《多模态机器学习》

第一章 引言

黄文炳

中国人民大学高瓴人工智能学院

hwenbing@126.com

2024年秋季

考核方式

<table>
<tr><td colspan="2">平时考核
（50%）</td><td rowspan="2">期末分组项目展示
（50%）</td></tr>
<tr><td>课程作业
（70%）</td><td>研讨交流
（30%）</td></tr>
</table>

- **所占学分：**2学分
- **课程作业：**习题、编程等，每两个星期一次
- **研讨交流：**随机抽查、课堂积极问答、随堂小测试等
- **编程语言：**Python
- **助教：**张岳霖

参考资料

- 书籍
  - Multimodal Deep Learning, https://arxiv.org/pdf/2301.04856.pdf
- 网上课程
  - CMU 11-777 Multimodal Machine Learning, https://cmu-multicomp-lab.github.io/mmml-course/fall2022/
  - CMU 11-877 Advanced Topics in Multimodal Machine Learning, https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/
  - https://cmu-multicomp-lab.github.io/mmml-tutorial/schedule/
- 综述
  - Multimodal Machine Learning: A Survey and Taxonomy, https://arxiv.org/pdf/1705.09406.pdf
  - Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, https://arxiv.org/pdf/2209.03430.pdf

内容提纲

① 为什么要学习《多模态机器学习》

② 什么是“多模态机器学习”

③ 本课程将要学习的内容

④ 本章小结

内容提纲

① 为什么要学习《多模态机器学习》

② 什么是“多模态机器学习”

③ 本课程将要学习的内容

④ 本章小结

视觉问答 (VQA) by GPT-4

GPT-4 visual input example, Extreme Ironing:

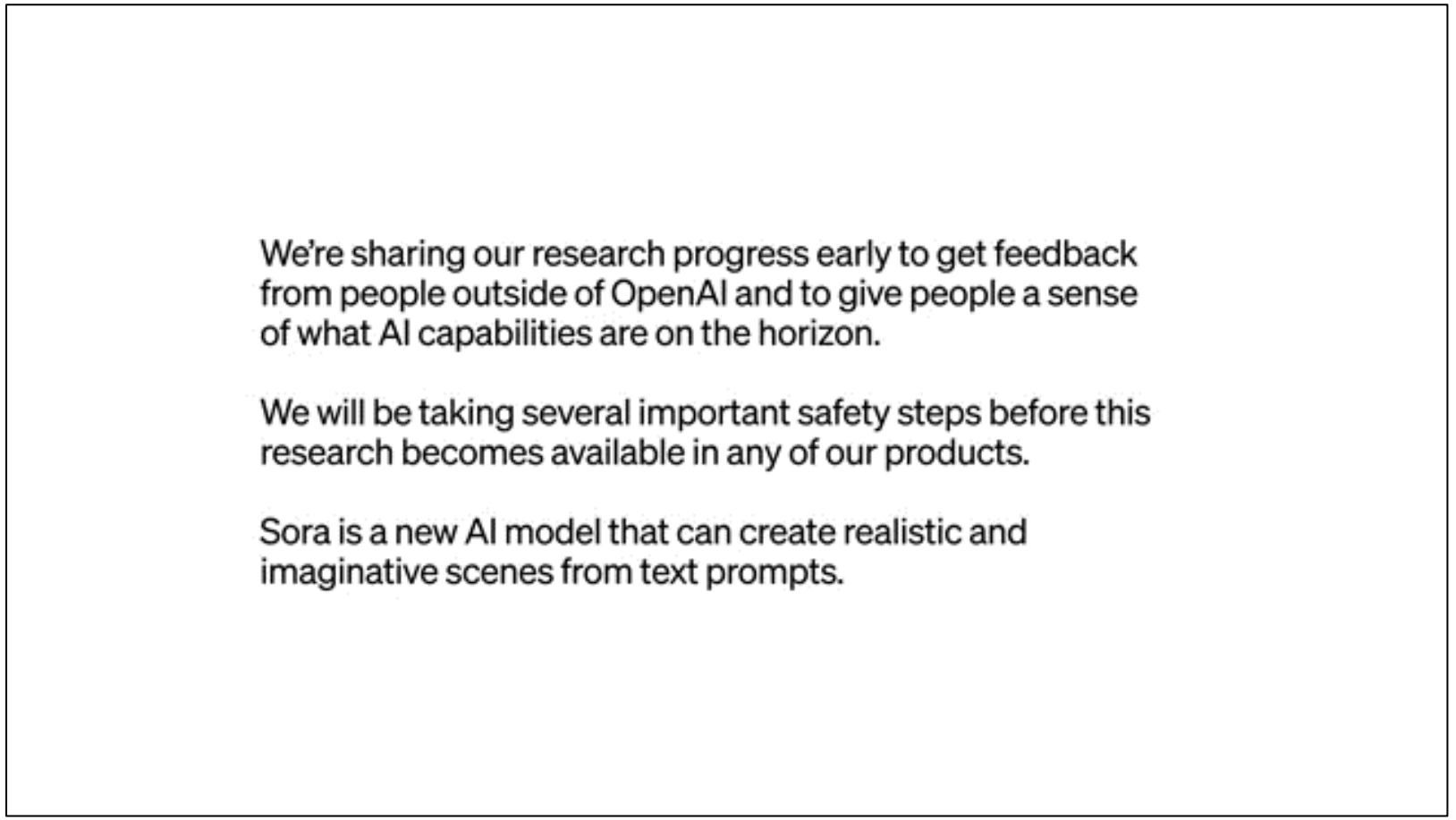
User What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

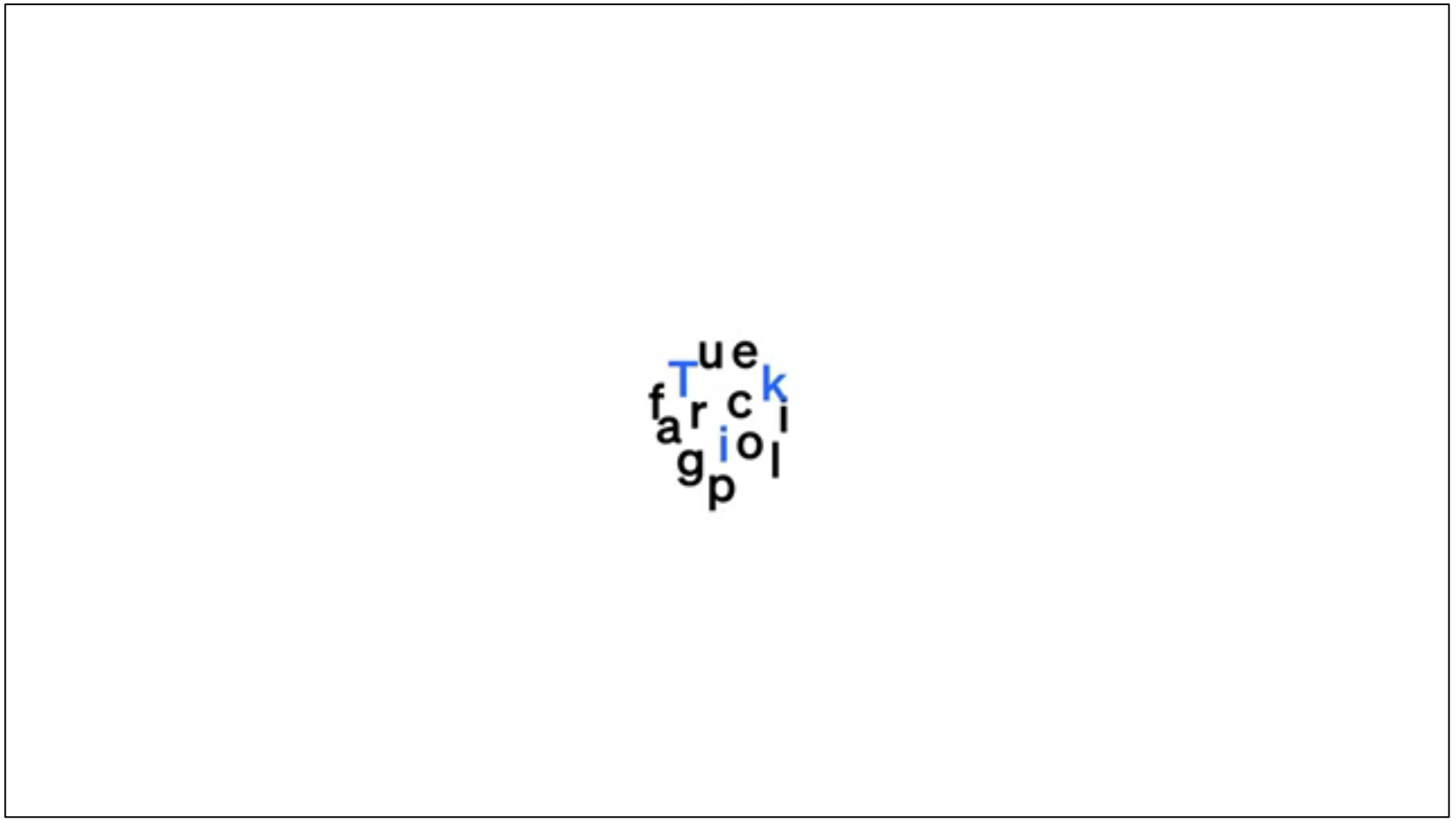
文生视频

Introducing Sora — OpenAI’s text-to-video model



文本、图像、语音

Hello GPT-4o

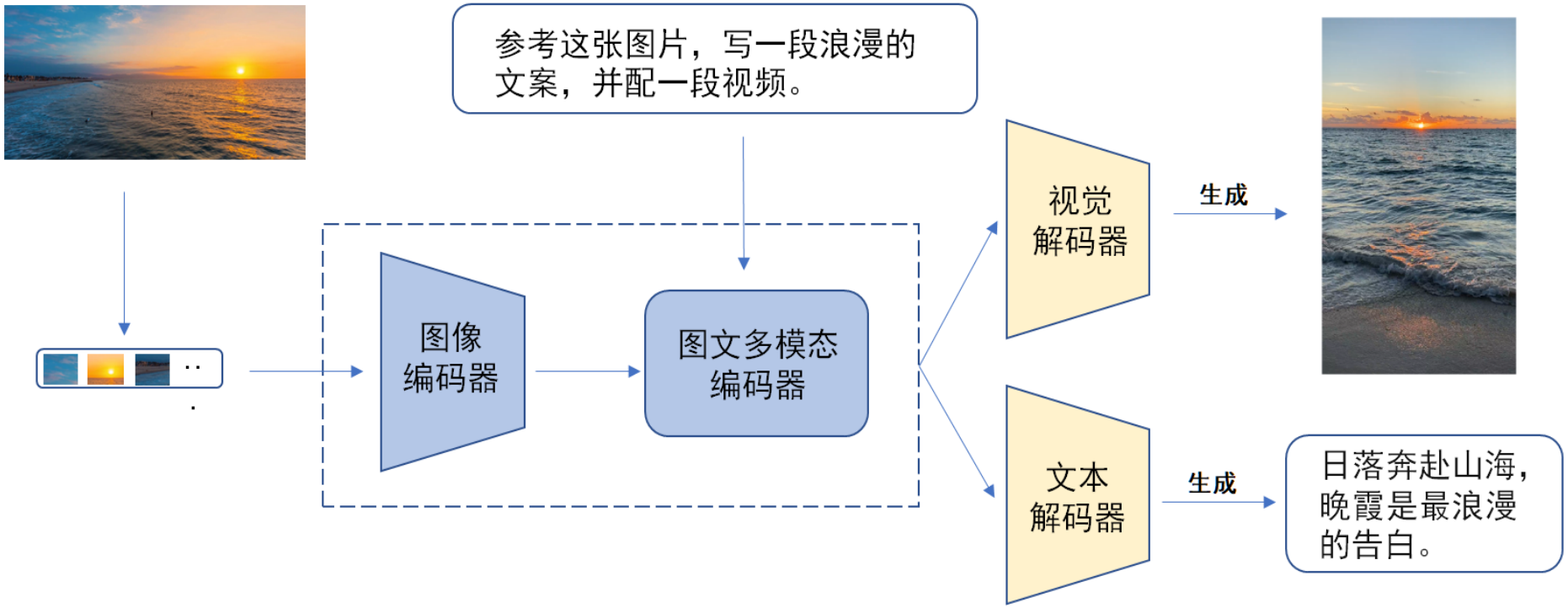


DeepMind 推出 RT-2：机器人的视觉 - 语言 - 动作（VLA）模型

DeepMind 研究人员跟机械臂说选择「已灭绝的动物」，机械臂手臂伸出，爪子张开落下，它抓住了恐龙玩偶。

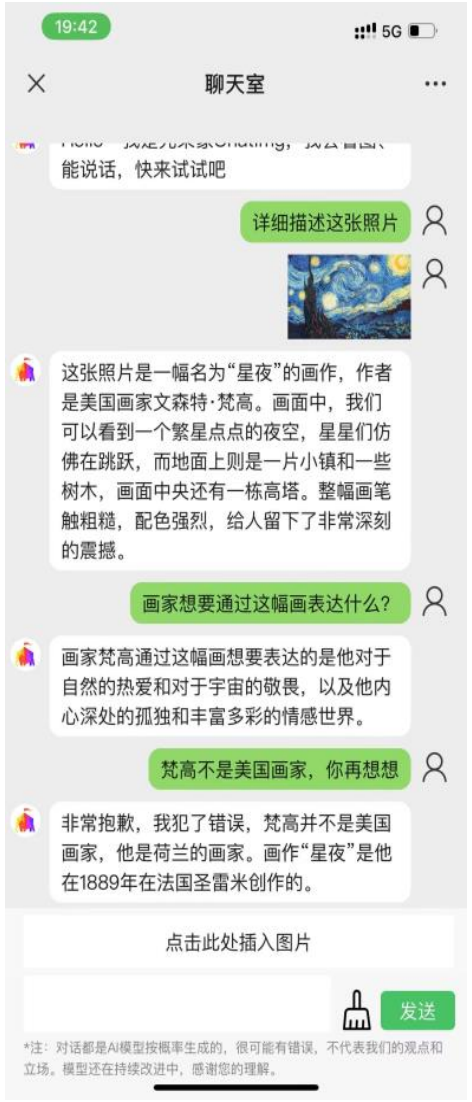
跟机器人说，把可乐罐给泰勒・斯威夫特

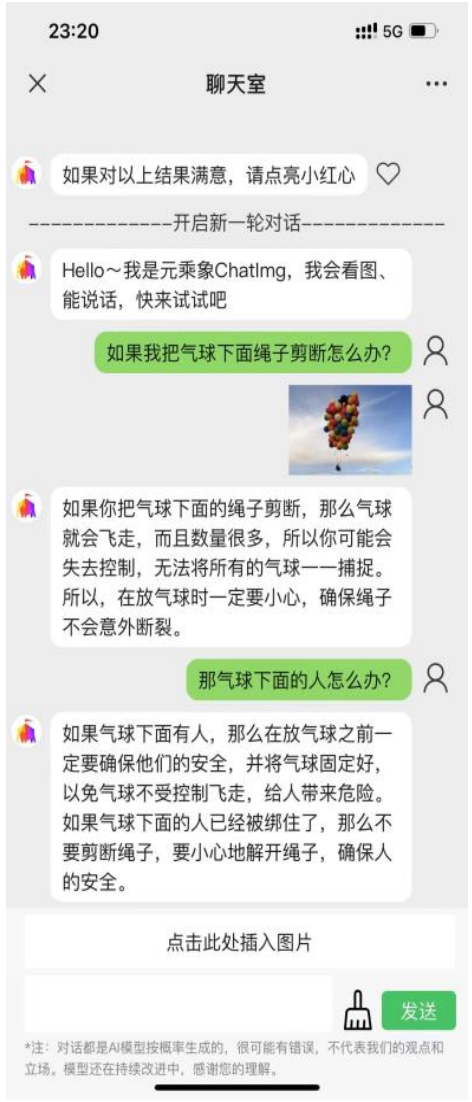
中国人民大学卢志武教授团队提出多模态通用生成模型：元成象 ChatImg



- 通用生成的定义：多模态输入，多模态输出（语言生成、图像生成、视频生成）

多模态通用生成模型：元成象 ChatImg





内容提纲

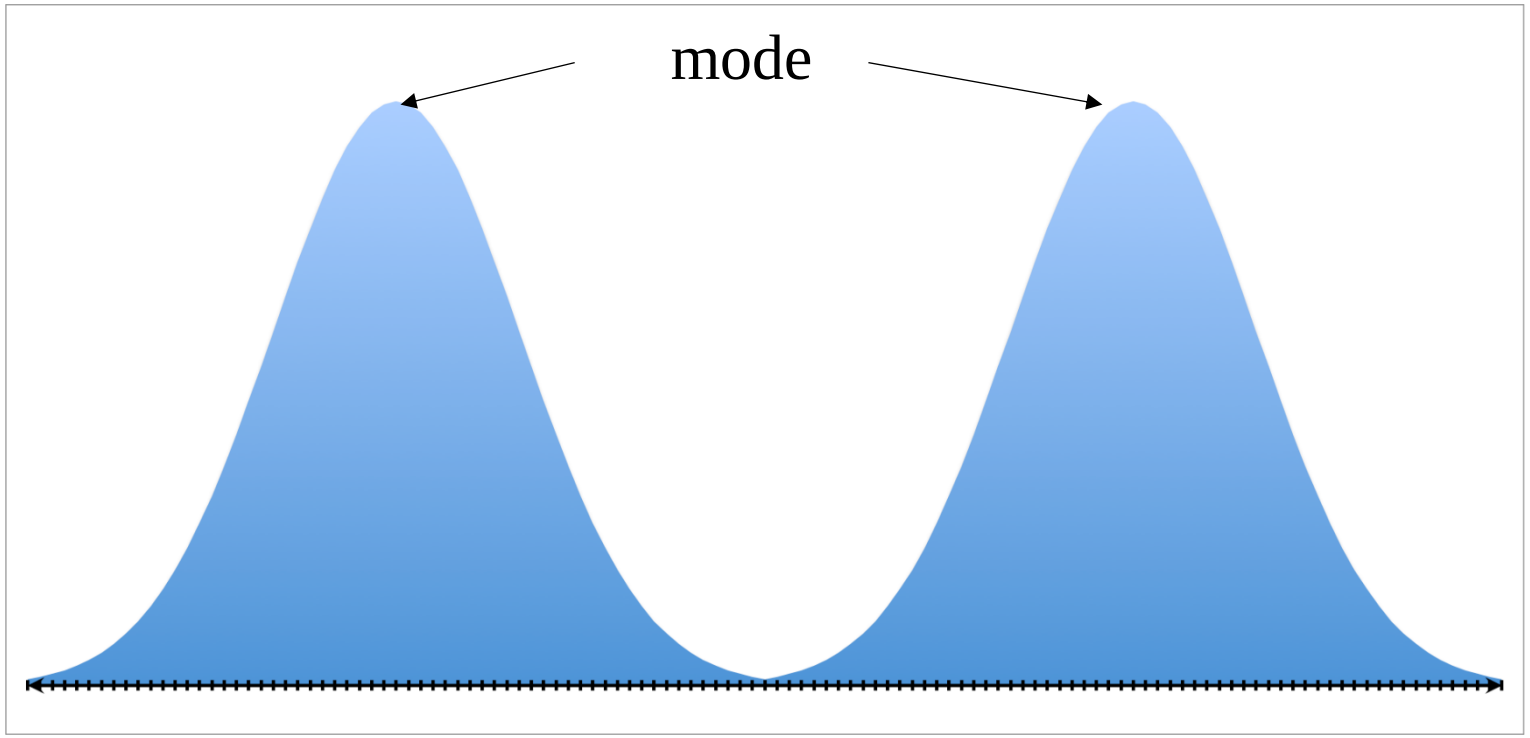
① 为什么要学习《多模态机器学习》

② 什么是“多模态机器学习”

③ 本课程将要学习的内容

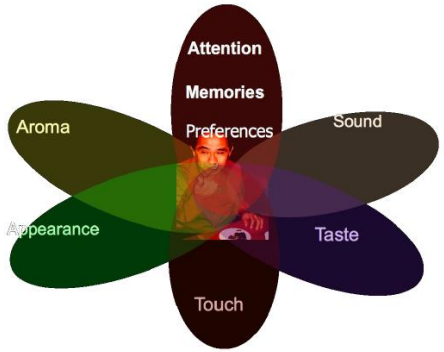
④ 本章小结

What is Multimodal?

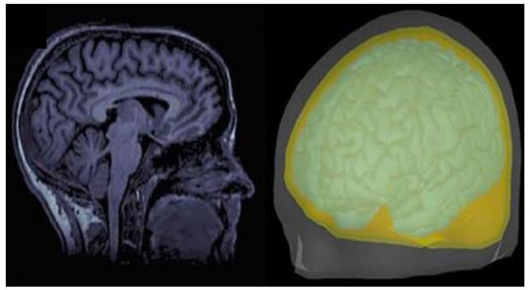


In statistics, a multimodal distribution is a probability distribution with more than one mode

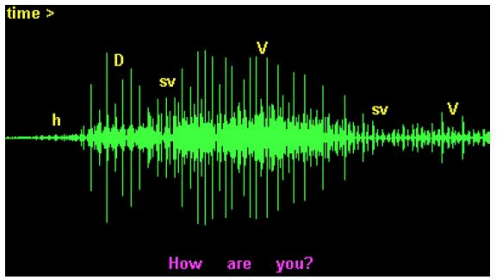
What is Multimodal?



Psychology

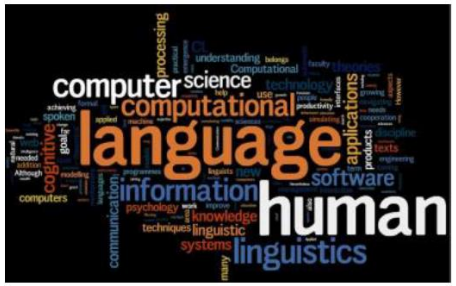


Medical



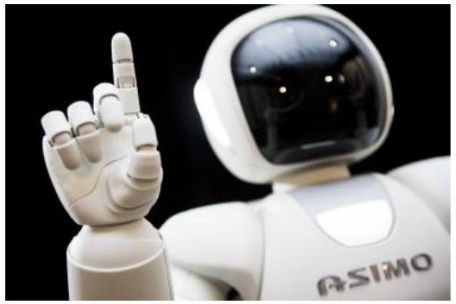
Speech

Vision

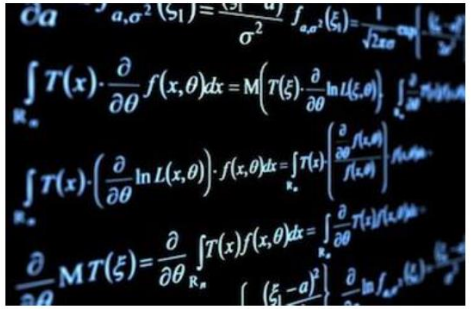


Language

Multimedia



Robotics



Learning

What is Multimodal?

Language

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

Acoustic

- **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

Touch

- **Haptics**
- **Motion**

Physiological

- **Skin conductance**
- **Electrocardiogram**

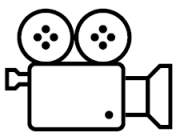
Mobile

- **GPS location**
- **Accelerometer**
- **Light sensors**

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	Language	Sentiment intensity
Image	Detected objects	Object categories

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

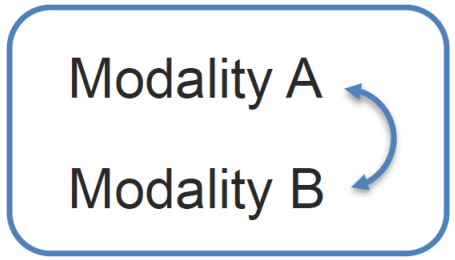
A research-oriented definition…

***Multimodal* is the science of**

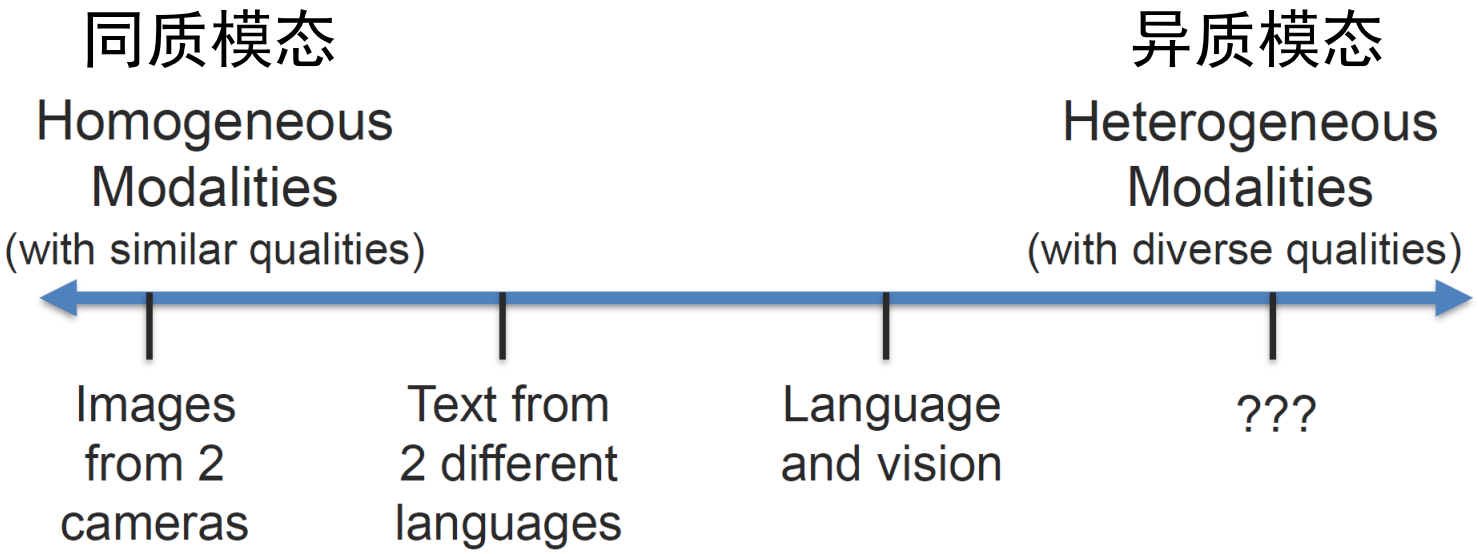
heterogeneous and interconnected data

Heterogeneous Modalities

Information present in different modalities will often show diverse qualities, structures and representations.



Examples:



Abstract modalities are more likely to be homogeneous

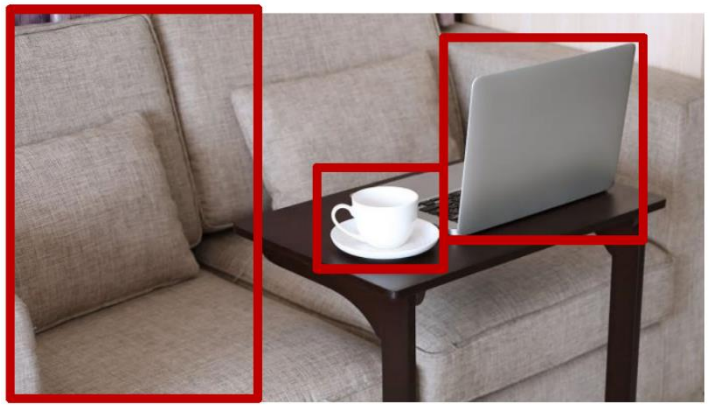
Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures and representations.

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

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures and representations.



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

1. **Element representations:** discrete, continuous, granularity

▲ {  ,  ,  }

● *{teacup, right, laptop, clean, room}*

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures and representations.

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

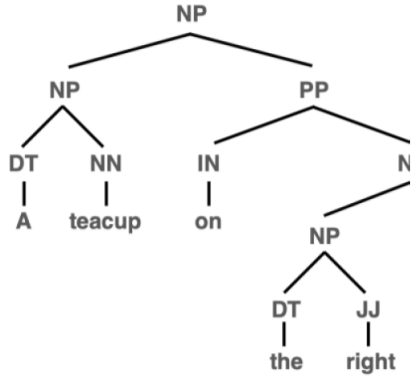
2 **Element distributions:** density, frequency

▲ ▲ ▲ objects per image

●●●● words per minute

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures and representations.



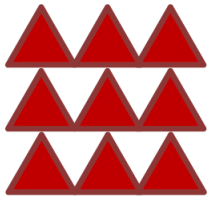
...

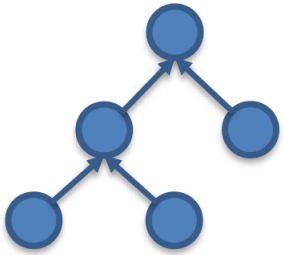
Latent (implicit)

A teacup on the right …

Explicit (observable)

3 **Structure:** temporal, spatial, hierarchical, latent, explicit





Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures and representations.

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

4 **Information:** abstraction, entropy

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures and representations.

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

5 **Noise:** uncertainty, signal-to-noise ratio, missing data

teacup → teacip

right → rihjt

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures and representations.

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

6 **Relevance:** task relevance, context dependence

→ recreational
→ living room
→ right-handed

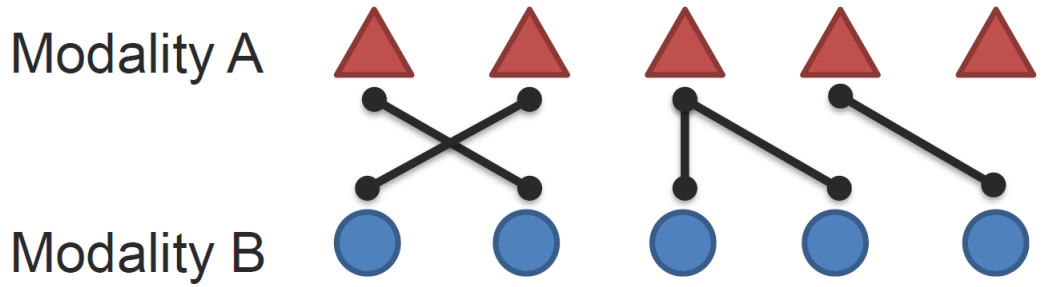
A teacup on the right of a laptop in a clean room.

→ workspace
→ study room

Interconnected Modalities

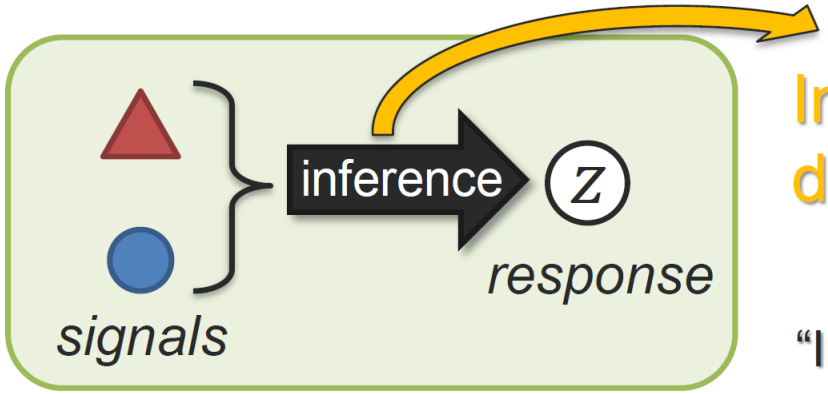
① Modality connections

Modalities are often related and share commonality



② Modality interactions

Modality elements often interact during inference



Interactions happen during inference!

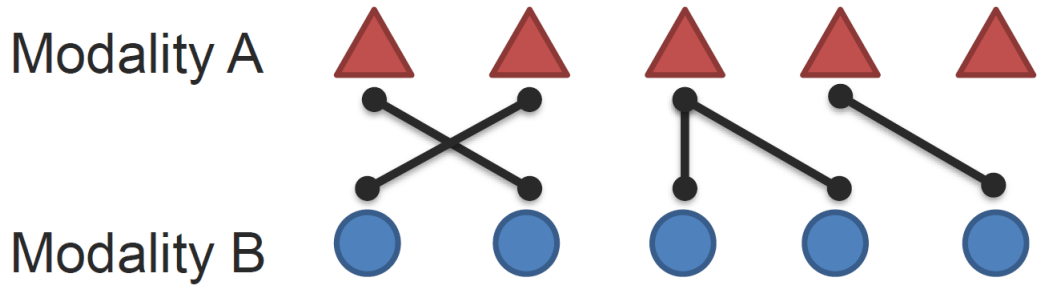
"Inference" examples:

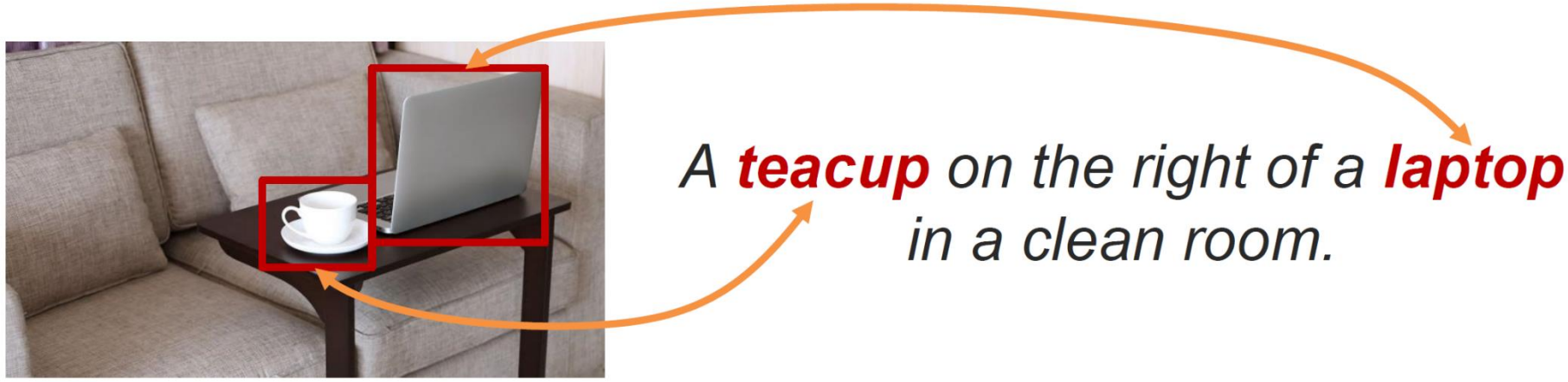
- Behavior perception
- Recognition task
- Modality translation

Interconnected Modalities

① Modality connections

Modalities are often related and share commonality

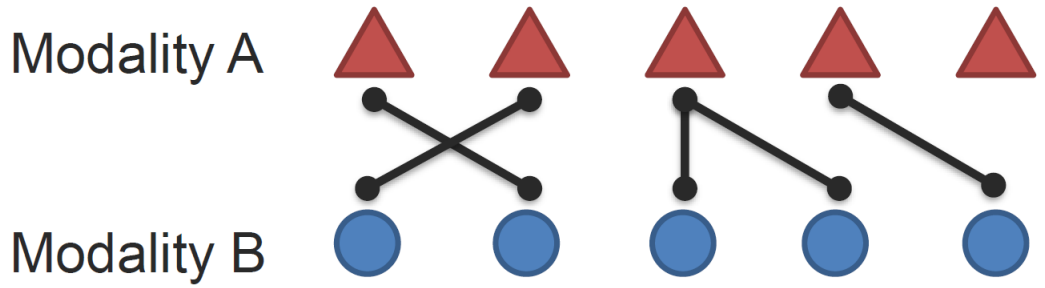




Interconnected Modalities

① Modality connections

Modalities are often related and share commonality

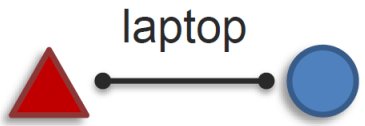


Statistical → **Semantic**

Association

e.g., correlation, co-occurrence

Correspondence

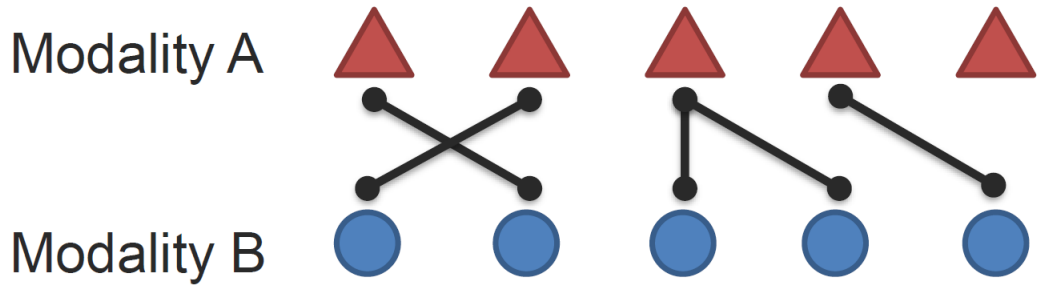


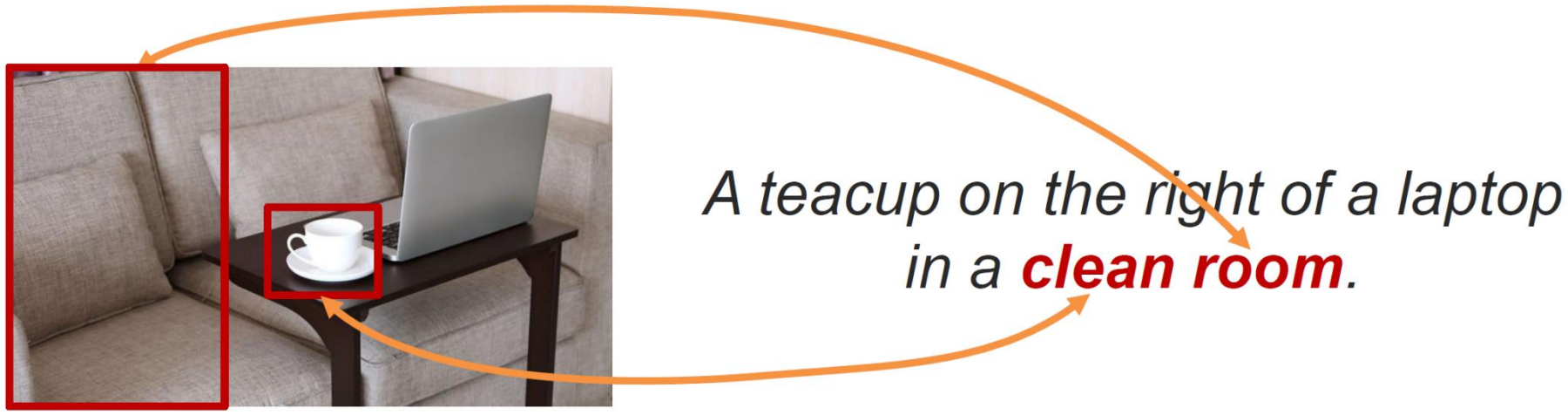
e.g., grounding

Interconnected Modalities

① Modality connections

Modalities are often related and share commonality

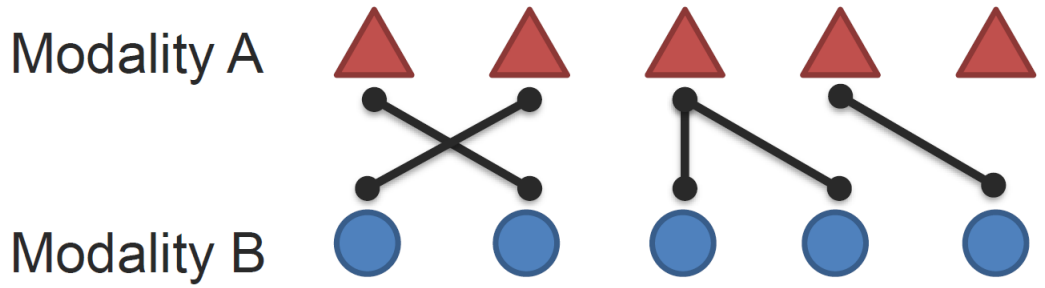




Interconnected Modalities

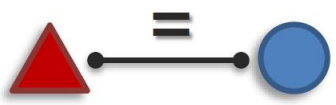
① Modality connections

Modalities are often related and share commonality



Statistical

Association



e.g., correlation, co-occurrence

Dependency

e.g., causal, temporal

Semantic

Correspondence

e.g., grounding

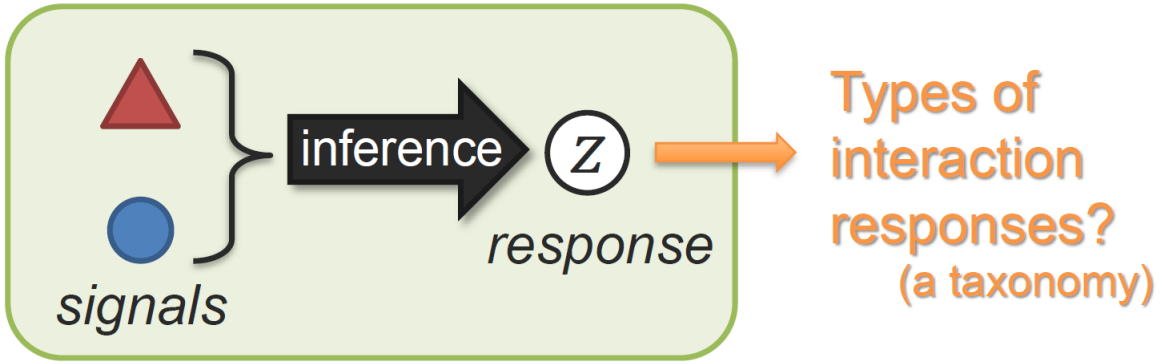
Relationship

e.g., function

Interconnected Modalities

② Modality interactions

Modality elements often interact during inference



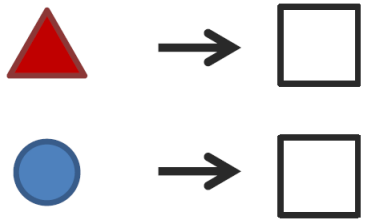
Is this indoors?

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

inference *Yes!*

inference *Yes!*

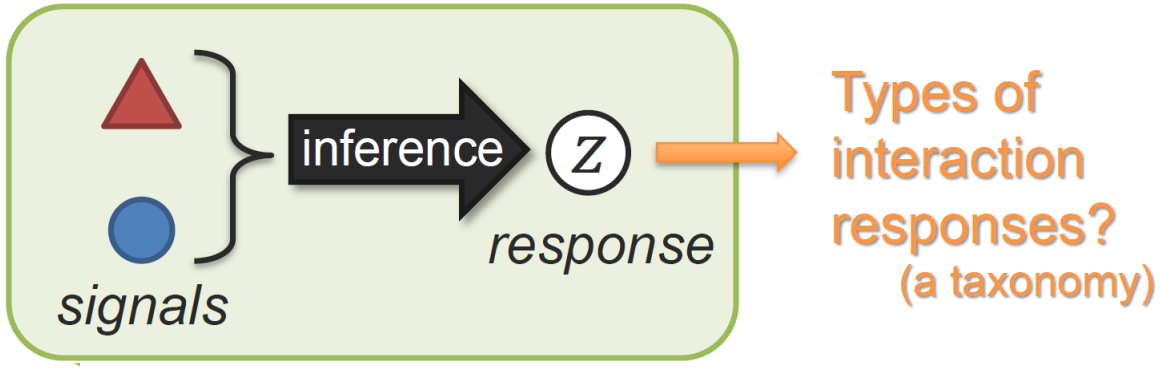
Unimodal redundancy



Interconnected Modalities

② Modality interactions

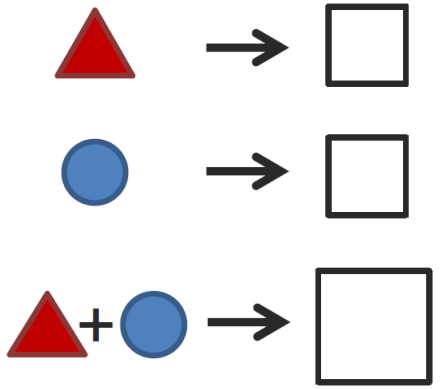
Modality elements often interact during inference



Is this indoors?

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

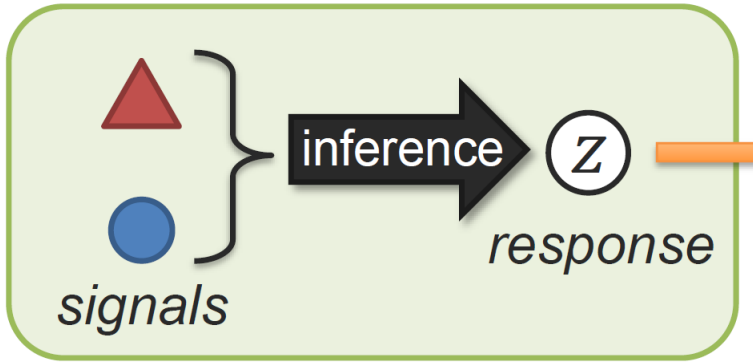
inference ***Yes!***



Interconnected Modalities

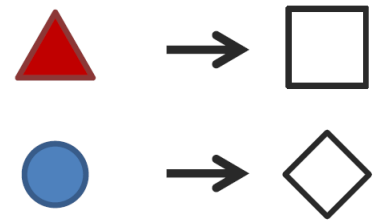
② Modality interactions

Modality elements often interact during inference



Types of interaction responses?
(a taxonomy)

Unimodal Non-redundancy



Is this a living room?

inference → *Yes!*

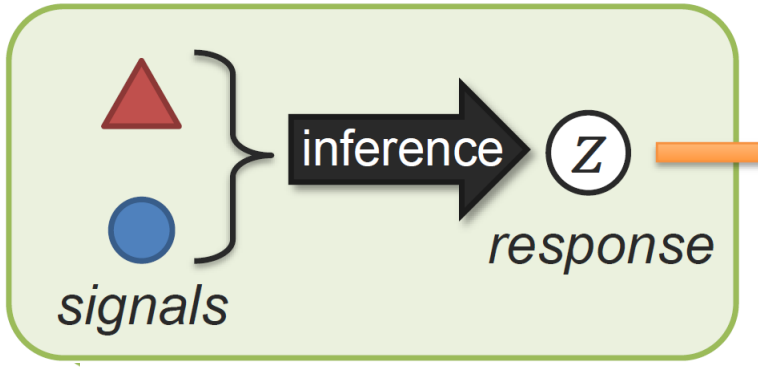
A teacup on the right of a laptop in a clean room.

inference → *No, probably study room.*

Interconnected Modalities

② Modality interactions

Modality elements often interact during inference



Types of interaction responses?
(a taxonomy)

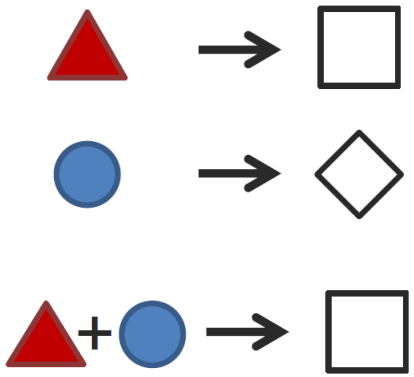
Is this a living room?

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

inference

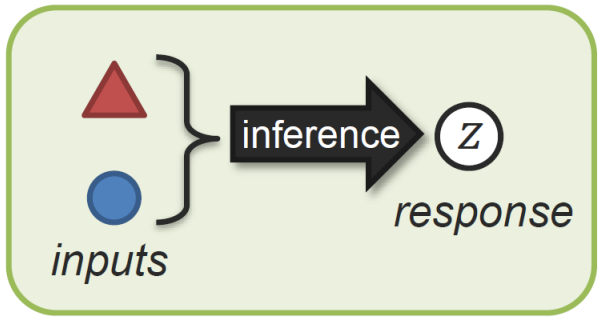
Yes!

Unimodal Non-redundancy



Multimodal dominance

Taxonomy of Interaction Responses: A Behavioral Science View



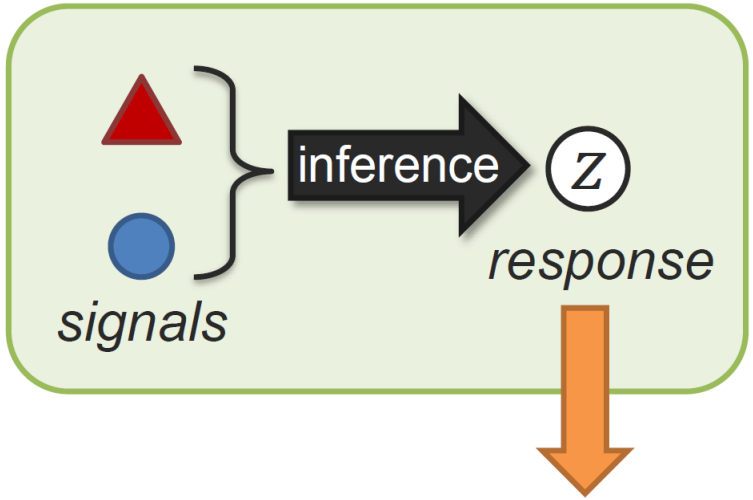
Multimodal Communication

	signal	response	signal	response	
Redundancy	a →	□	a+b →	□	Equivalence
	b →	□	a+b →	▭	Enhancement
Nonredundancy	a →	□	a+b →	□ and ○	Independence
	b →	○	a+b →	□	Dominance
			a+b →	▭ (or □)	Modulation
			a+b →	△	Emergence

Partan and Marler (2005). *Issues in the classification of multimodal communication signals*. American Naturalist, 166(2)

Dimensions of Modality Interactions

What are the dimensions for **digitally-represented** modalities?

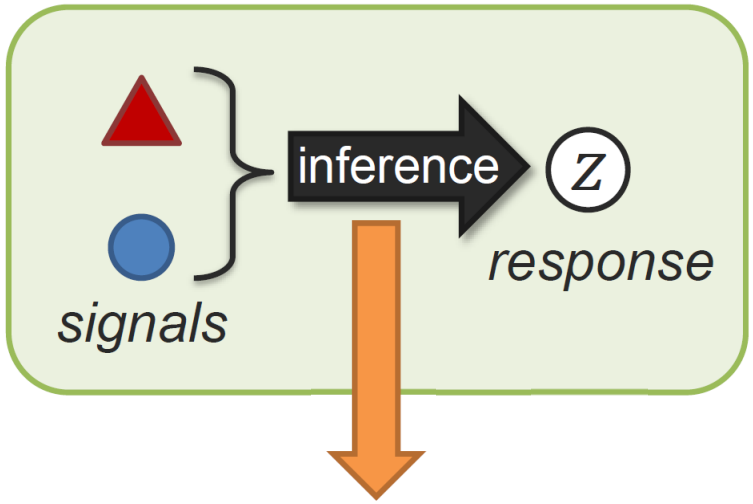


1 Interaction Responses:

- Redundancy
- Non-redundancy
- Dominance
- Emergence…

Dimensions of Modality Interactions

What are the dimensions for **digitally-represented** modalities?

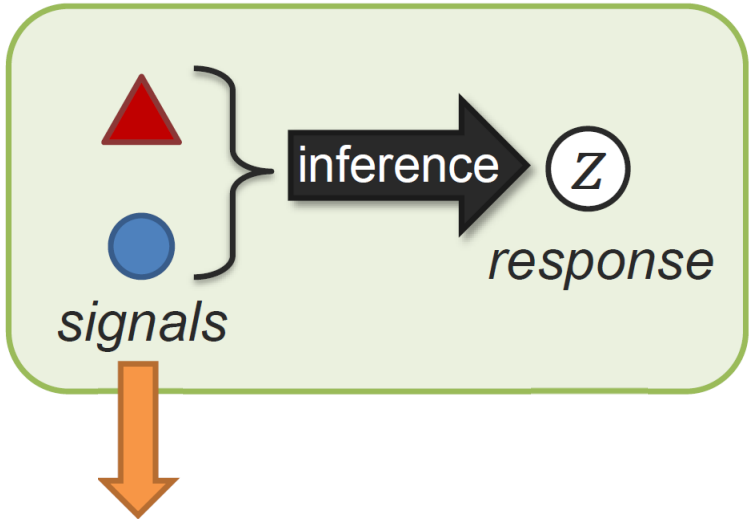


2 Interaction Mechanics:

- Additive
- multiplicative
- Nonlinear
- Causal,
- Logical, …

Dimensions of Modality Interactions

What are the dimensions for **digitally-represented** modalities?

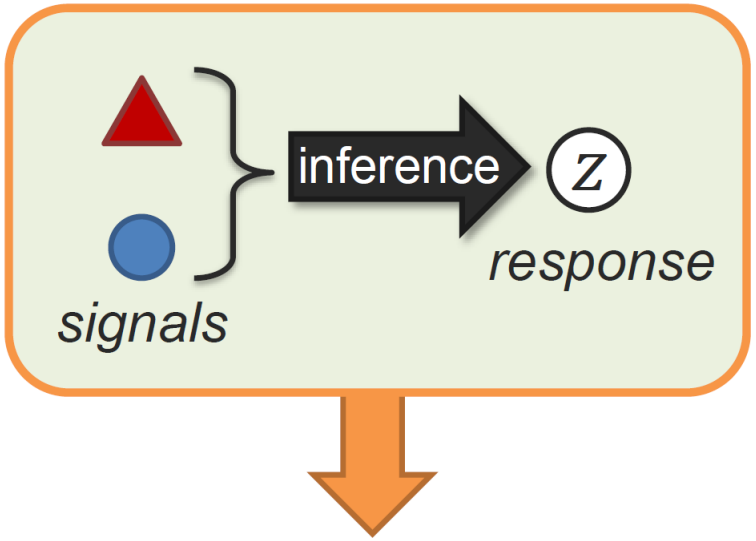


3 Input modalities:

- Unimodal
- Bimodal
- Trimodal
- High-modal, ...

Dimensions of Modality Interactions

What are the dimensions for **digitally-represented** modalities?



4 Context:

- Structure context
- Task relevance
- Context dependence
- High-modal, ...

What is Multimodal?

***Multimodal* is the science of**

heterogeneous and interconnected data ☺

What is Multimodal Machine Learning?

Multimodal Machine Learning (ML) is the study of computer algorithms that learn and improve through the use and experience of data from multiple modalities

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

Multimodal AI is a superset of **Multimodal ML**

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 …)
  - Main focus of this tutorial: last 5 years

Behavioral Study of Multimodal

Language
and gestures

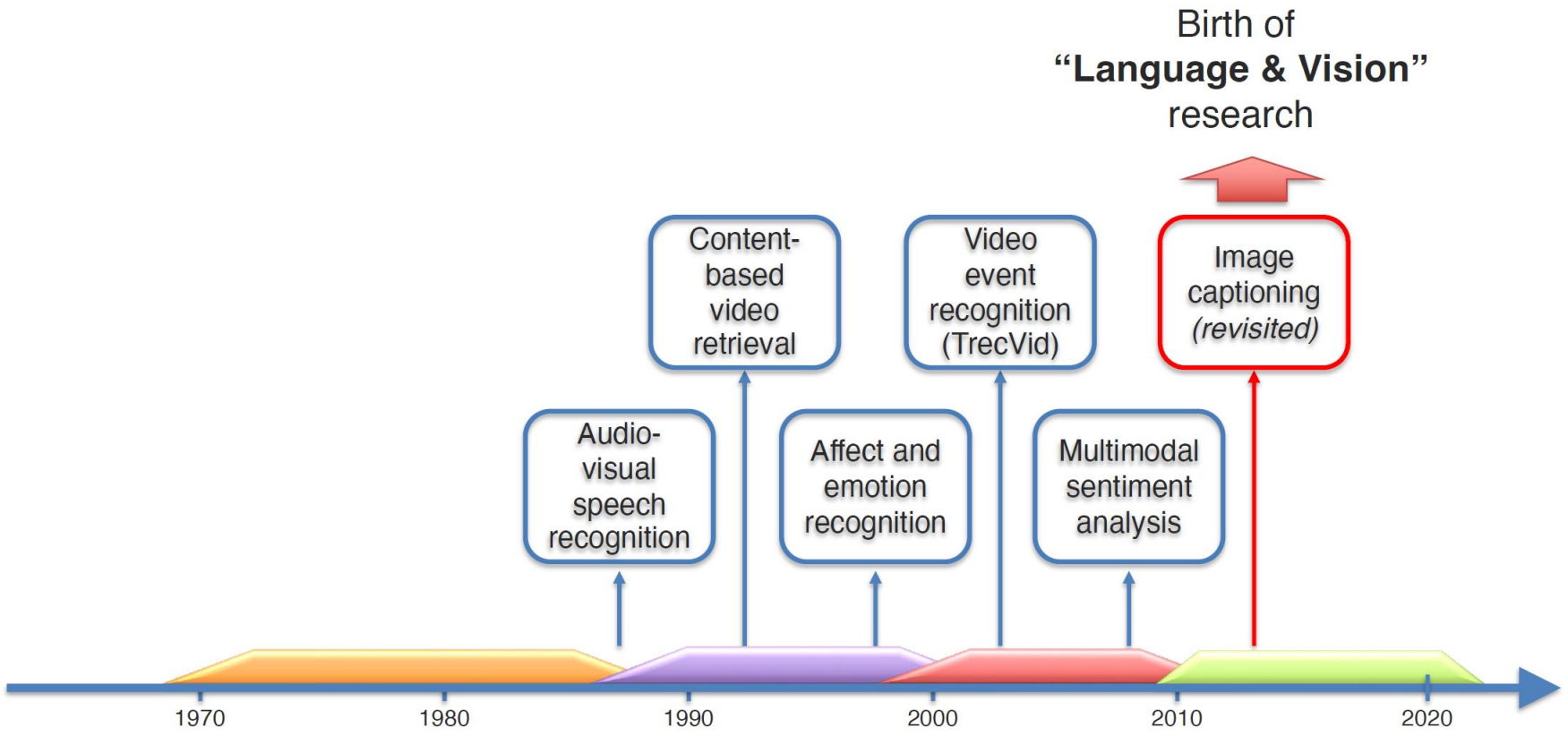
David McNeill

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

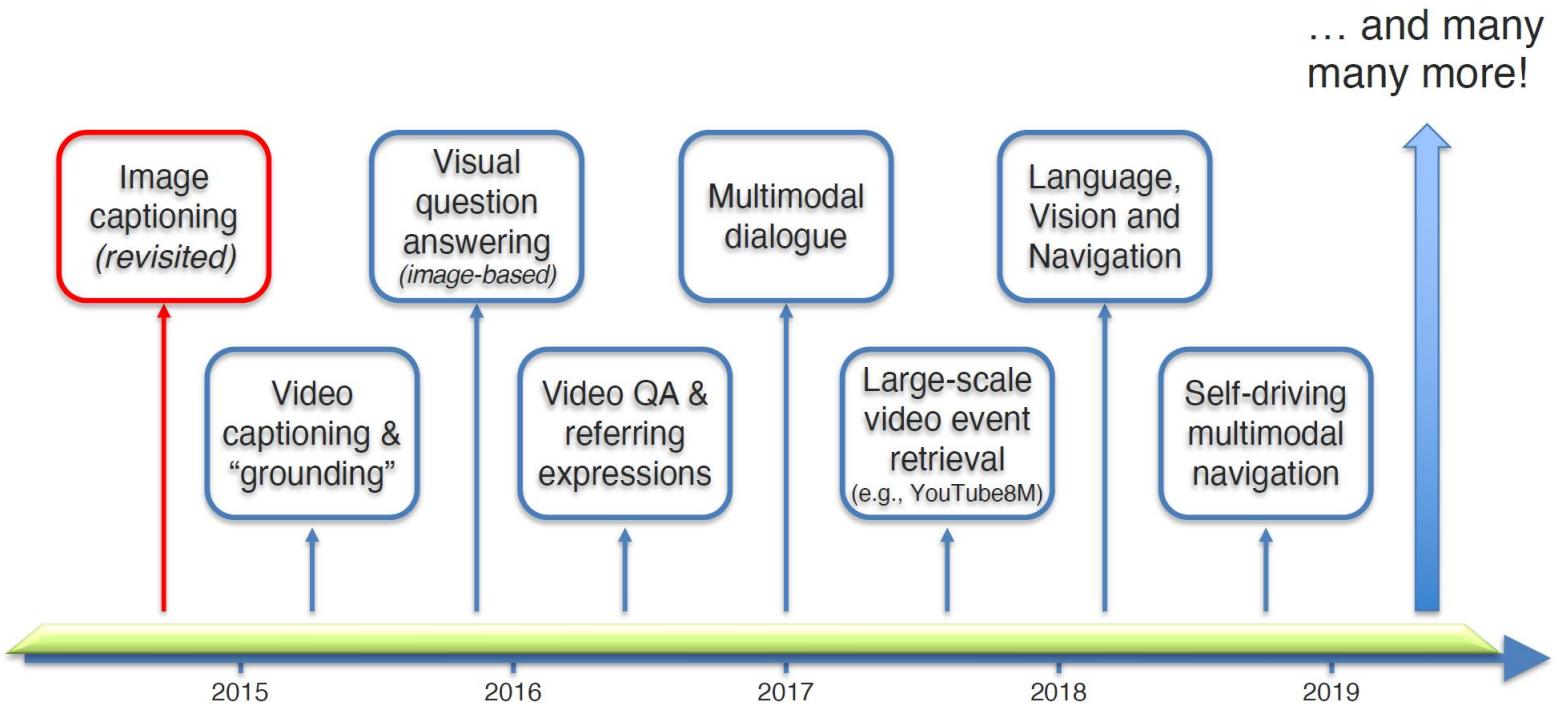
McGurk effect

Behavioral Study of Multimodal

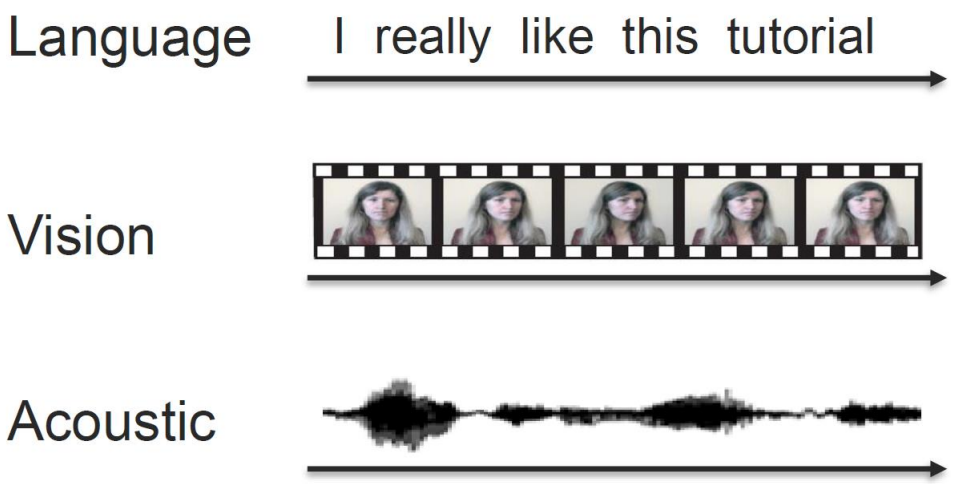
Multimodal Research Tasks



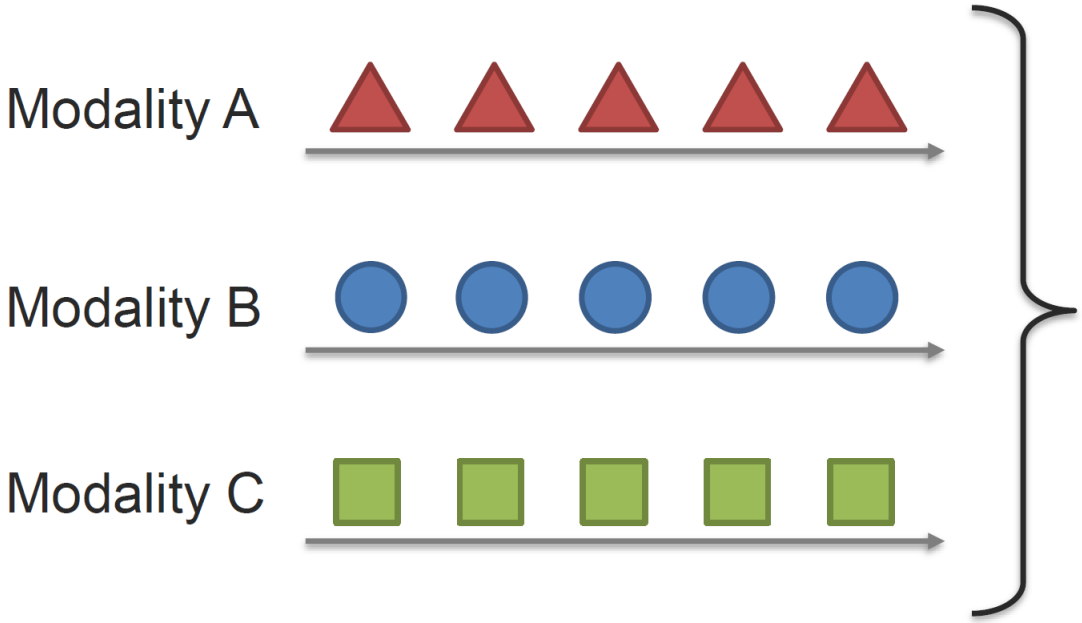
Multimodal Research Tasks



Multimodal Machine Learning



Multimodal Machine Learning



Multimodal ML

- Unsupervised,
- Self-supervised,
- Supervised,
- Reinforcement,

…

- Representation
- Alignment
- Reasoning
- Generation
- Transference

内容提纲

① 为什么要学习《多模态机器学习》

② 什么是“多模态机器学习”

③ 本课程将要学习的内容

④ 本章小结

课程提纲

单模态表示

- 视觉模态
- 文本模态
- 三维点云
- 声音模态

基本概念

- 神经网络及其优化

经典多模态机器学习

- 多模态表示
- 多模态对齐
- 多模态推理
- 多模态生成
- 多模态迁移

通用多模态机器学习

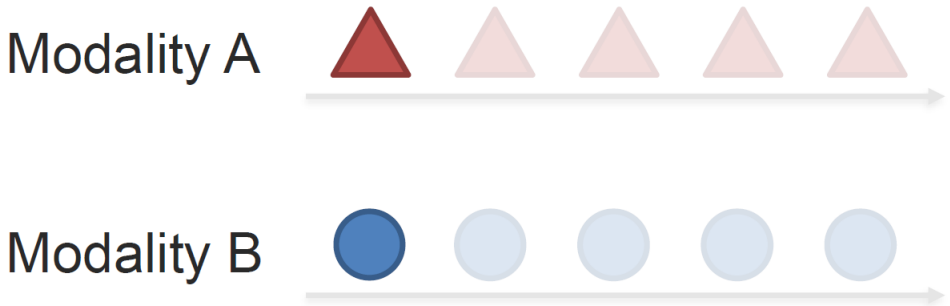
- 通用多模态（大）模型
- 多模态预训练
- 多模态典型应用

Task 1: Representation (表示)

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:



It can be seen as a "local" representation

or

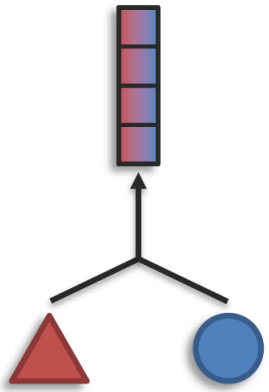
representation using holistic features

Task 1: Representation (表示)

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

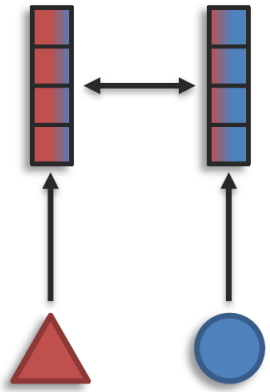
Sub-challenges:

Fusion



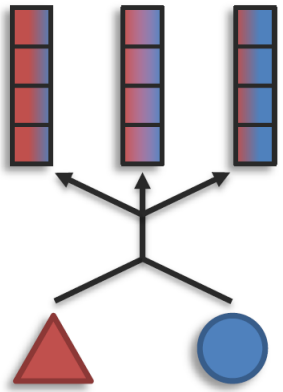
modalities **>** # representations

Coordination



modalities **=** # representations

Fission



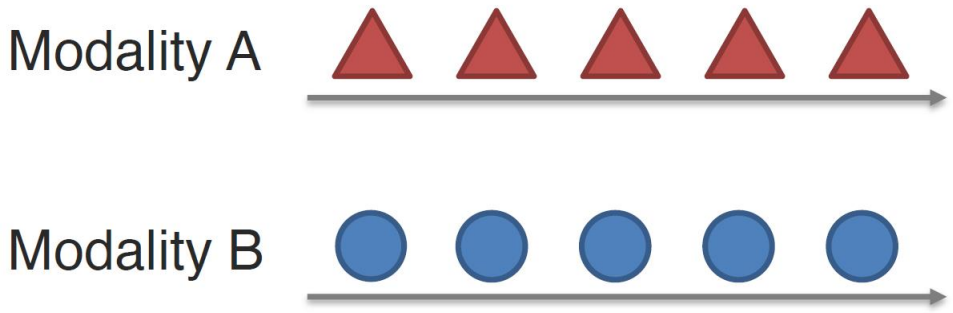
modalities **<** # representations

Task 2: Alignment (对齐)

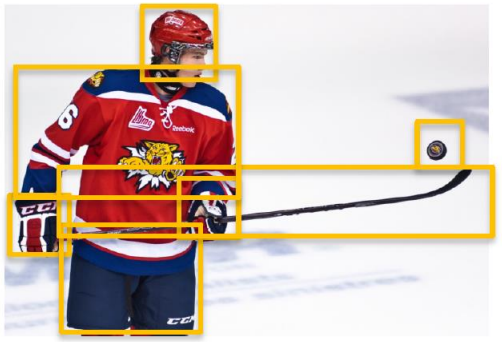
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

➡ Most modalities have internal structure with multiple elements

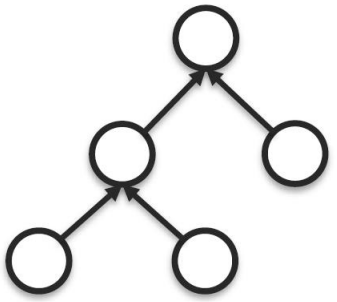
Elements with temporal structure:



Other structured examples:



Spatial



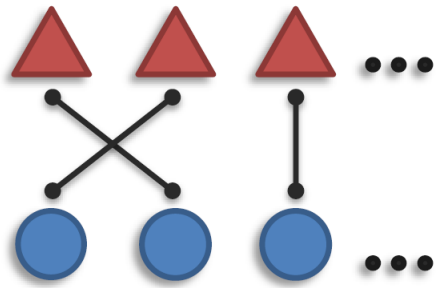
Hierarchical

Task 2: Alignment (对齐)

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

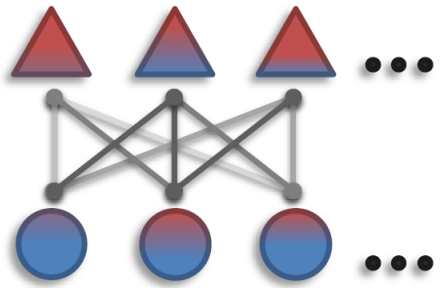
Sub-challenges:

Connections



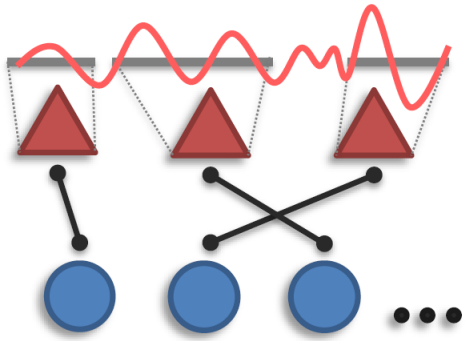
Explicit alignment
(e.g., grounding)

Aligned Representation



Alignment + representation
(aka, contextualized representation)

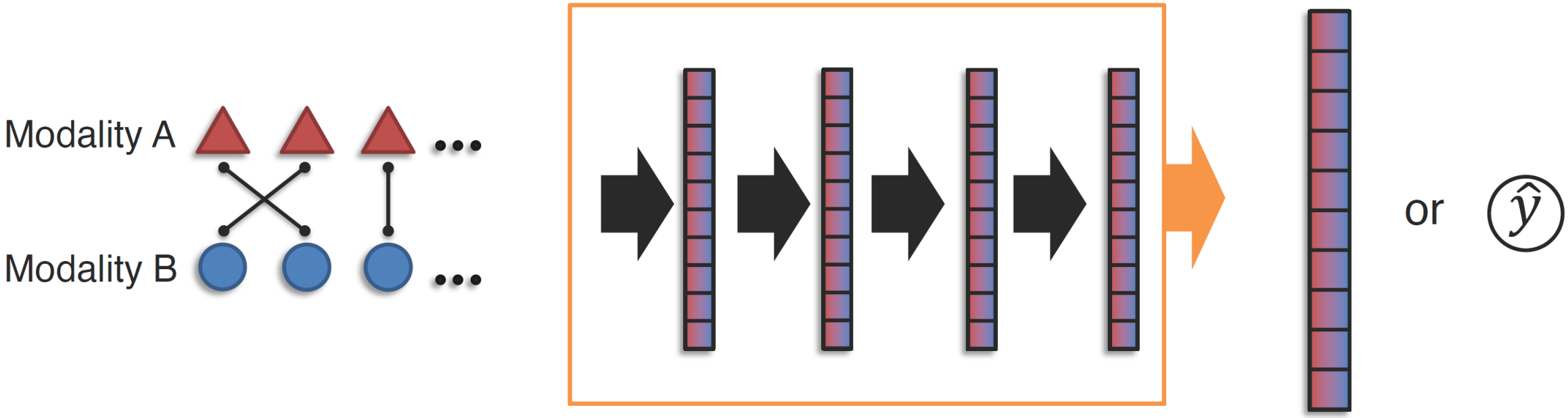
Elements



Segmentation of individual elements

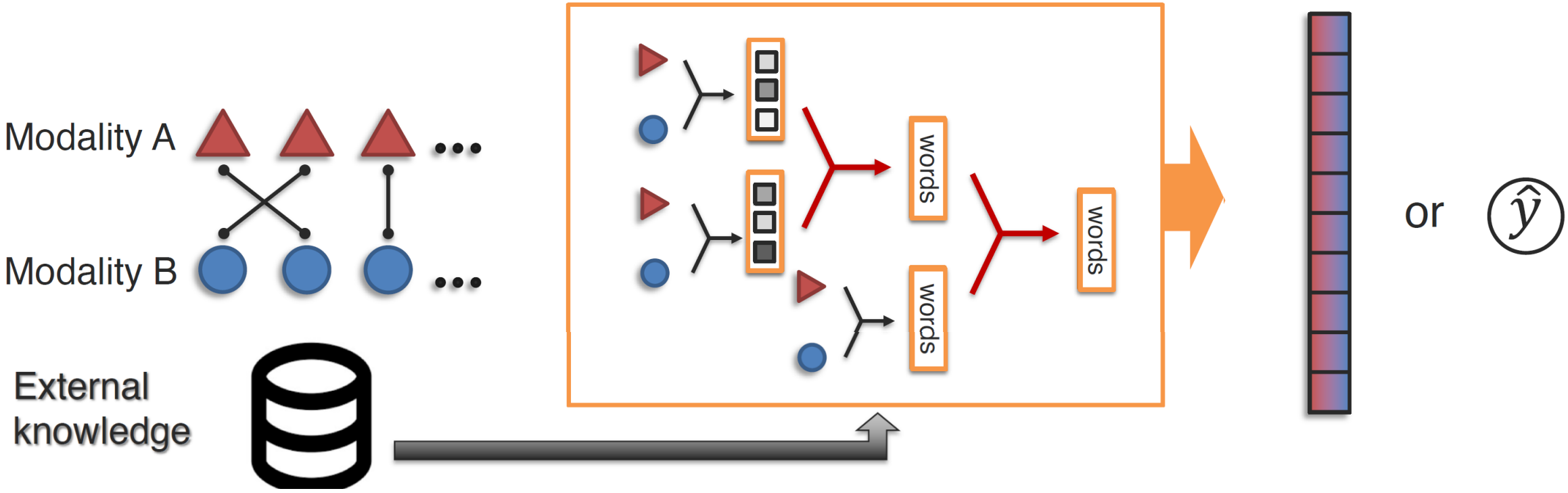
Task 3: Reasoning (推理)

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



Task 3: Reasoning (推理)

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

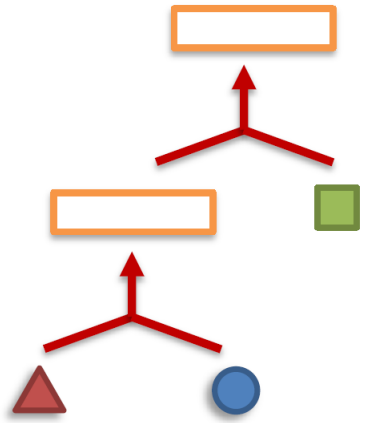


Task 3: Reasoning (推理)

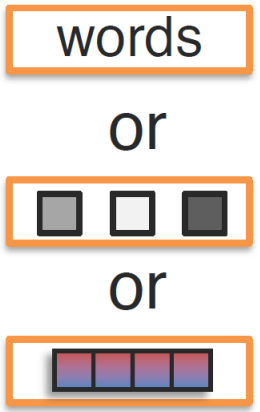
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

Sub-challenges:

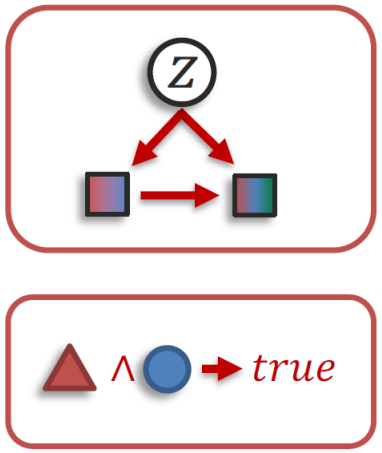
Structure Modeling



Intermediate concepts



Inference Paradigm

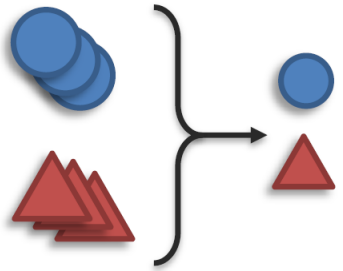
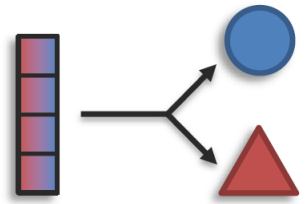


External Knowledge

Task 4: Generation (生成)

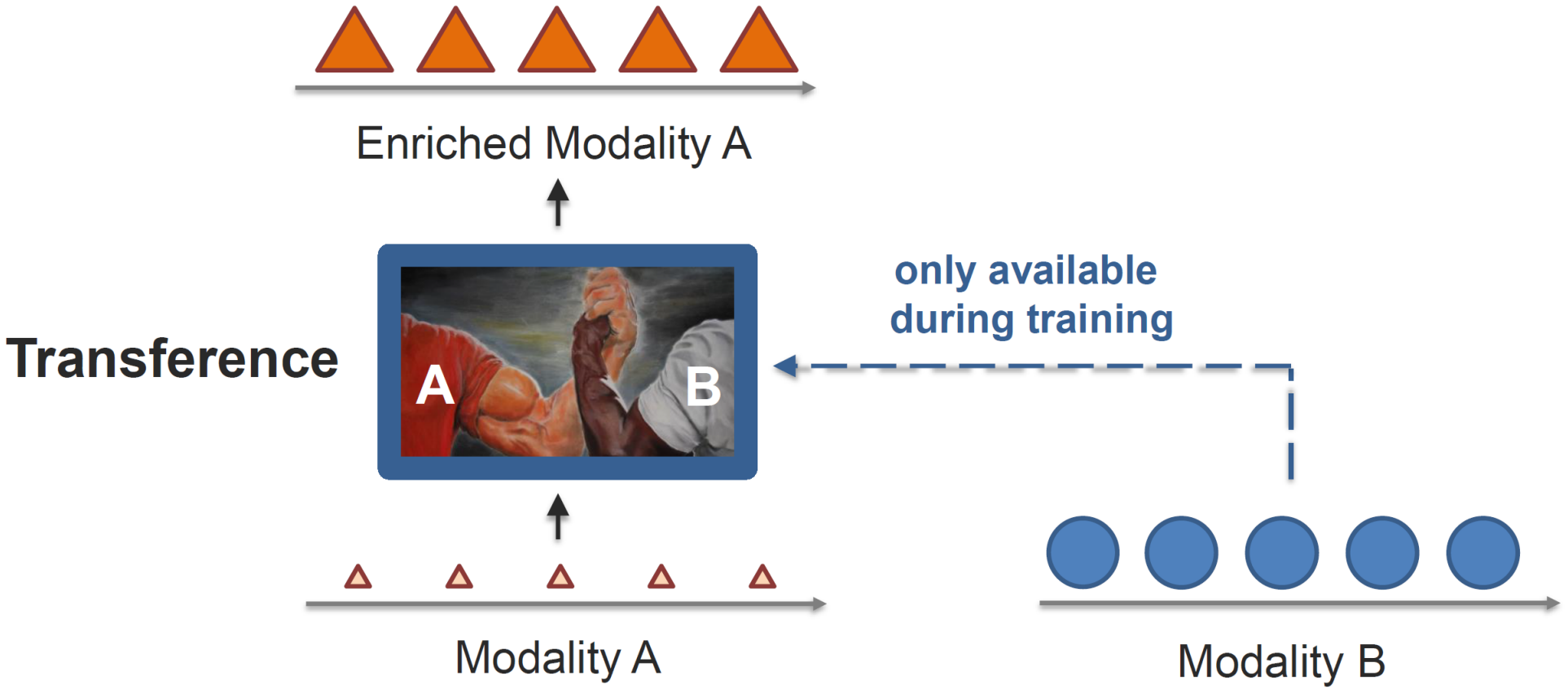
Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence

Sub-challenges:

	Summarization	Translation	Creation
			
Information: (content)	Reduction	Maintenance	Expansion
			

Task 5: Transference (迁移)

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

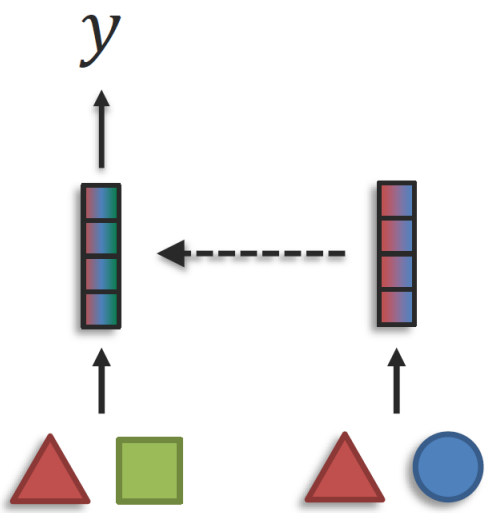


Task 5: Transference (迁移)

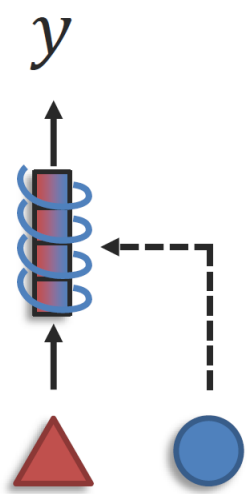
Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources

Sub-challenges:

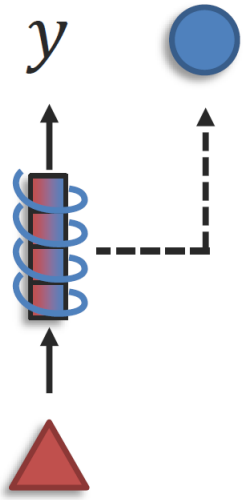
Transfer



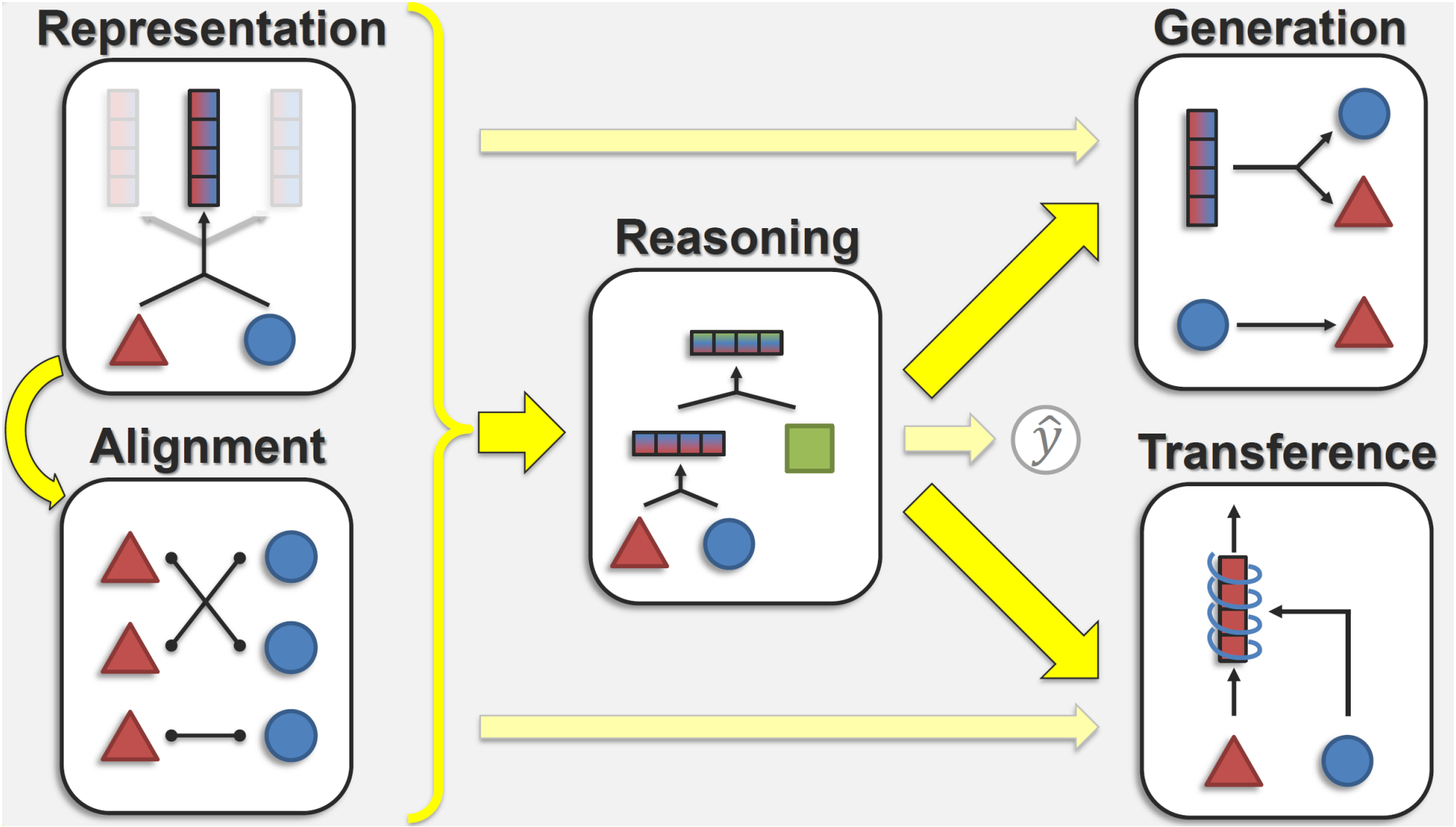
Co-learning via representation



Co-learning via generation



Multimodal Machine Learning Tasks



内容提纲

① 为什么要学习《多模态机器学习》

② 什么是“多模态机器学习”

③ 本课程将要学习的内容

④ 本章小结

总结

- 了解Multimodal Machine Learning 的意义与整体学习内容
- 掌握Modality, Multimodal的定义
- 掌握 Dimensions of Heterogeneity and Modality Interactions
- 掌握Multimodal Machine Learning Tasks

考核方式

<table>
<tr><td colspan="2">平时考核
（50%）</td><td rowspan="2">期末分组项目展示
（50%）</td></tr>
<tr><td>课程作业
（70%）</td><td>研讨交流
（30%）</td></tr>
</table>

- **所占学分：**2学分
- **课程作业：**习题、编程等，每两个星期一次
- **研讨交流：**随机抽查、课堂积极问答、随堂小测试等
- **编程语言：**Python
- **助教：**张岳霖

思考

Which word can describe this figure in the language of modality interaction?

(a) Equivalence.

(b) Dominance.

(c) Emergence.

(d) Independence.