Multimodal AI

Lecture 2.2 – Data & Heterogeneity

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

mit media lab

MIT EECS

Assignments for This Coming Week

HW1: reading assignment + homework due next Tuesday 2/17

For project:

- Project proposal instructions released, due Tuesday (2/24). Submit on canvas
- Meet with me 4-5pm if need feedback about proposal ideas.

Clarification: 2-4 students per group. Listeners can help out on projects and will not be counted in the group size.

No 1 student groups, unless exceptional circumstances.

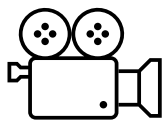
Lecture Outline

1. Vision, language, audio, sensing, set, graph modalities
2. Modality profile
3. Types of data and labels
4. Common learning objectives and generalization

What is a Sensory Modality?

Sensory 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:

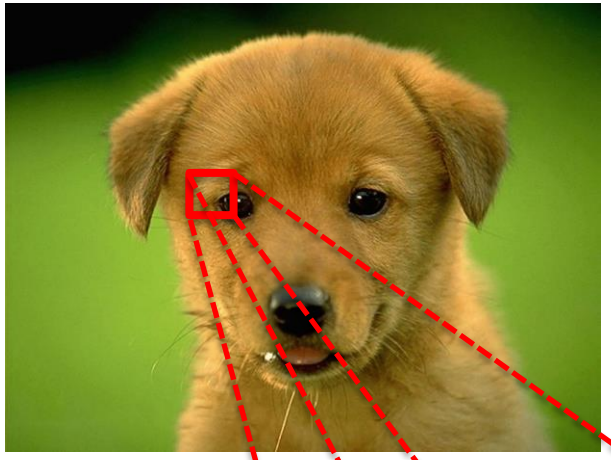
Speech signal	Language	Sentiment intensity
Image	Detected objects	Object categories

Most of AI is about learning abstractions, or representations, from data.

multisensory intelligence

Visual Modality

Color image



88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

Input observation x_i

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Binary classification problem

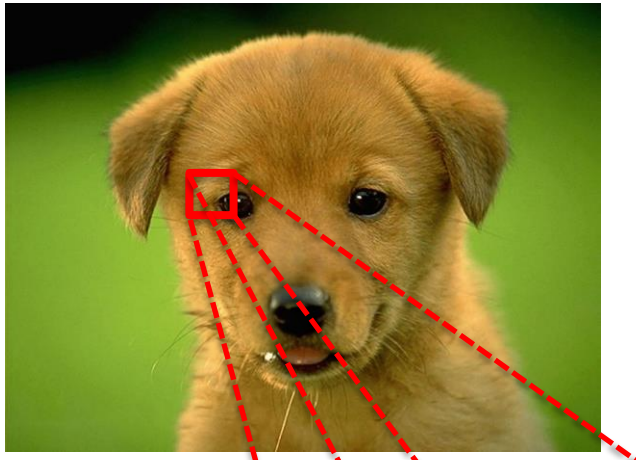
Dog ?

label $y_i \in \mathcal{Y} = \{0,1\}$

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Visual Modality



Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation x_i

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-class classification problem

Duck
-or-
Cat
-or-
Dog
-or-
Pig
-or-
Bird ?

label $y_i \in \mathcal{Y} = \{0,1,2,3, \dots\}$

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Language Modality

Written language

★★★★★ **Masterful!**

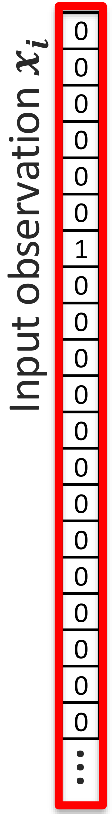
By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

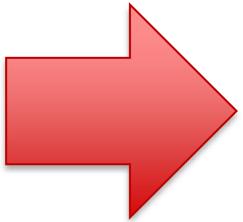
Spoken language

```
              MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

              CLARK
But you've only put up the good
things they say about me.

              MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```



“one-hot” vector

$|x_i|$ = number of words in dictionary



Word-level classification

Part-of-speech ?
(noun, verb,…)

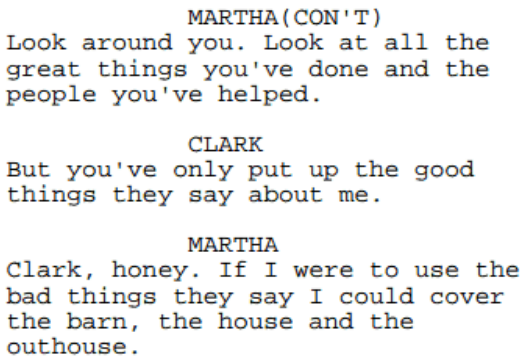
Sentiment ?
(positive or negative)

Named entity ?
(names of person,…)

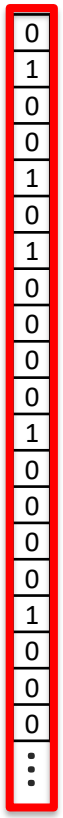
Language Modality

Written language

Spoken language



Input observation x_i



“bag-of-word” vector

$|x_i|$ = number of words in dictionary

Document-level classification

Sentiment ?
(positive or negative)

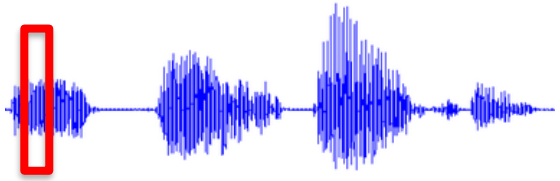
Response?

What happens with word ordering?

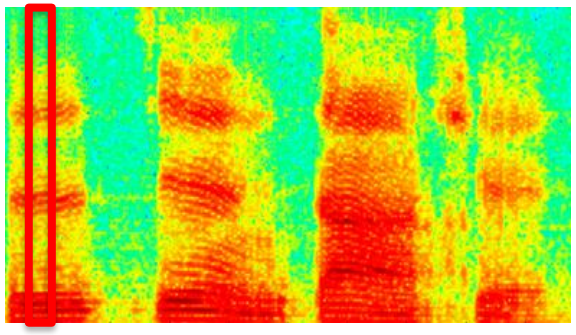
multisensory intelligence

Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectrogram

Input observation x_i

0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02

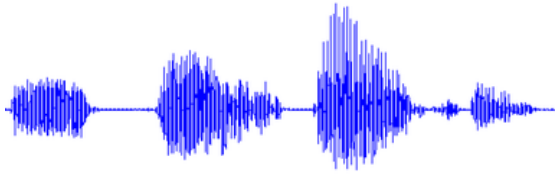
Spoken word ?

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

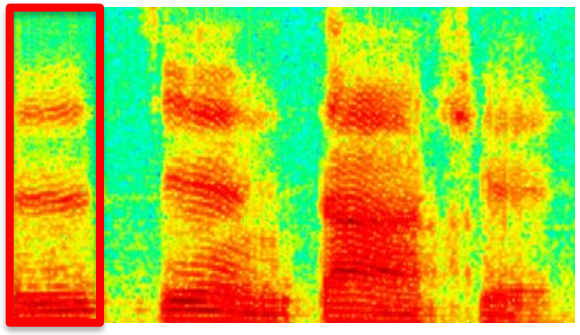
multisensory intelligence

Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectrogram

Input observation x_i

0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02
0.24
0.26
0.58
0.9
0.99
0.79
0.45
0.34
0.24
...

Emotion ?

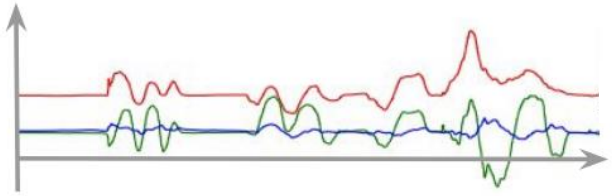
Spoken word ?

Voice quality ?

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Sensor Modality



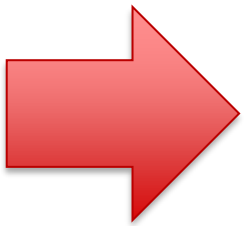
Force-Torque Sensor

Time series data across six-axis Force-Torque sensor:
T × 6 signal.

Proprioception

Measure values internal to the system (robot); e.g. motor speed, wheel load, **robot arm joint angles**, battery voltage.

Time series data across current position and velocity of the end-effector:
T × 2 signal.

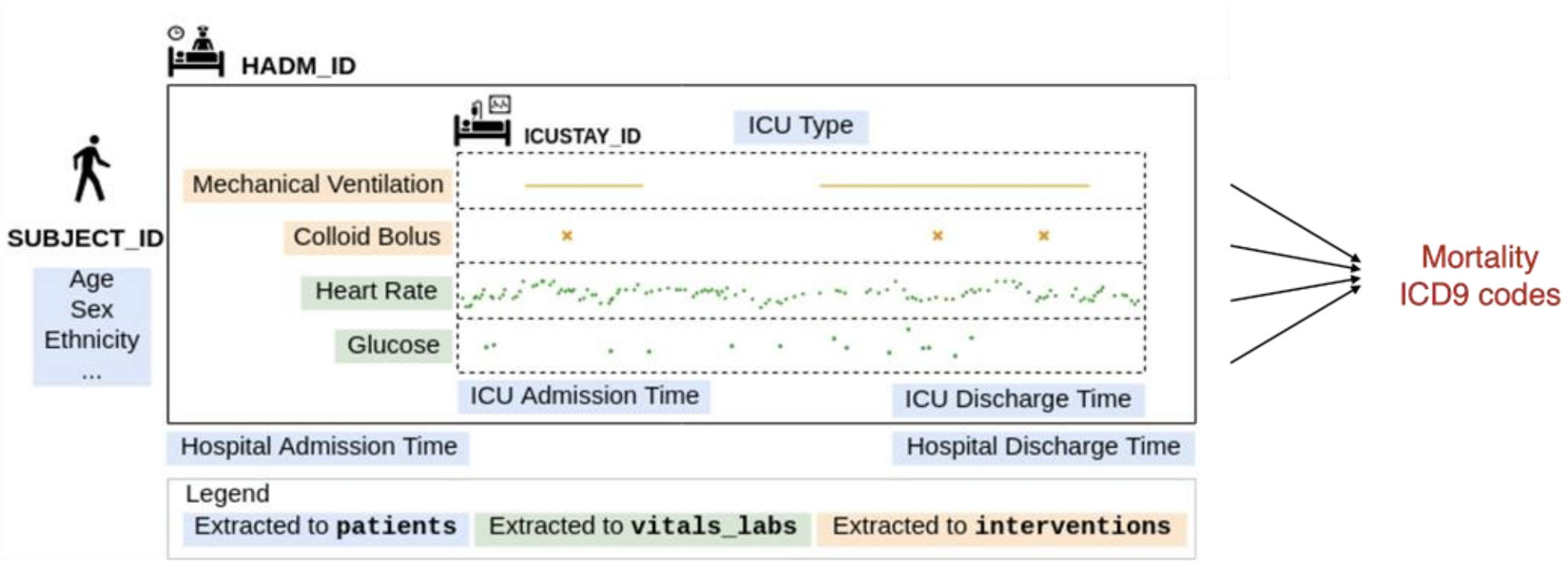


Object property
Next action

[Lee et al., Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. ICRA 2019]

multisensory intelligence

Tabular Modality

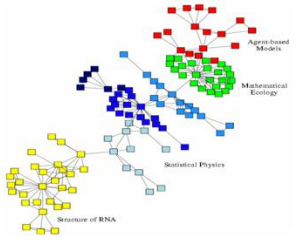


[Johnson et al., MIMIC-IV: A Freely Accessible Electronic Health Record Dataset. Scientific Data 2023]

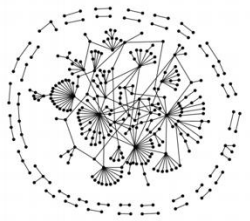
multisensory intelligence

Graph Modality

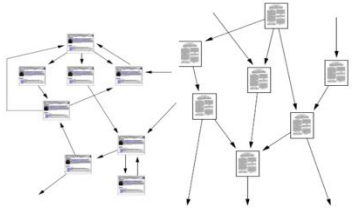
Social networks



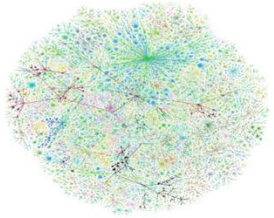
Economic networks



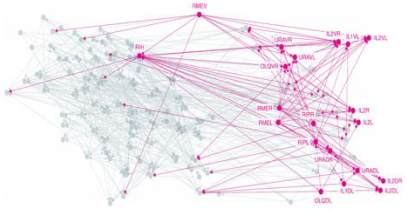
Biomedical networks



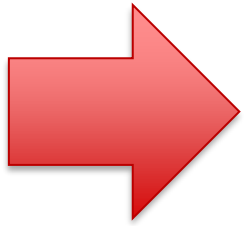
Information networks:
Web & citations



Internet



Networks of neurons



Tasks on graphs:
Node classification
Link prediction
…

Using graphs:
Knowledge graphs for QA
Social network for sentiment analysis
…

[Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

multisensory intelligence

Set Modality

Sets

Point clouds

Set anomaly detection
Set expansion
Set completion
Point cloud classification
Point cloud generation

[Zaheer et al., DeepSets. NeurIPS 2017 and Li et al., Point Cloud GAN. arxiv 2018]

multisensory intelligence

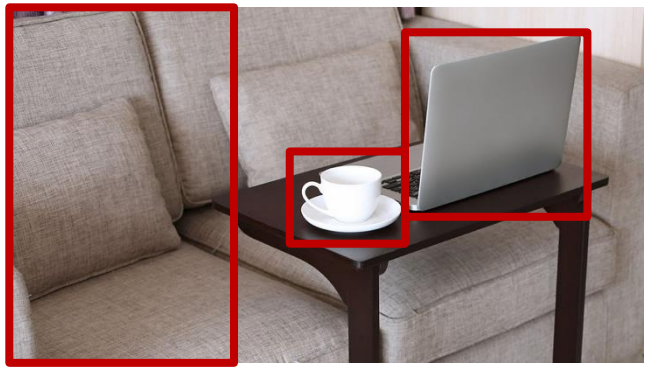
Modality Profile

The qualities and structures that are unique to a data modality.

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

Modality Profile

The distribution of individual elements within that modality.



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

1 **Distribution:** discrete or continuous, support

▲ {  ,  ,  }

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

Modality Profile

The frequency at which elements appear or are sampled.

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

2 **Granularity:** sampling rate and frequency

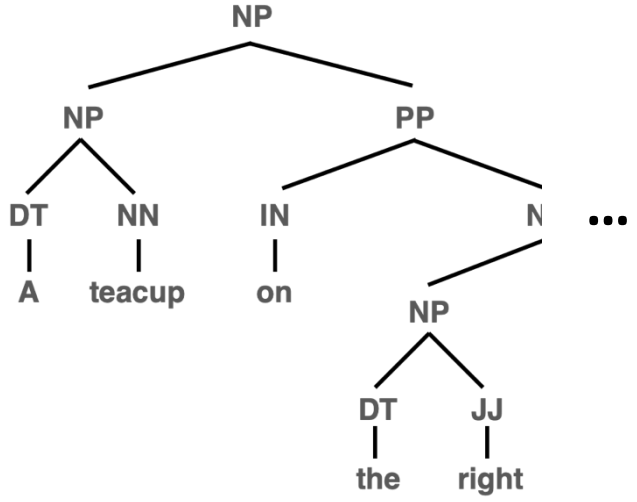
objects per image

words per minute

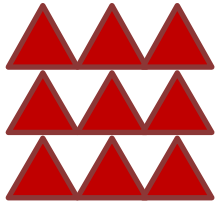
multisensory intelligence

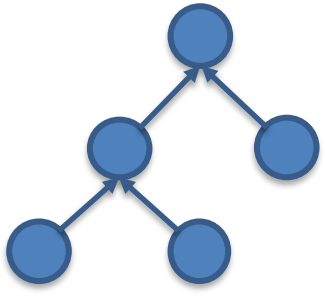
Modality Profile

The way elements compose with each other to form entire data.



3 **Structure:** static, temporal, spatial, hierarchical





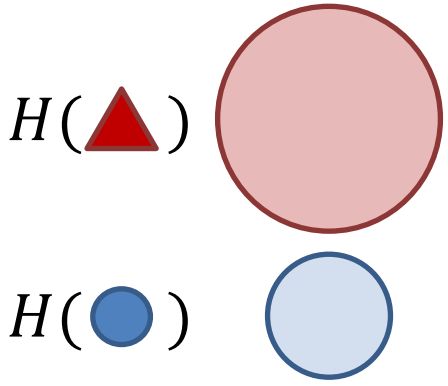
multisensory intelligence

Modality Profile

The total information contained in the elements and their composition.

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

4 **Information:** entropy and density



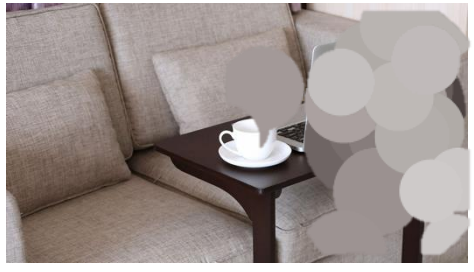
multisensory intelligence

Modality Profile

The natural imperfections in the data modality.

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

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

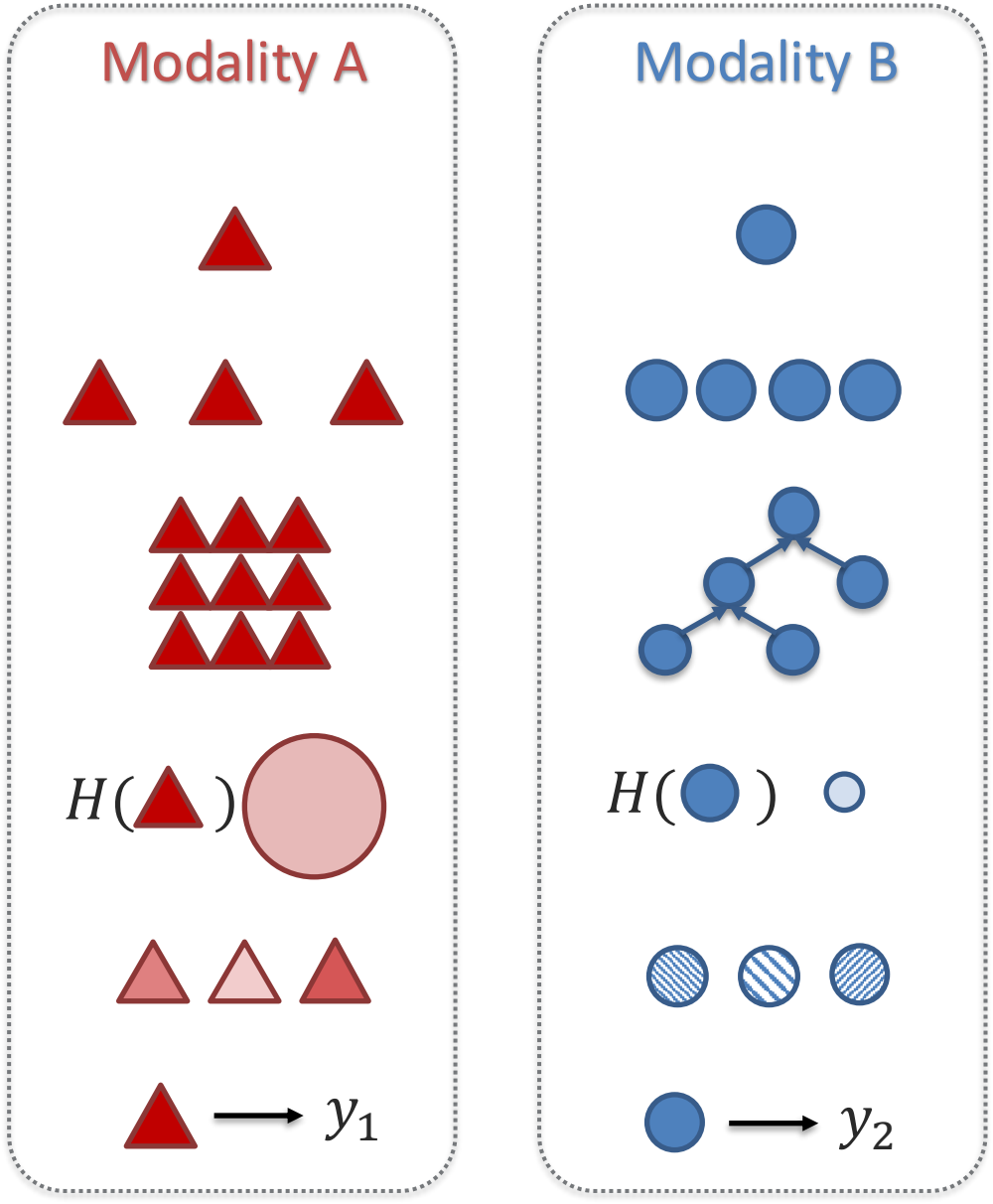


teacup → teacip

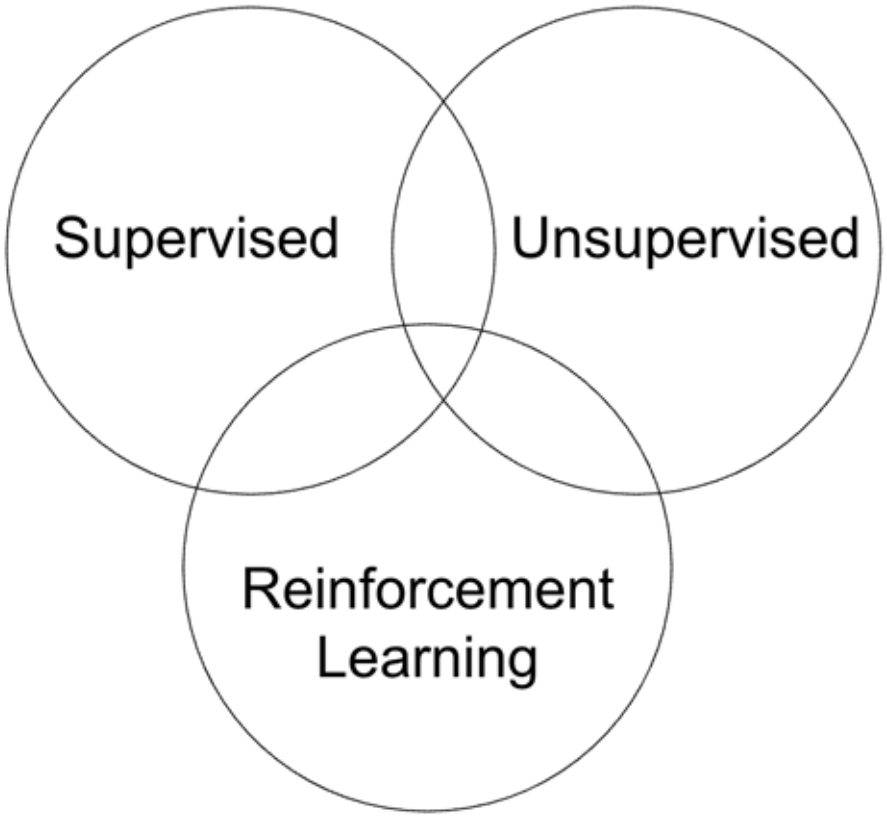
right → rihjt

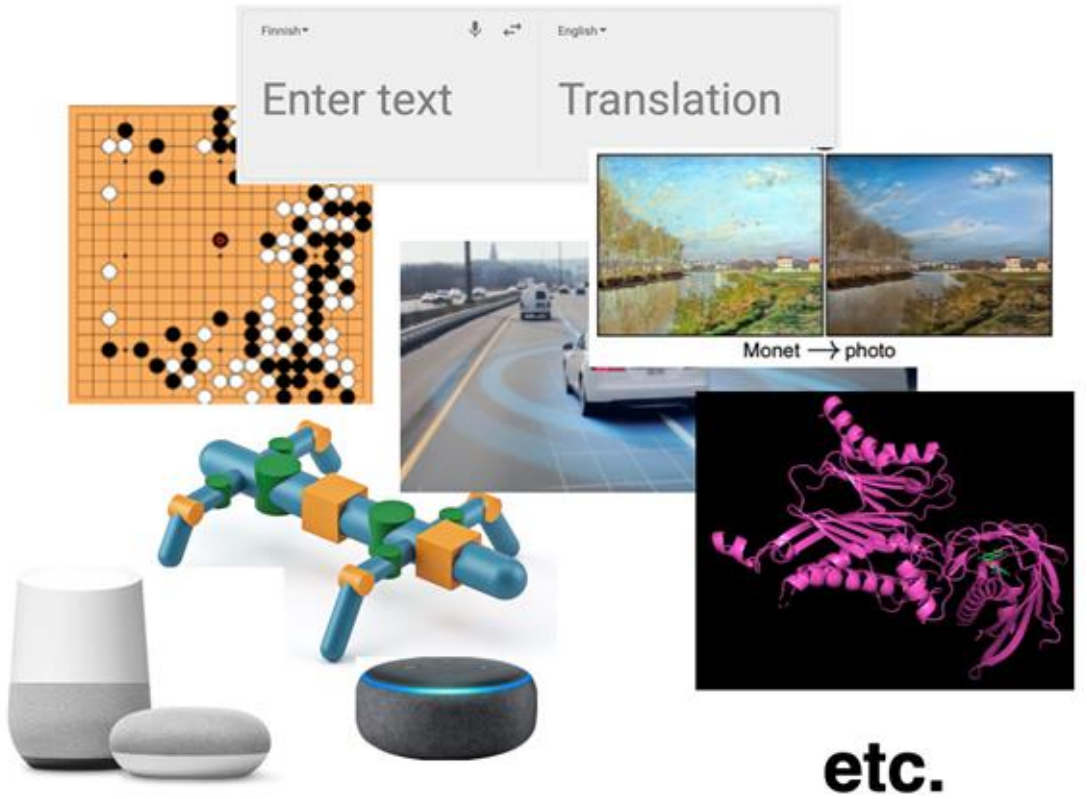
Modality Profile

1. **Element representations:**
   Discrete, continuous, granularity
2. **Element distributions:**
   Density, frequency
3. **Structure:**
   Temporal, spatial, latent, explicit
4. **Information:**
   Abstraction, entropy
5. **Noise:**
   Uncertainty, noise, missing data
6. **Relevance:**
   Task, context dependence



Types of Learning Paradigms





(the categorization can be refined, e.g. there are active learning, semi-supervised, selective, contrastive, few-shot, inverse reinforcement learning…)

[Slides adapted from 6.790]

Supervised Learning

Goal: correctly classify so far unseen test images

Goal: predict to what degree a drug candidate binds to the intended target protein (based on a dataset of already-screened molecules against the target)

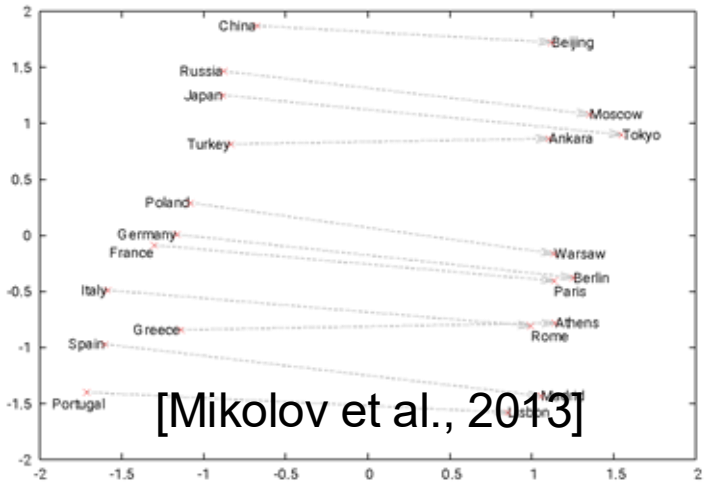
- Learning a machine translation system from pairs of sentences

Spanish (input)	**English (output)**
Aquí tienes un bolígrafo	Here's a pen
Las conferencias de ML son divertidas	ML conferences are fun
Todo el mundo debería estudiar AI	Everyone should study AI

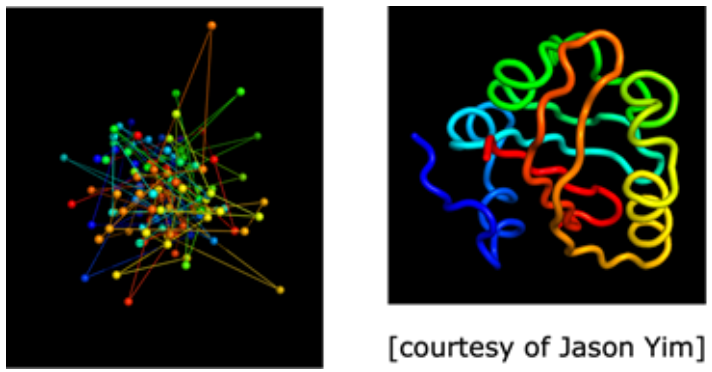
[Slides adapted from 6.790]

Unsupervised Learning

dimensionality reduction, embedding



[Mikolov et al., 2013]



[courtesy of Jason Yim]

Over 3D protein structures, etc.

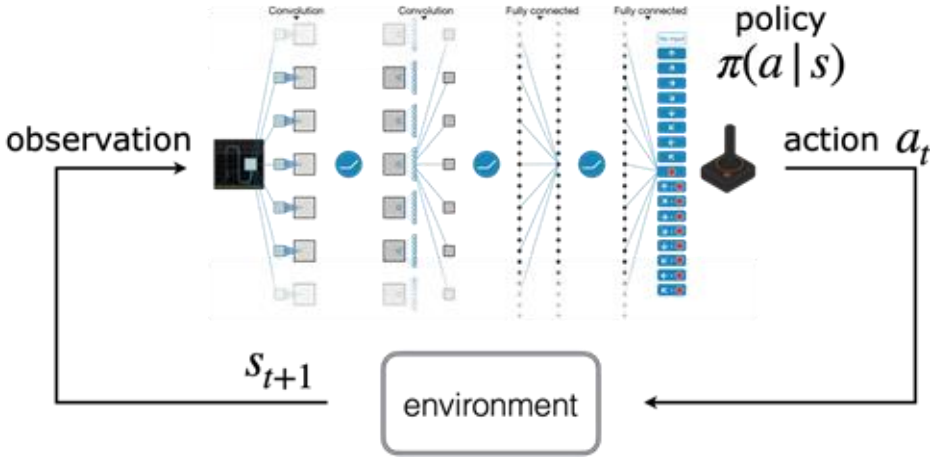
+Self-Supervised paradigm

de-noising diffusion models over images

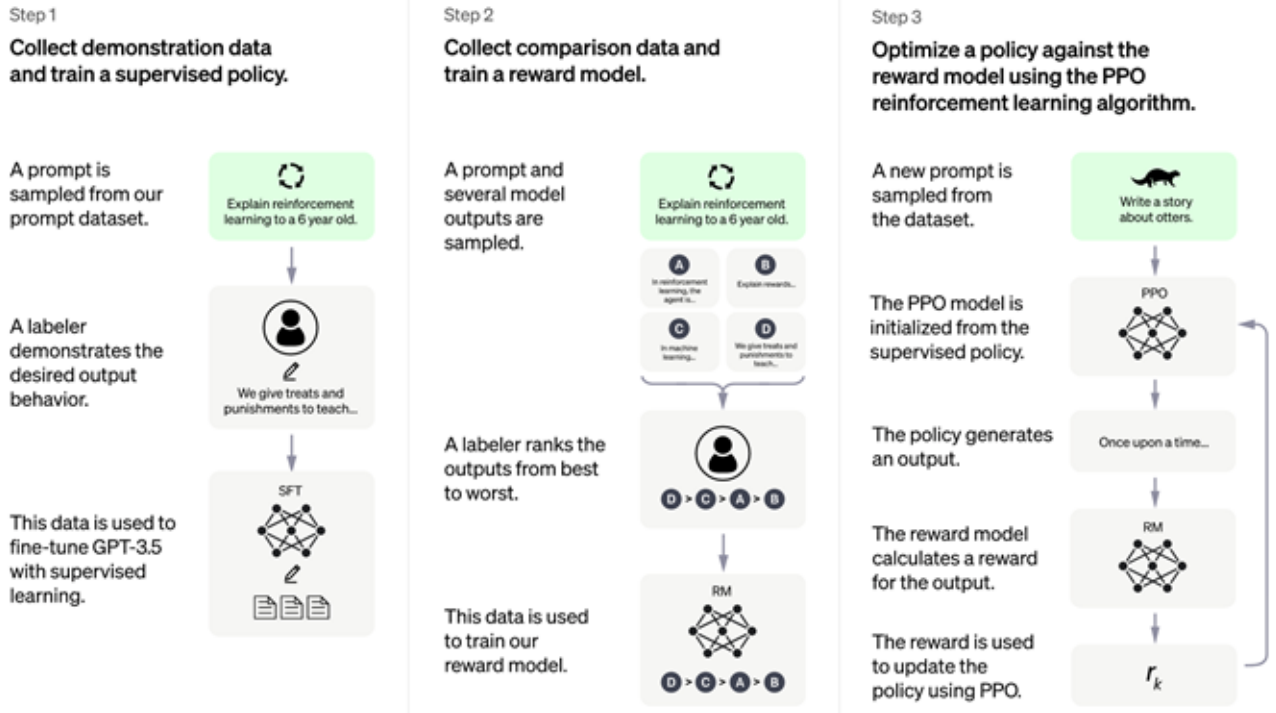
[image from Rissanen et al 2022]

[Slides adapted from 6.790]

Reinforcement Learning



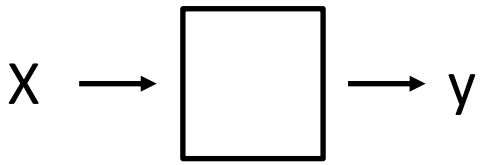
ChatGPT



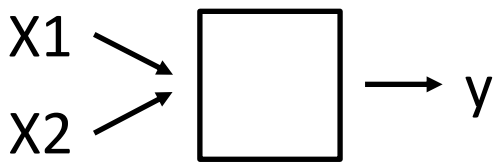
[Slides adapted from 6.790]

More Learning Paradigms

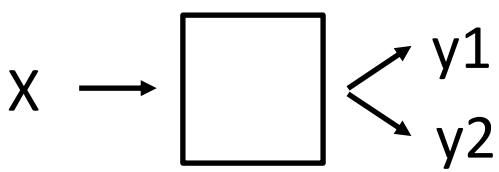
Supervised learning



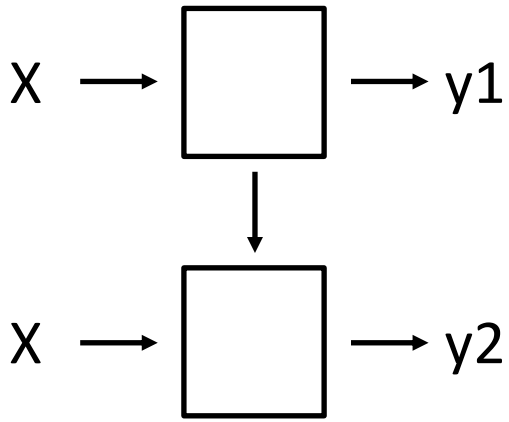
Multimodal (supervised) learning



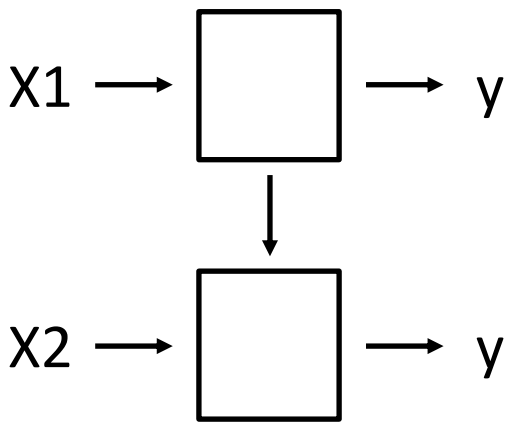
Multitask learning



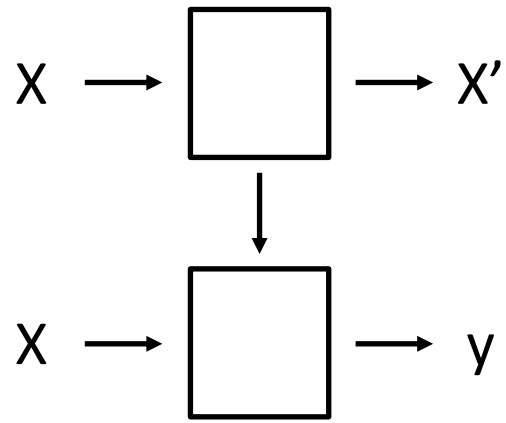
Transfer learning



Cross-modal learning

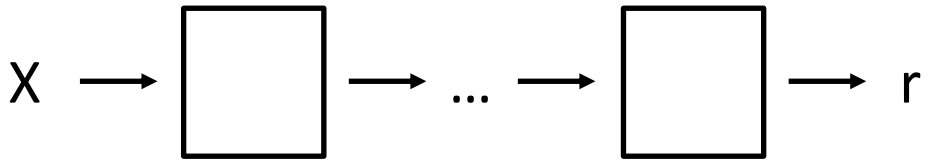


Unsupervised/self-supervised pre-training

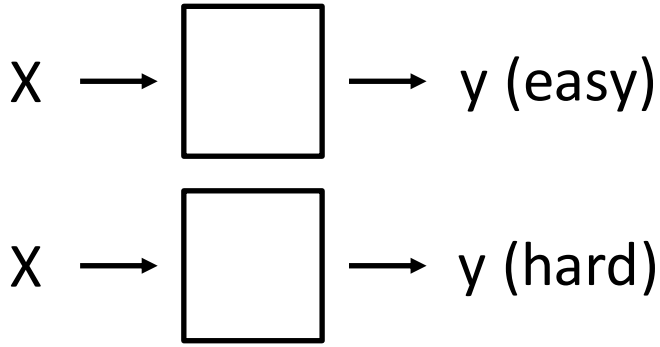


More Interactive Learning Paradigms

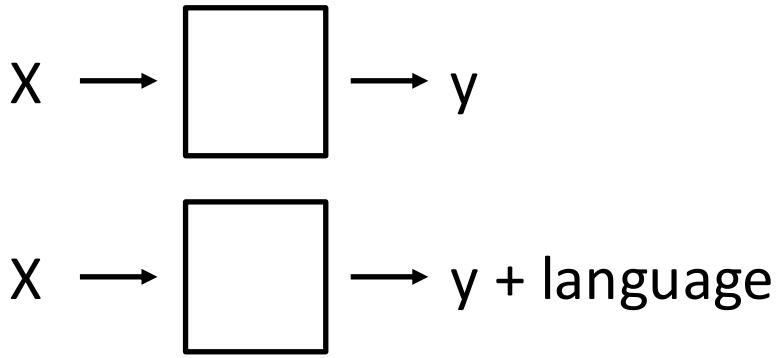
Reinforcement learning



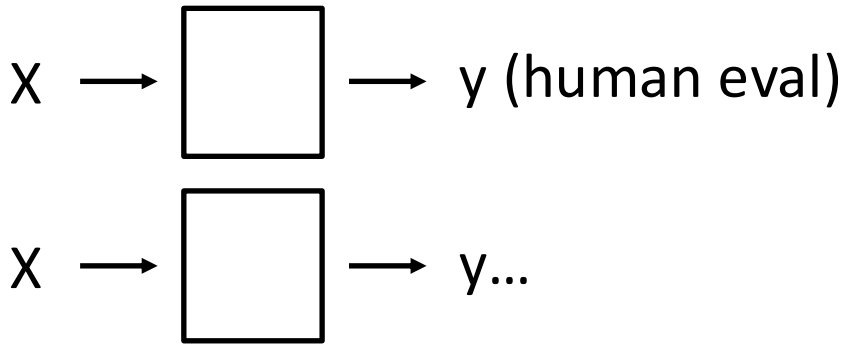
Curriculum/active learning



LLM adaptation



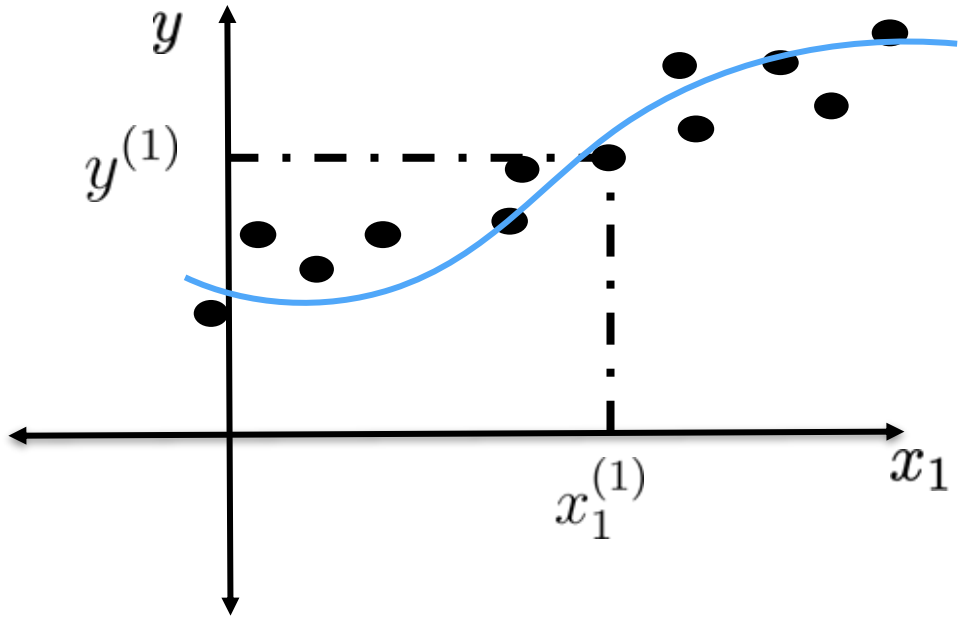
Human-in-the-loop learning



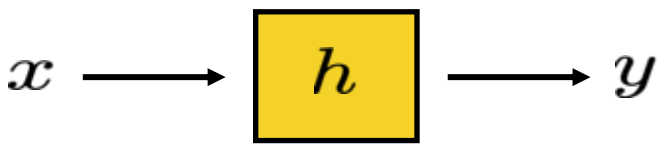
Learning Process

We want a "good" way to label our data

- How to label? Learn a model
- We typically consider a class of possible models



Input:
Data



Output:
Label

how well our model labels new data depends largely on how good the chosen model class is

[Slides adapted from 6.390]

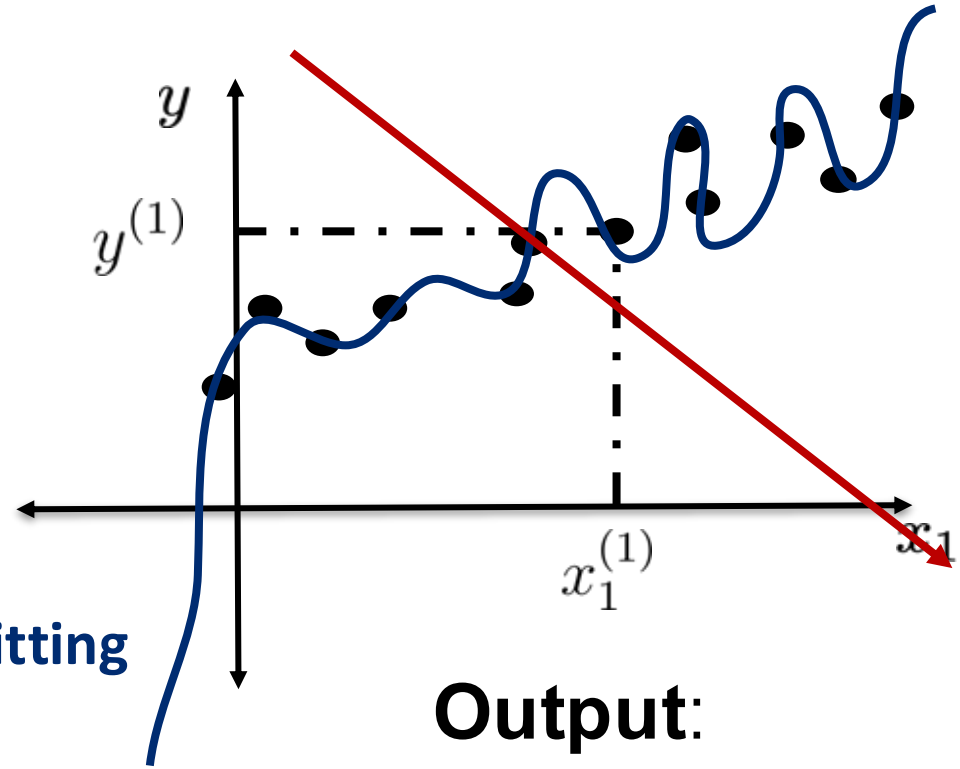
Overfitting vs Generalization

What we really want is to generalize to **future data**!

What we don't want:

- Model does not capture the input-output relationship → **Underfitting**
- Model too specialized to training data → **Overfitting**

Split collected data into training, validation, and testing. Critical to make sure test data conditions match real-time deployment conditions.

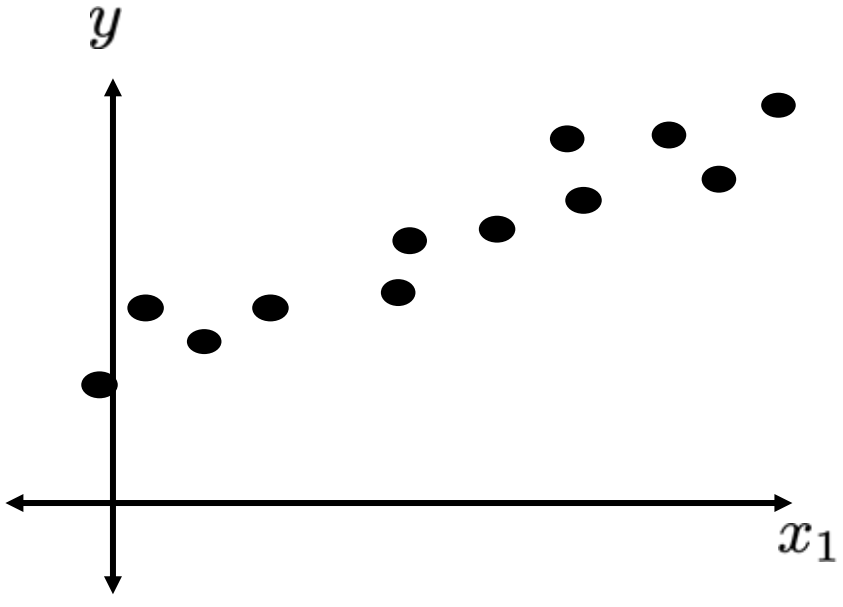


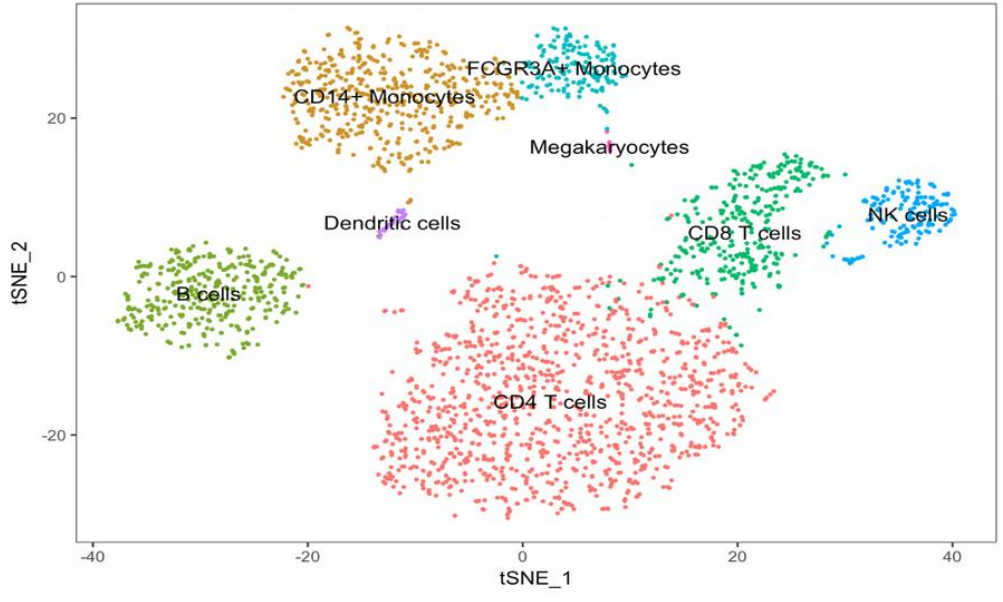
Output:
Label

[Slides adapted from 6.390]

Summary: How To Data

1. Decide how much data to collect, and how much to label (costs and time)
2. Clean data: normalize/standardize, find noisy data, anomaly/outlier detection
3. Visualize data: plot, dimensionality reduction (PCA, t-sne), cluster analysis
4. Decide on evaluation metric (proxy + real, quantitative and qualitative)
5. Choose model class and learning algorithm (more coming up)



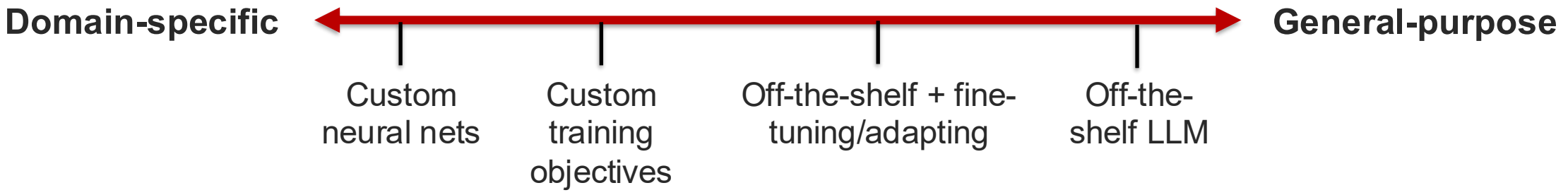


multisensory intelligence

Lecture Outline

1. A unifying paradigm of model architectures
2. Temporal sequence models
3. Spatial convolution models
4. Models for sets and graphs

Two General Modeling Paradigms



Your decision will depend on many factors.

multisensory intelligence

Designing Models for Data

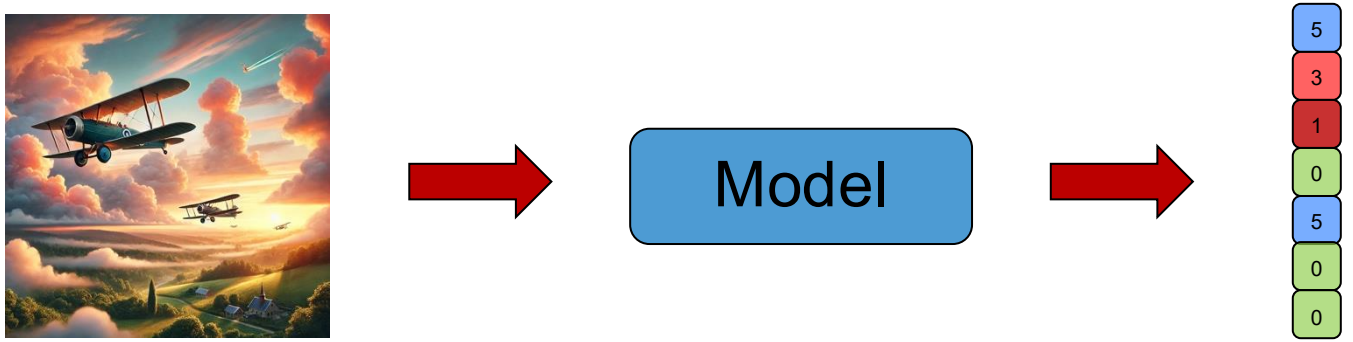
What is a good model?

One that captures the:
- right semantic information
- at the right granularity
- using an appropriate amount of data and labels
- with the right resource constraints
- with the right level of usability (explainability, accessibility, etc.)
- and more…

Domain-specific  **General-purpose**

Unified View of Deep Learning Models

1. Learning representations



2. Combining representations (information aggregation)

Unified View of Deep Learning Models

Composing differentiable functions and training objectives.

1. Basic representation building blocks for each element

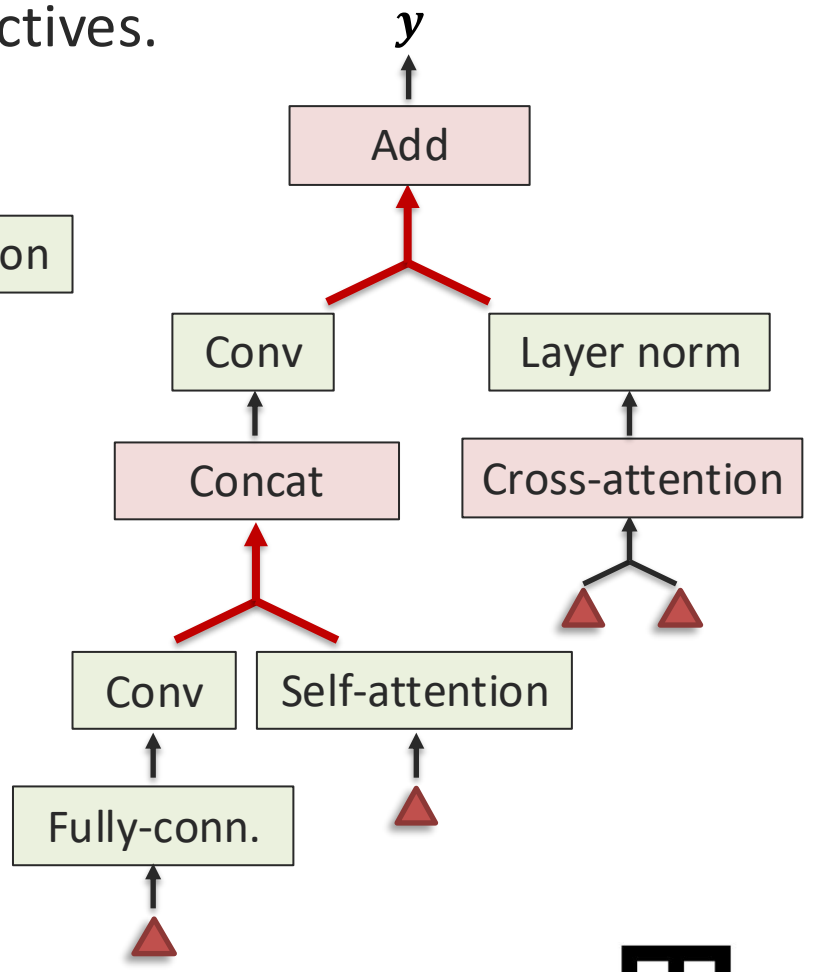
   Fully-connected | ReLU | Layer norm | Conv | Self-attention

2. Basic information aggregation blocks

   Concat | Cross-attention | Add | Max

3. Compute loss function

4. Take gradients, update with stochastic gradient descent

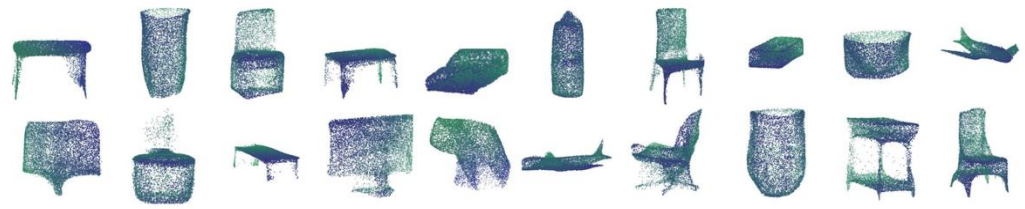


multisensory intelligence

A Simple Classification Example

Sets and point clouds

Sets



Point clouds

Set anomaly detection
Set expansion
Set completion
Point cloud classification
Point cloud generation

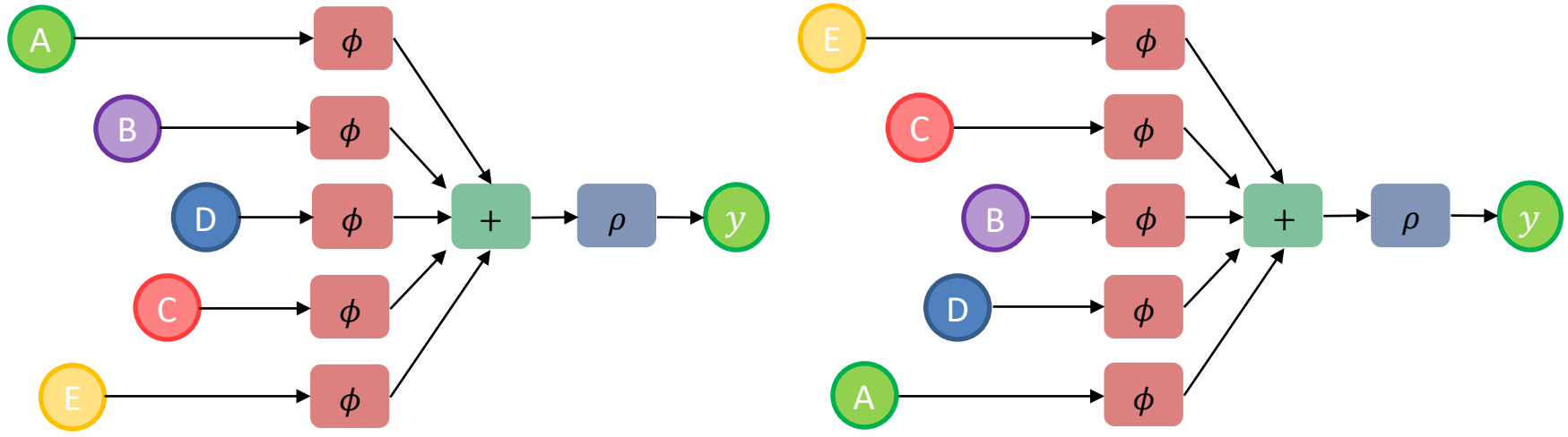
[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

multisensory intelligence

A Simple Classification Example

Models for set-based data must be invariant to element order.

1. Parameter sharing for each set element
2. Permutation invariant aggregation function



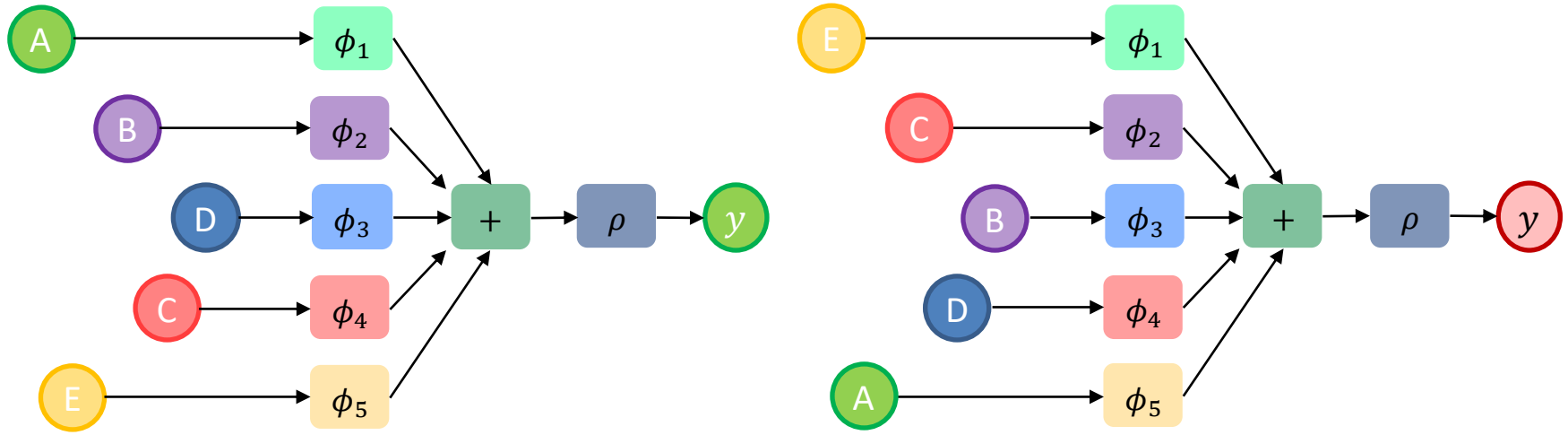
Giving ABDCE also gives ECBDA, BCAED etc…

[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

A Simple Classification Example

Models for set-based data must be invariant to element order.

1. No parameter sharing for each set element
2. Permutation invariant aggregation function



Need to give ABDCE, then ECBDA, and more…. 5! more samples needed

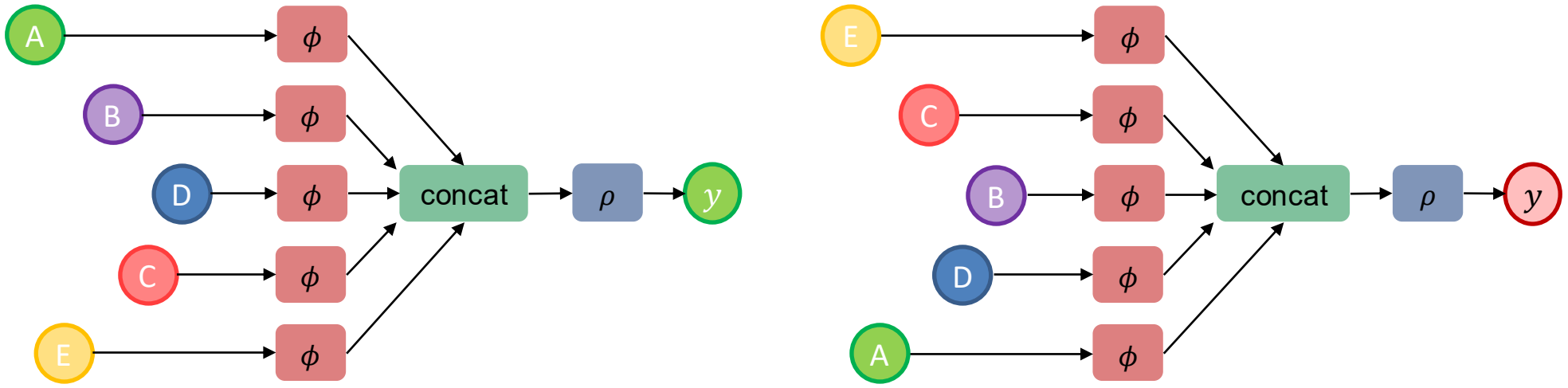
[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

multisensory intelligence

A Simple Classification Example

Models for set-based data must be invariant to element order.

1. Parameter sharing for each set element
2. Not permutation invariant aggregation function



Need to give ABDCE, then ECBDA, and more…. 5! more samples needed

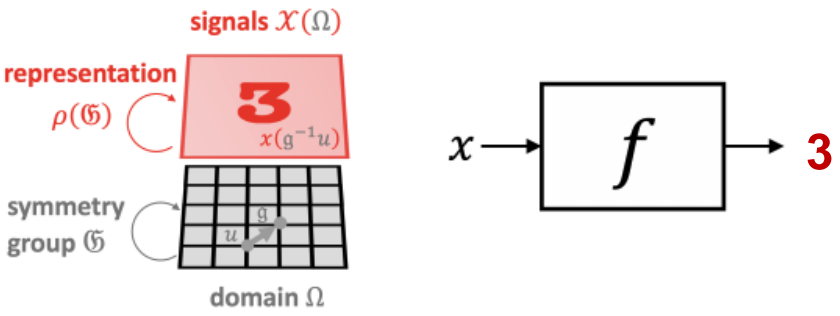
[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

multisensory intelligence

Structure

Data invariances – example of image classification

A function $f : \mathcal{X}(\Omega) \to \mathcal{Y}$ is $\mathfrak{G}$-*invariant* if $f(\rho(\mathfrak{g})x) = f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$ and $x \in \mathcal{X}(\Omega)$, i.e., its output is unaffected by the group action on the input.



sunset

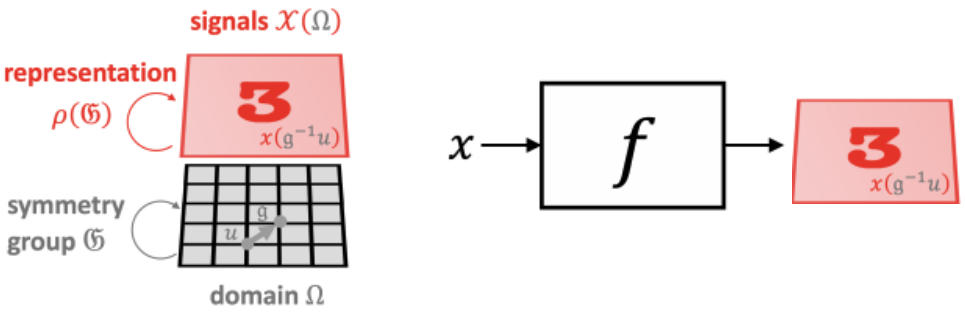
[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

multisensory intelligence

Structure

Data equivariances – example of image segmentation

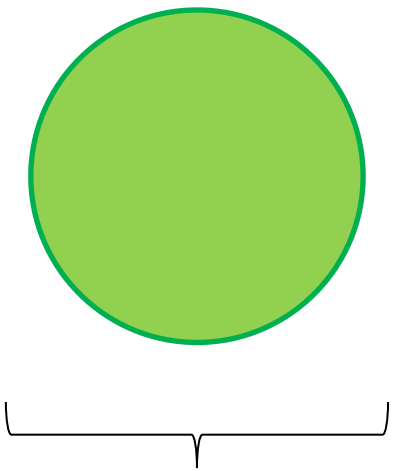
A function $f : \mathcal{X}(\Omega) \to \mathcal{X}(\Omega)$ is $\mathfrak{G}$-*equivariant* if $f(\rho(\mathfrak{g})x) = \rho(\mathfrak{g})f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$, i.e., group action on the input affects the output in the same way.



[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

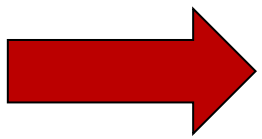
multisensory intelligence

Elements



X_i

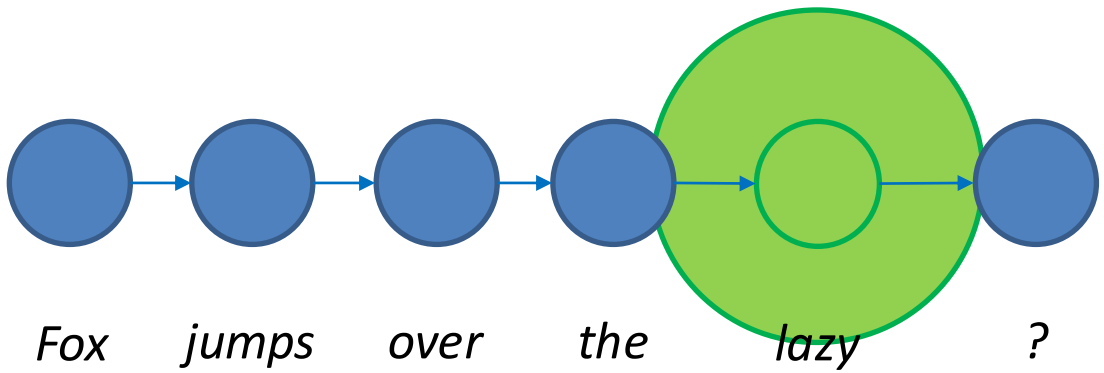
Lazy



Y_i

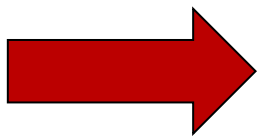
Sentiment?

From Token to Sequences



X_i

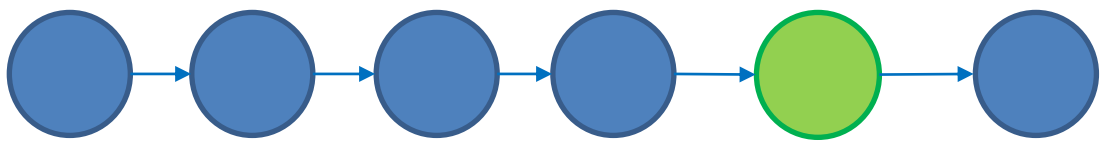
Lazy



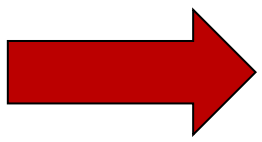
Y_i

Next word?

From Token to Sequences



Fox *jumps* *over* *the* *lazy* *?*

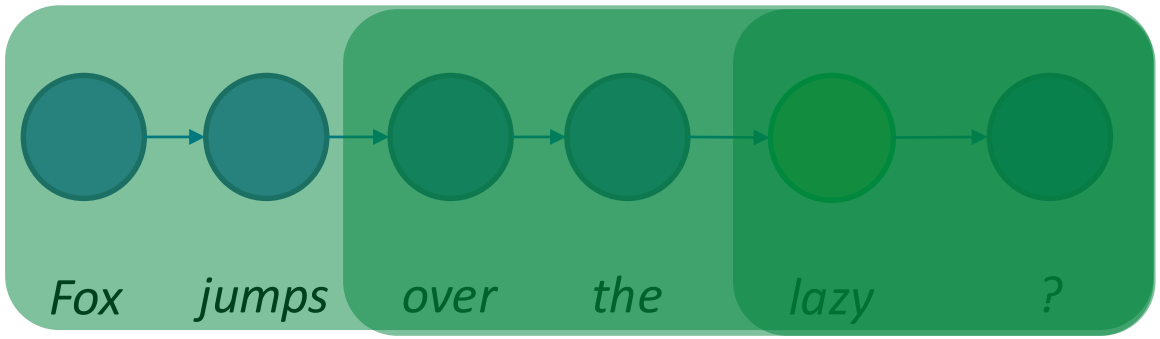


Y_i

Next word?

How do we aggregate information?

From Token to Sequences



Y_i

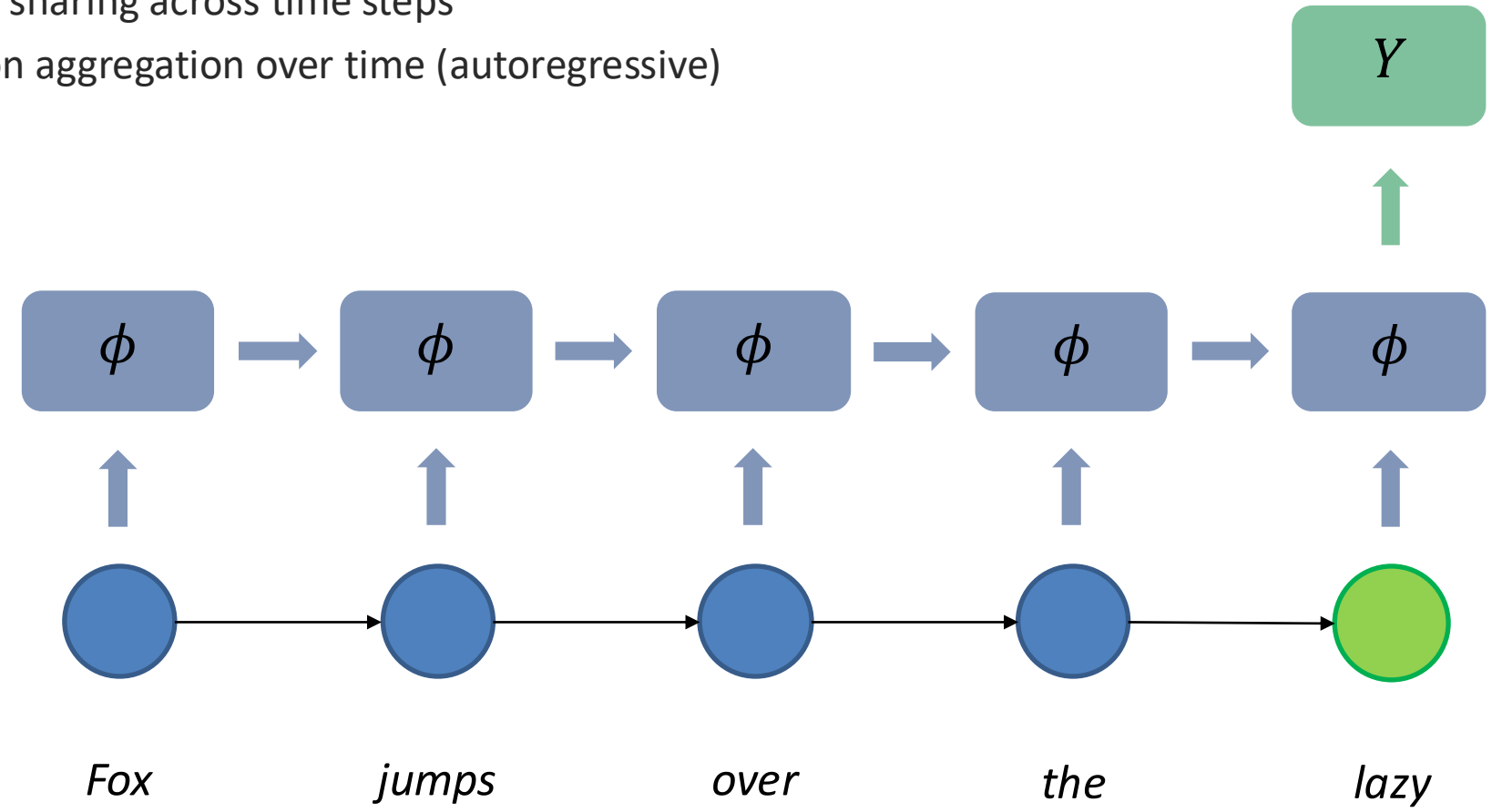
Next word?

How do we aggregate information?

Sequence Classification

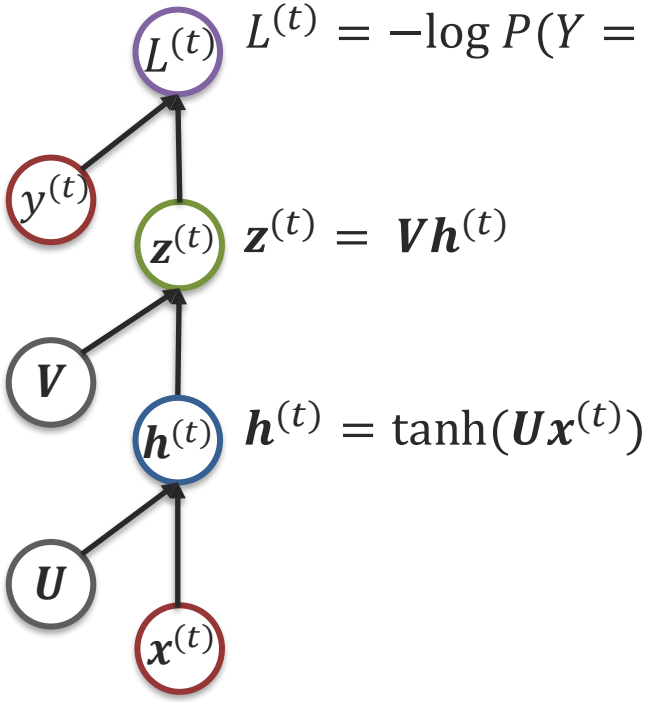
Models for sequential data must be invariant to time, but equivariant to word order.

1. Parameter sharing across time steps
2. Information aggregation over time (autoregressive)



Sequence Classification

Feedforward Neural Network



$$L^{(t)} = -\log P(Y = y^{(t)} | \boldsymbol{z}^{(t)})$$

$$\boldsymbol{z}^{(t)} = \boldsymbol{V}\boldsymbol{h}^{(t)}$$

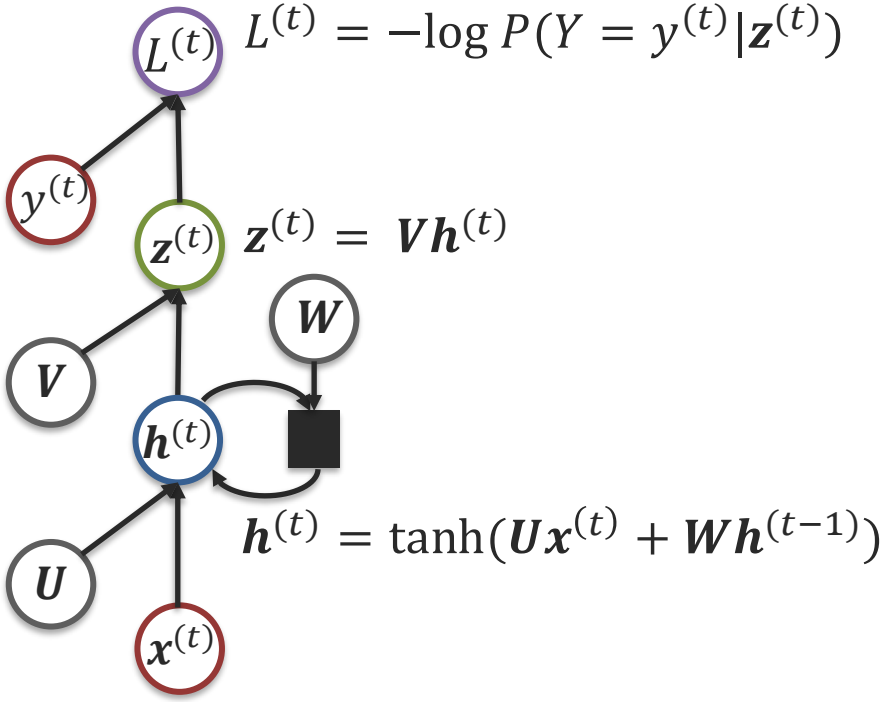
$$\boldsymbol{h}^{(t)} = \tanh(\boldsymbol{U}\boldsymbol{x}^{(t)})$$

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Sequence Classification

$$L = \sum_t L^{(t)}$$

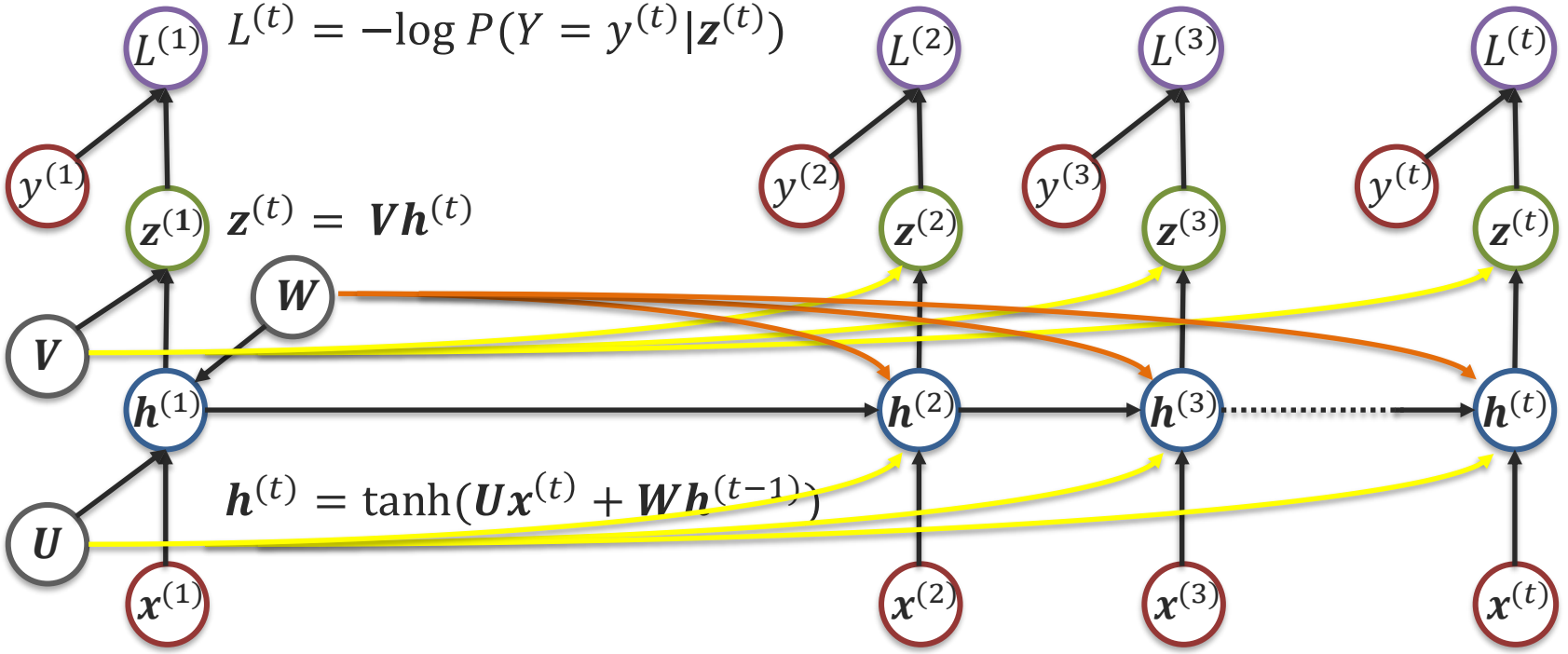


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Sequence Classification

$$L = \sum_t L^{(t)}$$

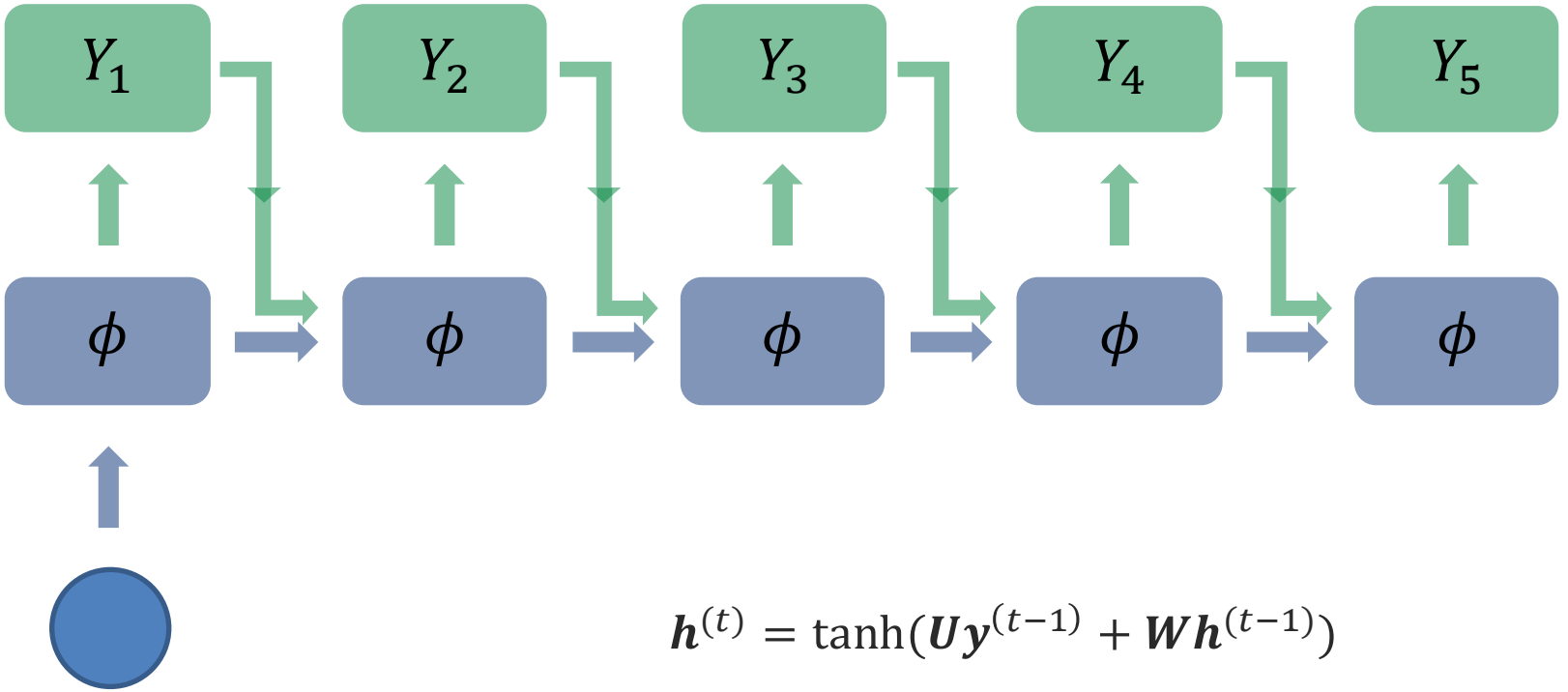


Same model parameters are used for all time steps.

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Sequence Generation

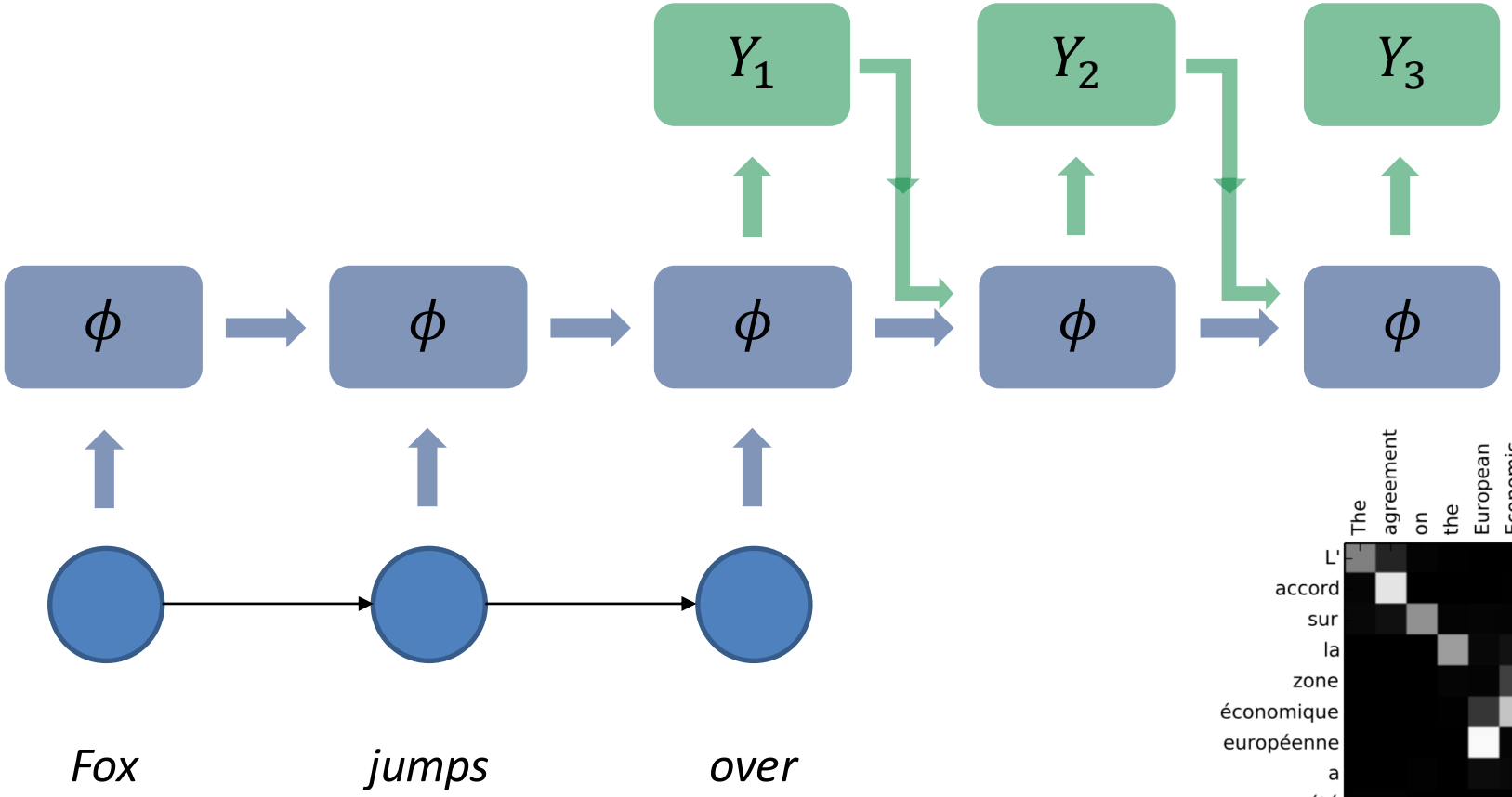


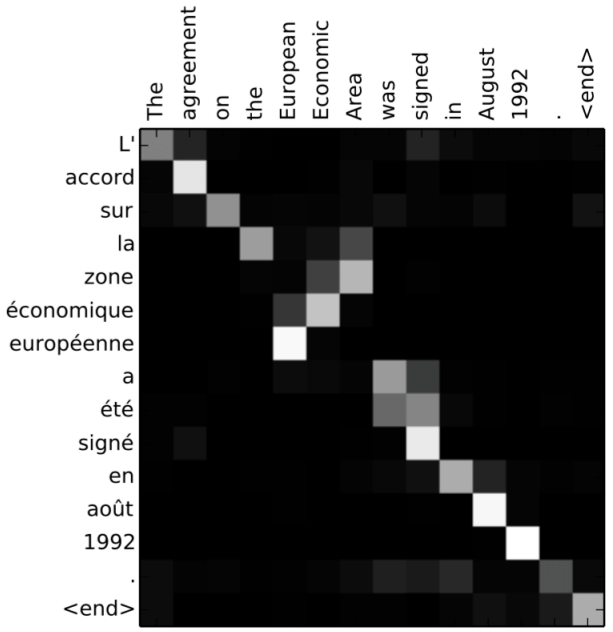
$$\boldsymbol{h}^{(t)} = \tanh(\boldsymbol{U}\boldsymbol{y}^{(t-1)} + \boldsymbol{W}\boldsymbol{h}^{(t-1)})$$

e.g, text or music generation

Modern versions: RNN -> LSTM -> TCN -> State space models

Sequence-to-Sequence Models





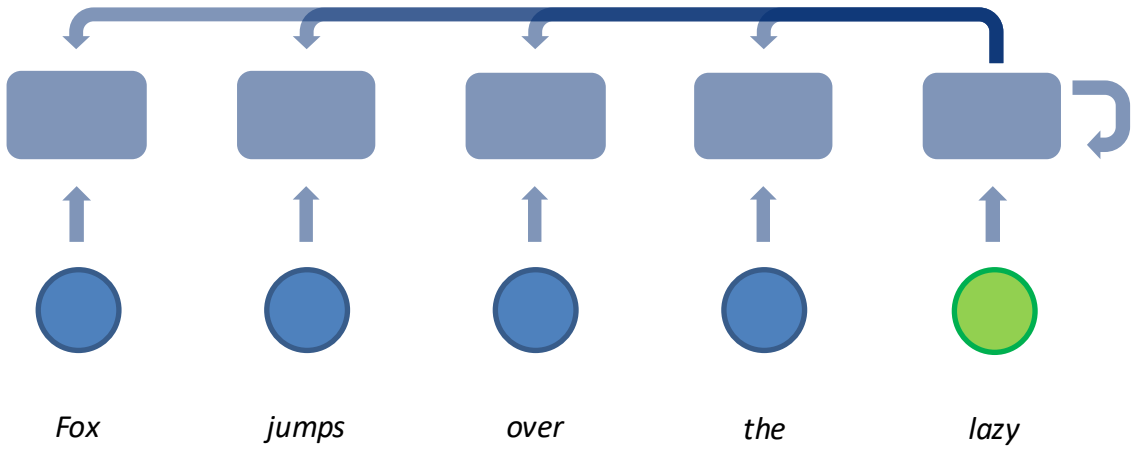
e.g, machine translation -> birth of attention-based models

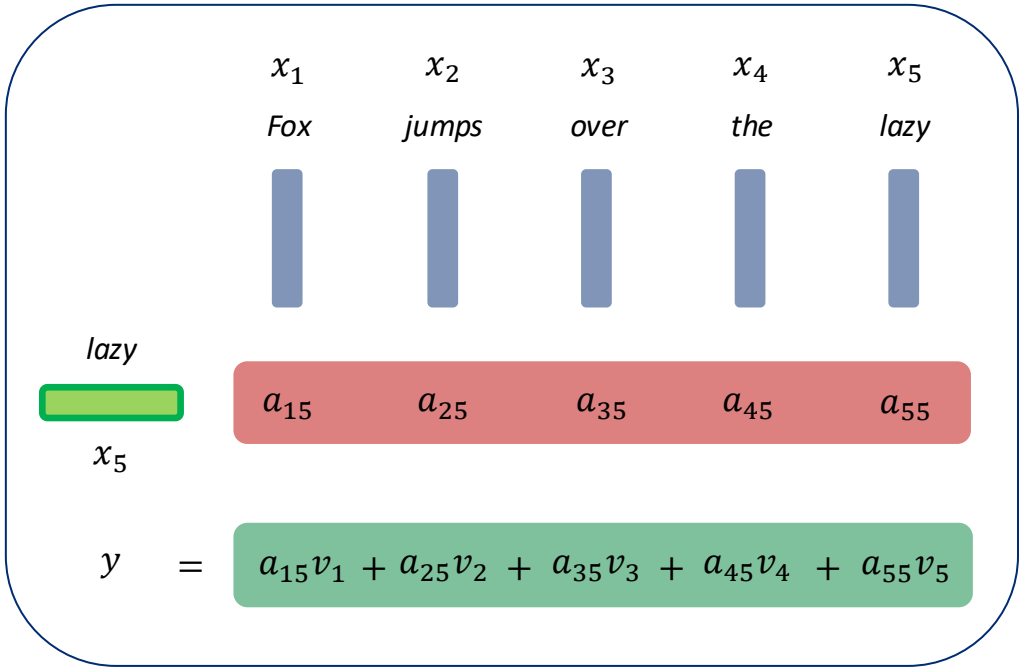
[Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015]

Modern Sequence Models

Birth of attention-based models

– Dynamic weights for different elements





[Vaswani et al., Attention is All You Need. NeurIPS 2017]

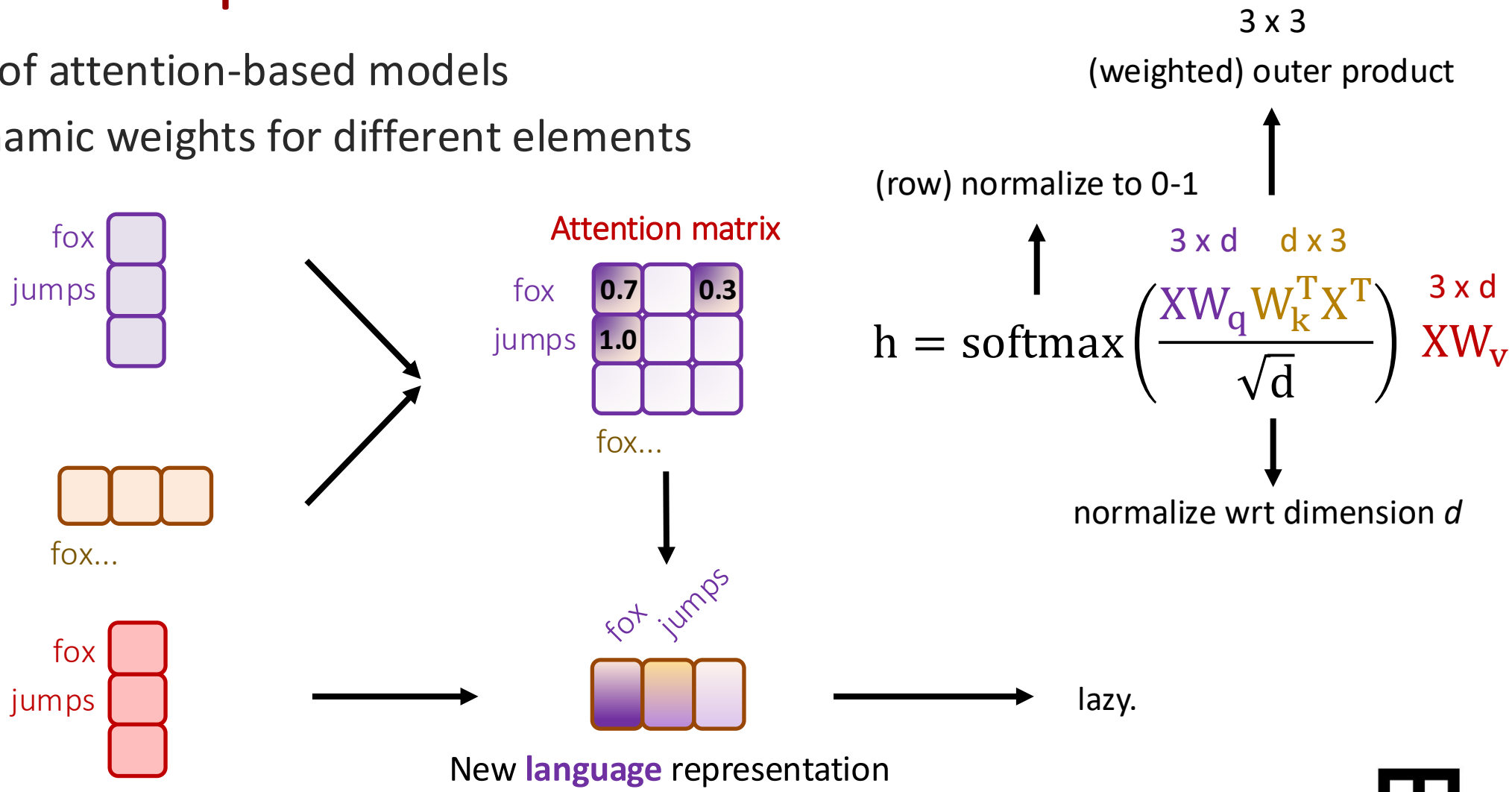
multisensory intelligence

53

Modern Sequence Models

Birth of attention-based models

– Dynamic weights for different elements



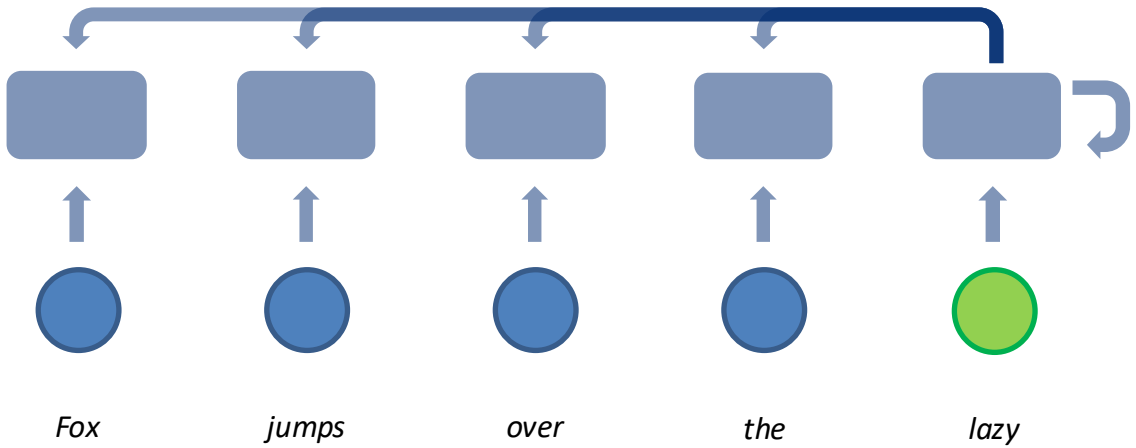
[Vaswani et al., Attention is All You Need. NeurIPS 2017]

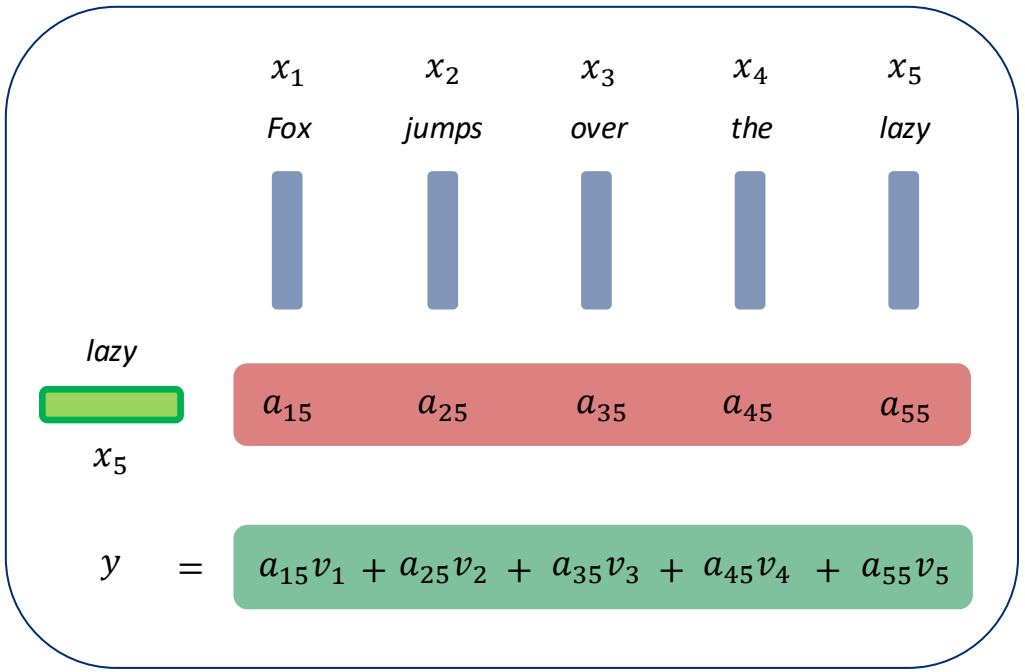
multisensory intelligence

Modern Sequence Models

Models for sequential data must be invariant to time, but equivariant to word order.

1. Parameter sharing across time steps
2. Information aggregation over time (in parallel)





$$\mathrm{h} = \mathrm{softmax}\left(\frac{\overset{\mathrm{T\,x\,d}}{\mathrm{X}\mathrm{W}_\mathrm{q}}\overset{\mathrm{d\,x\,T}}{\mathrm{W}_\mathrm{k}^\mathrm{T}\mathrm{X}^\mathrm{T}}}{\sqrt{\mathrm{d}}}\right)\overset{\mathrm{T\,x\,d}}{\mathrm{X}\mathrm{W}_\mathrm{v}}$$

[Vaswani et al., Attention is All You Need. NeurIPS 2017]

multisensory intelligence

Spatial Data

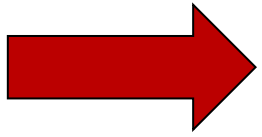
Y_i

Is there a fox?

How do we aggregate information?

multisensory intelligence

Spatial Data



Y_i

Is there a fox?

How do we aggregate information?

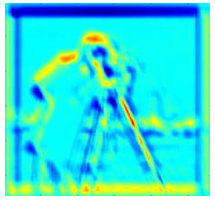
Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translations.

Input

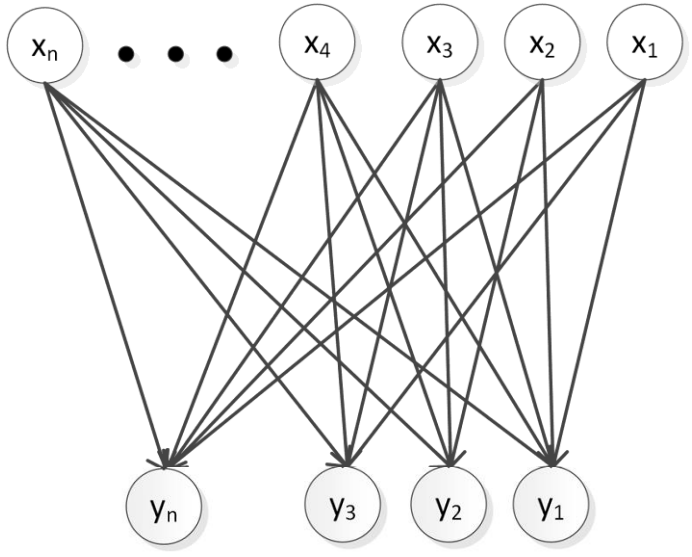
(image)

Output



(response map)

Input: all pixels



Output: responses

Not efficient!

200×200 image requires $40{,}000 \times n$ parameters

(where n is size of output)

And it may learn different outputs for different pixel positions

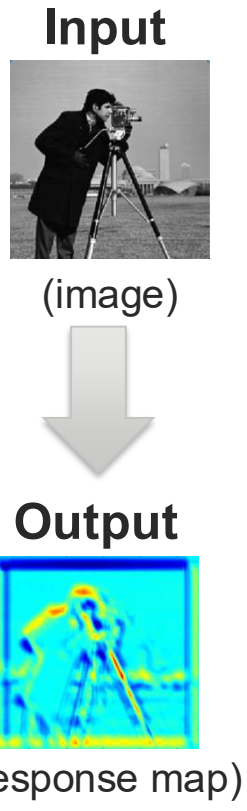
⇨ Not spatial invariant

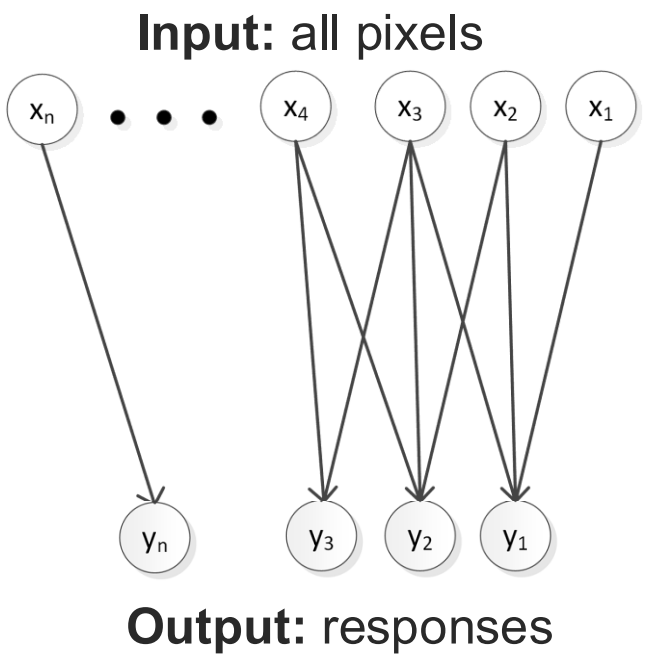
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Convolutional Neural Networks

Modification 1: Only apply the filter to a small sliding window
-> for efficiency and locality





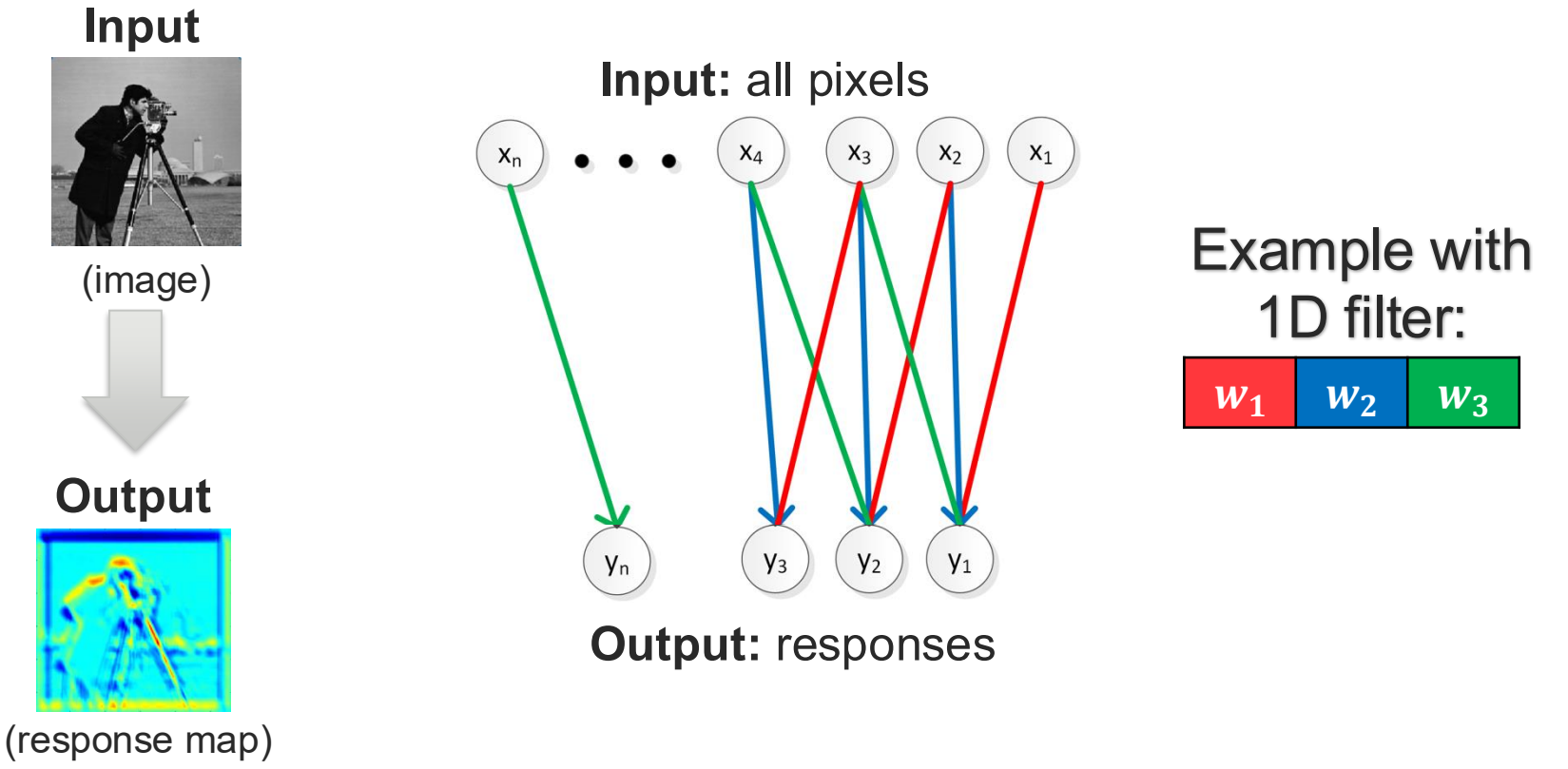
Example with
1D filter:

w_1	w_2	w_3

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Convolutional Neural Networks

Modification 2: Same filter applied to all sliding windows
-> for spatial invariance



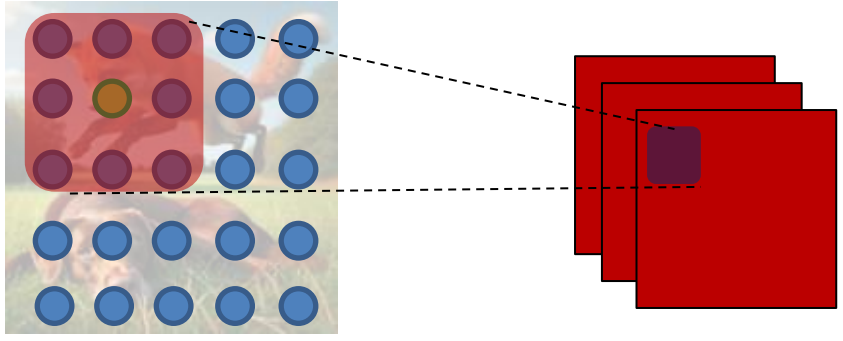
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

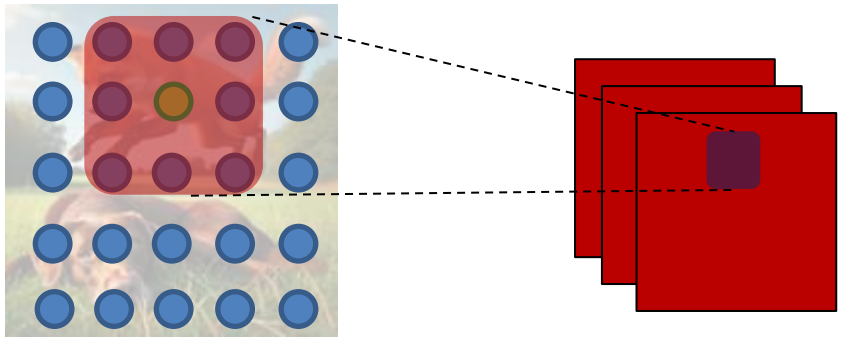
1. Parameter sharing across k x k convolutional filter



Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

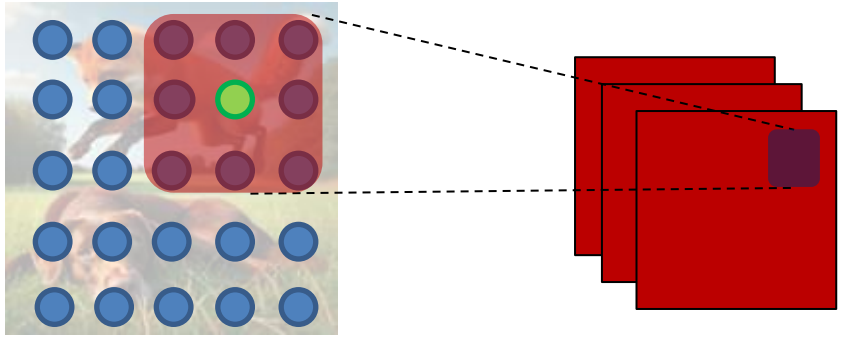
1. Parameter sharing across k x k convolutional filter



Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

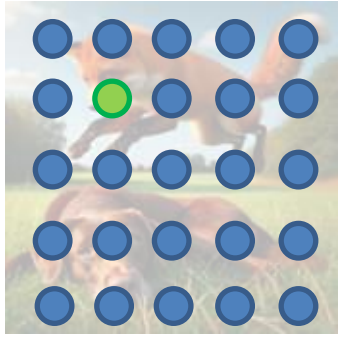
1. Parameter sharing across k x k convolutional filter

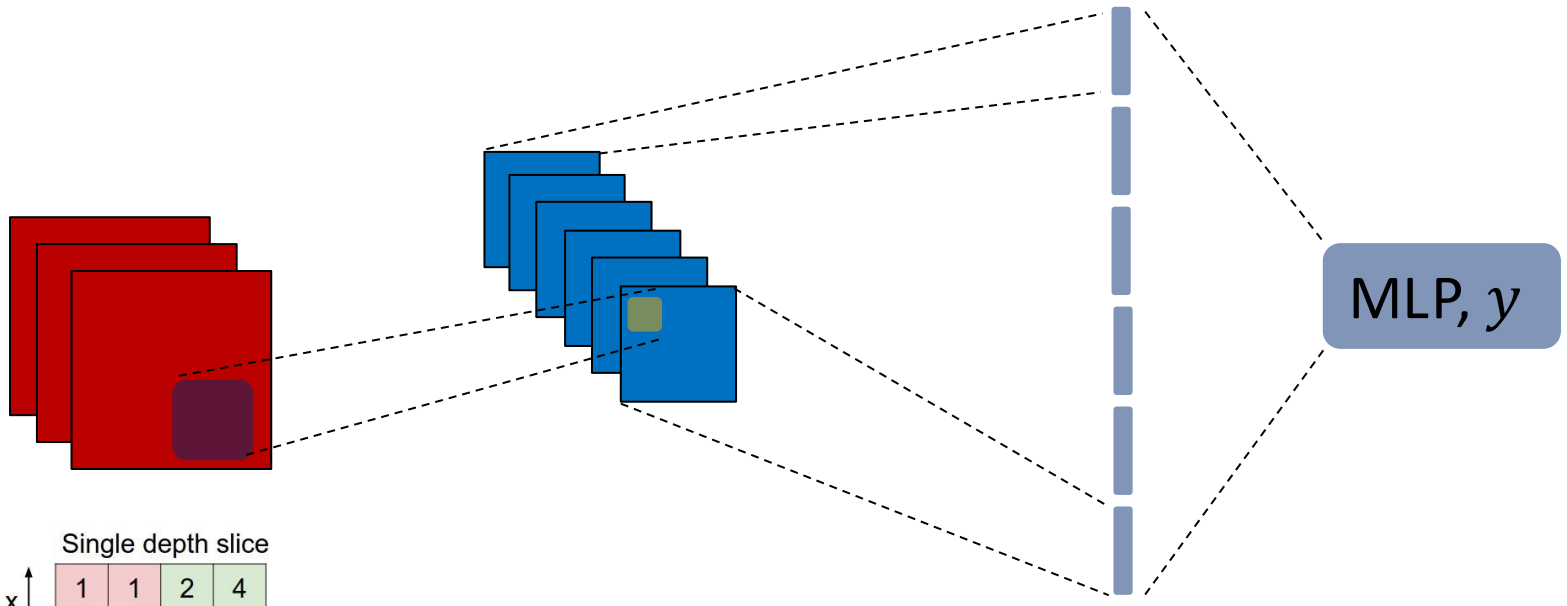


Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

1. Parameter sharing across k x k convolutional filter
2. Information aggregation over k x k pooling region





multisensory intelligence

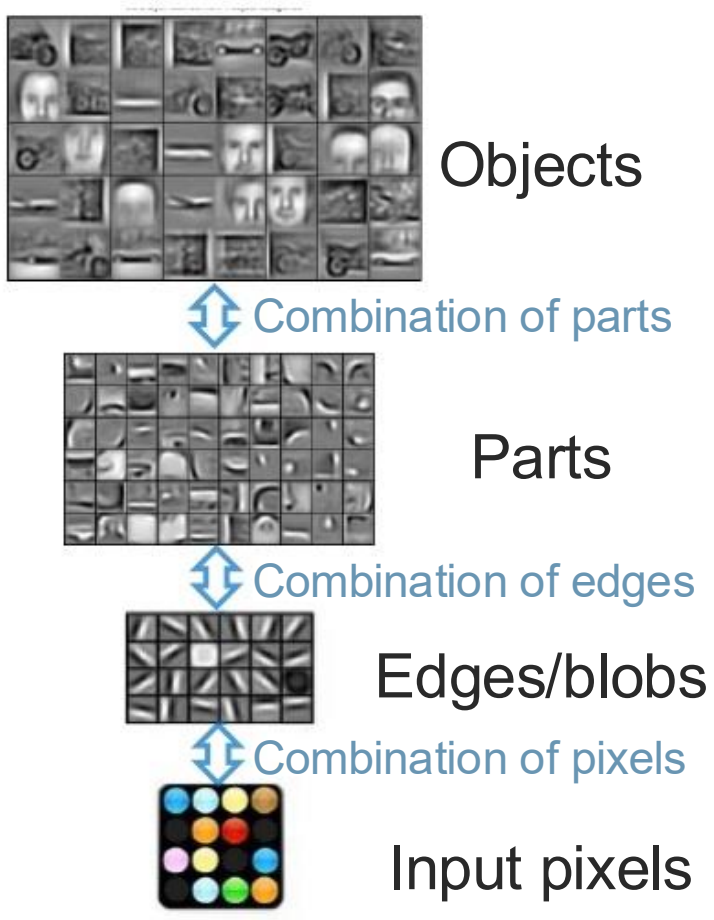
Convolutional Neural Networks

Multiple convolutional layers

- Allows the network to learn combinations of sub-parts, to increase complexity

Multiple pooling layers

- Allows the network to learn increasingly abstract & summarized information



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

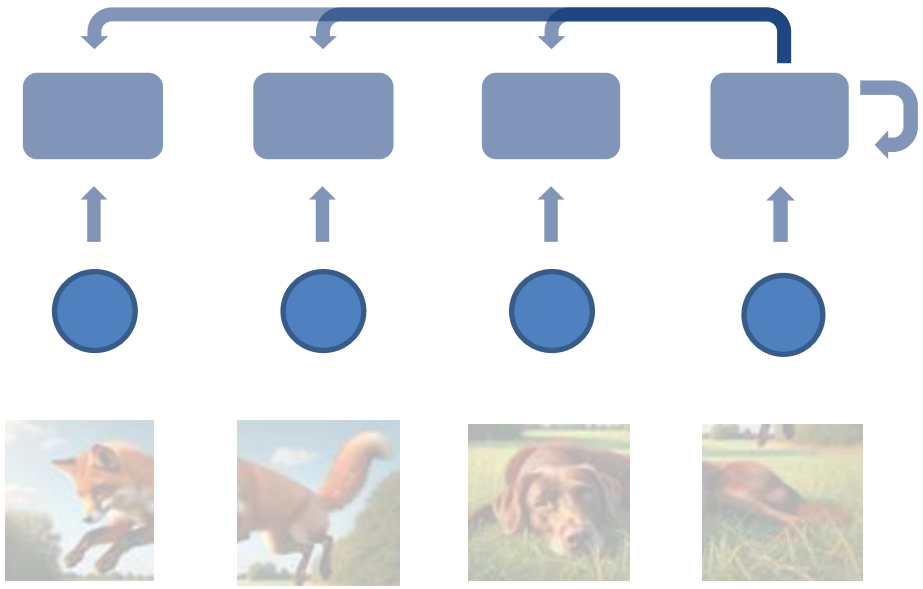
Vision Transformer

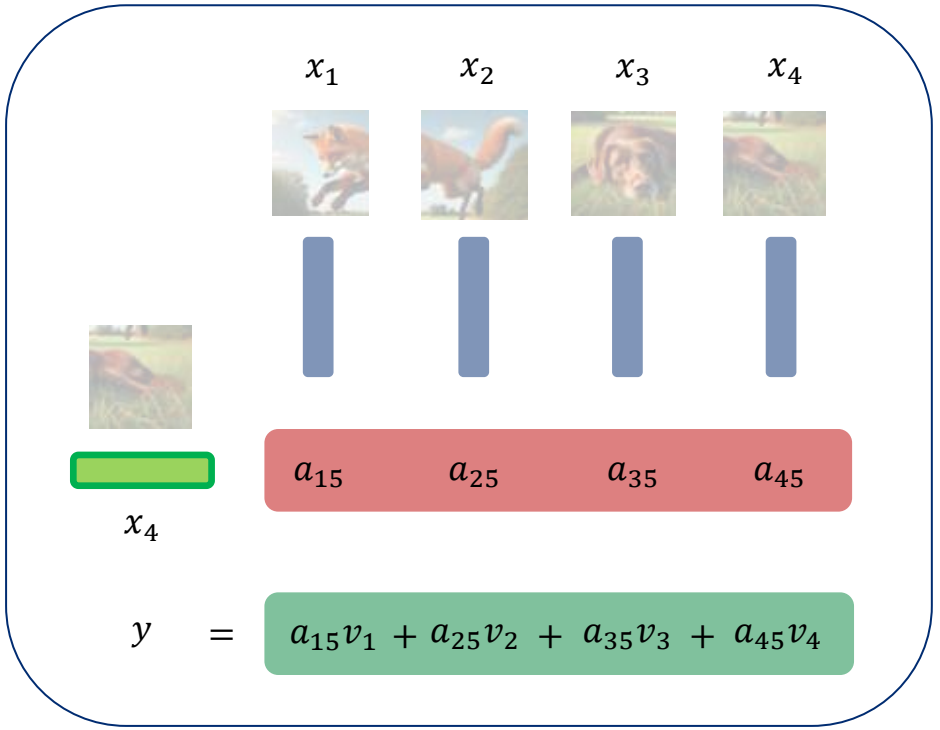
$$h = \text{softmax}\left(\frac{XW_q W_k^T X^T}{\sqrt{d}}\right) XW_v$$

(T x d, d x T, T x d)

Models for spatial data need to be invariant to spatial translation

1. Parameter sharing across k x k self-attention region
2. Information aggregation over k x k patch region





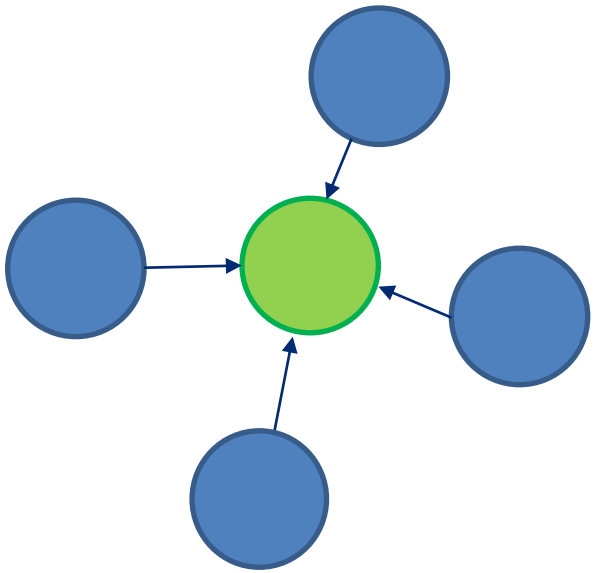
[Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021]

multisensory intelligence

Vision Transformer

[Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021]

Graphs



Y_i

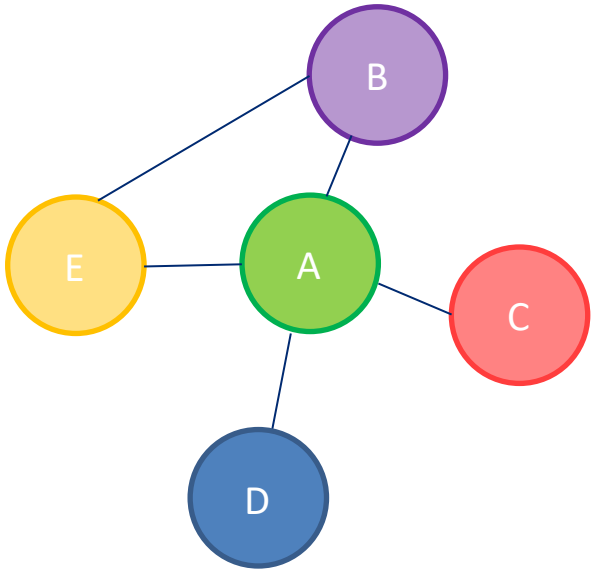
What molecule is this?

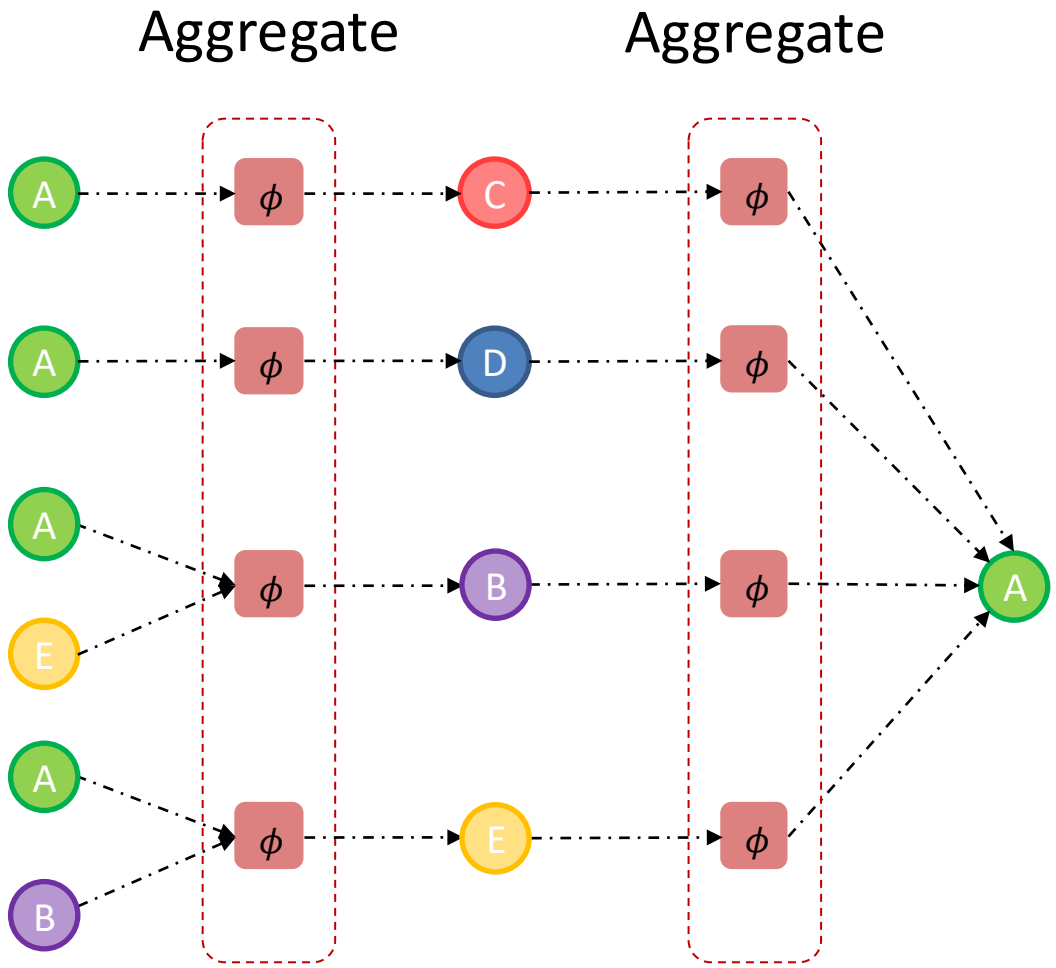
How do we aggregate information?

Graph Neural Networks

Models for graph data:

1. Parameter sharing across nodes
2. Information aggregation over neighbors (edges)





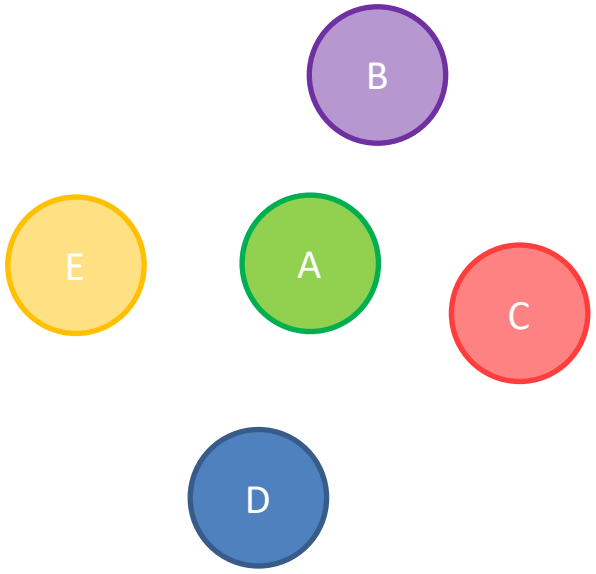
[Velickovic et al., Graph Attention Networks. ICLR 2018]
[Yun et al., Graph Transformer Networks. NeurIPS 2019]

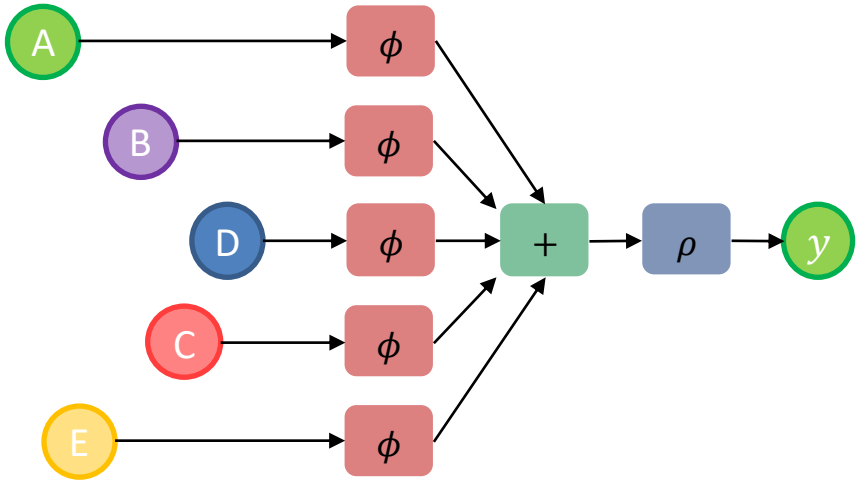
multisensory intelligence

Graph Recover Sets

Sets are graphs with only nodes, no edges

1. Parameter sharing across nodes -> set elements
2. Information aggregation over neighbors -> no neighbors





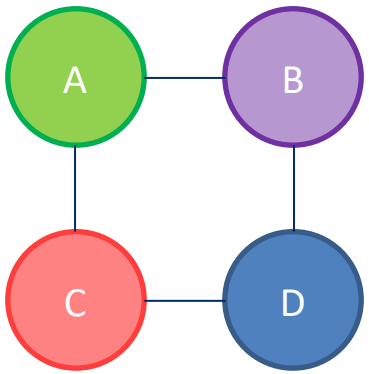
[Zaheer et al., DeepSets. NeurIPS 2017]
[Lee et al., Set Transformer. ICML 2019]

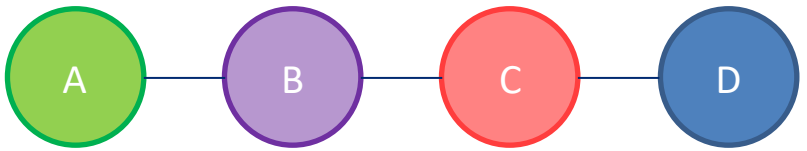
multisensory intelligence

Graph Recover Spatial and Temporal Data

Spatial data and sequential data

1. Parameter sharing across nodes
2. Information aggregation over neighbors

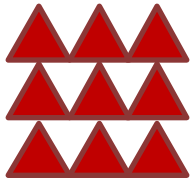


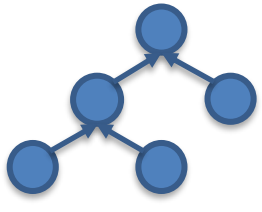


[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

Summary: How To Model

1. Decide how much data to collect, and how much to label (costs and time)
2. Clean data: normalize/standardize, find noisy data, anomaly/outlier detection
3. Visualize data: plot, dimensionality reduction (PCA, t-sne), cluster analysis
4. Decide on evaluation metric (proxy + real, quantitative and qualitative)
5. Choose modeling paradigm - domain-specific vs general-purpose
6. Figure out base elements and their representation
7. Figure out data invariances & equivariances (+other parts of modality profile)
8. Iterate between data collection, model design, model training, hyperparameter tuning etc. until satisfied.





multisensory intelligence

Lecture Summary

1. A unifying paradigm of model architectures
2. Temporal sequence models
3. Spatial convolution models
4. Models for sets and graphs

Assignments for This Coming Week

HW1: reading assignment + homework due next Tuesday 2/17

For project:

- Project proposal instructions released, due Tuesday (2/24). Submit on canvas
- Meet with me 4-5pm if need feedback about proposal ideas.