

# UVA CS 4774: Machine Learning

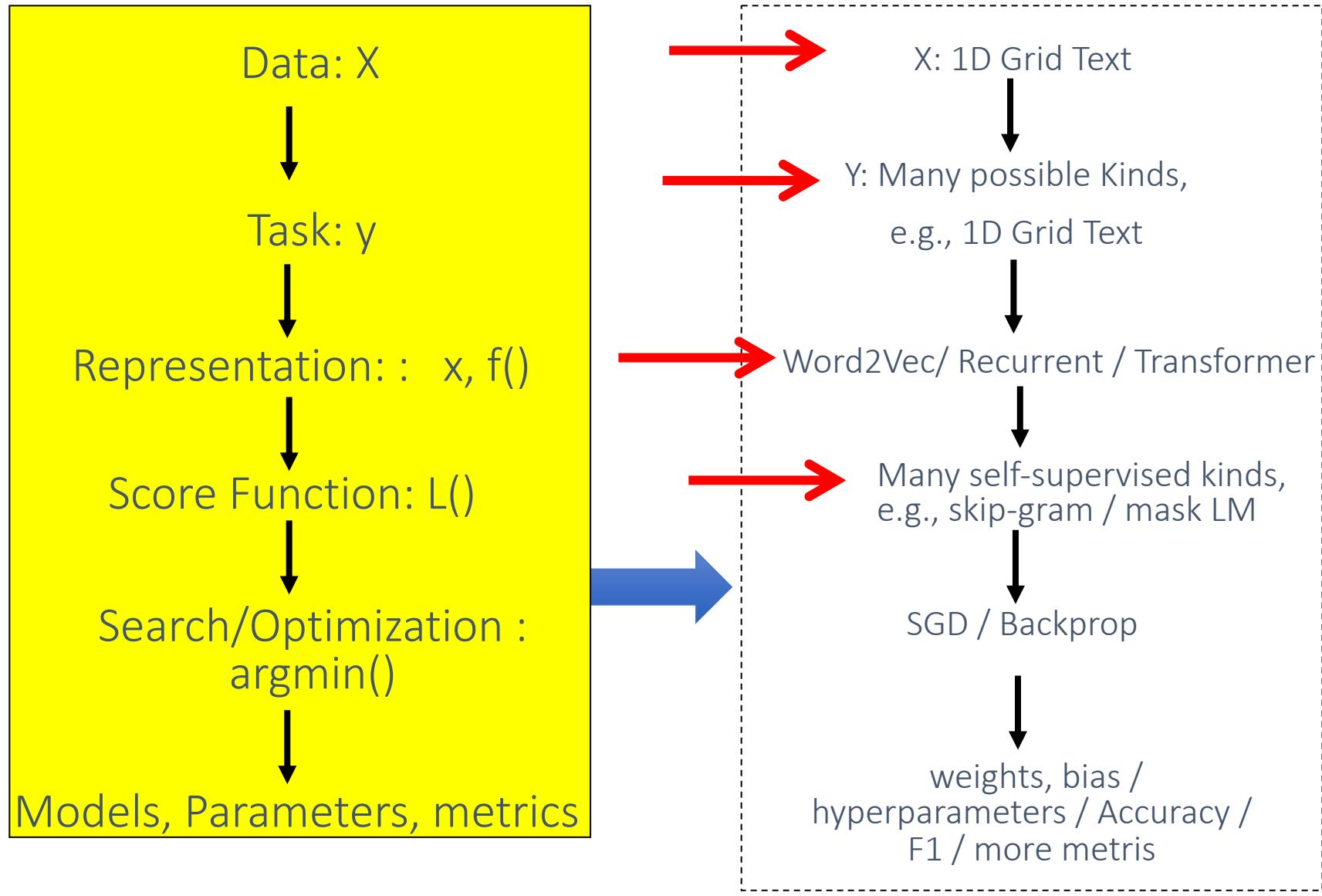
## S3: Lecture 18: Deep Neural Networks for Natural Language Processing

Dr. Yanjun Qi

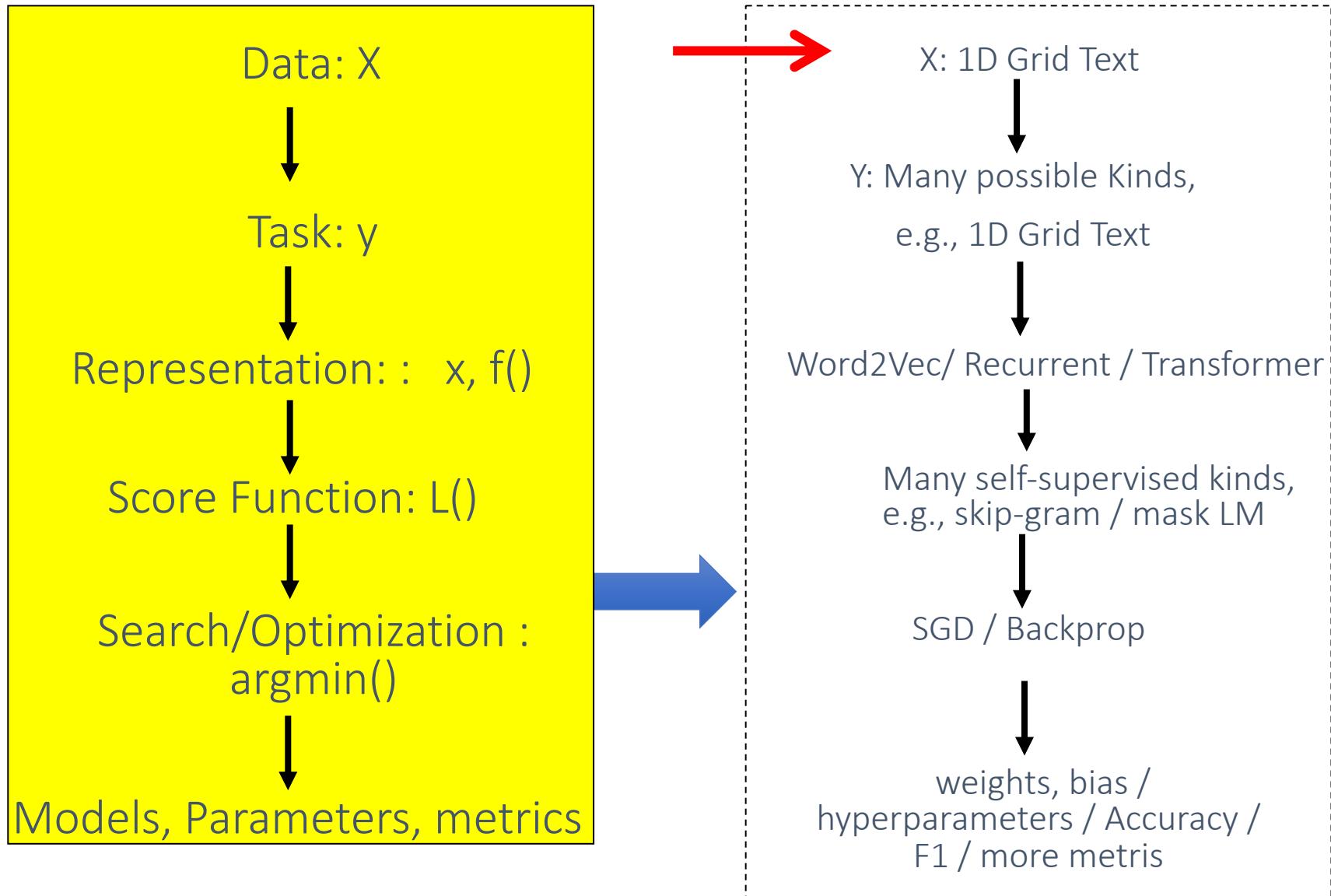
University of Virginia  
Department of Computer Science

Module I

# Today: Neural Network Models on 1D Grid / Language Data



# Today: Neural Network Models on 1D Grid / Language Data



# What is NLP

- **Wiki: Natural language processing (NLP)** is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages.



# Go beyond the keyword matching



- Identify the **structure** and **meaning** of words, sentences, texts and conversations
- Deep understanding of broad language
- NLP is all around us

# Machine translation

A screenshot of a Google search results page. The search query "buenas noches" is entered in the search bar. The results are filtered by "All". It shows approximately 20,800,000 results found in 0.54 seconds. A prominent feature is a machine translation box comparing "buenas noches" (Spanish) on the left and "Goodnight" (English) on the right. Below this, there is a link to "3 more translations" and an option to "Open in Google Translate".

Google

buenas noches

All Images Shopping Apps Videos More Search tools

About 20,800,000 results (0.54 seconds)

Spanish▼  English▼

buenas noches Edit Goodnight

3 more translations

Open in Google Translate

# Dialog Systems

**Gift shop**

Items such as caps, t-shirts, sweatshirts and other miscellanea such as buttons and mouse pads have been designed. In addition, merchandise for almost all of the projects is available.

CD or DVD

There is a series of CDs/DVDs with selected Wikipedia content being produced by Wikipedians and SOS Children.

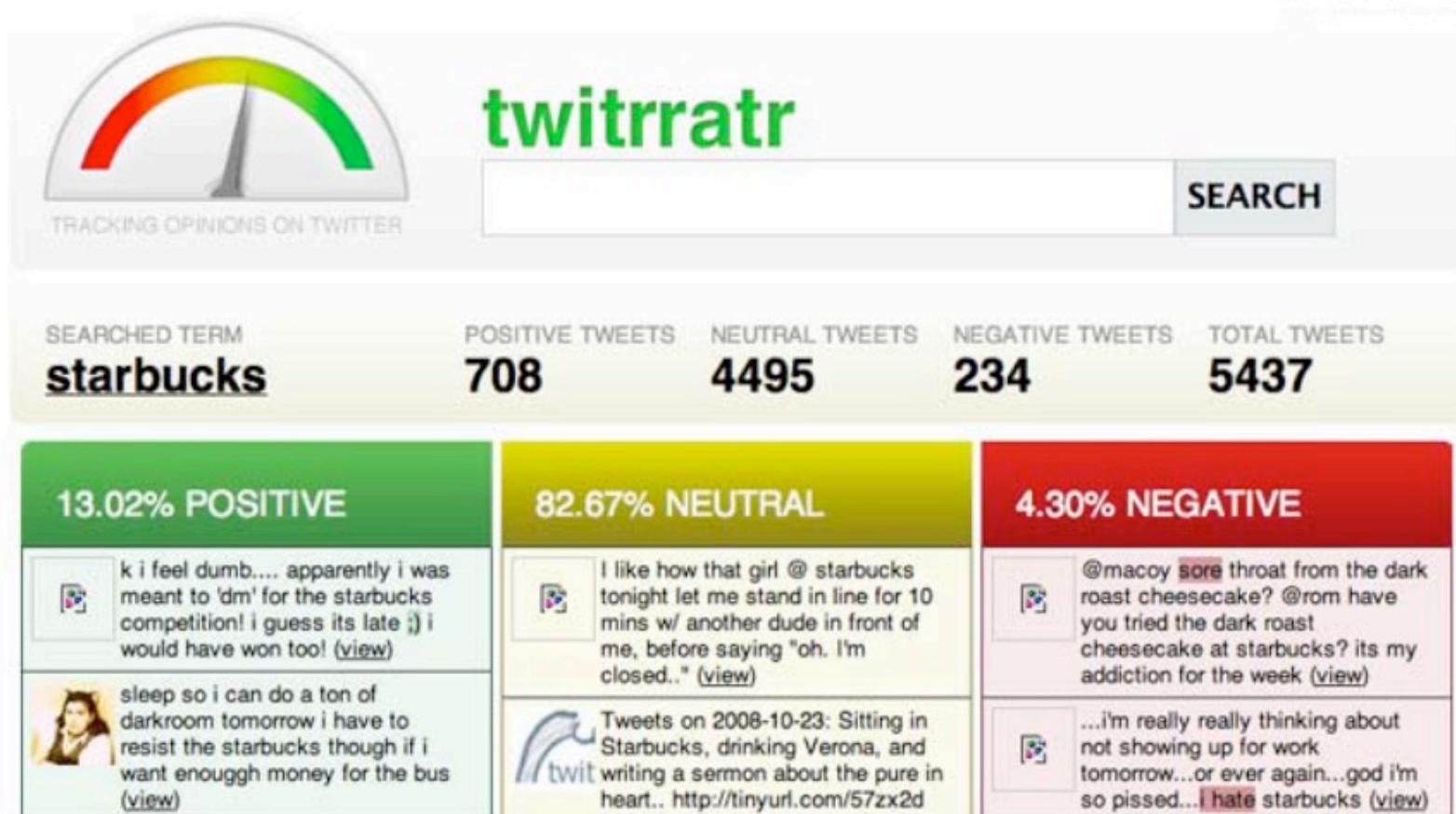
Downloading

Downloading content from Wikipedia is free of charge.

All text content is licensed under the GNU Free Documentation License

(GFDL). Images and other files are available under different terms, as detailed on

# Sentiment/Opinion Analysis



# Text Classification



BIDNESS ETC

The page at <https://mail.google.com/> says:

Did you mean to attach files?  
You wrote "is attached" in your message, but there are no files attached. Send anyway?

OK Cancel

1–21 of 21 < > ⚙️

Primary Social 1 new Promotions 2 new Updates 1 new

James, me (2) Hiking trip on Saturday - Yay - so glad you can join. We should leave from I 3:14 pm

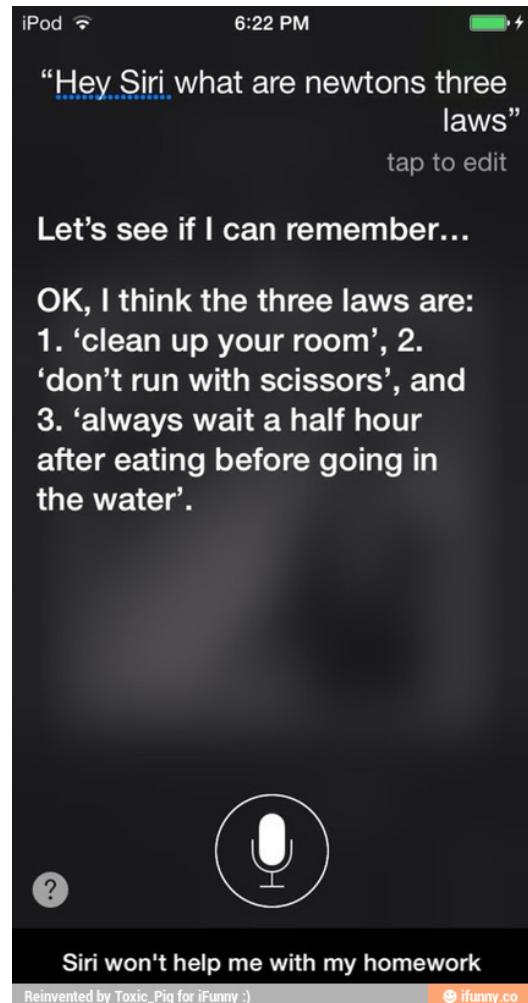
Hannah Cho Thank you - Keri - so good that you and Steve were able to come over. Thank you : 3:05 pm

Jay Birdsong School Upcoming school conference dates Hello everyone. A few people have [www.wired.com](http://www.wired.com)

# Question answering



['Watson' computer wins at 'Jeopardy'](#)



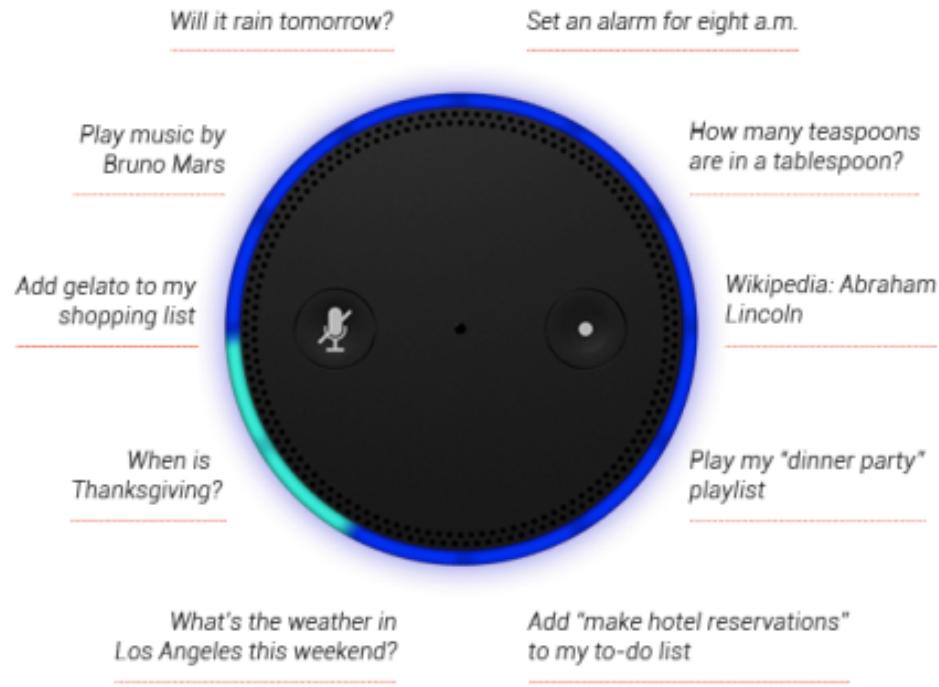
credit: ifunny.com

# Language Comprehension

**Christopher Robin** is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

- Q: who wrote Winnie the Pooh?
- Q: where is Chris lived?

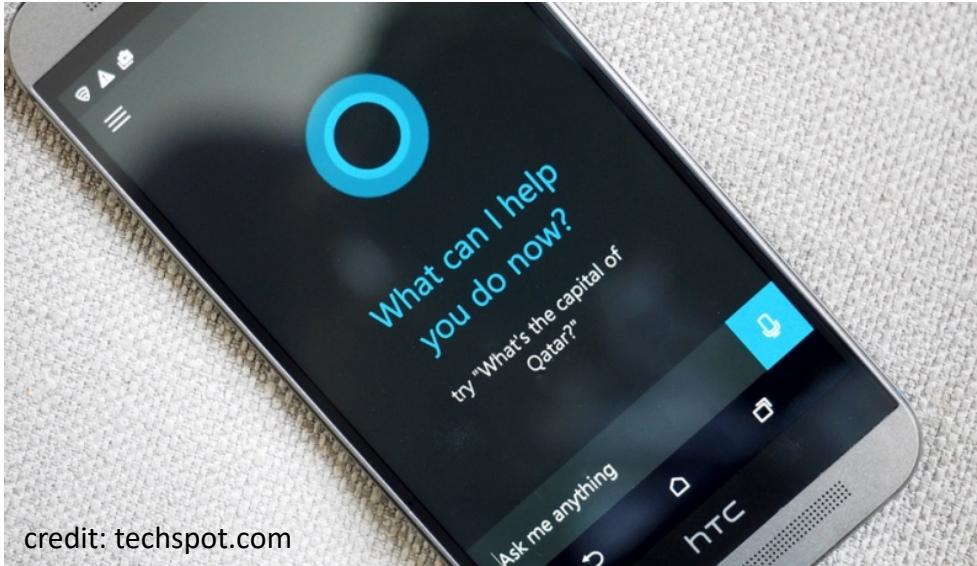
# Natural language instruction



<https://youtu.be/KkOCeAtKHlc?t=1m28s>

# More on natural language instruction

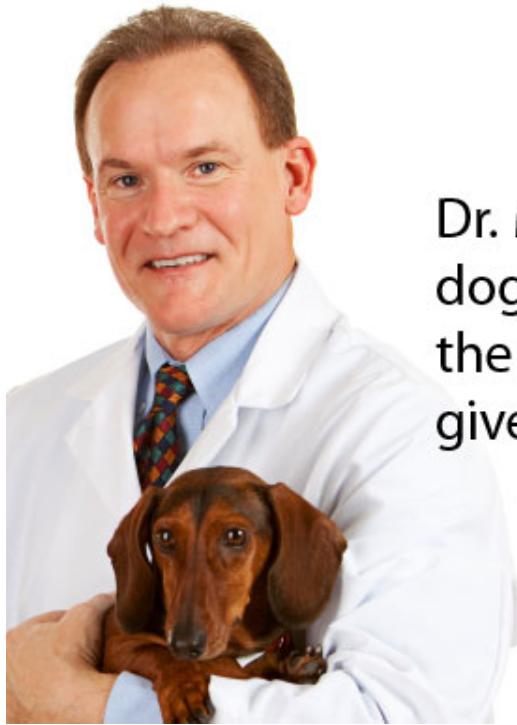
## Digital personal assistant



- Semantic parsing – understand tasks
- Entity linking – “my wife” = “Kellie” in the phone book

# Challenges – ambiguity

- Pronoun reference ambiguity

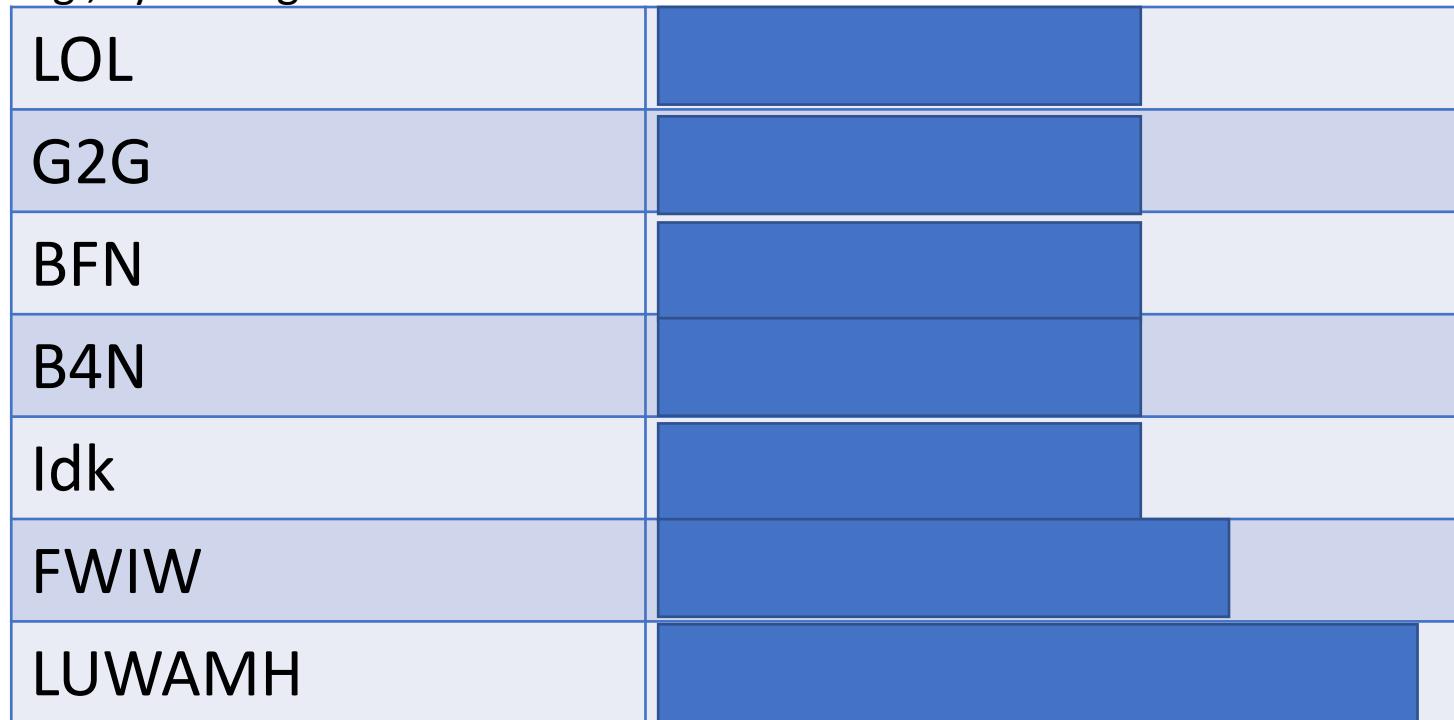


Dr. Macklin often brings his dog Champion to visit with the patients. **He** just loves to give big, wet, sloppy kisses!

Credit: <http://www.printwand.com/blog/8-catastrophic-examples-of-word-choice-mistakes>

# Challenges – language is not static

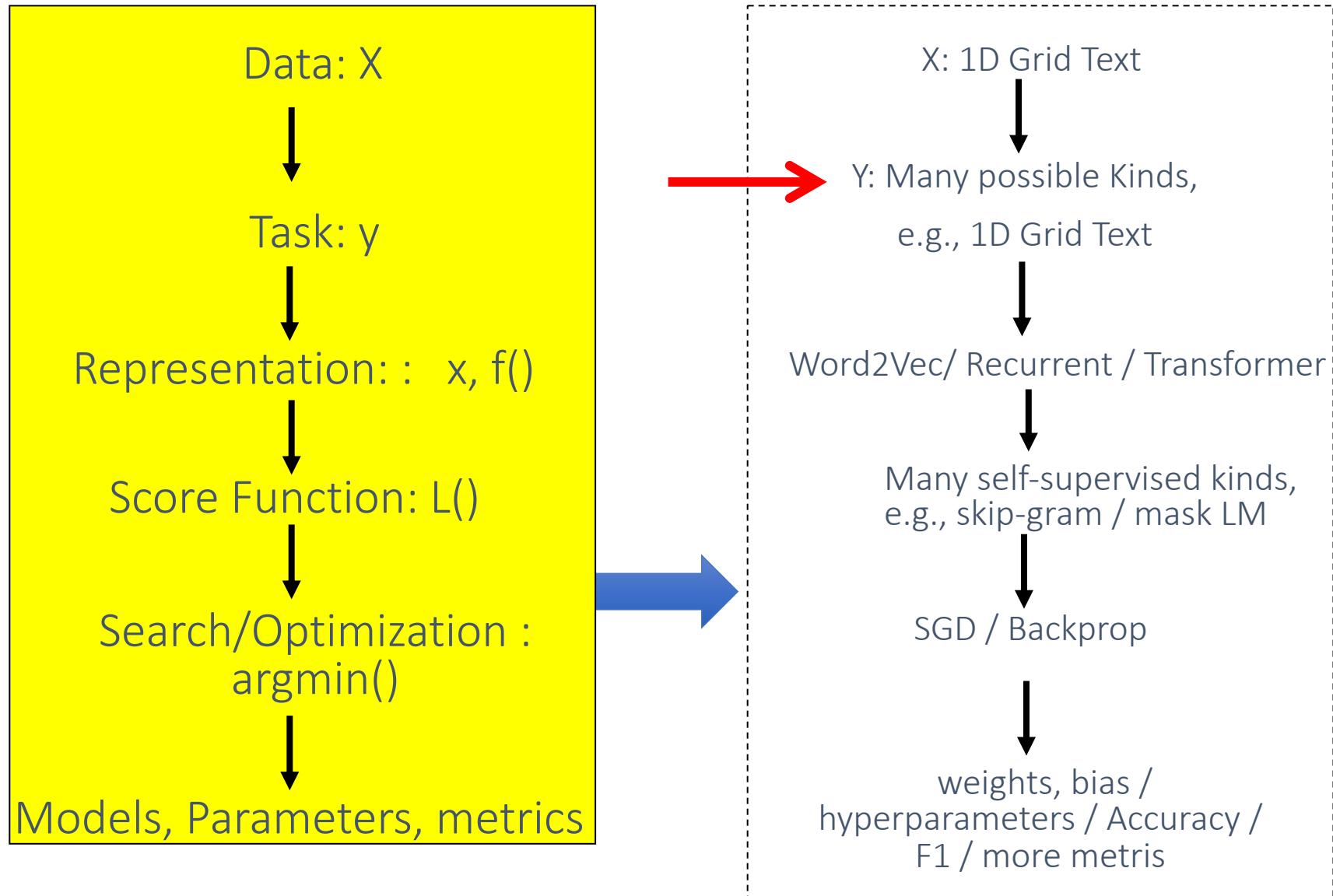
- Language grows and changes
  - e.g., cyber lingo



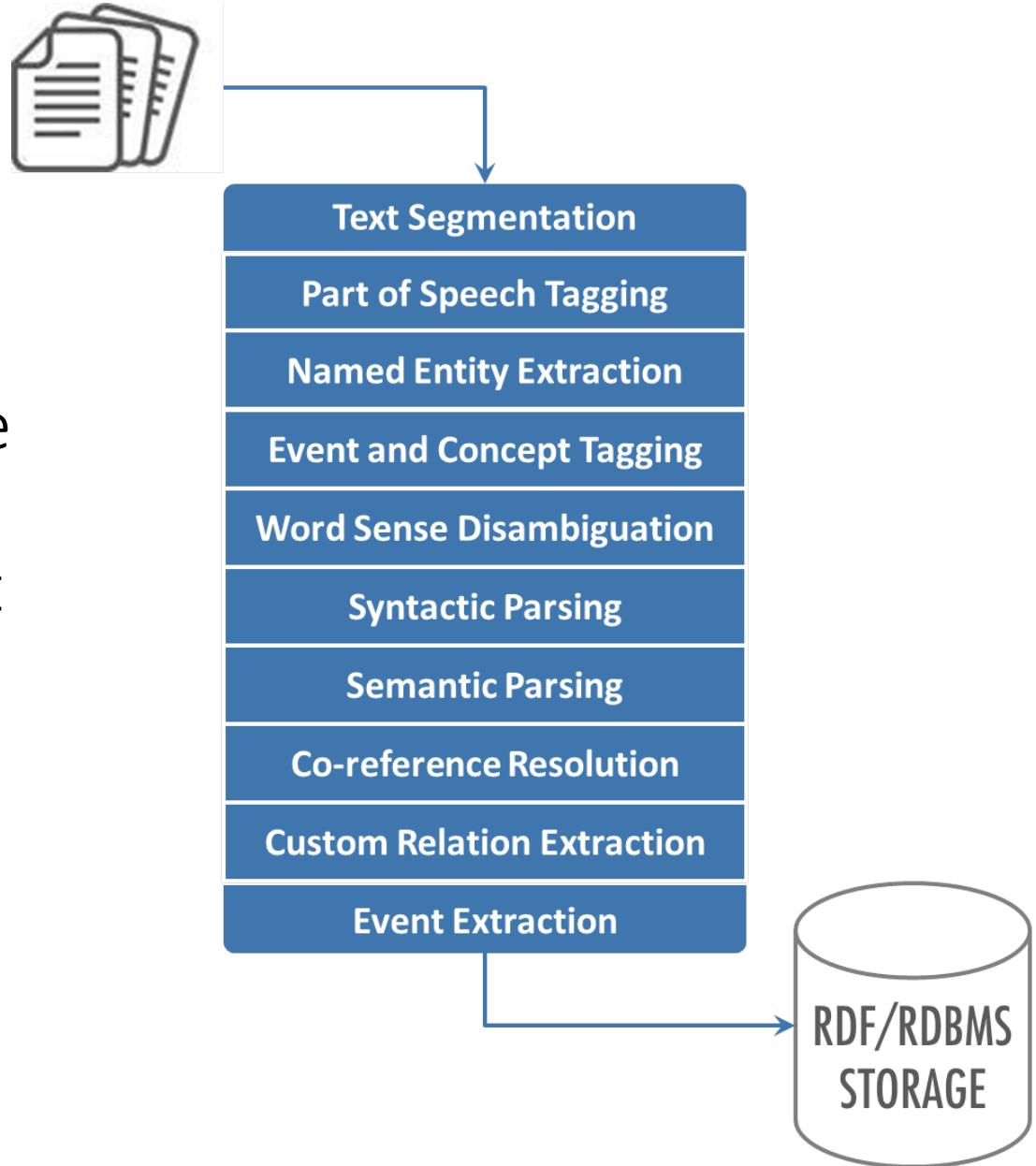
# Challenges – scale

- Examples:
  - Bible (King James version): ~700K
  - Penn Tree bank ~1M from Wall street journal
  - Newswire collection: 500M+
  - Wikipedia: 2.9 billion word (English)
  - Web: several billions of words

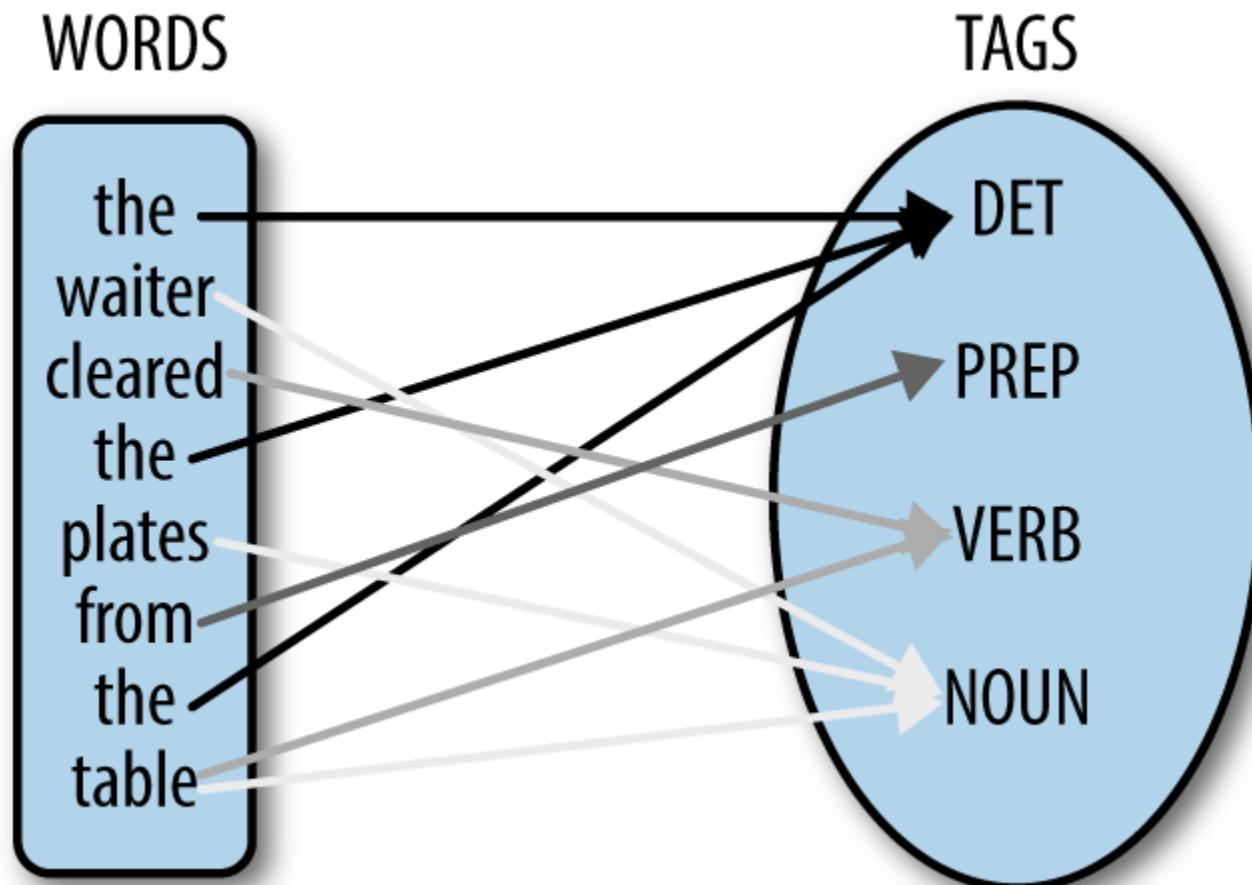
# Today: Neural Network Models on 1D Grid / Language Data



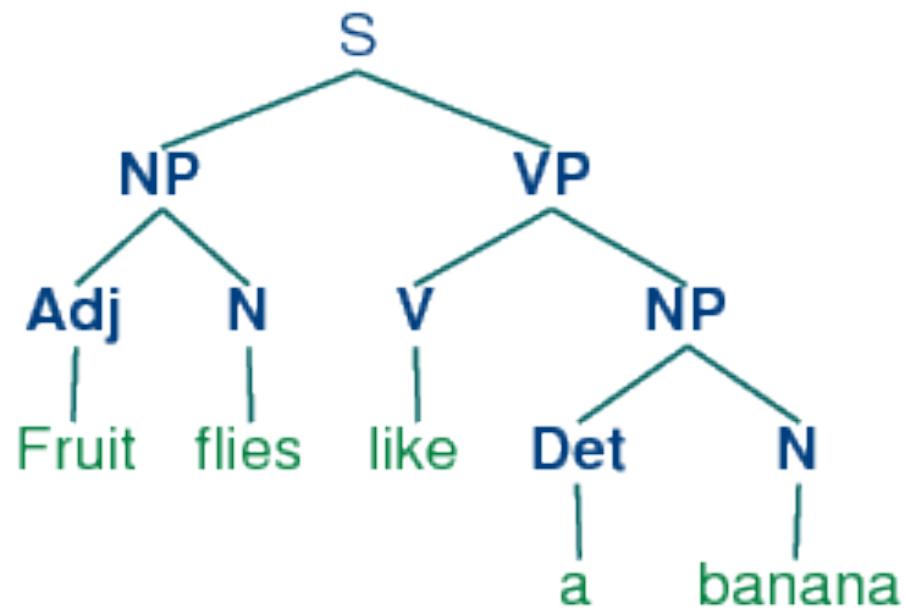
# Classic NLP Pipeline Components for Understanding Text



# Part of speech tagging



# Syntactic (Constituency) parsing



# Syntactic structure => meaning

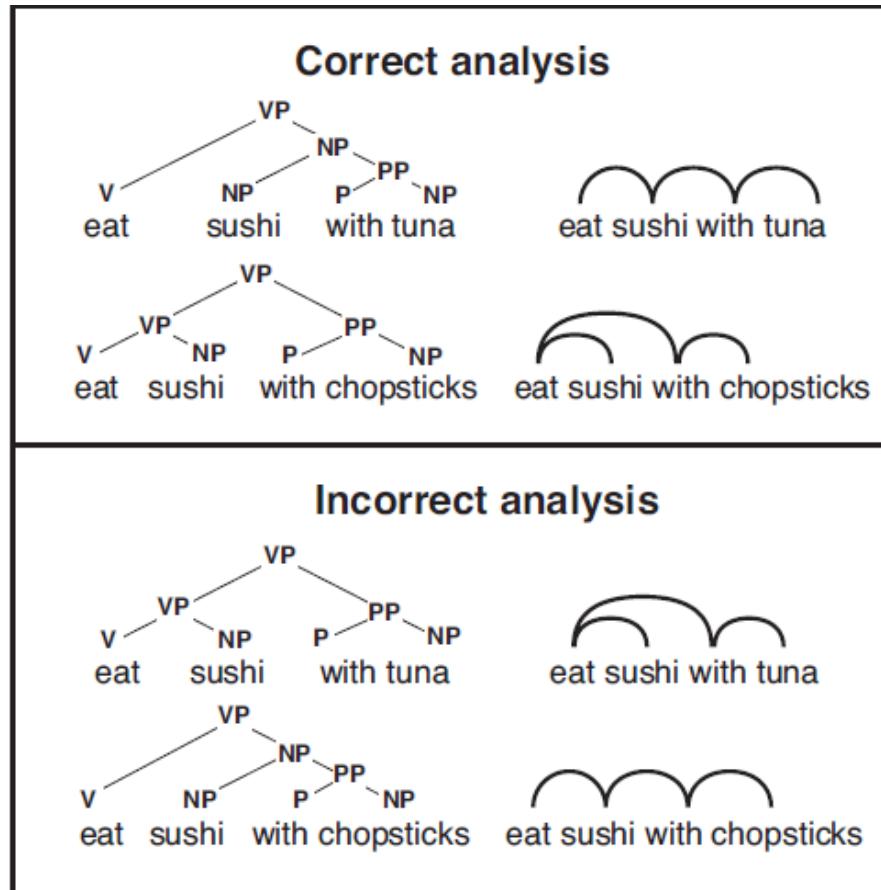
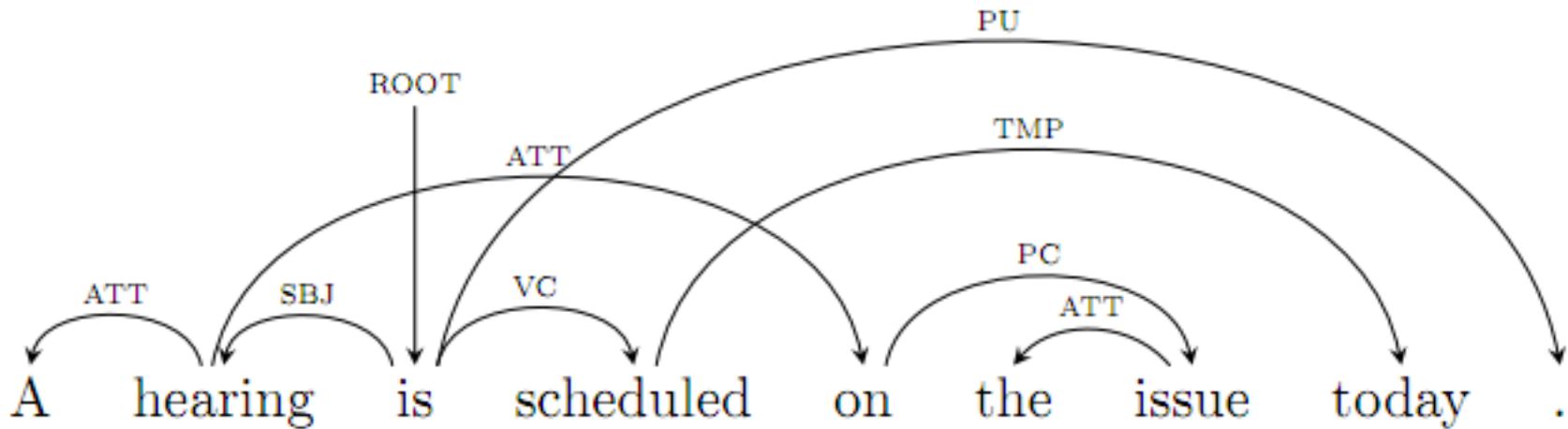


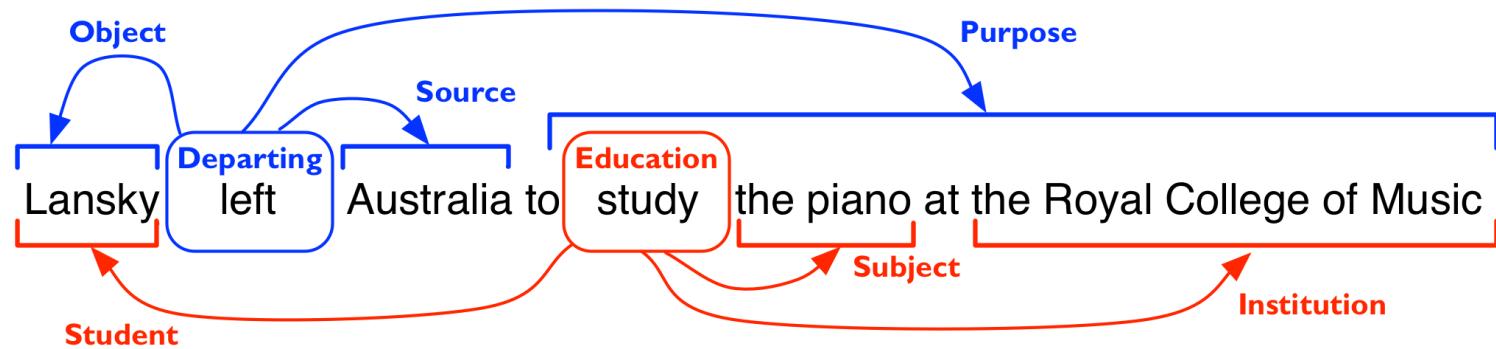
Image credit: Julia Hockenmaier, Intro to NLP

# Dependency Parsing



# Semantic analysis

- Word sense disambiguation
- Semantic role labeling



Credit: Ivan Titov

# Information Extraction

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

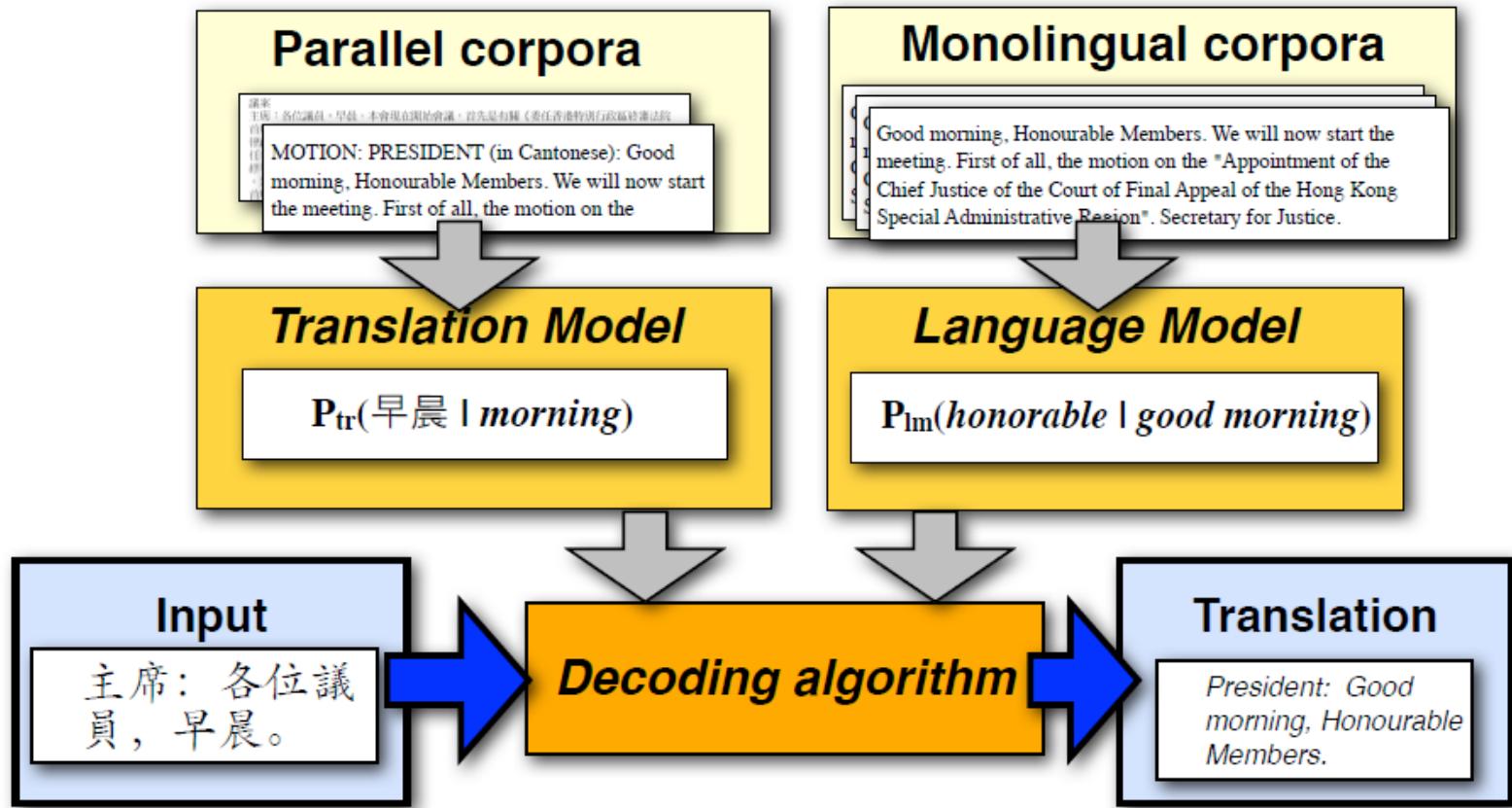
Q: [Chris] = [Mr. Robin] ?

**Christopher Robin** is alive and well. He is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

# Co-reference Resolution

**Christopher Robin** is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a **boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

# Statistical machine translation



# UVA CS 4774: Machine Learning

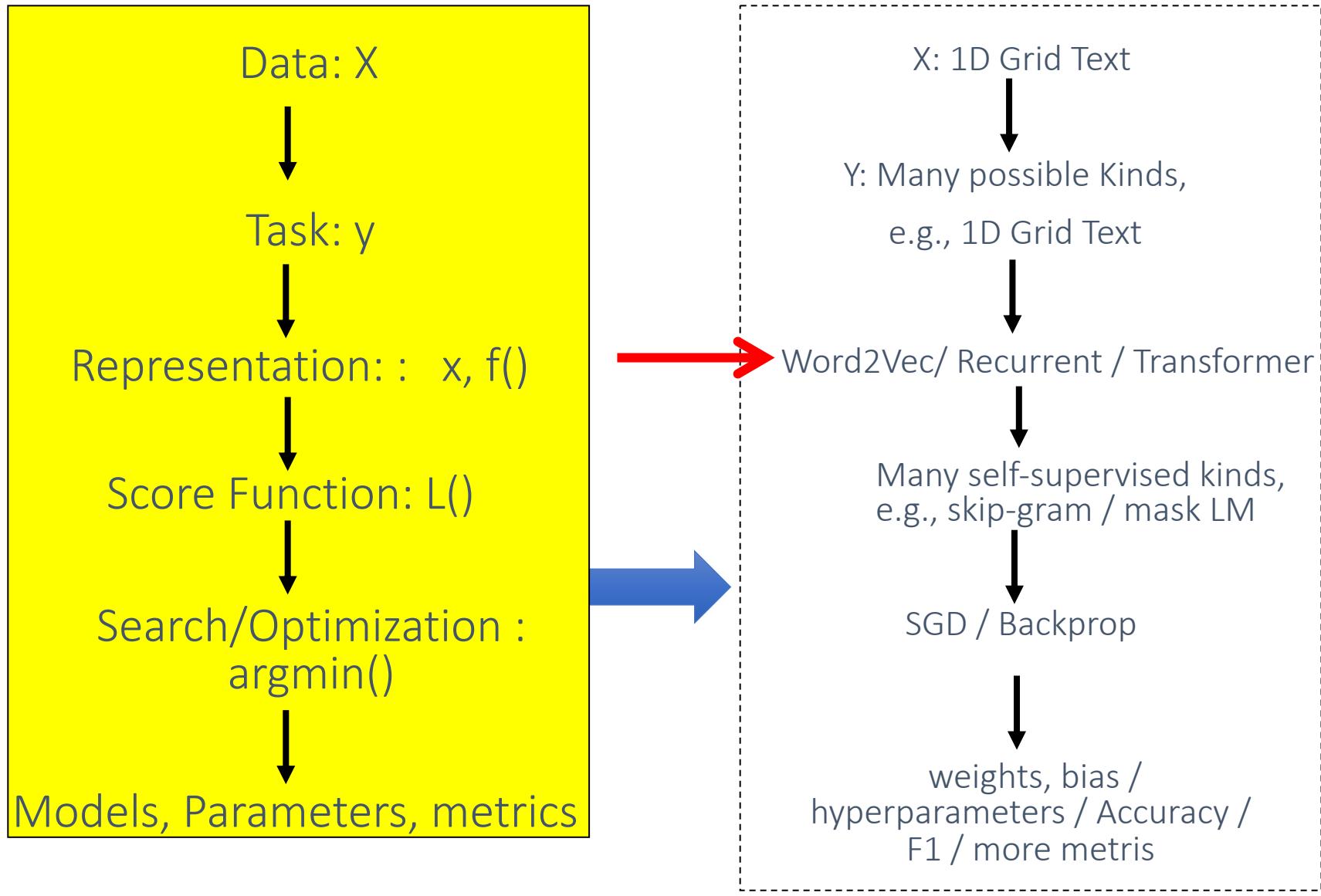
## S3: Lecture 18: Deep Neural Networks for Natural Language Processing

Dr. Yanjun Qi

University of Virginia  
Department of Computer Science

Module  
II

# Today: Neural Network Models on 1D Grid / Language Data



# Roadmap : f() on natural language



- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA )
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now )
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 ...

# Variable Length Issue in Natural Language Data:

X

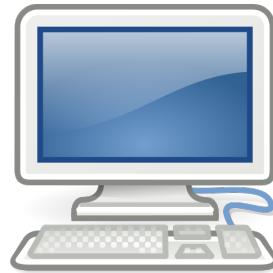
This Food is not good.



Y



This wonderful book is  
a pleasure to read.



# Recap: The bag of words representation

f( I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet. ) = c

## Recap: The bag of words representation

$$f(\begin{array}{|c|c|} \hline \text{great} & 2 \\ \hline \text{love} & 2 \\ \hline \text{recommend} & 1 \\ \hline \text{laugh} & 1 \\ \hline \text{happy} & 1 \\ \hline \dots & \dots \\ \hline \end{array}) = c$$

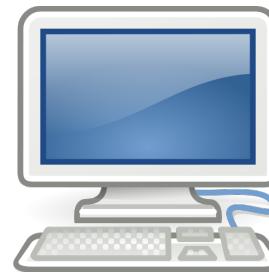
# BOW NOT Applicable to many NLP tasks:

- removes position information and can not (or hard to) represent word compositions

Y: French Translation

X

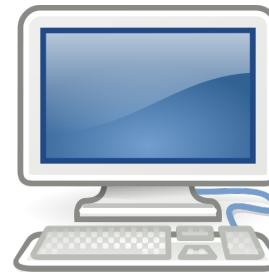
This Food is good.



FRENCH SPANISH ENGLISH

Cette nourriture est bonne.

This Food is very  
very good.



FRENCH SPANISH ENGLISH

Cette nourriture est très très bonne.

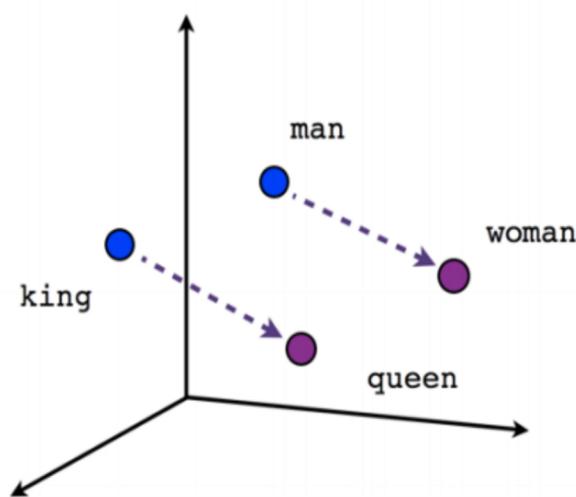
# Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA )
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now )
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 ...

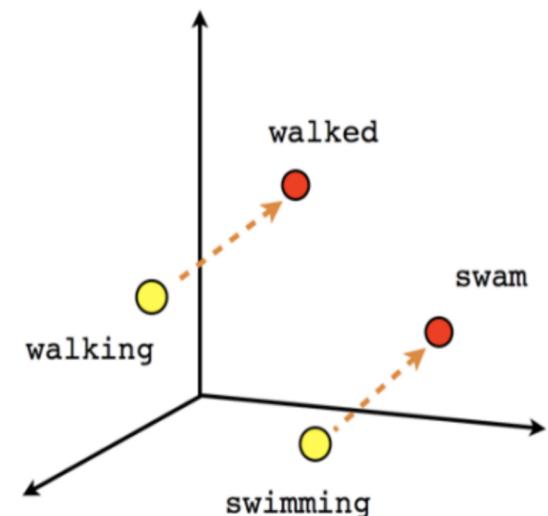
# How to Represent A Word in DNN

- Basic approach – “one hot vector”

- Binary vector
- Length = | vocab |
- 1 in the position of the word id, the rest are 0
- However, does not represent word meaning
- **Extremely high dimensional** (there are over 200K words in the English language)
- **Extremely sparse**



Male-Female



Verb tense

- Solution:  
**Distributional Word Embedding Vectors**

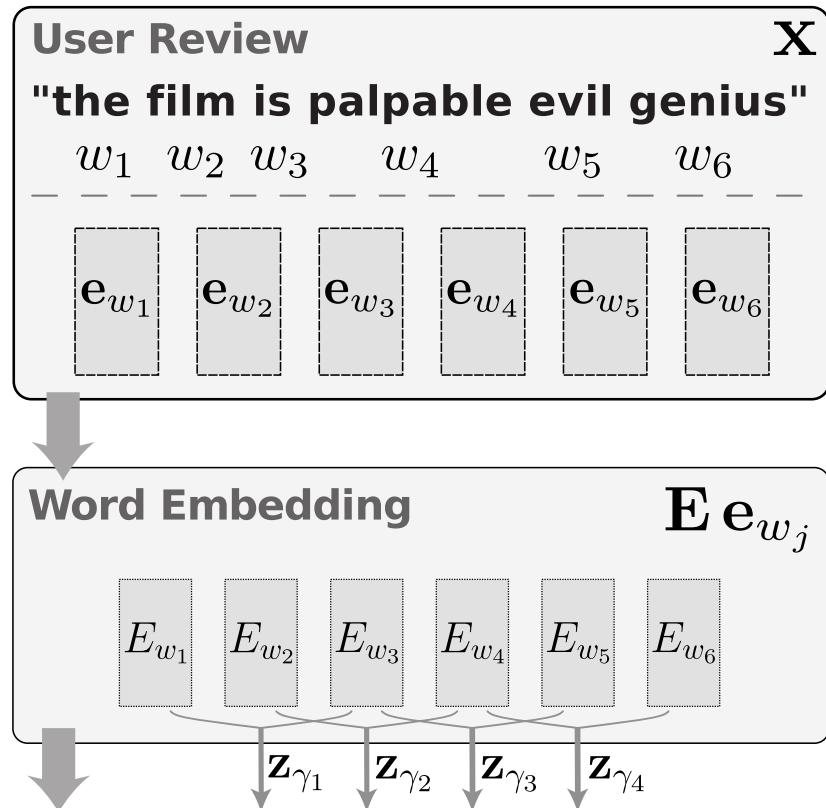
# Popular word embeddings

- **GloVe (Global Vectors)**
  - Pennington et al., 2014
- **fasttext**
  - Bojanowski et al., 2017

However, Natural language is

- Variable-length
- Composition of multiple words
- Word meaning is contextual

- Elmo
  - Peters, 2018
- BERT
  - Devlin et al., 2018

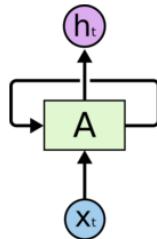


# Roadmap : f() on natural language

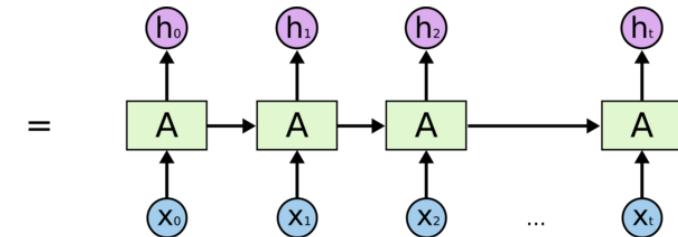
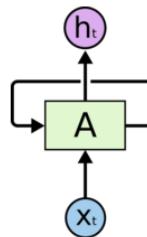
- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA )
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- • Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now )
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 ...

# Recurrent Neural Networks

- Allow us to operate over sequences of vectors (with variable length)
- Allow Sequences in the input, as the output, or in the most general case both



Recurrent Neural Networks have loops.



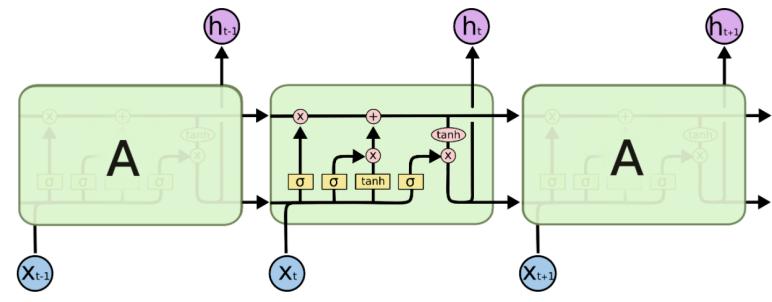
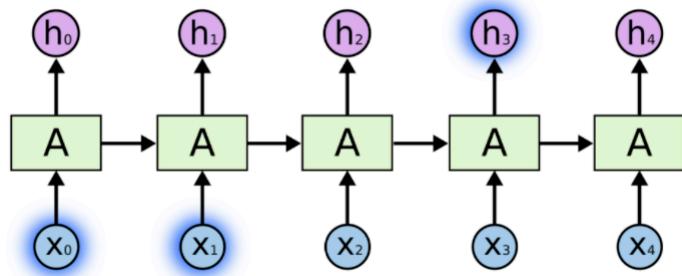
An unrolled recurrent neural network.

Recurrent Neural Networks are networks with loops in them, allowing information to persist.

Image Credits from Christopher Olah

# Deep RNN in the 90's

- Prof. Schmidhuber invented "Long short-term memory" – Recurrent NN (LSTM-RNN) model in 1997



The repeating module in an LSTM contains four interacting layers.

Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780.

Image Credits from Christopher Olah

# Recurrent Neural Networks Got Popular

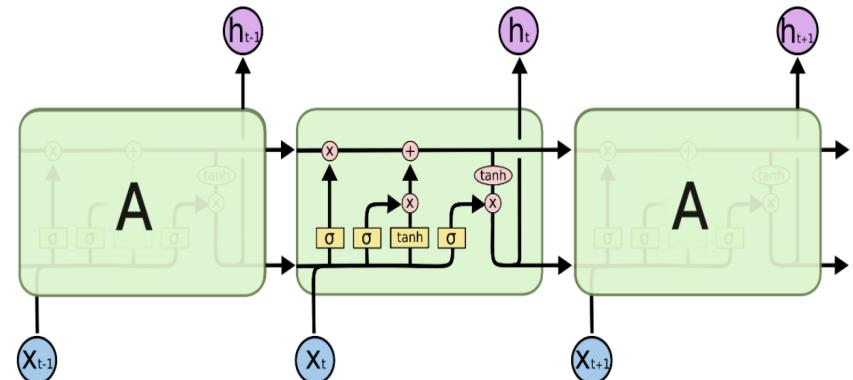
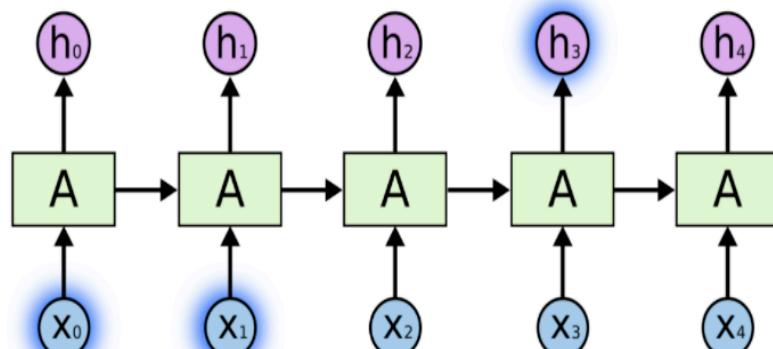
- Incredible success applying RNNs to language modeling and sequence learning problems

Task	Input Sequence	Output Sequence
Machine translation (Sutskever et al. 2014)	English	French
Question answering (Bordes et al. 2014)	Question	Answer
Speech recognition (Graves et al. 2013)	Voice	Text
Handwriting prediction (Graves 2013)	Handwriting	Text
Opinion mining (Irsoy et al. 2014)	Text	Opinion expression

# LSTM

- "Long short-term memory" – Recurrent NN (LSTM-RNN)

$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}) = \overrightarrow{LSTM}(\mathbf{x}_t)$$



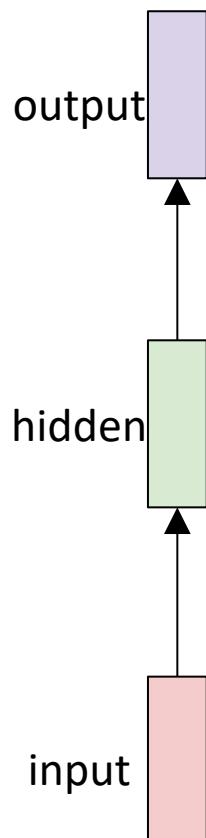
The repeating module in an LSTM contains four interacting layers.

Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780.

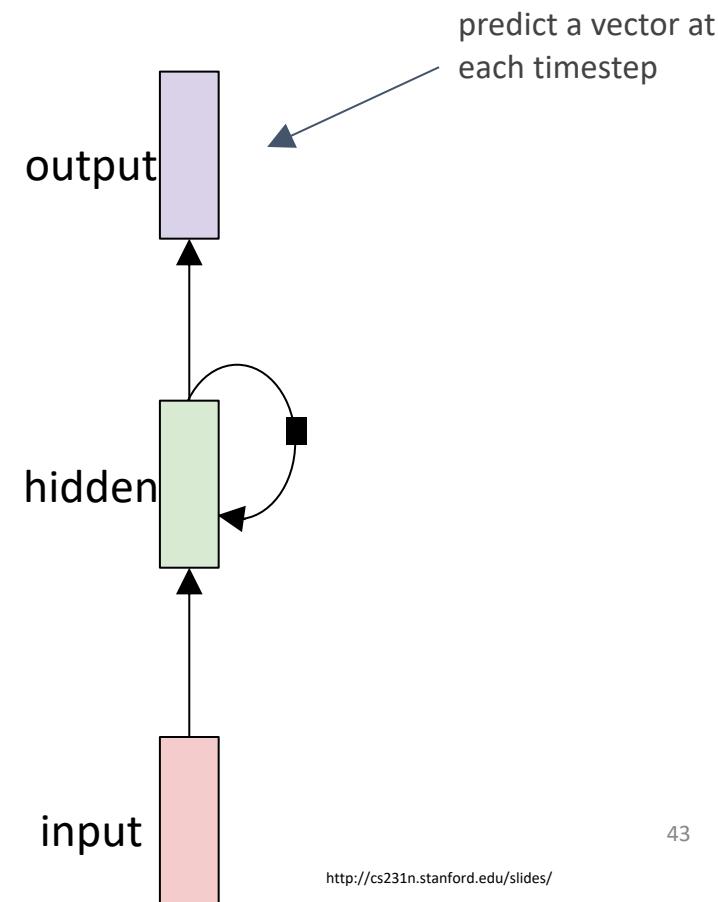
# RNN models dynamic temporal dependency

- Make **fully-connected** layer model each unit recurrently
- Units form a **directed chain graph** along a sequence
- Each unit uses **recent history** and current input in modeling

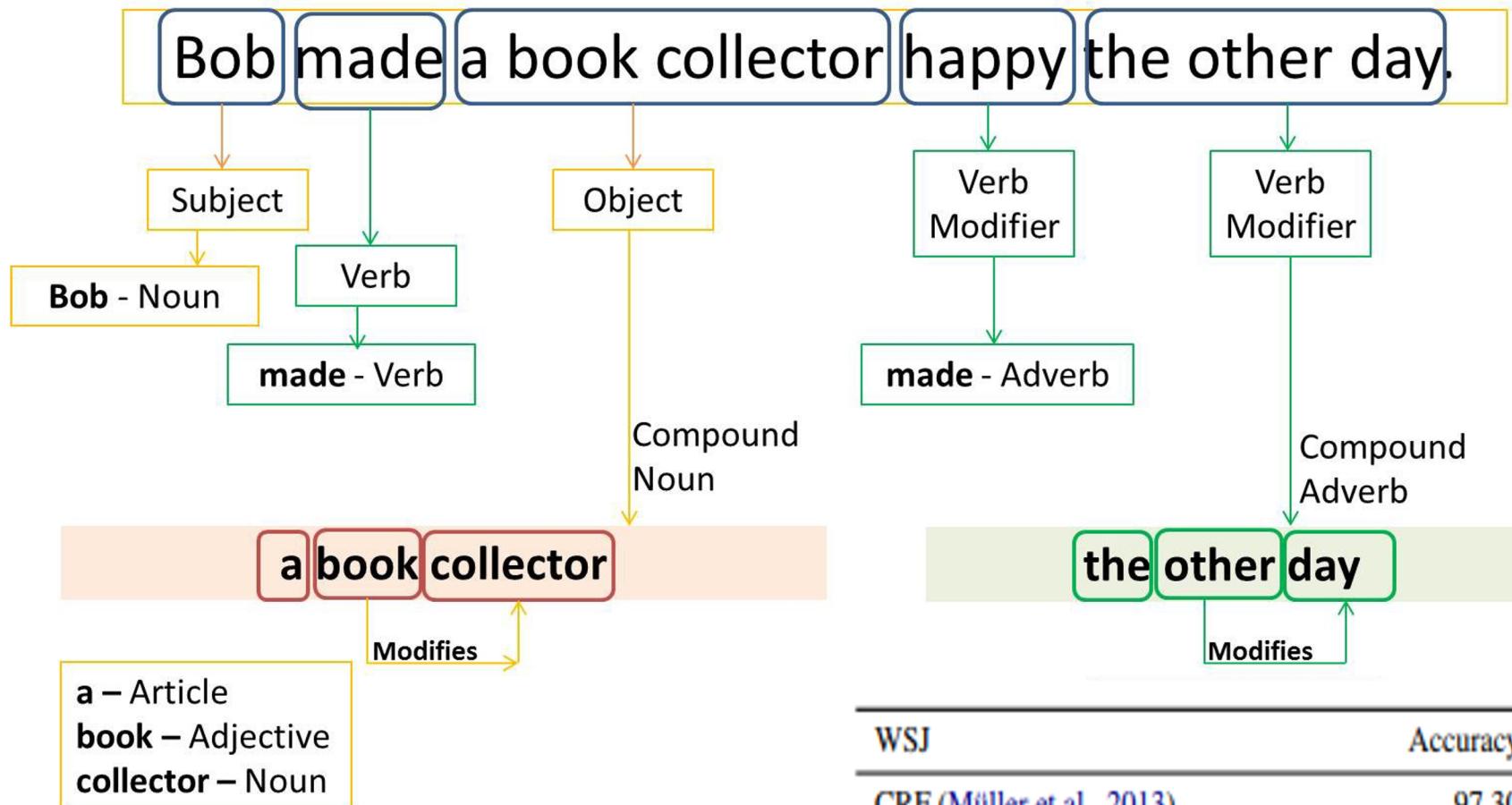
**Traditional “Feed Forward”  
Neural Network**



**Recurrent Neural Network**



# POS tagging (solved by CovNet or RNN-LSTM)



<https://nlp.stanford.edu/software/tagger.shtml>

<https://www.nltk.org/book/ch05.html>

Table 2: Tagging accuracy on the WSJ test set.

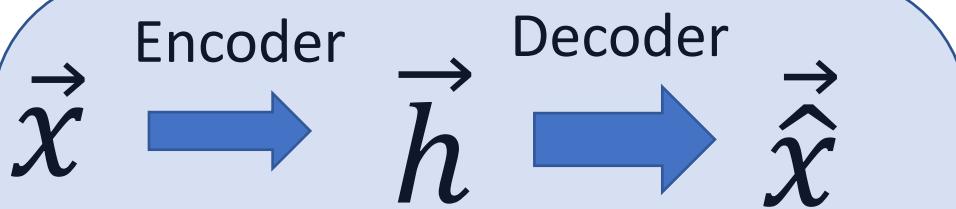
# RNN can models variable-length input / output

Anything requiring  
long-range patterns

- Question detection
- Natural language context understanding
- Entity disambiguation
- Sentence embedding

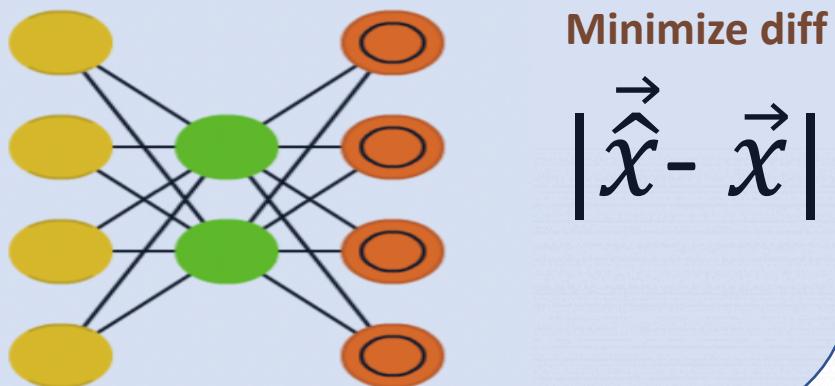
Anything generative

- Machine translation
- Natural language generation
- Question answering
- Skip-thoughts

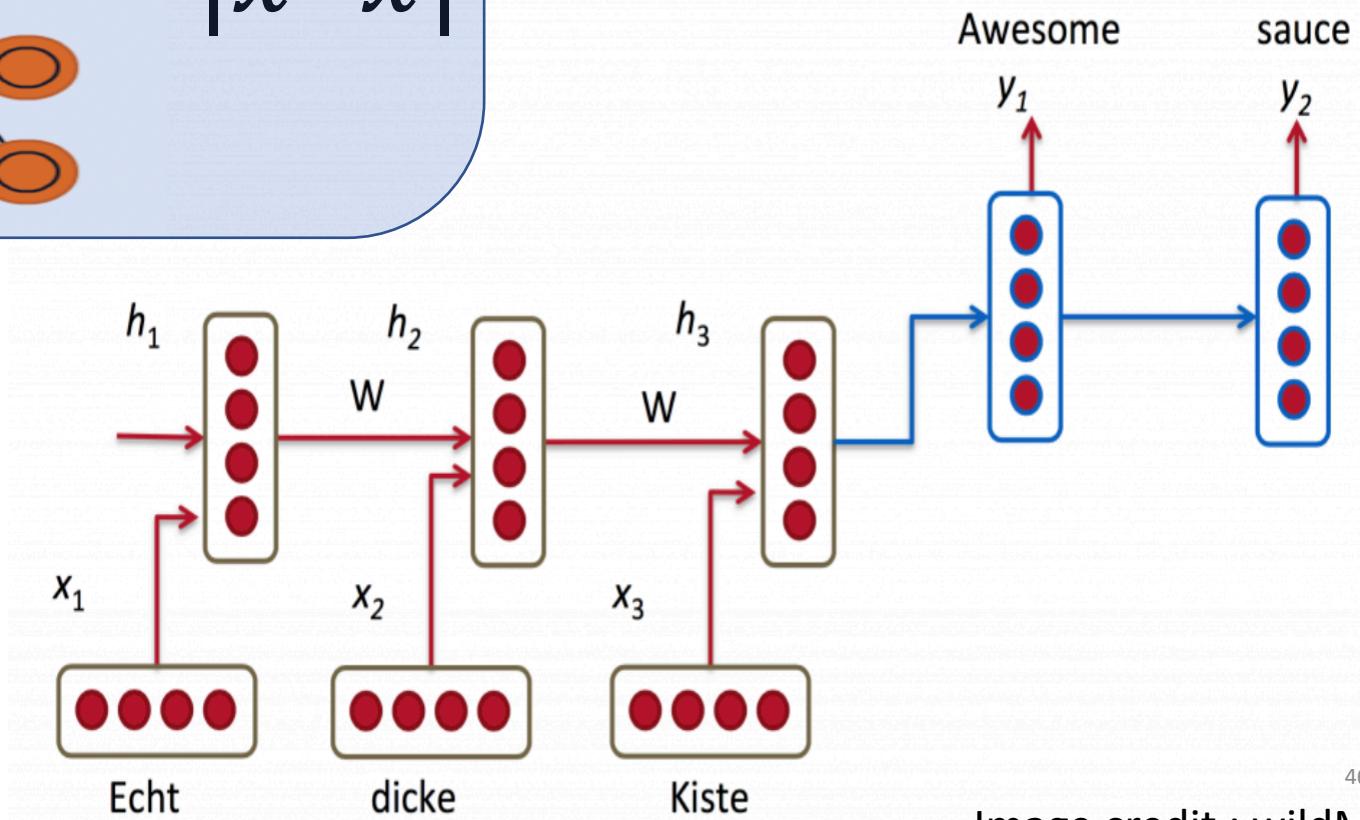


RNN can models variable-length input / output

Auto Encoder (AE)



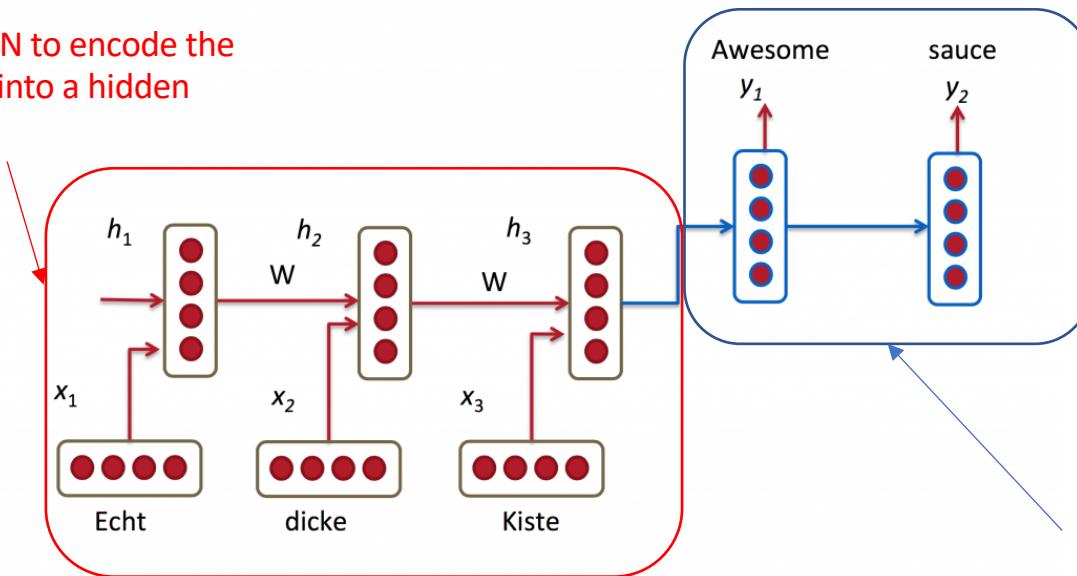
## Seq2Seq



# Seq2Seq for Machine Translation

In machine translation, the input is a sequence of words in source language, and the output is a sequence of words in target language.

Encoder: An RNN to encode the input sentence into a hidden state (feature)



Encoder-decoder architecture for machine translation

Two LSTMs for Machine Translation (German to English)

- Encoder LSTM (on German)
- Decoder LSTM (on English)

Decoder: An RNN with the hidden state of the sentence in source language as the input and output the translated sentence

# Seq2Seq for more Sequence-to-Sequence Generation Tasks

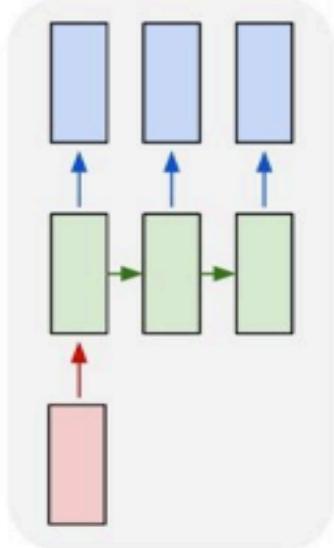
Given source sentences, learn an optimal model to automatically generate accurate and diversified target sentences that look like human generated sentences.



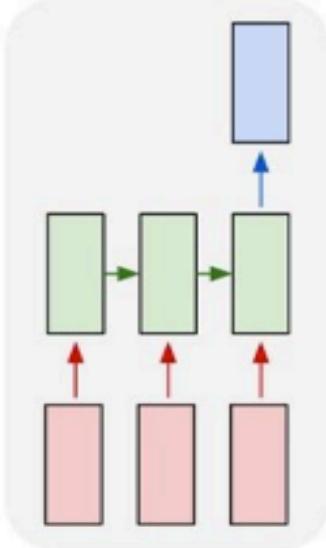
- **Paraphrase generation:** “How did Trump win the election?” → “How did Trump become president?”
- **Dialogue generation:** “You know French?” → “Sure do ... my Mom's from Canada”
- **Question answering:** “What was the name of the 1937 treaty?” → “Bald Eagle Protection Act”
- **Style Transfer:** “Just a dum funny question hahahaha” → “Just a senseless , funny question.”

# Recurrent Neural Networks (RNNs) can handle

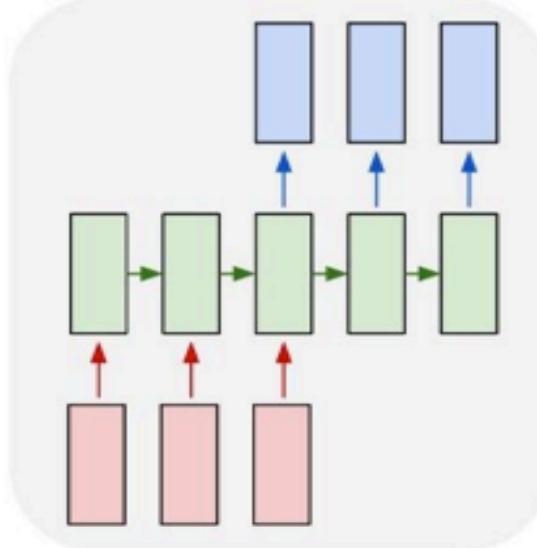
one to many



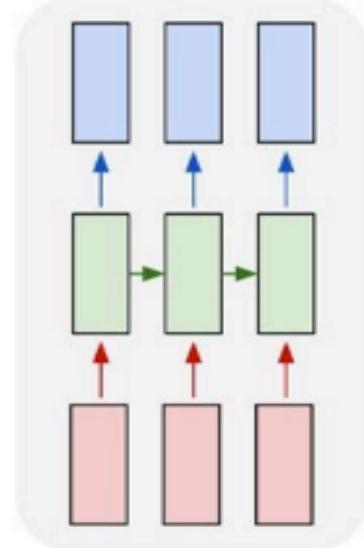
many to one



many to many

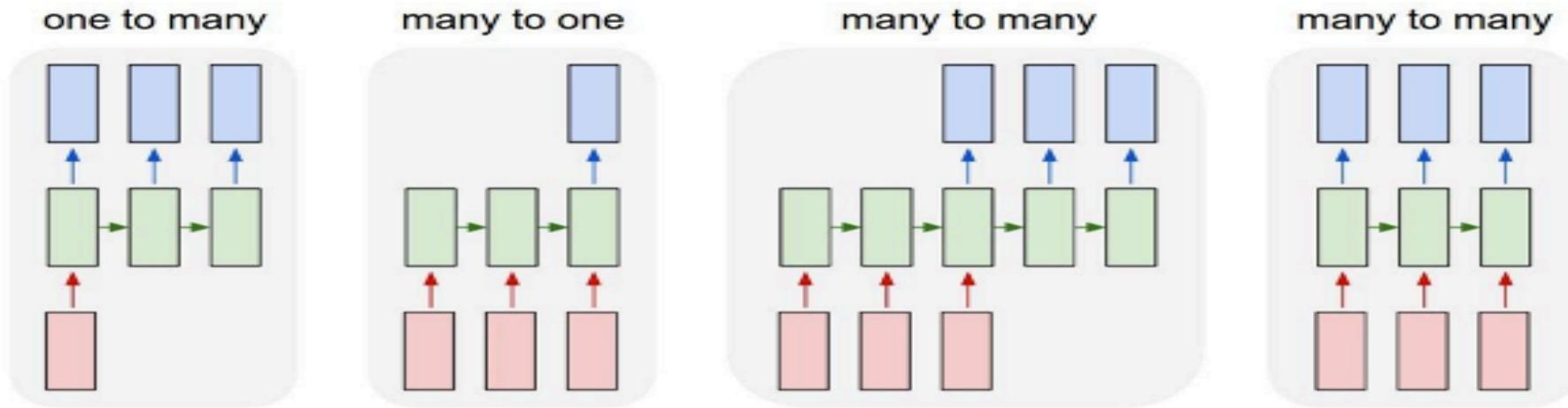


many to many



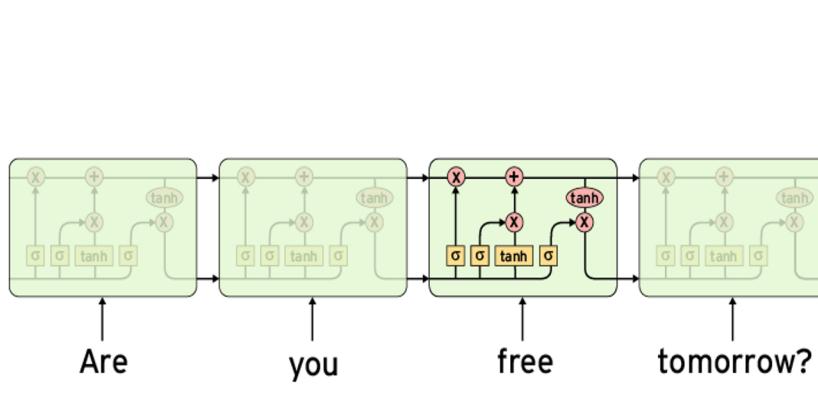
e.g. **Sentiment Classification**  
sequence of words -> sentiment

# Recurrent Neural Networks (RNNs) can handle



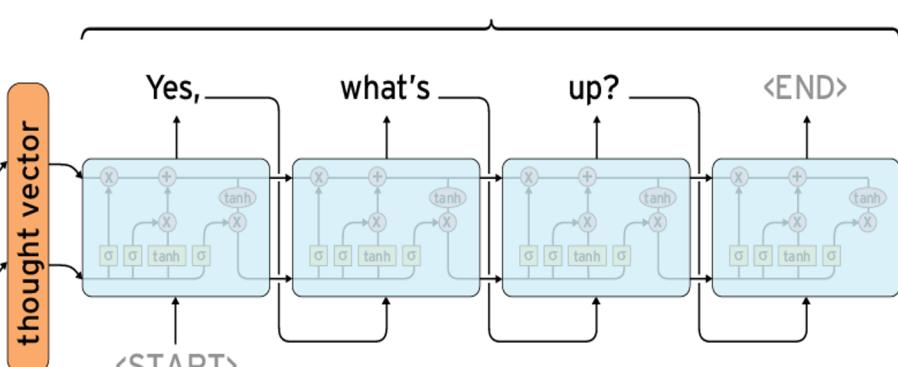
e.g. **Machine Translation**  
seq of words → seq of words

ENCODER



Incoming Email

Reply



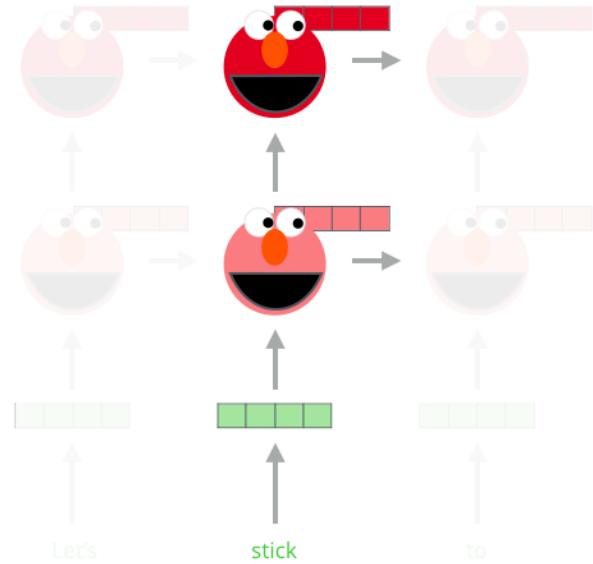
DECODER

## Embedding of “stick” in “Let’s stick to” - Step #2

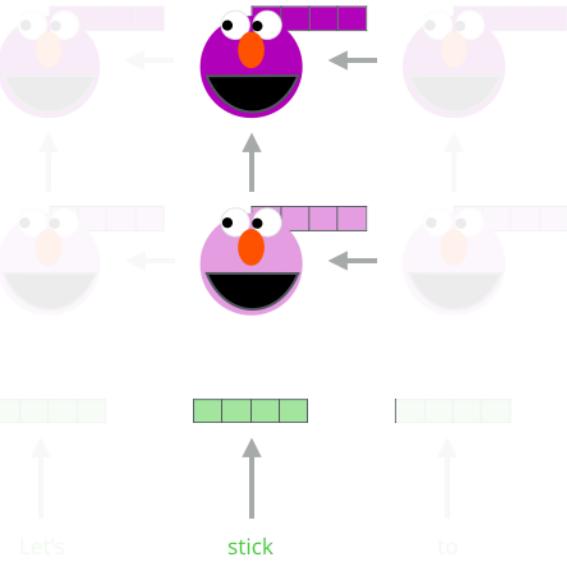
1- Concatenate hidden layers



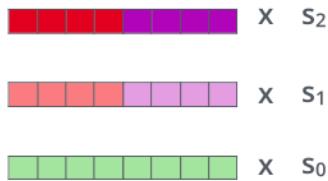
Forward Language Model



Backward Language Model



2- Multiply each vector by a weight based on the task



3- Sum the (now weighted) vectors



ELMo embedding of “stick” for this task in this context

**contextual embedding**

ELMo's embedding of a word given the sentence is the concatenation of its biLSTM's hidden states for the word.

# UVA CS 4774: Machine Learning

## S3: Lecture 18: Deep Neural Networks for Natural Language Processing

Dr. Yanjun Qi

University of Virginia  
Department of Computer Science

Module  
III

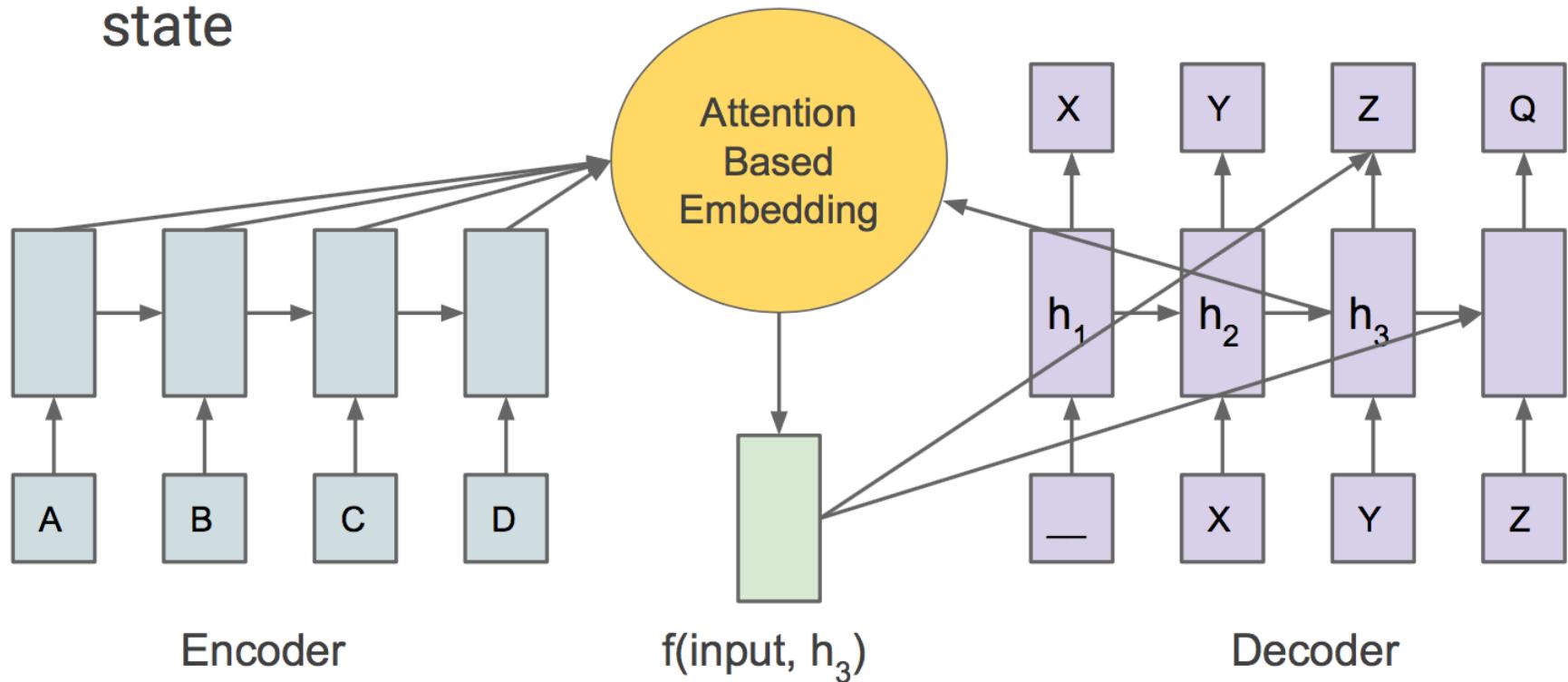
# Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA )
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now )
  - • Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 ...

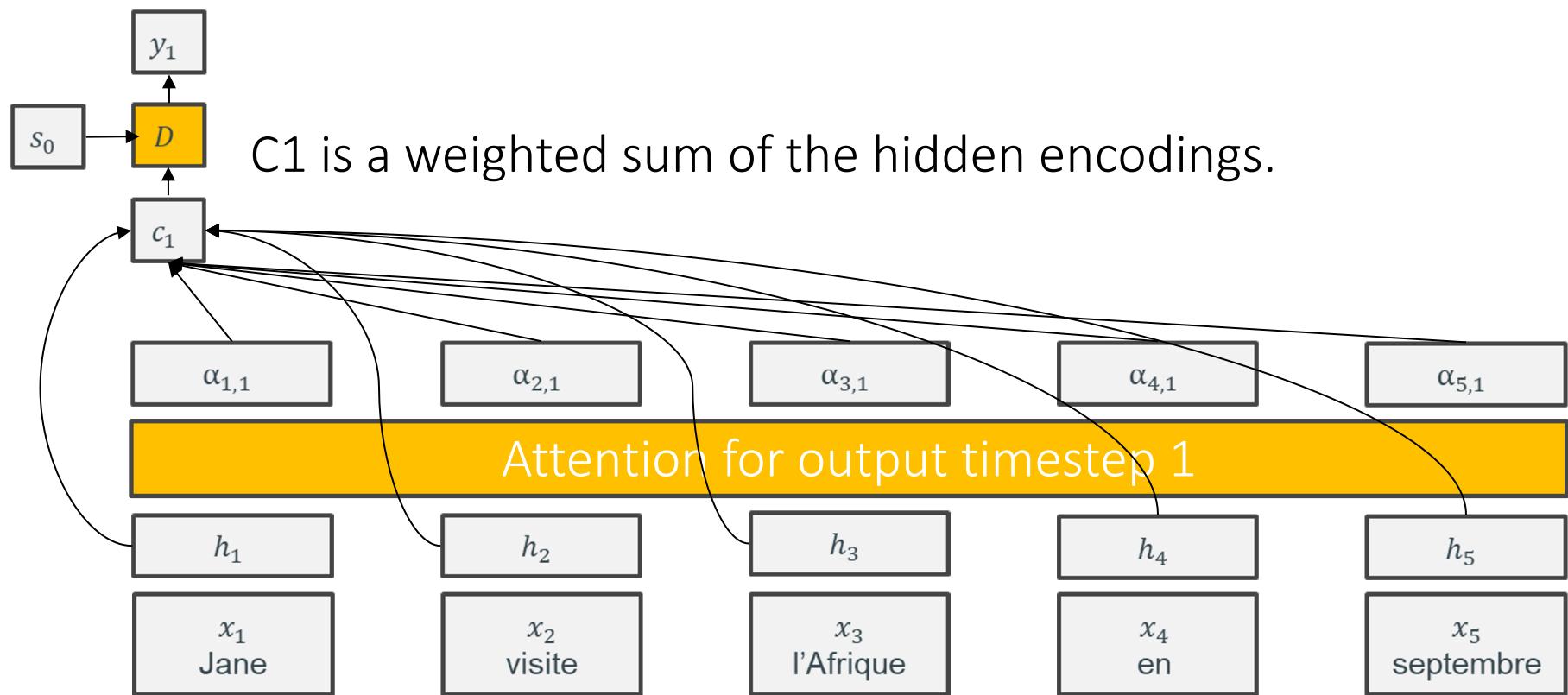
## Attention Trick:

# Seq2Seq with Attention

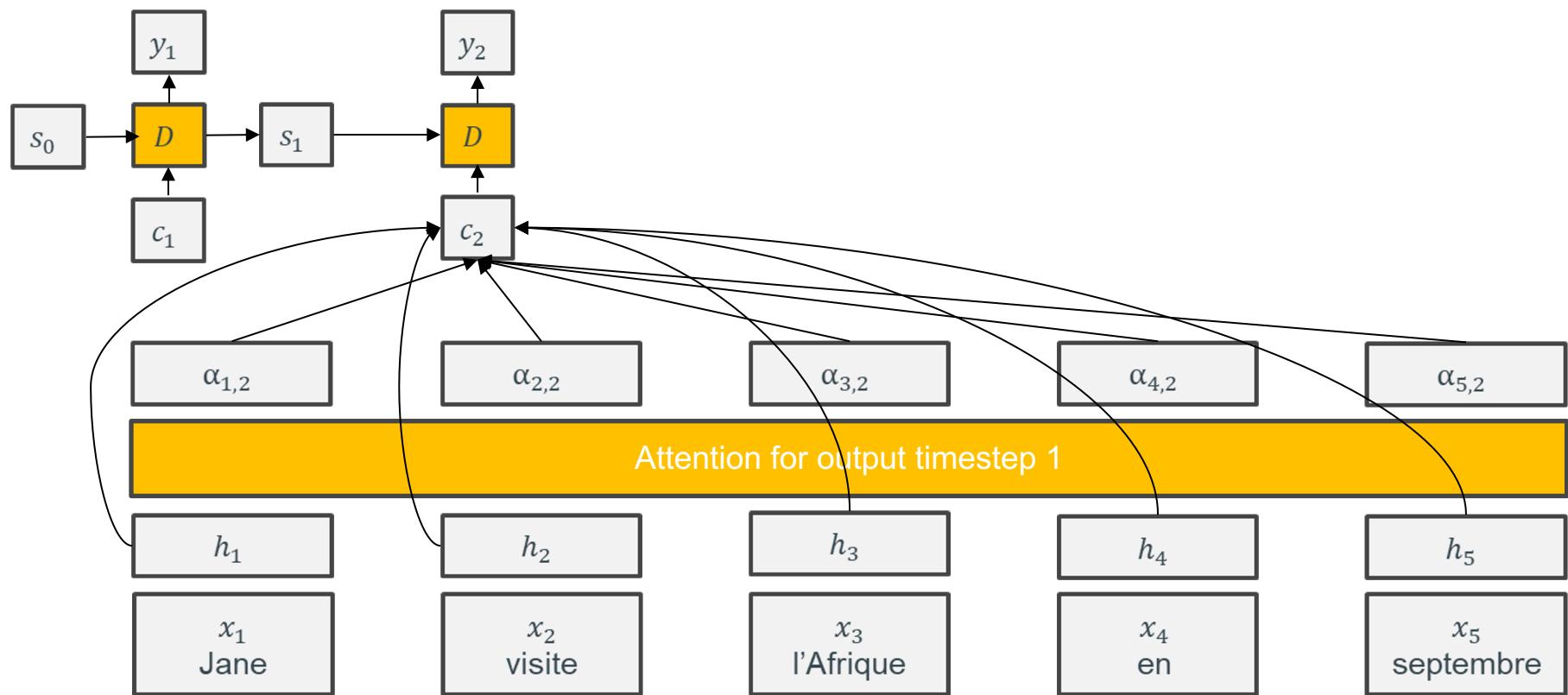
- Embedding used to predict output, and compute next hidden state



The attention module gives us a weight for each input.



We then repeat for future timesteps.



# Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA )
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now )
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 ...



Self-attention creates attention layers mapping from a sequence to itself.

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

# Transformer: Exploiting Self Attentions

- A Google Brain model.
  - Variable-length input
  - Fixed-length output (but typically extended to a variable-length output)
  - **No recurrence**
  - Surprisingly not patented.
- Uses 3 kinds of attention
  - Encoder self-attention.
  - Decoder self-attention.
  - Encoder-decoder multi-head attention.

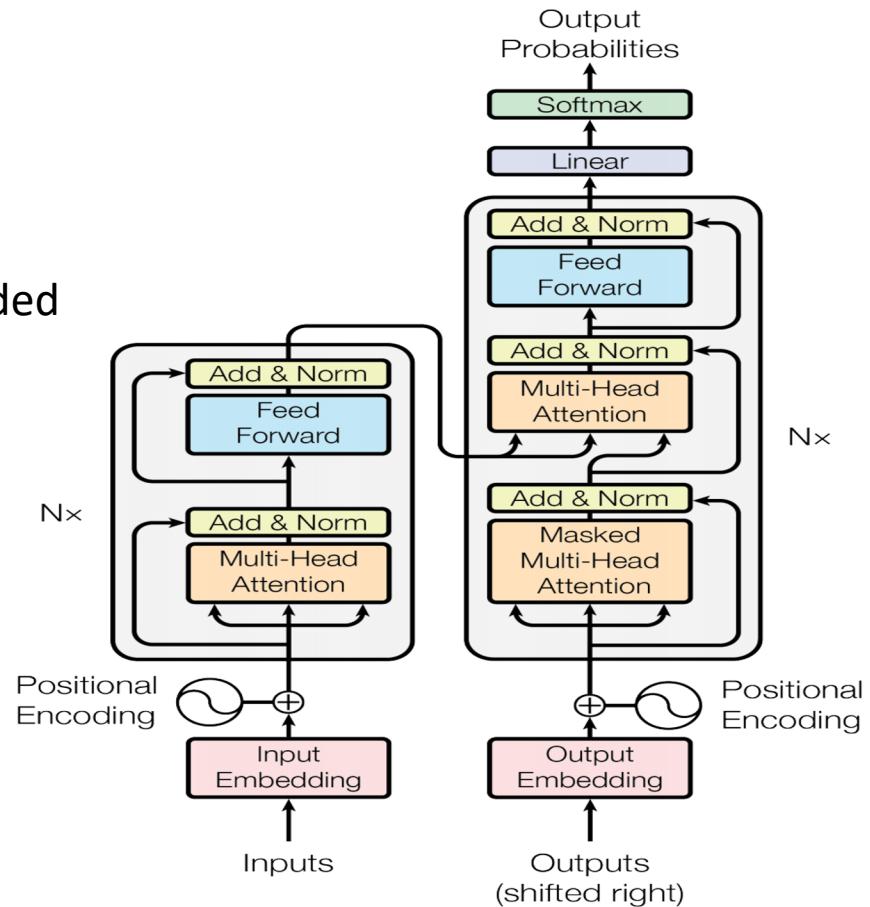
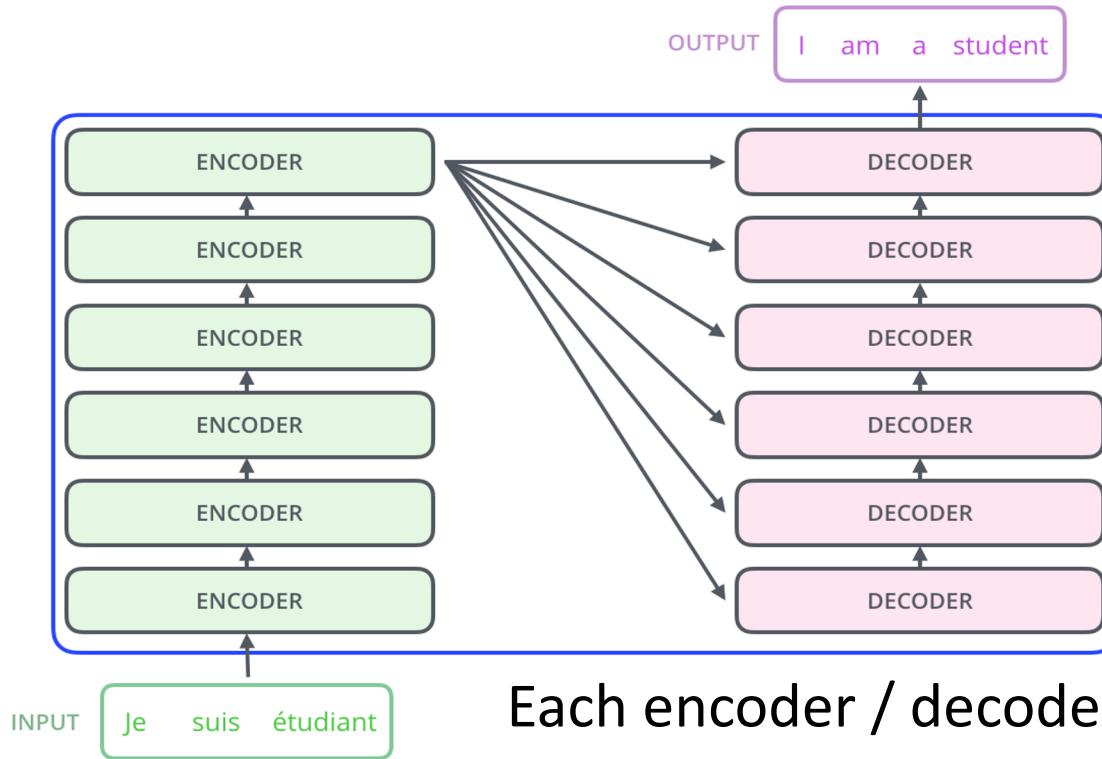
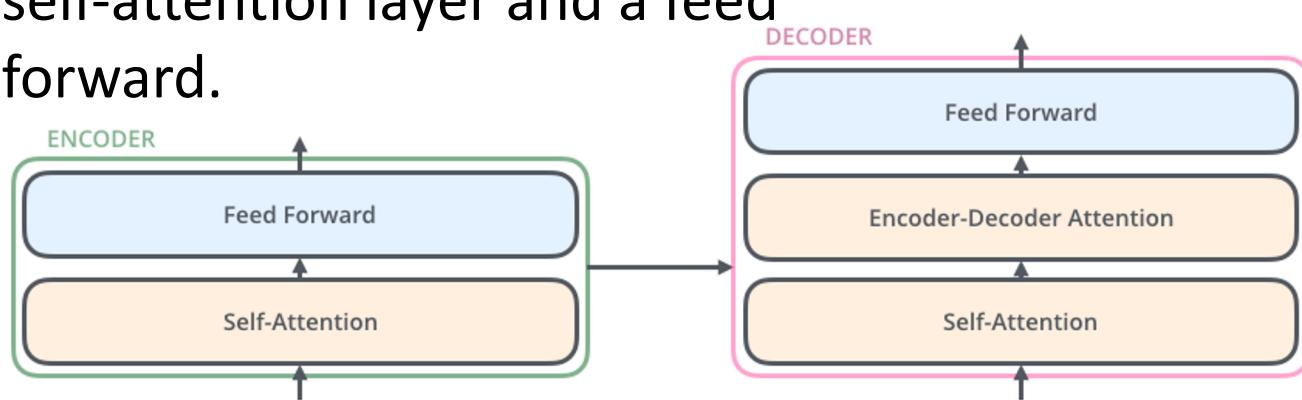


Figure 1: The Transformer - model architecture.

# Transformer is Seq2Seq model



Each encoder / decoder layer has a self-attention layer and a feed forward.



# Roadmap : f() on natural language

- Before Deep NLP (Pre 2012)
  - (BOW / LSI / Topic LDA )
- Word2Vec (2013-2016)
  - (GloVe/ FastText)
- Recurrent NN (2014-2016)
  - LSTM
  - Seq2Seq
- Attention / Self-Attention (2016 – now )
  - Attention
  - Transformer (self-attention, attention only)
  - BERT / XLNet/ GPT-2 / T5 ...

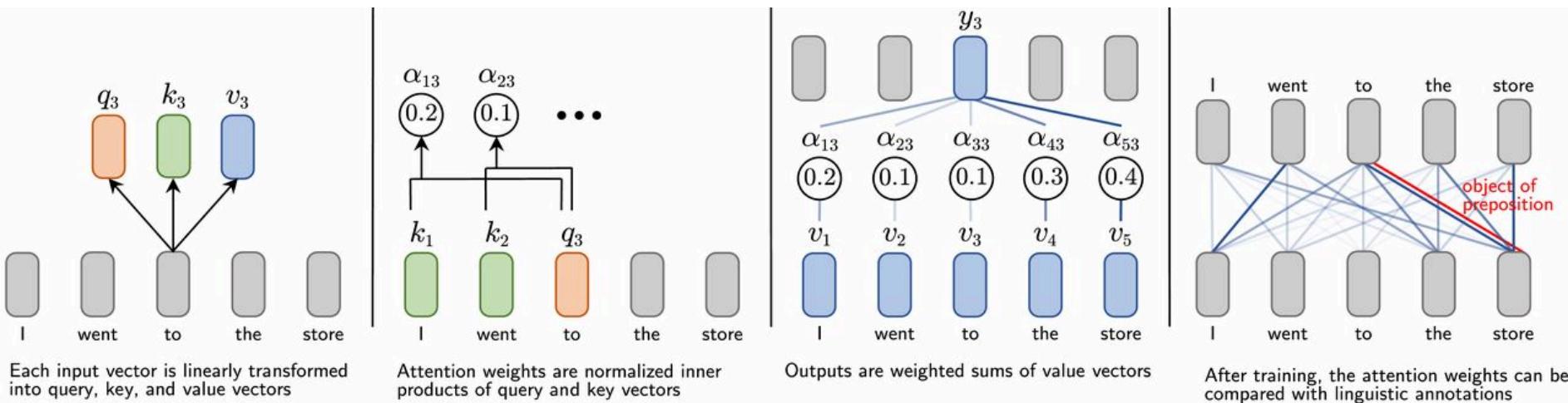


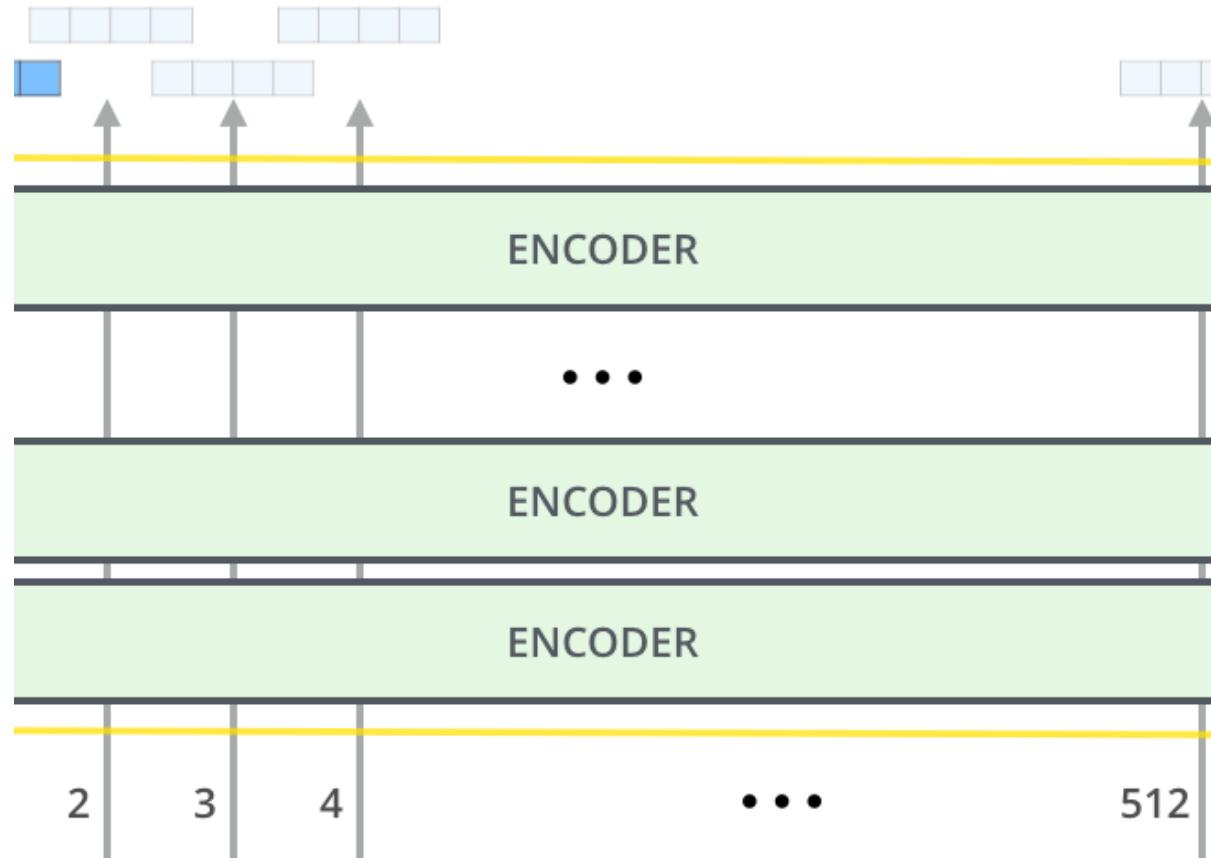


BERT: Bidirectional Encoder Representations from Transformers  
Pre-trained transformer encoder for sentence embedding



## Notable pre-trained NLP models





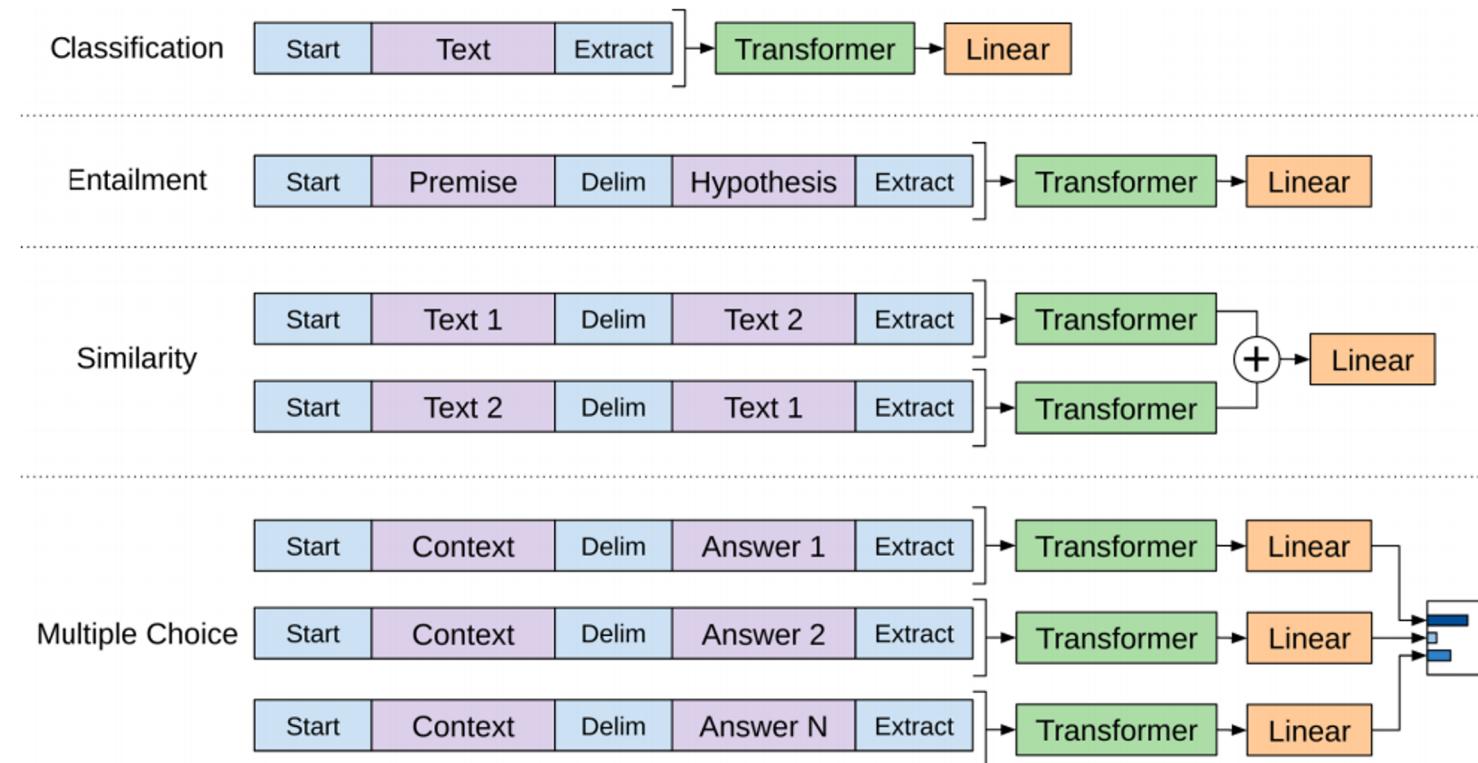
] Help Prince Mayuko

BERT

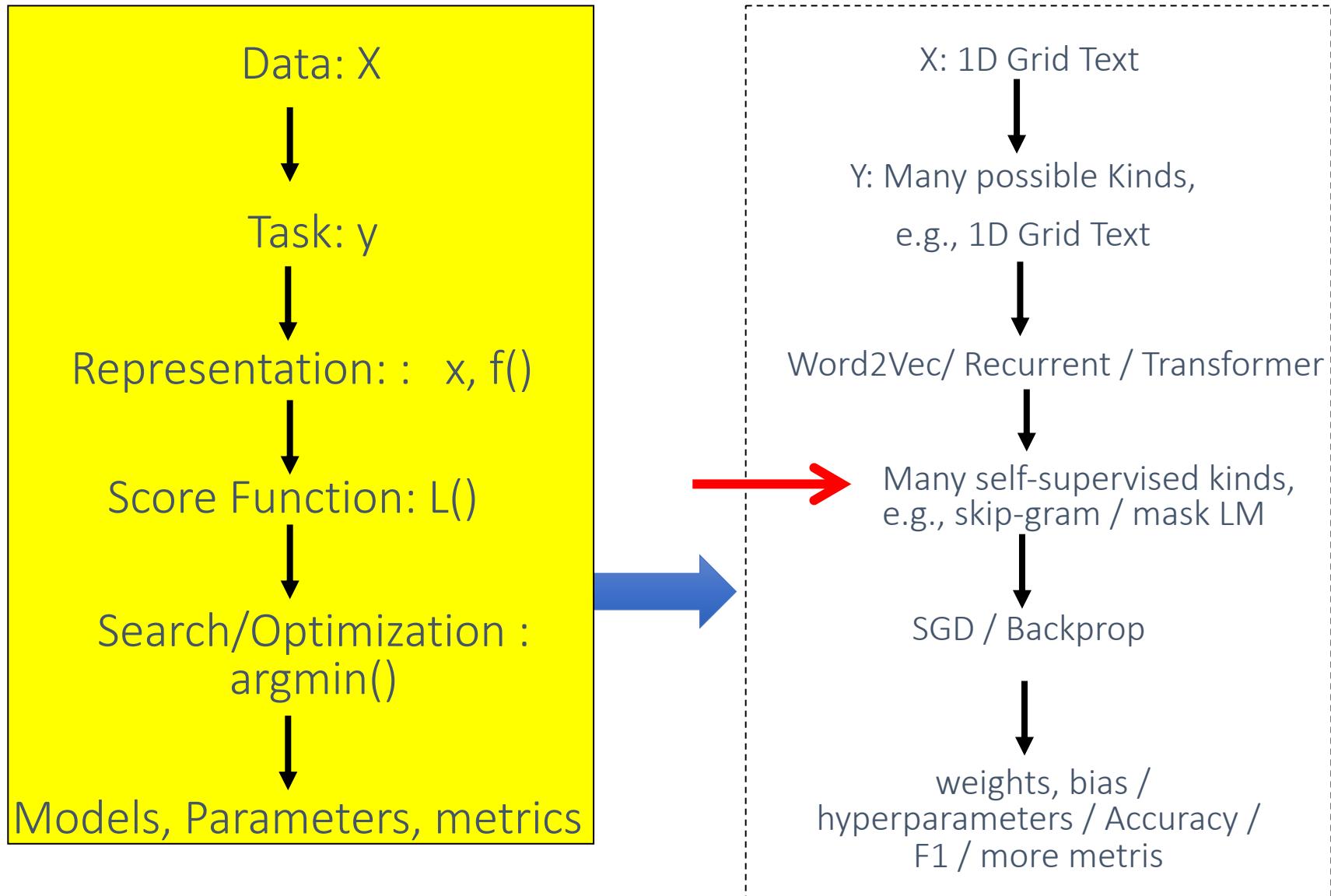
BERT's  
architecture is  
just a  
transformer's  
encoder stack.

Open AI's GPT-2 is a really large transformer: 1.5 billion parameters! Trained on 8M pages

As with BERT, you can use the pretrained model for any task.  
Different tasks use the OpenAI transformer in different ways.

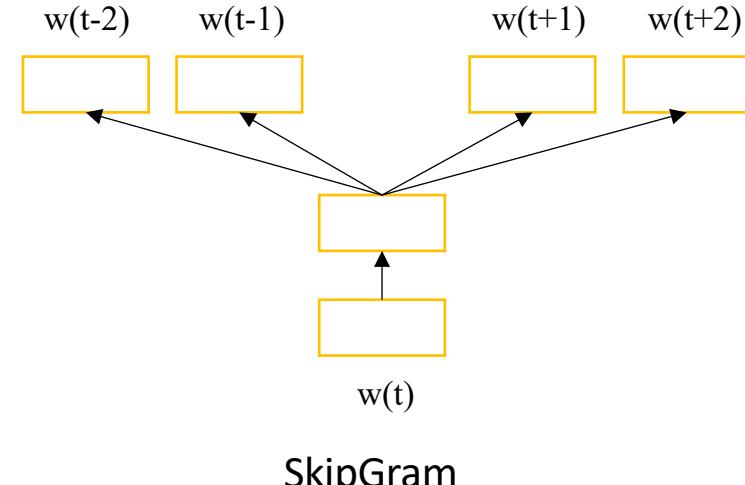
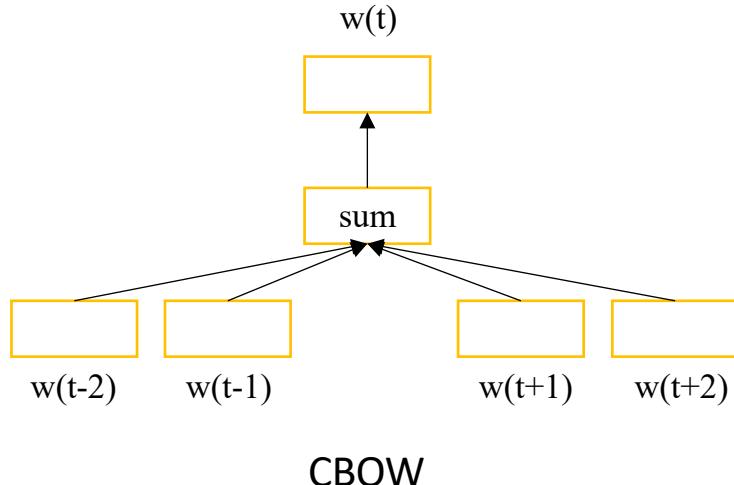


# Today: Neural Network Models on 1D Grid / Language Data



# Word2vec: CBOW / SkipGram (Basic Word2Vec)

- Distributed representations of words and phrases and their compositionality (NIPS 2013, Mikolov et al.)
- CBOW
  - predict the input tokens based on context tokens
- SkipGram
  - predict context tokens based on input tokens



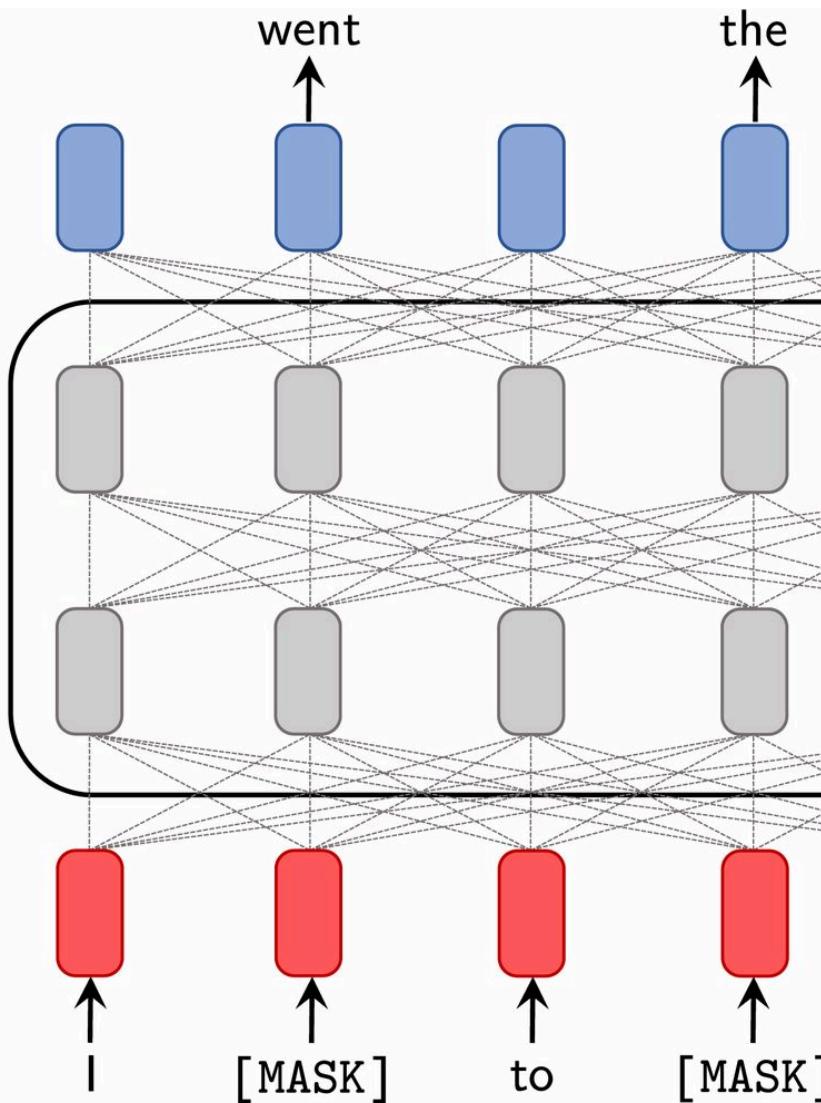
## Predictions

Contextualized  
embeddings

Transformer  
Network

Word  
embeddings

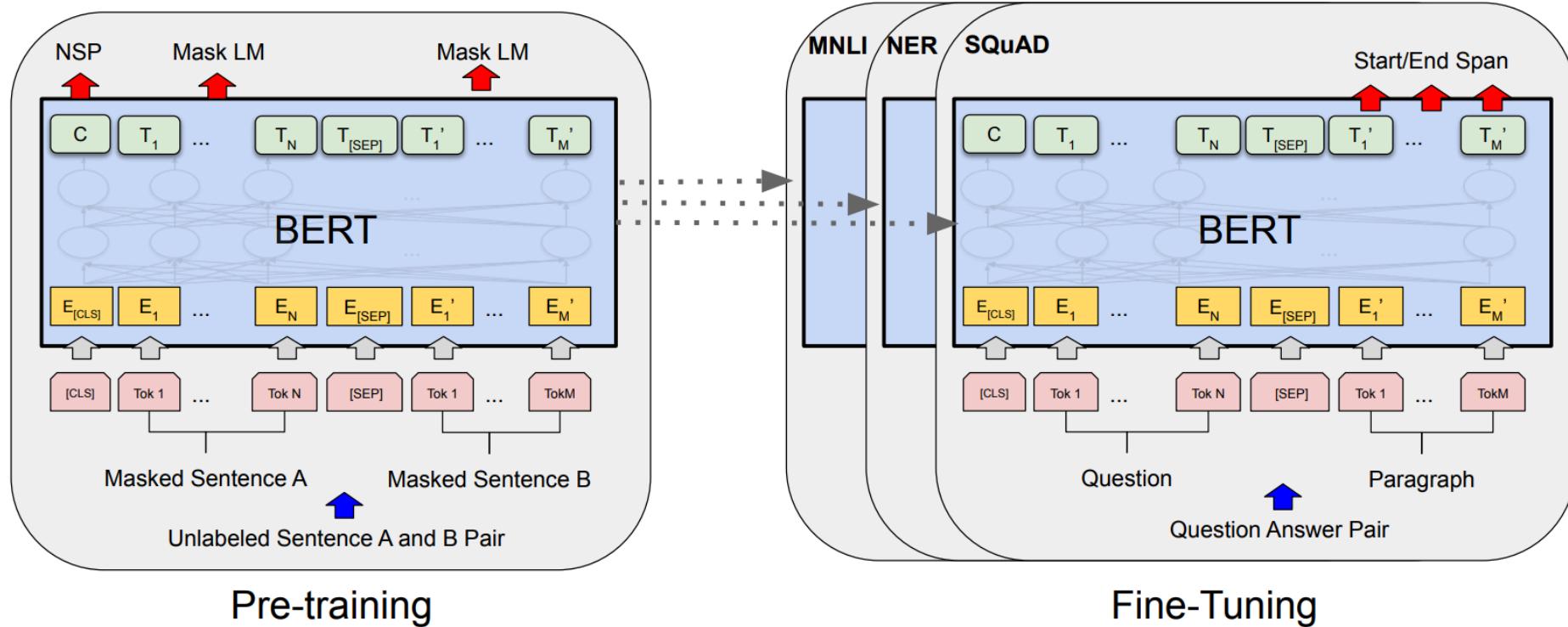
Masked Input



BERT is  
trained  
just like a  
skip-gram  
model.

# BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding (NAACL 2019, Devlin et al.)

- Denoising Auto Encoder
- [MASK]: a unique token introduced in the training process to mask some tokens
- Predict masked tokens based on their context information,
- Pre-train and fine-tune
- Intuition: representation should be robust to the introduction of noise
  - Masked Language Model (MLM)



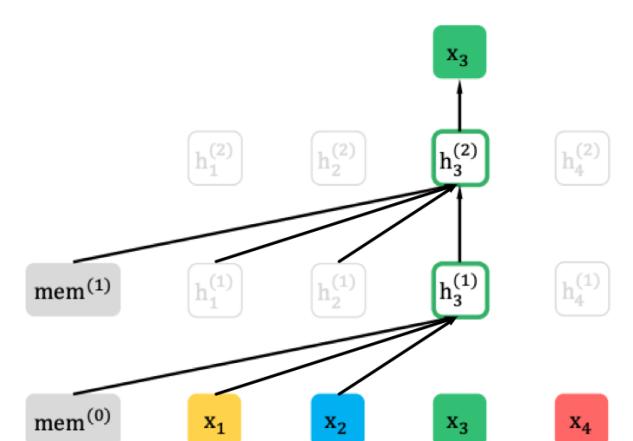
# XLNet (Generalized autoregressive pretraining for language understanding (NIPS 2019, Yang et al.)

- **XLNetPLM: Permutation Language Model**

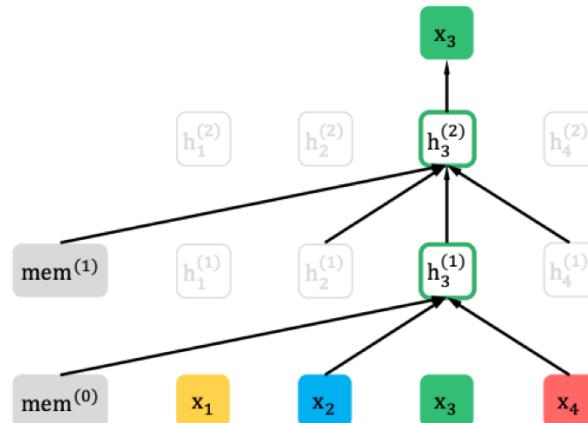
- Based-on AutoRegressive model
- learning bidirectional contexts by permutation
- reparameterization and a special two-stream self-attention for target-aware prediction

- **Transformer-XL: Extra Long Transformer**

- segment recurrence mechanism + relative encoding scheme



Original Sequence:  $[x_1, x_2, x_3, x_4]$



Permutated Sequence:  $[x_2, x_4, x_3, x_1]$

# ALBERT: A lite BERT (2019, Lan et al.)

- proposes Sentence Order Prediction (SOP) task to replace Next Sentence Prediction (NSP)
- in NSP, the negative next sentence is sampled from other passages that may have different topics with the current one, turning the NSP into a far easier topic model problem.
- in SOP, two sentences that exchange their position are regarded as a negative sample, making the model concentrate on the coherence of the semantic meaning.

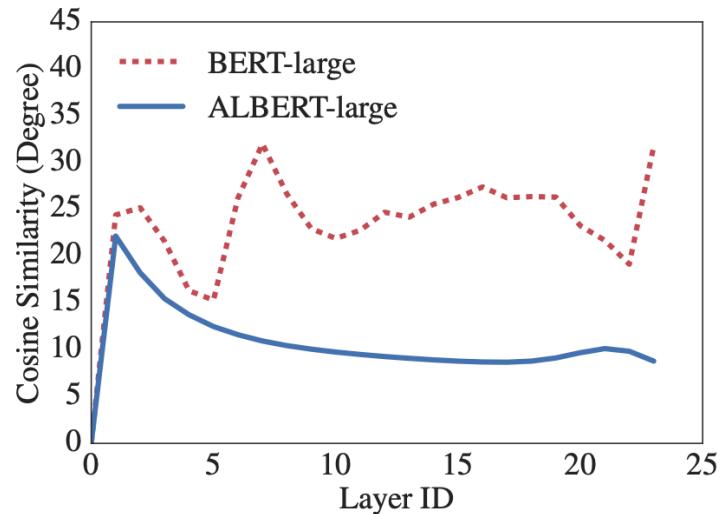
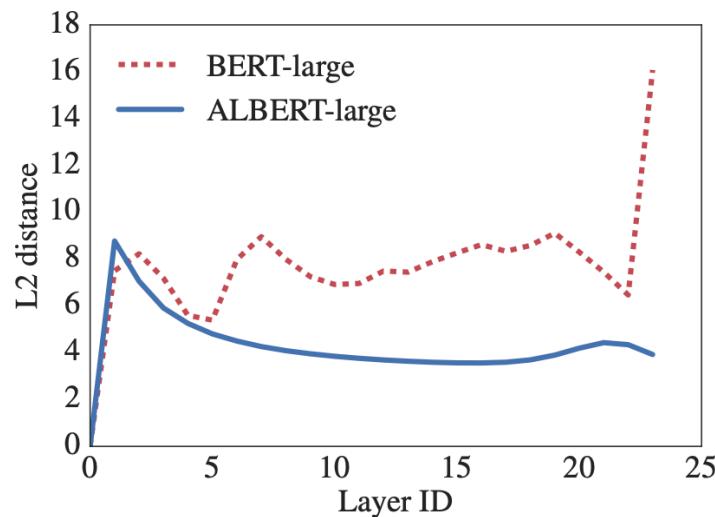
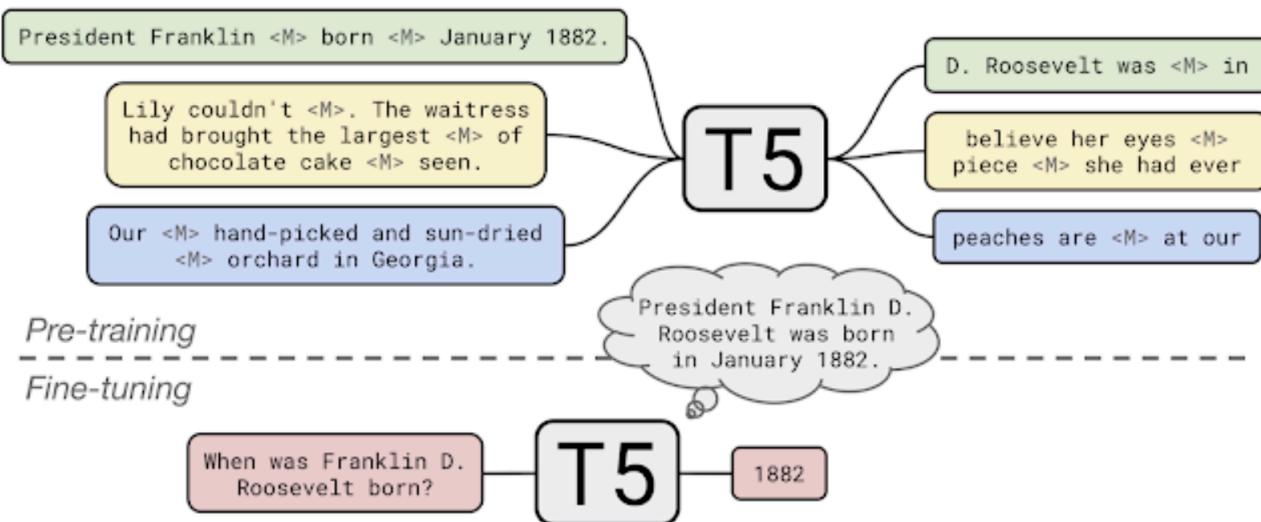


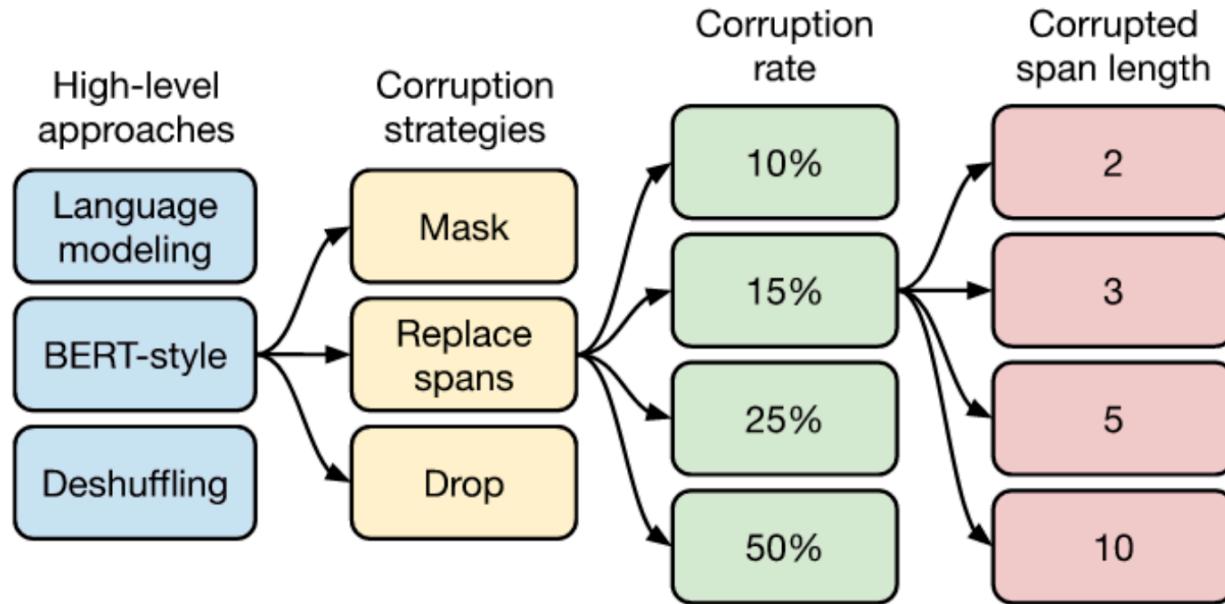
Figure 1: The L2 distances and cosine similarity (in terms of degree) of the input and output embedding of each layer for BERT-large and ALBERT-large.

Based: Dr. Yangqiu Song's slides

# T5: Even more noise

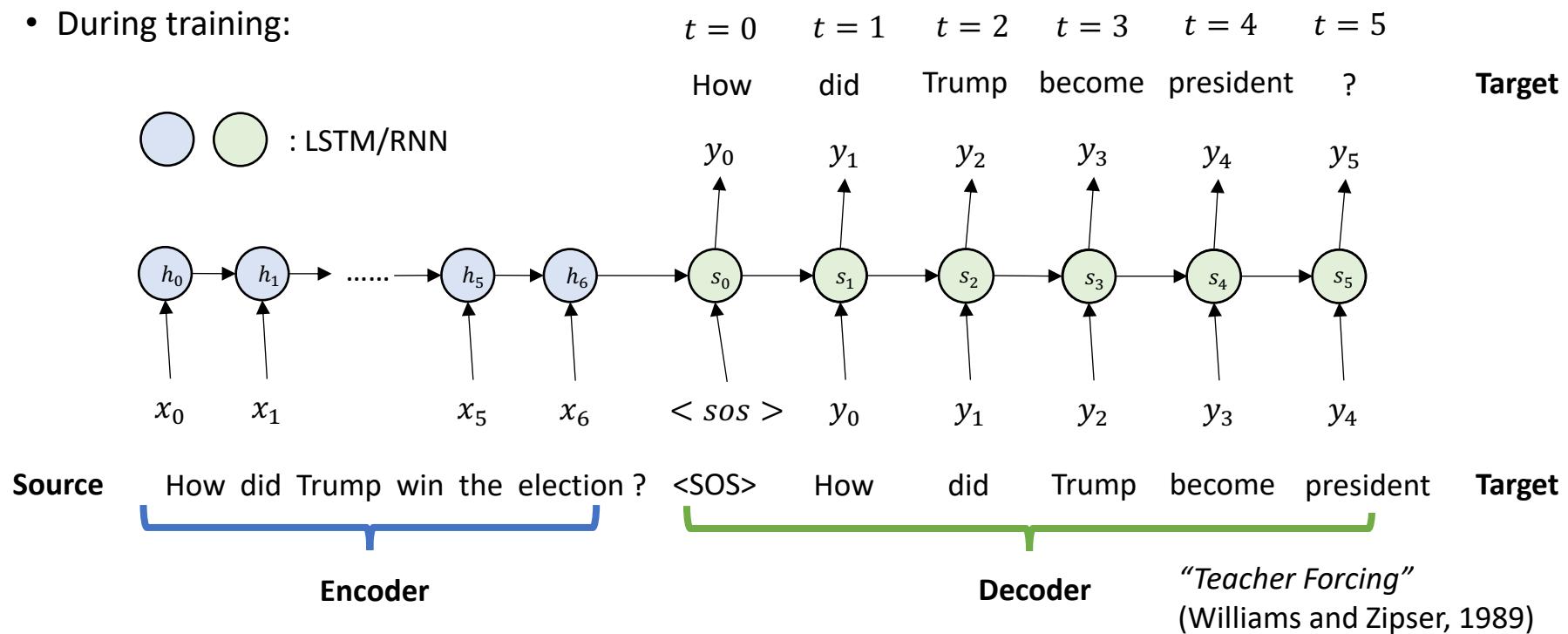


During pre-training, T5 learns to fill in dropped-out spans of text (denoted by <M>) from documents in C4. To apply T5 to closed-book question answer, we fine-tuned it to answer questions without inputting any additional information or context. This forces T5 to answer questions based on “knowledge” that it internalized during pre-training.

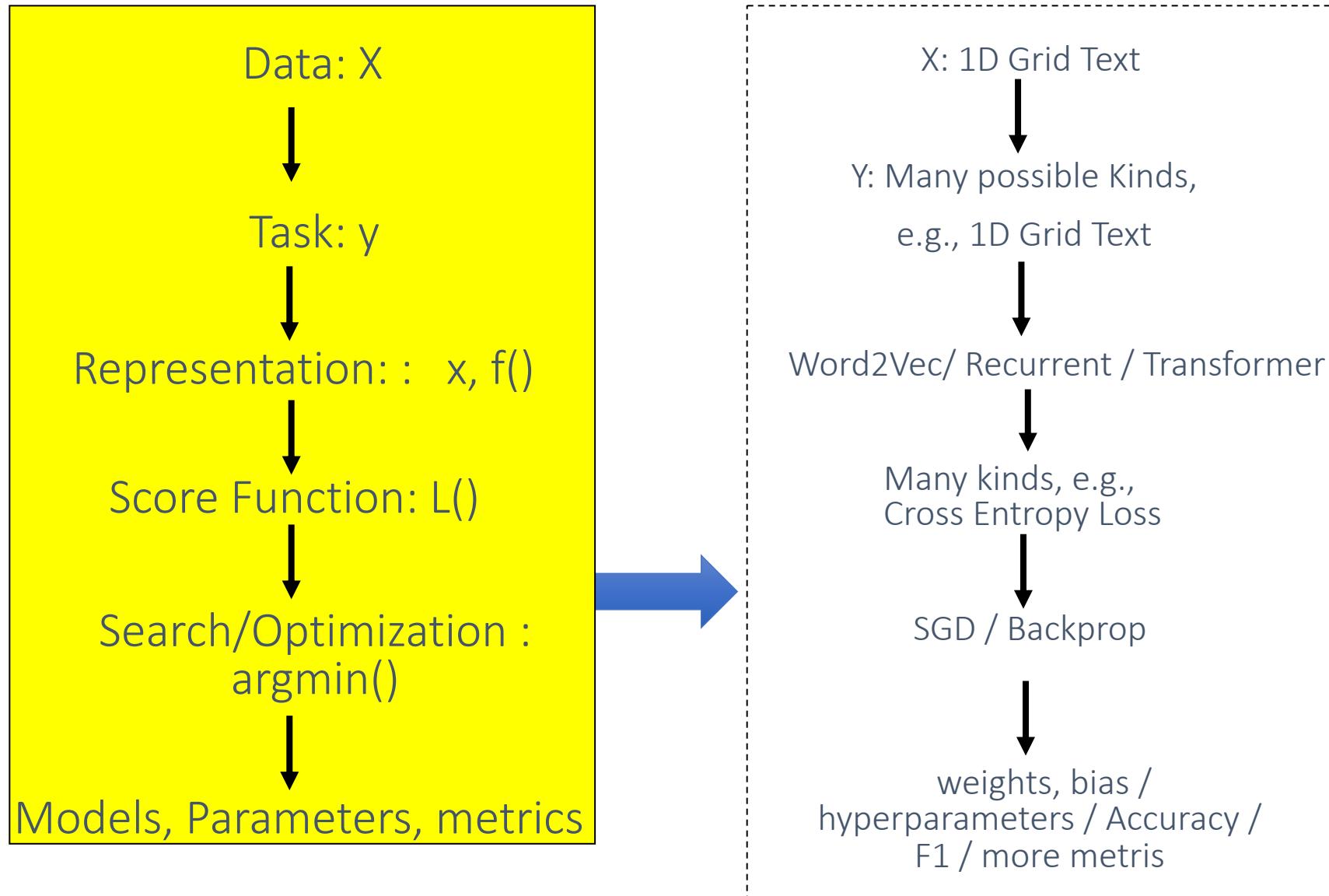


# Teaching Forcing: for training general Seq2Seq

- During training:



# Today Recap: Neural Network Models on 1D Grid / Language Data



# References

- Dr. Yann Lecun's deep learning tutorials
- Dr. Li Deng's ICML 2014 Deep Learning Tutorial
- Dr. Kai Yu's deep learning tutorial
- Dr. Rob Fergus' deep learning tutorial
- Prof. Nando de Freitas' slides
- Olivier Grisel's talk at Paris Data Geeks / Open World Forum
- Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.
- Dr. Hung-yi Lee's CNN slides
- NIPS 2017 DL Trend Tutorial