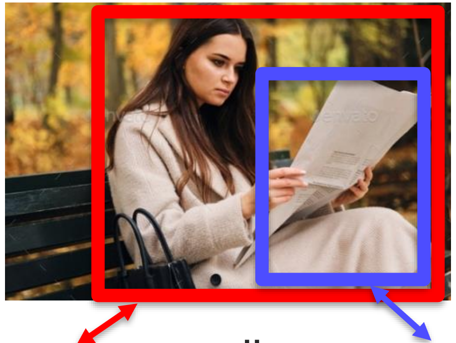
Definition: Identify and model connections between elements of multiple modalities



Modality Connections



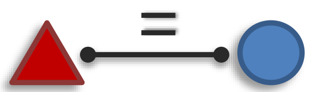
Definition: Tying language (words, phrases,…) to non-linguistic elements, such as the visual world (objects, people, …)



A woman reading newspaper

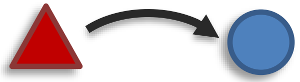
Statistical

Association



e.g., correlation, co-occurrence

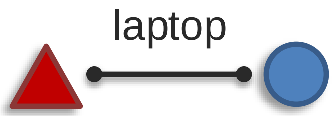
Dependency



e.g., causal, temporal

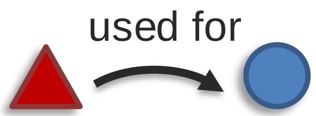
Semantic

Correspondence



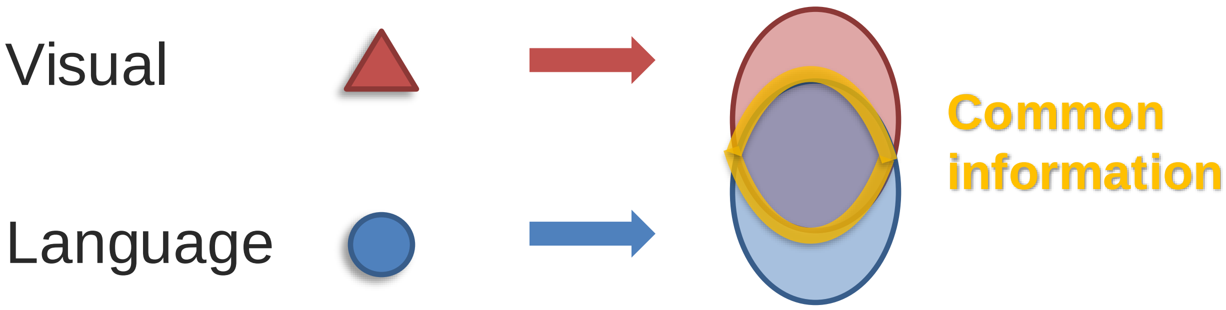
e.g., grounding

Relationship

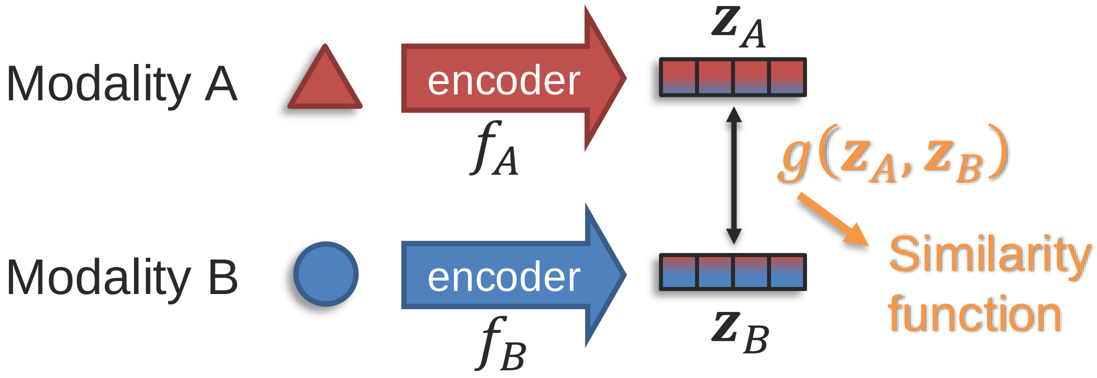


e.g., function

Modality Connections

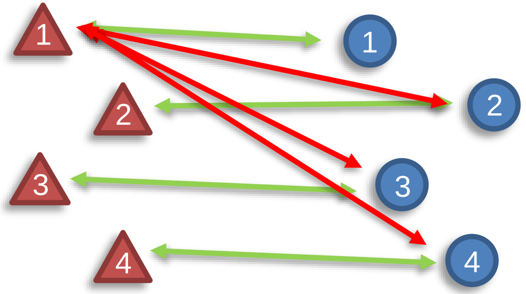


Learning aligned representations:

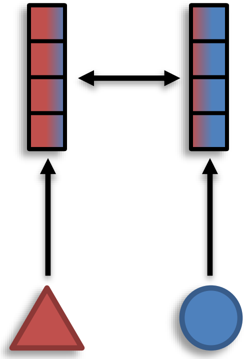


or contrastive learning

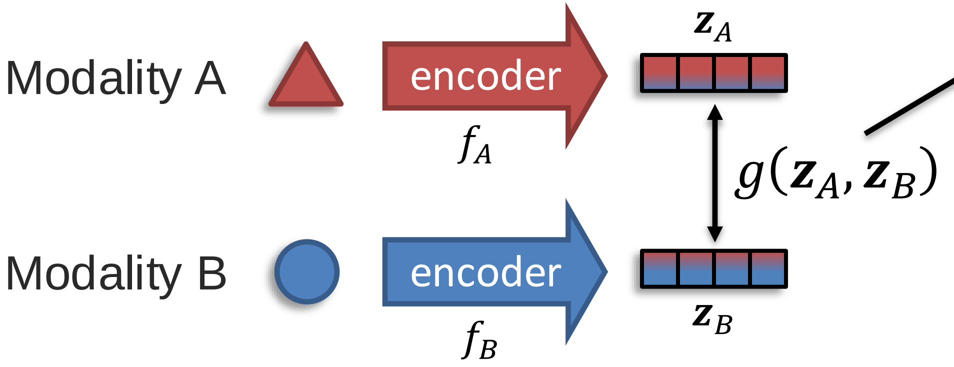
Supervision: Paired data



Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



2. Alignment function captures **connections**

1. Specialized encoders capture **heterogeneity**

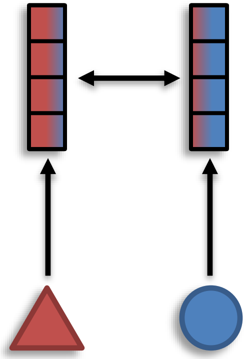
Learning with alignment function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

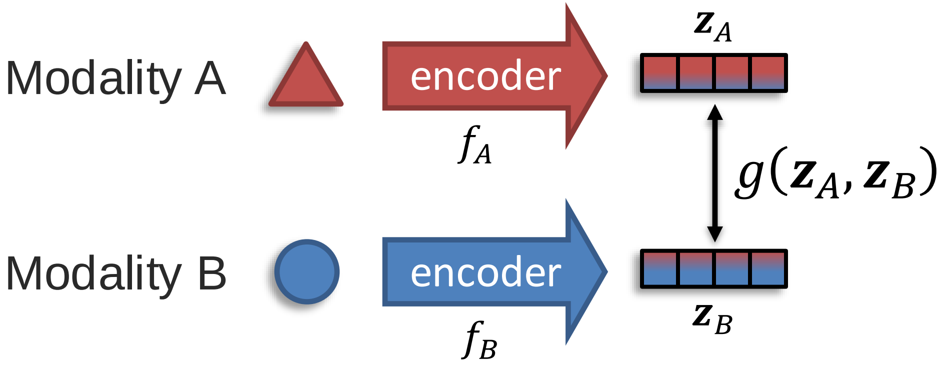
with model parameters θ_g, θ_{f_A} and θ_{f_B}

multisensory intelligence

Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



Learning with alignment function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

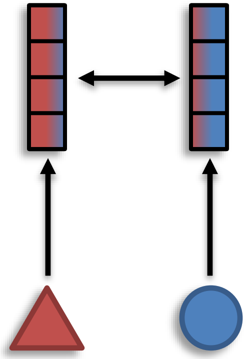
with model parameters θ_g, θ_{f_A} and θ_{f_B}

1 Cosine similarity:

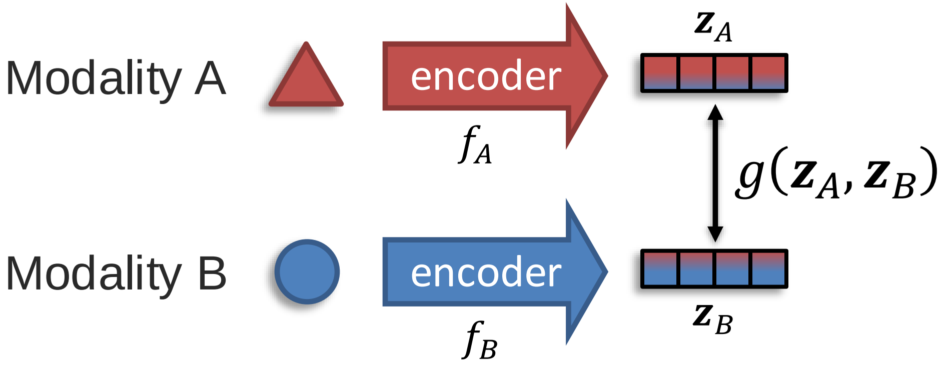
$$g(\boldsymbol{z}_A, \boldsymbol{z}_B) = \frac{\boldsymbol{z}_A \cdot \boldsymbol{z}_B}{\|\boldsymbol{z}_A\|\|\boldsymbol{z}_B\|}$$

multisensory intelligence

Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



Learning with alignment function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

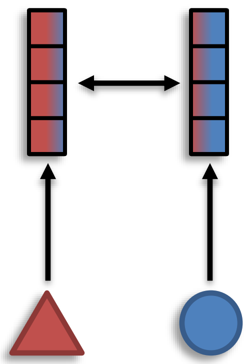
② Kernel similarity functions:

$$g(\boldsymbol{z}_A, \boldsymbol{z}_B) = k(\boldsymbol{z}_A, \boldsymbol{z}_B)$$

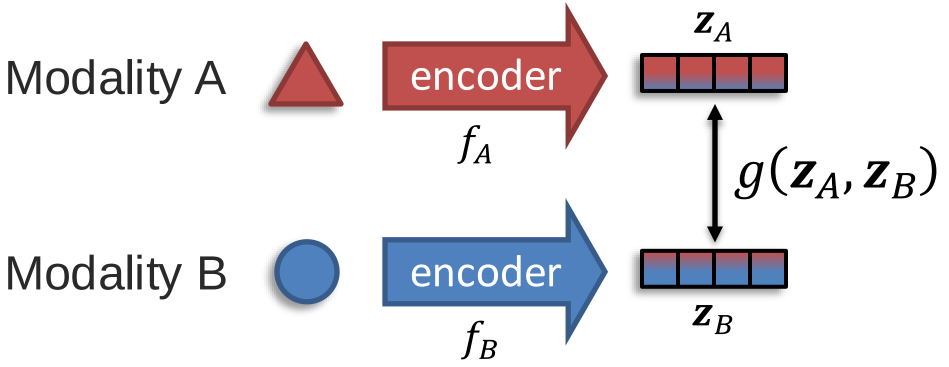
- Linear
- Polynomial
- Exponential
- RBF

multisensory intelligence

Aligned Representations



Definition: Learn multimodal representations aligned through their connections.



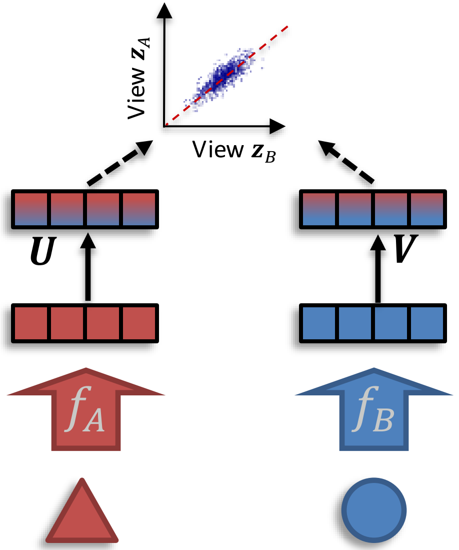
Learning with alignment function:

$$\mathcal{L} = g(f_A(\triangle), f_B(\bigcirc))$$

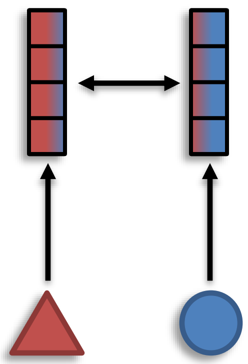
with model parameters θ_g, θ_{f_A} and θ_{f_B}

③ Canonical Correlation Analysis (CCA):

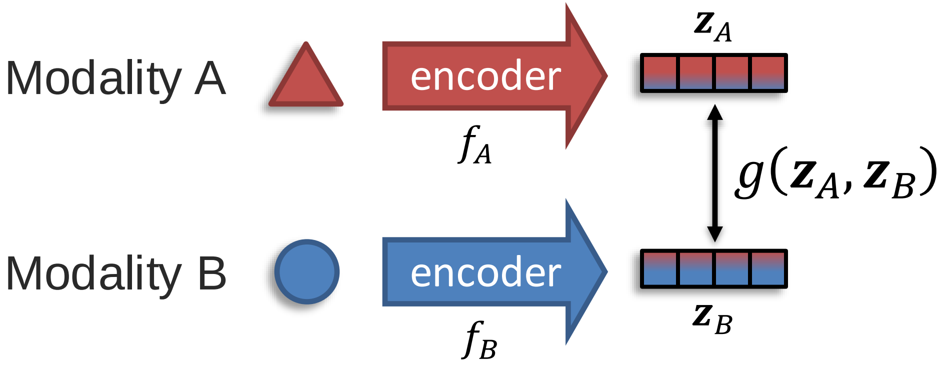
$$\underset{\boldsymbol{V},\boldsymbol{U},f_A,f_B}{\operatorname{argmax}} \; corr(\boldsymbol{z}_A, \boldsymbol{z}_B)$$



Aligned Representations



Definition: Learn multimodal representations aligned through their connections.

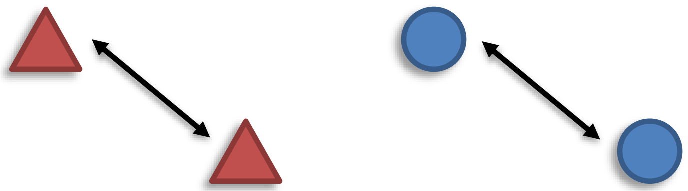


Learning with alignment function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

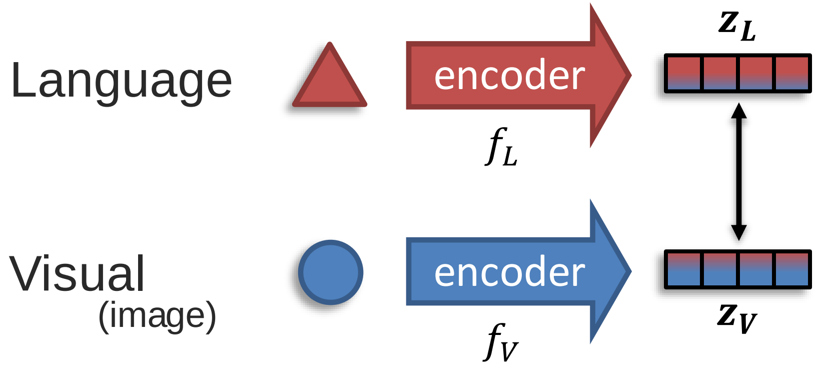
with model parameters θ_g, θ_{f_A} and θ_{f_B}

4 Order, hierarchy, pairwise relationships.

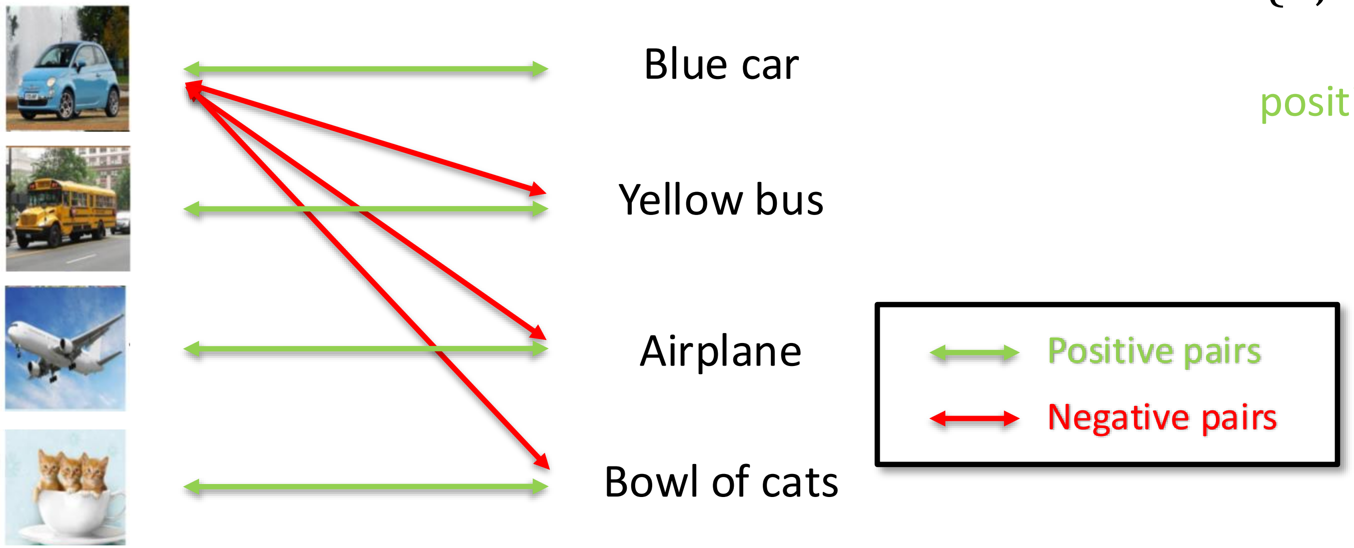


multisensory intelligence

Alignment with Contrastive Learning



Paired data: { ▲ , ● } (e.g., images and text)



Contrastive loss:

brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

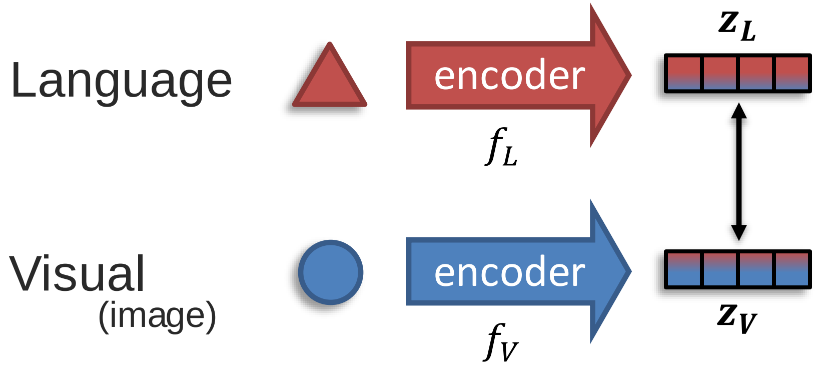
$$\max\{0, \alpha + \boxed{g(\boldsymbol{z}_A, \boldsymbol{z}_B^+)} - \boxed{g(\boldsymbol{z}_A, \boldsymbol{z}_B^-)}\}$$

positive pairs

negative pair

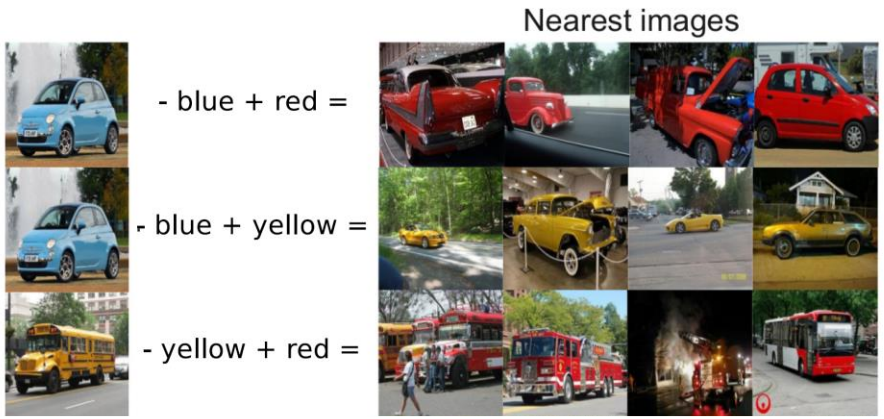
Coordination function
(e.g., cosine similarity)

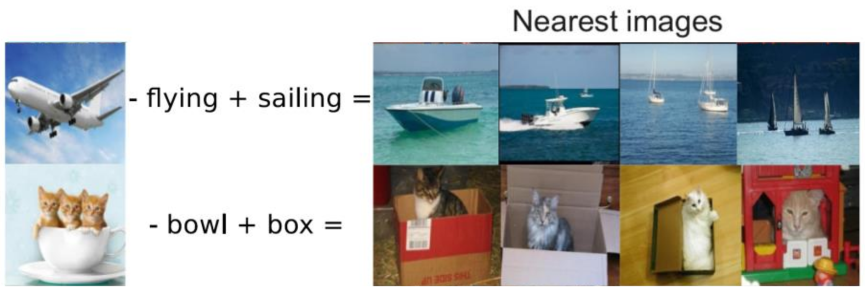
Visual-Semantic Embeddings



Contrastive loss:

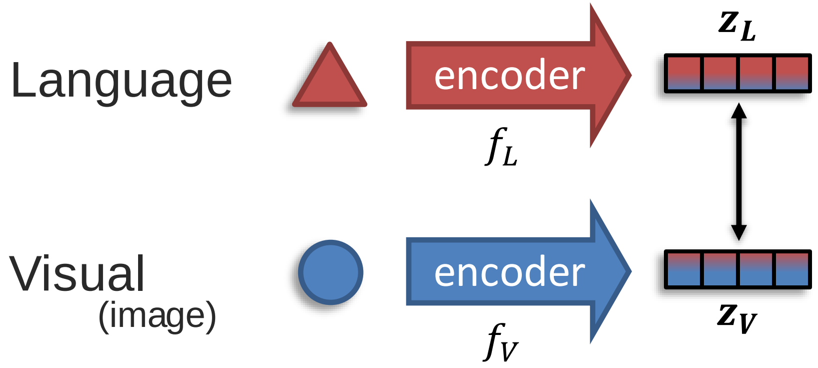
brings positive pairs closer and pushes negative pairs apart





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

Contrastive Language Image Pretraining



Popular contrastive loss: InfoNCE

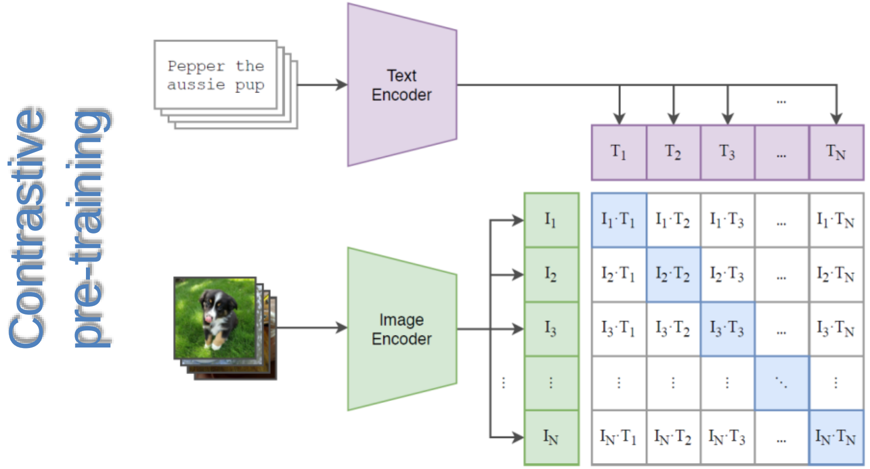
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\boxed{\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}}{\boxed{\sum_{j=1}^{N} \mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}}$$

positive pairs

negative pairs
and positive pairs

Similarity function can
be cosine similarity

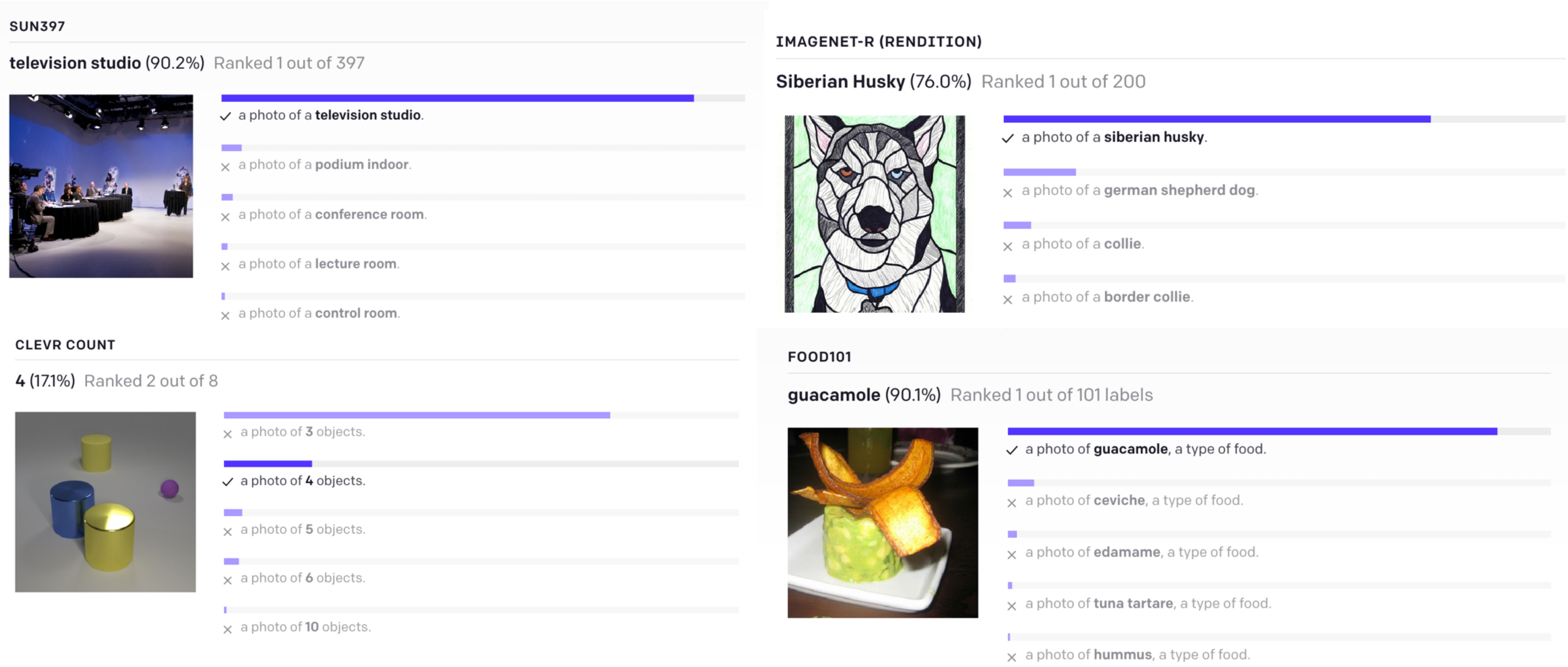
Positive and negative pairs:



- CLIP encoders (f_L and f_V) are great for language-vision tasks
- $\boldsymbol{z}_L$ and $\boldsymbol{z}_V$ are coordinated but not identical representation spaces

[Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021]

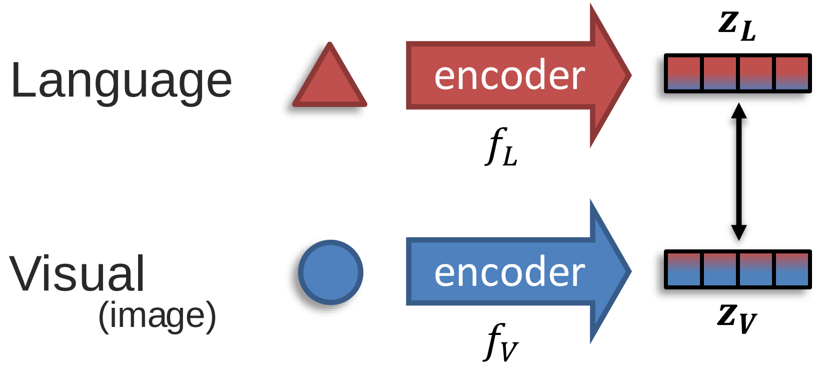
CLIP (Contrastive Language–Image Pre-training)



[Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021]

multisensory intelligence

Contrastive Learning and Connected Modalities

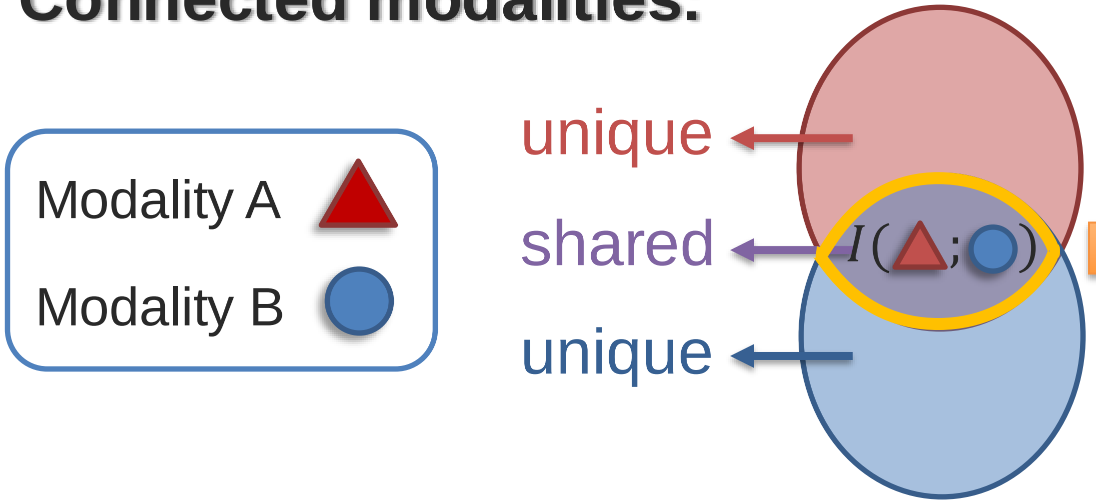


Popular contrastive loss: InfoNCE

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}{\sum_{j=1}^{N} \mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

CLIP focuses on shared connections

Connected modalities:

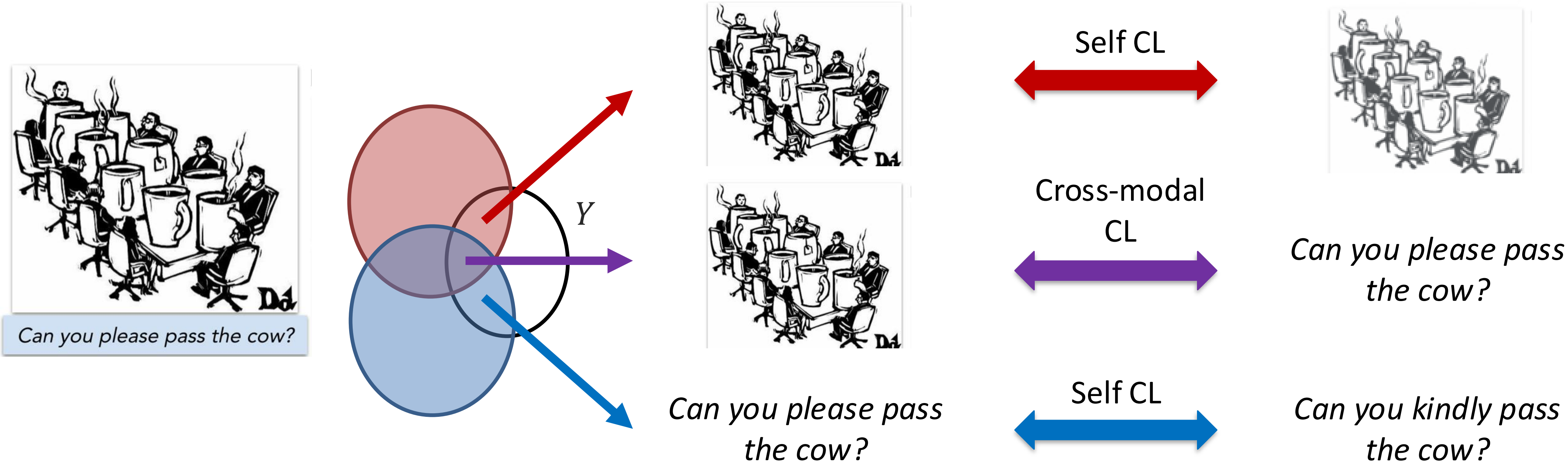


Mutual information $I(X;Y)$

$$\mathbb{E}_{X,Y}\left[\log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right]$$

[Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021]

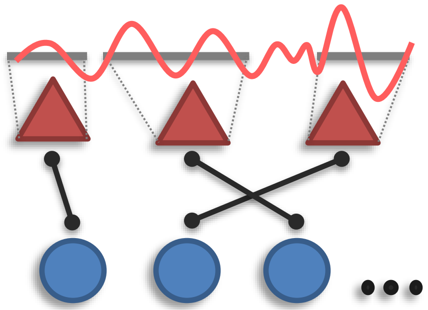
Factorized Contrastive Learning



Learns both shared and unique information.

[Liang et al., Factorized Contrastive Learning: Going Beyond Multi-view Redundancy, NeurIPS 2023]

Challenge 2b: Continuous Alignment

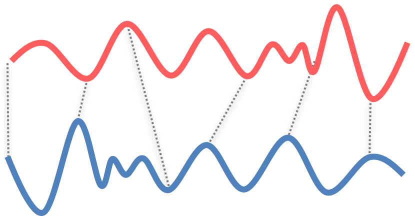


Definition: Model alignment between modalities with continuous signals and no explicit elements

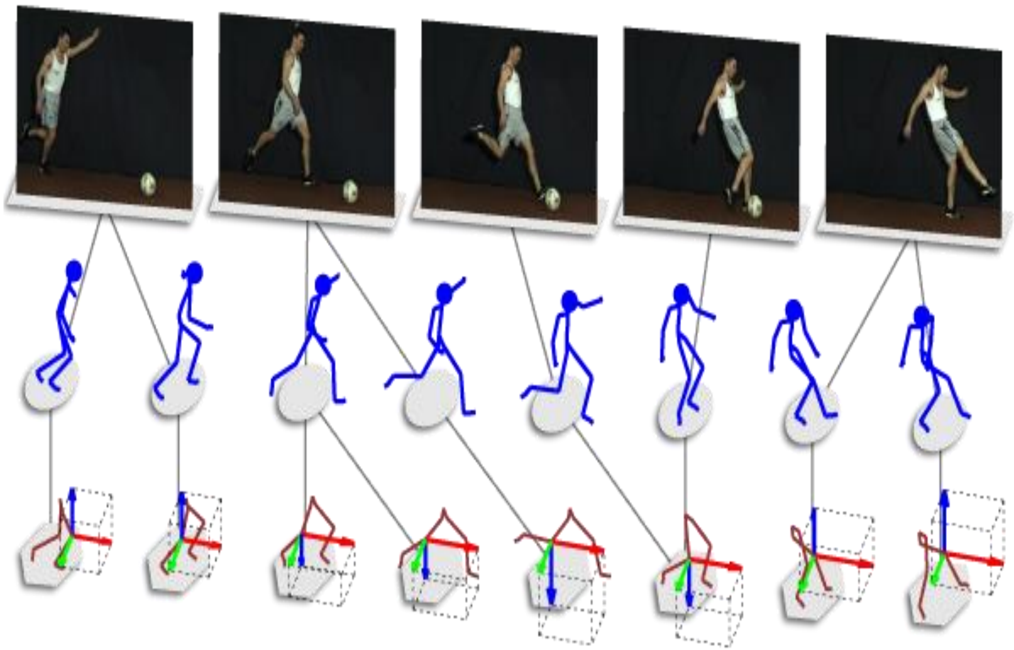
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Challenge 2b: Continuous Alignment

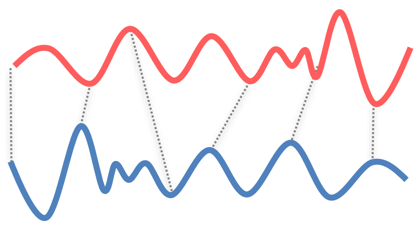


Aligning video sequences

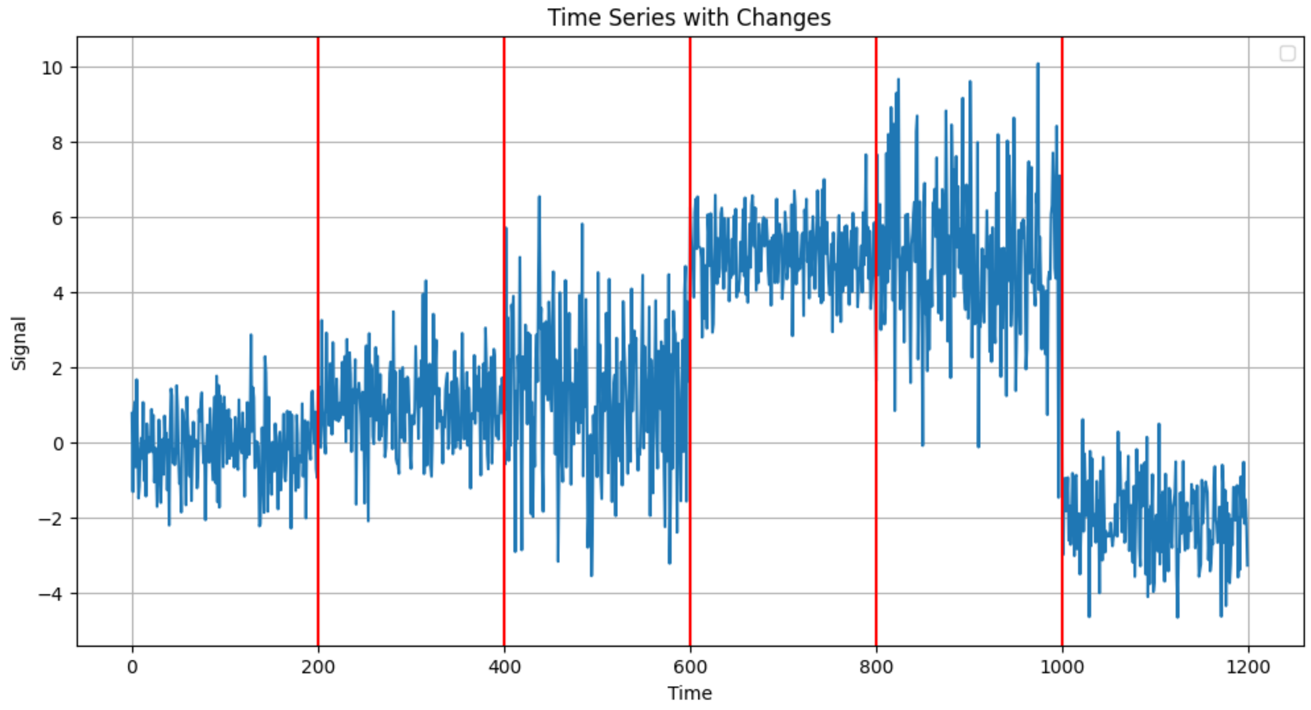


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]
[Zhou and De la Torre. Canonical Time Warping for Alignment of Human Behavior, NeurIPS 2009]

Challenge 2b: Continuous Alignment



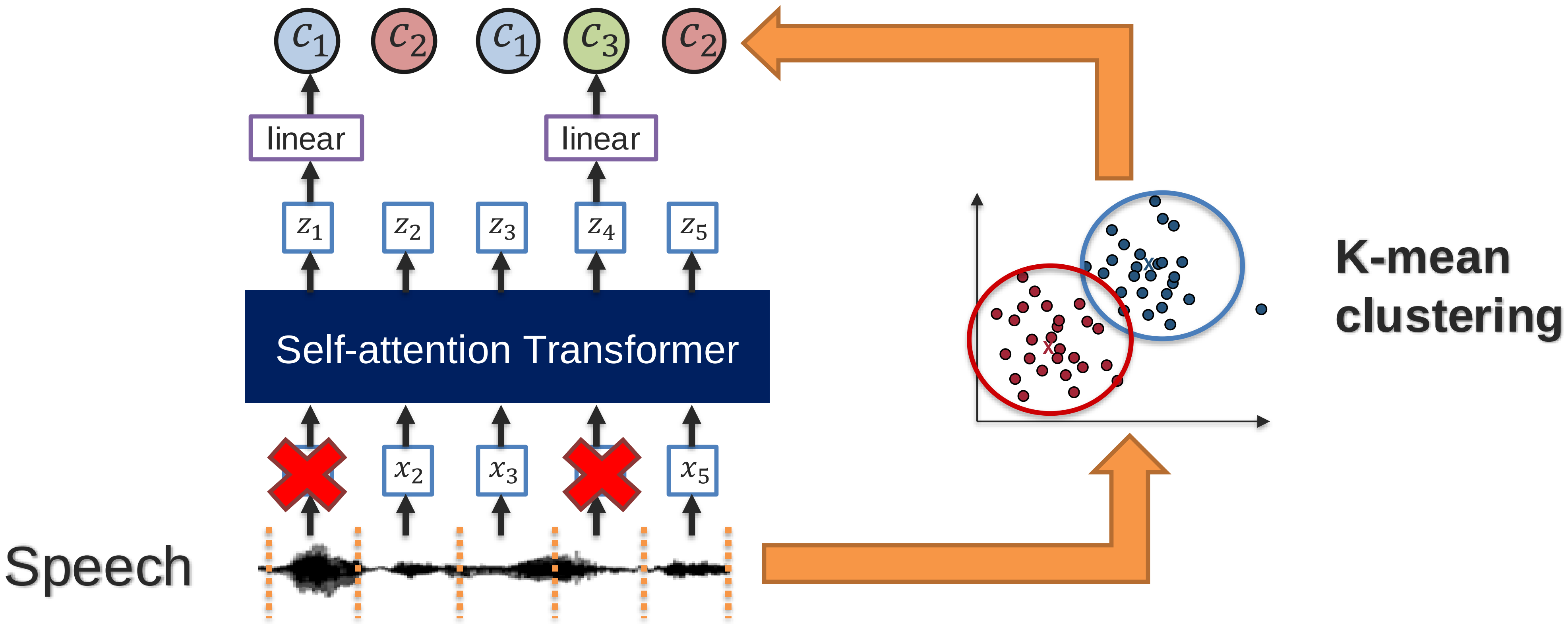
Changepoint detection



multisensory intelligence

Challenge 2b: Continuous Alignment

HUBERT: Hidden-Unit BERT



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

[Hsu et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, IEEE TASLP 2021]

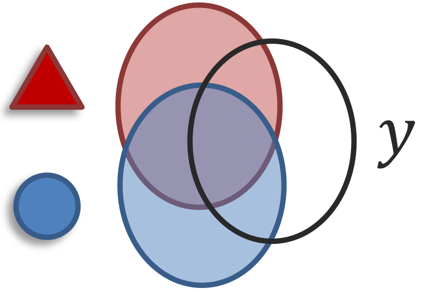
multisensory intelligence

Today's lecture

1. Introduction to multimodal AI
2. Principles of heterogeneity, connections, interactions
3. Core multimodal challenges
4. Multimodal alignment

Summary: How To Multimodal

1. Think about data heterogeneity, connections, interactions.
2. Decide how much data in each modality to collect, and how much to label (costs and time).
3. Clean data: normalize/standardize, find noisy data, anomaly/outlier detection
4. Visualize data: plot, dimensionality reduction (PCA, t-sne), cluster analysis
5. Decide on evaluation metric (proxy + real, quantitative and qualitative)
6. Figure out what challenge and sub-challenge, and latest work in that space.
7. Decide whether to build on prior work, try general-purpose or domain-specific models, top-down vs bottom-up research etc.



multisensory intelligence

Assignments for This Coming Week

Reading assignment due tomorrow Wednesday (3/5).

This Thursday (3/6): second reading discussion on **modern AI architectures**.

1. Scaling laws for multimodal models
2. Not all tokens are all you need?

For project:

- I gave feedback and assigned primary TA.
- Meet with me and primary TA every other week.

Next Tuesday: lecture on **multimodal representation fusion**