UiT The Arctic University of Norway

NORA summer school on multimodal learning

Fundamentals of multimodal classification

Rwiddhi Chakraborty

UiT Machine Learning Group and Visual Intelligence

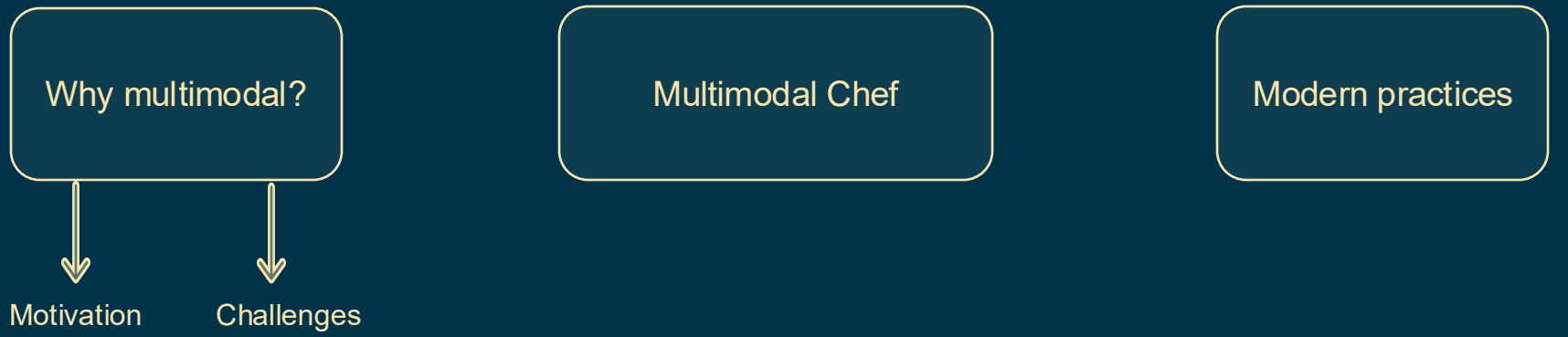
In this talk

Why multimodal?

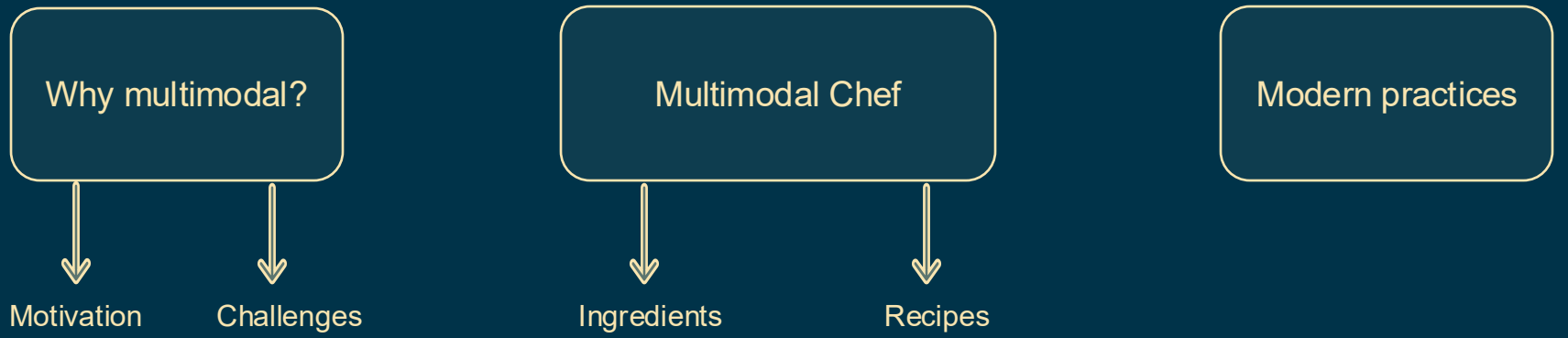
Multimodal Chef

Modern practices

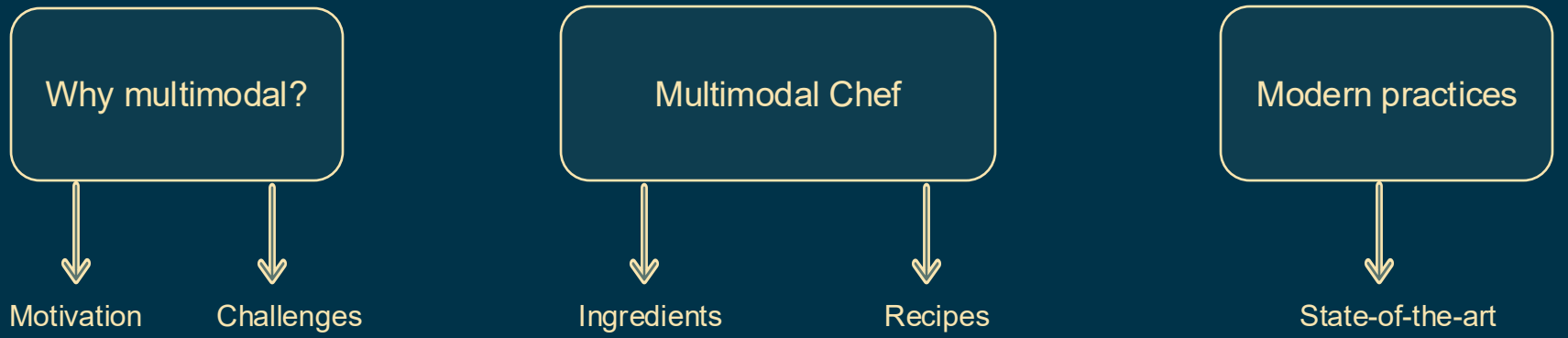
In this talk


Why multimodal?
Motivation
Challenges
Multimodal Chef
Modern practices

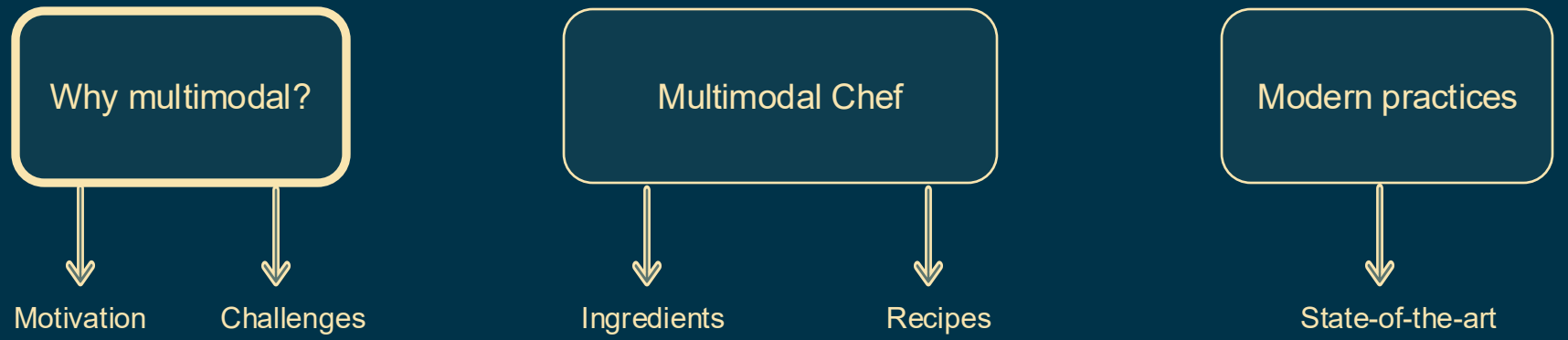
In this talk


Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices

In this talk


Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices
State-of-the-art

In this talk


Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices
State-of-the-art

Why Multimodal?

Data exists in many forms

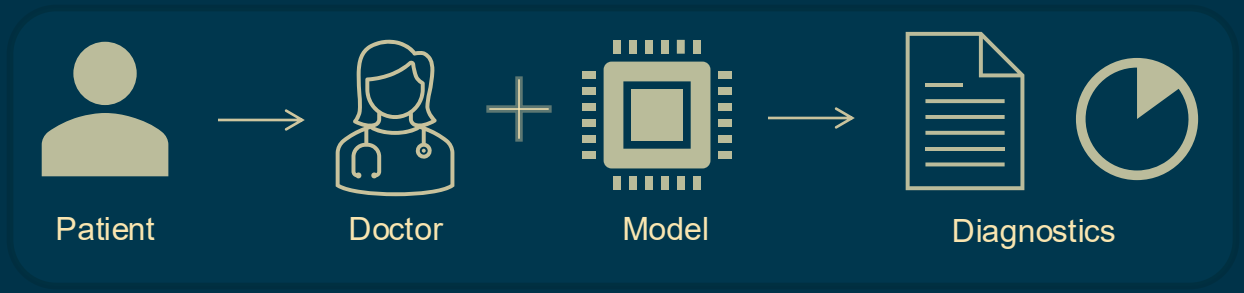
Why Multimodal?

Why Multimodal?

Diversity of forms ⇒ diversity of information ⇒ better learning

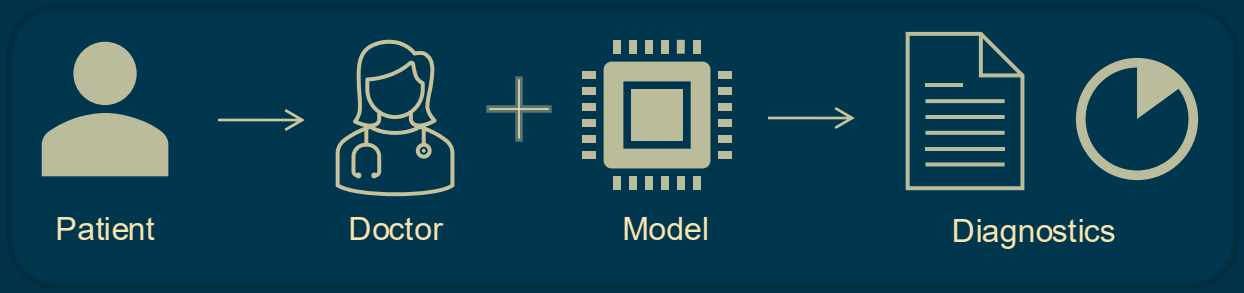
In the wild

Patient X visits an AI-assisted health service to generate a diagnostic report.



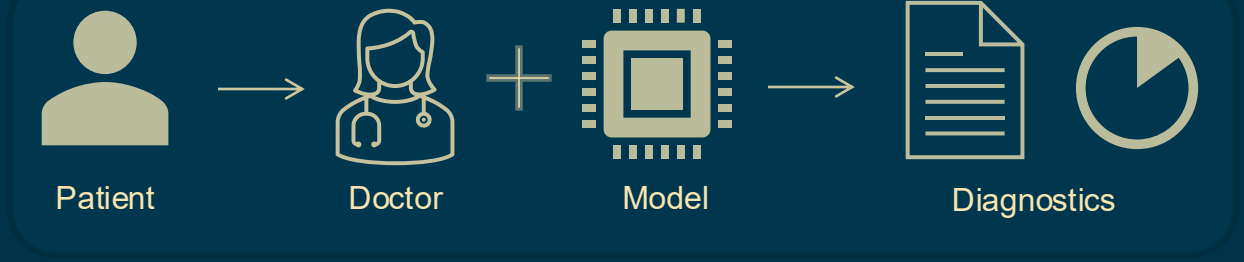
In the wild

Patient X visits an AI-assisted health service to generate a diagnostic report.
What kind of data can we access?



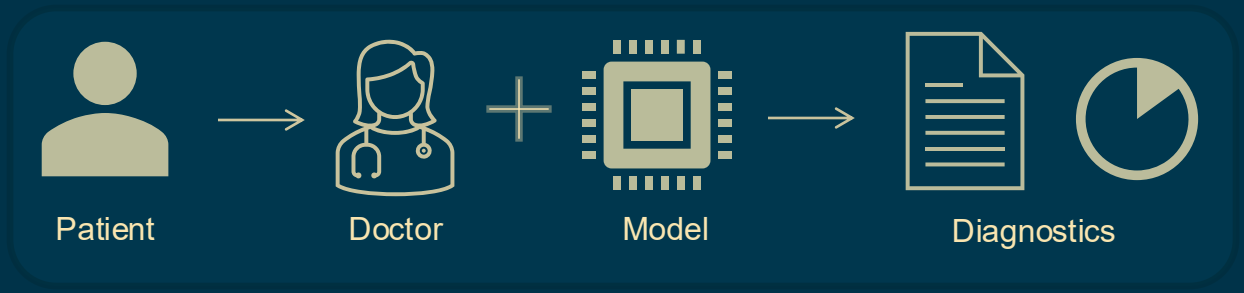
In the wild

Patient X visits an AI-assisted health service to generate a diagnostic report.
What kind of data can we access?



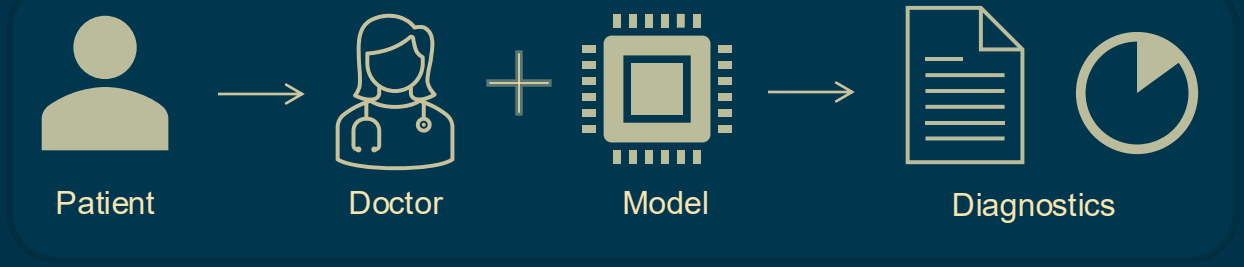
In the wild

Patient X visits an AI-assisted health service to generate a diagnostic report.
What kind of data can we access?



In the wild

Patient X visits an AI-assisted health service to generate a diagnostic report.
What kind of data can we access?



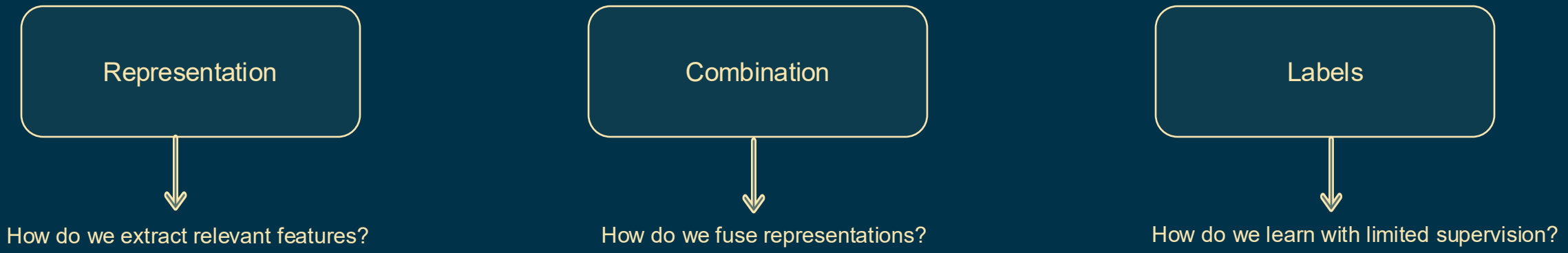
Challenges

How do we extract relevant features?

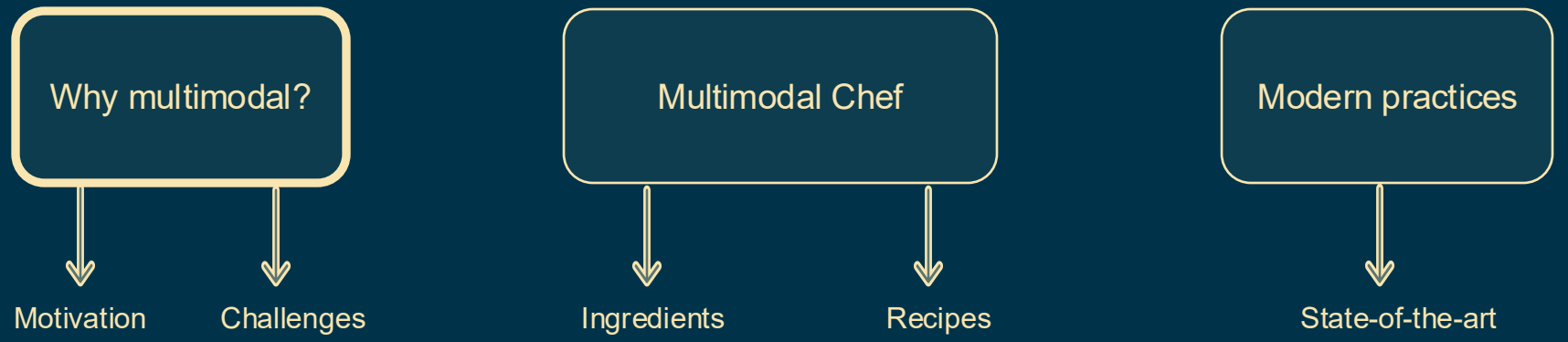
Challenges


Representation
How do we extract relevant features?
Combination
How do we fuse representations?

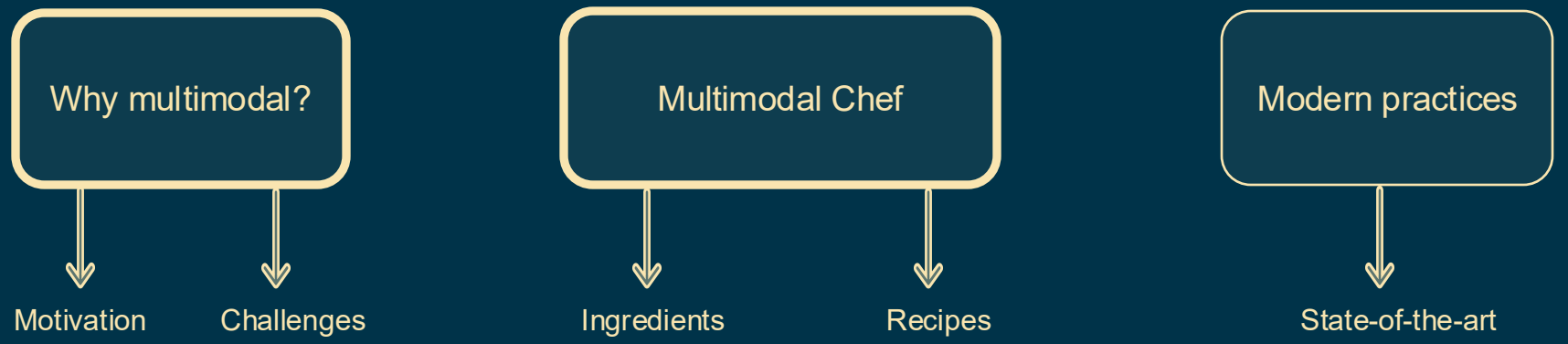
Challenges


Representation
How do we extract relevant features?
Combination
How do we fuse representations?
Labels
How do we learn with limited supervision?

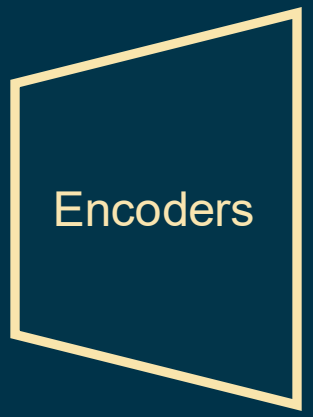
In this talk


Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices
State-of-the-art

In this talk

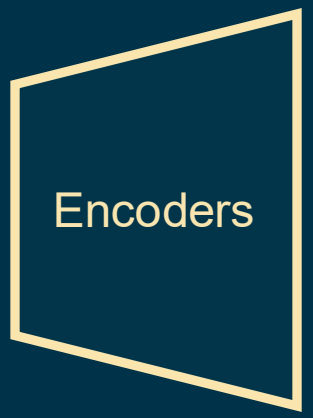

Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices
State-of-the-art

Multimodal Chef: The Ingredients

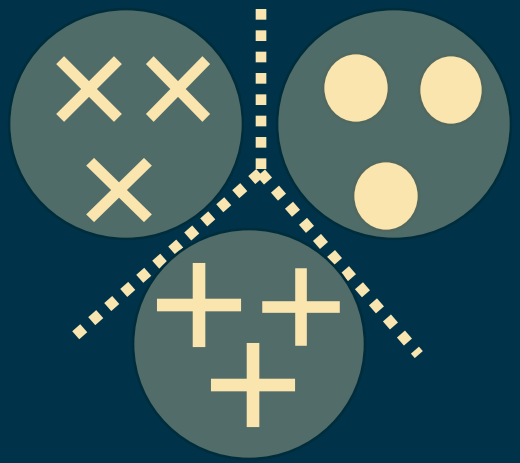


Feature Extraction

Multimodal Chef: The Ingredients

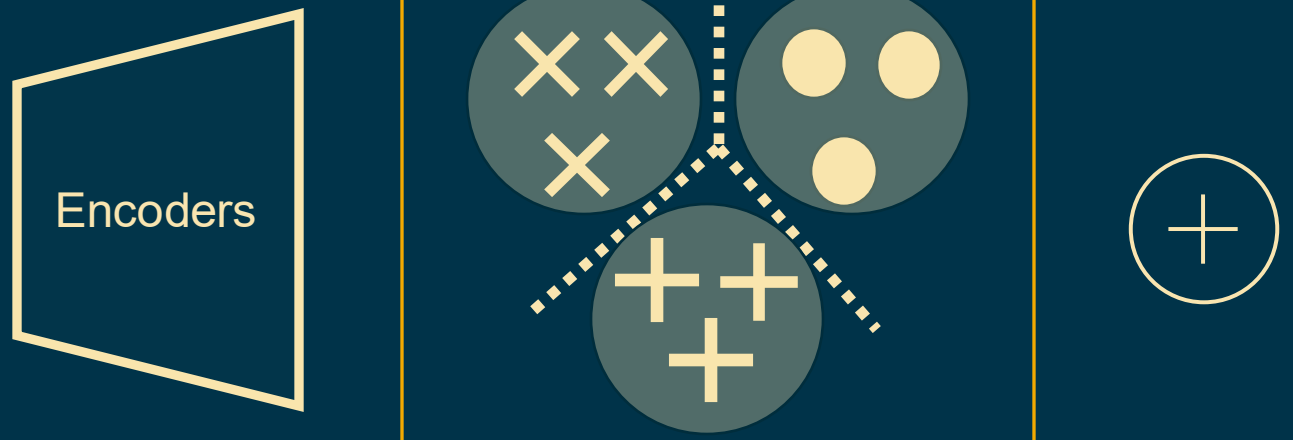


Feature Extraction



Representations

Multimodal Chef: The Ingredients

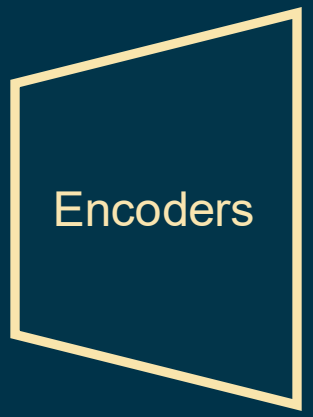


Feature Extraction

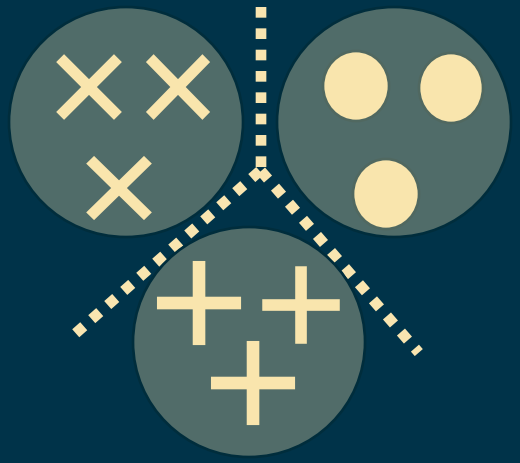
Representations

Fuse & Learn

Multimodal Chef: The Ingredients

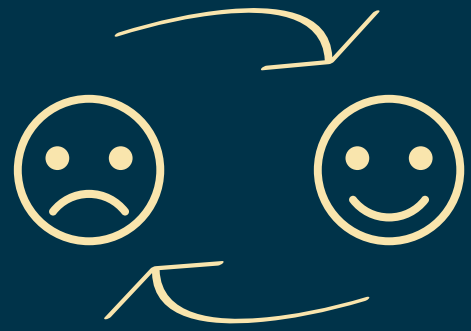


Feature Extraction



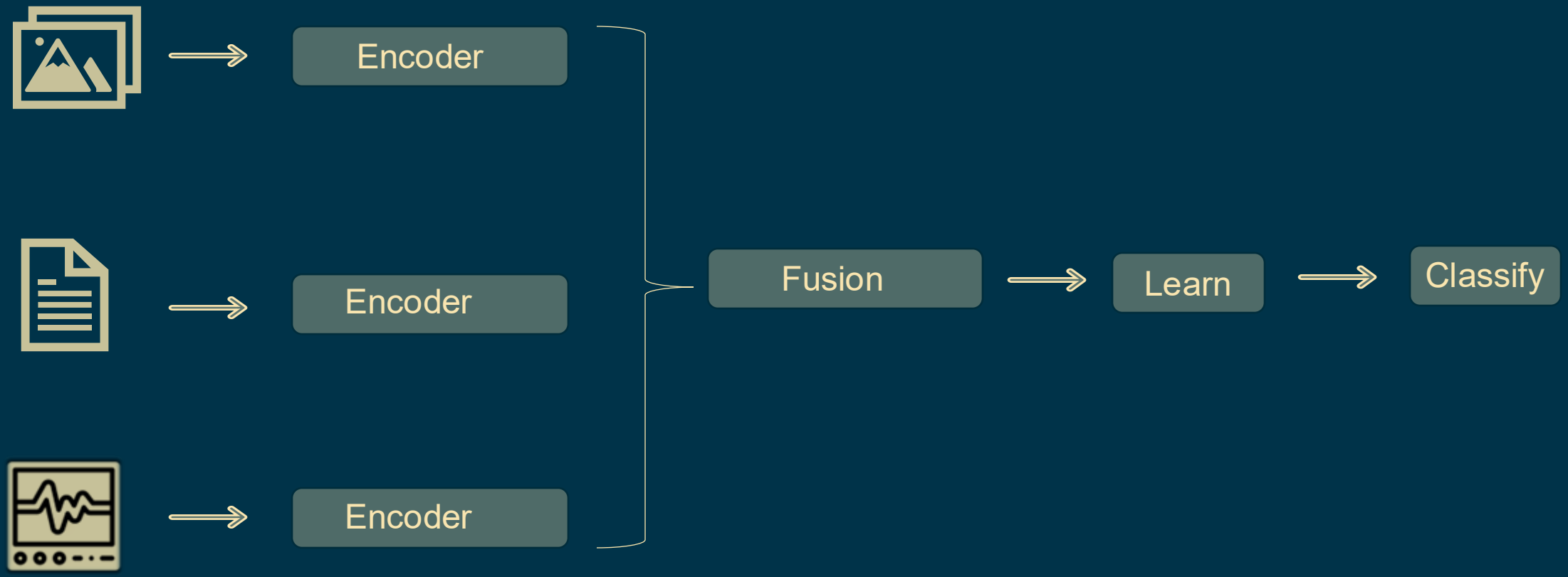
Representations

Fuse & Learn

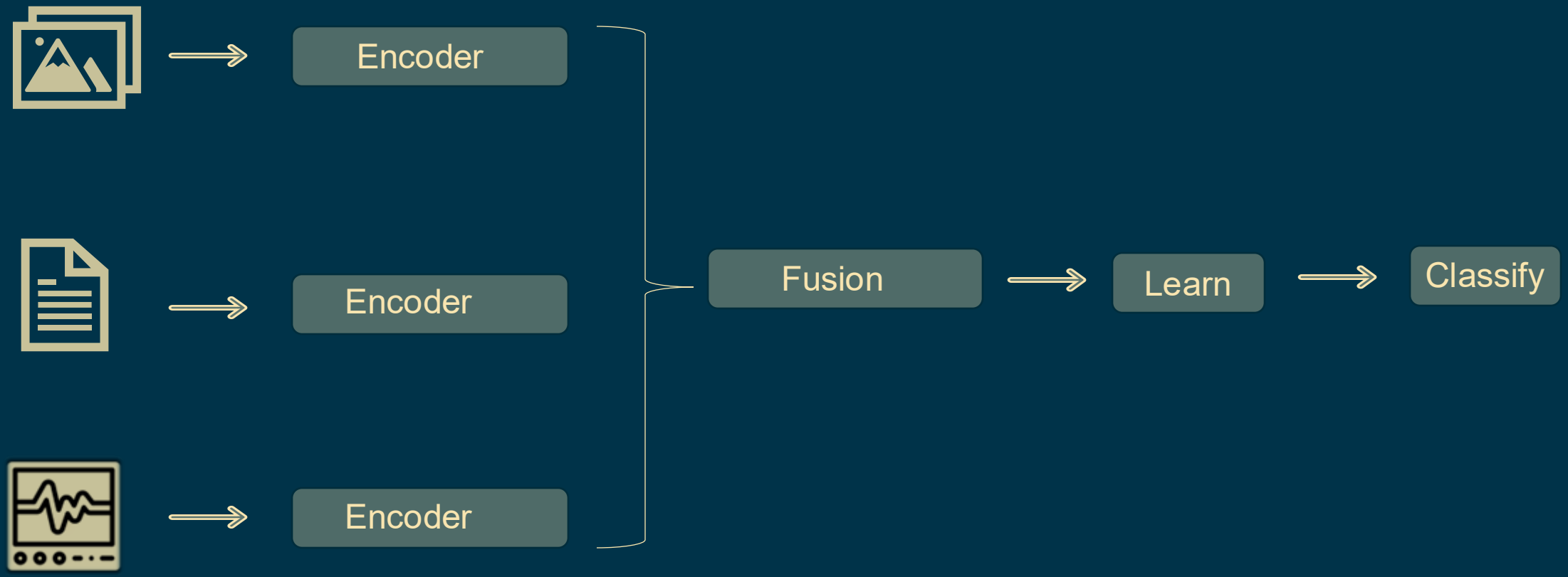


Classify

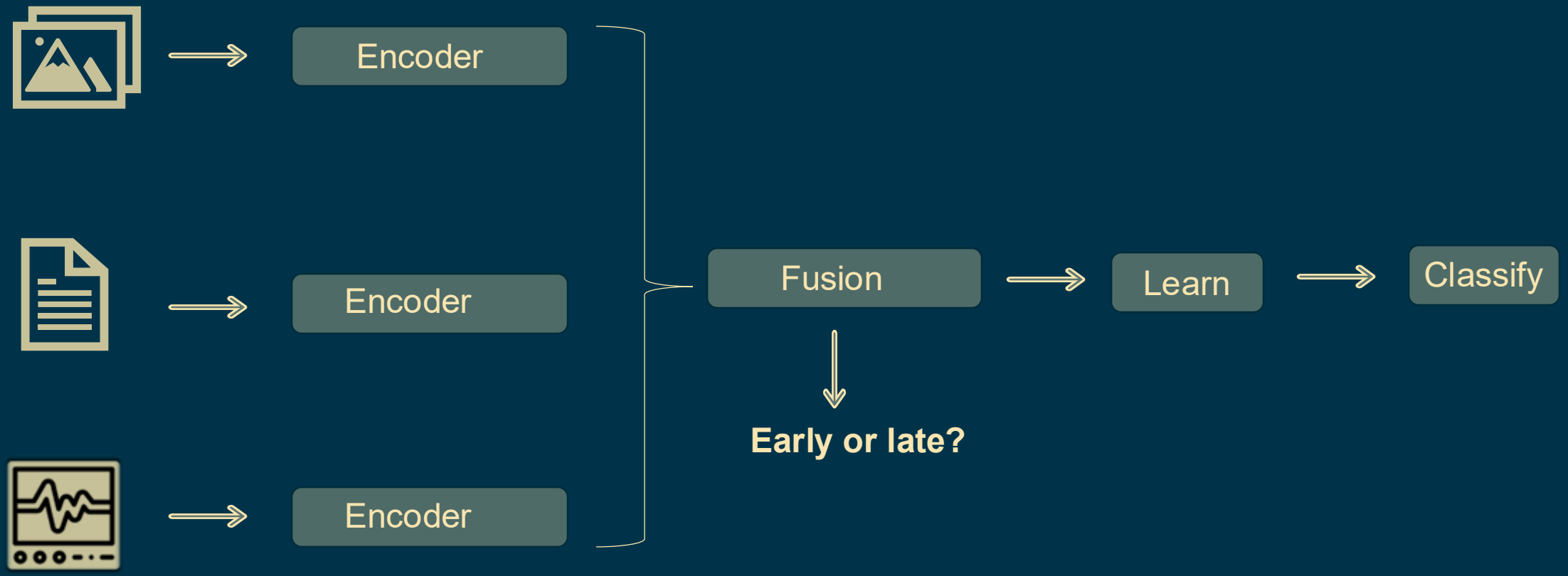
Multimodal Chef: The Ingredients



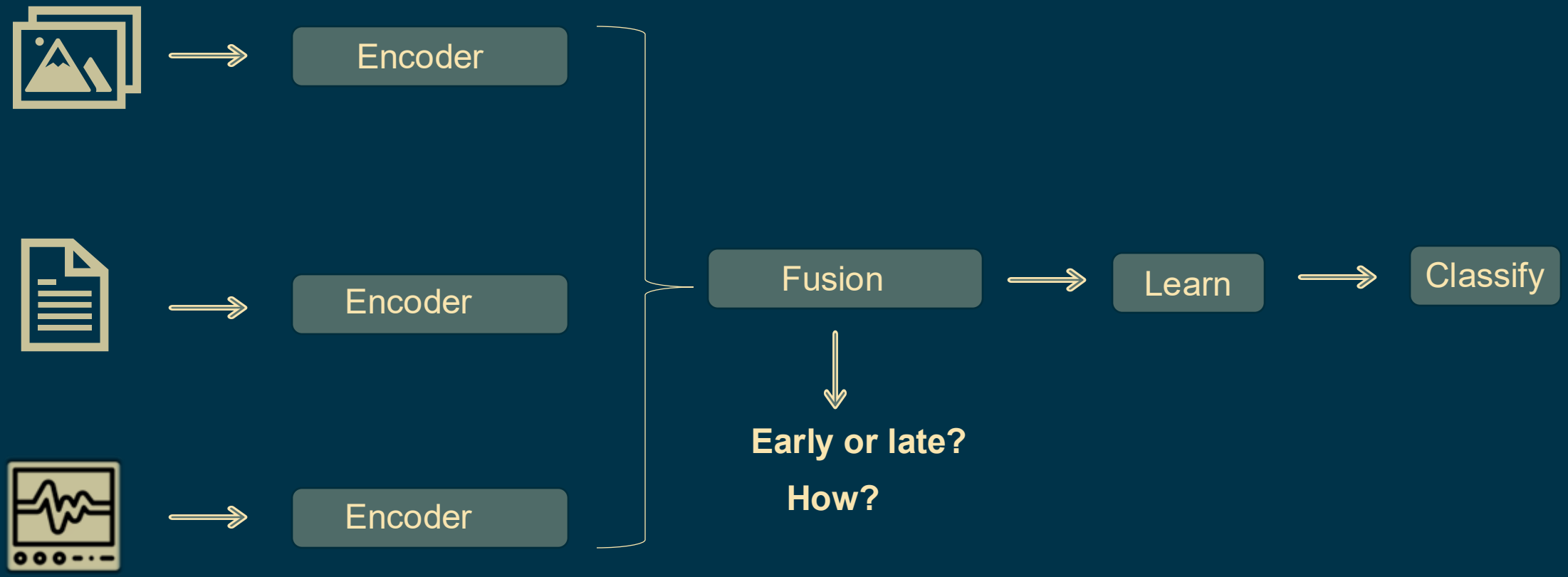
Multimodal Chef: Recipes



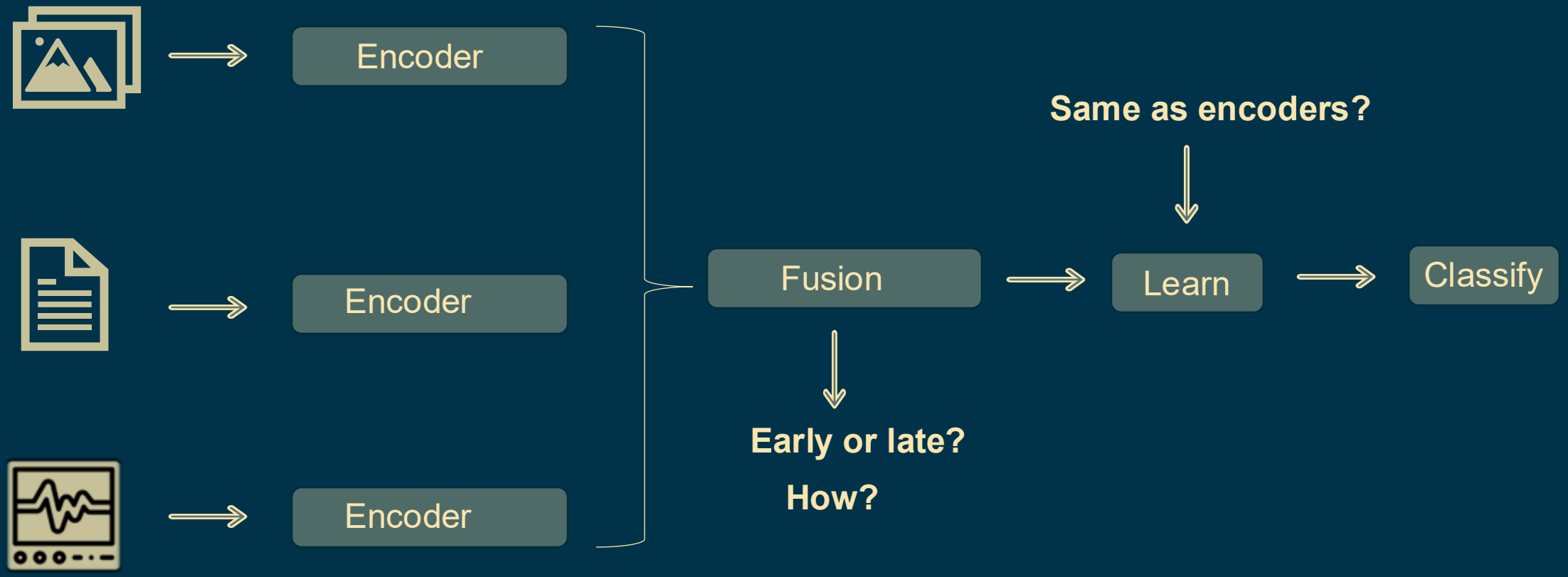
Multimodal Chef: Recipes



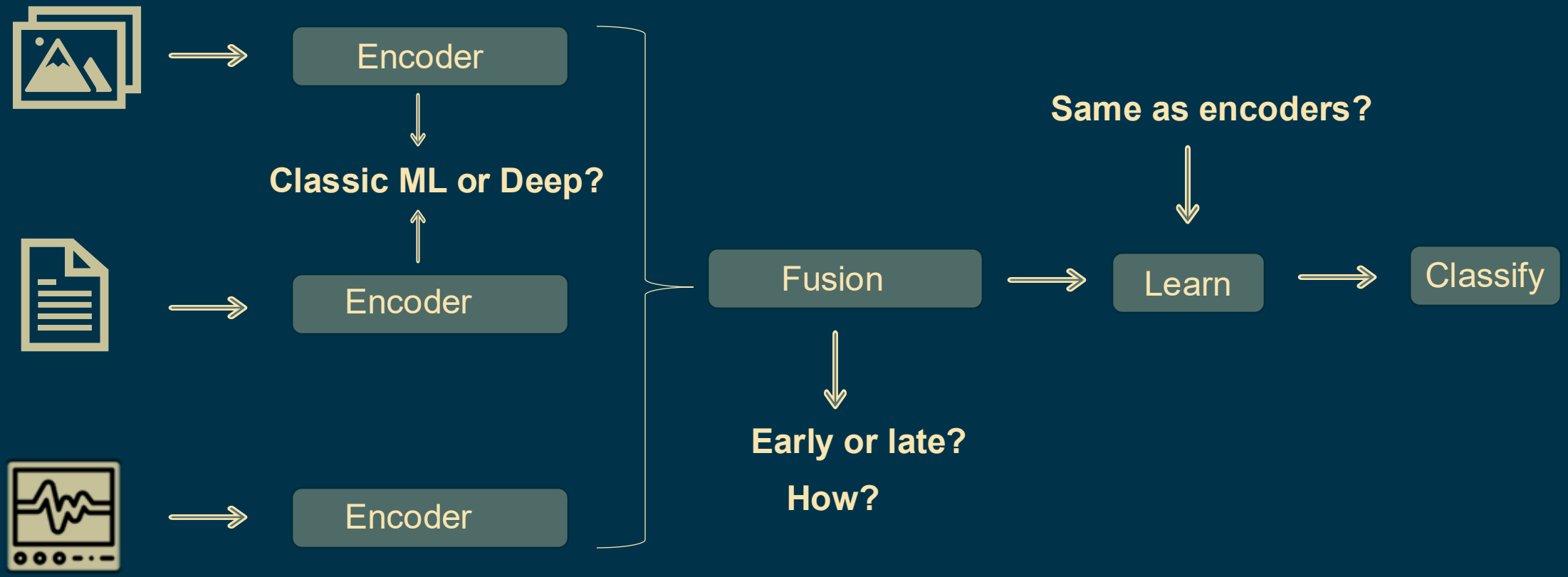
Multimodal Chef: Recipes



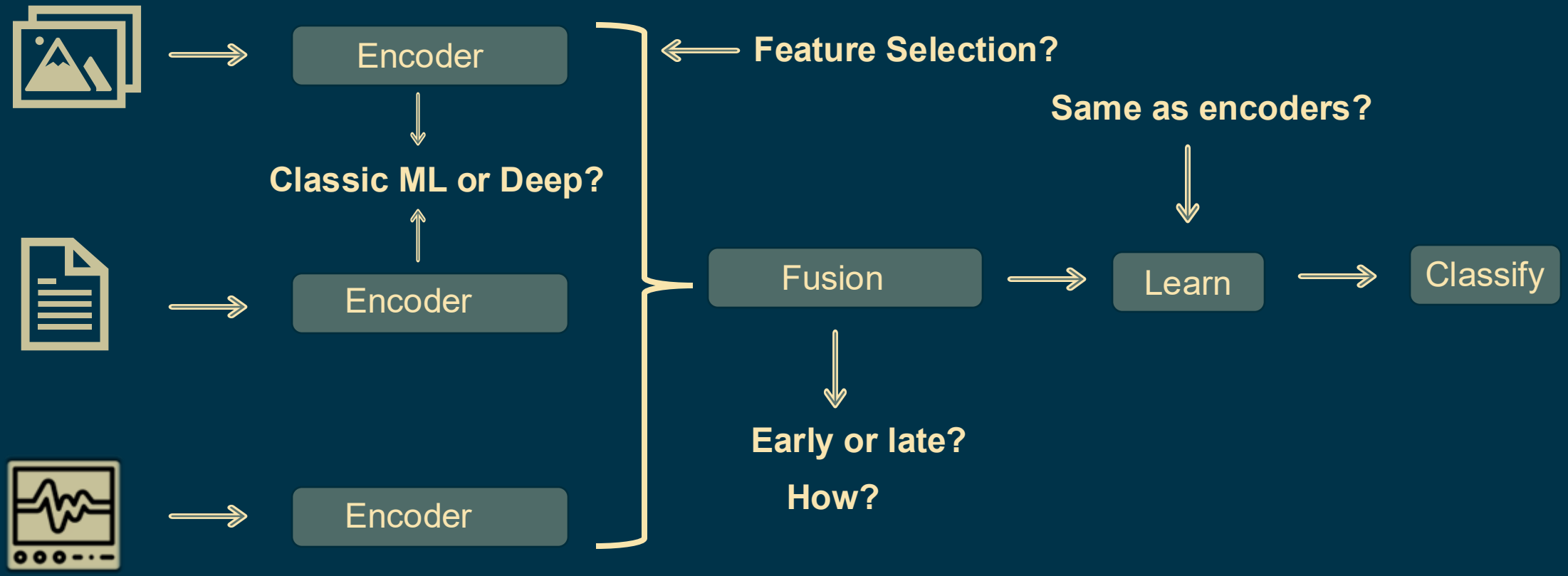
Multimodal Chef: Recipes



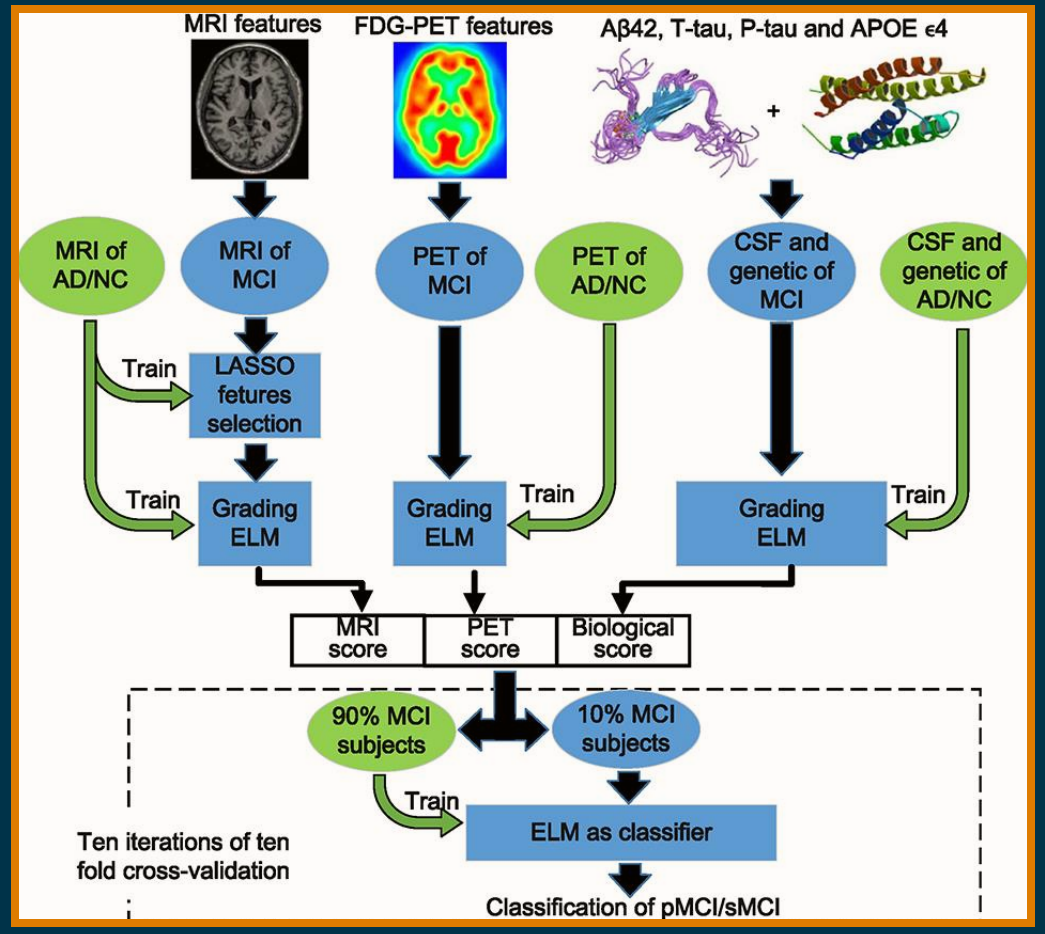
Multimodal Chef: Recipes



Multimodal Chef: Recipes

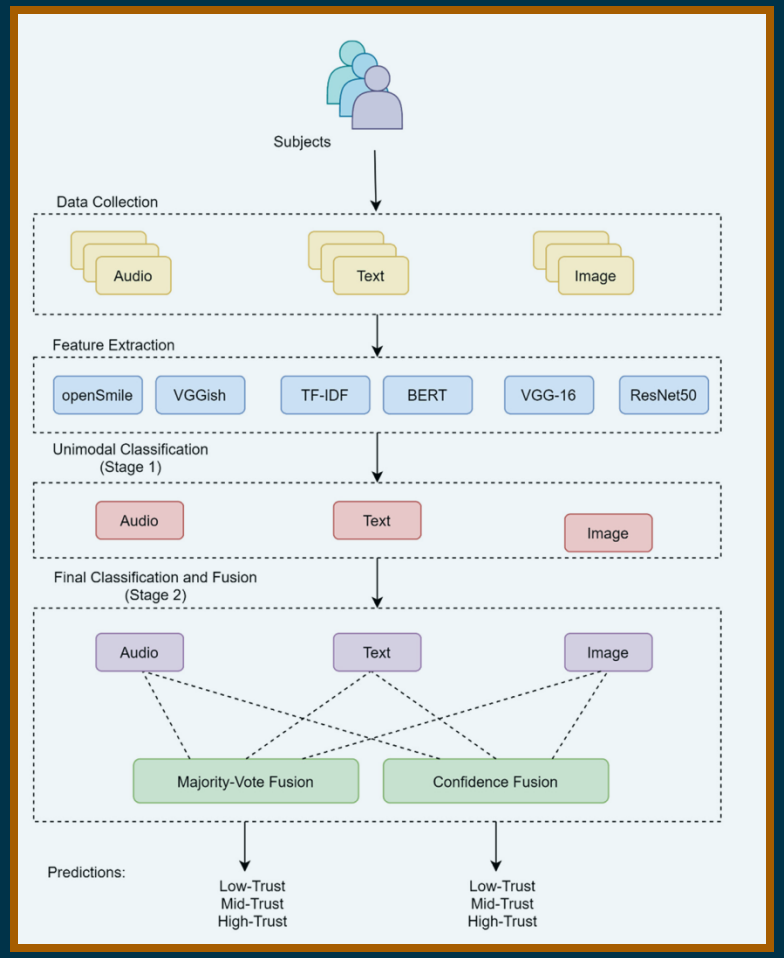


Multimodal Chef: Recipes in action



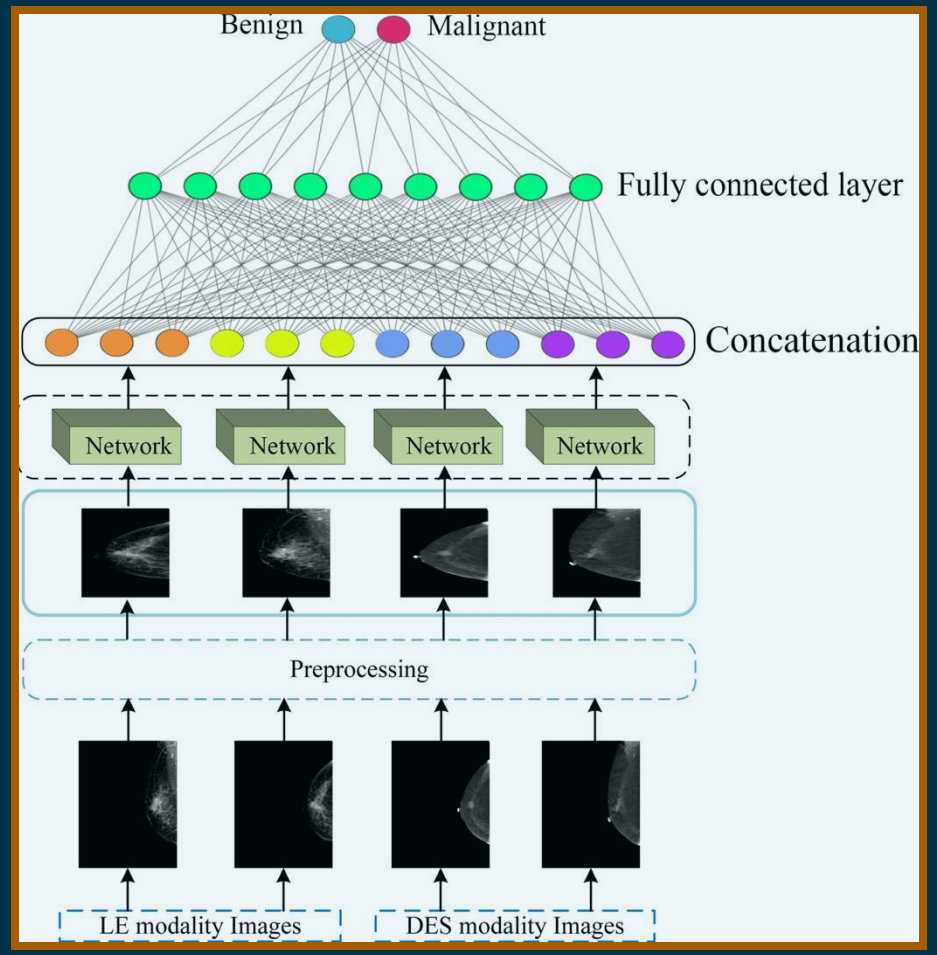
Lin W, Gao Q, Yuan J, Chen Z, Feng C, Chen W, Du M and Tong T (2020) Predicting Alzheimer's Disease Conversion From Mild Cognitive Impairment Using an Extreme Learning Machine-Based Grading Method With Multimodal Data. *Front. Aging Neurosci.* 12:77. doi: 10.3389/fnagi.2020.00077

Multimodal Chef: Recipes in action



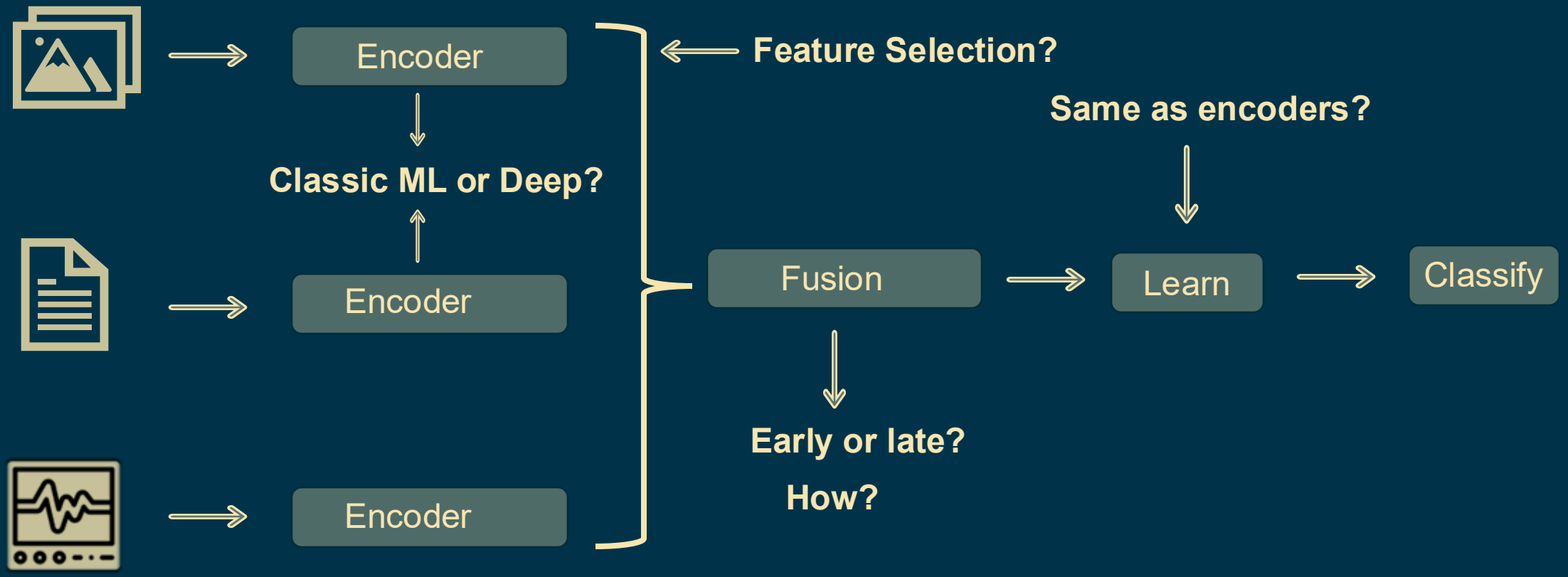
Syed, M.S.S.; Pirogova, E.; Lech, M. Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach. *Electronics* **2021**, *10*, 1259. https://doi.org/10.3390/electronics10111259

Multimodal Chef: Recipes in action

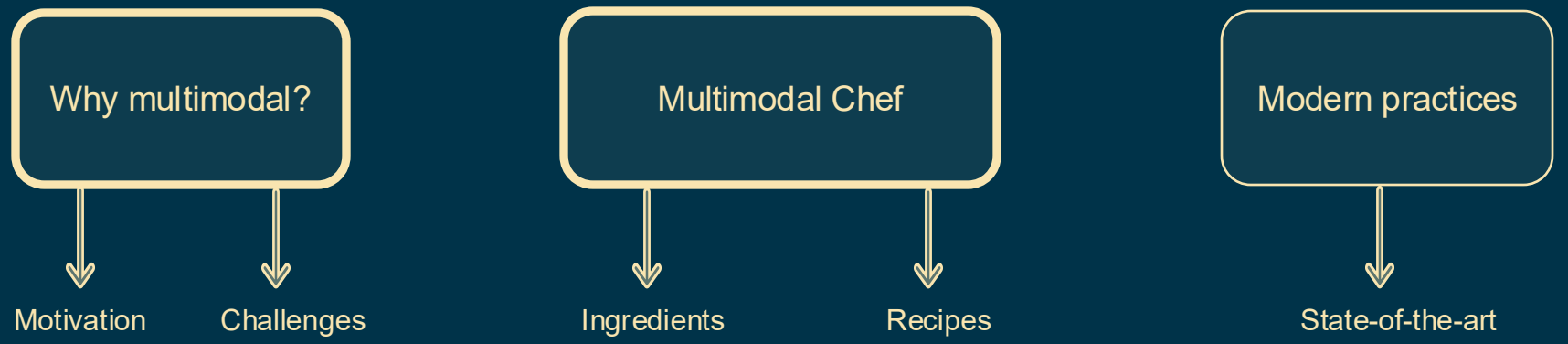


Song, J., Zheng, Y., Zakir Ullah, M. *et al.* Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images. *Int J CARS* **16**, 979–988 (2021). https://doi.org/10.1007/s11548-021-02391-4

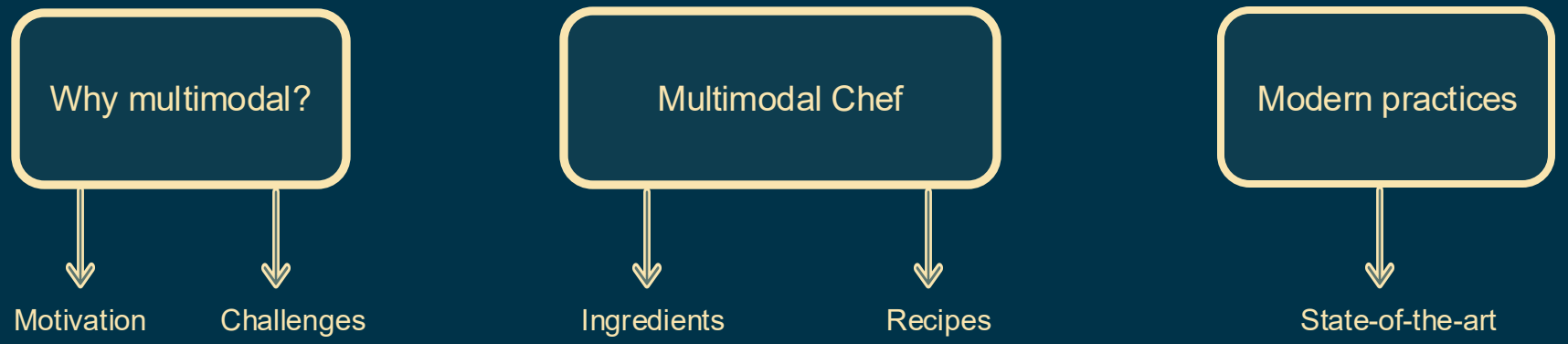
Multimodal Chef: Recipes



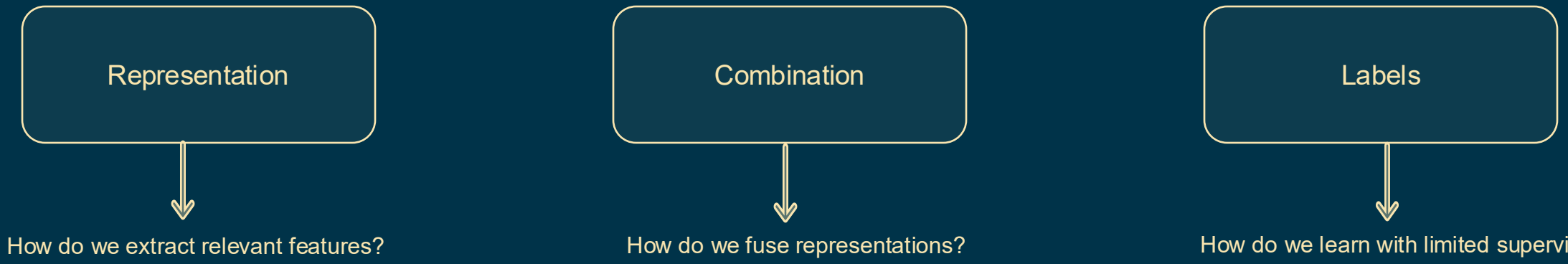
In this talk


Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices
State-of-the-art

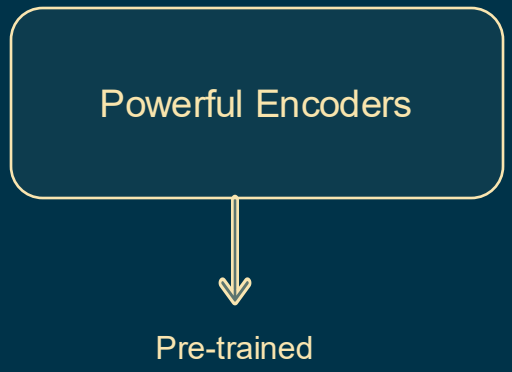
In this talk


Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices
State-of-the-art

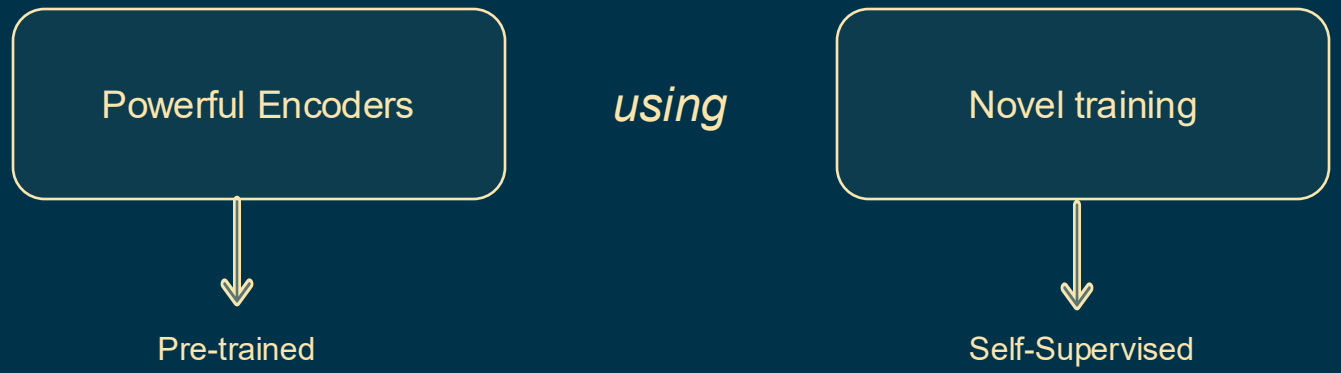
Recall the challenges



Modern Practice



Modern Practice



Modern Practice

Modern Practice

Great because

Powerful representations

”Fusion” baked into training

Modern Practice

Great because

- Powerful representations
- ”Fusion” baked into training

BUT

- Need high quality data and good compute

Modern Practice: Contrastive Pre-training

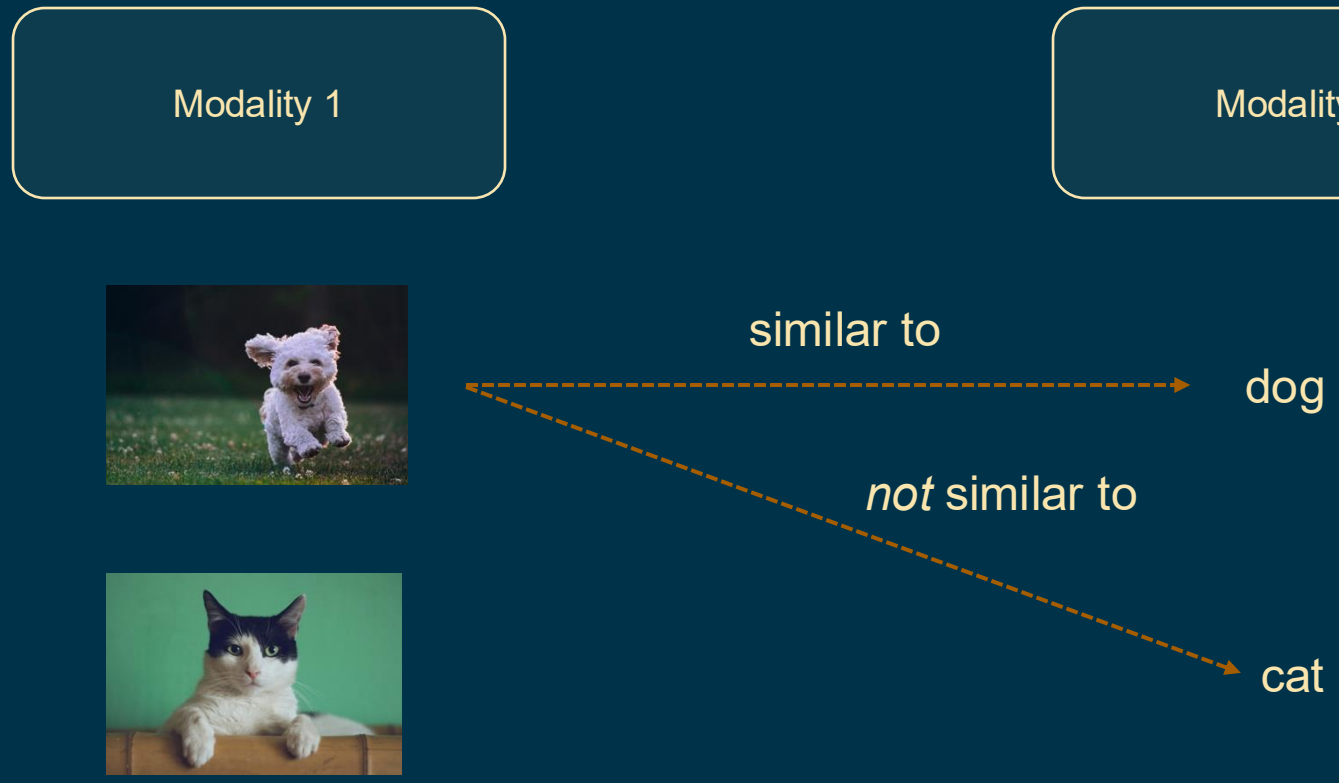
Modality 1

Modality 2

dog

cat

Modern Practice: Contrastive Pre-training



Modern Practice: Contrastive Pre-training

You're good if you have 400 million such examples

Modality 1

Modality 2

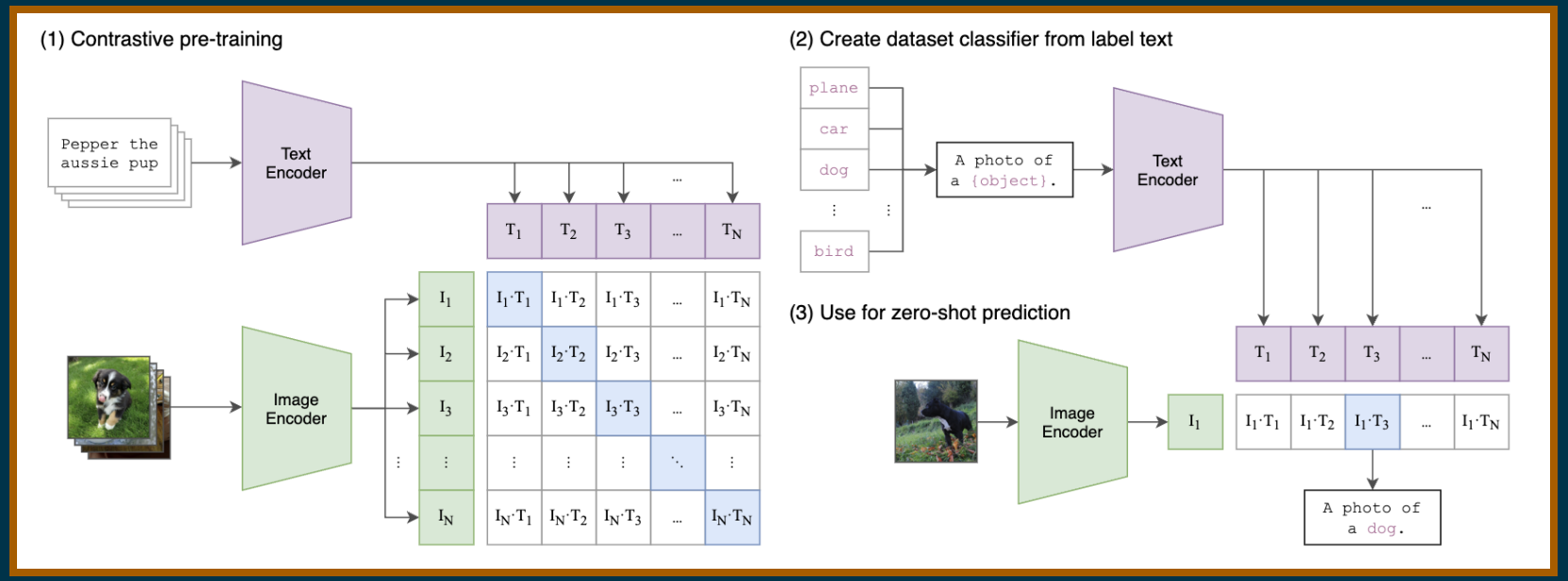
similar to

dog

not similar to

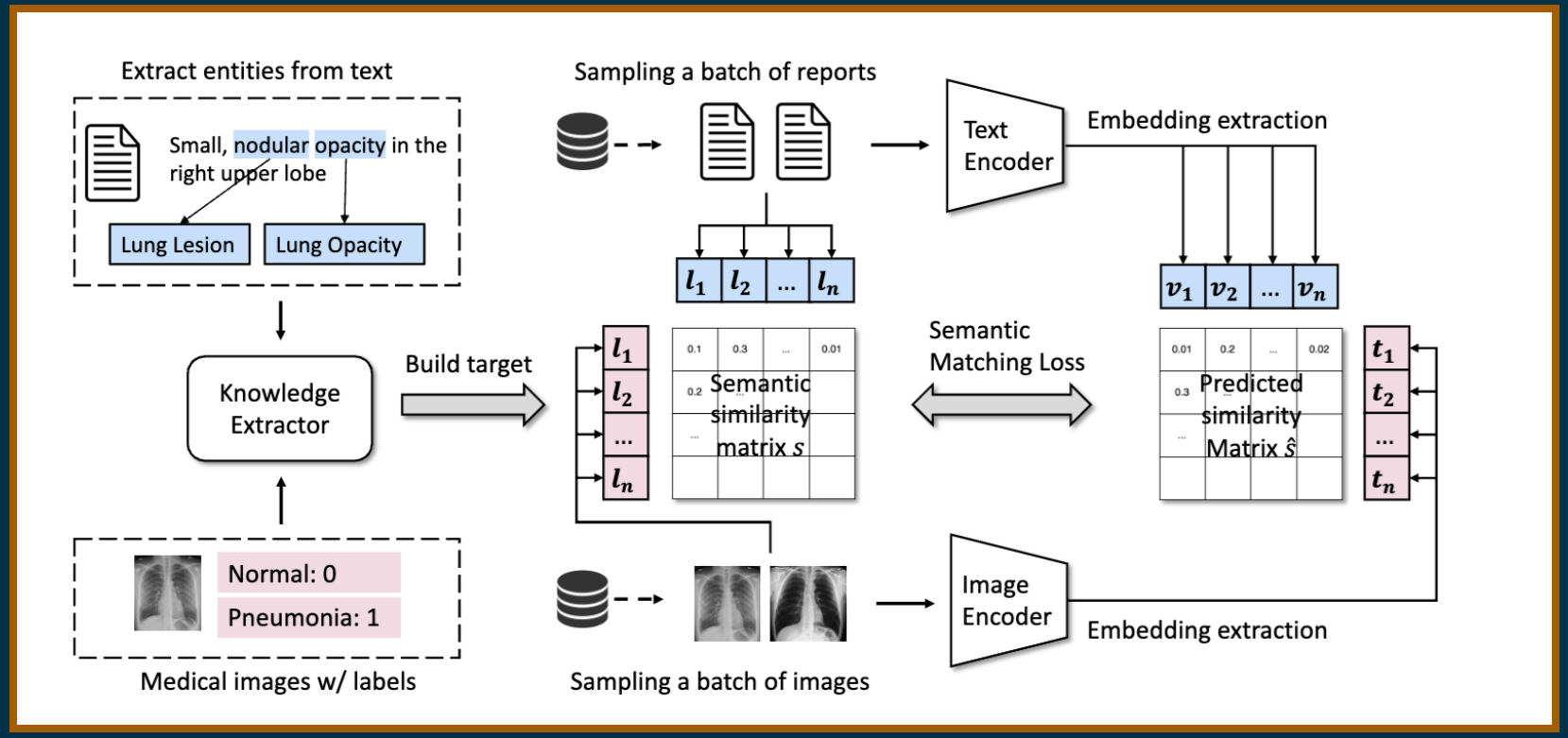
cat

Modern Practice: Contrastive Pre-training



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

Modern Practice: MedCLIP



Wang, Zifeng, et al. "Medclip: Contrastive learning from unpaired medical images and text." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. Vol. 2022. 2022.

Modern Practice: The Contrastive Zoo!

RemoteCLIP

MedCLIP

FashionCLIP

AudioCLIP

BioCLIP

ECG-CLIP

Modern Practice: The Contrastive Zoo!

RemoteCLIP

MedCLIP

FashionCLIP

AudioCLIP

BioCLIP

ECG-CLIP

and many more..

Modern Practice: The Contrastive Zoo!

RemoteCLIP

MedCLIP

FashionCLIP

AudioCLIP

BioCLIP

ECG-CLIP

and many (MANY) more..

Modern Practice: CLIP is all you need?

Modern Practice: CLIP is all you need?

Not exactly

Modern Practice: CLIP is all you need?

Not exactly

CLIP is simply a method of training similar representations

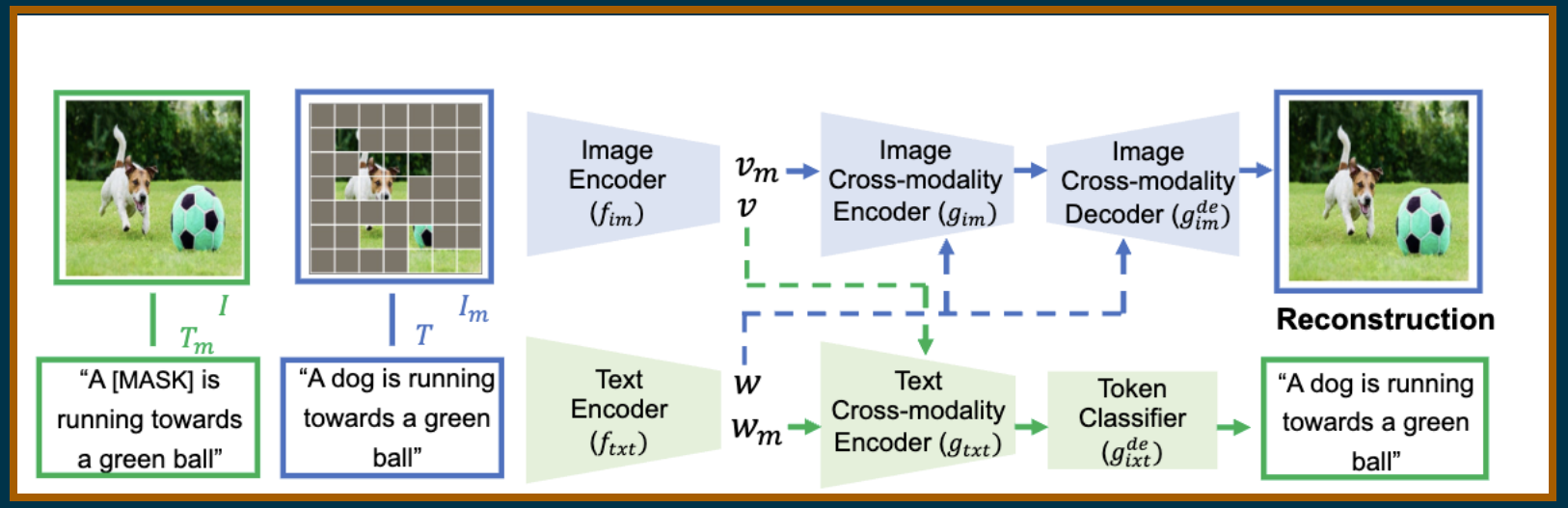
Modern Practice: CLIP is all you need?

Not exactly

CLIP is simply a method of training similar representations

There are other quite a few popular alternatives

Modern Practice: Masked Modelling



Kwon, Gukyeong, et al. "Masked vision and language modeling for multi-modal representation learning." arXiv preprint arXiv:2208.02131 (2022).

Modern Practice: CLIP is all you need?

Not exactly

CLIP is simply a method of training similar representations

There are other quite a few popular alternatives

Modern Practice: CLIP is all you need?

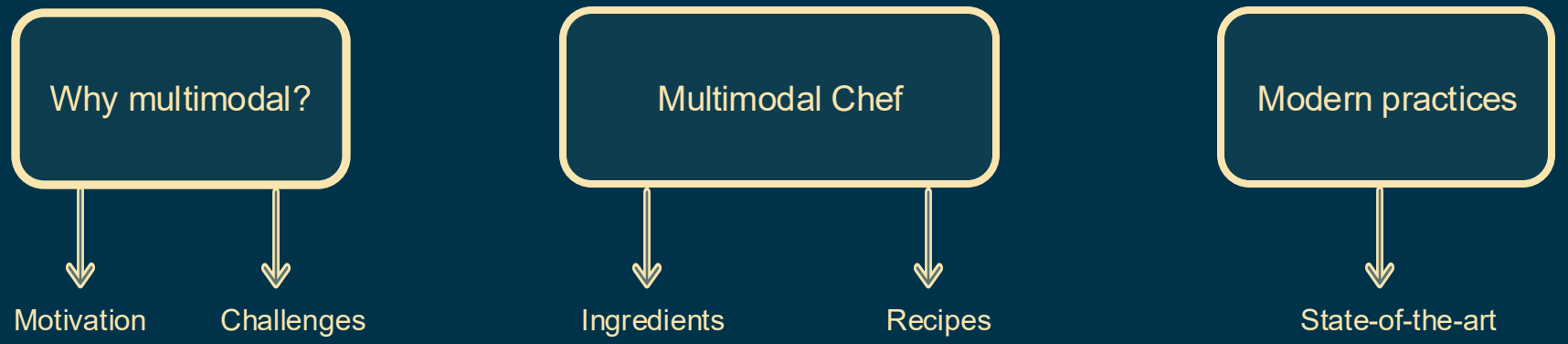
Not exactly

CLIP is simply a method of training similar representations

There are other quite a few popular alternatives

You could even pair of a powerful image-only encoder with a large language model!

Summary


Why multimodal?
Motivation
Challenges
Multimodal Chef
Ingredients
Recipes
Modern practices
State-of-the-art