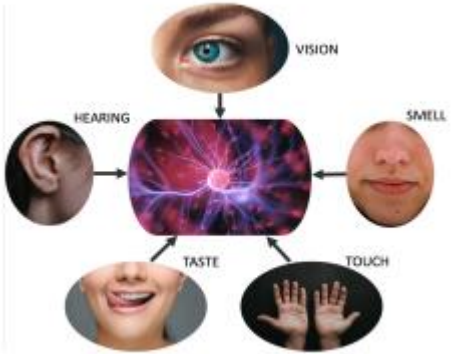
Data representation for multimodal models

Lidia Pivovarova

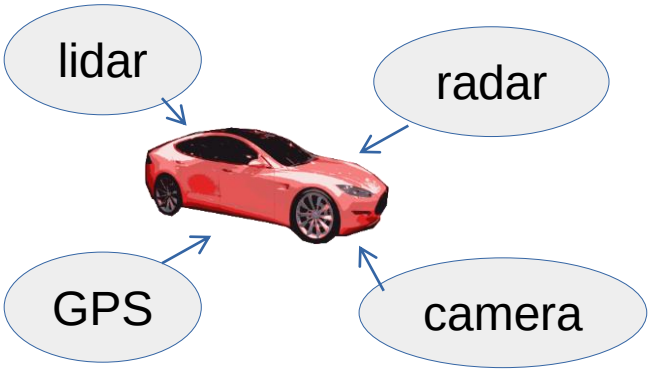
28.08.2024

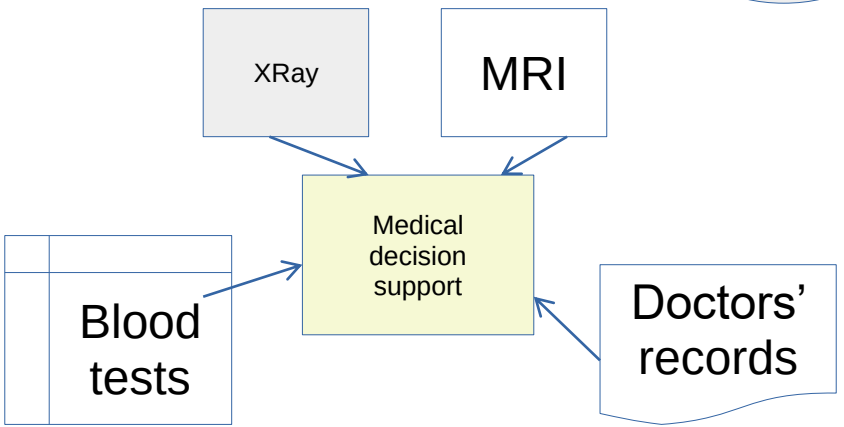
Summer study group

Multimodality in machine learning



(Zhao et al., 2024)





Application domains

- Robotics
- Healthcare
- Multimedia
- Human-computer interaction
- Interactive agents
- etc

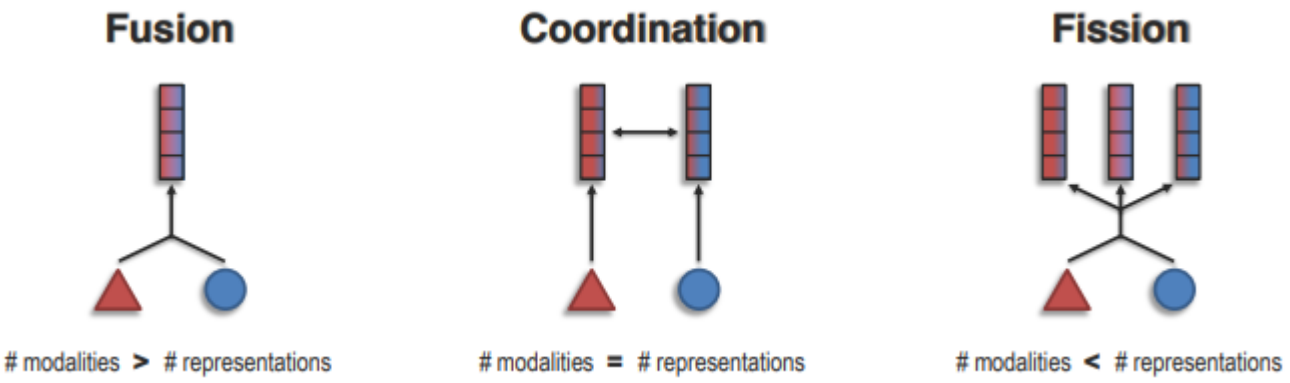
Principles of multimodality

1. Modalities are heterogenious
   - diverse in quality, structure, distributions, noise, and relevance
2. Modalities are connected
   - statistical and semantic relations, redundancy and uniqueness of information
3. Modalities interact
   - modalities processed together result in a novel

Unimodal representations

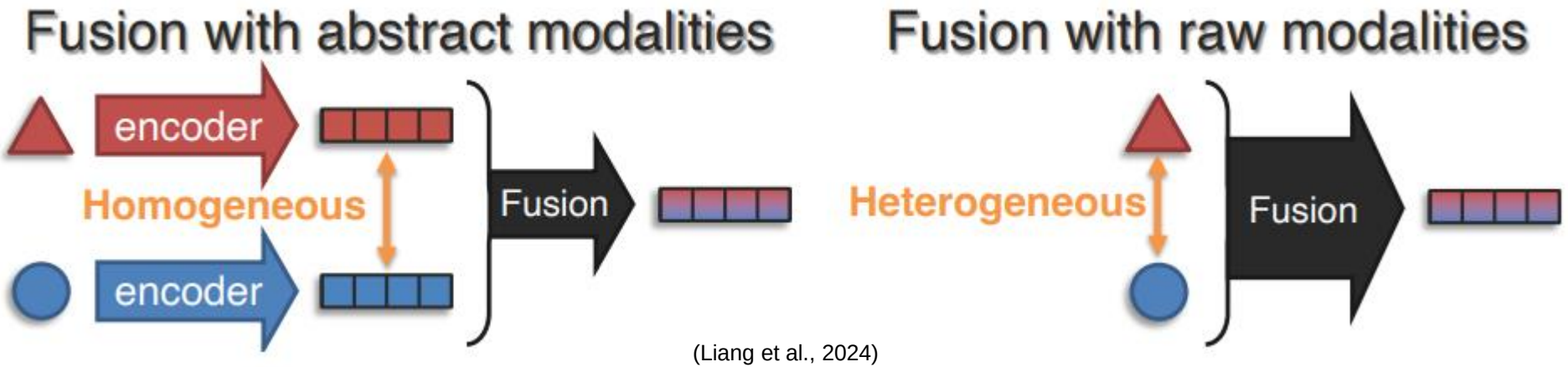
- Shift from hand-crafted features to data-driven, obtained as a result of self-supervised pretraining
  - Visual features: convolutional networks, e.g. ResNet,
  - Textual features: emebddings, LLMs
  - Audio: neural acoustic models
  - Graphs: either linearization and LLMs or graph networks

Representation



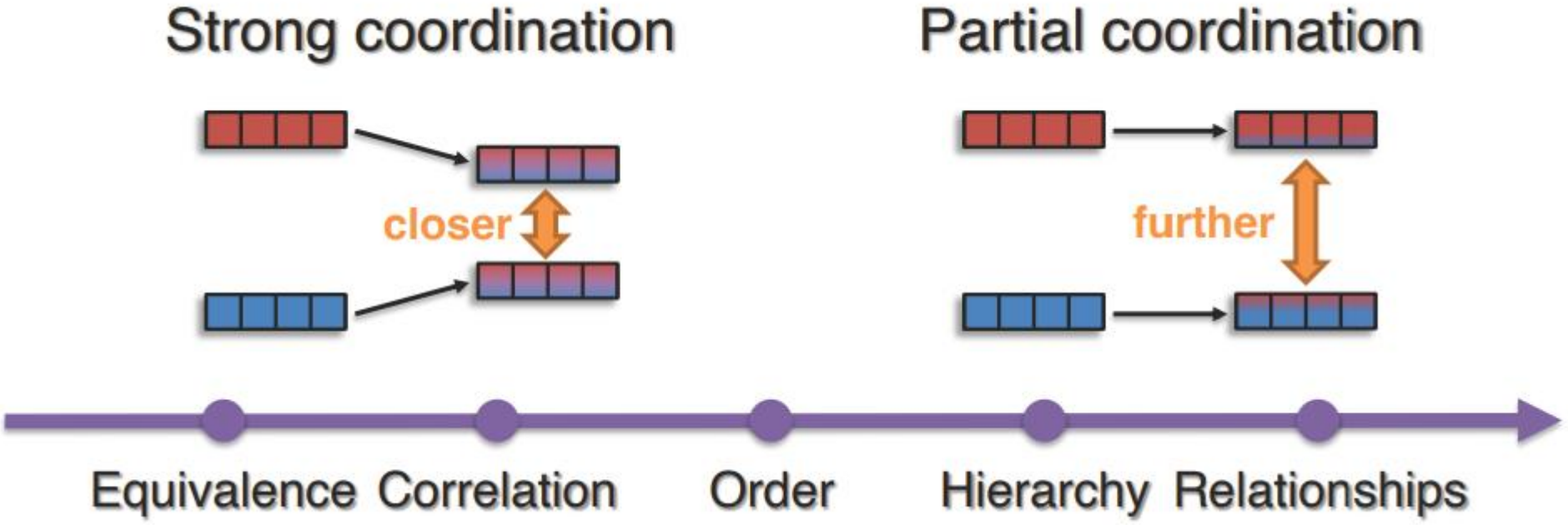
(Liang et al., 2024)

Fusion



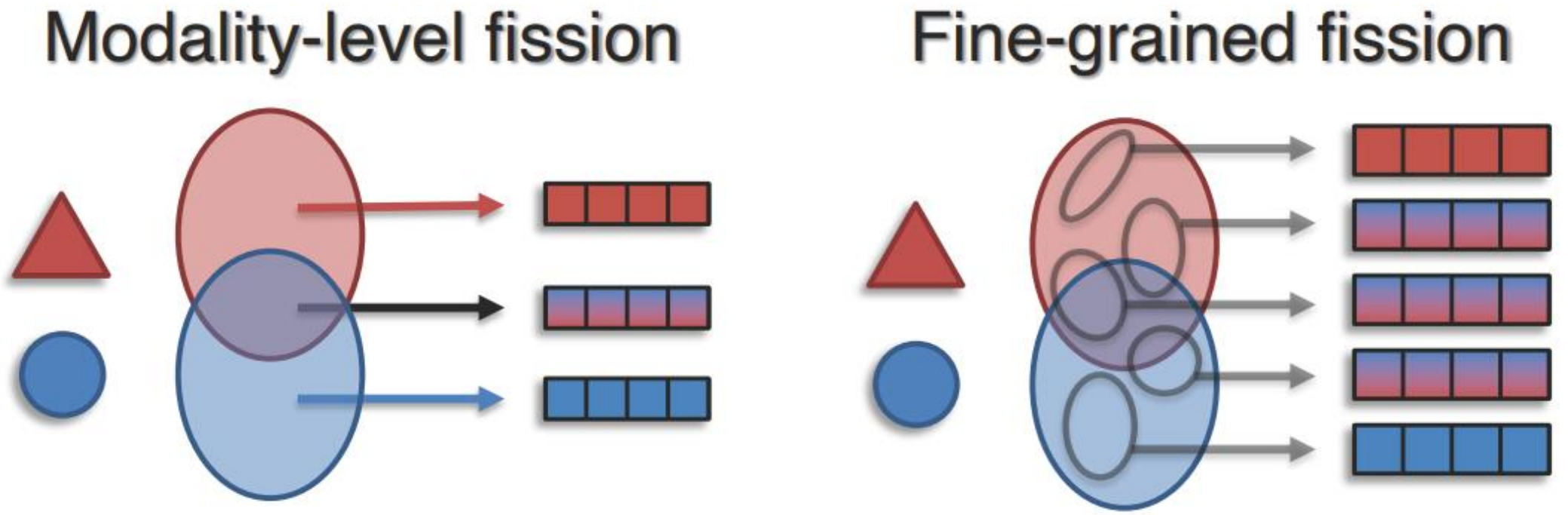
(Liang et al., 2024)

Coordination



(Liang et al., 2024)

Fission



(Liang et al., 2024)

Typical tasks

Frameworks	Applications	Modalities
Joint representation	Video classification	A + V
		A + V + T
	Event detection	A + V
		A + V + T
	Sentiment analysis	A + V + T
	Visual question answering	I + T
	Emotion recognition	A + V
		A + V + T
	Speech recognition	A + V
Coordinated representation	Cross-modal retrieval	I $\sim$ T
	Image caption	I $\sim$ T
	Cross-modal embeddings	V $\sim$ T
		I $\sim$ T
	Transfer learning	I $\sim$ T
Encoder-decoder	Image caption	I $\rightarrow$ T
	Video description	V $\rightarrow$ T
	Text to image synthesis	T $\rightarrow$ I

(Guo et al., 2019)

Document understanding

THE

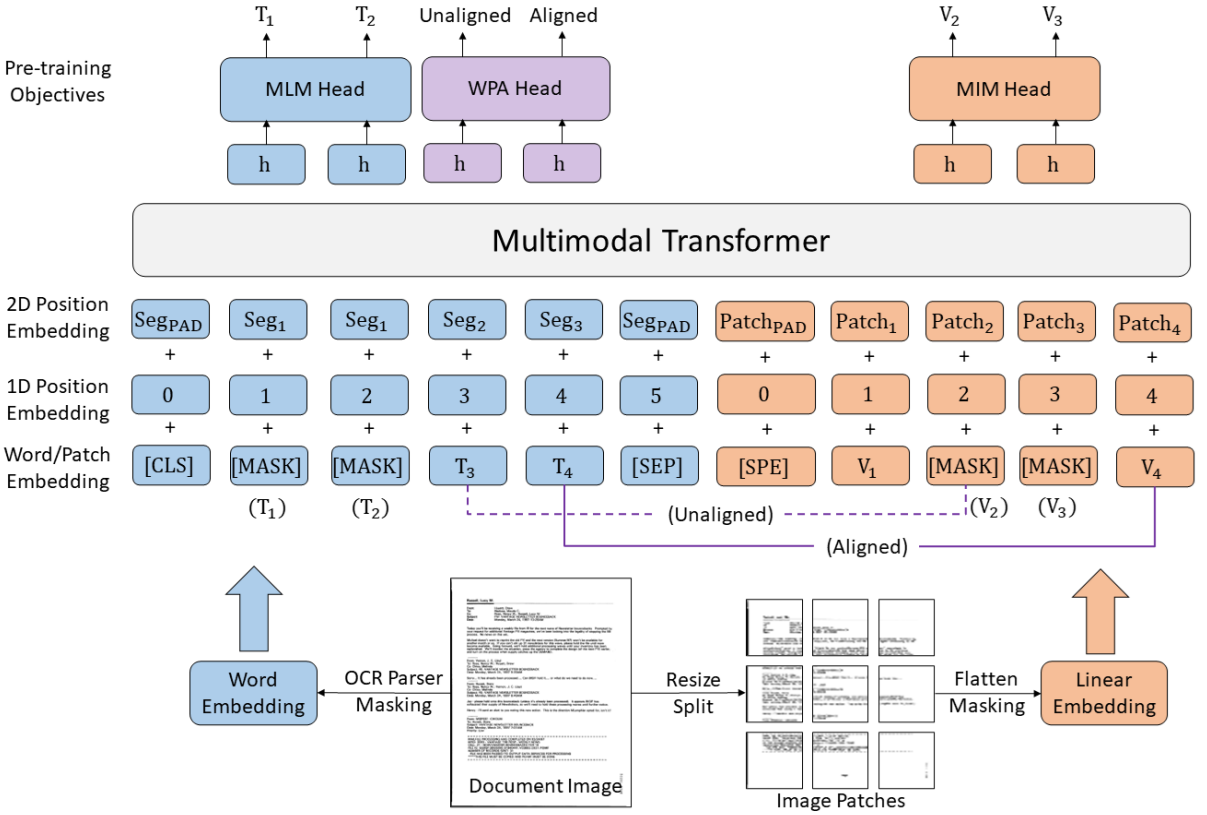
CHARTER *of* FORESTS (*a*)

Granted by King JOHN *to his Subjects in the Year* 1215 *.

JOHN, by the Grace of God, King of *England*, &c. Know ye, that for the Honour of God, and the Health of our Soul, and the Souls of our Anceſtors and Succeſſors, and for the Exaltation of *Holy-Church*, and for the Reformation of our Kingdom, We have of our free and good Will given and granted for Us and our Heirs, theſe *Liberties* hereafter ſpecified, to be had and obſerv'd in our Kingdom of *England* for ever.

(*a*) The *Foreſts* belong'd originally to the Crown, and the Kings had granted ſeveral Parts and Parcels to private Men, who had grubb'd them up and made them Arable, or Paſture. But yet all that was thus grubb'd was ſtill call'd *Foreſt*. Theſe Foreſts belonging to the King as his own *Demeſns*, or as the Sovereign Lord, were a continual Source of vexatious Suits, as well againſt thoſe which held them of the King, as againſt the neighbouring Freemen under pretence of the Rights of the Crown.

* As it is to be found in *Matthew Paris*, p. 250.



Sources

- Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, 56(10), 1-42.
- Zhao, F., Zhang, C., & Geng, B. (2024). Deep Multimodal Data Fusion. ACM Computing Surveys, 56(9), 1-36.
- Guo, W., Wang, J., & Wang, S. (2019). Deep