Introduction to Deep Learning for Dialogue Systems

이 화 란

Hwaran Lee

SK T-Brain, AI Center

October 10, 2019

Outline

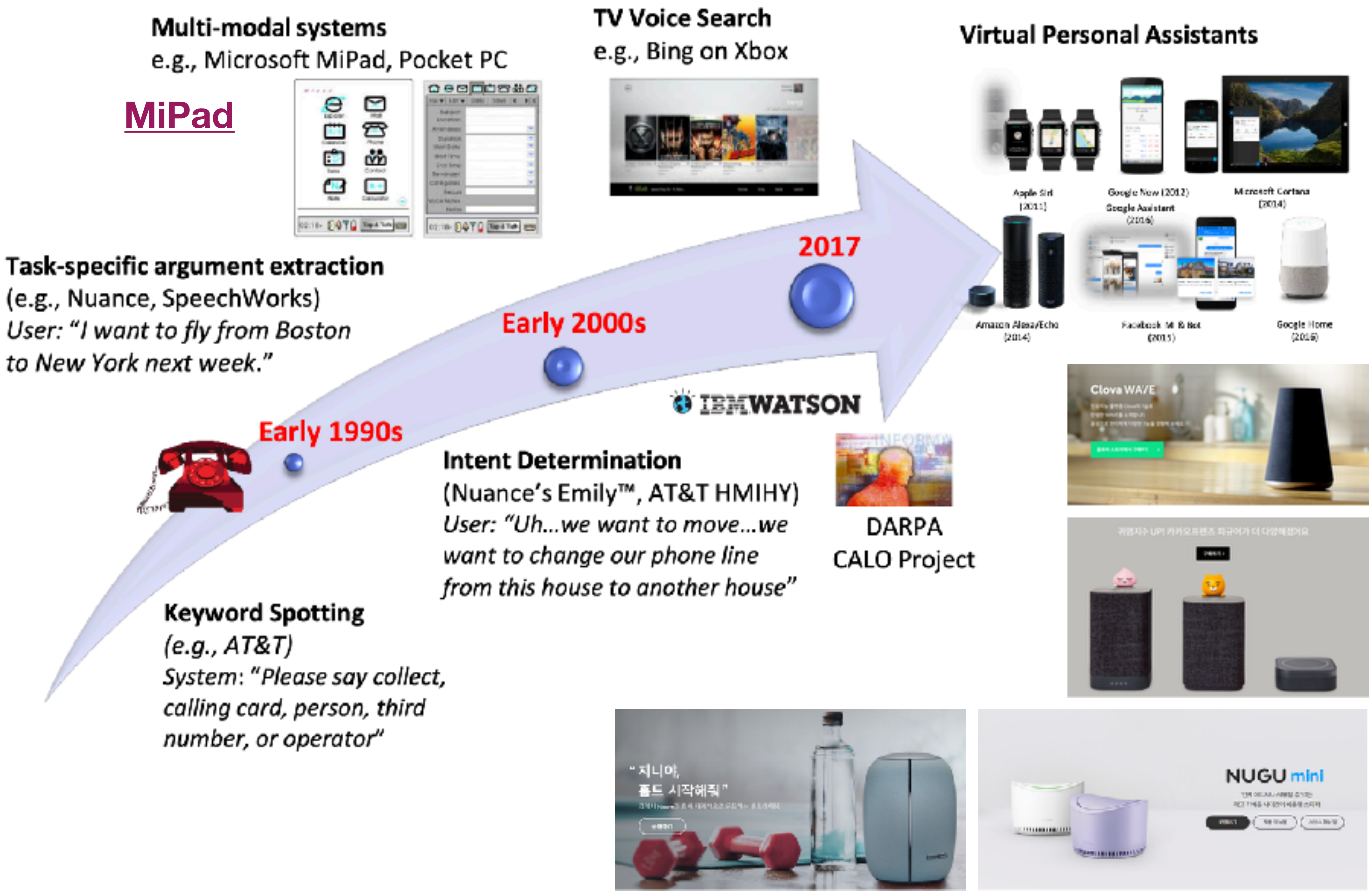
I. Introduction to dialog systems

II. Background
- Machine learning
- Deep learning and Neural networks

III. Deep learning for Natural Language
- Word embedding
- Language models

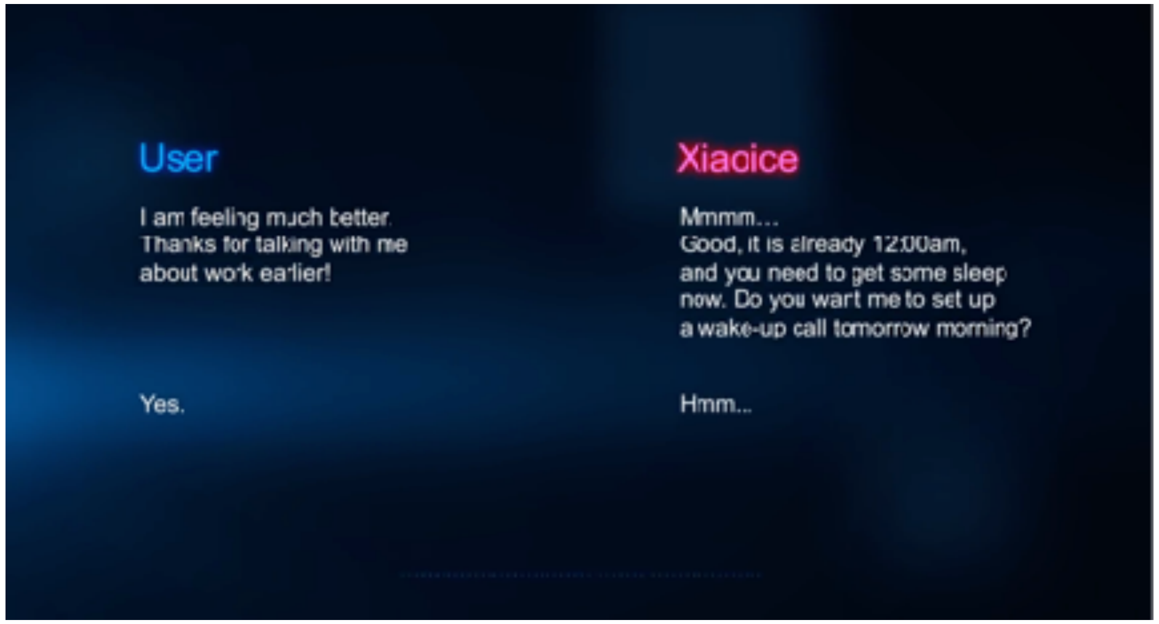
IV. Deep learning for Dialog systems
- SUMBT
- LaRL
- Challenges

Brief History of Dialogue Systems



Brief History of Dialogue Systems

Google, Duplex (2018)

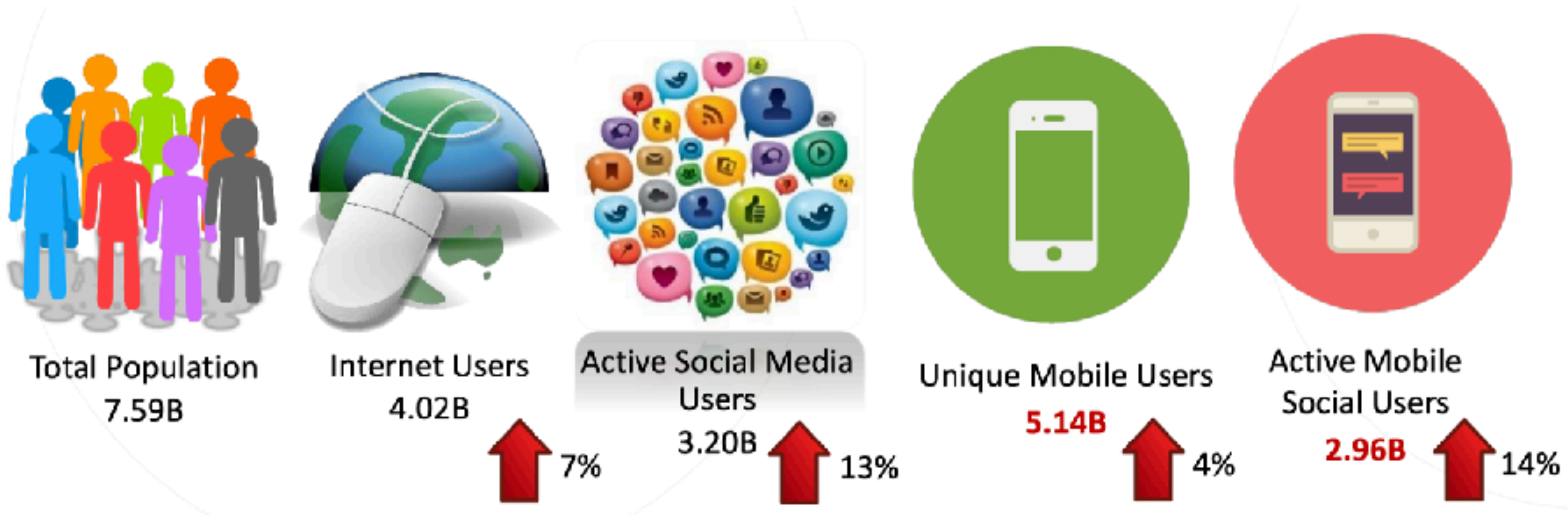


Microsoft, Xiaoice (2018)

Naver Line, Duet (2019)

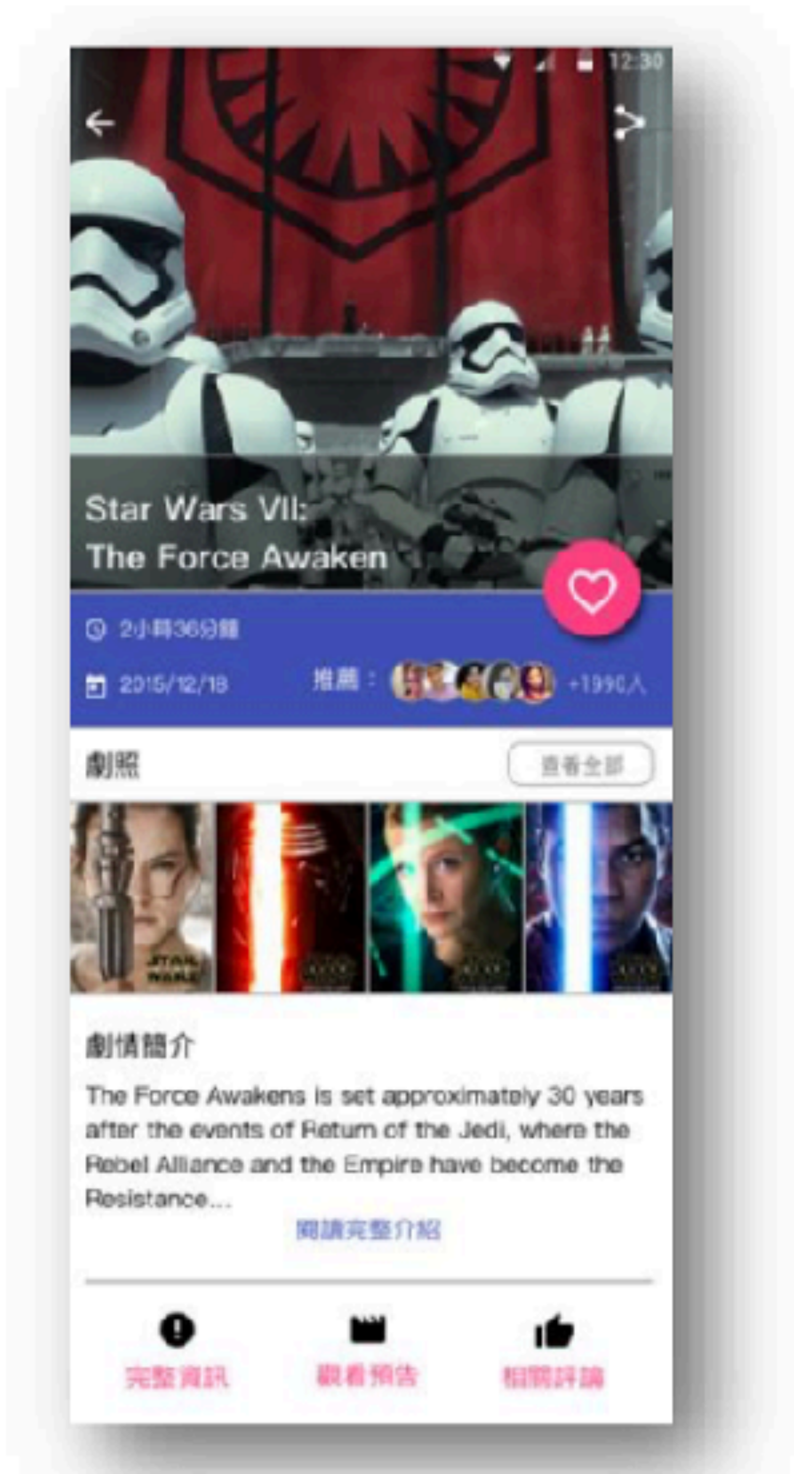
Why Natural Language?

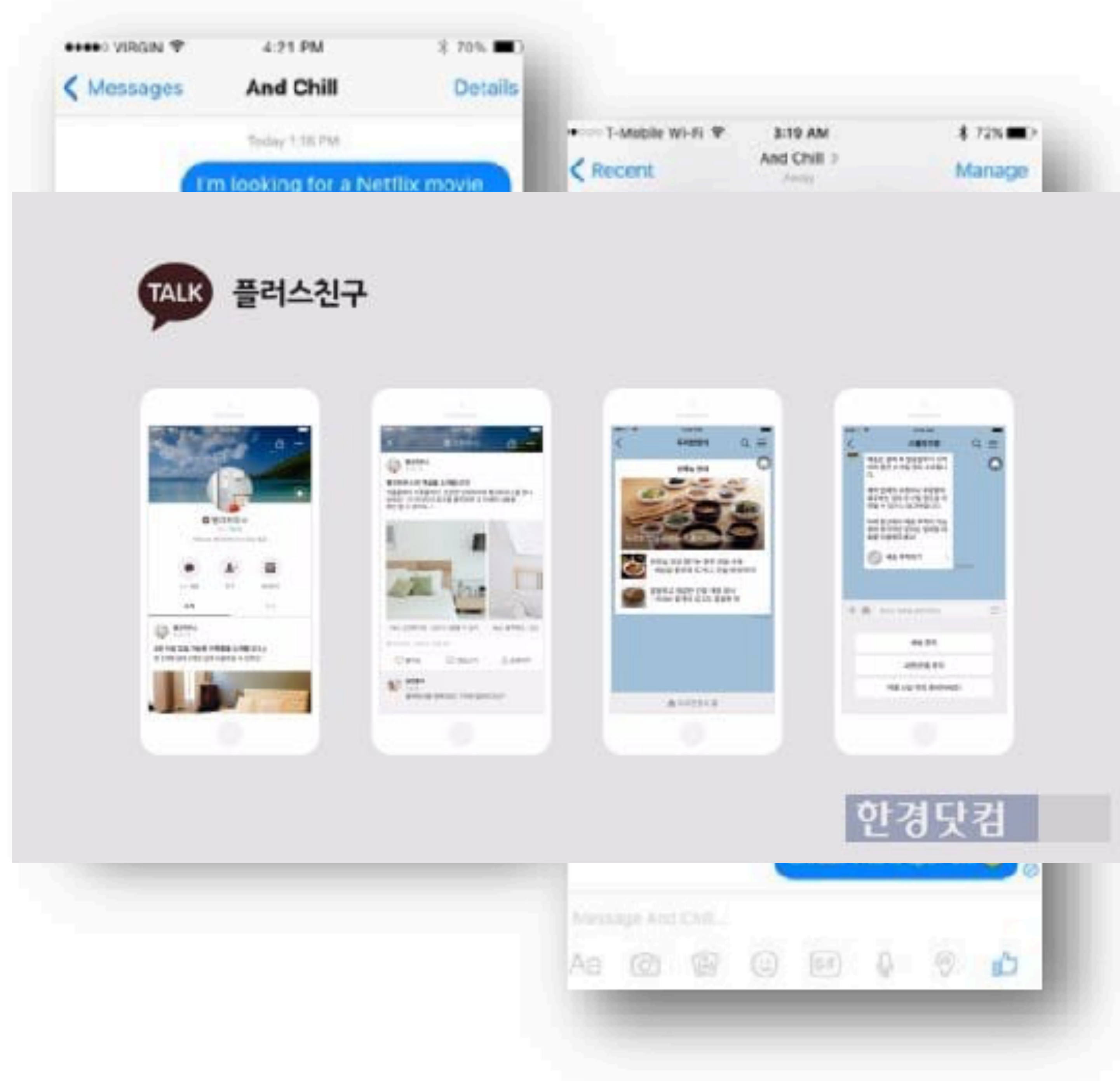
- Global Digital Statistics (2018 January)



The more **natural** and **convenient** input of devices evolves towards **speech**

GUI v.s. CUI (Conversational UI)





GUI v.s. CUI (Conversational UI)

	Website/APP's GUI	Msg's CUI
Situation	Navigation, no specific goal	Searching, with specific goal
Information Quantity	More	Less
Information Precision	Low	High
Display	Structured	Non-structured
Interface	Graphics	Language
Manipulation	Click	mainly use texts or speech as input
Learning	Need time to learn and adapt	No need to learn
Entrance	App download	Incorporated in any msg-based interface
Flexibility	Low, like machine manipulation	High, like converse with a human

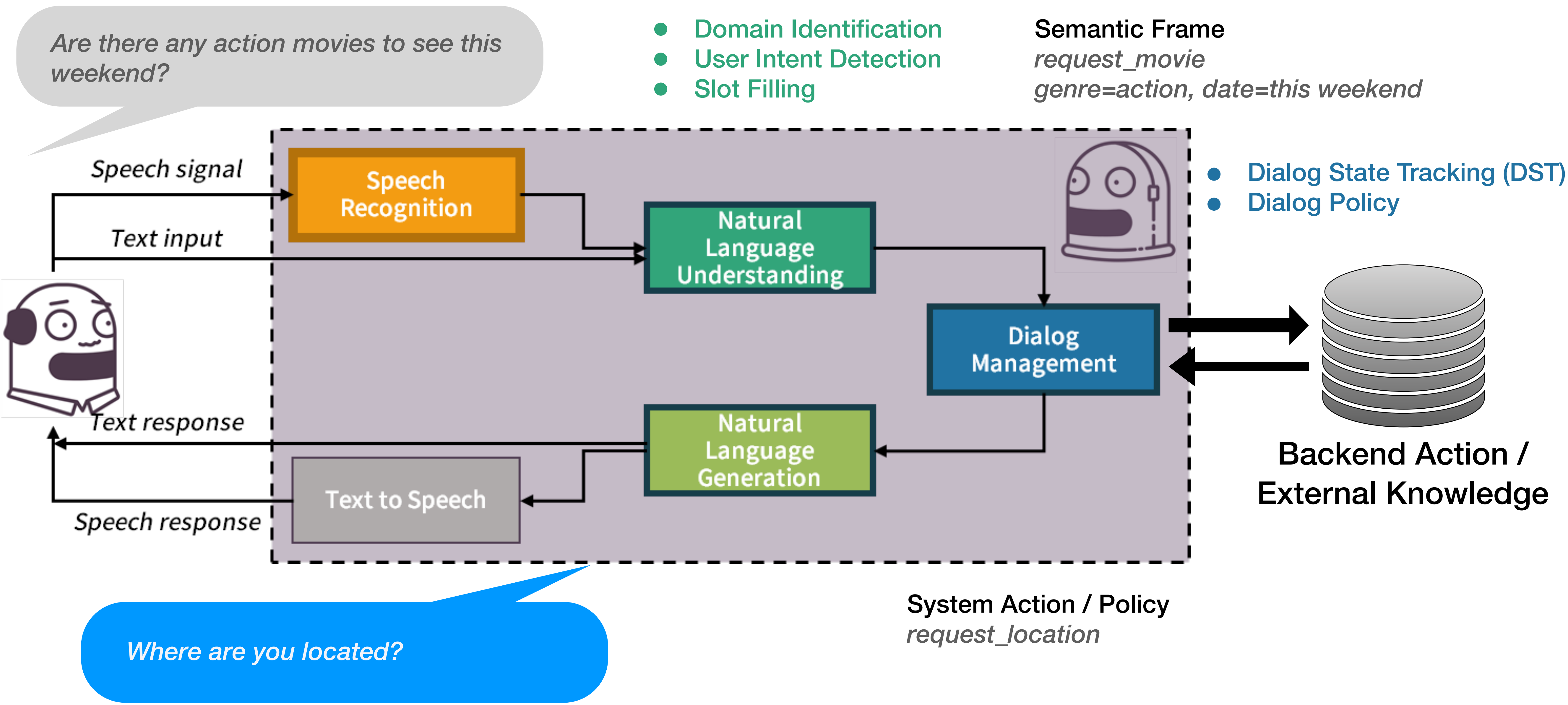
Material: https://deepdialogue.miulab.tw

Category of Dialogue Systems

User says:	Dialogue Category
• **I am smart**	→ **Chitchat**
• **I have a question** ***When Iron Man is dead?***	→ **QA**
• **I need to get this done** ***I want to book a restaurant***	→ **Goal-oriented**

Material: Jianfeng Gao & Michel Galley, Tutorial, ICML, 2019

Spoken Dialog Systems



Transition of NLP to Neural Approaches

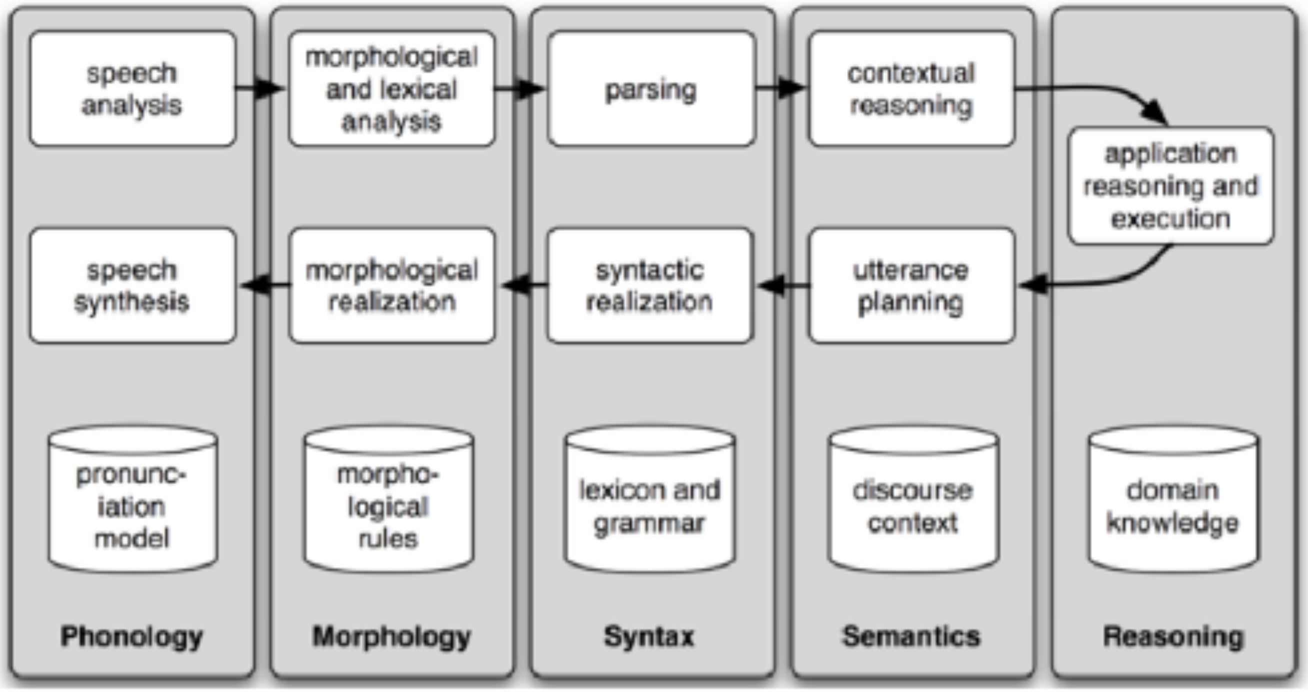
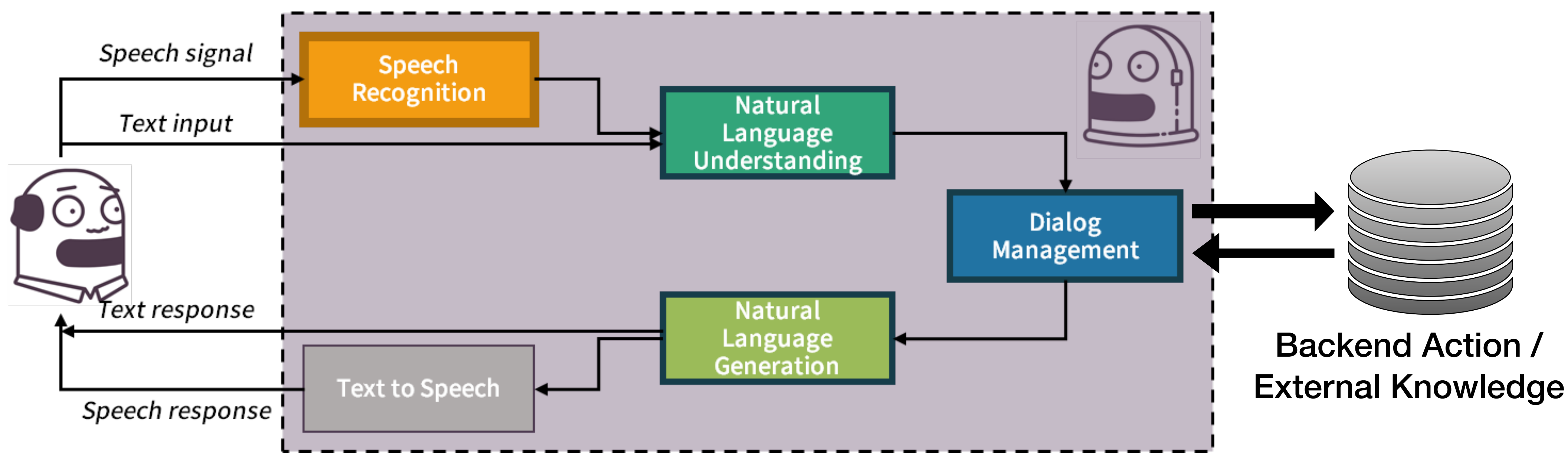


Figure 1.3: Traditional NLP Component Stack. Figure credit: Bird et al. (2009).

Neural Model for Each Module



Transition of NLP to Neural Approaches

Symbolic Space

- Knowledge is explicitly represented using words/relations/templates
- Reasoning is based on keyword matching, sensitive to paraphrase alternations
- Interpretable and efficient in execution but difficult to train E2E.

Input: Query

Symbolic → Neural
Encoding the query/knowledge

E2E training via back propagation

Errors

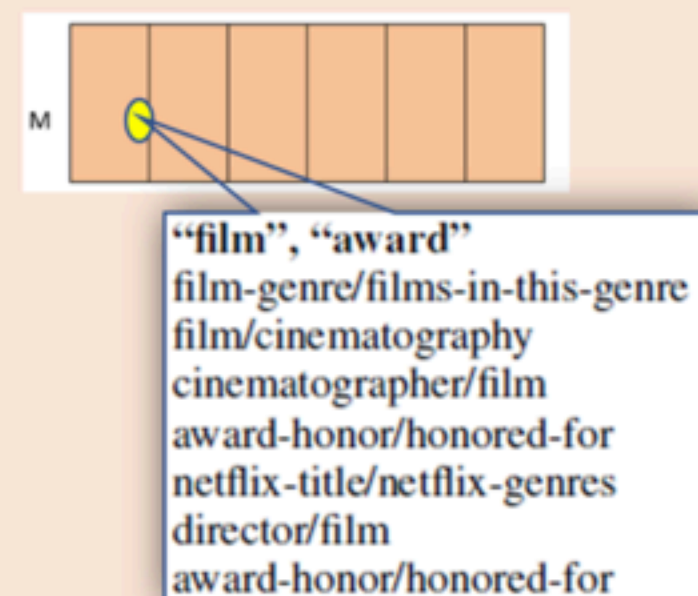
Reasoning in neural space to generate answer vector

Output: Answer

Neural → Symbolic
Decoding the answer in NL

Neural Space

- Knowledge is implicitly represented by semantic classes as cont. vectors
- Reasoning is based on semantic matching, robust to paraphrase alternations
- Easy to train E2E, but uninterpretable and inefficient in execution



Outline

I. Introduction to dialog systems

II. Background

- Machine learning
- Deep learning and Neural networks

III. Deep learning for Natural Language

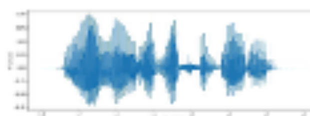
- Word embedding
- Language models

IV. Deep learning for Dialog systems

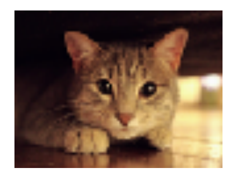
- SUMBT
- LaRL
- Challenges

Machine Learning ≈ Find appropriate function

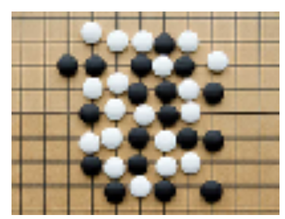
- Speech Recognition

$f($  $)$ = 안녕하세요

- Image classification

$f($  $)$ = Cat

- Go Playing

$f($  $)$ = 5-5 (next movement)

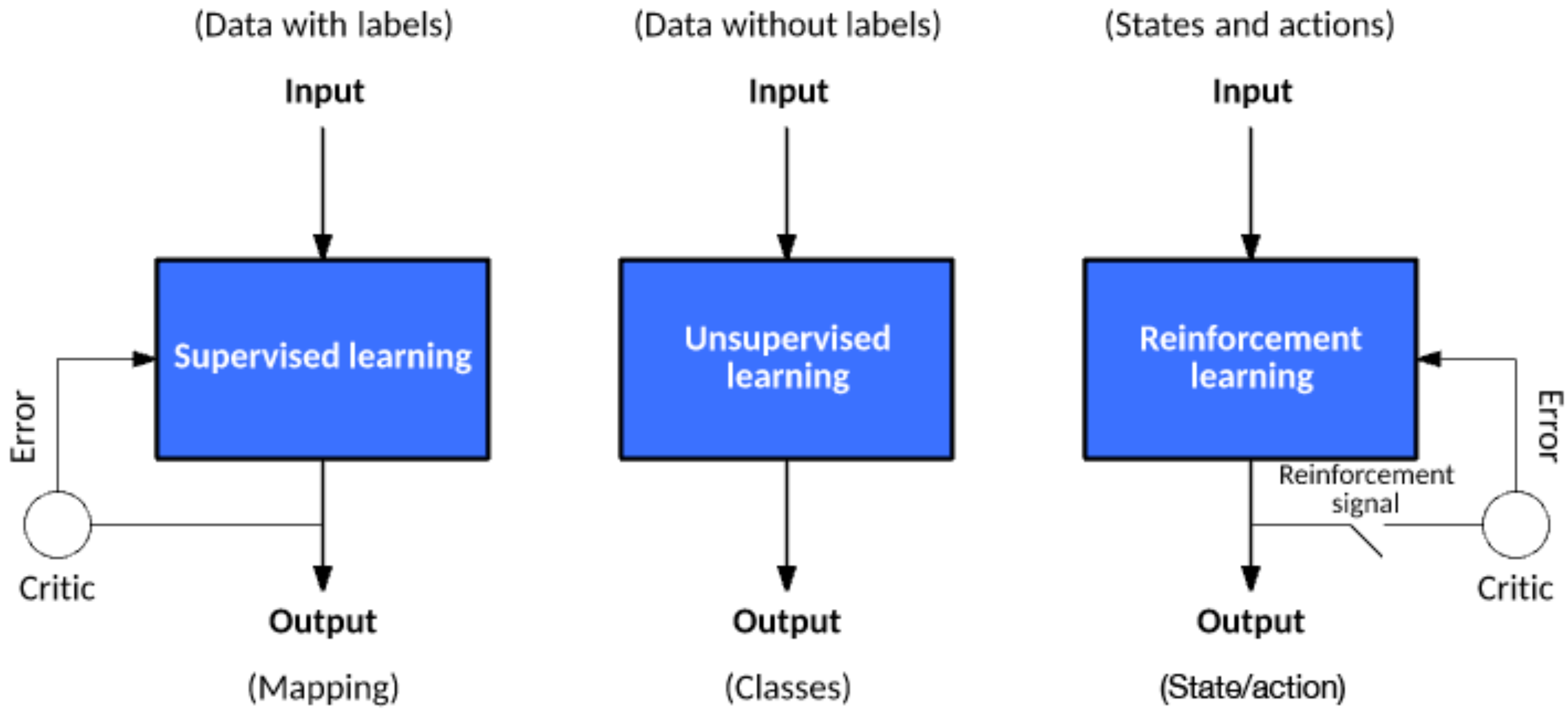
- Chat Bot

$f($ “오늘 점심 메뉴 뭐지?” $)$ = “오늘 점심 식단은…”

Input → **Model f** → **Output**

Material: https://deepdialogue.miulab.tw

Types of Machine Learning

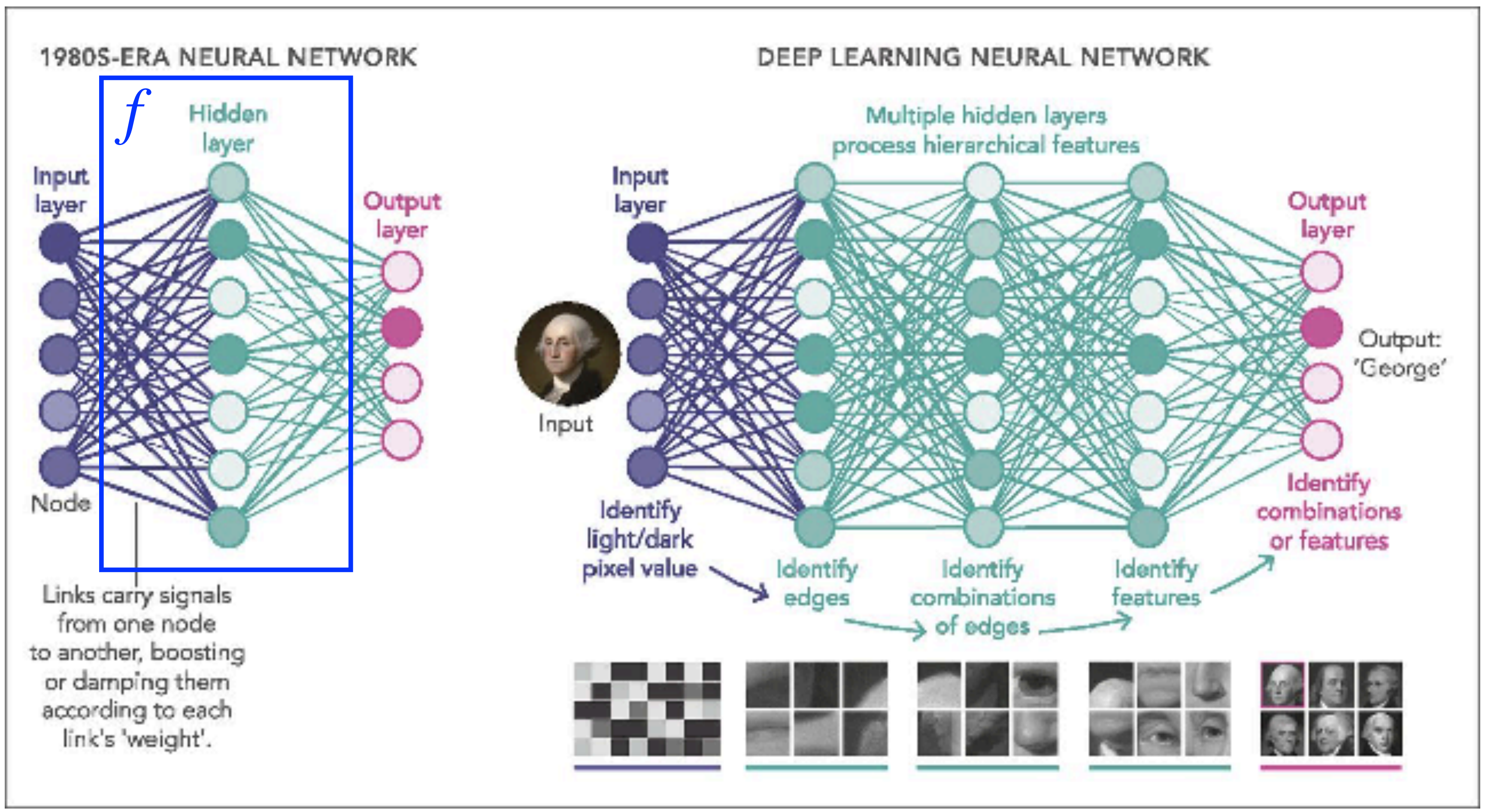


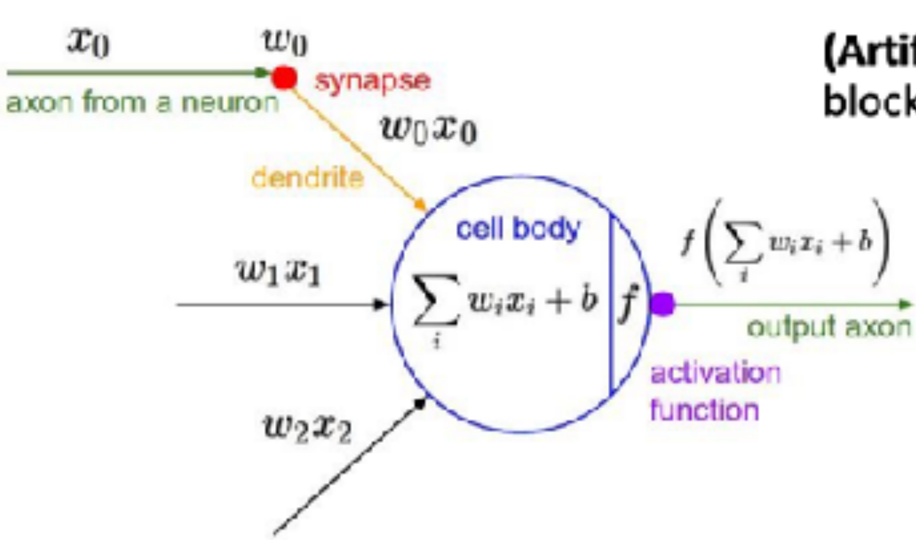
- **Classification**
  - **Image classification**
  - **Sentiment text classification**
- **Regression**
  - **Weather forecasting**
  - **Market forecasting**

- **Clustering**
  - **Recommender system**
- **Dimension reduction**
  - **Meaningful compression**
  - **Feature extraction**
- **Topic modeling**

- **Robot Navigation**
- **Game AI**
- **Dialog Policy Learning**

Neural Networks and Deep Learning

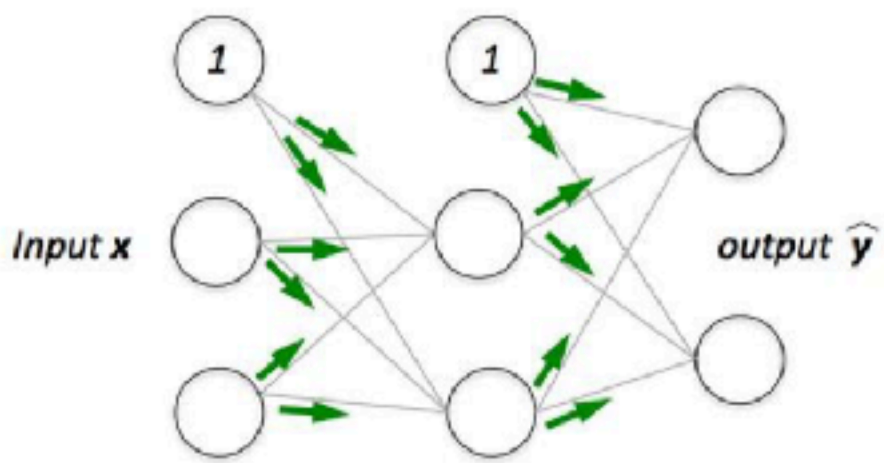




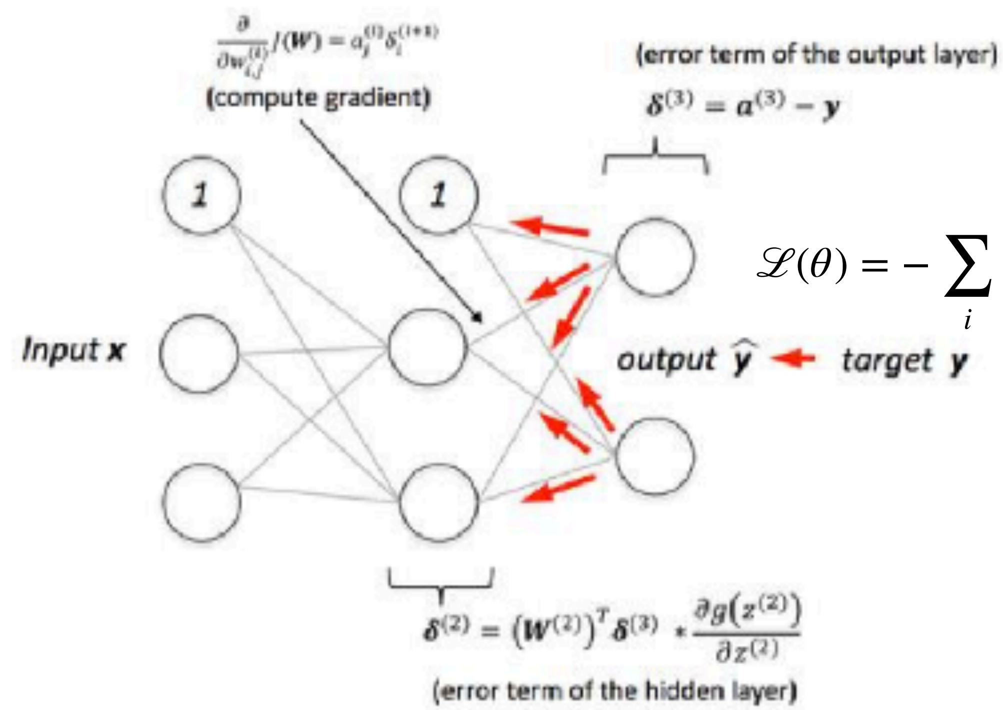
(Artificial) Neuron: computational building block for the "neural network"

Training Neural Nets = Optimization

Forward pass



Error Back Propagation



$$\mathscr{L}(\theta) = -\sum_i (y_i - f(x_i))^2$$

- Update the **weights** and **biases** to decrease **loss function**
  1. Forward pass: compute network output and error
  2. Backward pass: compute gradient by EBP
  3. Update weights by gradient descent

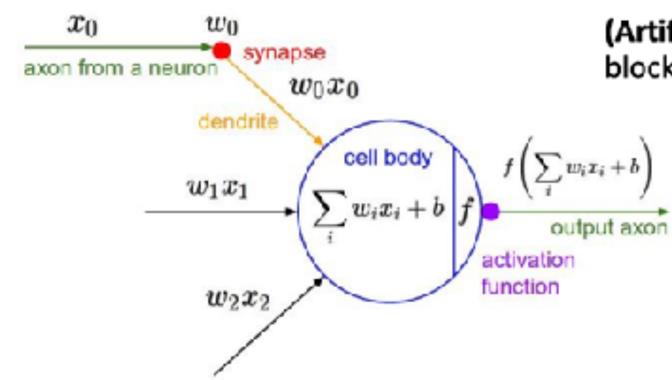
$$w^{(t+1)} \leftarrow w^{(t)} - \eta \frac{\partial \mathscr{L}}{\partial w^{(t)}}$$

Three things defining deep learning

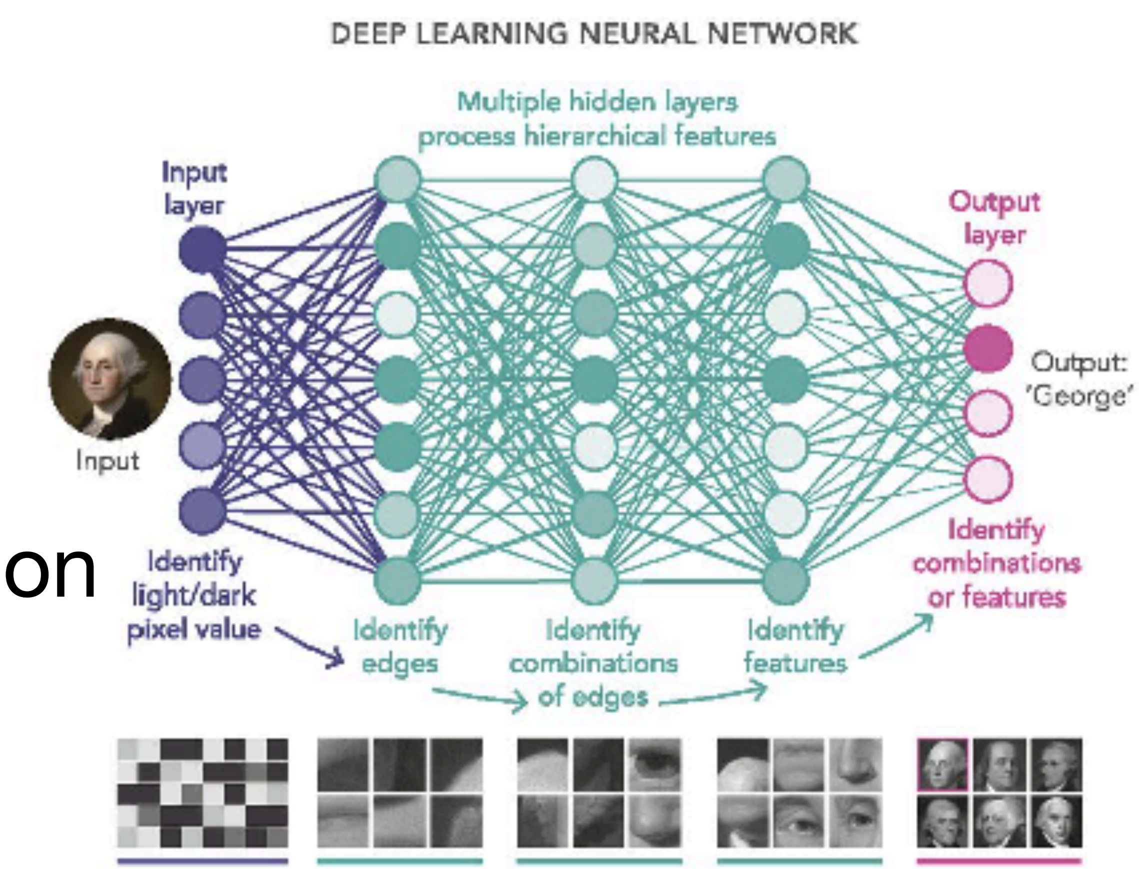
1. Neuron type
   (activation function)

2. Architecture

3. Learning algorithm:
   Loss function & Optimization

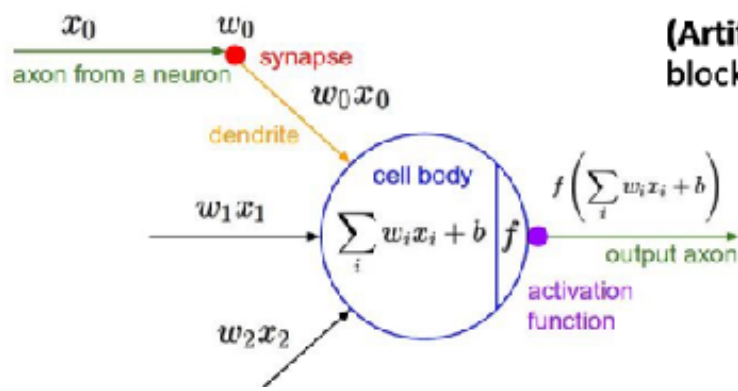


(Artificial) Neuron: computational building block for the "neural network"

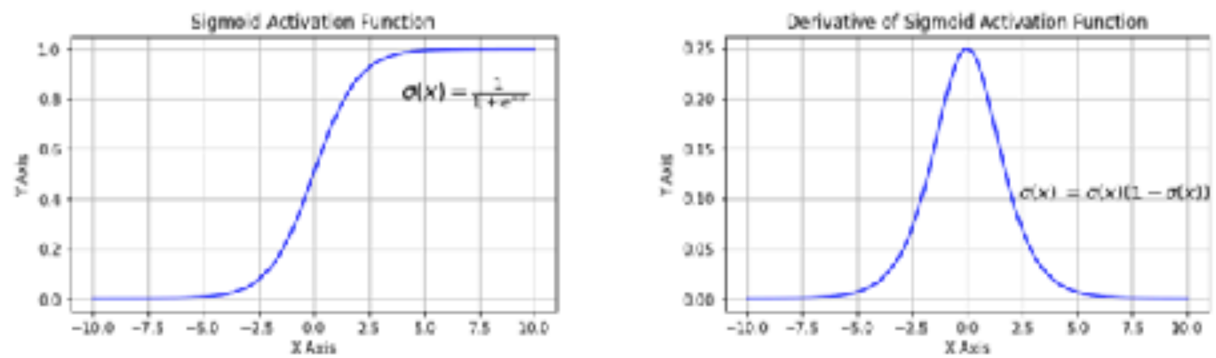


$$\mathscr{L}(\theta) = -\sum_i (y_i - f(x_i))^2$$

1. Neuron type (activation function)

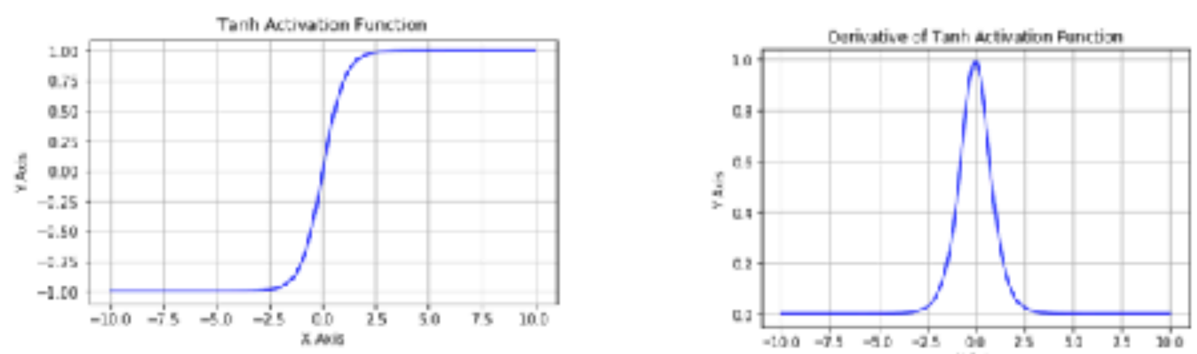


(Artificial) Neuron: computational building block for the "neural network"



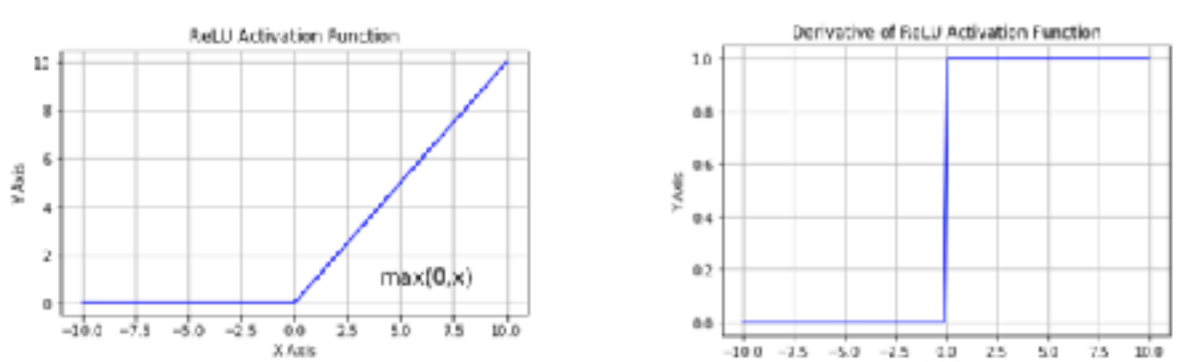
Sigmoid

- Vanishing gradients
- Not zero centered



Tanh

- Vanishing gradients

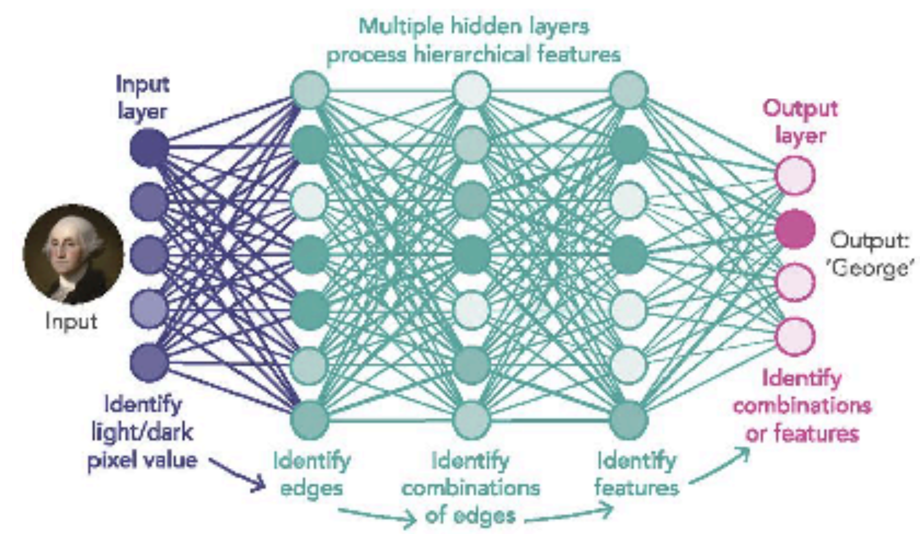


ReLU

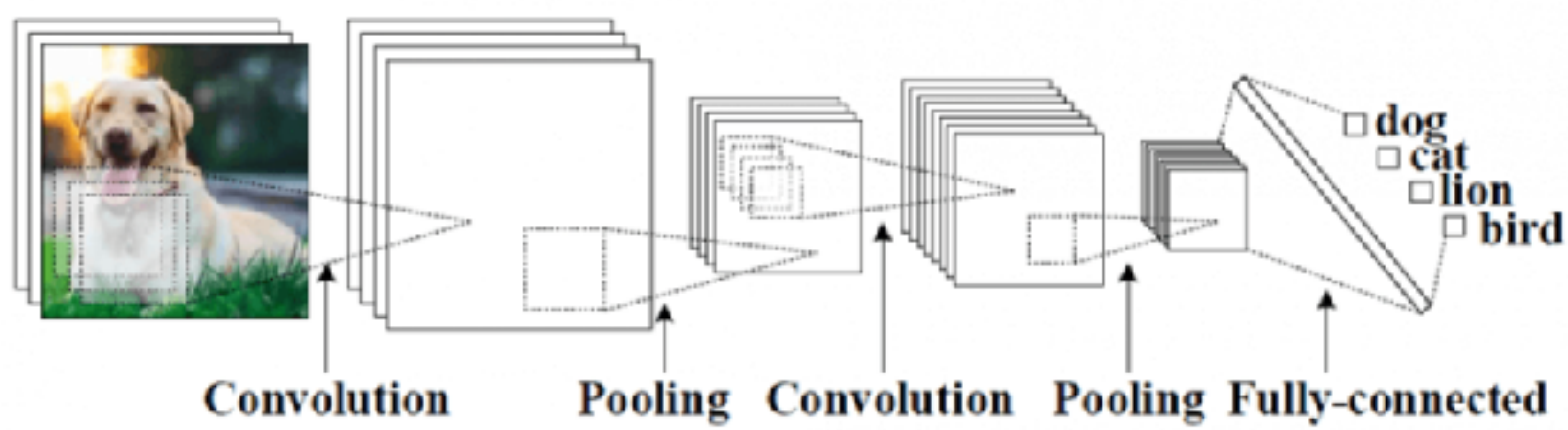
- Not zero centered

2. Architecture

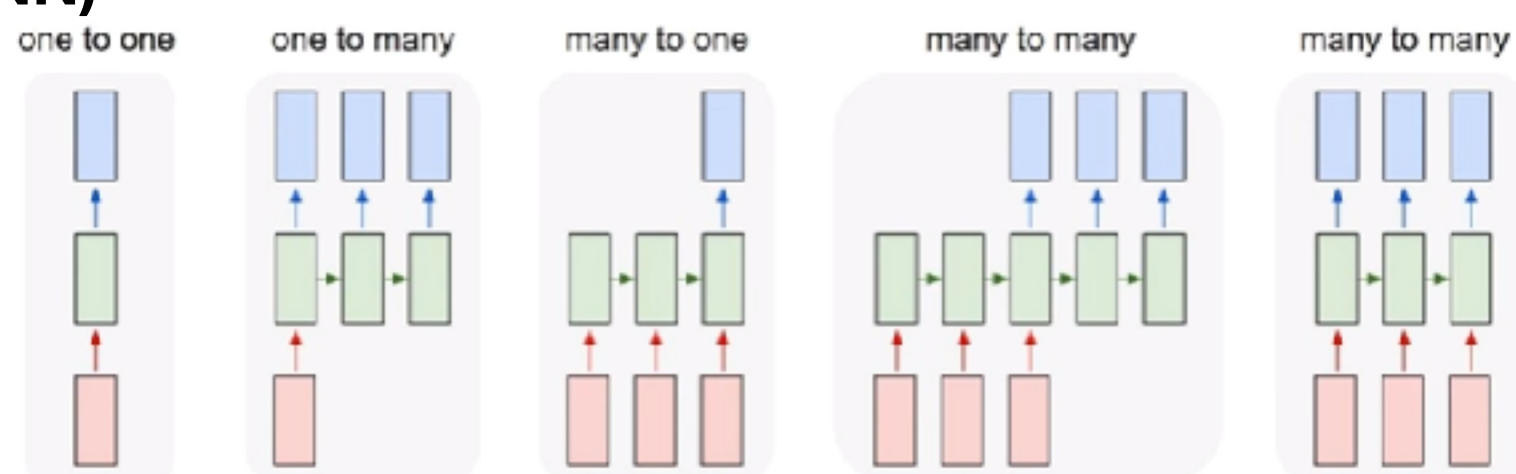
1. **Deep Neural Networks (DNN)**
   - **Fully Connected Layers**



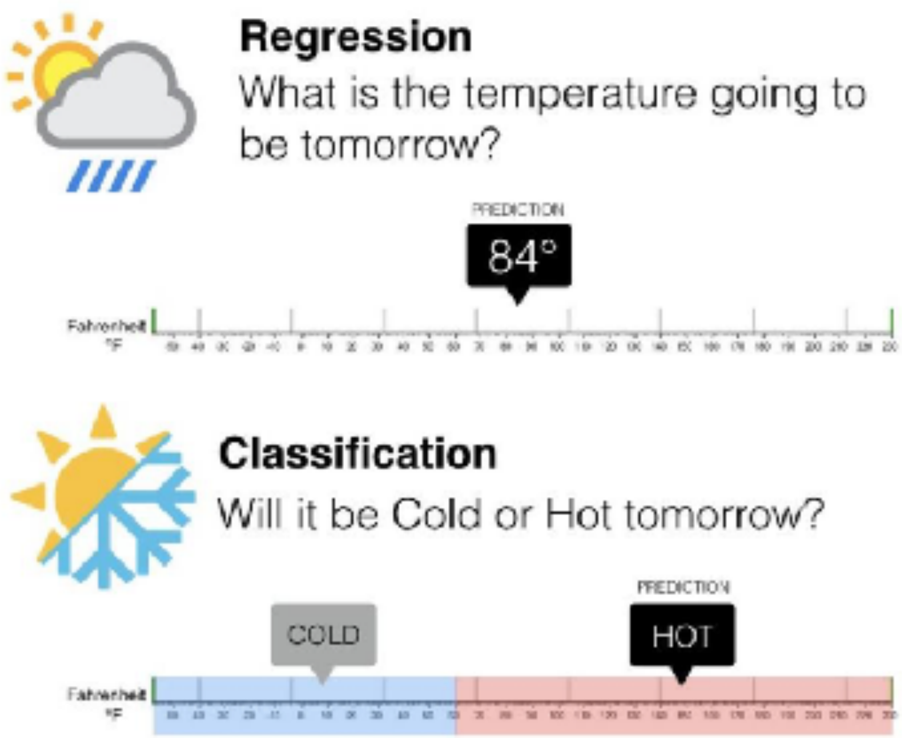
2. **Convolutional Neural Networks (CNN)**
   - **Weight sharing and pooling**
   - **Spatial data: Image**



3. **Recurrent Neural Networks (RNN)**
   - **Time series data**
   - **Speech, Language, Video**



3. Learning algorithm: Loss function & Optimization



- Loss function quantifies gap between **prediction** and **ground truth (labels)**
- For regression:
  - Mean Squared Error (MSE)
- For classification:
  - Cross Entropy Loss
    (a.k.a. Negative Log Likelihood)

Mean Squared Error

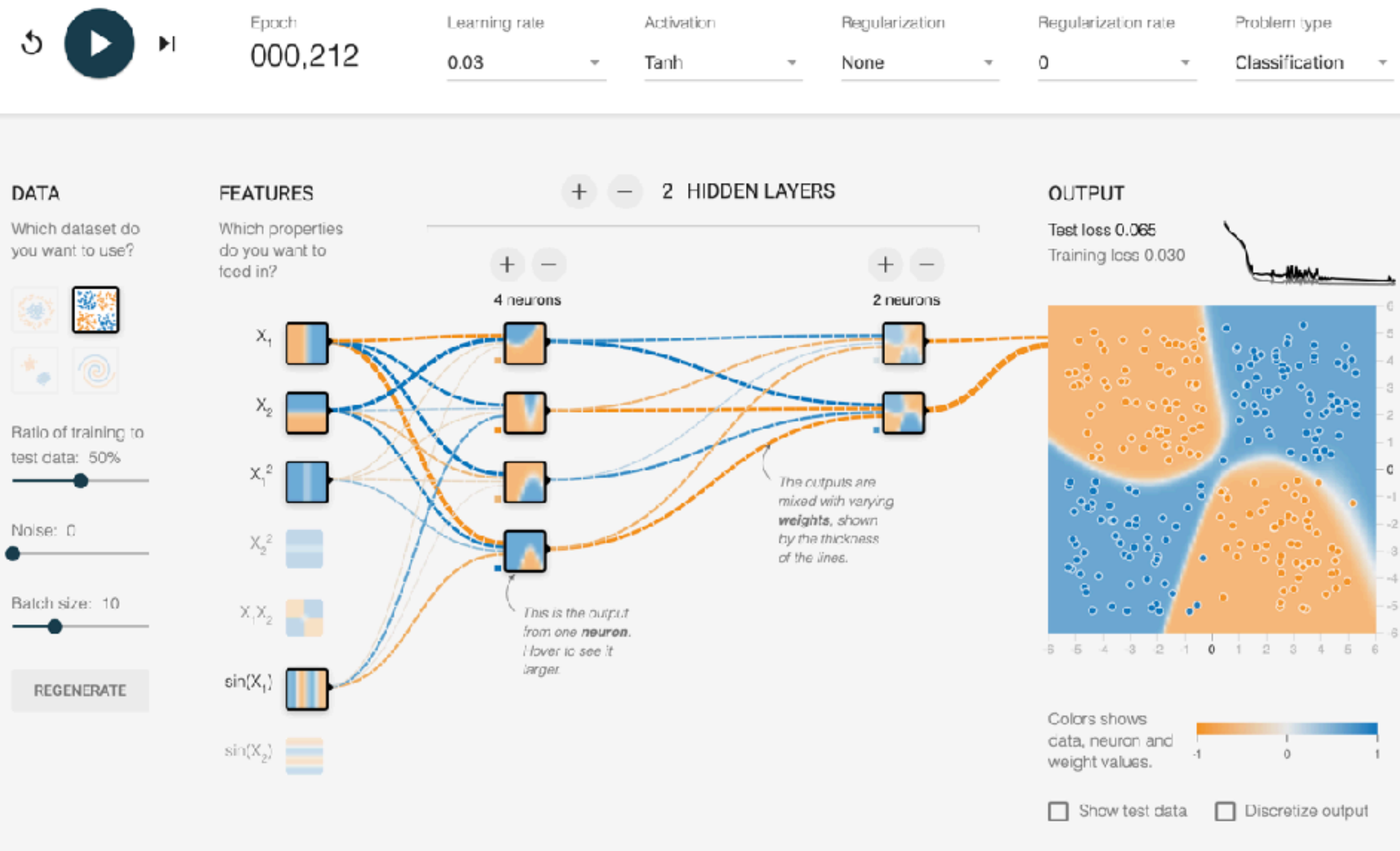
$$\mathscr{L}(\theta) = -\frac{1}{N}\sum_{i}(t_i - f(x_i))^2$$

Cross Entropy Loss

$$\mathscr{L}(\theta) = -\sum_{i}^{C} t_i \log(p(y \mid x_i))$$

- **Optimization: Stochastic gradient descent (SGD)**

Neural Network Playground



http://playground.tensorflow.org

Outline

I. Introduction to dialog systems

II. Background

- Machine learning
- Deep learning and Neural networks

III. Deep learning for Natural Language

- Word embedding
- Language models

IV. Deep learning for Dialog systems

- SUMBT
- LaRL
- Challenges

Natural Language Process Tasks

- Text Classification
  - Sentiment classification: *I love it* → *positive? negative?*
- Language Generation
  - Machine translation: 사랑합니다 → *Love it*
  - Image captioning:

"man in black shirt is playing guitar."

- Question-answering
  (Machine reading comprehension)
- POS tagging
- Chungking
- …

Airport

The Stanford Question Answering Dataset

An airport is an aerodrome with facilities for flights to take off and land. Airports often have facilities to store and maintain aircraft, and a control tower. An airport consists of a landing area, which comprises an aerially accessible open space including at least one operationally active surface such as a runway for a plane to take off or a helipad, and often includes adjacent utility buildings such as control towers, hangars and terminals. Larger airports may have fixed base operator services, airport aprons, air traffic control centres, passenger facilities such as restaurants and lounges, and emergency services.

What is an aerodome with facilities for flights to take off and land?
airport

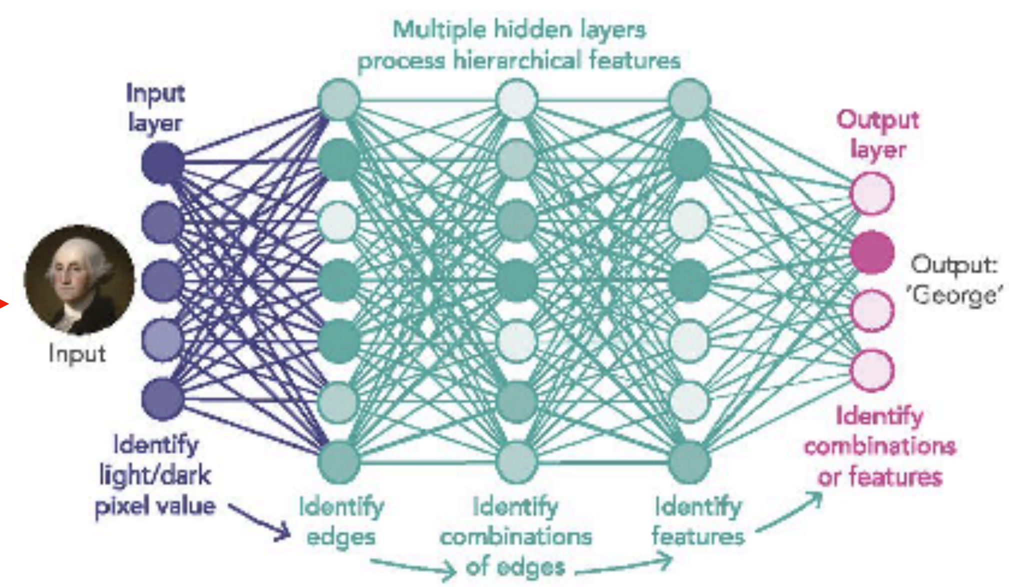
What is an aerially accessible open space that includes at least one active surface such as a runway or a helipad?
landing area

What is an airport?
aerodrome with facilities for flights to take off and land

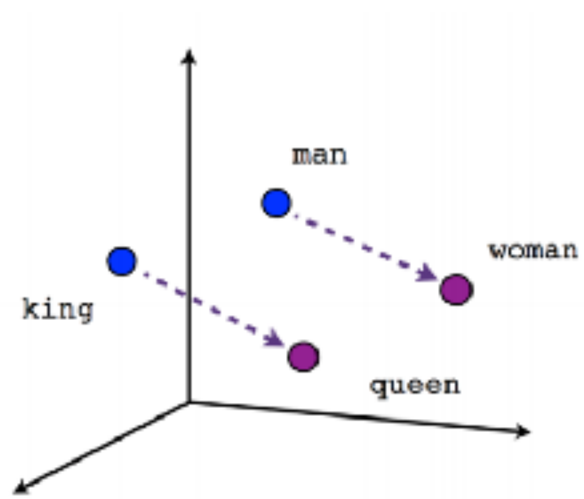
Word Embeddings (word2vec)

- How to represent word symbols as (semantic) vectors?

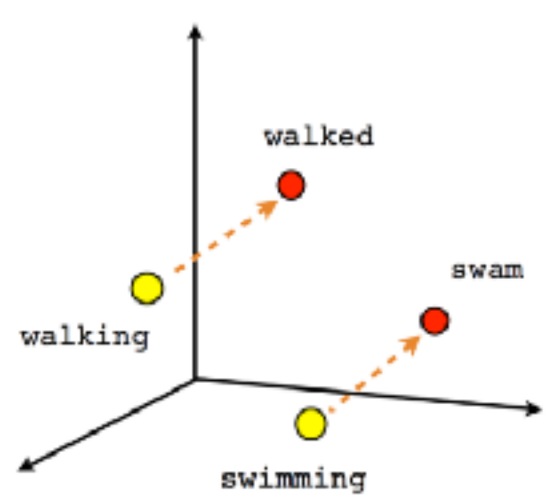
love
it
!
?



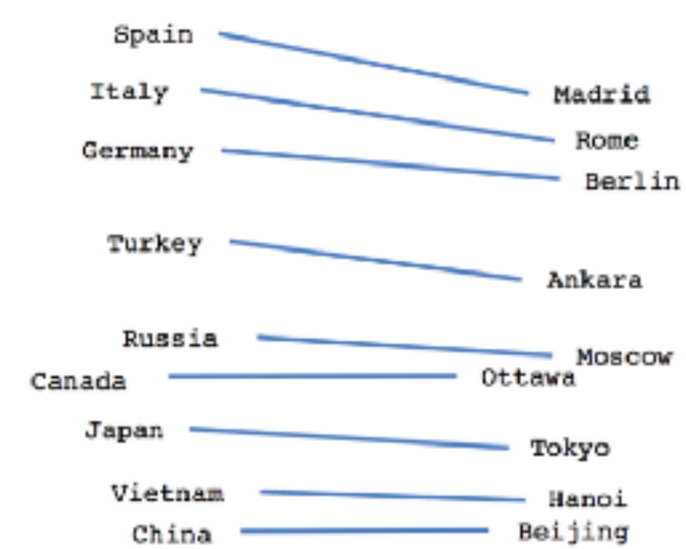
Pos
Neg



Male-Female



Verb tense

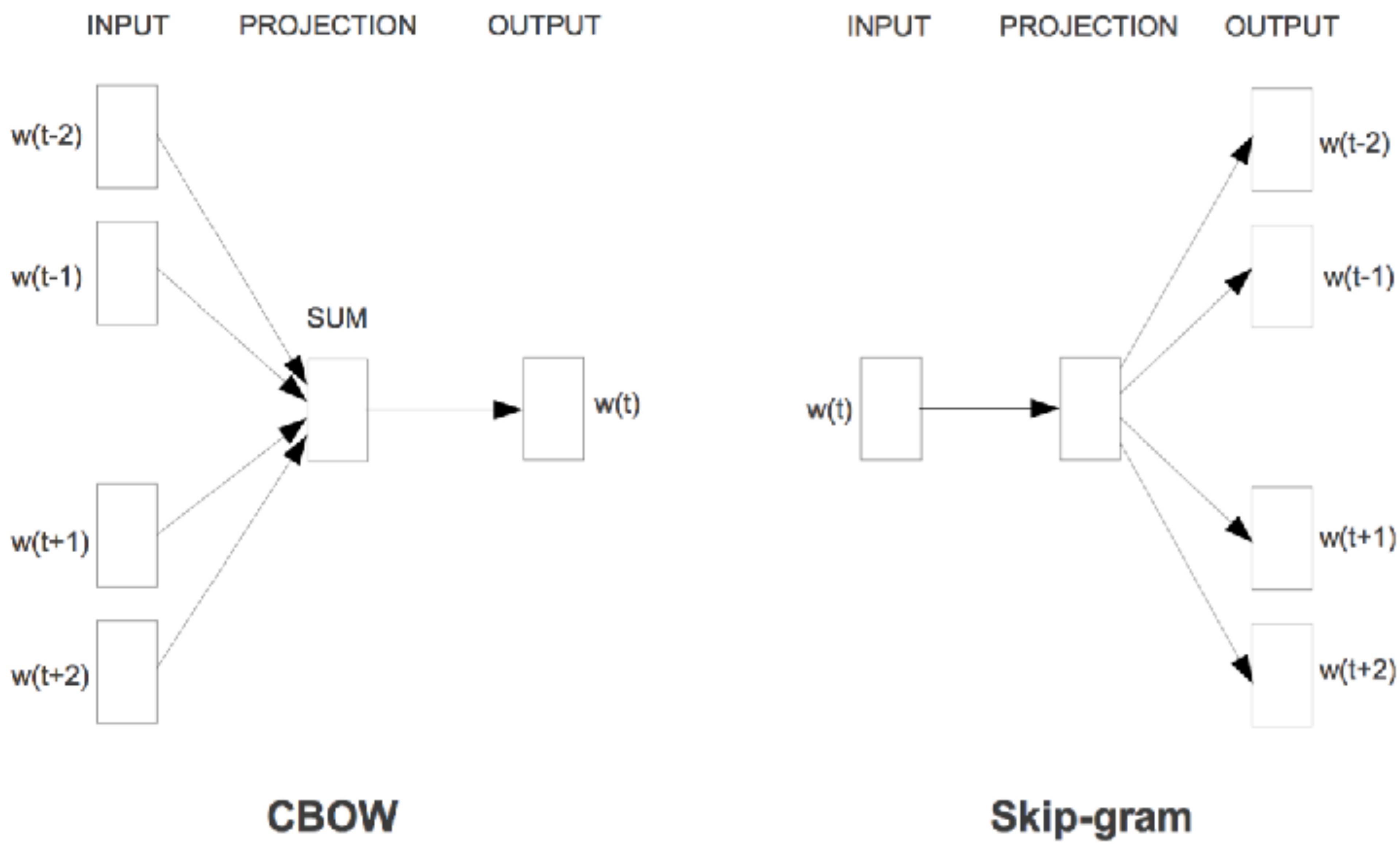


Country-Capital

https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space

Word Embeddings (word2vec)

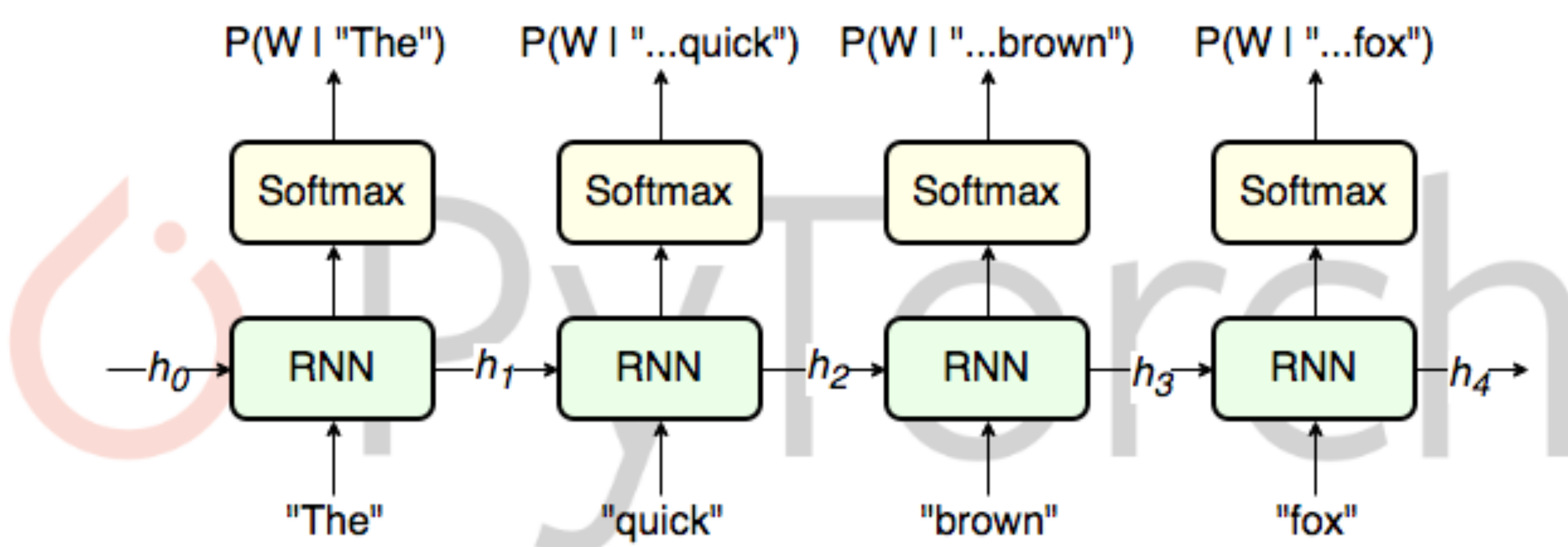
- Learn the meaning of a word from its neighborhoods!



T. Mikolove et al., Efficient Estimation of Word representations is vector space, 2013.

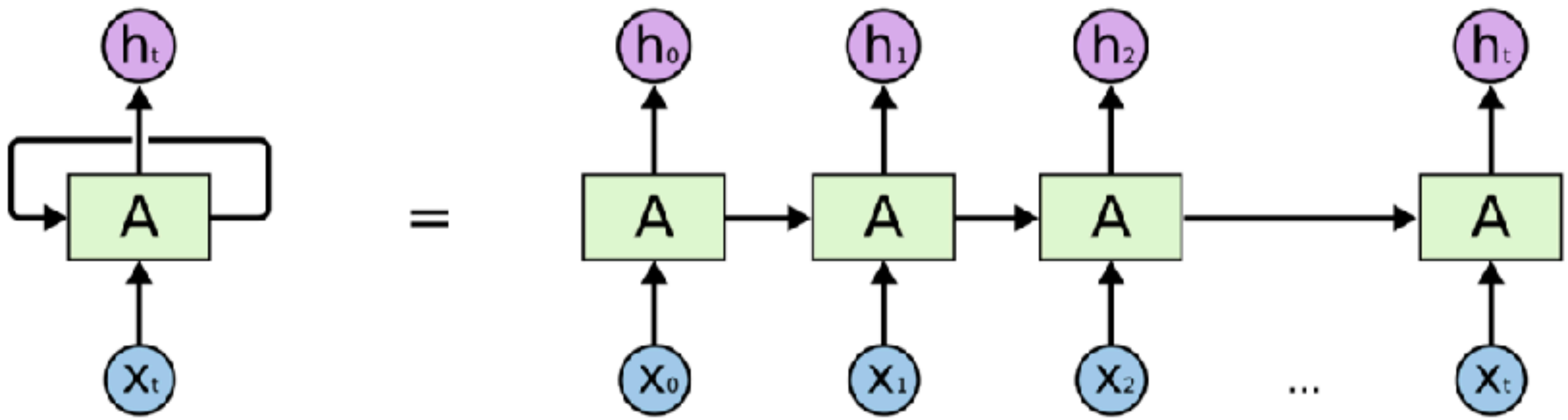
Language Model

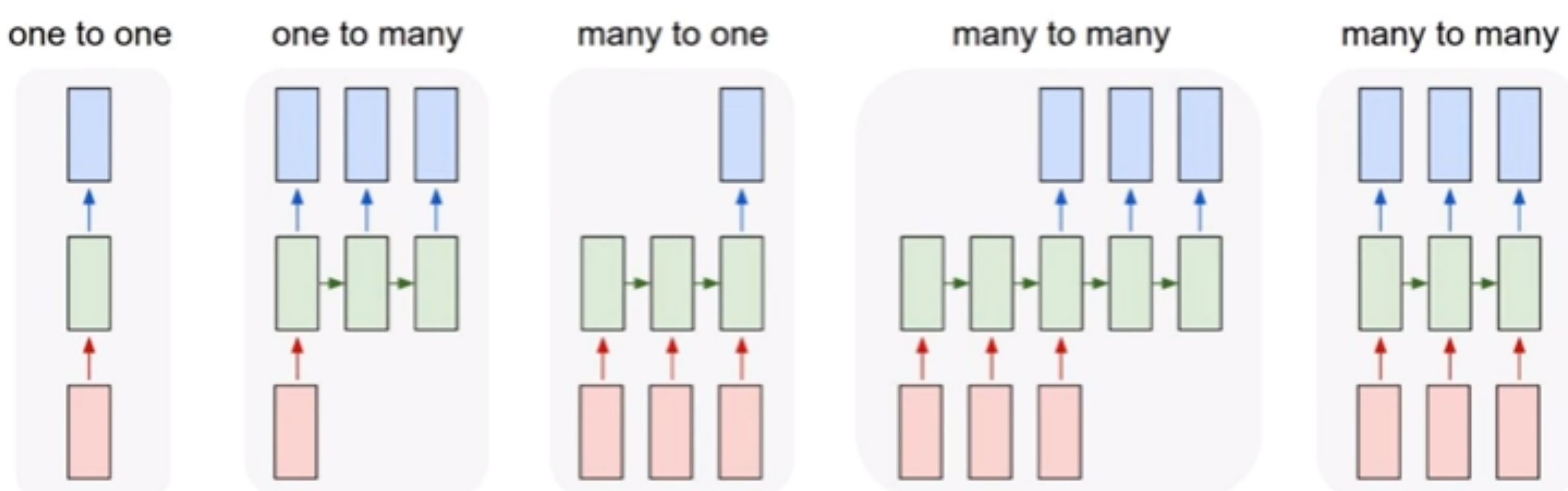
- Probability of a sequence of m words: $p(w_1, w_2, \ldots w_m)$
  - Application: Choose the next word: $p(w_{m+1} \mid w_{1,\ldots,m})$
- N-Gram LM
  - $p(w_{m+1} \mid w_{m,m-1}) = \dfrac{count(w_{m+1}, w_m, w_{m-1})}{count(w_m, w_{m-1})}$ (tri-gram)
  - Count based approach has weakness on *unseen word sequence*
  - Fixed width context
- Neural Language Model
  - RNNLM (Mikolov, 2010)



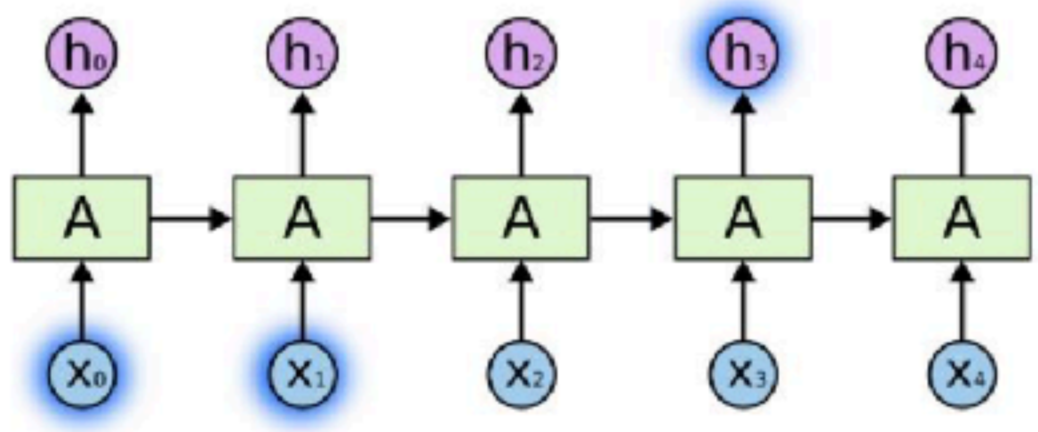
Recurrent Neural Networks

$$\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1}) = \sigma(W_x \mathbf{x}_t + W_h \mathbf{h}_{t-1} + \mathbf{b})$$



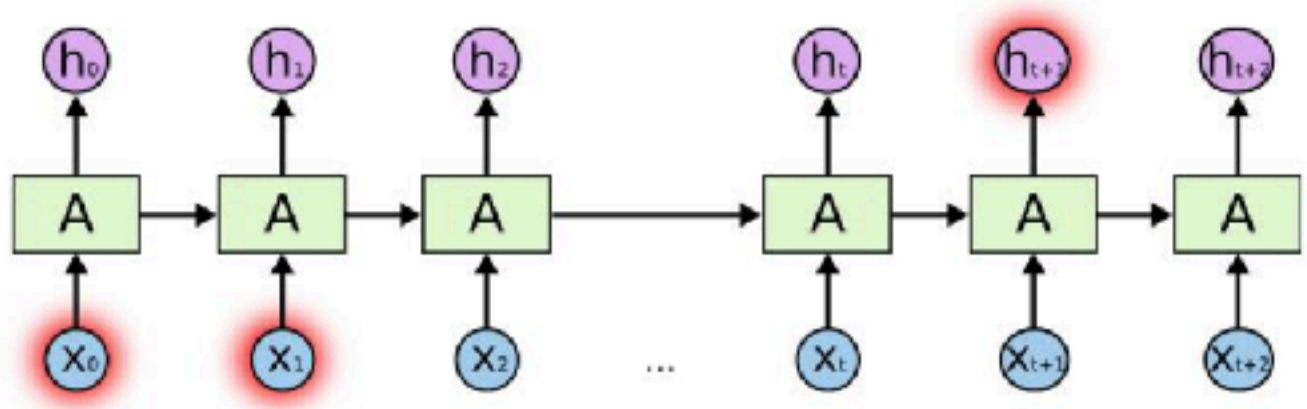


Long-Term Dependency



Context

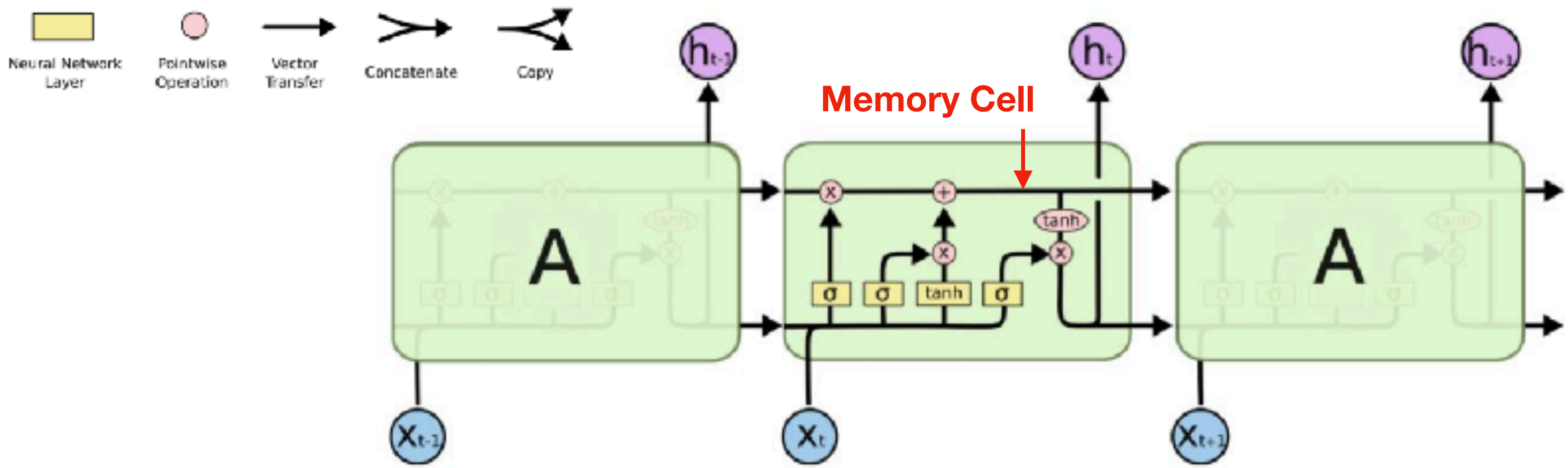
- Short-term dependence:
  Bob is eating an **apple.**
- Long-term dependence:
  Bob likes **apples**. He is hungry and decided to have a snack. So now he is eating an **apple.**



In theory, vanilla RNNs can handle arbitrarily long-term dependence.

In practice, it's difficult.

Long Short-Term Memory (LSTM) Networks

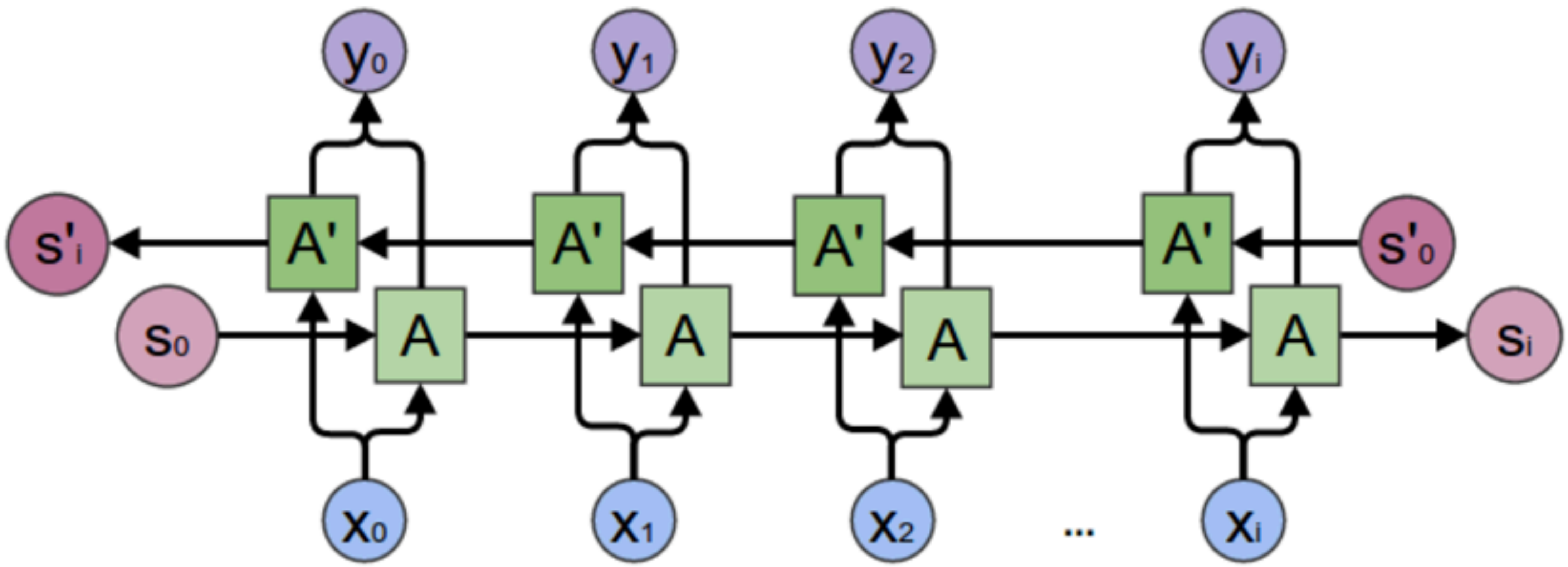


Pick what to **forget** and what to **remember**!

Previous output (h_{t-1}) and **new data** (x_t) are fed into the LSTM layers

1. Decide what to ***forget*** from the memory cell (forget gate)
2. Decide what to ***remember*** from the data and previous output (input gate)
3. Decide what to ***output*** (output gate)

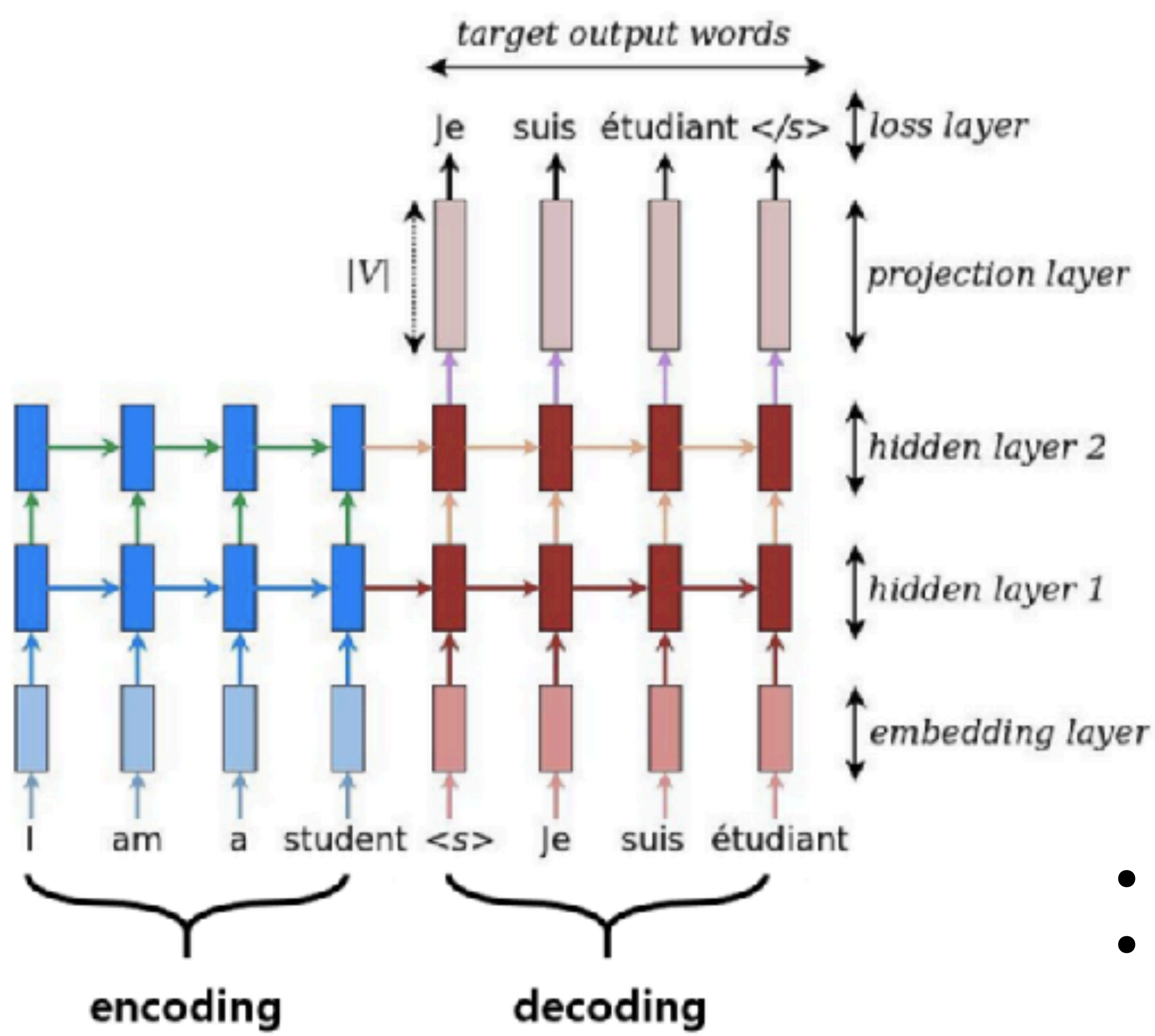
Bi-directional RNN



- Learn representations from both past and future time steps

Encoder-decoder architecture

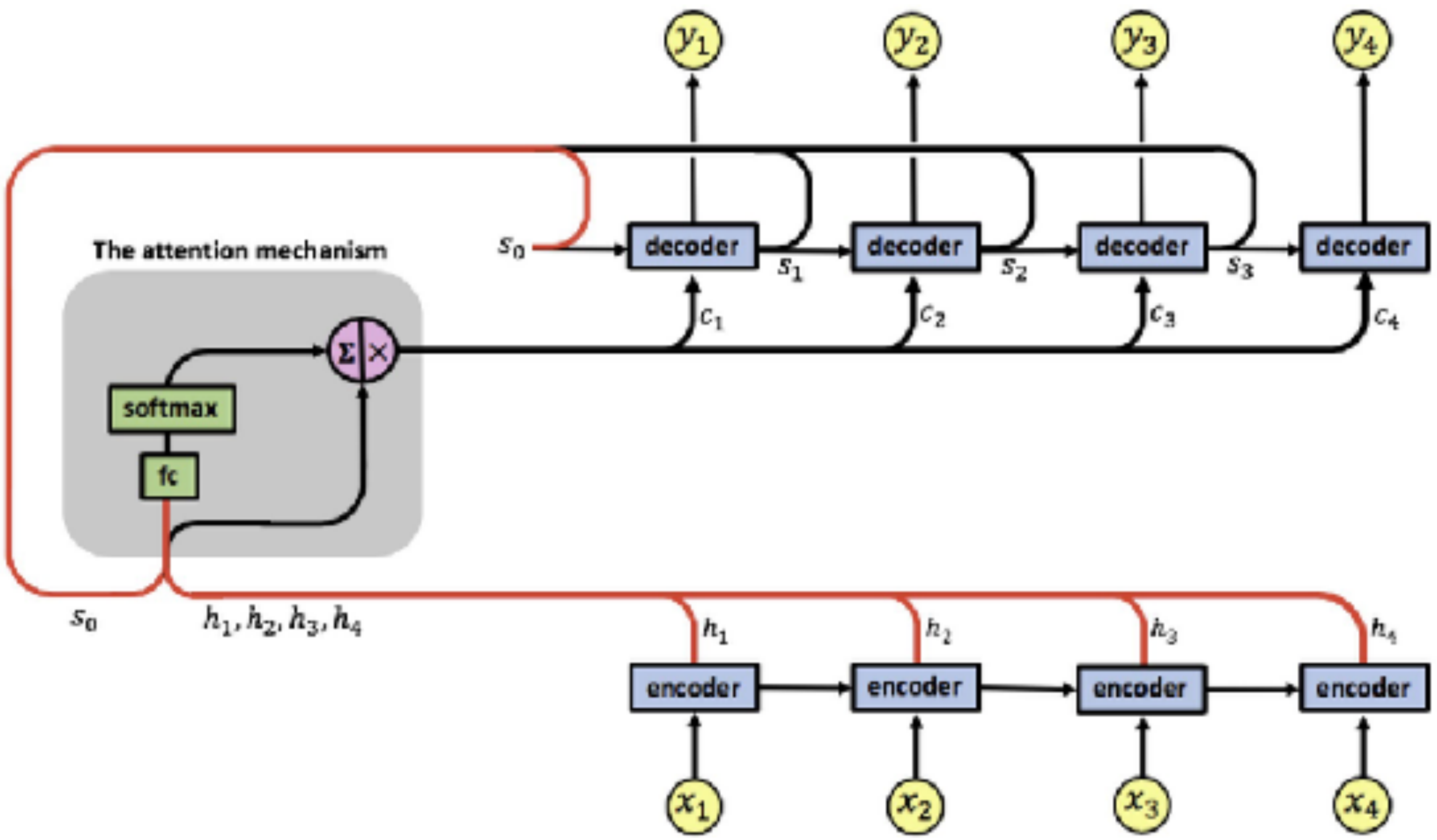
- Sequence-to-sequence (Seq2seq)

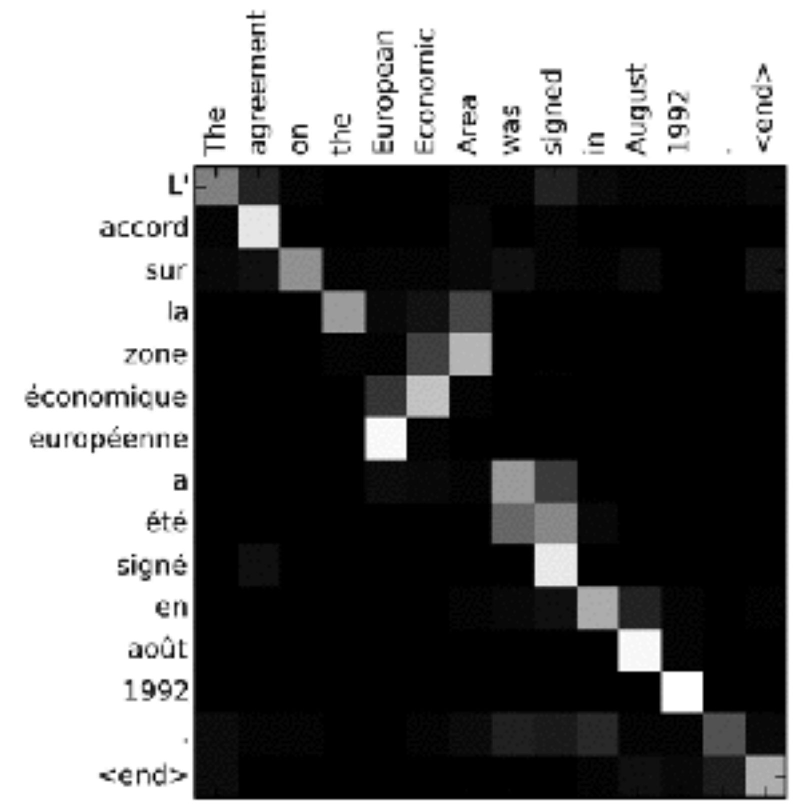


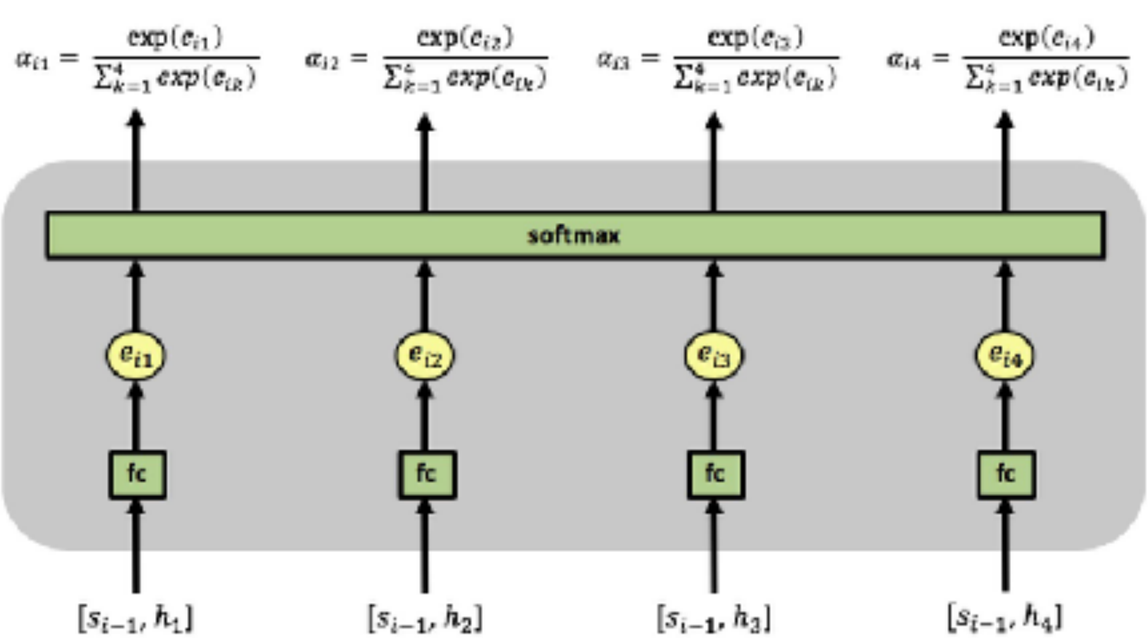
- **Machine translation**
- **Dialog Response generation**

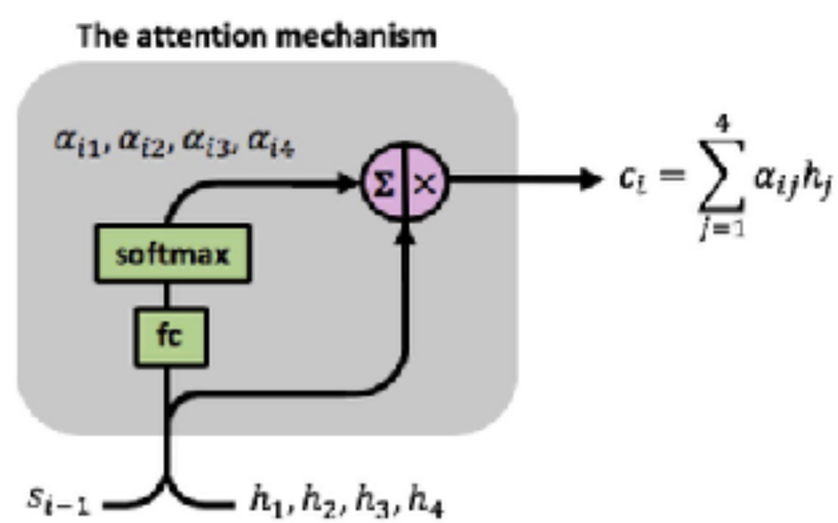
Attention Mechanism

Focus on certain parts **of the input sequence when predicting a certain part of the output sequence**

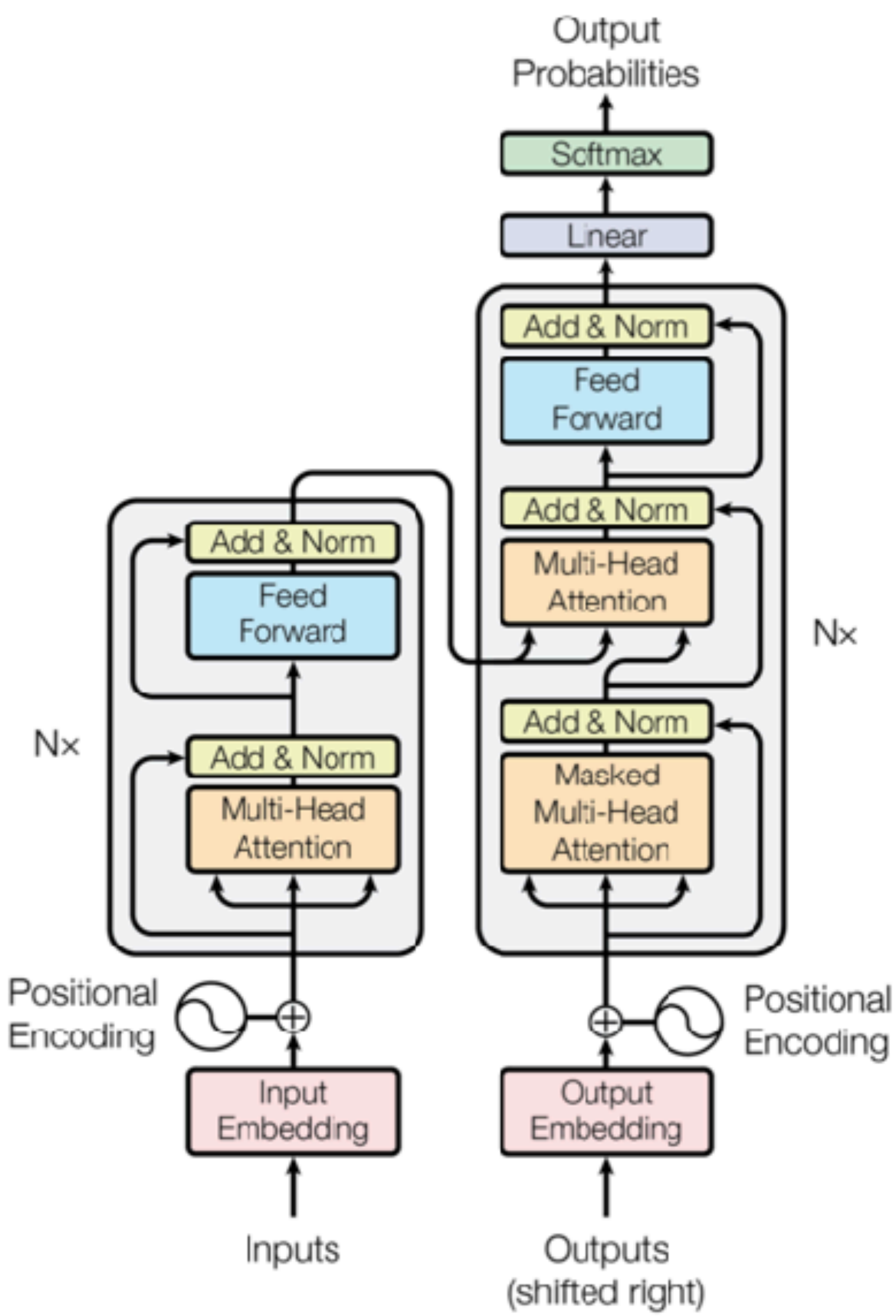








Transformer, Attention is all you need

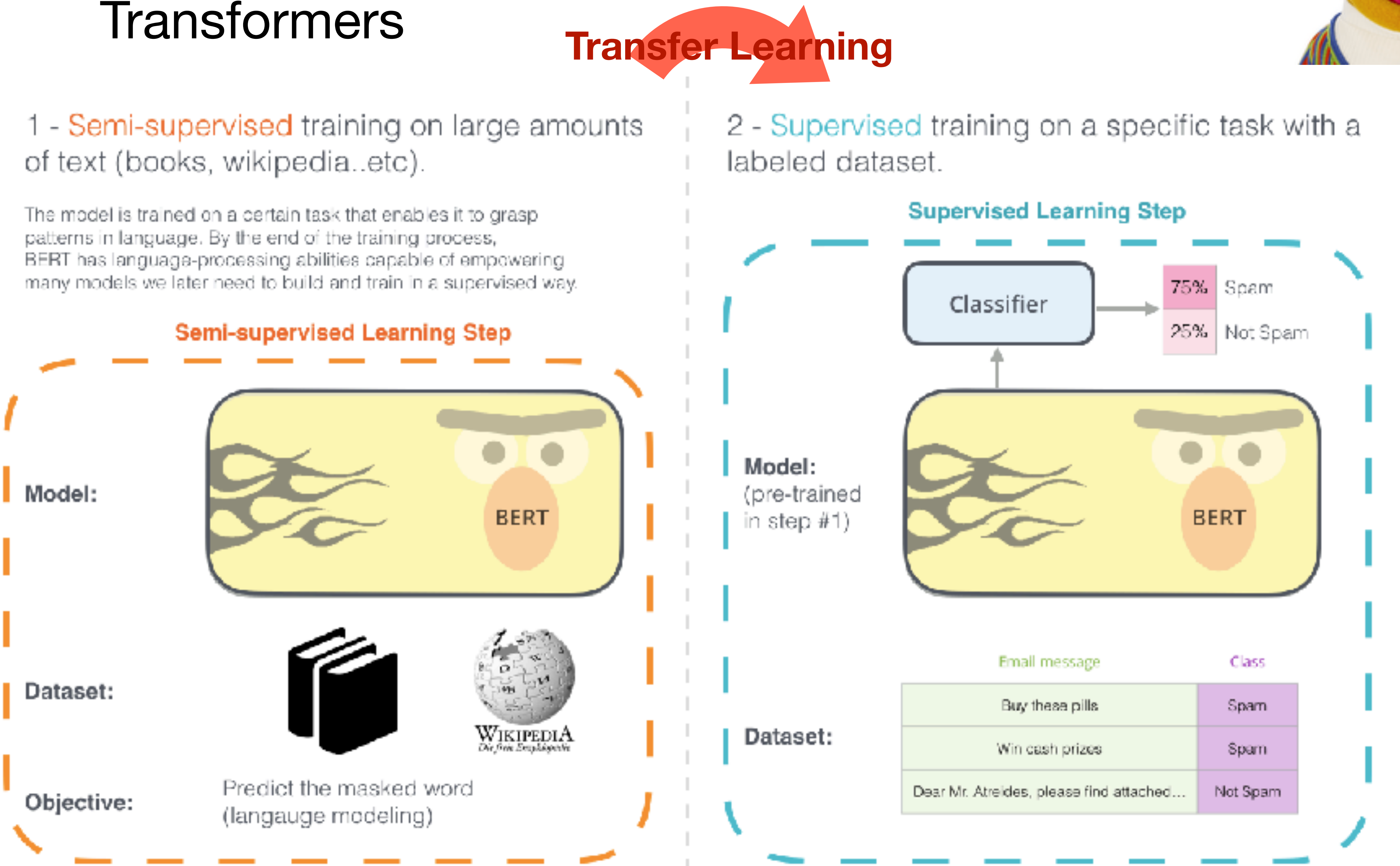


- **Without RNNs, only attention mechanism is used!**
  - **Self-attention**
  - **Multi-head attention**
  - **Positional encoding**

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Recent Word and Sentence Representation

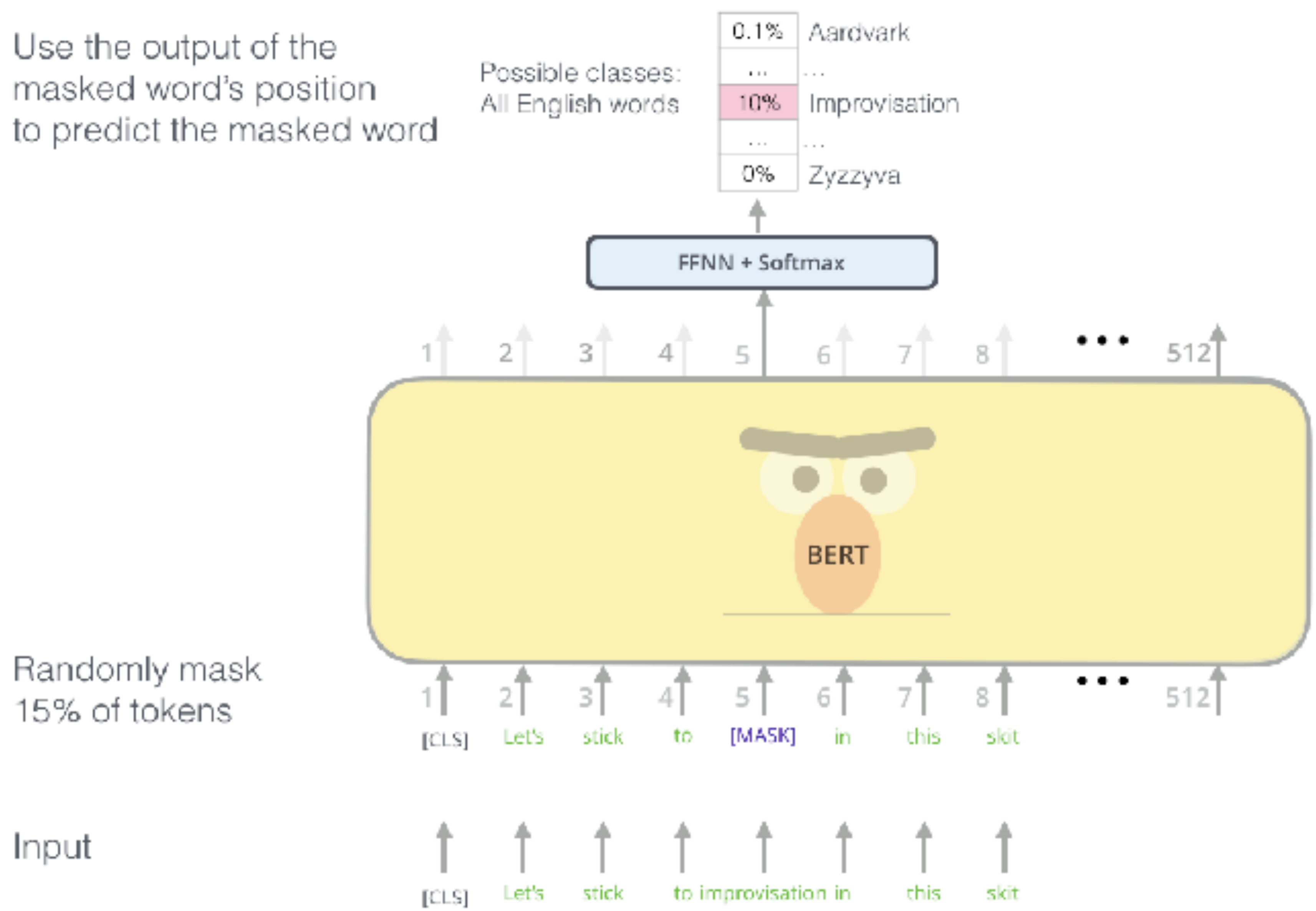
- BERT: Bi-directional Encoder Representations from Transformers



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

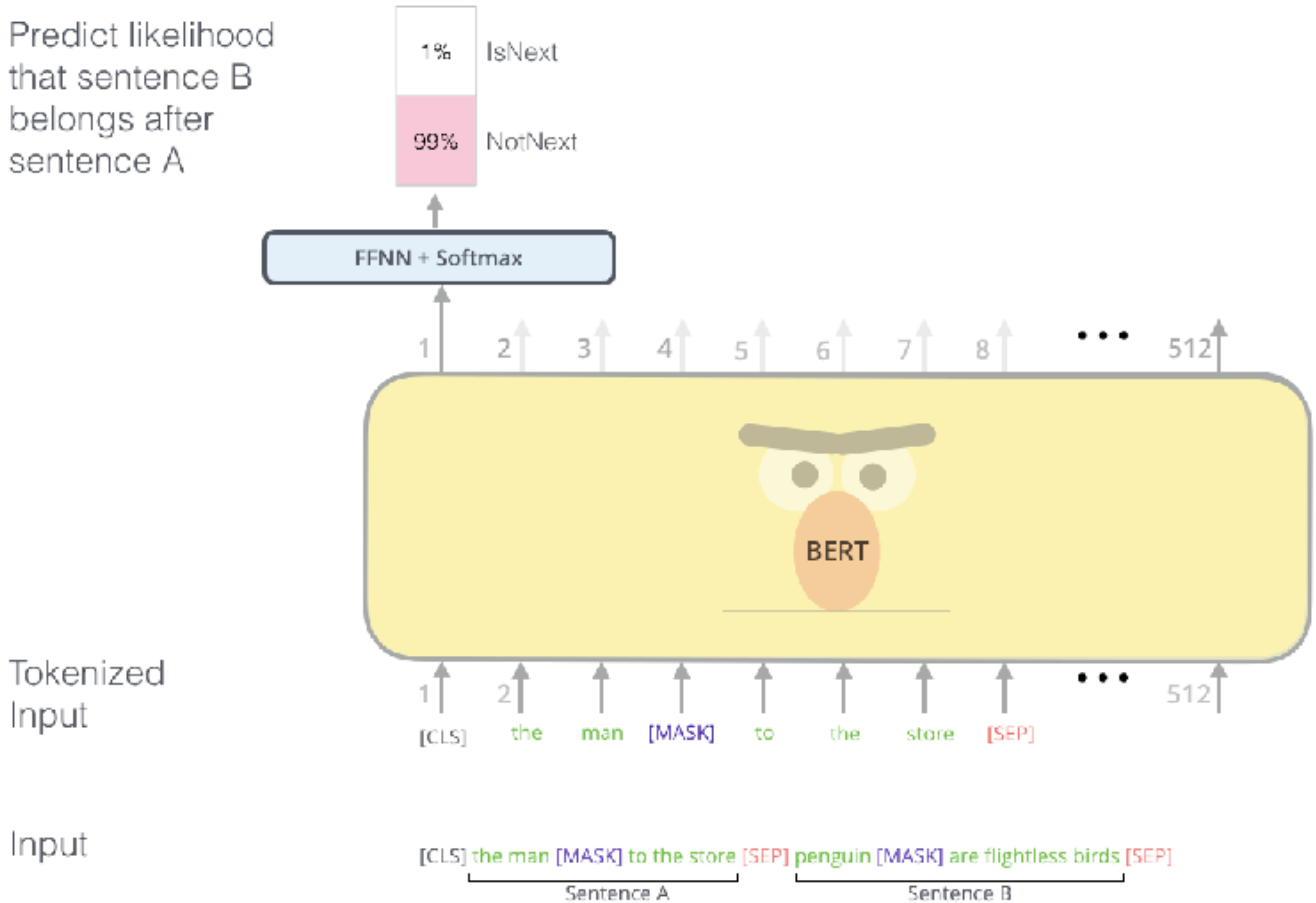
BERT

- Pretraining: Masked Language Model



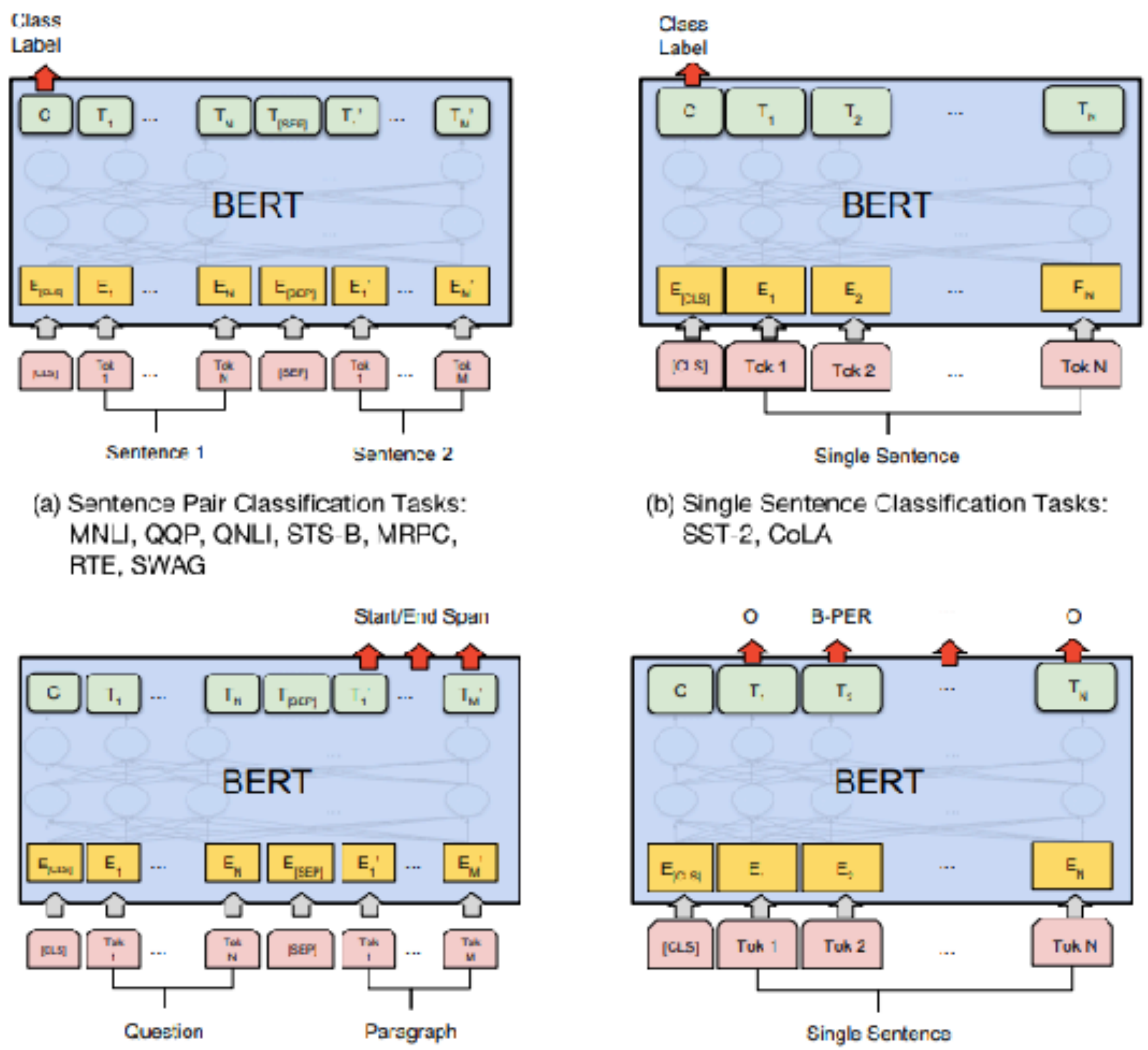
BERT

- Pretraining: Two-sentence Classification



BERT

- Fine-tuning for downstream tasks



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

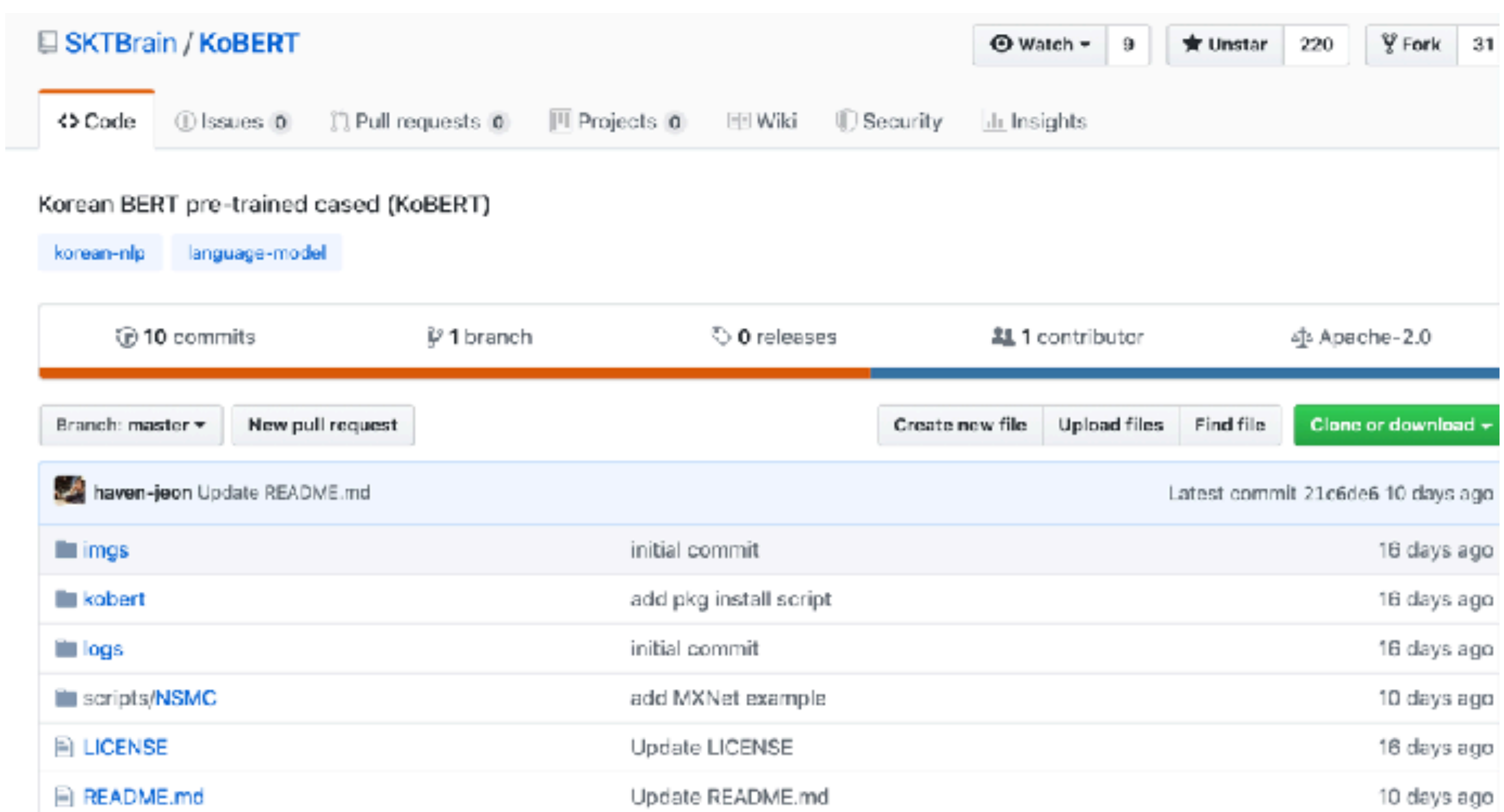
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Pre-trained BERT for Korean

- Google Bert (Multilingual) : https://github.com/google-research/bert/blob/master/multilingual.md
- ETRI, KorBert: http://aiopen.etri.re.kr/service_dataset.php
- SK T-Brain, **KoBERT**: https://github.com/SKTBrain/KoBERT



Outline

I. Introduction to dialog systems

II. Background
- Machine learning
- Deep learning and Neural networks

III. Deep learning for Natural Language
- Word embedding
- Language models

IV. Deep learning for Dialog systems
- SUMBT
- LaRL
- Challenges

Conversational Agents

Chit-Chat

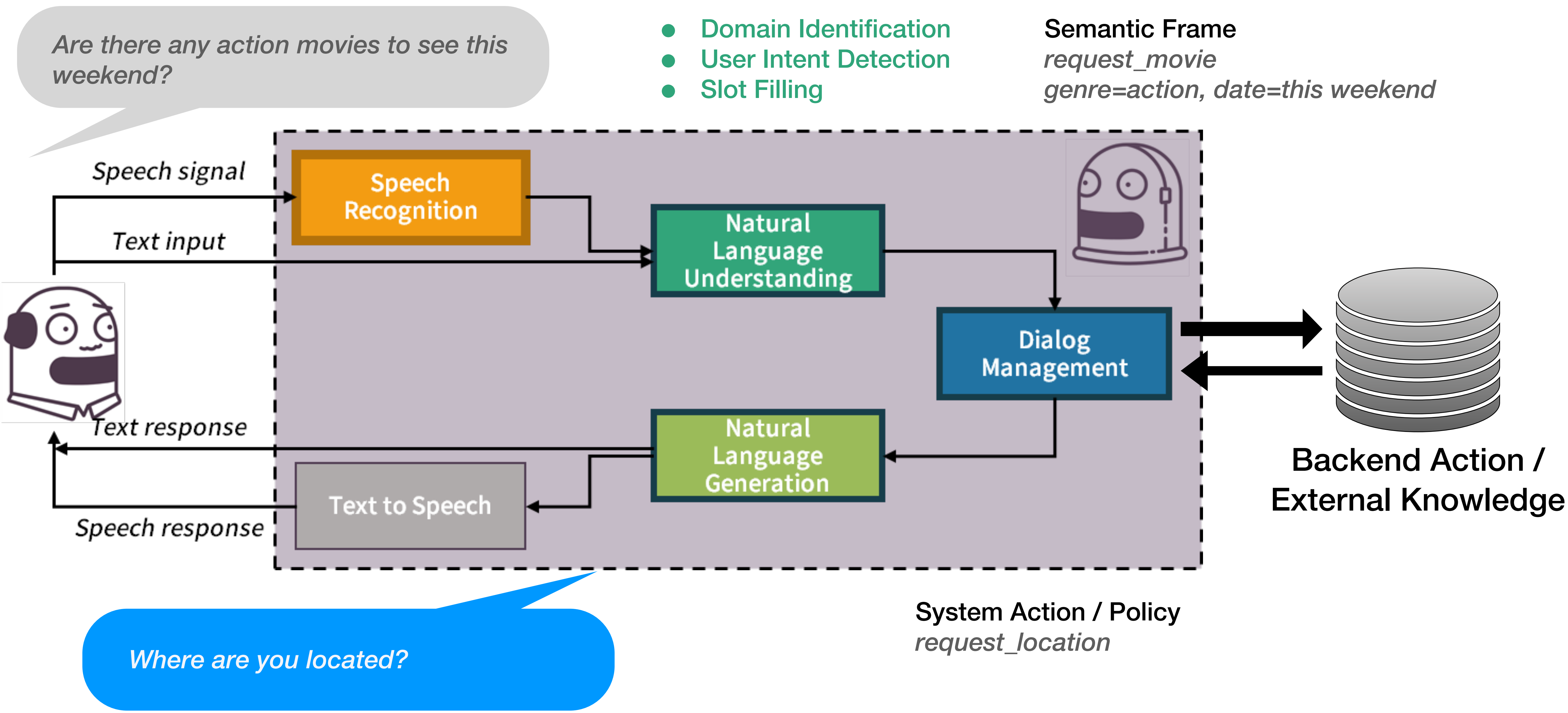
seq2seq models → Seq2seq with conversation contexts → Knowledge-grounded seq2seq models

Task-Oriented

Single-domain, system-initiative → Multi-domain, contextual, mixed-initiative → End-to-end learning, massively multi-domain

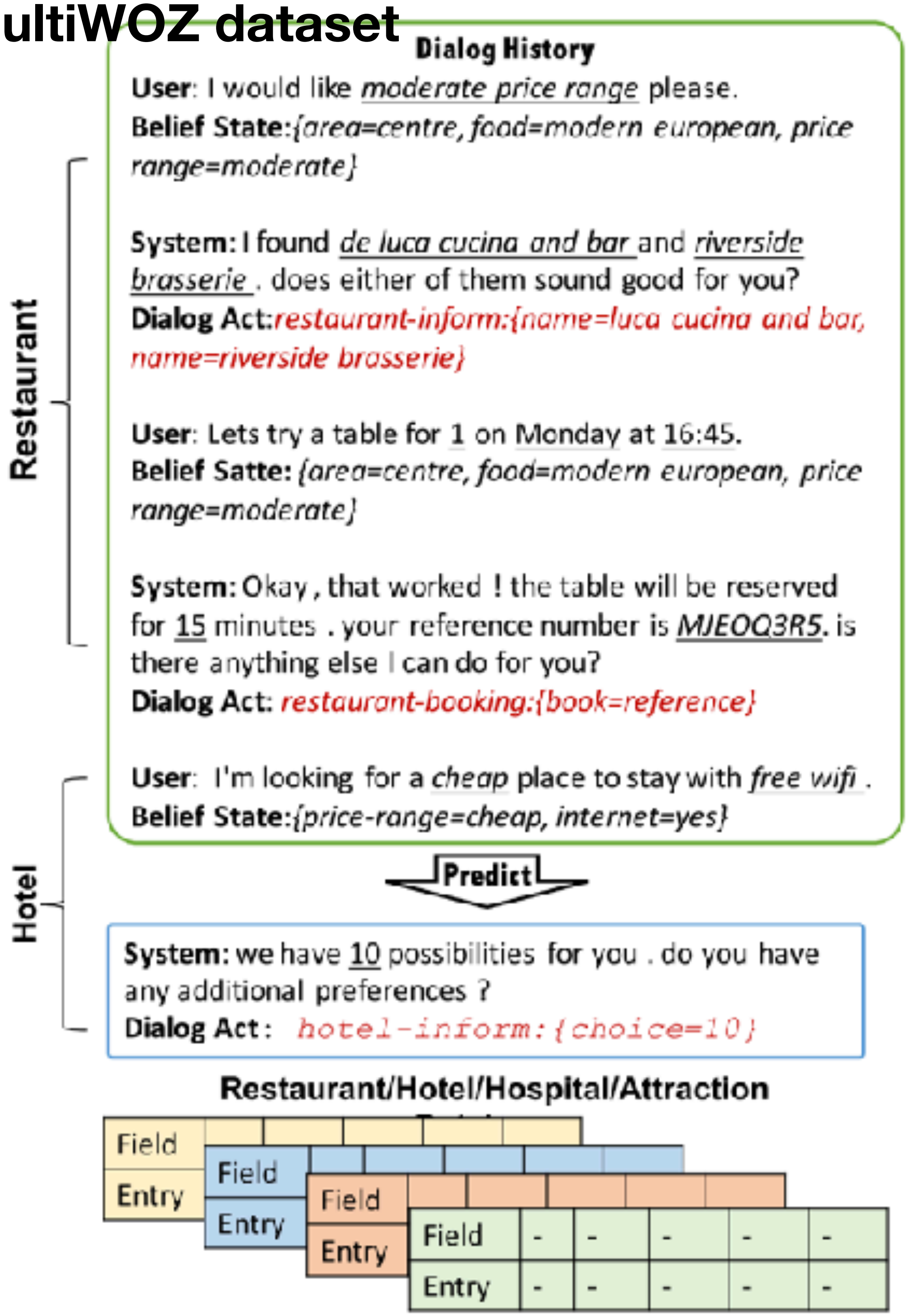
Material: https://deepdialogue.miulab.tw

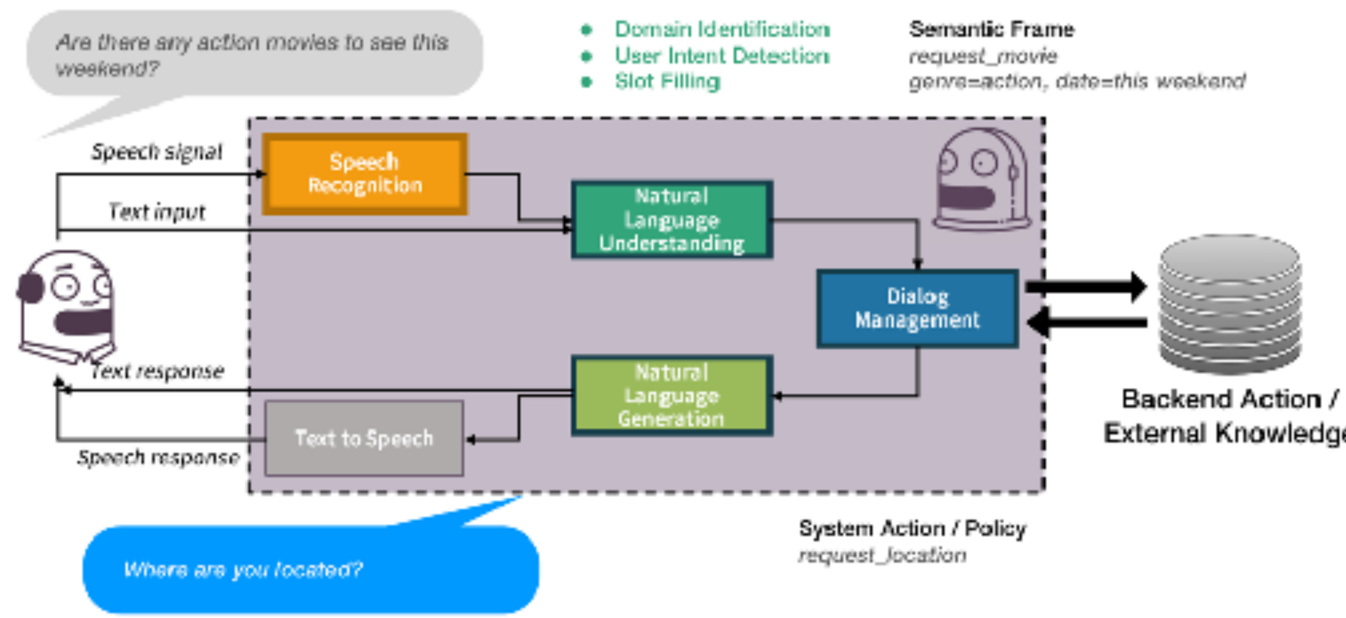
Spoken Dialog Systems



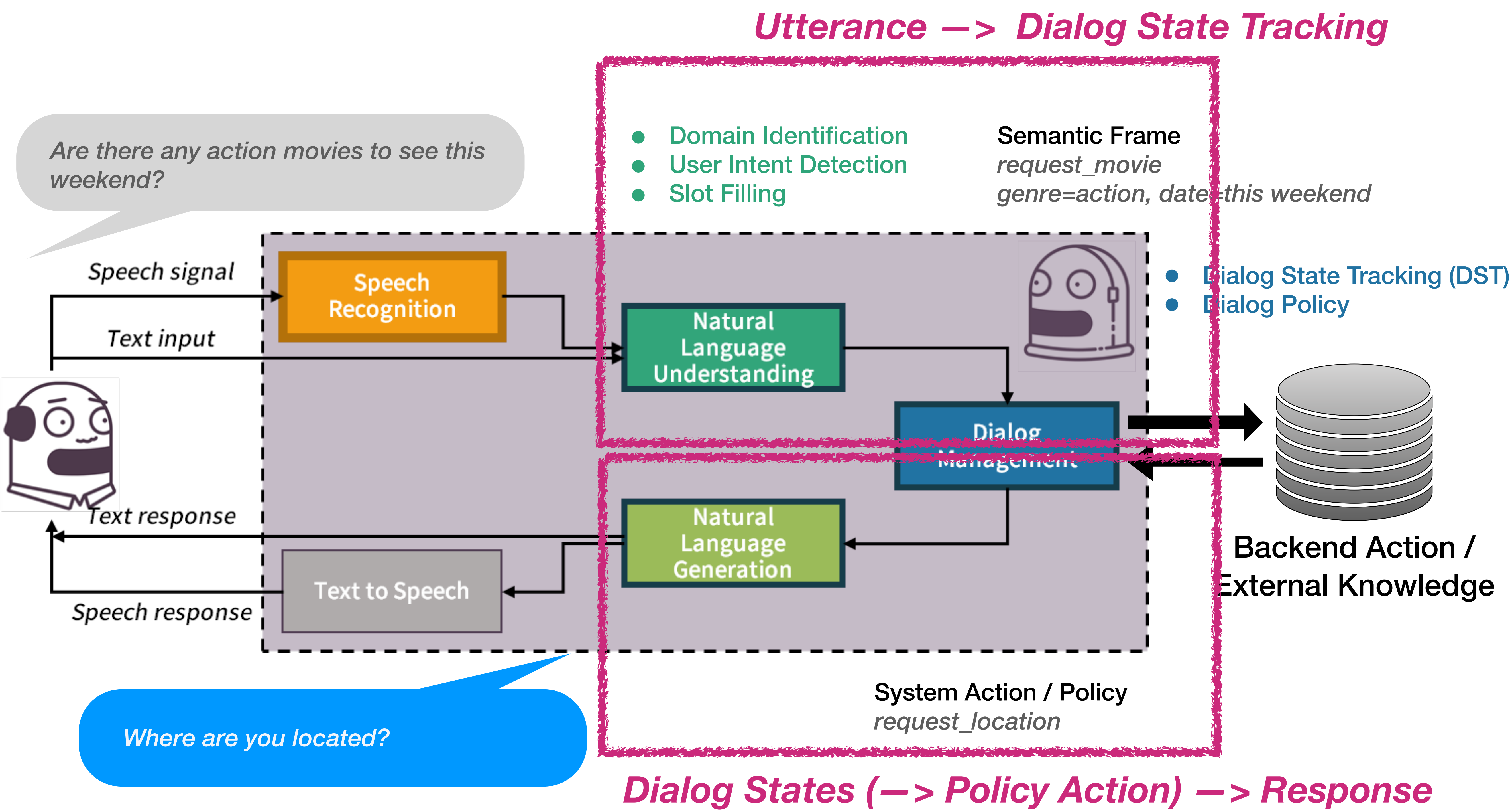
Multi-domain Goal-Oriented Dialogue System

MultiWOZ dataset



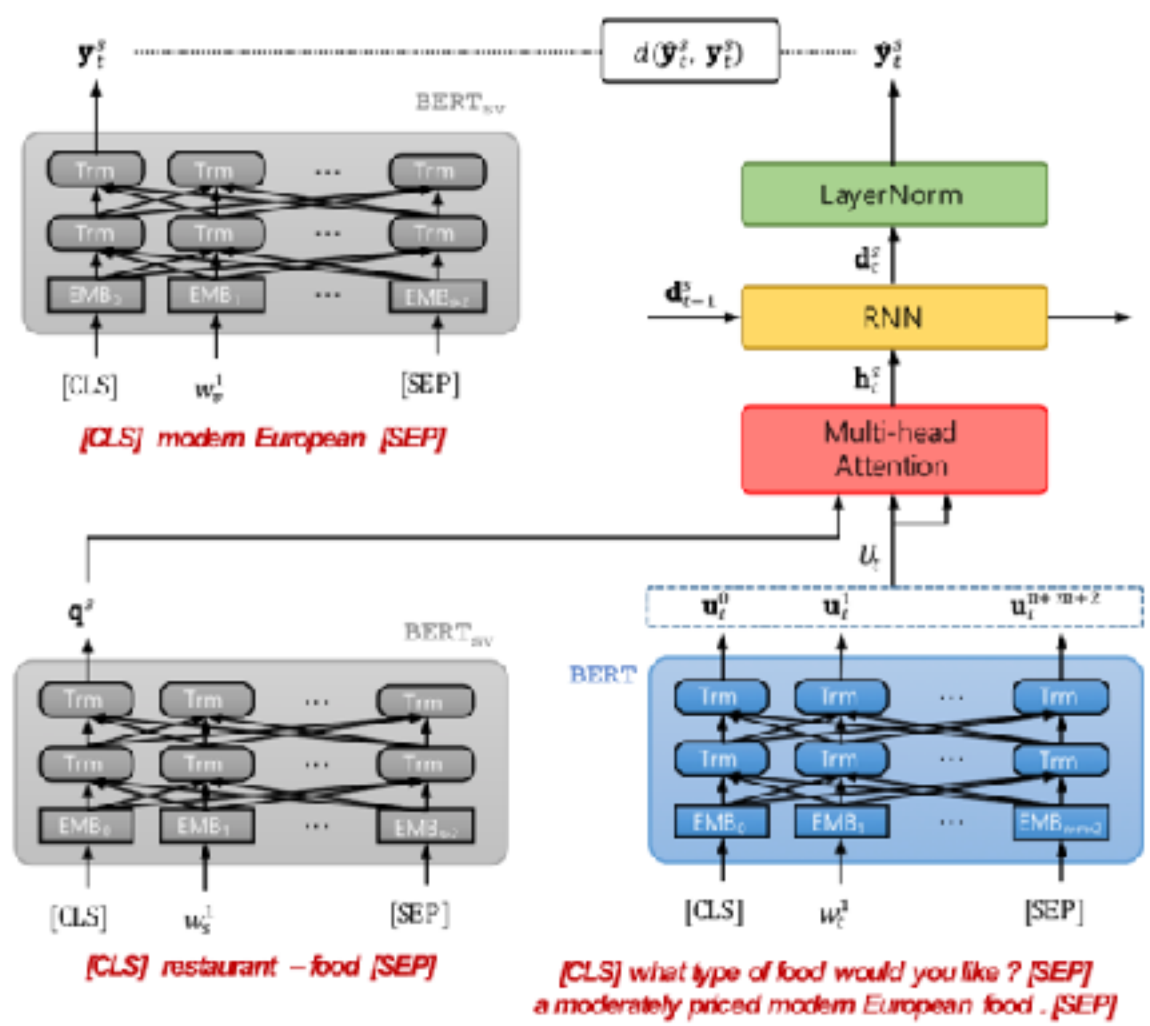


Toward End-to-End Multi-Domain Goal-oriented Dialogue systems



SUMBT: Slot-Utterance Matching Belief Tracker

- **Problem: Domain *independent* belief tracker**
- **Key Idea: Find the slot-value of a domain-slot type from user and system's utterances using attention mechanism like question-answering problems**

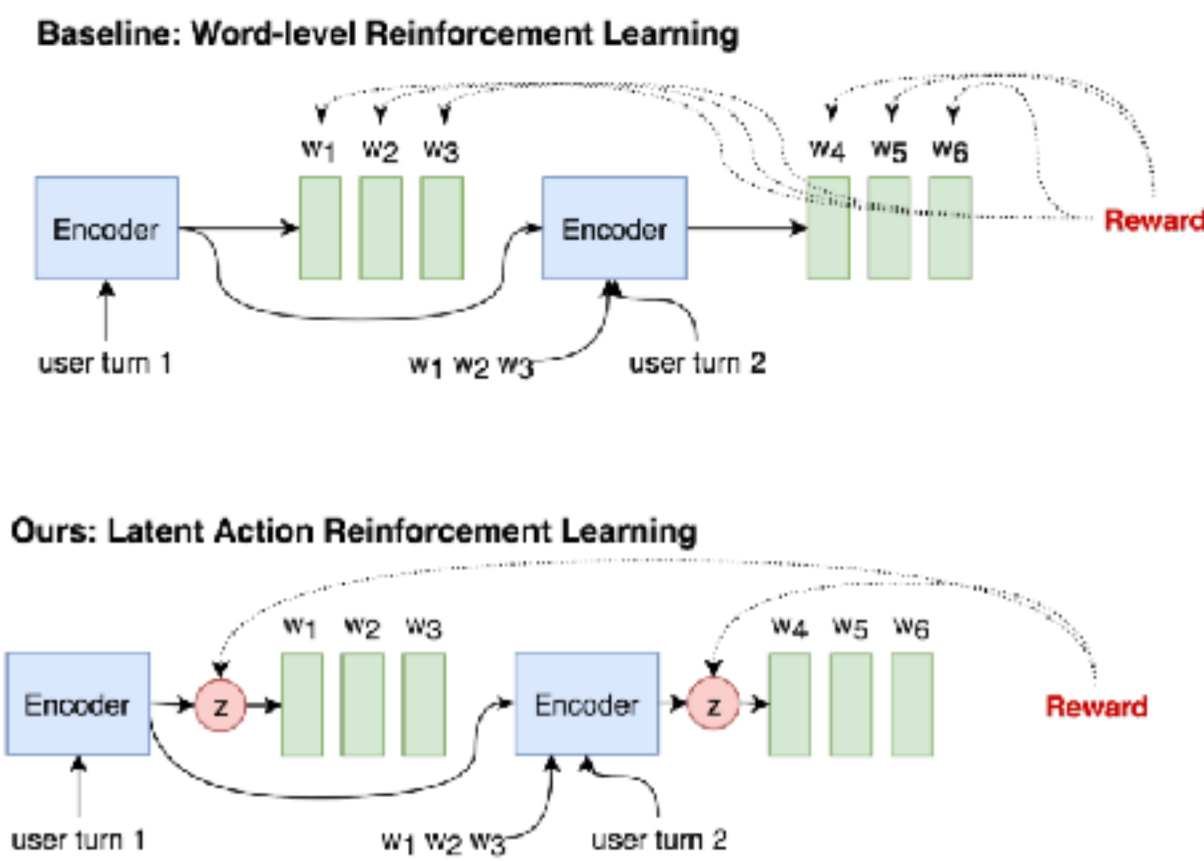


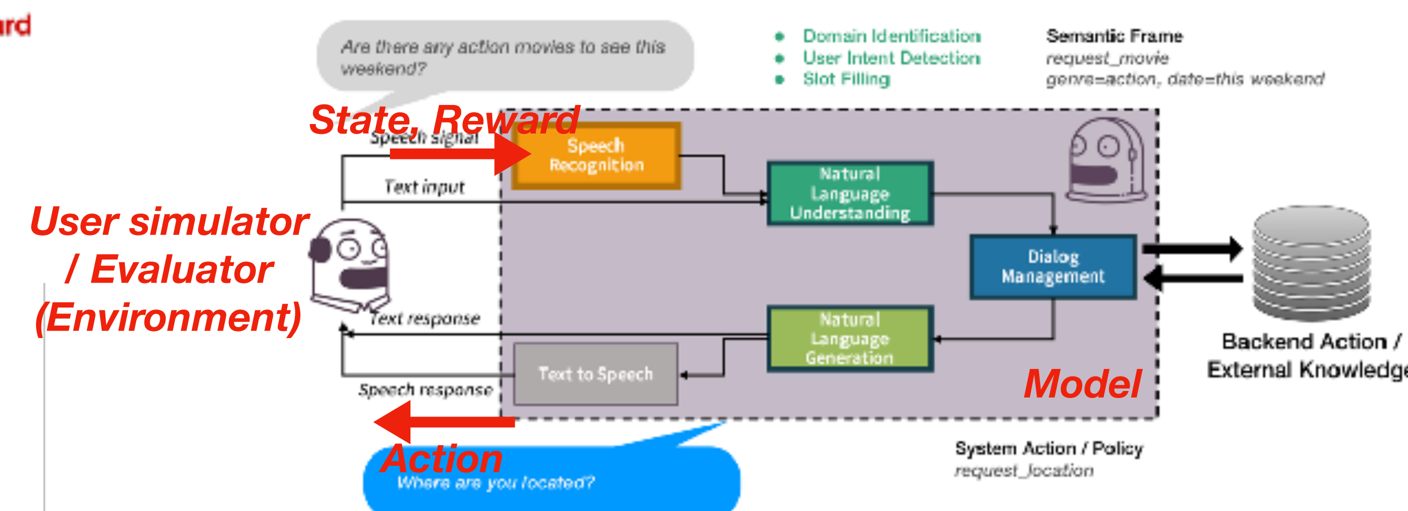
Model	MultiWOZ		MultiWOZ (Only Restaurant)	
	Joint	Slot	Joint	Slot
MDBT (Ramadan et al., 2018)*	0.1557	0.8953	0.1789	0.5499
GLAD (Zhong et al., 2018)*	0.3557	0.9544	0.5323	0.9654
GCE (Nouri et al., 2018)*	0.3627	**0.9842**	0.6093	0.9585
TRADE (Wu et al., 2019)	**0.4862**	0.9692	0.6535	0.9328
SUMBT	**0.49065**	0.97290	**0.82840**	**0.96475**

Lee H, Lee J, Kim TY. SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking. ACL, 2019.

LaRL: Latent Action Reinforcement Learning

- **Problems:**
  - **Simple hand-crafted system action space**
  - **Word-level RL suffers from credit assignment**
- **Key Idea: Latent action spaces, decoupling the discourse-level decision-making from natural language generation**

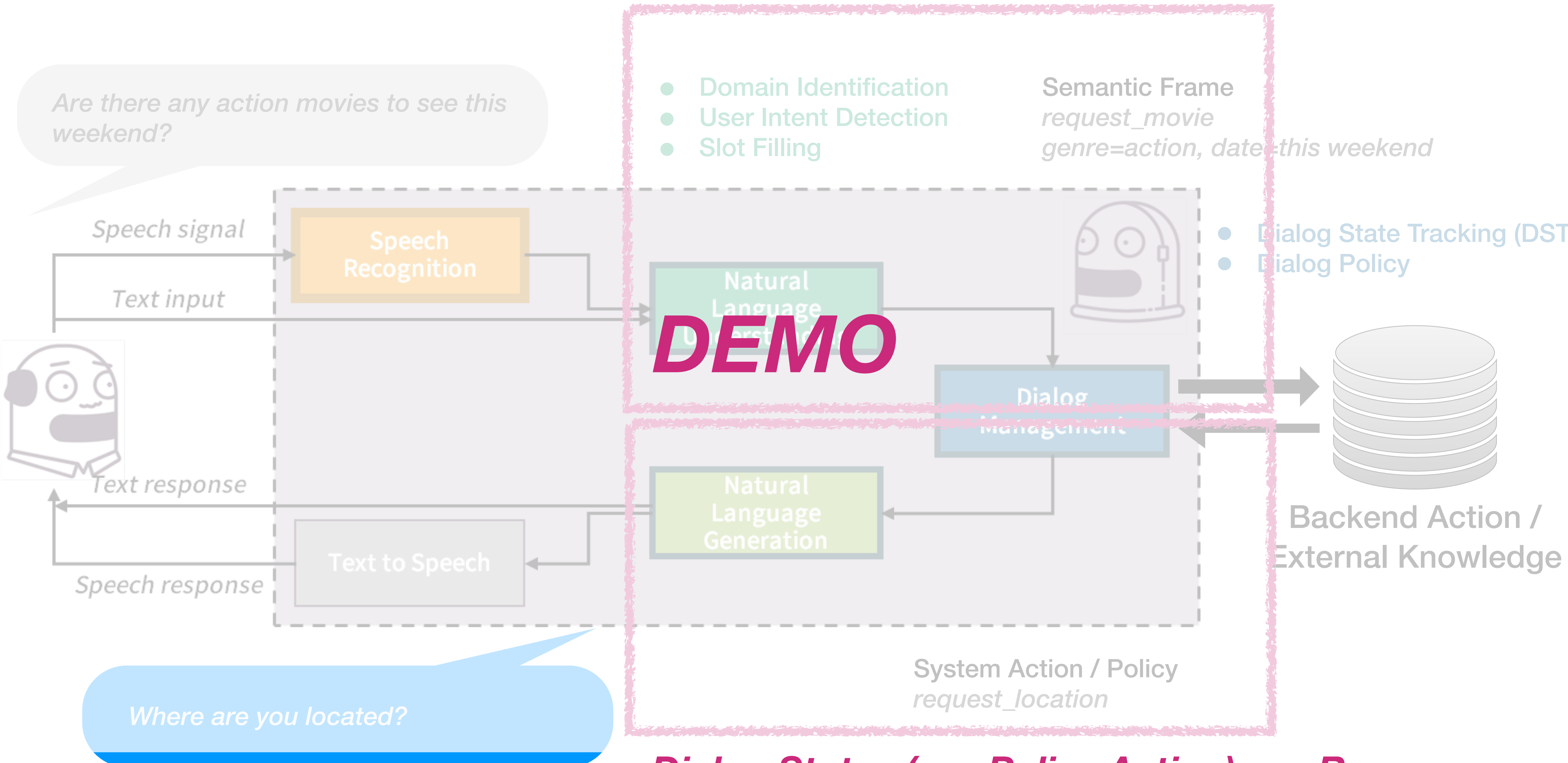




Zhao, T., Xie, K., & Eskenazi, M. Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models. NAACL-HLT 2019

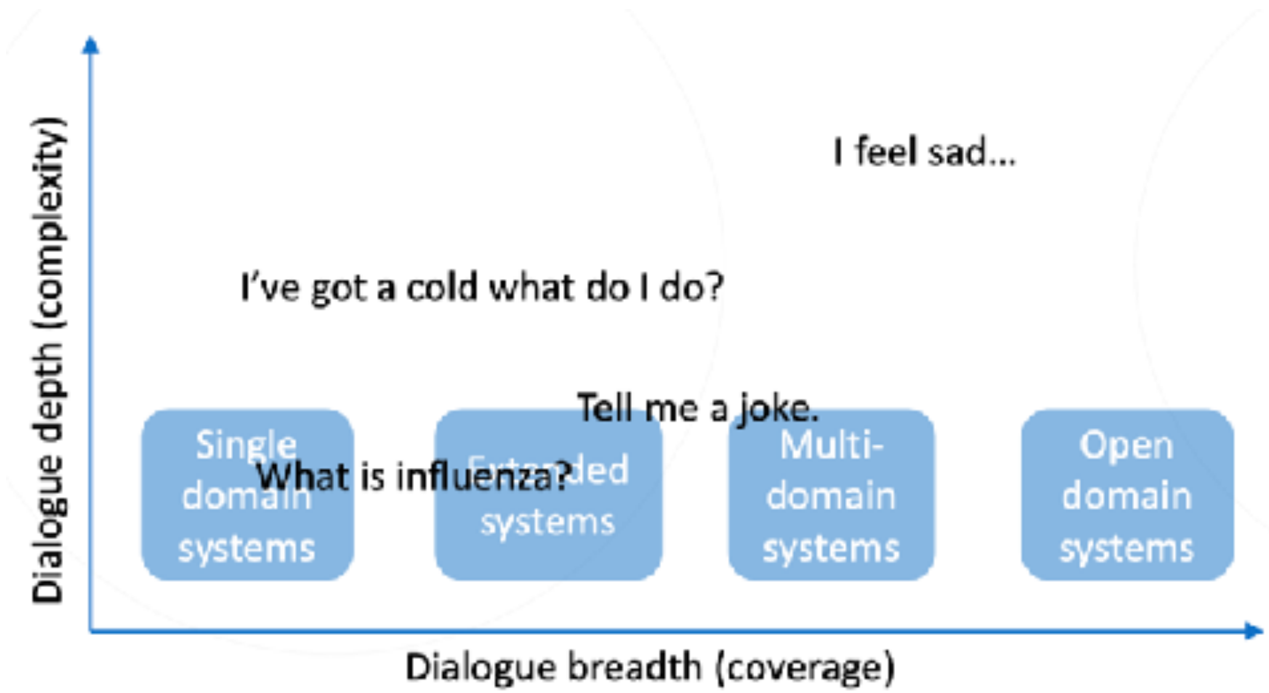
Toward End-to-End Multi-Domain Goal-oriented Dialogue systems

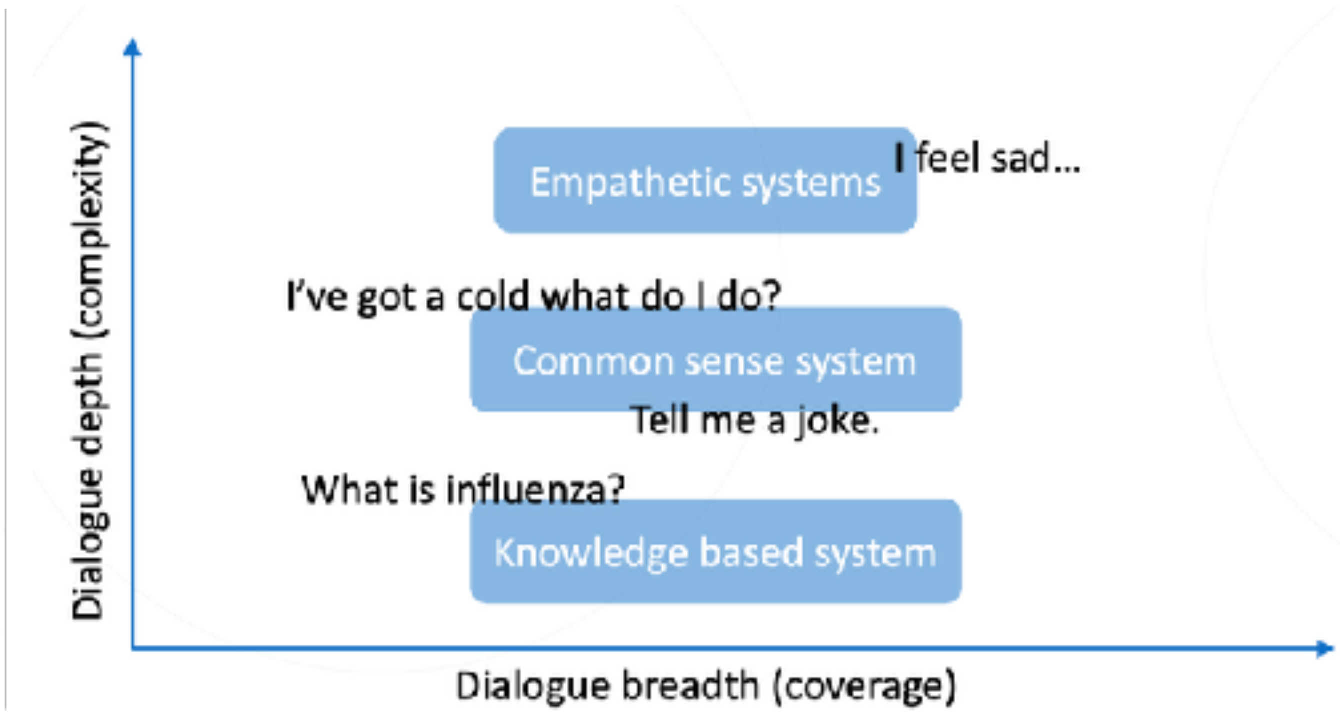
Utterance —> Dialog State Tracking



Dialog States (—> Policy Action) —> Response

Evolution Roadmap

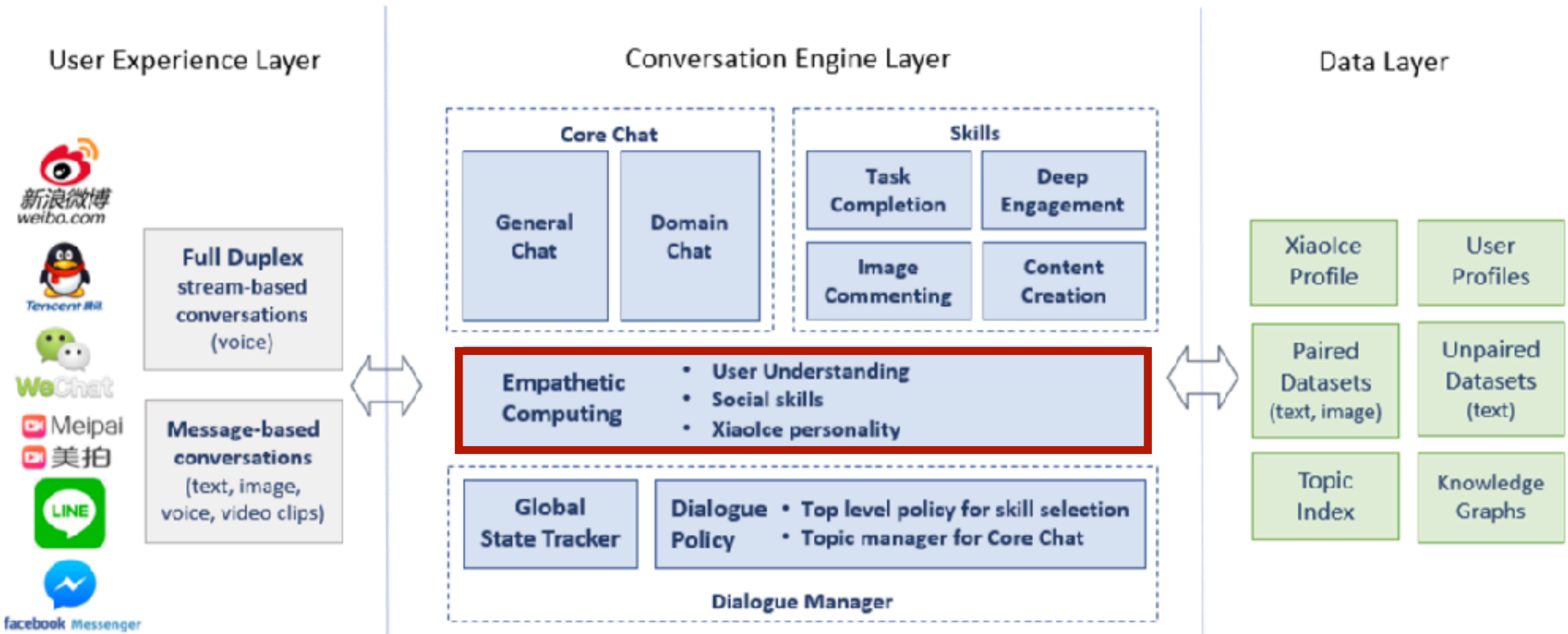




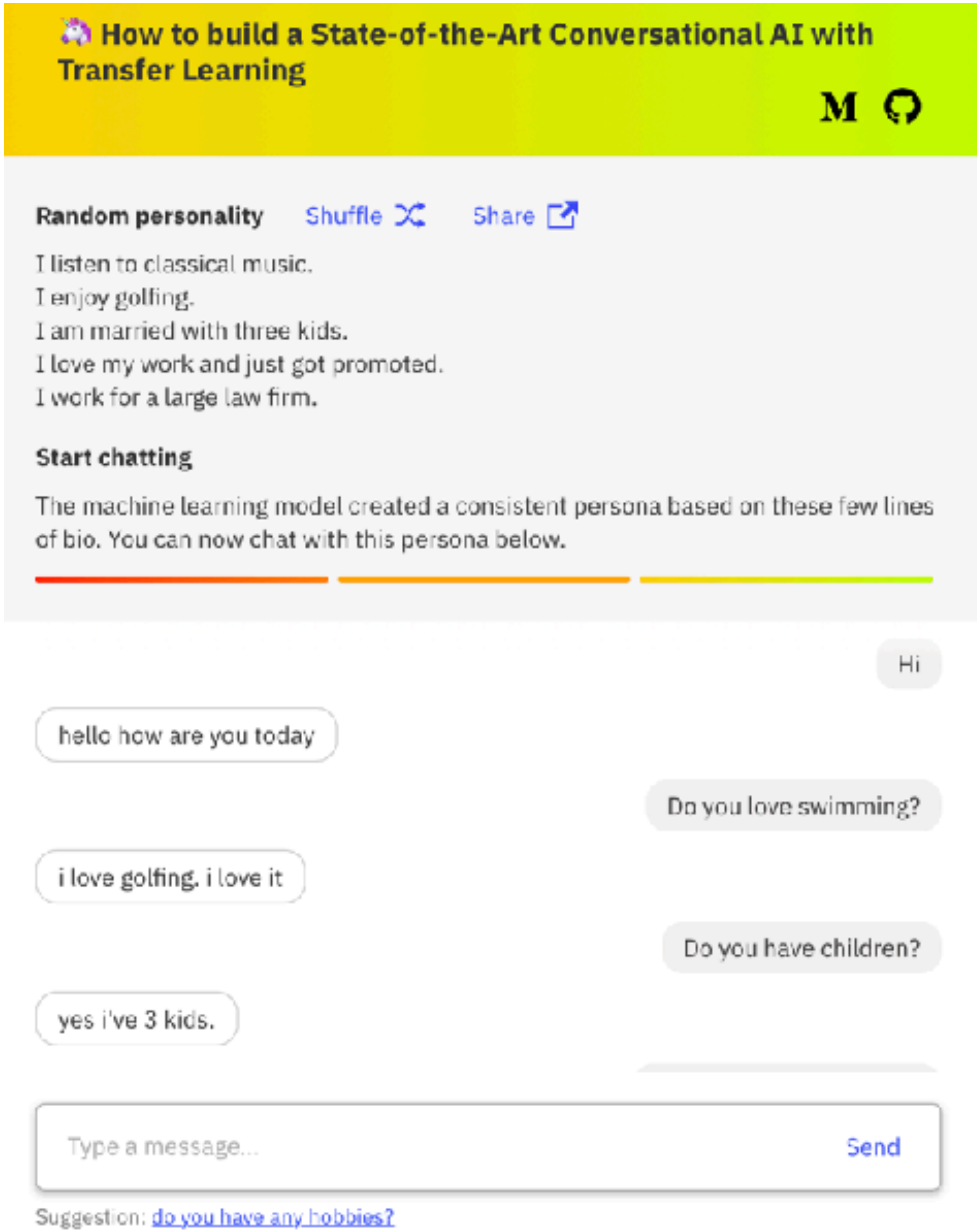
Material: https://deepdialogue.miulab.tw

XiaoIce
System Architecture

Microsoft, Xiaoice (2018)



Dialogue System with Personality



https://convai.huggingface.co

Summary

I. Introduction to dialog systems
  - Brief history, components and categories of dialogue systems

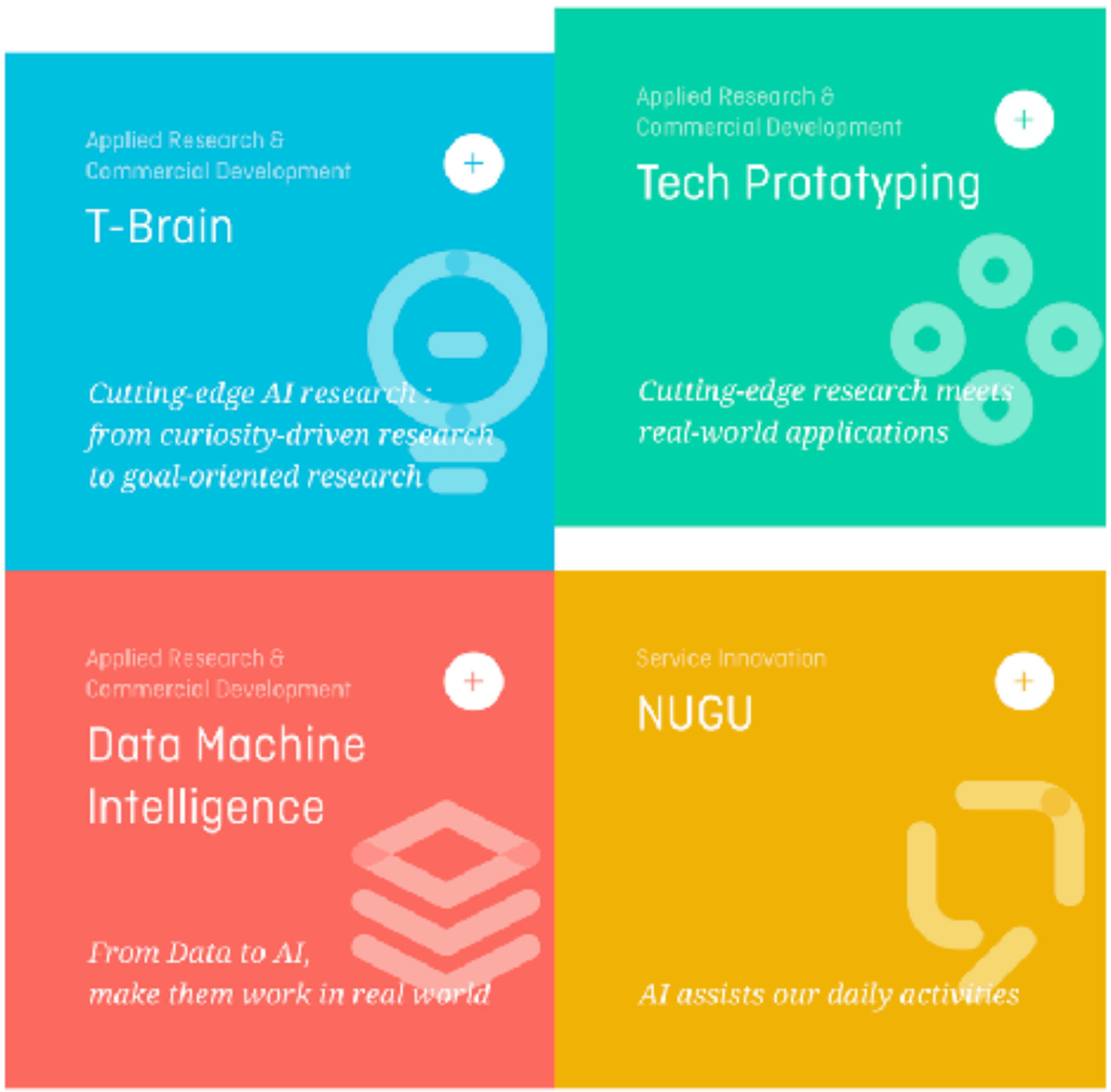
II. Background
  - Machine learning:
    Supervised, Unsupervised, Reinforcement Learning
  - Deep learning and Neural networks:
    Neuron, Architecture, Learning Algorithm

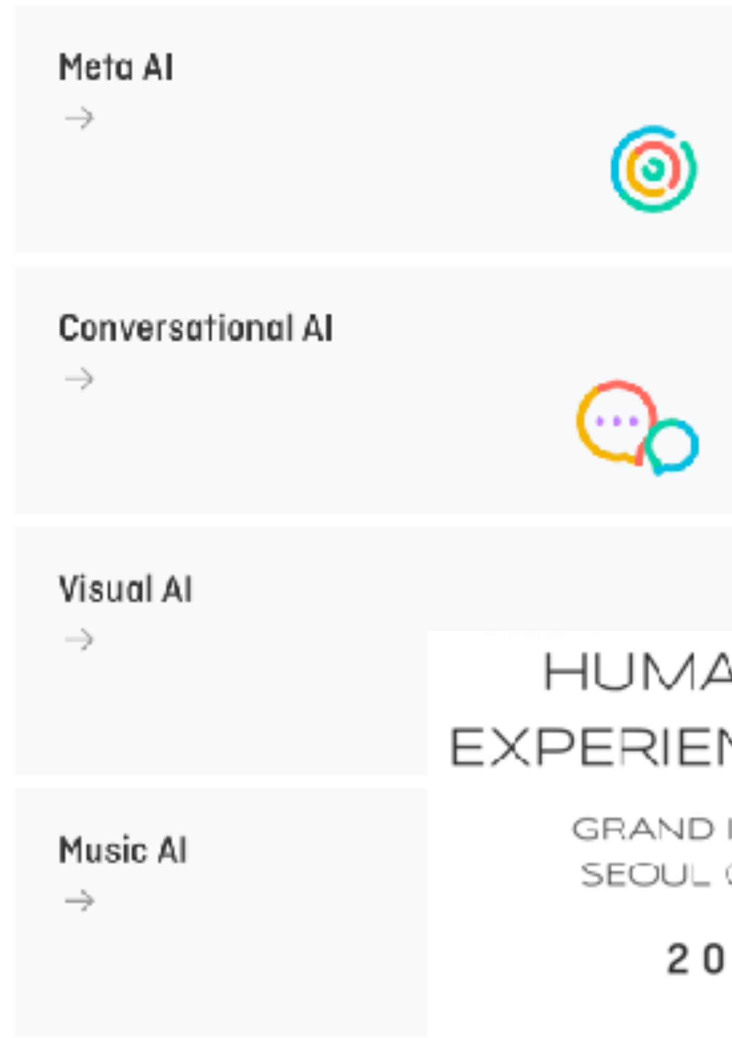
III. Deep learning for Natural Language
  - Word embedding: Skip-gram, CBOW
  - Language models: RNN, BERT (Attention, Transformer)

IV. Deep learning for Dialog systems
  - E2E Multi-domain Goal-oriented Dialog System
  - Future direction
    - Empathic, Personality, Open domain, Common sense …

SK T-Brain, AI Center

- https://skt.ai





Thank you

hwaran.lee@gmail.com