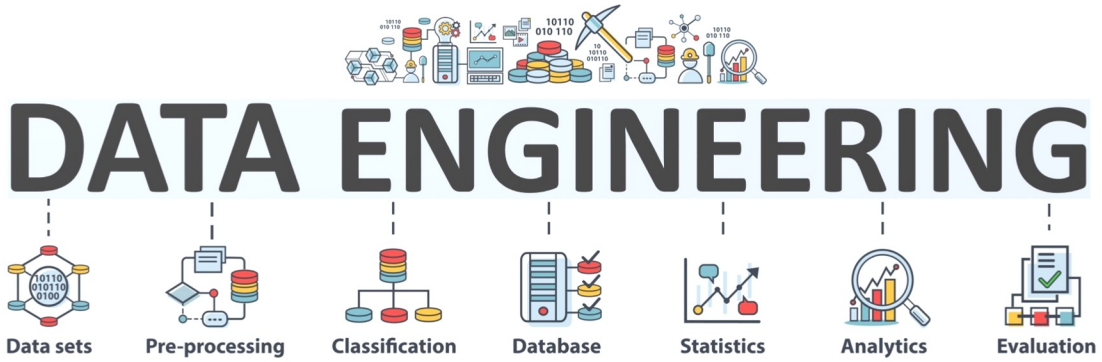


Lecture 10:

Advanced Topics and Recent Trends

CS5481 Data Engineering

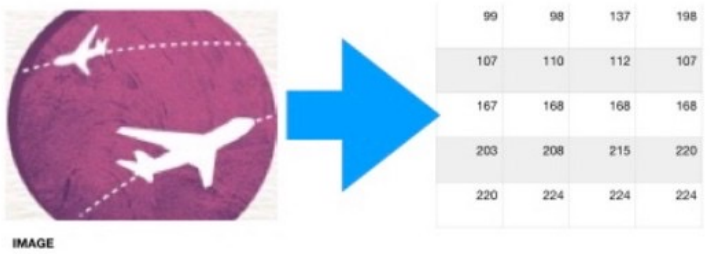
Instructor: Yifan Zhang

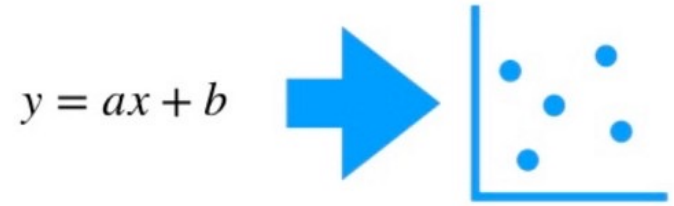
Outline

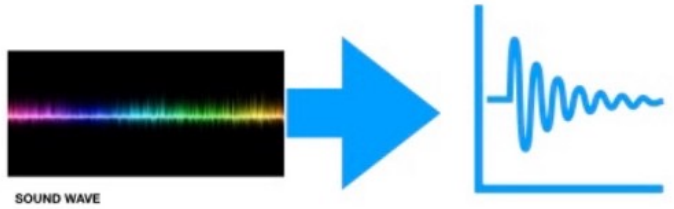
1. Representation learning and multimodal learning
2. Data bias issues
3. Explainability

What is representation learning?

- Representation learning: a set of techniques that allows a system to automatically discover the representations needed for feature detection, clustering, or classification, etc. from raw data.
- Representations are features in embedding (or latent, representation) space.
- The very hot technique, deep learning, is a non-linear method to learn representations of images, texts, etc.

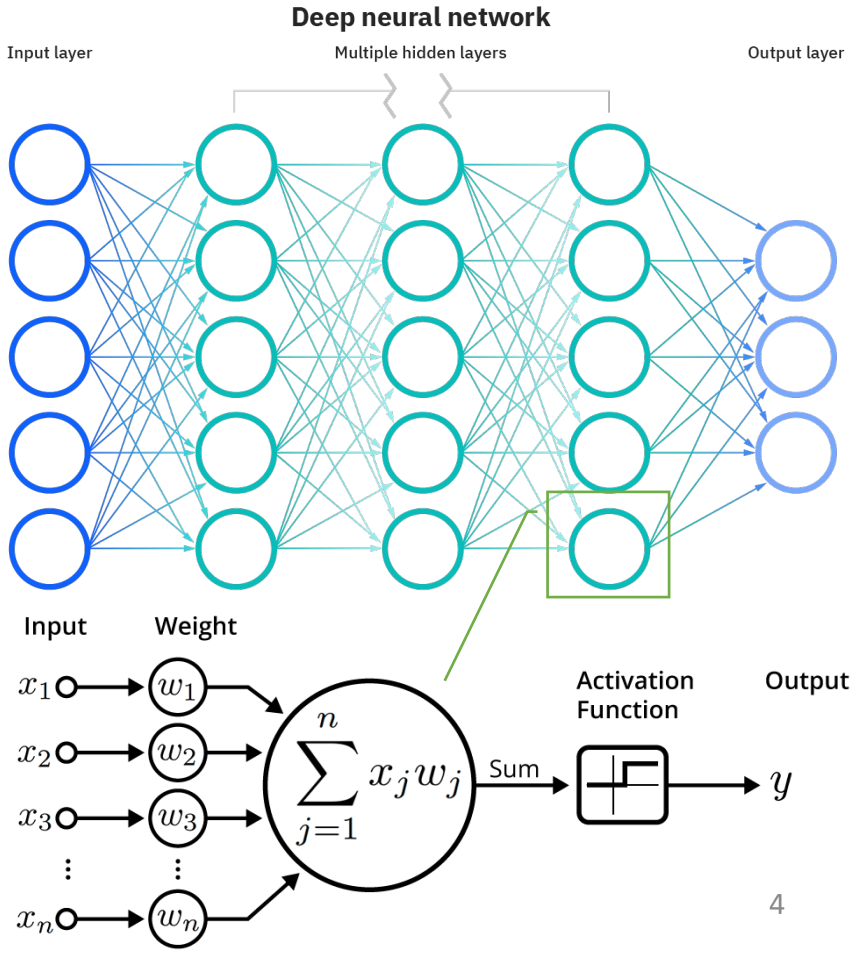






Deep learning for representation learning (1)

- **Deep Learning** is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks, usually with more layers**.**
- **Neurons** are computing nodes.
- **Lines** connecting neurons representing weights
- Usually used for image processing, automatic speech recognition, text processing, and many other tasks.



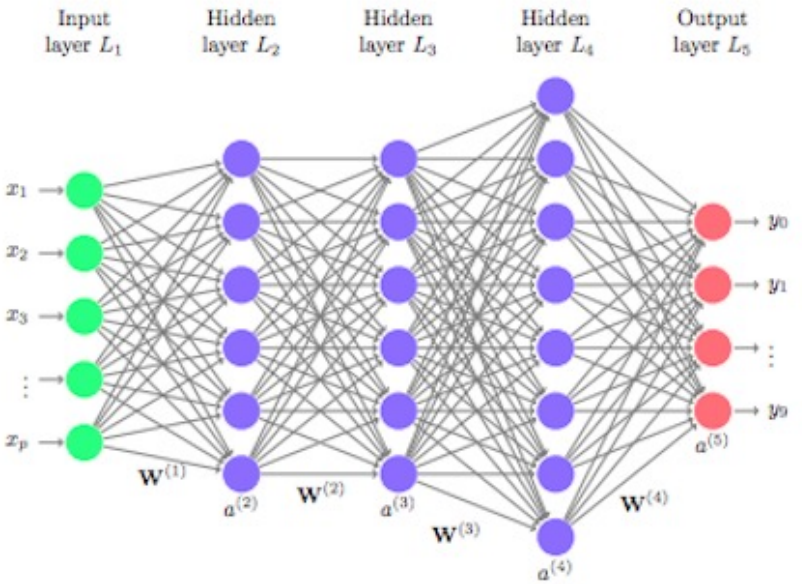
Deep learning for representation learning (2)

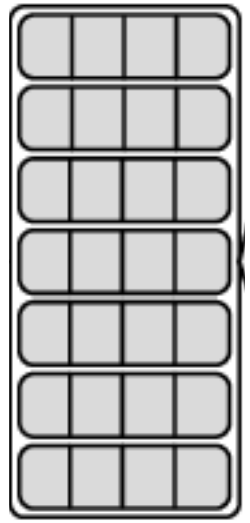
- **Deep learning maps** images, texts, speeches, etc. **into vectors** in representation spaces. If downstream tasks are trained together, then it is end-to-end.

Image pixels

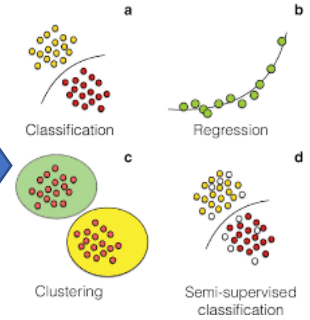
Speech signal waves

Text tokens (words)





Representations (latent vectors)



Downstream tasks

Deep learning for representation learning (3)

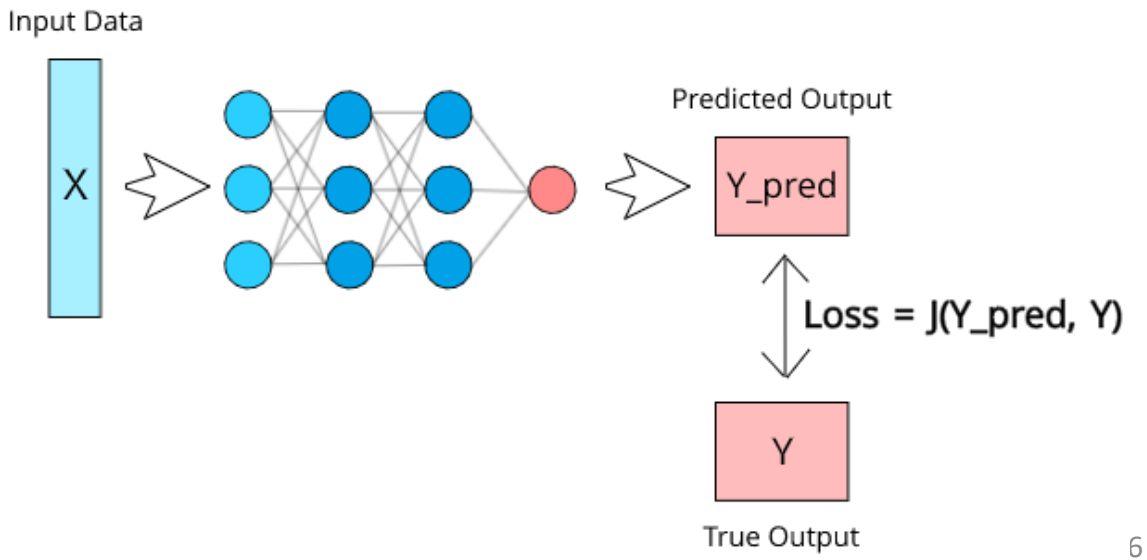
- **Loss function** is important to guide the training of neural networks.
- Commonly used loss functions: Mean Squared Error (MSE), Mean Absolute Error (MAE), Huber Loss, Cross-Entropy Loss

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|y^{(i)} - \hat{y}^{(i)}|$$

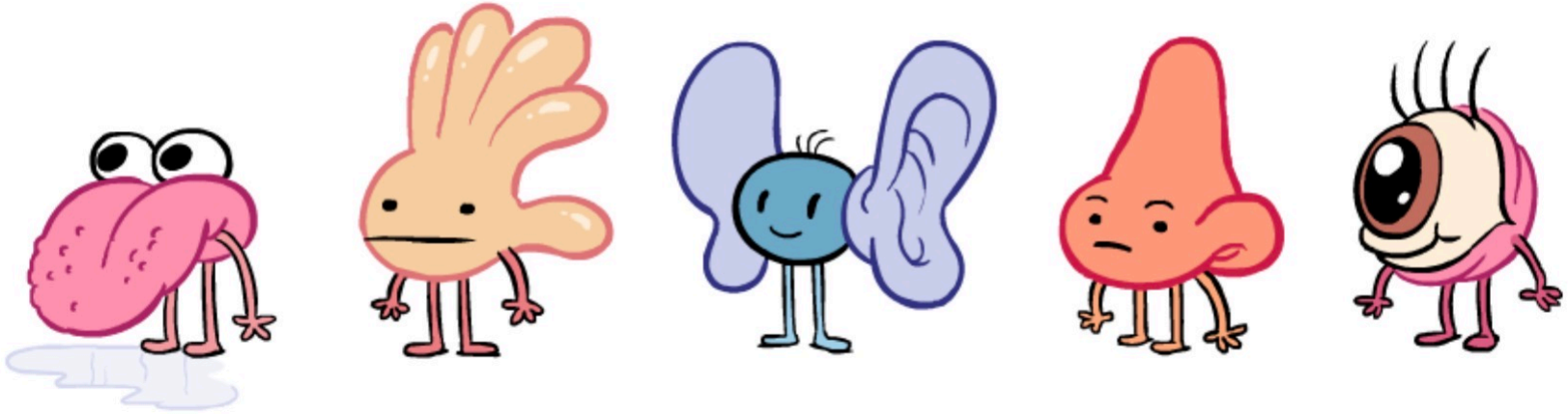
$$Huber\ Loss = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 & |y^{(i)} - \hat{y}^{(i)}| \le \delta \\ \frac{1}{n}\sum_{i=1}^{n}\delta(|y^{(i)} - \hat{y}^{(i)}| - \frac{1}{2}\delta) & |y^{(i)} - \hat{y}^{(i)}| > \delta \end{cases}$$

$$L_{\text{cross-entropy}}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_i y_i \log(\hat{y}_i)$$

- Training: stochastic gradient descent based method

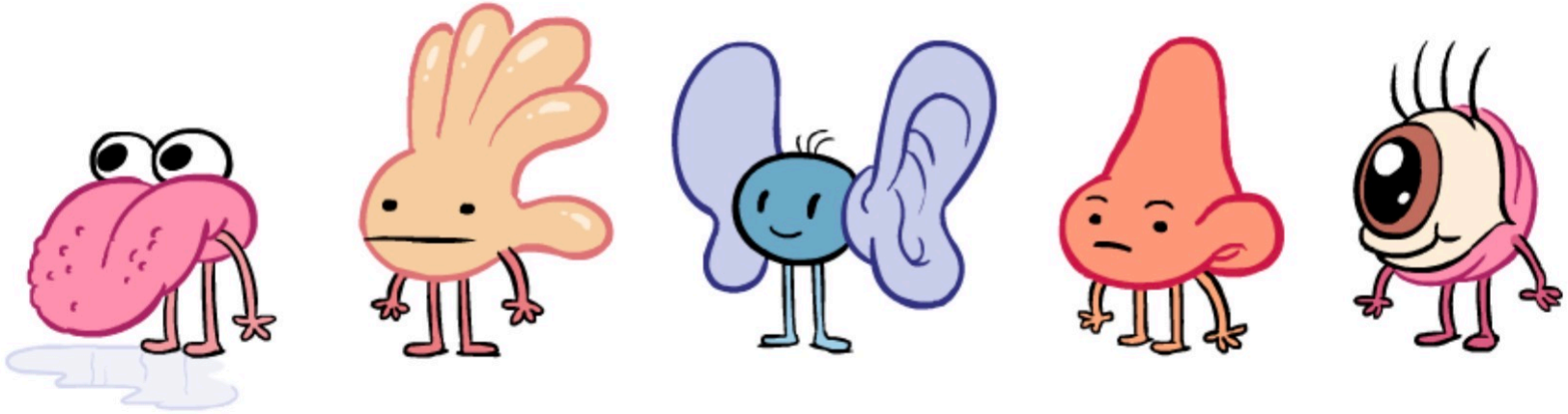


Why do we need multimodal learning?



- “Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors.”
- “In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together.
- Multimodal machine learning aims to build models that can process and relate information from multiple modalities.”

What is modality?



The way in which something happens or is experienced.

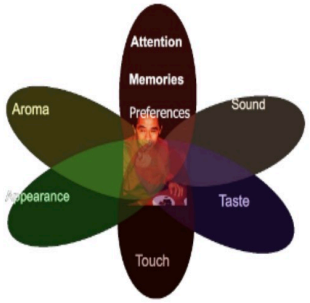
- Modality refers to a certain type of information and/or the representation format in which information is stored.
- Sensory modality: one of the primary forms of sensation, as vision or touch; channel of communication.

Multiple modalities

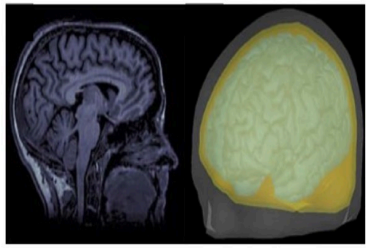
There are multiple types of modalities.

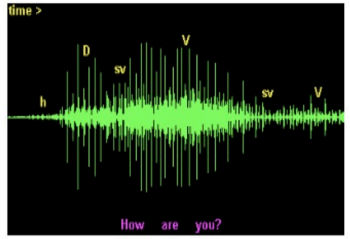
Natural language (both spoken or written)
Visual (from images or videos)
Auditory (including voice, sounds and music)
Haptics / touch
Smell, taste and self-motion
Physiological signals
- Electrocardiogram (ECG), skin conductance

Other modalities
- Infrared images, depth images, fMRI

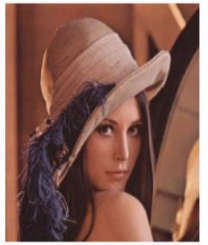


Psychology


Medical

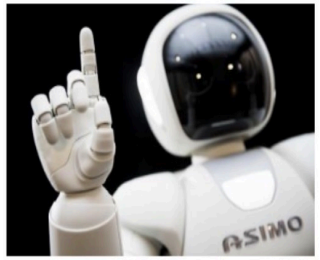


Speech

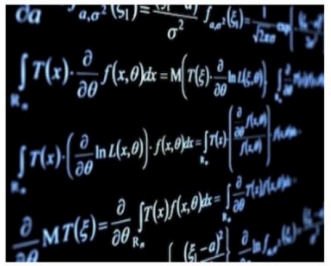

Vision

Language


Multimedia



Robotics

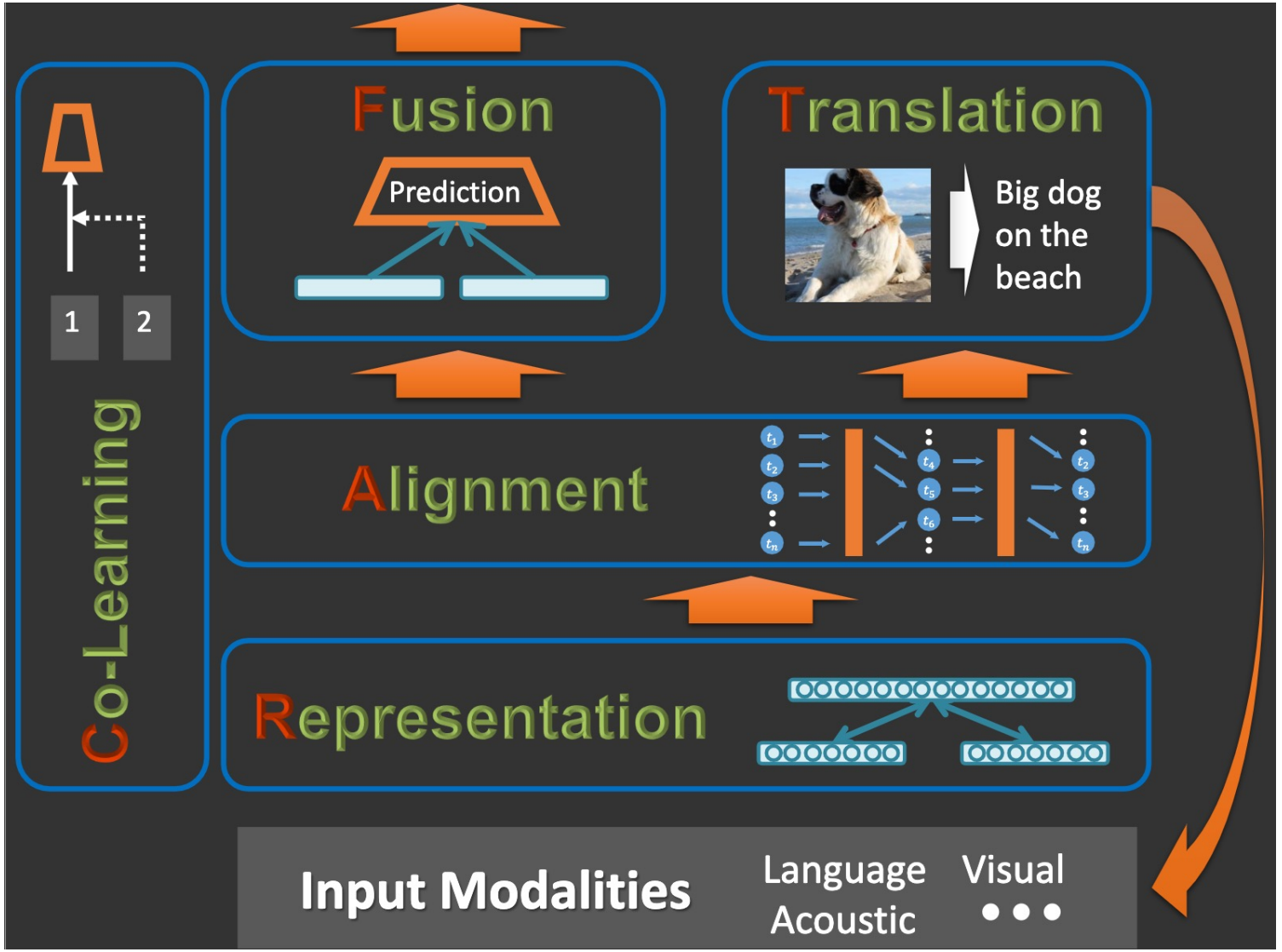

Learning

Real-world examples

- Affect recognition
  - Emotion
  - Persuasion
  - Personality traits
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

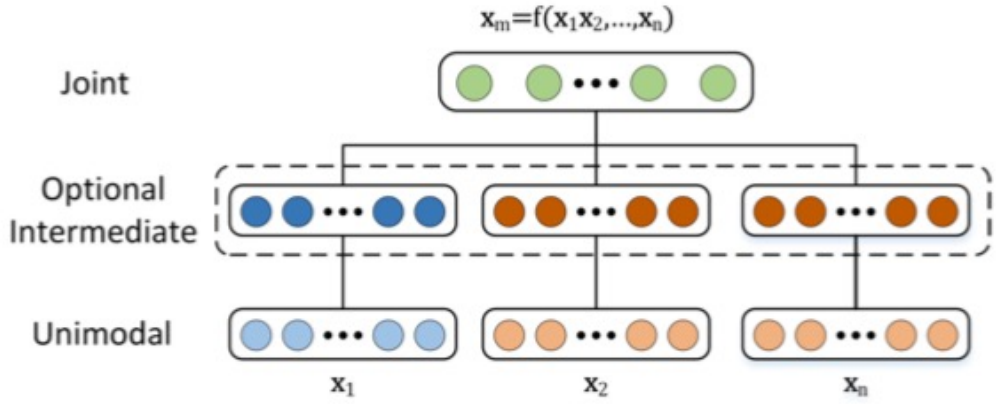
Core technical parts

- Representation
- Alignment
- Fusion
- Translation
- Co-learning

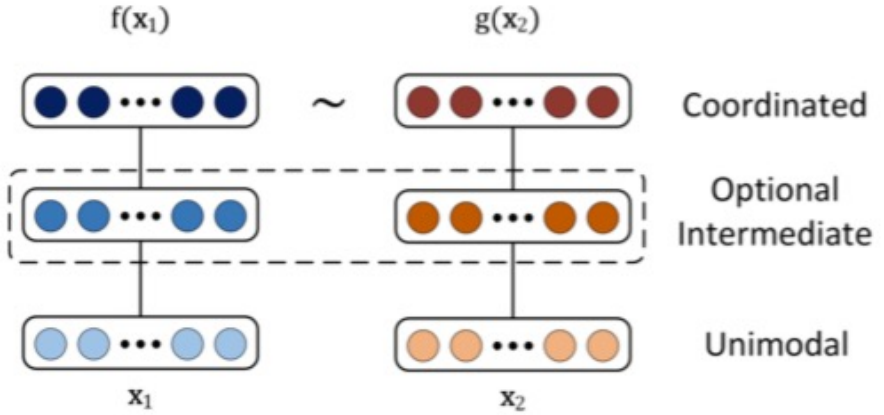


Multimodal representation

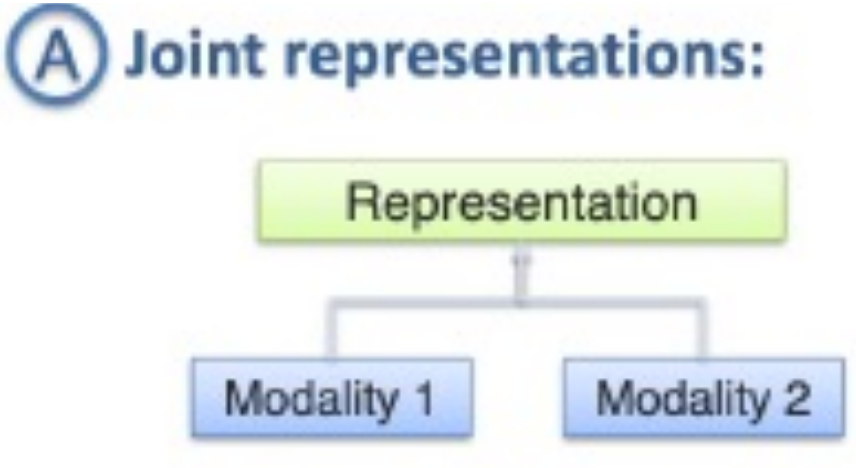
- **Learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy.**

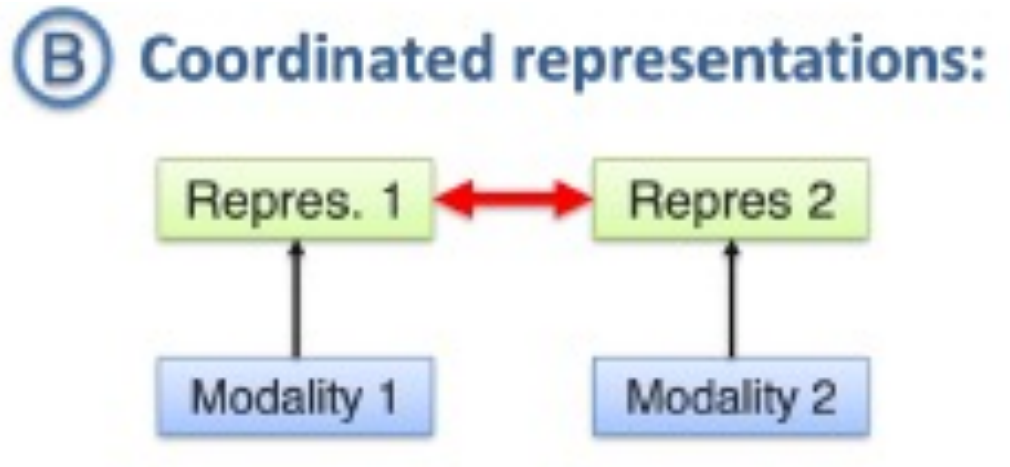


(a) Joint representation



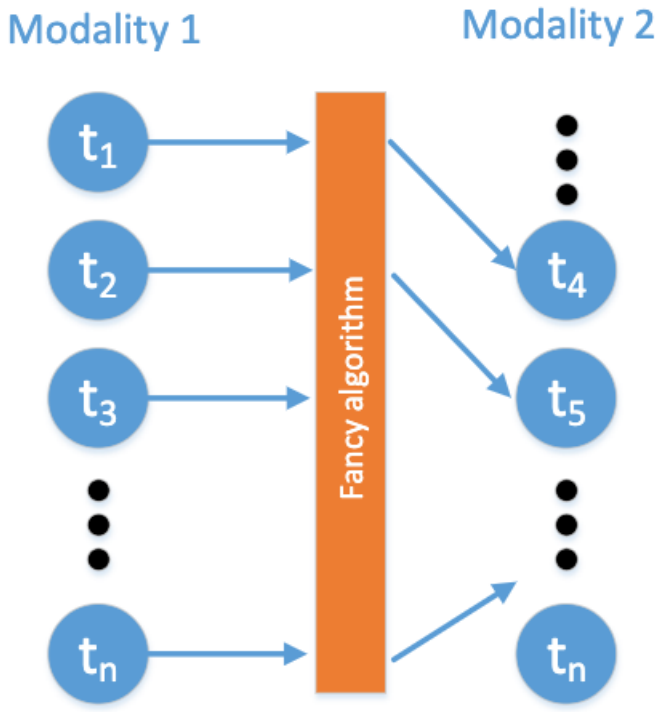
(b) Coordinated representations





Multimodal alignment

- **Identify the direct relations between sub-elements from two or more different modalities**



A Explicit Alignment

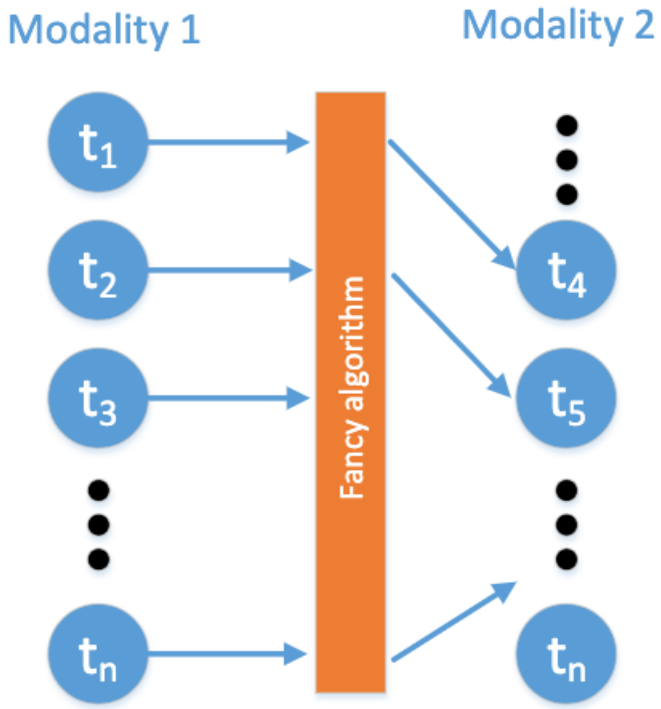
The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

Uses internally latent alignment of modalities in order to better solve a different problem

Multimodal alignment

- **Identify the direct relations between sub-elements from two or more different modalities**



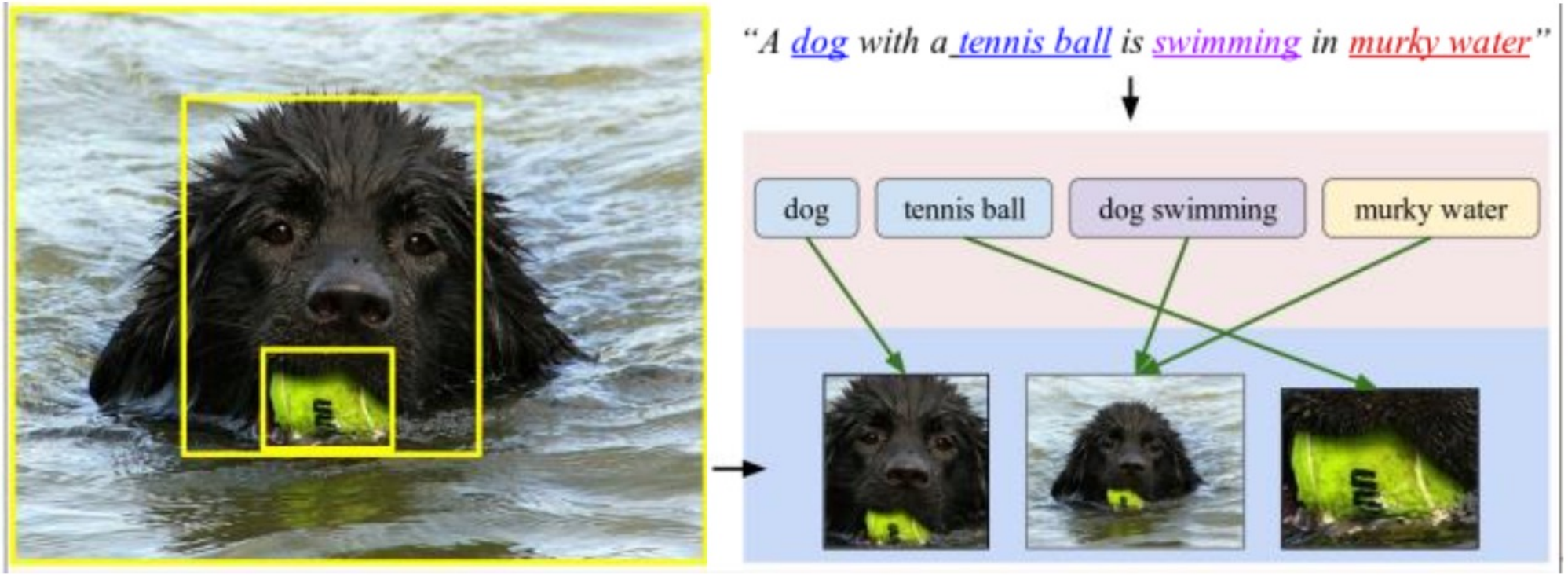
A Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

Uses internally latent alignment of modalities in order to better solve a different problem

Multimodal alignment example



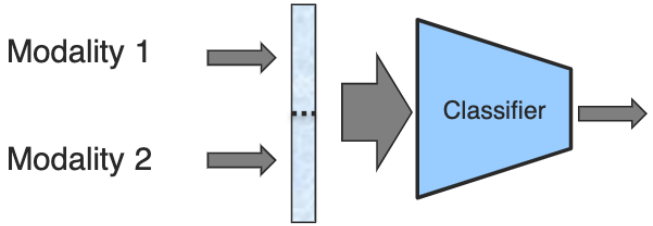
Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

Multimodal fusion

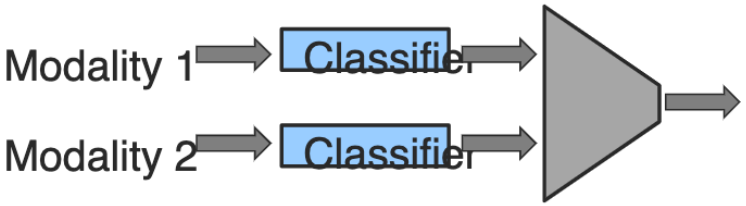
- To join information from two or more modalities to perform a task, e.g., prediction, classification.

A Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion

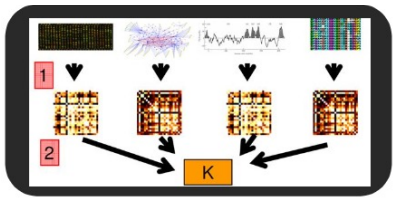


B Model-Based (Intermediate) Approaches

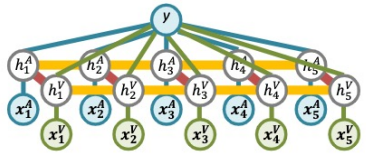
1) Deep neural networks

2) Kernel-based methods

3) Graphical models

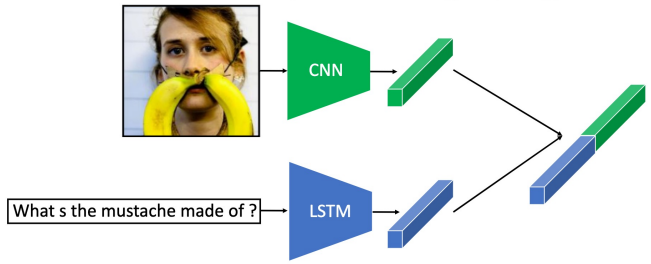


Multiple kernel learning



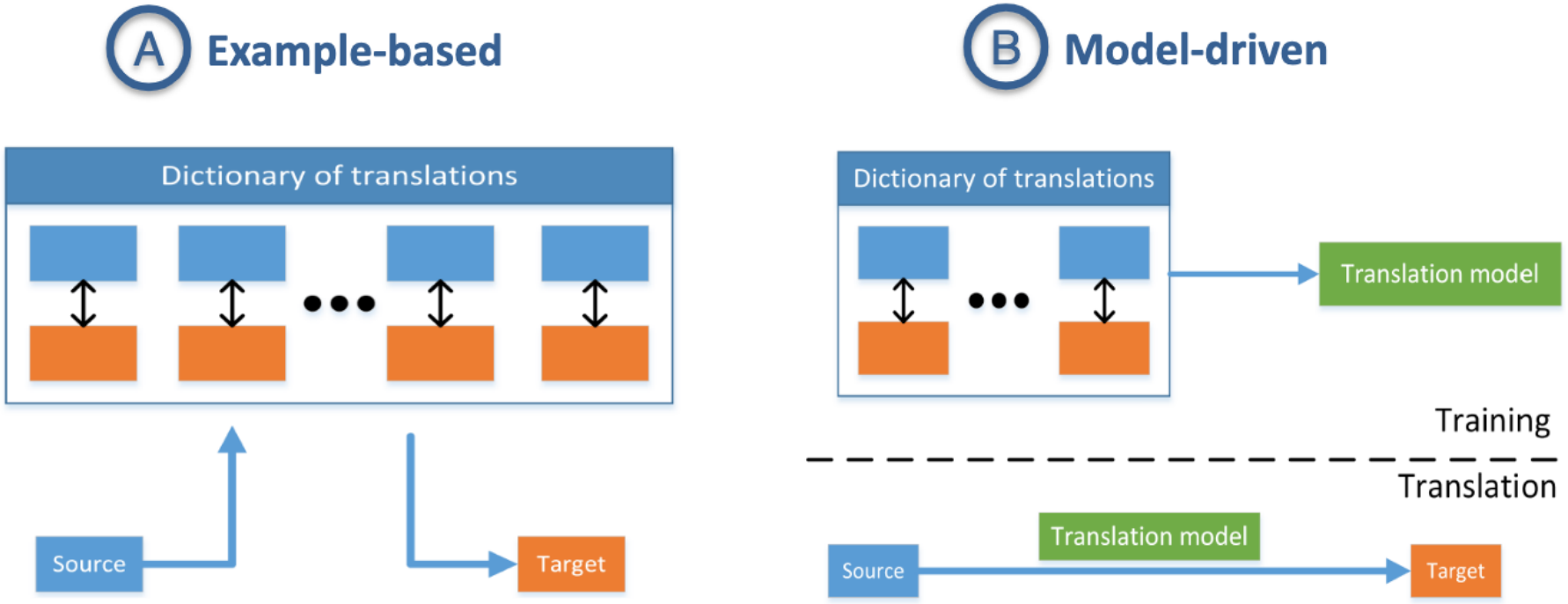
Multi-View Hidden CRF

$$\mathbf{y} = MLP([\mathbf{x}^1, \mathbf{x}^2])$$



Multimodal translation

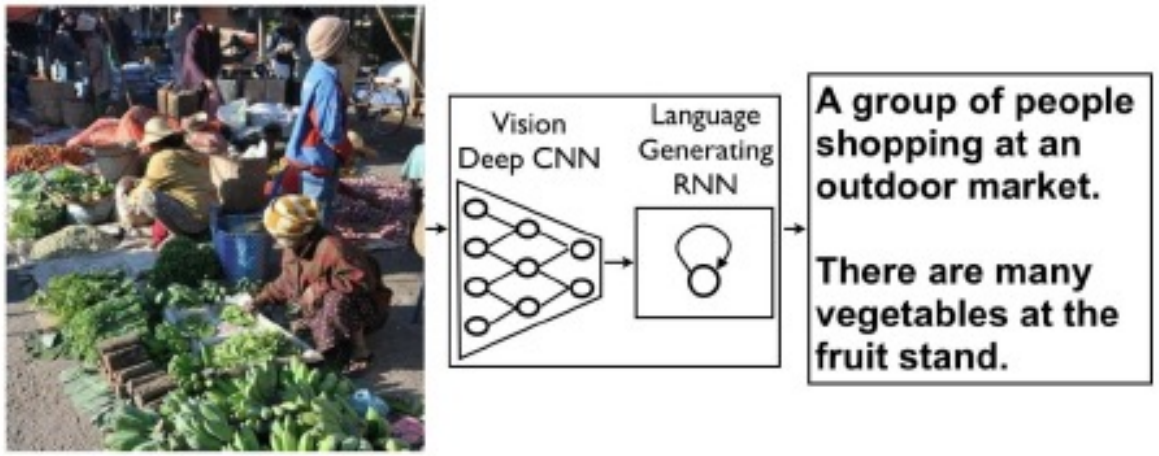
- Process of **changing data from one modality to another**, where the translation relationship can often be open-ended or subjective.



Multimodal translation example (1)

Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013

Translating transcriptions and audio streams to visual gestures



[Vinyals et al., "Show and Tell: A Neural Image Caption Generator", CVPR 2015]

Translating images to texts (captions)

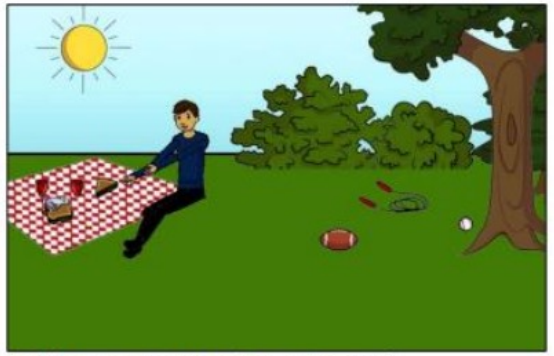
Multimodal translation example (2)

Visual Question Answering

- A very new and exciting task created in part to address evaluation problems with the above task
- Task - Given an image and a question answer the question (http://www.visualqa.org/)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

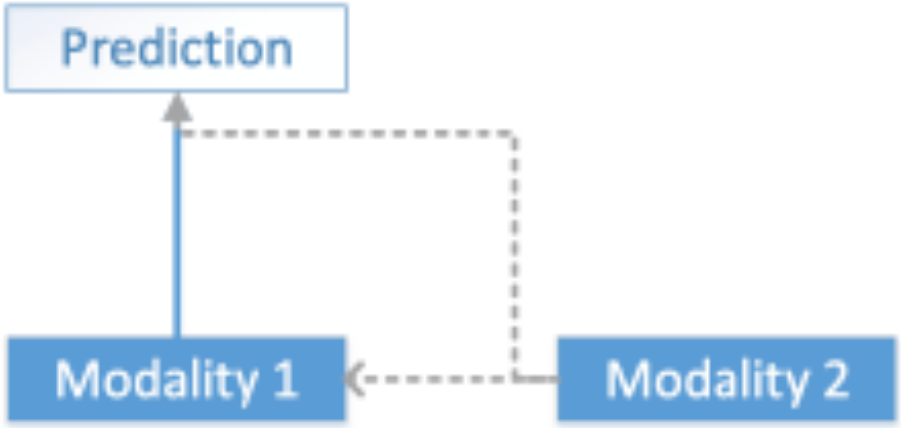


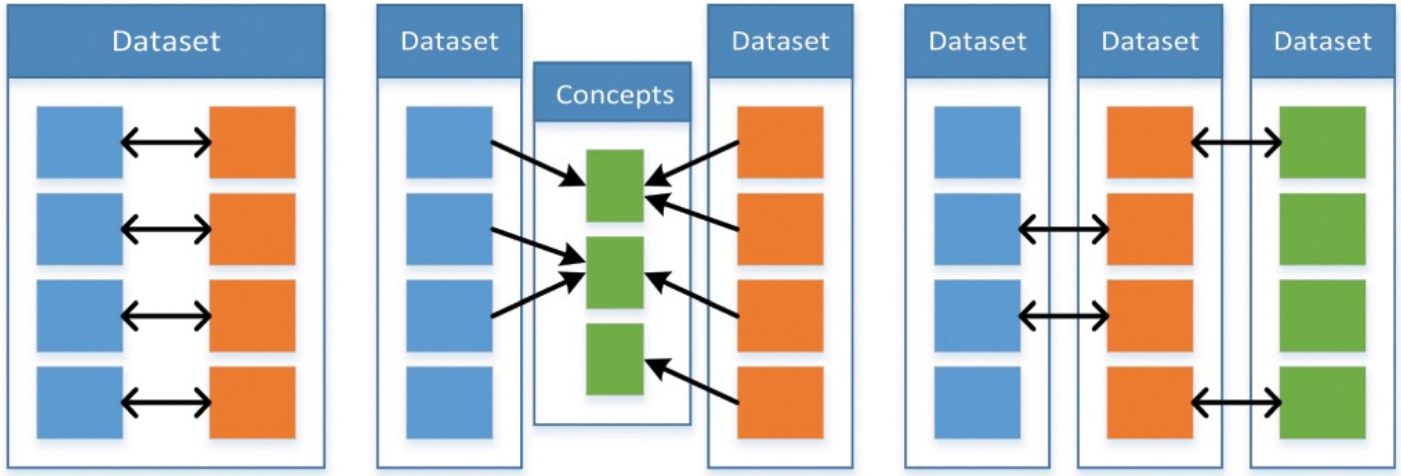
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Multimodal co-learning

- Transfer learning knowledge between modalities, including their representations and predictive models. Three types: parallel, non-parallel, hybrid.





(a) Parallel (b) Non-parallel (c) Hybrid

Data bias may result in discrimination (1)

- Unbalanced data.
  - Unfair to users with specific features

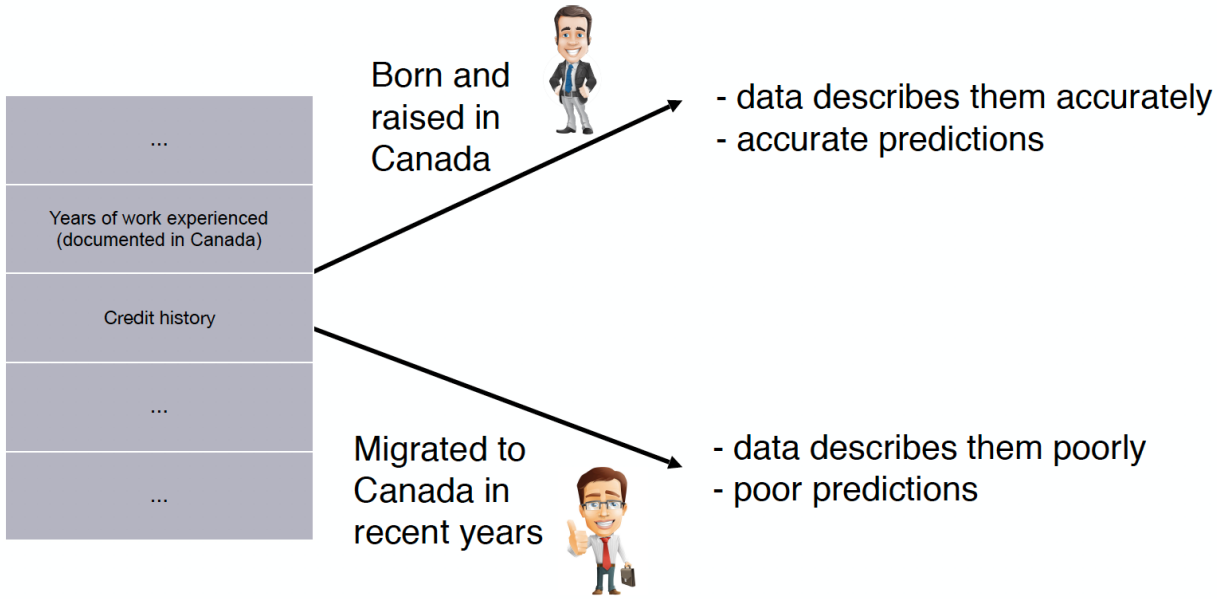
They both apply for a loan with a high amount

Lots of data about similar (male) applicants

no data about similar (female) applicants

Data bias may result in discrimination (2)

- Missing attributes
  - Overall accuracy is still high, but unfair to a small group of population



A hypothetical (extreme) situation:

Born and raised in Canada

- data describes them accurately
- accurate predictions (95% accurate)

90% of population

The model is still 90% accurate!

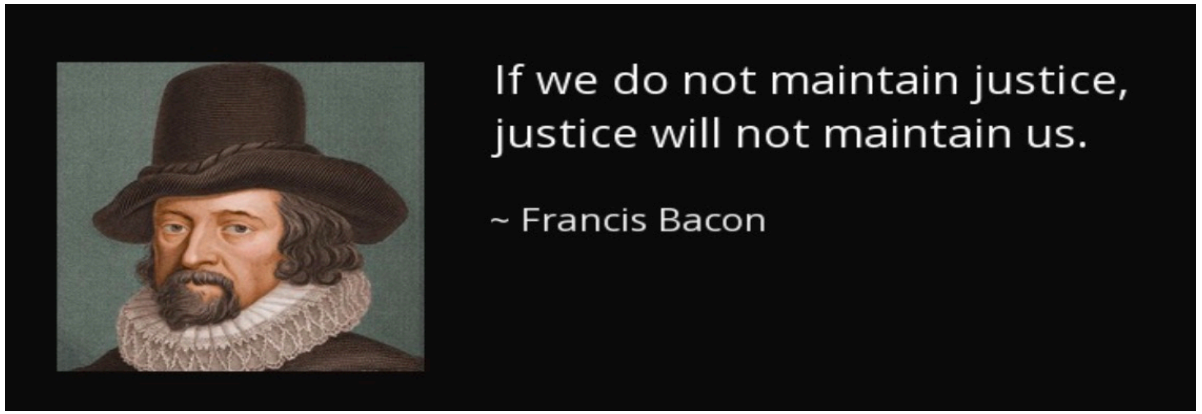
Migrated to Canada in recent years

- data describes them poorly
- poor predictions (50% accurate)

10% of population

Why we should care about fairness?

- Basic social principle!
- Law Against Discrimination!



Legally recognized 'protected classes'

Race (Civil Rights Act of 1964)
Color (Civil Rights Act of 1964)
Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
Religion (Civil Rights Act of 1964)
National origin (Civil Rights Act of 1964)
Citizenship (Immigration Reform and Control Act)
Age (Age Discrimination in Employment Act of 1967)
Pregnancy (Pregnancy Discrimination Act)
Familial status (Civil Rights Act of 1968)
Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Regulated domains

Credit (Equal Credit Opportunity Act)
Education (Civil Rights Act of 1964; Education Amendments of 1972)
Employment (Civil Rights Act of 1964)
Housing (Fair Housing Act)
Public Accommodation (Civil Rights Act of 1964)
Extends to marketing and advertising; not limited to final decision
This list sets aside complex web of laws that regulates the government

Themis: Greek goddess of justice, divine order, fairness, law, and custom.

How to address fairness issues in data-related applications?

- Bias is across the data engineering pipeline!

Pre-processing

Data is noisy
Biases
Encodes protected attributes

e.g.,

Discrimination Discovery
Un-bias the data
Sampling
Embedding
Dimension reduction

In-processing

Data scientists do not build the models

e.g.,

Learning subject to constraints
Ranking
Inference

Post-processing

unfair outcome
no user feedback

e.g.,

Causal discovery
Transparency & Interpretability
Verification

Why do we need explainability?

- Black-box models achieve the SOTA performance on many data-driven tasks using machine learning, especially deep learning algorithms.

- However, it is hard for people to know how the prediction are made and the base of a prediction.

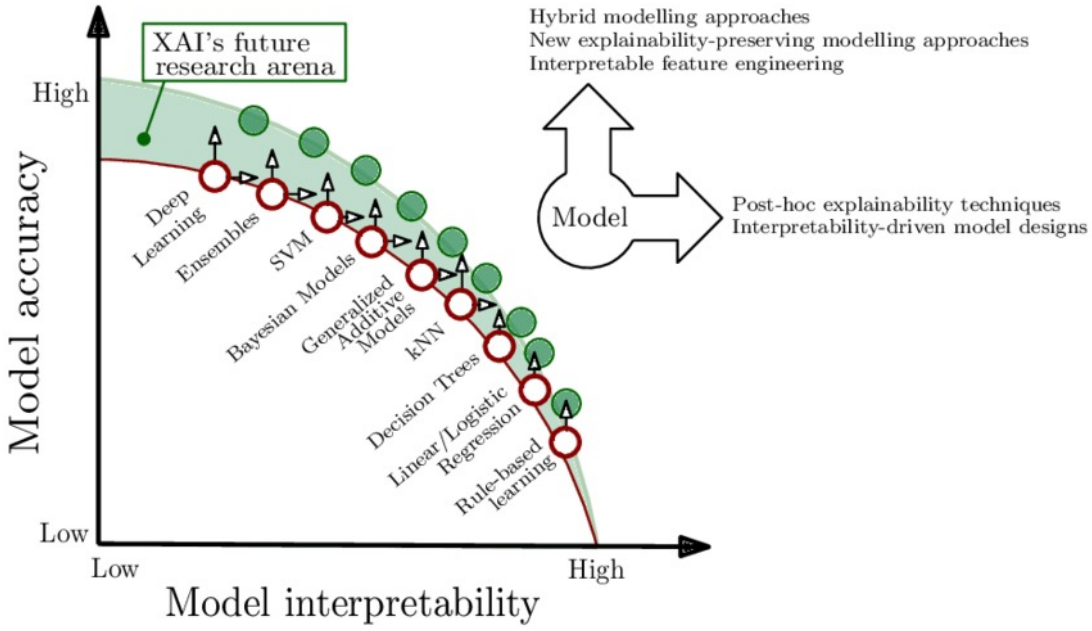
- This problem hinders the further application of these models and researchers are required to make models and their predictions more human-understandable.

Black-box vs. white-box models

- Black-box models (not easily interpretable models): deep learning approaches and the use of representation (embedding) features.

- White-box models (easily interpretable models): decision trees, rule-based models, Markov models, logistic regression, linear regression, and so on.

- Reduced interpretability comes reduced trust.



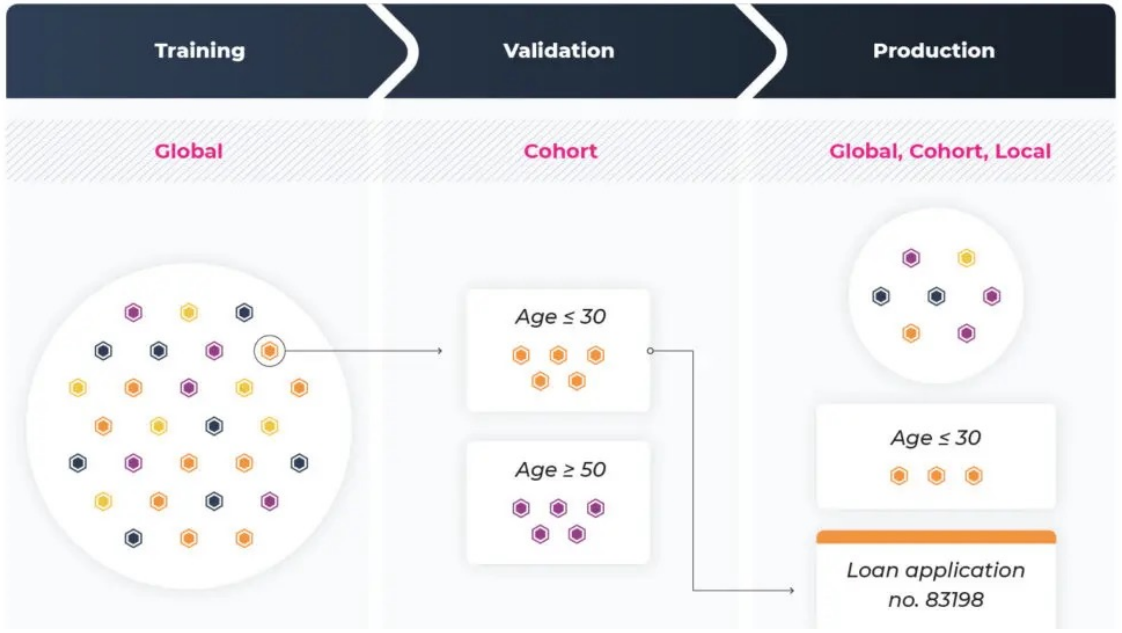
Explainability definition

- Explainability can be defined as to “understand how a model arrives at its result, also referred to as the outcome explanation problem.” If end-users are able to understand the reasoning behind the prediction, then it is assumed more trust will grow.

- Furthermore, this allows for a positive feedback loop from user to development interactions. Some characterizations of explainability include explanations for individual predictions or the model’s prediction process as a whole.

Local explainability vs. global explainability

- **Local explainability** refers to the ability of the system to tell a user why a particular decision was made.
- **Global explainability** is about the explanation of the learning algorithm and model as a whole, including the training data used, appropriate uses of the algorithms, and warnings regarding weaknesses of the algorithm and inappropriate uses. Global explainability is like a "user's manual".



Self-explainability vs. post-hoc explainability

- **Self-explainability** (intrinsic explainability) refers to machine learning models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models. In deep learning, self-explainability refers to using interpretable modules as basic building blocks to replace existing ones.

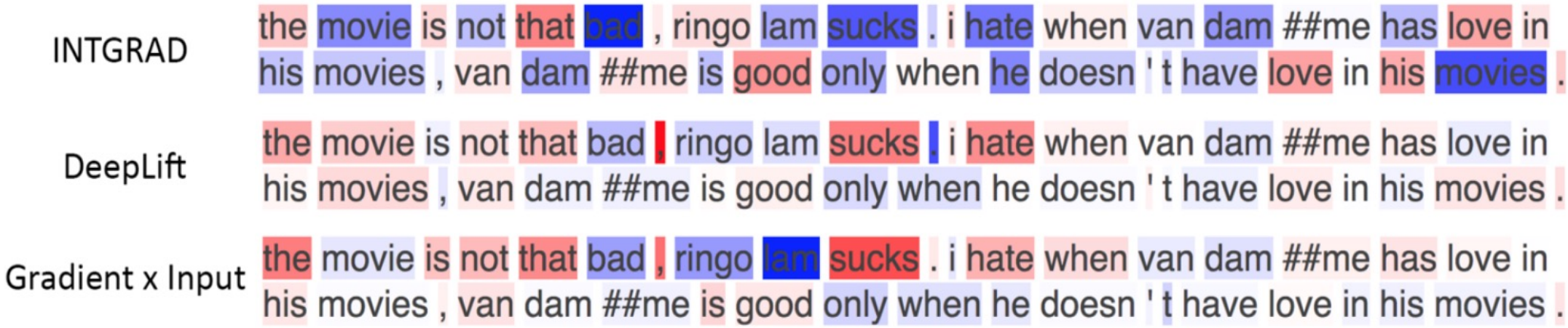
- **Post hoc explainability** refers to the application of interpretation methods after model training.

Explainability techniques: feature importance

- To derive explanation by investigating the importance scores of different features used to output the final prediction.
  - manual features obtained from feature engineering
  - latent features learned by NNs
- Feature importance metrics
  - Attention score
  - First-derivative saliency
  - Shapley value
  - Mutual information

Explanation example

In a sentiment analysis task with different explanation algorithms, the color indicates which words are more important to determine the sentiment of the sentence, to the positive or negative side.



Explainability techniques: surrogate model

- Model predictions are explained by learning a second, usually more explainable model, as a proxy.
  - E.g., LIME, an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- These proxy models are model-agnostic and can achieve both local or global explanations.
- One drawback of this method is that the learned surrogate models and the original models may have completely different mechanisms to make predictions, leading to concerns about the fidelity of surrogate model-based approaches.

LIME: surrogate model

- LIME for Local Interpretable Model-agnostic Explanations

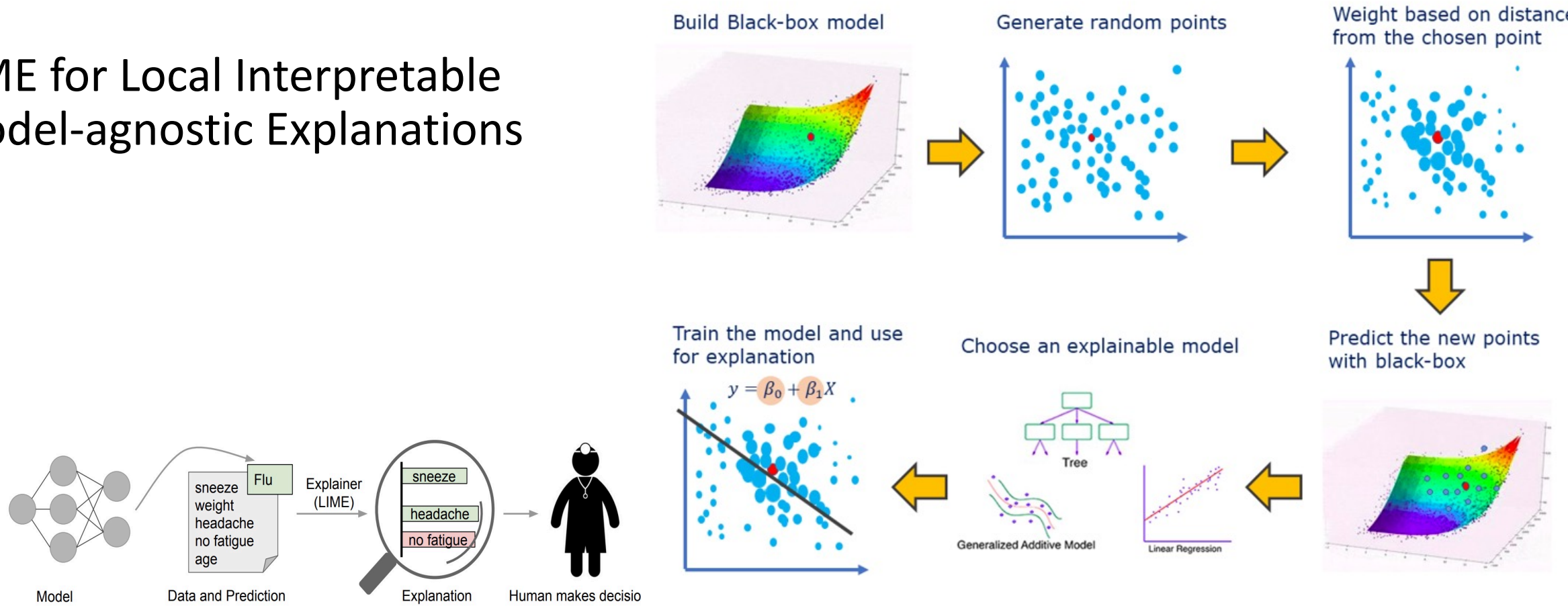


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

Open problems

- What interpretability exactly means and how to define interpretability
- How to evaluate interpreting methods
- Whose interpretability are we talking about
- Go beyond classification tasks
- The inconsistency between different interpreting methods
- The cost of model performance in exchange for interpretability
- The utility of interpretability

Thanks for your attention!

Appendix

1. Multimodal Machine Learning: A Survey and Taxonomy,

https://arxiv.org/abs/1705.09406