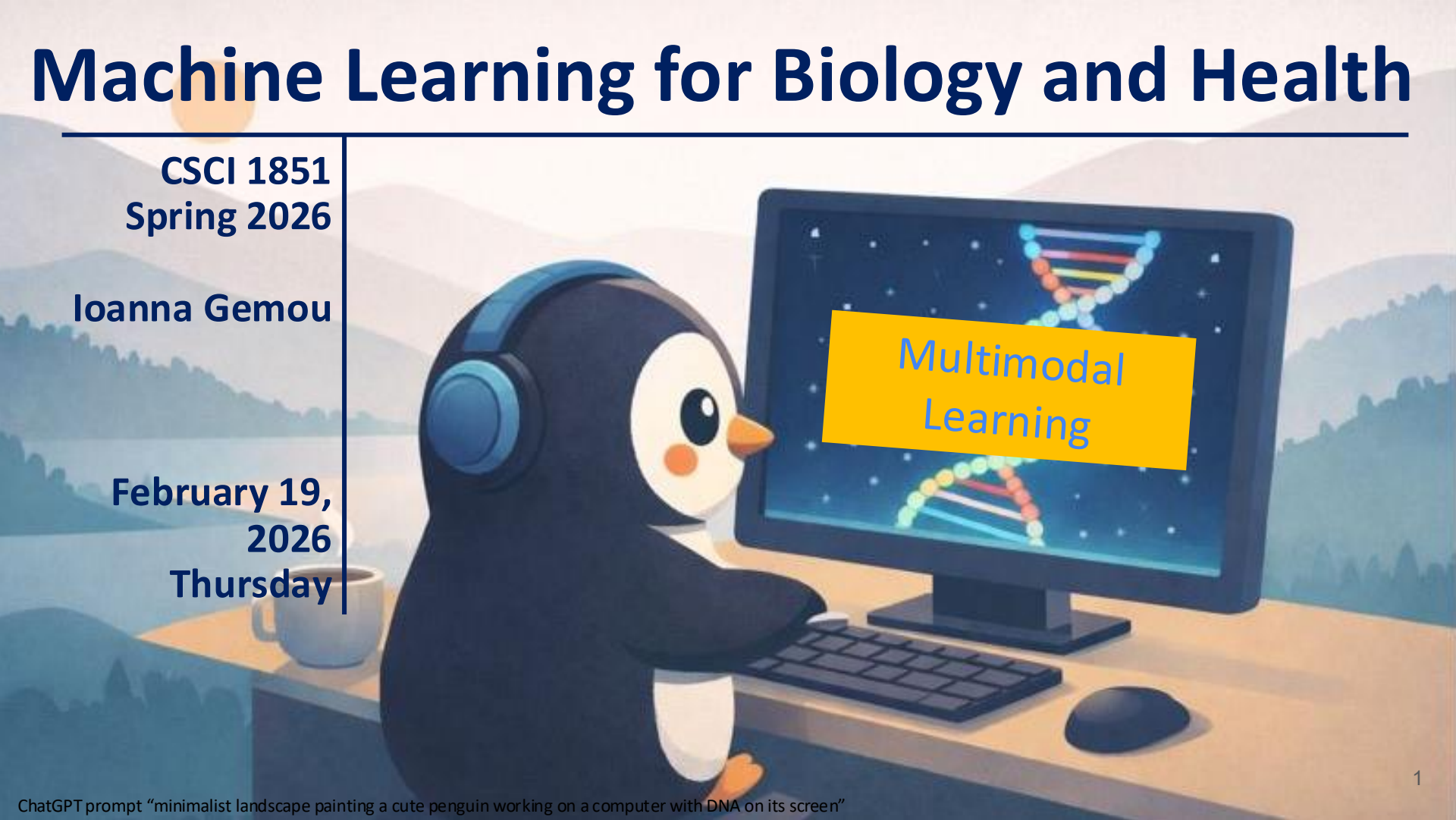
Machine Learning for Biology and Health

CSCI 1851
Spring 2026

Ioanna Gemou

February 19, 2026
Thursday



ChatGPT prompt “minimalist landscape painting a cute penguin working on a computer with DNA on its screen”

About me

PhD at Brown working with Ritambhara!

Research Interests: Deep learning, multimodal learning, interpretable ML

Email: ioanna_gemou@brown.edu
Website: www.igemou.github.io

Office: CIT 423

Right now I am working on…

Counterfactual Explanations: minimal change to input that flips prediction!

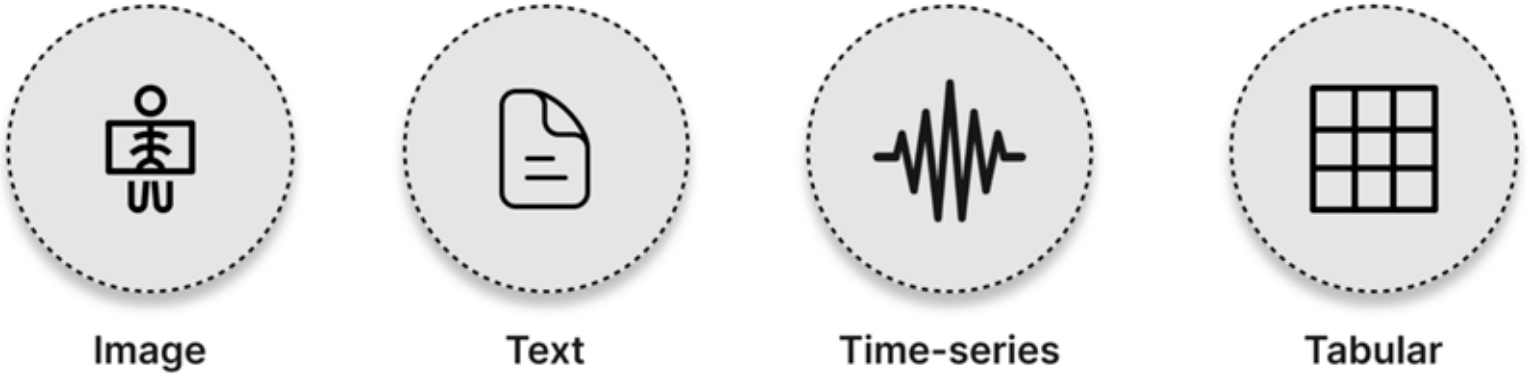
Today’s agenda

- What is multimodal learning?
- Why do we need it?
  - Both a human and machine perspective
- How can we use it?
- Main steps for multimodal learning
- In class activity
- Established multimodal models: CLIP
- Final Project

What is multimodal learning?

Learning from more than 1 data types

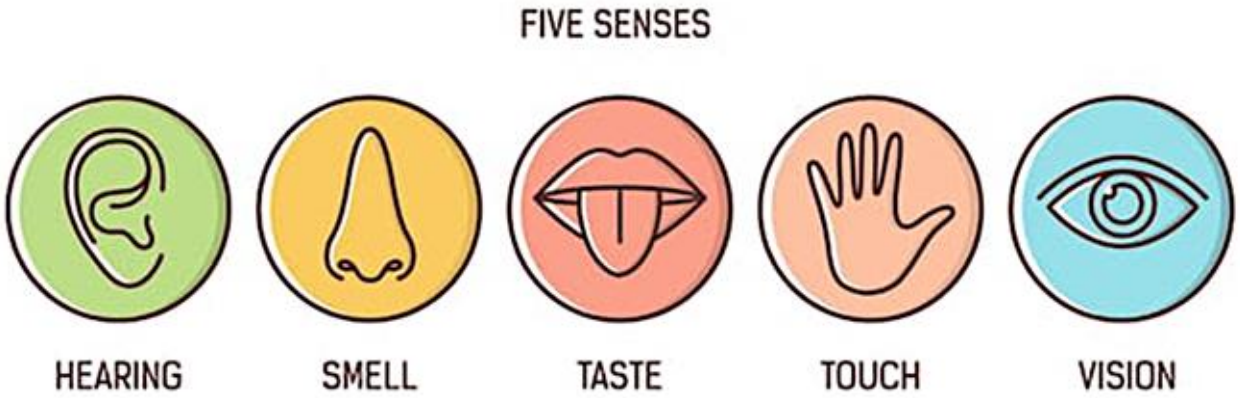
For example:



\+ more!

Why do we need it?

We, as humans, perceive the world with **multiple sensory systems**—vision, audition, touch, smell…



https://oodo-optical.com/the-five-senses/

Why do we need it?

1. **Degeneracy in neural structure**

- A system functions even with the loss of one component
- Eg. Spatial properties are developed even in the blind, using touch etc.

2. **Different subsystems can educate each other**

- Learn to associate multiple representations
- Children spend hours touching and feeling objects

From a machine learning perspective

1. One modality rarely contains enough information: a single input is often an incomplete signal.

A chest X-ray alone may not distinguish pneumonia from pulmonary edema. Adding vitals or a short clinical note may resolve the ambiguity.

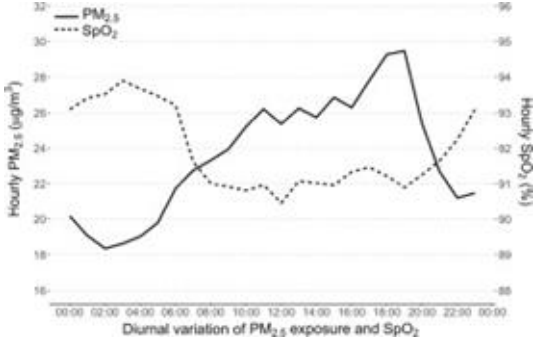
From a machine learning perspective

2. Different modalities capture different kinds of structure: each modality has its own inductive bias.

Images capture spatial patterns

Fever *to 39°C, productive cough,* ***hypoxia*** *on room air.*

Text captures abstract concepts

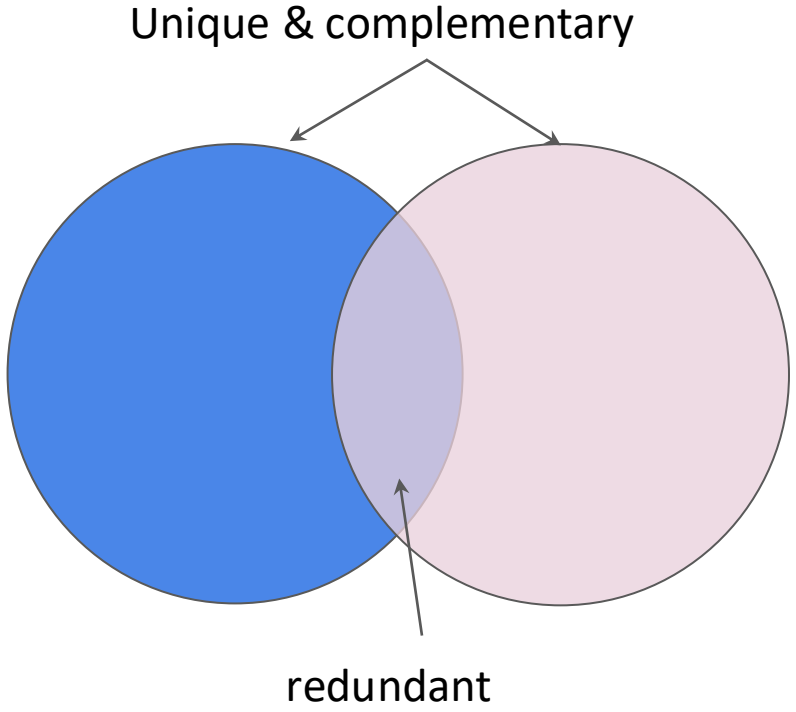


Time series capture trends

Xi Xia et. al., 2020

How can we use Multimodal Learning?

1. Cross-modal supervision: Use one modality to help learn in another (exploits “redundancy”)
2. Fusion: Combine multiple modalities that adding new knowledge (exploits “complementarity”)



What is multimodal learning?

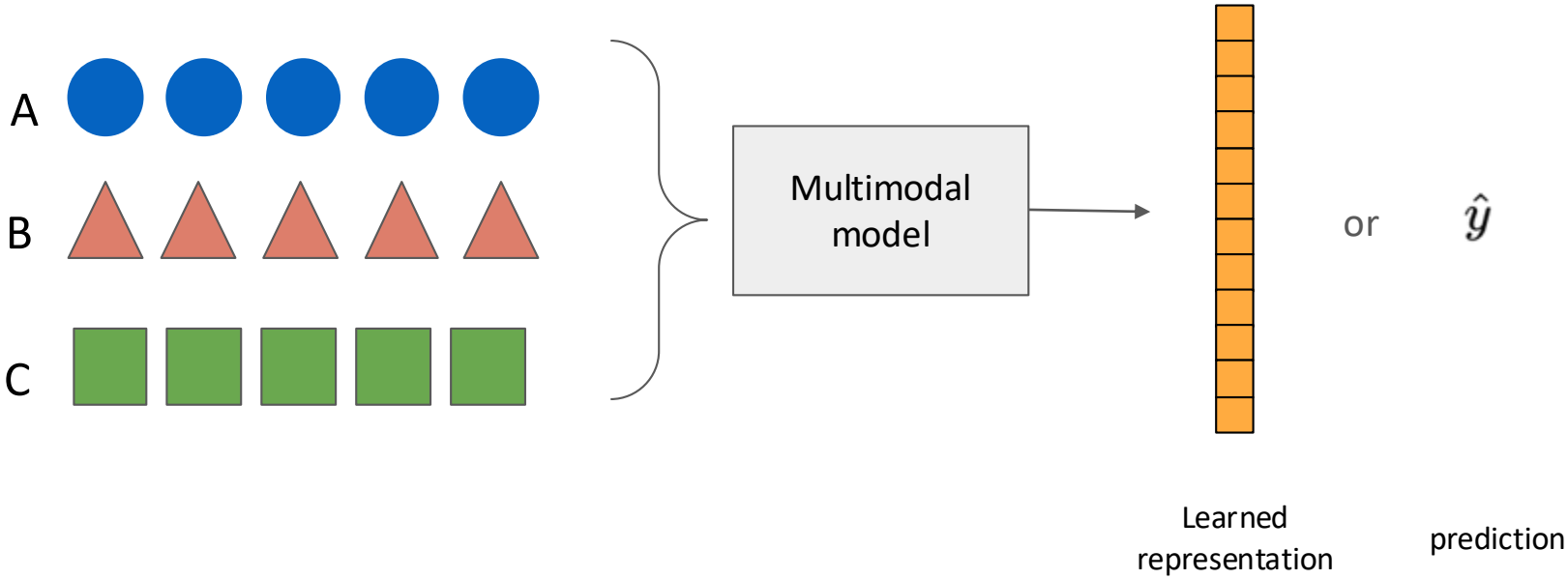
Multimodal learning aims to train a model to process multiple modalities

The goal is to:

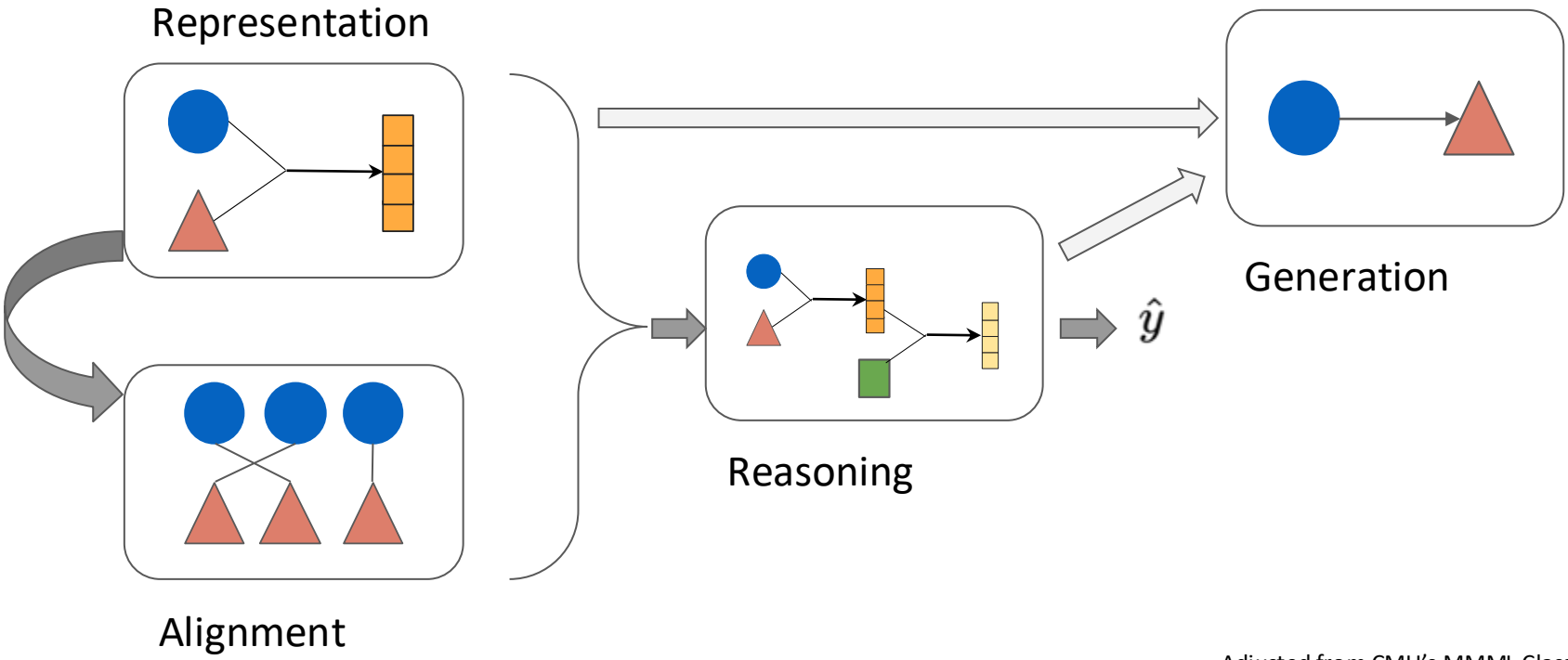
- Understand multiple modalities jointly
- Reason about one modality using other modalities as reference
- Retrieve one modality given other modalities as reference
- Generate one modality given other modalities as reference
- ……

Most of the examples we will see focus on vision and language

Multimodal learning



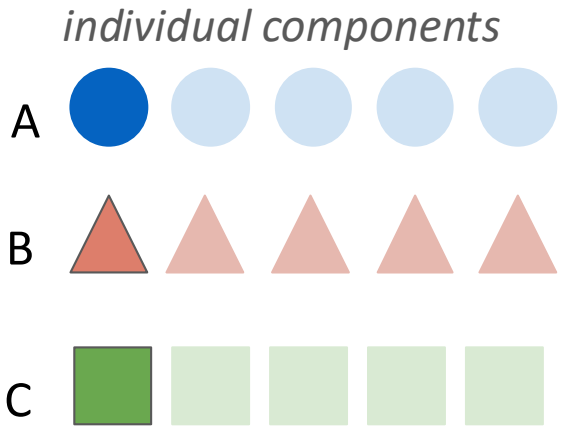
Main steps for multimodal learning



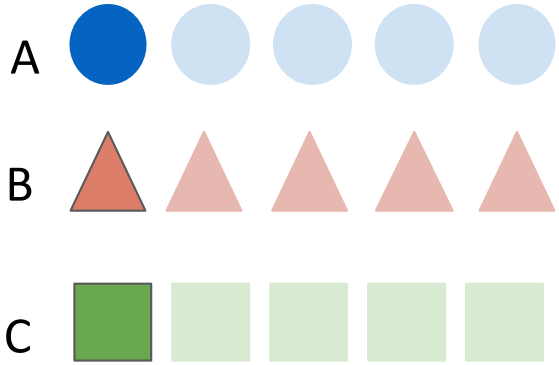
Adjusted from CMU's MMML Class

Representation

Learning representations that reflect **cross-modal** interactions between individual components, across different modalities



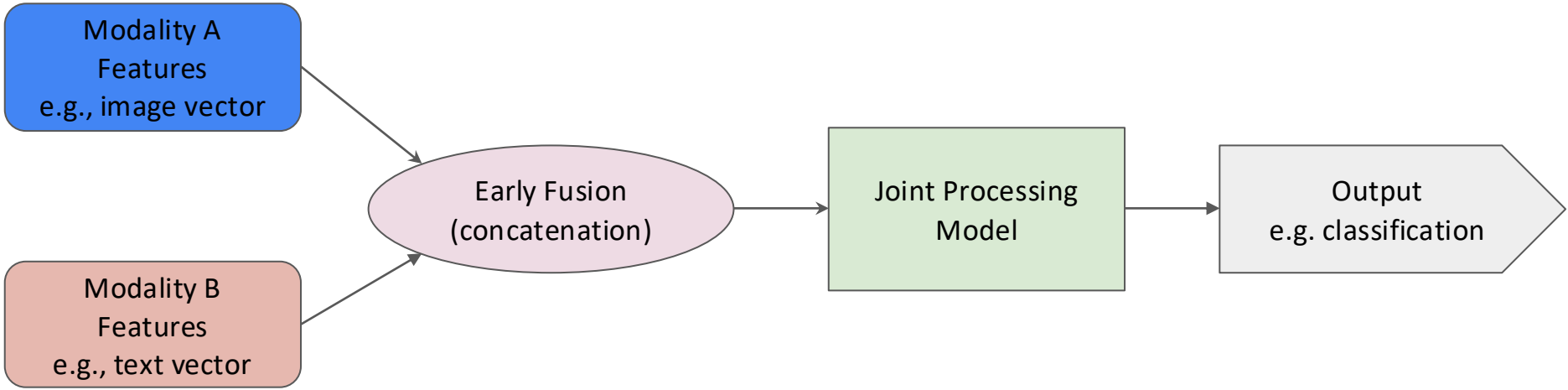
We have multiple data inputs…now what?



When and how do we combine these representations?

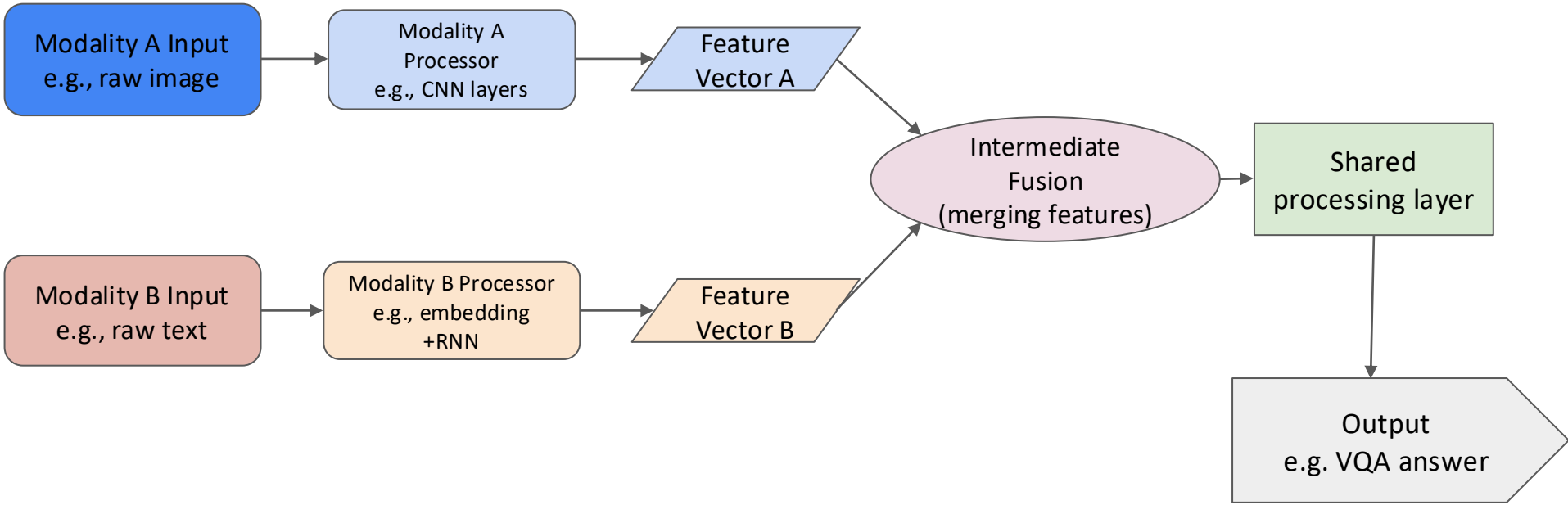
This is a fundamental building block in most multimodal modeling tasks.

Early Fusion of Representations

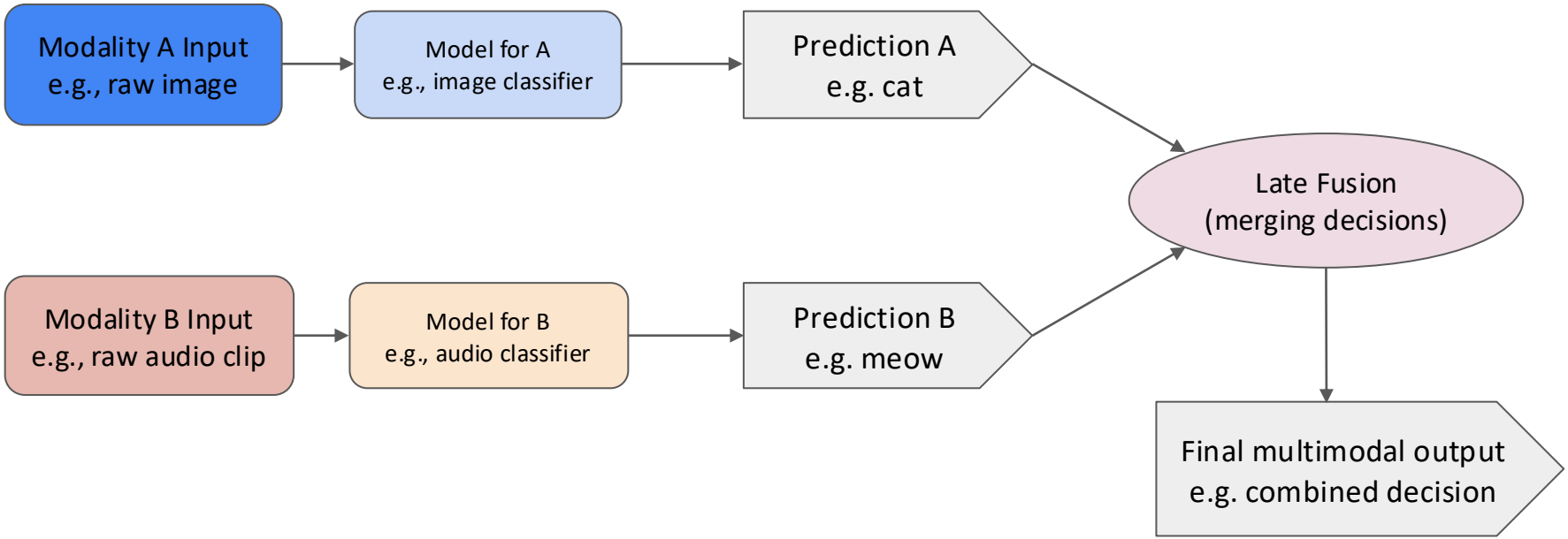


https://apxml.com/courses/intro-to-multimodal-ai/chapter-3-techniques-integrating-modalities/approaches-multimodal-fusion

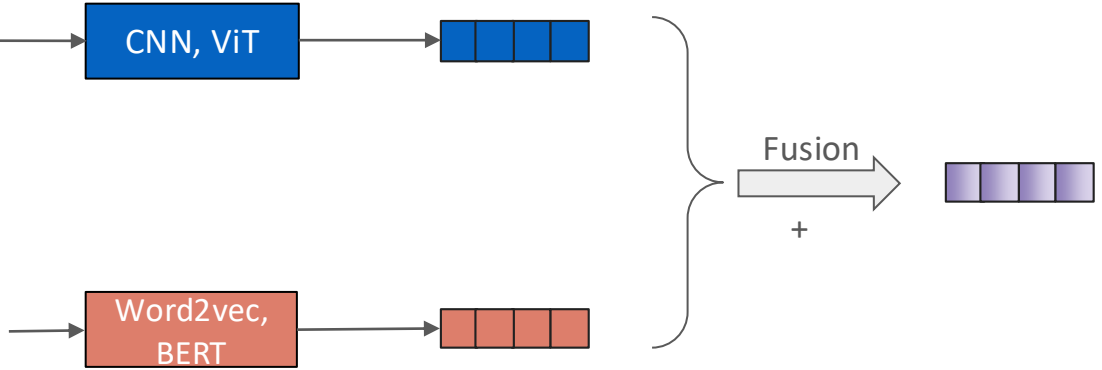
Intermediate Fusion of Representations



Late Fusion of Representations



Representation Fusion

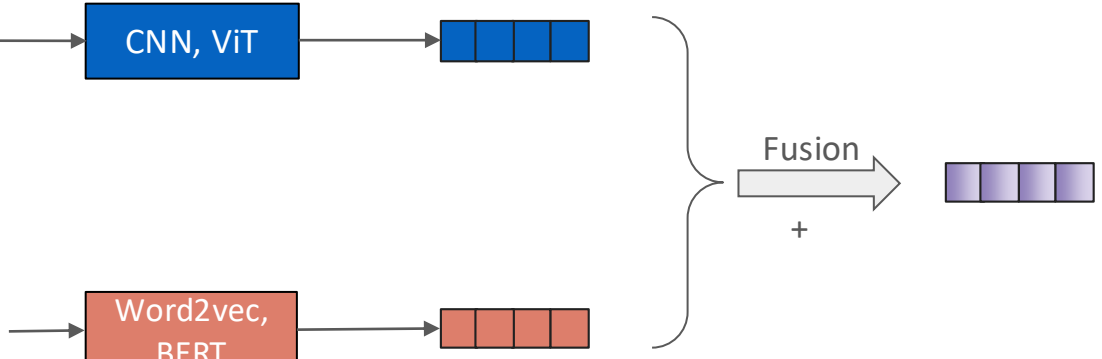


The trachea is central and no foreign body can be seen, the carina is visible. Lung markings can be seen in both the left and right sides, the lungs appear clear. The heart appears enlarged, occupying over 50% internal thoracic diameter suggesting cardiomegaly.

What kind of fusion is happening here?

Representation Fusion

The trachea is central and no foreign body can be seen, the carina is visible. Lung markings can be seen in both the left and right sides, the lungs appear clear. The heart appears enlarged, occupying over 50% internal thoracic diameter suggesting cardiomegaly.



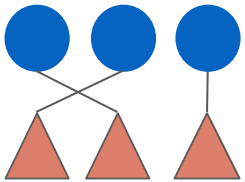
Intermediate fusion

Most multimodal systems use separate unimodal encoders, then combine representations

Alignment

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

Discrete
Alignment



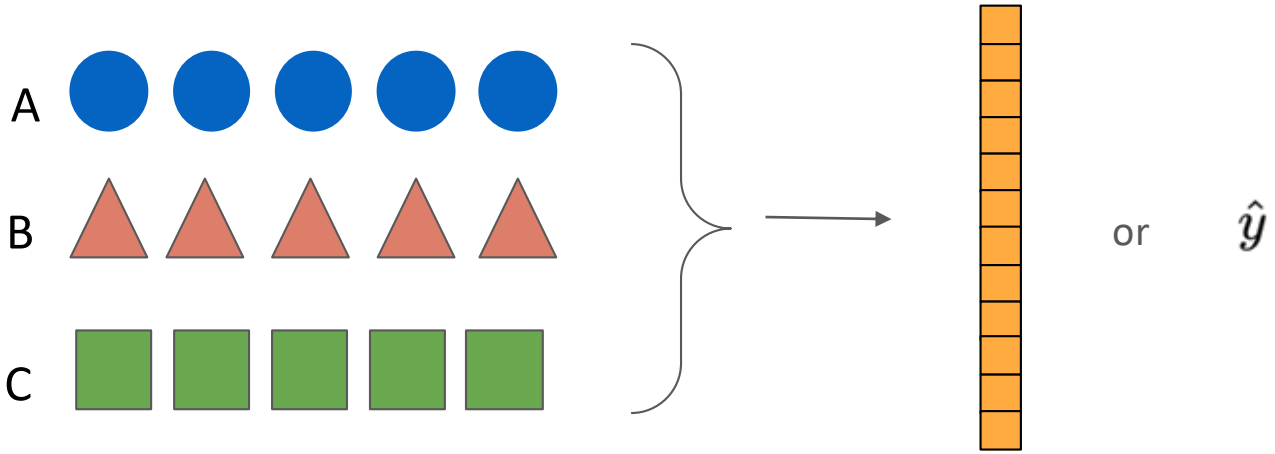
Discrete elements
and connections

e.g. words ↔ image regions
“dog” aligns with pixels of the dog
“pneumonia” aligns with opaque pixels

\+ Other types of alignment, such as continuous and contextualized

Reasoning

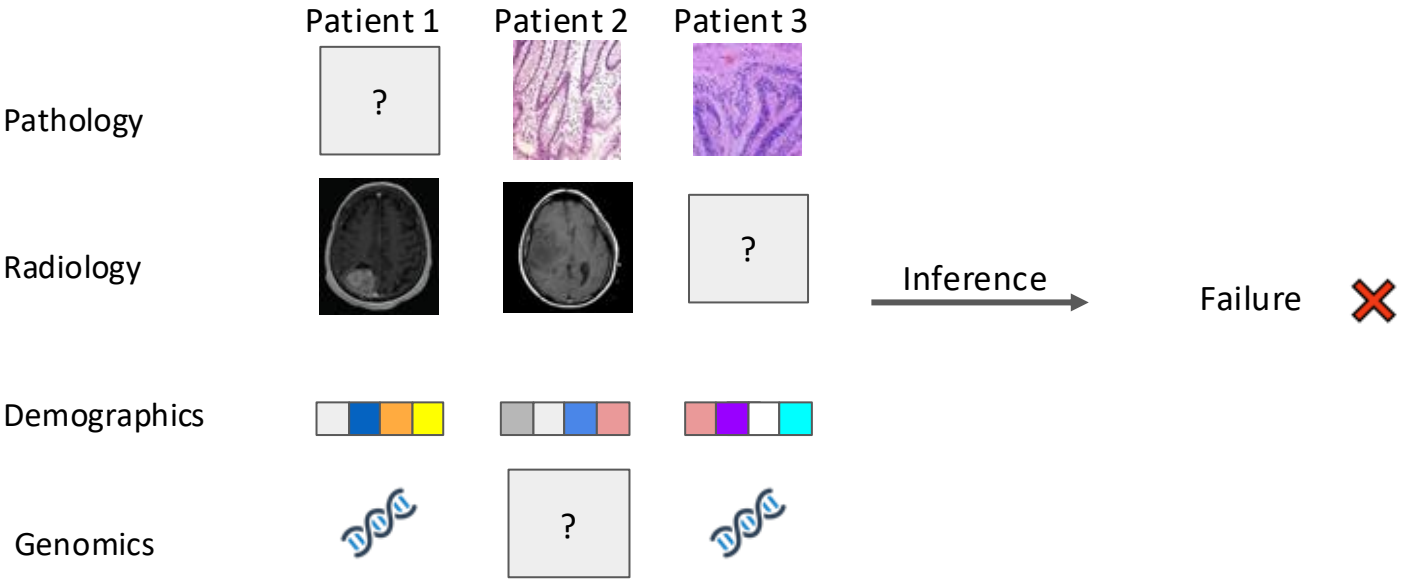
Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure



Challenges and failure modes of Multimodal Models

(1) Missingness: Models trained on complete data fail when a modality is absent at inference.

Task: cancer diagnosis and prognosis



Challenges and failure modes of Multimodal Models

(2) Imbalance: One modality dominates learning.

Text a lot of times dominates over images

\+ Doctor’s assessment: community-acquired pneumonia. Started ceftriaxone and azithromycin

Challenges and failure modes of Multimodal Models

(2) Imbalance: One modality dominates learning.

Text a lot of times dominates over images

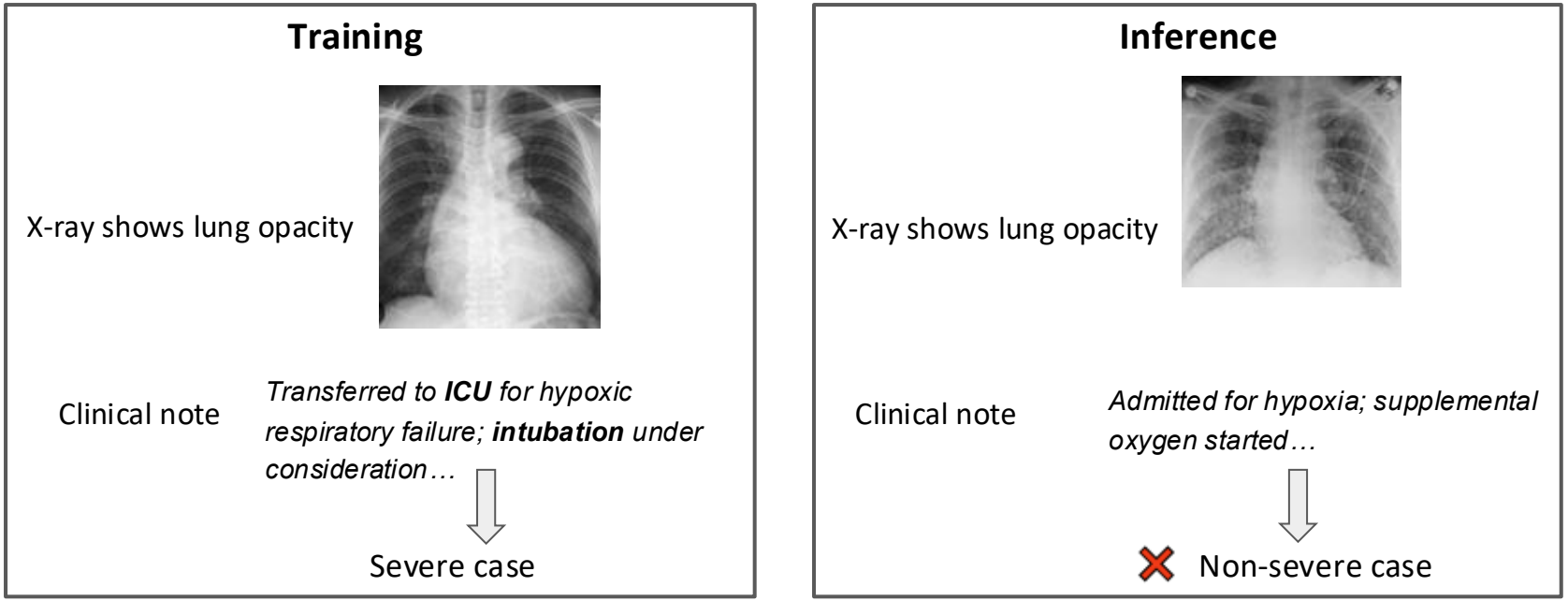
\+

Doctor's assessment: community-acquired pneumonia. Started ceftriaxone and azithromycin

The word *"pneumonia"* and the antibiotic names are perfectly predictive, so **the model does not need to look at the X-ray**.

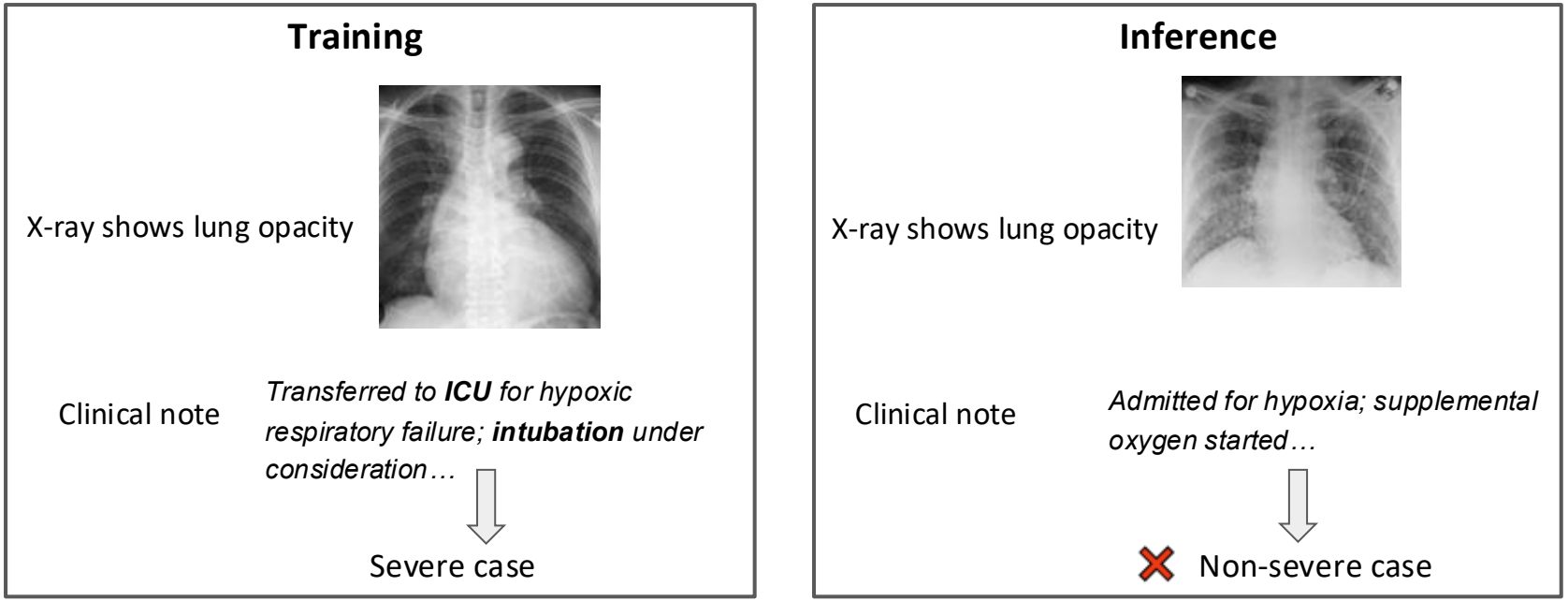
Challenges and failure modes of Multimodal Models

(3) Cross-modal correlations: Models exploit dataset shortcuts instead of learning relationships.



Challenges and failure modes of Multimodal Models

(3) Cross-modal correlations: Models exploit dataset shortcuts instead of learning relationships.



Over many samples, the model discovers an **easy shortcut**: Text mentions “ICU” ⟶ predict “severe”

In class activity!

Different goals for multimodal models

- Some multimodal models are trained to **solve a task**

  Example: Visual Question Answering (VQA) model

- Some are trained to **learn aligned representations** that can then be reused across many tasks

  Example: CLIP

Let's see how CLIP works!

CLIP - Contrastive Language-Image Pre-training

- CLIP is a multimodal (language-image) model
- Uses contrastive learning
- CLIP is a zero-shot classifier
- In 2021, CLIP beat unsupervised and supervised baselines on many datasets
- Leverages a huge amount of paired data (“web-scale”)
- While contrastive learning was not new at the time, it was never done at this multimodal scale

OpenAI

CLIP - Road Map

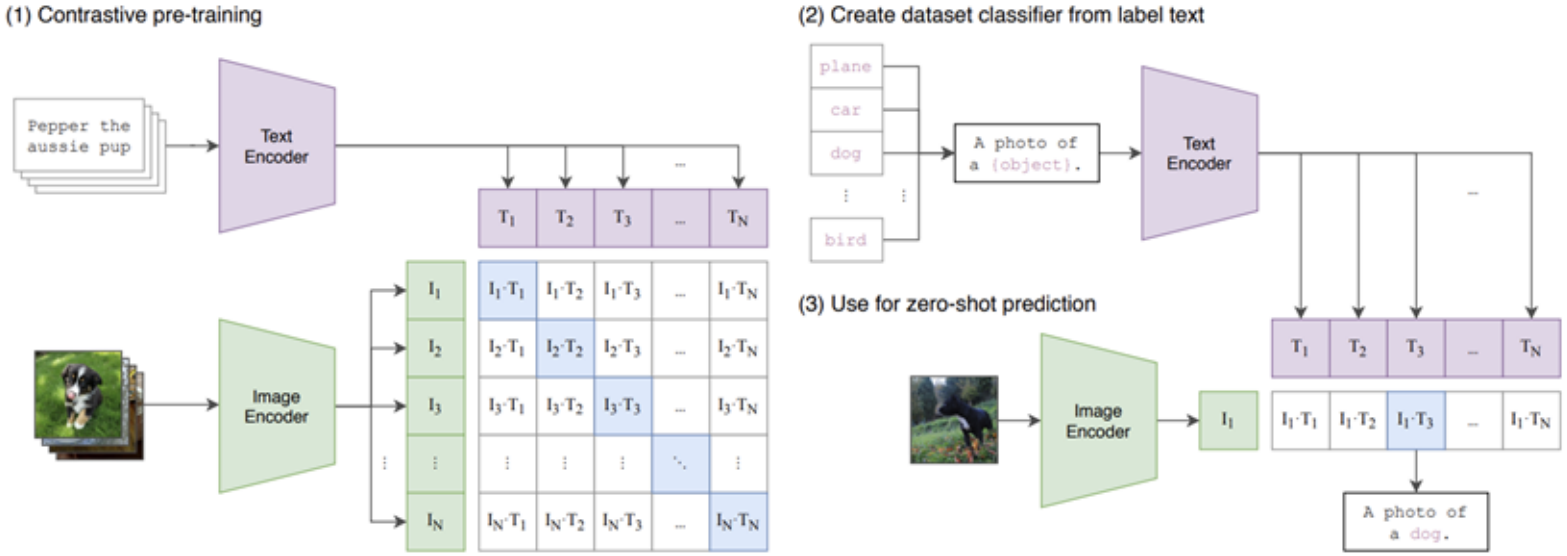
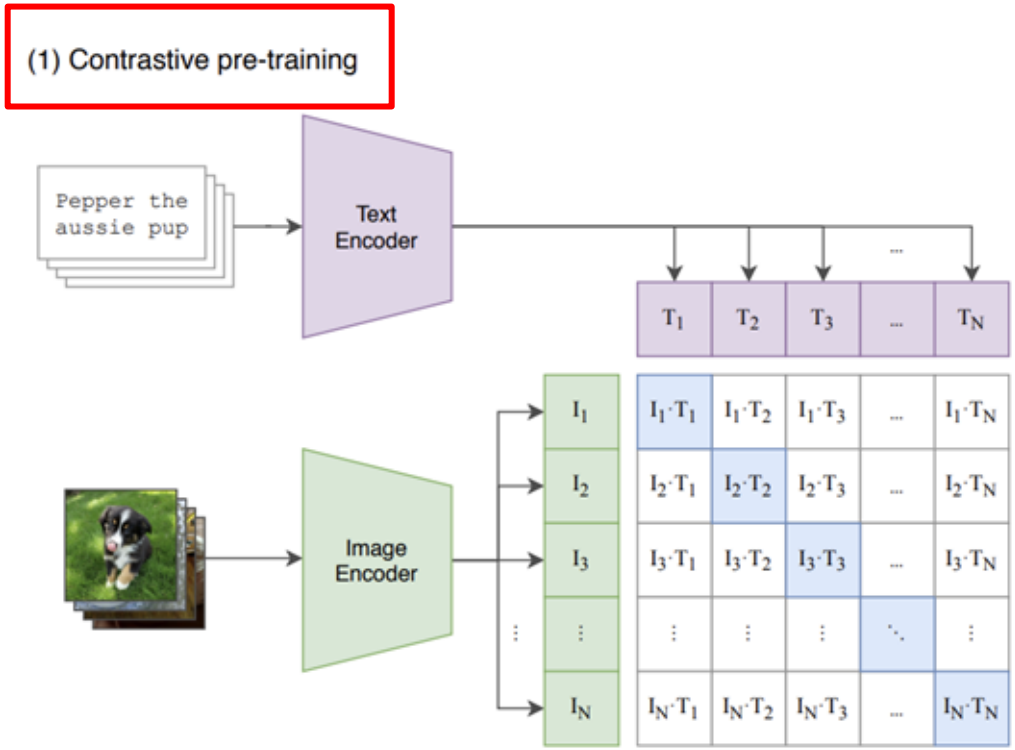


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

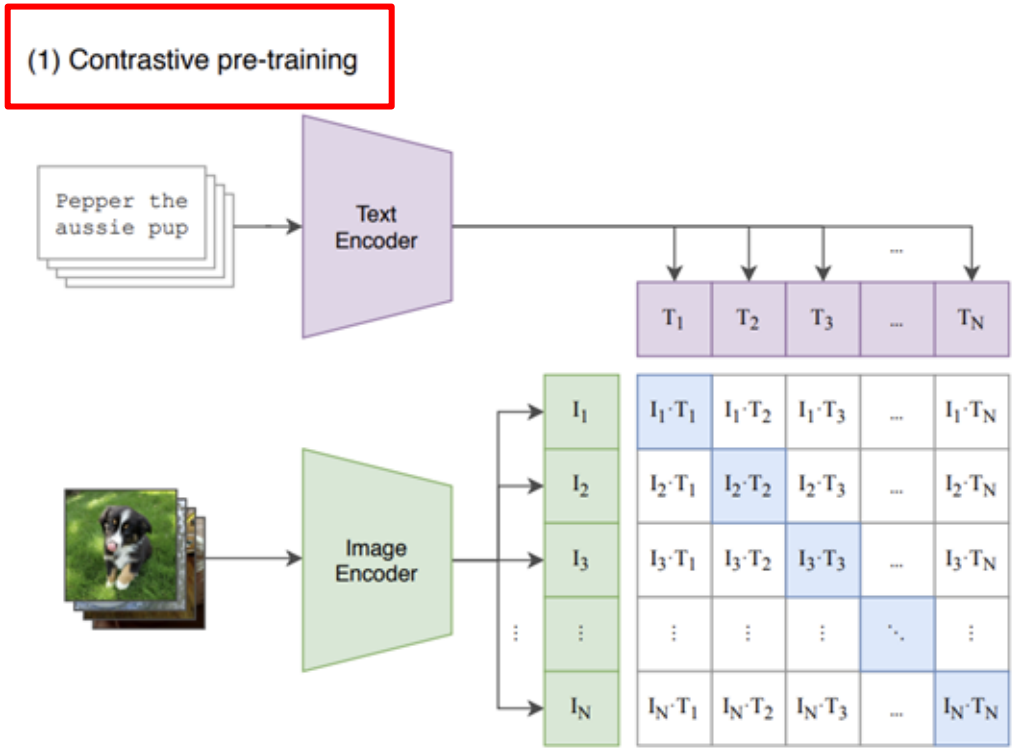
So.. what is “contrastive pre-training”?



Distinguishes between similar and dissimilar pairs!

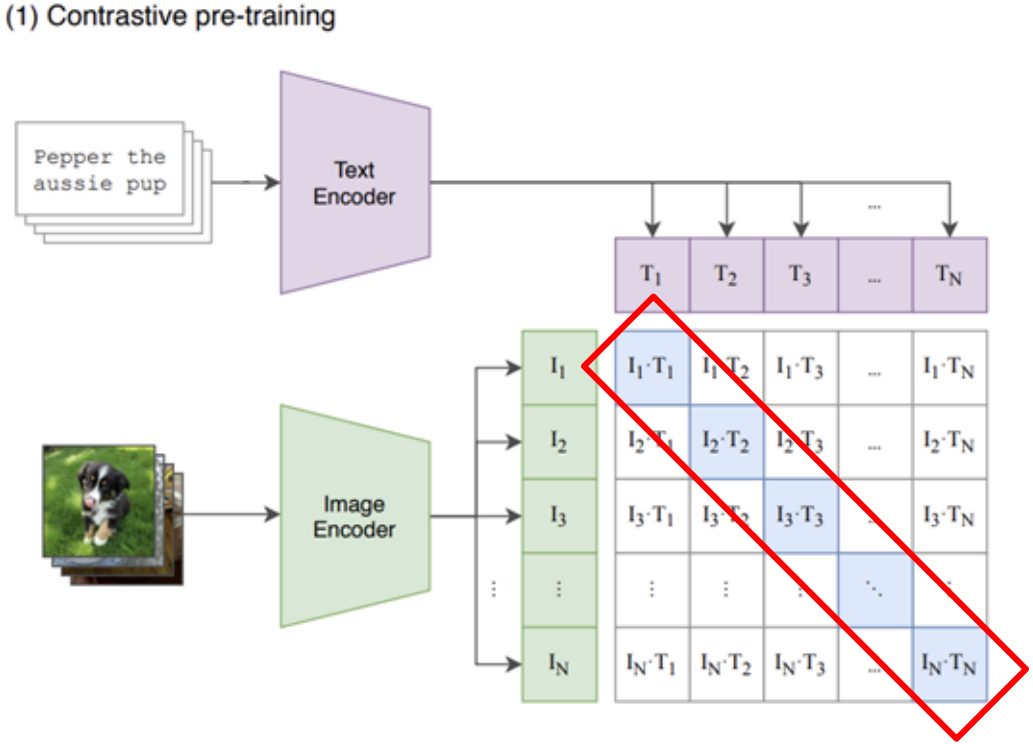
Important:
CLIP uses data-data pairs for training — no manual annotations!

So.. what is “contrastive pre-training”?



It focuses on contrasting features within pairs to learn discriminative representations, teaching the model what makes each data point unique or similar to others!

How does CLIP learn which image belongs to which caption?

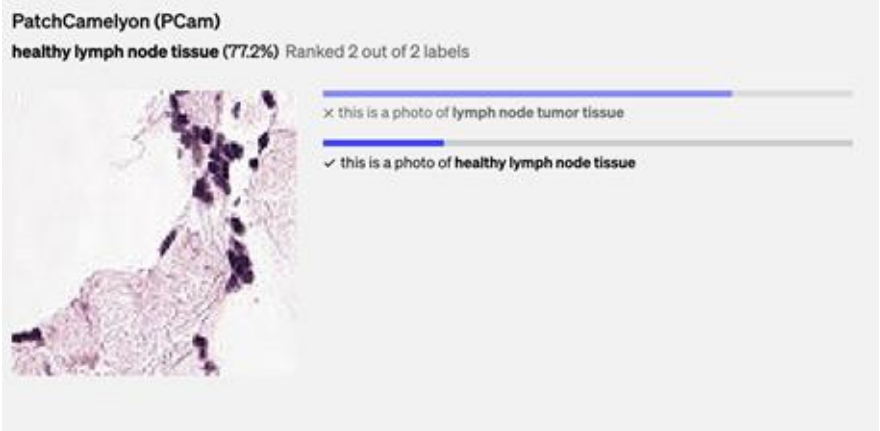


***Contrastive loss*:**
maximizes the similarity between correct pairs and minimizes the similarity between incorrect pairs

Zero-shot Learning

- Zero-shot learning allows the model to understand and relate text to images in ways it was not explicitly trained for

- This is possible because CLIP is trained on a vast amount of image-text pairs, learning a rich, multimodal space that generalizes well beyond its training data



Midterm & Final project

Multimodal **Skin Lesion Classification Dataset**

Contains:

- Image data (dermoscopic RGB images)
- Tabular data (clinical features)
- Metadata
- Label (diagnosis class, multiclass)

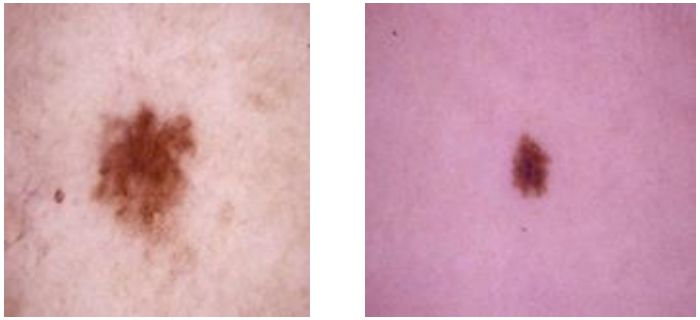
Goal: Predict the lesion class using both modalities.

Image data

Dermoscopic Images: high-resolution magnified skin images (224×224 RGB)

Common properties:

- Centered lesion
- High red channel mean (skin tone bias)
- Moderate contrast



Tabular Data

Three numerical features per sample, no missing values!

- Each row:
  [$feature_1$, $feature_2$, $feature_3$]
  [age, sex, anatomical_site]

Age: [15, 90]

Sex: {0,1}

Anatomical_site: {0,1,2,3,4,5}

Labels

Multiclass classification: 7 different classes

Code	Name	Type
mel	Melanoma	Malignant
bcc	Basal Cell Carcinoma	Malignant
akiec	Actinic Keratosis	Pre-cancerous
bkl	Benign Keratosis	Benign
nv	Melanocytic Nevus	Benign
df	Dermatofibroma	Benign
vasc	Vascular lesion	Benign

Metadata

Information accompanying the dataset

It includes:

- Class label names
- Class index mapping
- Anatomical site mapping
- Image size

Column	Description
class_names	label names
class_to_idx	mapping from class name → integer label
site_map	mapping from integer → body location
img_size	224

Challenges with the dataset

- **Class imbalance**: benign melanocytic nevus dominates the dataset, dermatofibroma, vascular lesions are rare

- **Multimodal alignment**: clinical features provide diagnostic priors, body site and age correlate with malignancy risk (fuse modalities correctly, check for shortcut learning)

Panneerselvam et el., 2025

Models

You may try & compare tabular-only, image-only, multimodal fusion

- Does adding clinical data improve performance?
- Which modality contributed most?
- Any shortcut learning?

Timeline

- February ~20th: 2500 train samples
- March 19th: 250 test samples, scoreboard on Kaggle
- March 31st: 250 test samples

  … Continue improving your models and go beyond accuracy: analyze which features or image regions drive predictions!

- April 28, 2026: scoreboard on Kaggle
- May 5th: final project presentations
- May 12th: project reports due

Recap: Multimodal learning

(1) What is it and why we need it

(2) Common challenges with multimodal models

(3) CLIP

(4) Final project