

# EN. 601.647/667

# INTRODUCTION TO HLT

# DEEP LEARNING

---

KENTON MURRAY

9/16/2021

# EN. 601.647/667

# INTRODUCTION TO HLT

# DEEP LEARNING

---

KENTON MURRAY

9/16/2021

Some slides were inspired by:  
Kevin Duh, Shinji Watanabe, Marine Carpaut,  
Graham Neubig, Jacob Eisenstein

# AGENDA

---

- Impact of Deep Learning on HLT
- Intro to Deep Learning
- What changed? Why recently?
- Deep Learning in HLT
- PyTorch

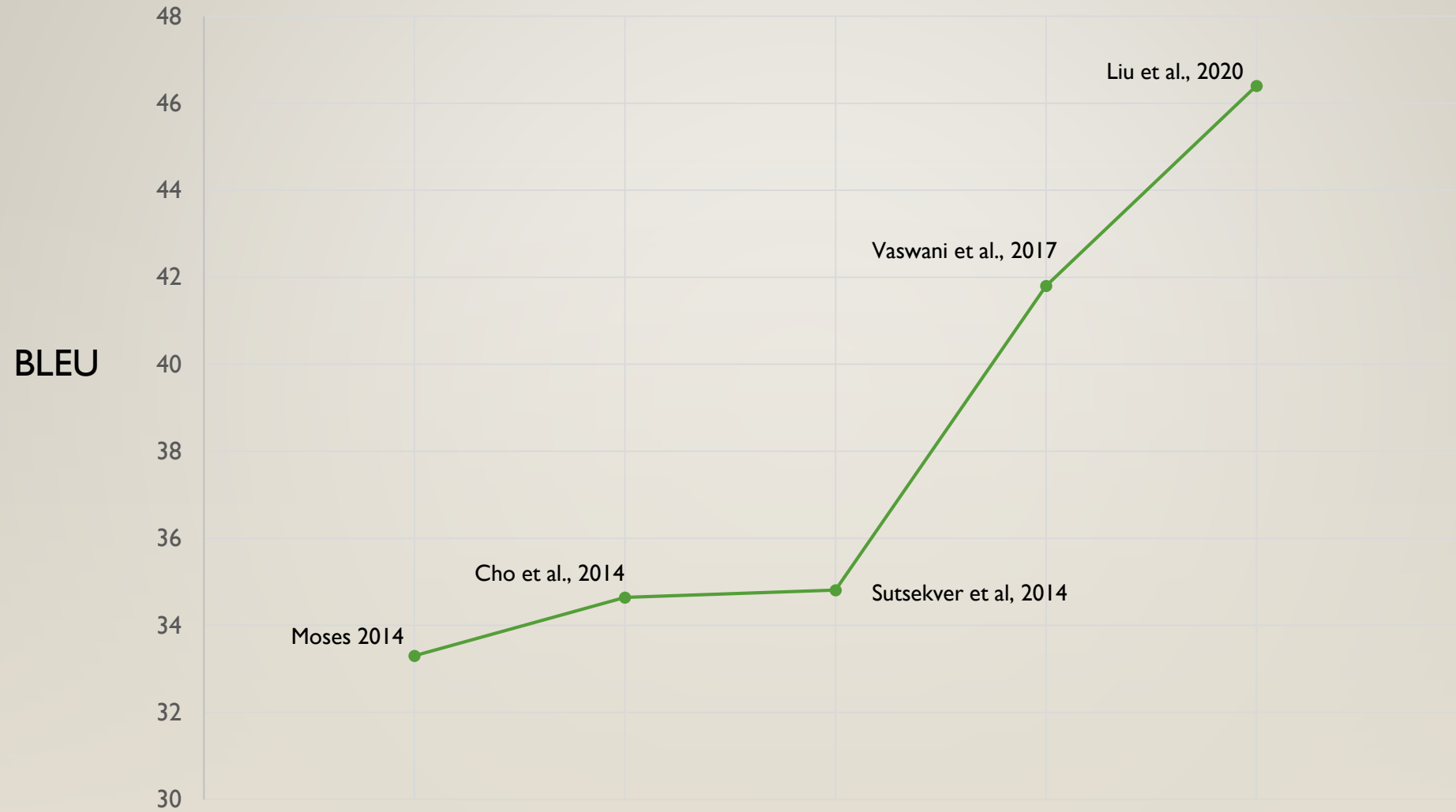
# QUICK ASIDE

---

- BLEU Scores
- Modified n-gram precision metric for Machine Translation
- 0 – 100.0 (higher is better)
- Reviewers generally like  $\sim +1.0$  gain over a baseline

# IMPACT OF DEEP LEARNING

WMT '14 En-Fr BLEU Score




# CATS!

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE


CORONAVIRUS HOW TO GET A COVID VACCINE BEST FACE MASKS COVID-19 FAQ NEWSLETTER LATEST NEWS

## Google's Artificial Brain Learns to Find Cat Videos

When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.



WATCH



Neuroscientist Explains One Concept in 5 Levels of Difficulty

Get WIRED

# ACL 2014 BEST PAPER

## Fast and Robust Neural Network Joint Models for Statistical Machine Translation

Jacob Devlin, Rabih Zbib, Zhongqiang Huang,  
Thomas Lamar, Richard Schwartz, and John Makhoul  
Raytheon BBN Technologies, 10 Moulton St, Cambridge, MA 02138, USA  
{jdevlin, rzbib, zhuang, tlamar, schwartz, makhoul}@bbn.com

### Abstract

Recent work has shown success in using neural network language models (NNLMs) as features in MT systems. Here, we present a novel formulation for a neural network *joint* model (NNJM), which augments the NNLM with a source context window. Our model is purely lexicalized and can be integrated into any MT decoder. We also present several variations of the NNJM which provide significant additive improvements.

Although the model is quite simple, it

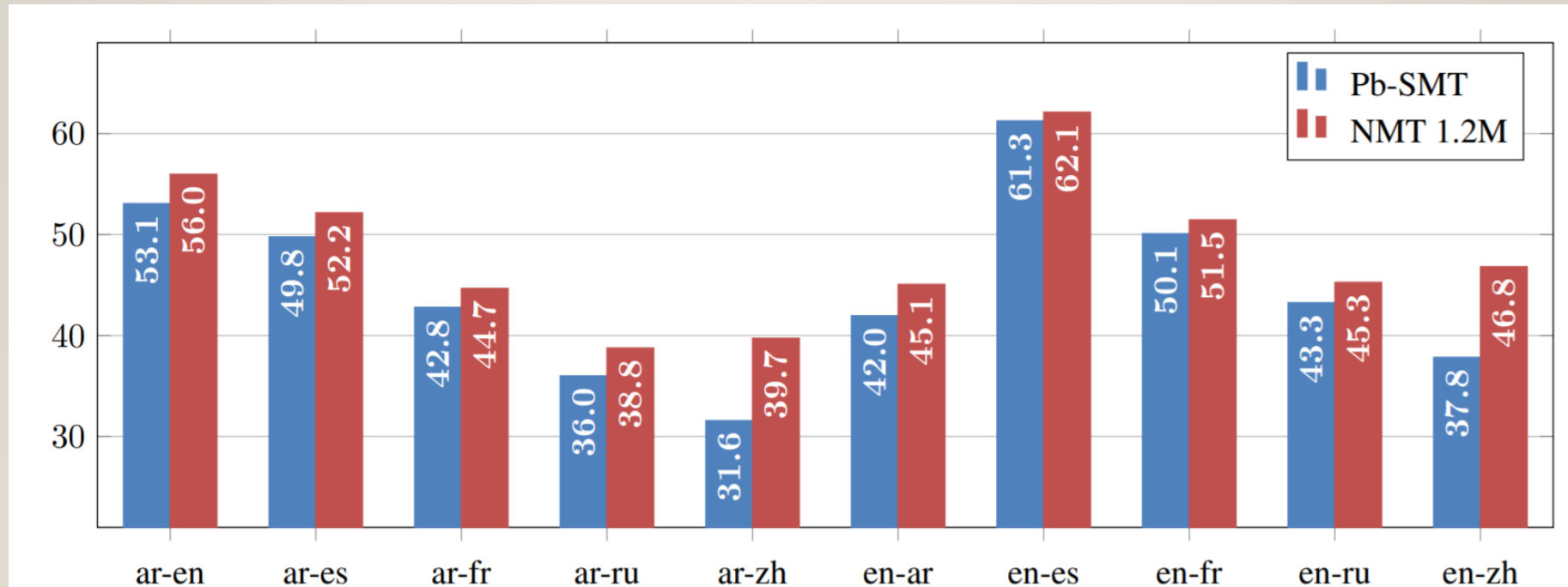
### 1 Introduction

In recent years, neural network models have become increasingly popular in NLP. Initially, these models were primarily used to create  $n$ -gram neural network language models (NNLMs) for speech recognition and machine translation (Bengio et al., 2003; Schwenk, 2010). They have since been extended to translation modeling, parsing, and many other NLP tasks.

In this paper we use a basic neural network architecture and a lexicalized probability model to create a powerful MT decoding feature. Specifically, we introduce a novel formulation for a neu-

# IS NEURAL MACHINE TRANSLATION READY FOR DEPLOYMENT?

- Junczys-Dowmunt et al., 2016

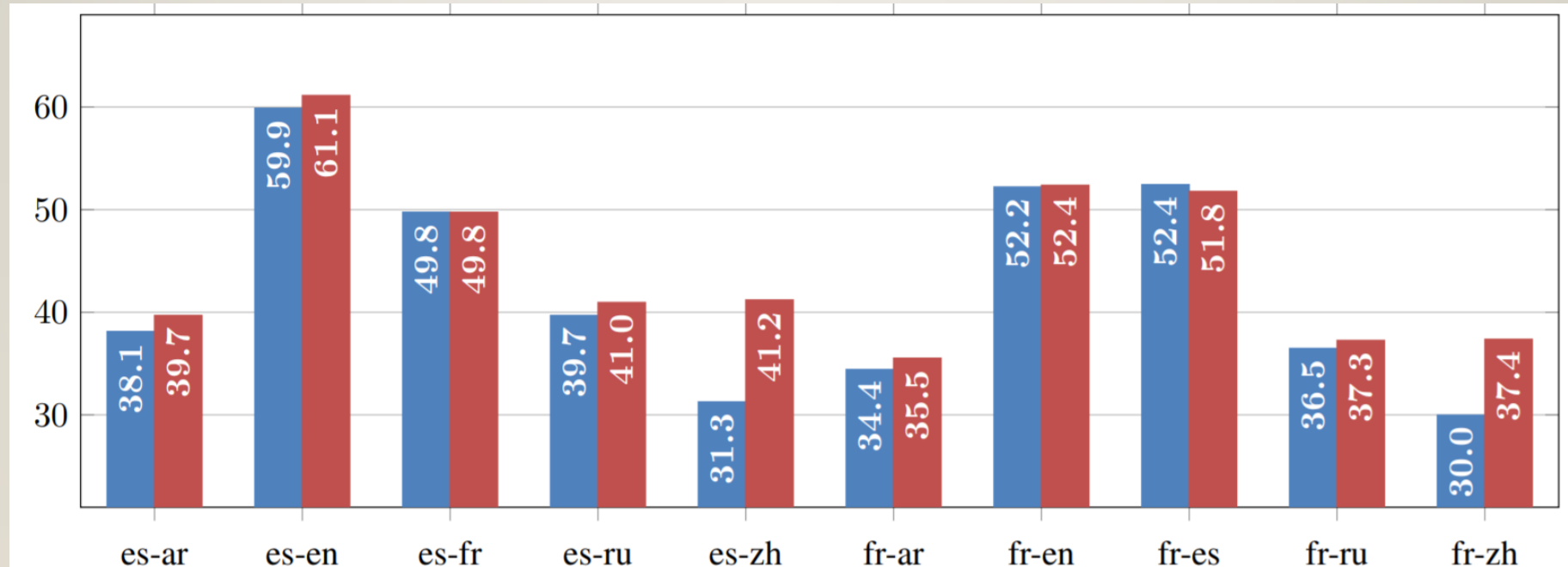




# IS NEURAL MACHINE TRANSLATION READY FOR DEPLOYMENT?

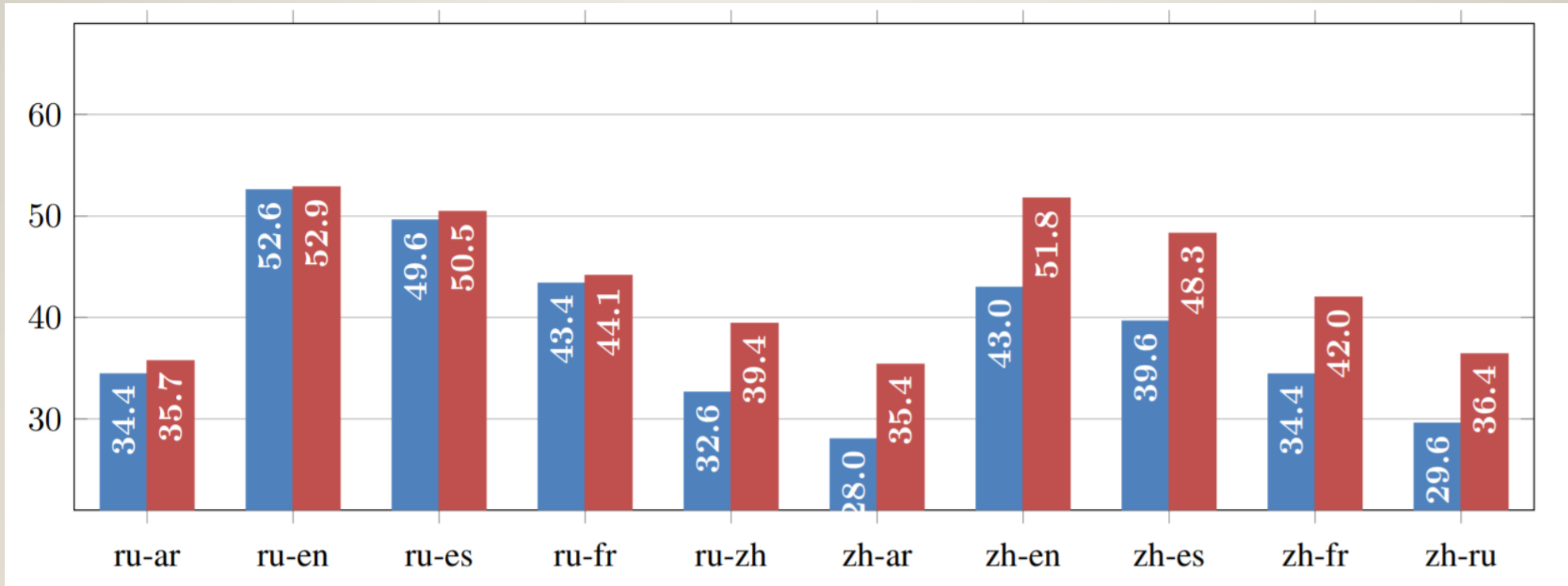
---

- Junczys-Dowmunt et al., 2016



# IS NEURAL MACHINE TRANSLATION READY FOR DEPLOYMENT?

- Junczys-Dowmunt et al., 2016



# DEEP LEARNING BACKGROUND

---

- Deep Learning is Neural Networks ....That are Deep!

# SUPERVISED LEARNING

---

- Model:  $\theta$
- Input:  $X$
- Output:  $Y$
- $P_{\theta}(Y|X)$

# SUPERVISED LEARNING (LANGUAGE ID)

---

- Model:  $\theta$
- Input:  $X$  (Hello my name is)
- Output:  $Y$  (English)
- $P_{\theta}(Y|X)$

# SUPERVISED LEARNING (LANGUAGE ID)

---

- Model:  $\theta$
- Input:  $X$  (Hola, mi nombre es)
- Output:  $Y$  (Spanish)
- $P_{\theta}(Y|X)$

# SUPERVISED LEARNING (LANGUAGE ID)

---

- Model:  $\theta$
- Input:  $X$  (Salut, je m'appelle)
- Output:  $Y$  (French)
- $P_{\theta}(Y|X)$

# SUPERVISED LEARNING (LANGUAGE ID)

---

- Model:  $\theta$
- Input:  $X$  (Imanalla, nuqap suti Murray kan)
- Output:  $Y$  (?)
- $P_{\theta}(Y|X)$



# SUPERVISED LEARNING (SPEECH RECOGNITION)

---

- Model:  $\theta$
- Input:  $X$  (🔊)
- Output:  $Y$  (?)
- $P_{\theta}(Y|X)$

# SUPERVISED LEARNING (SPEECH RECOGNITION)

---

- Model:  $\theta$
- Input: X (🔊)
- Output: Y (Hello World)
- $P_{\theta}(Y|X)$

# SUPERVISED LEARNING (MACHINE TRANSLATION)

---

- Model:  $\theta$
- Input:  $X$  ( مرحبا بالعالم )
- Output:  $Y$  (?)
- $P_{\theta}(Y|X)$

# SUPERVISED LEARNING (MACHINE TRANSLATION)

---

- Model:  $\theta$
- Input: X ( مرحبا بالعالم )
- Output: Y (Hello World)
- $P_{\theta}(Y|X)$

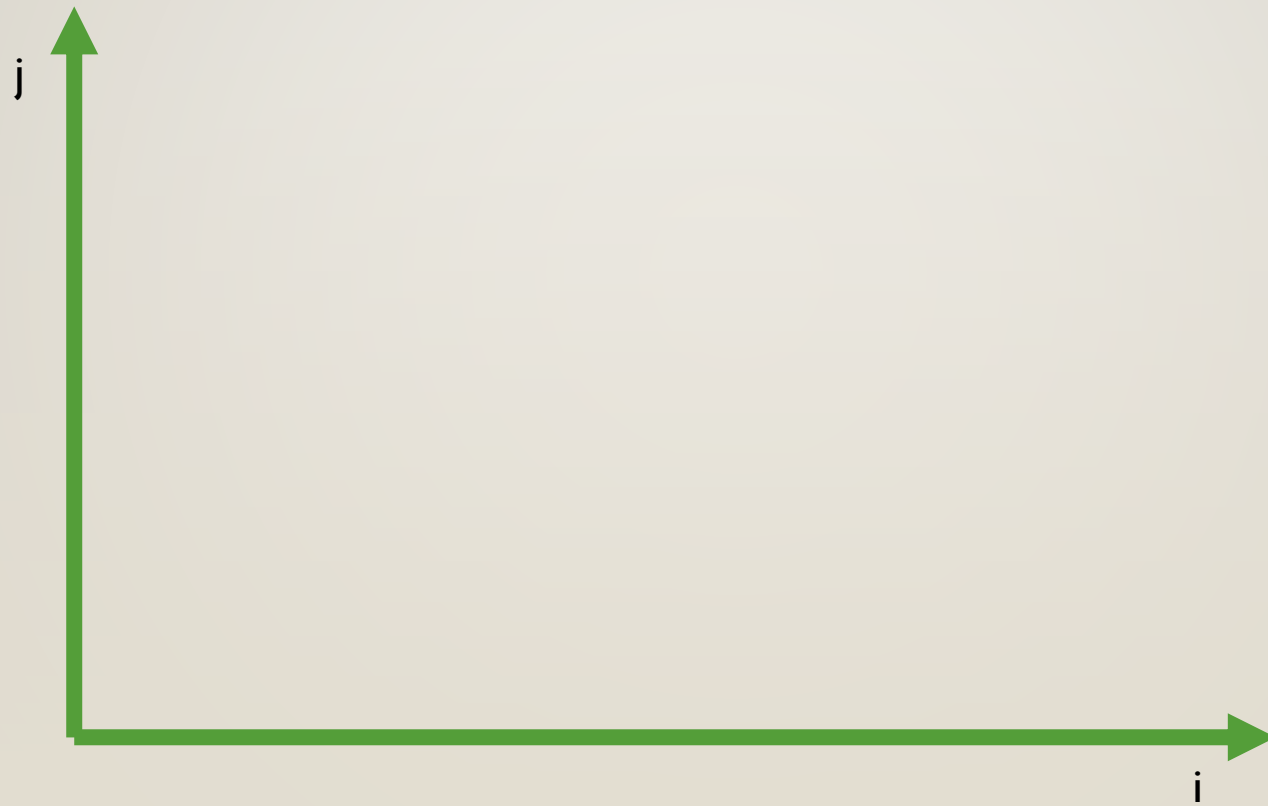
# CLASSIFICATION

---

- One of the first tasks public thinks of
- Sentiment Analysis
- Speaker/Writer ID
- Language ID
- Phoneme Recognition

# BINARY CLASSIFICATION

---



# BINARY CLASSIFICATION

---

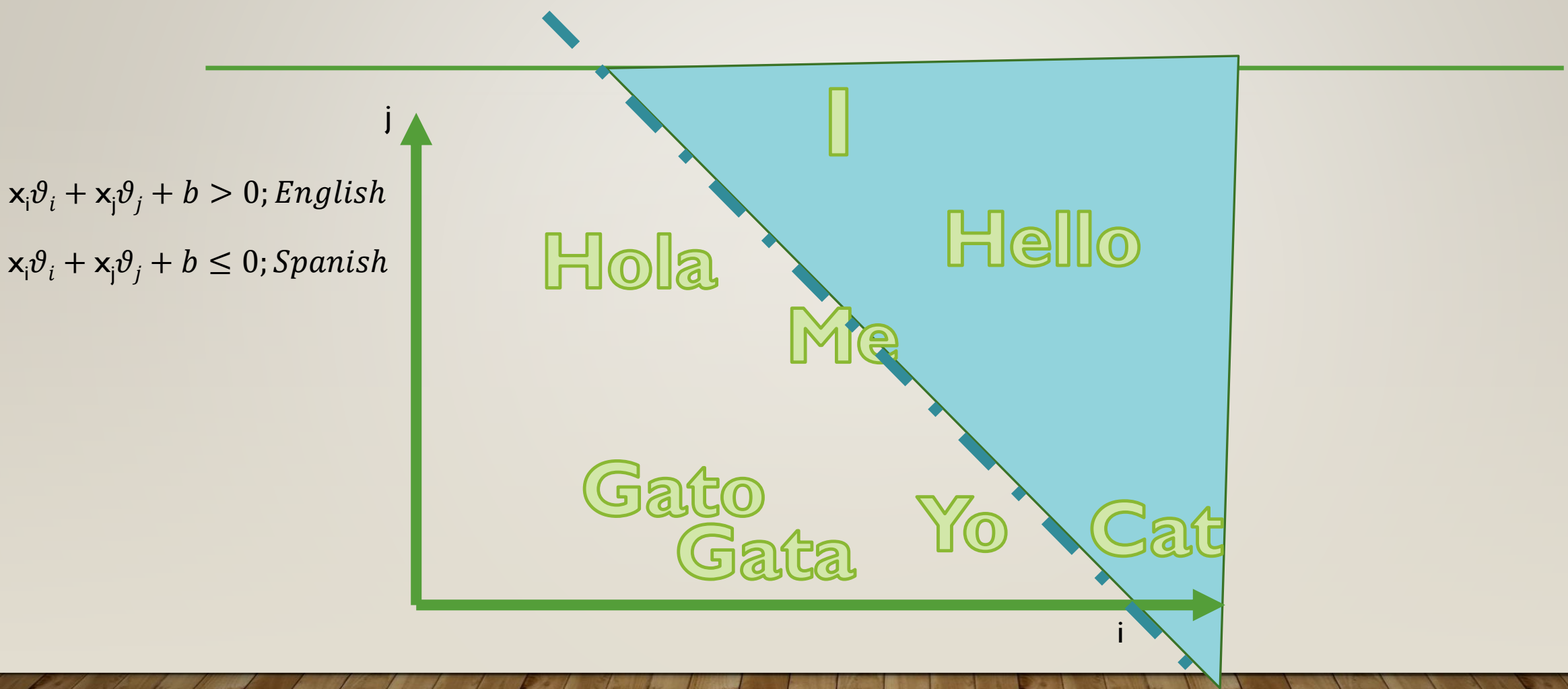


# BINARY CLASSIFICATION

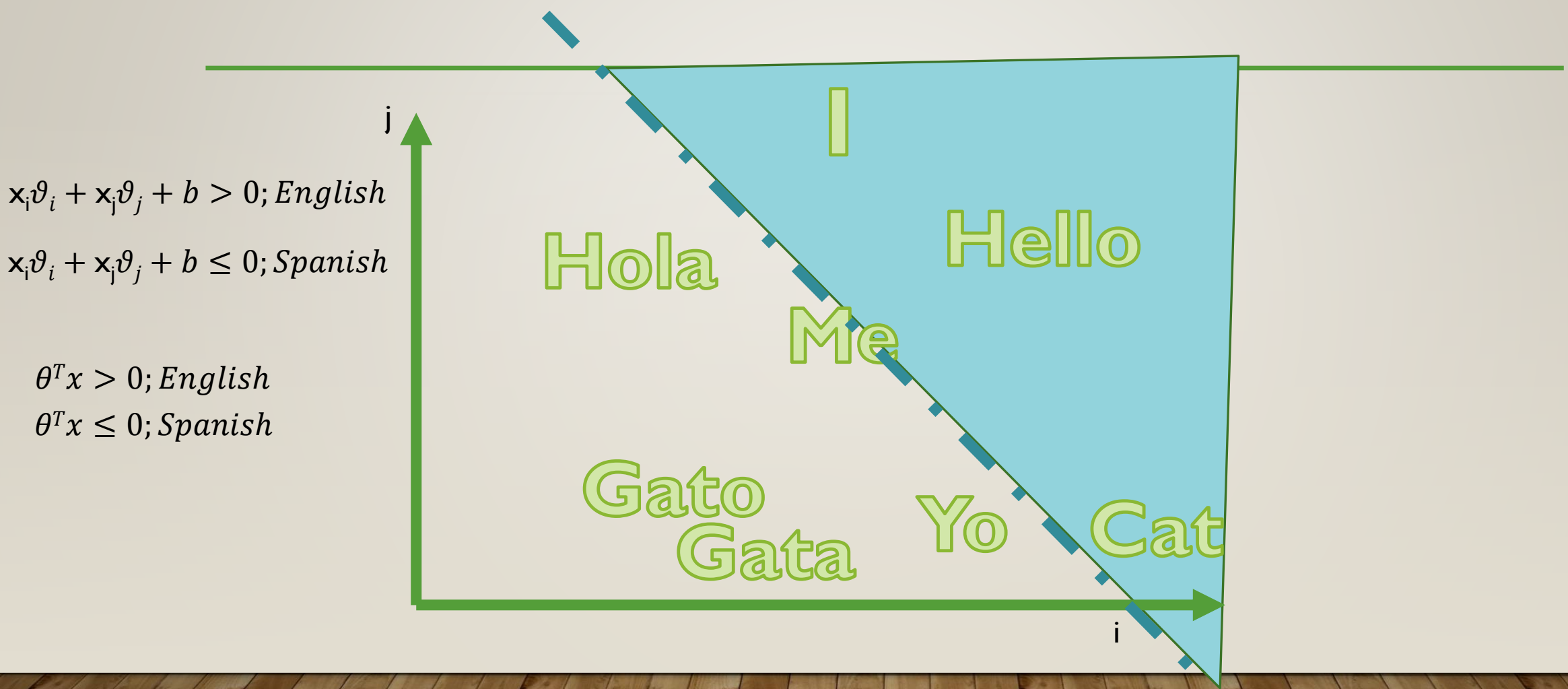




# LINEAR MODELS



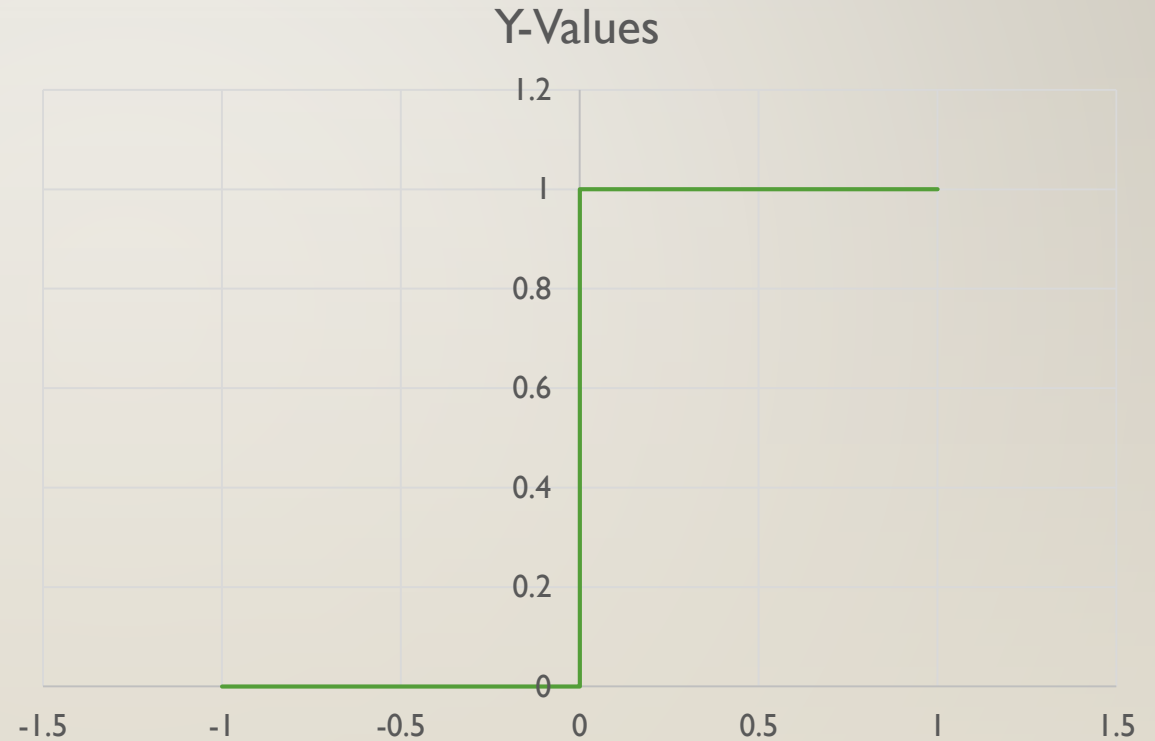
# LINEAR MODELS



# PERCEPTRON

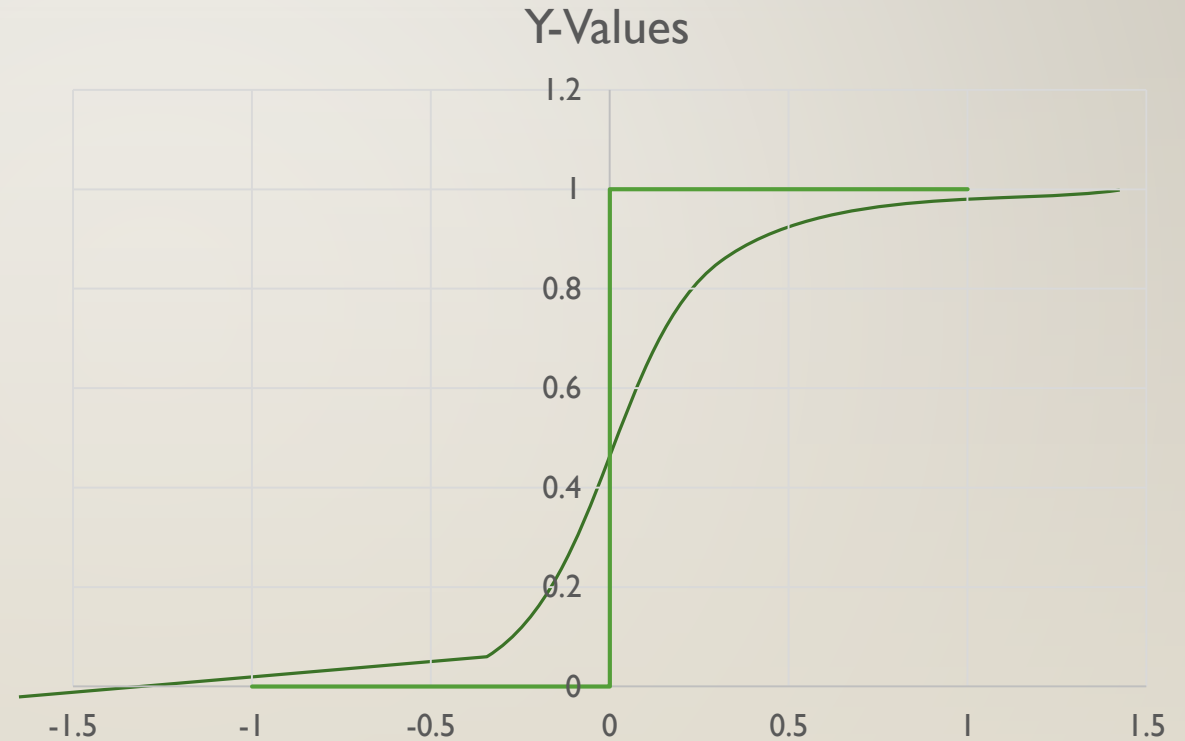
---

- Make it a probability
- $P(y=\text{English}|x) = 1.0$  if  $\theta^T x > 0$
- $P(y=\text{English}|x) = 0.0$  if  $\theta^T x \leq 0$



# LOGISTIC REGRESSION

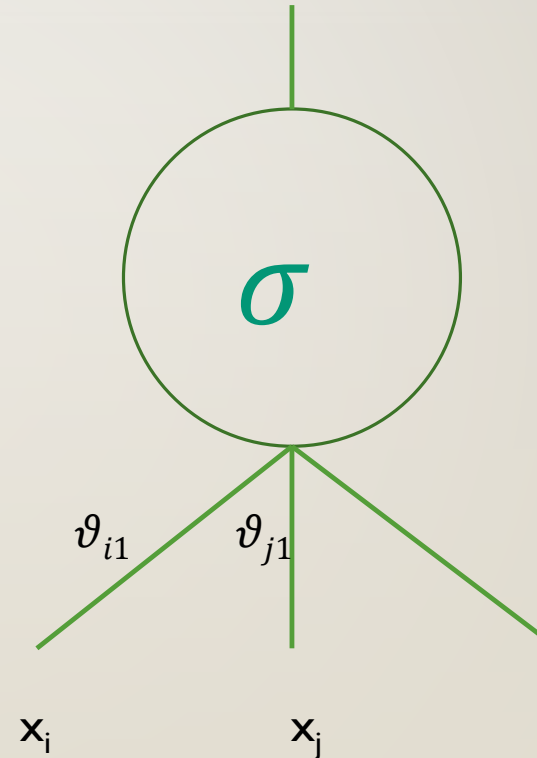
- Make it a probability
- $P(y=\text{English}|x) = \sigma(\theta^T x)$
- $P(y=\text{English}|x) = \frac{1}{1+e^{\theta^T x}}$
- Softer
- Differentiable



# LOGISTIC REGRESSION

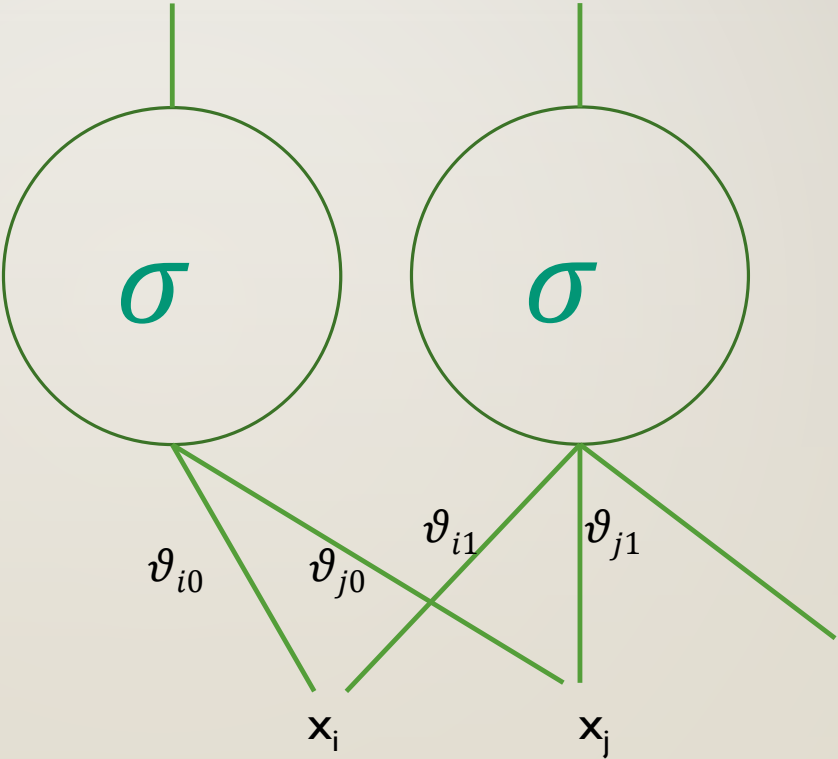
---

- Make it a probability
- $P(y=\text{English}|\mathbf{x}) = \sigma(\boldsymbol{\theta}^T \mathbf{x})$
- $P(y=\text{English}|\mathbf{x}) = \frac{1}{1+e^{\boldsymbol{\theta}^T \mathbf{x}}}$
- Softer
- Differentiable



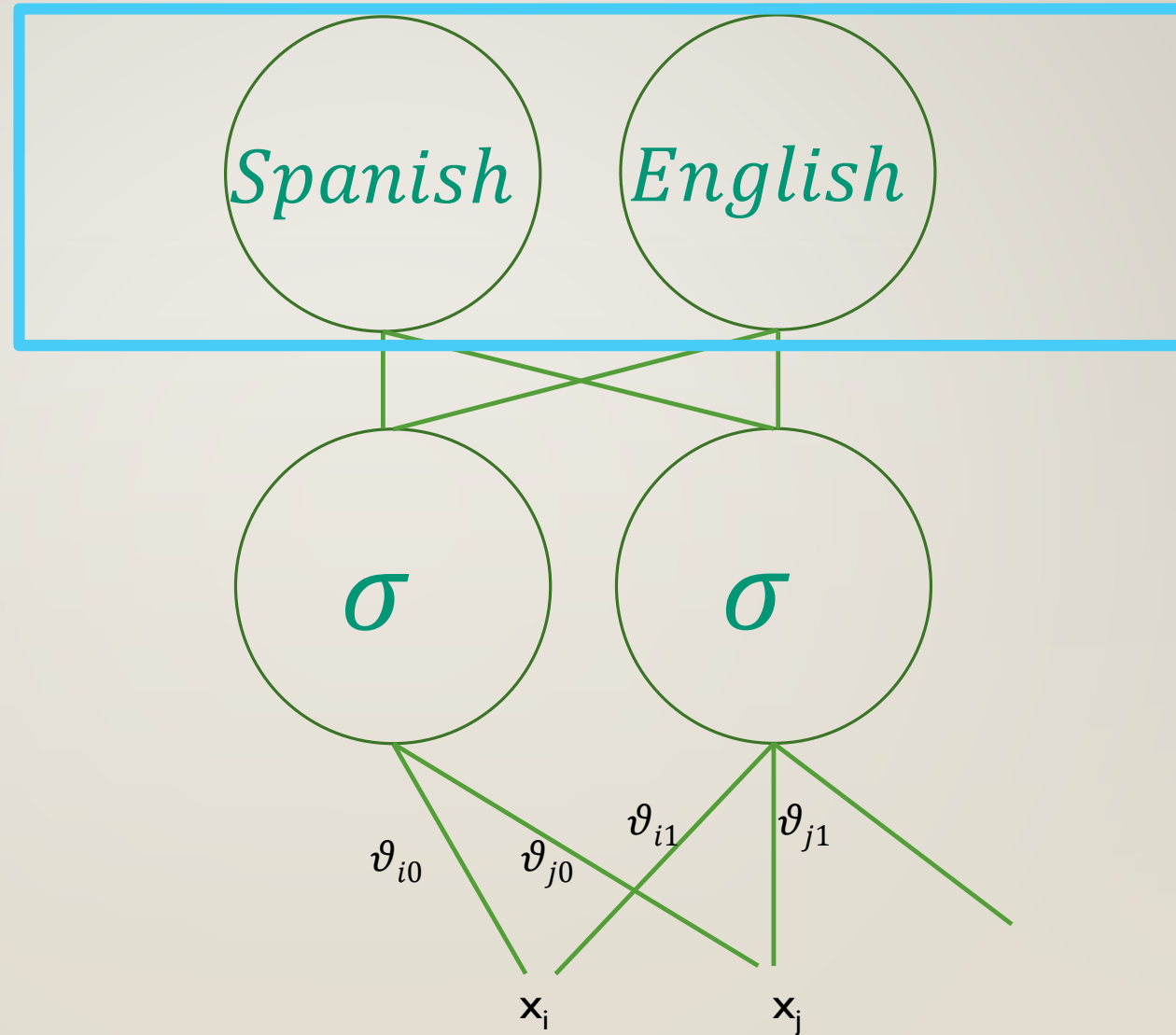
# 2 NEURON FEED-FORWARD

---



## 2 Neuron Feed-Forward

Softmax



# SOFTMAX

---

- Make the output a probability distribution

- $$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=0}^Z e^{z_j}}$$

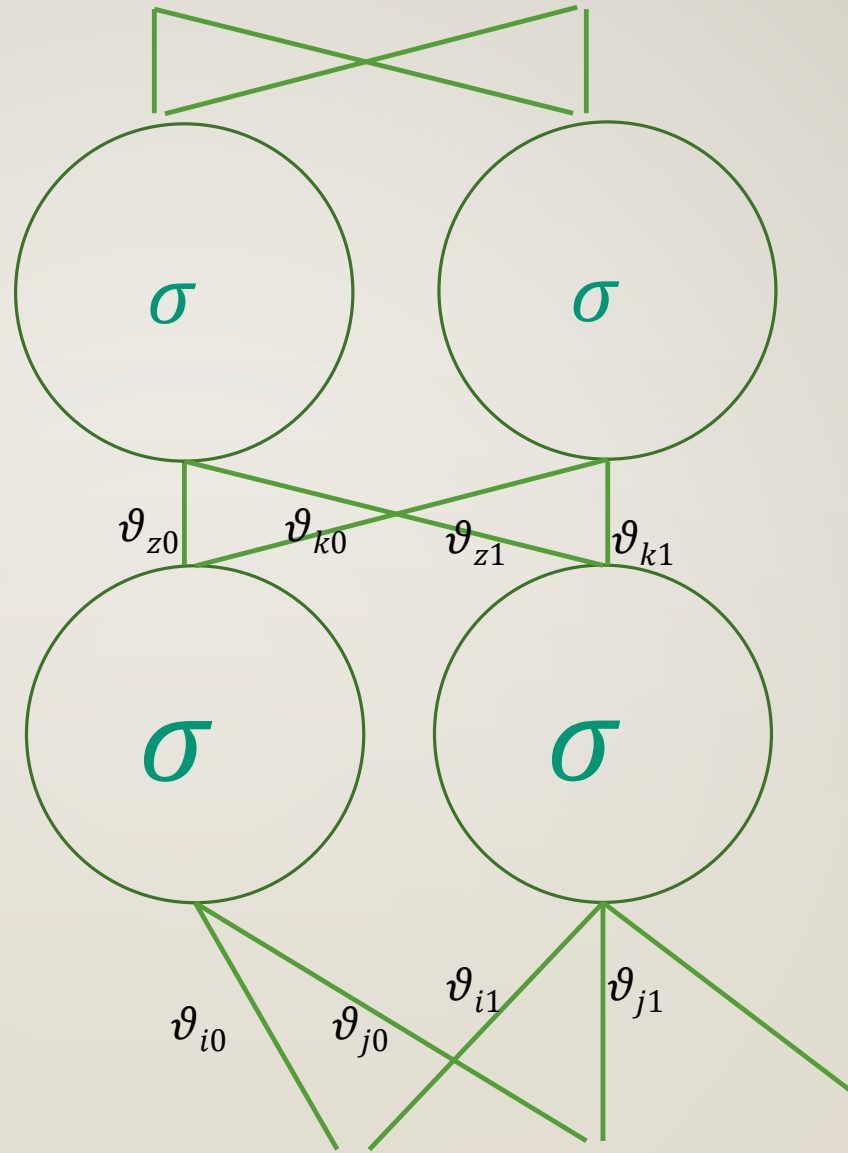


# SOFTMAX

---

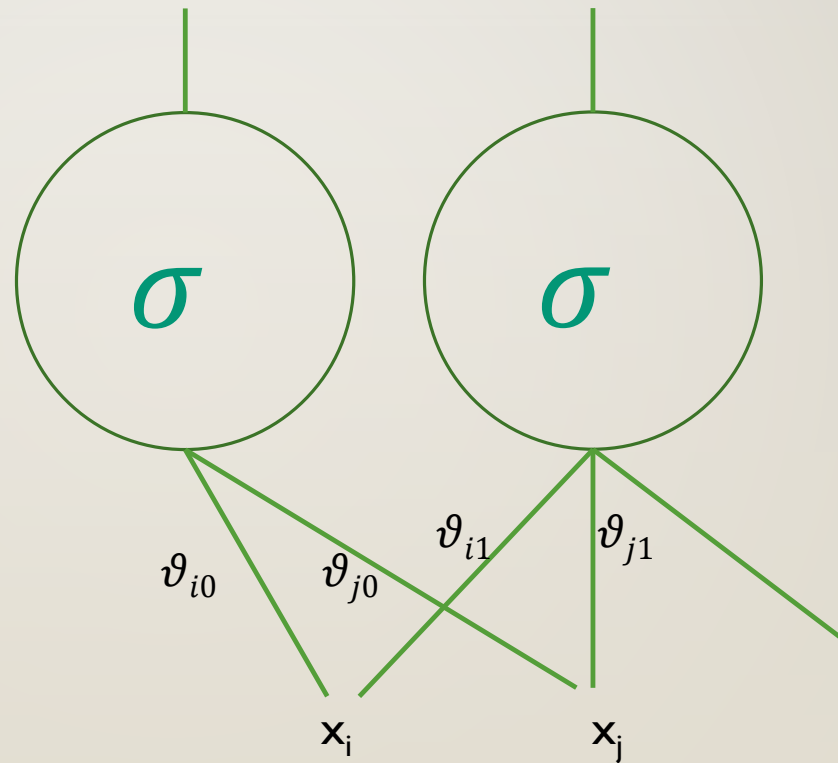
- Make the output a probability distribution
- $P(y=\text{English}|\mathbf{x}) = \frac{1}{1+e^{\theta^T \mathbf{x}}}$
- $\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=0}^Z e^{z_j}}$
- Training: Differentiable through it
- Testing: Take the Max

DEEP



# OTHER NON-LINEARITIES

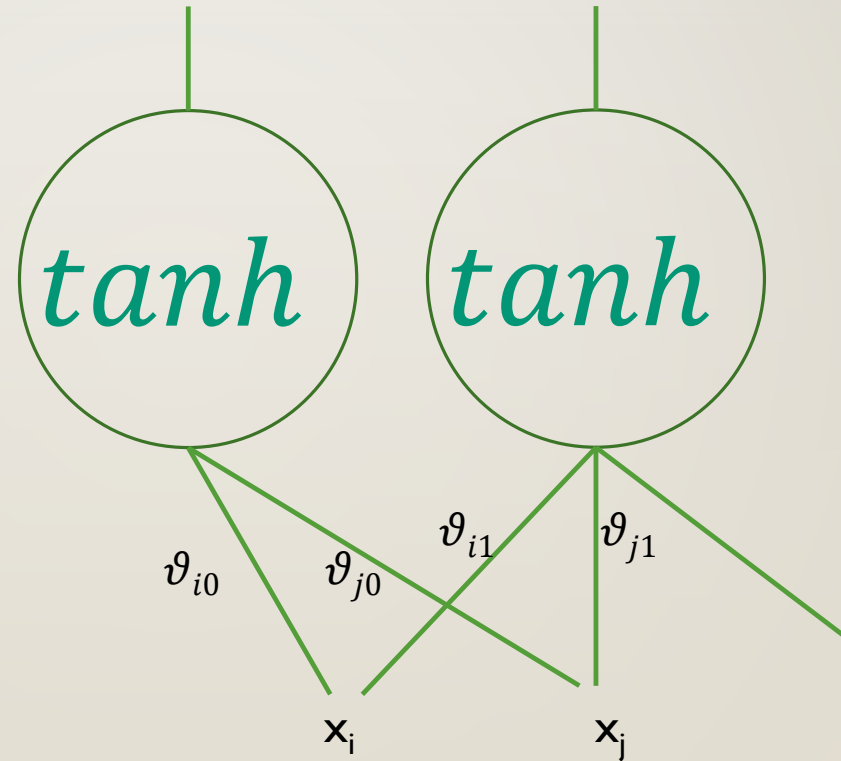
---



# OTHER NON-LINEARITIES

---

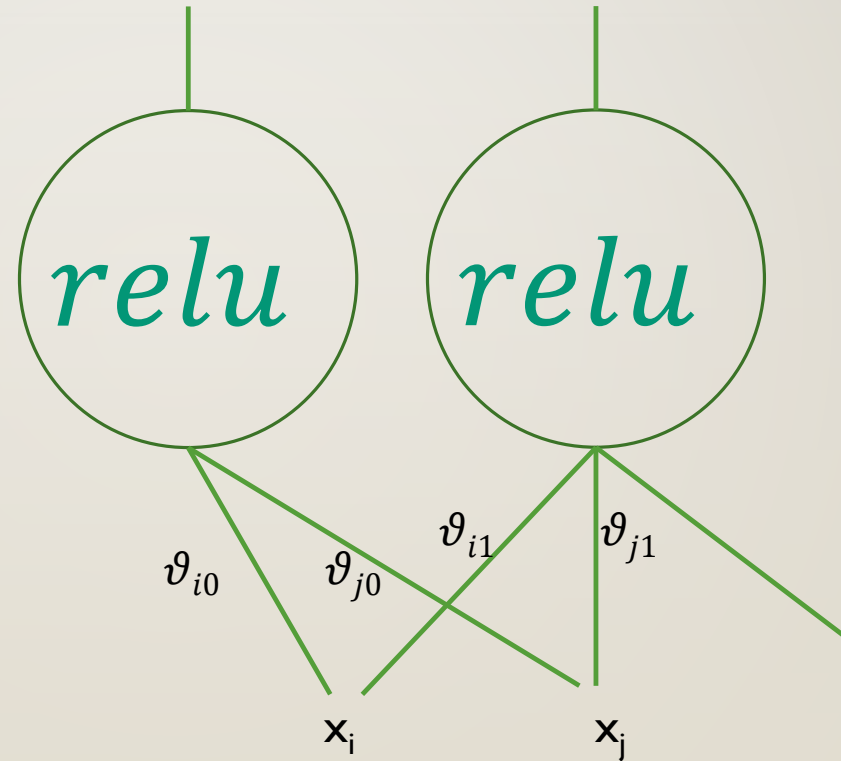
- A lot like sigmoid
- Range:  $(-1.0, 1.0)$



# OTHER NON-LINEARITIES

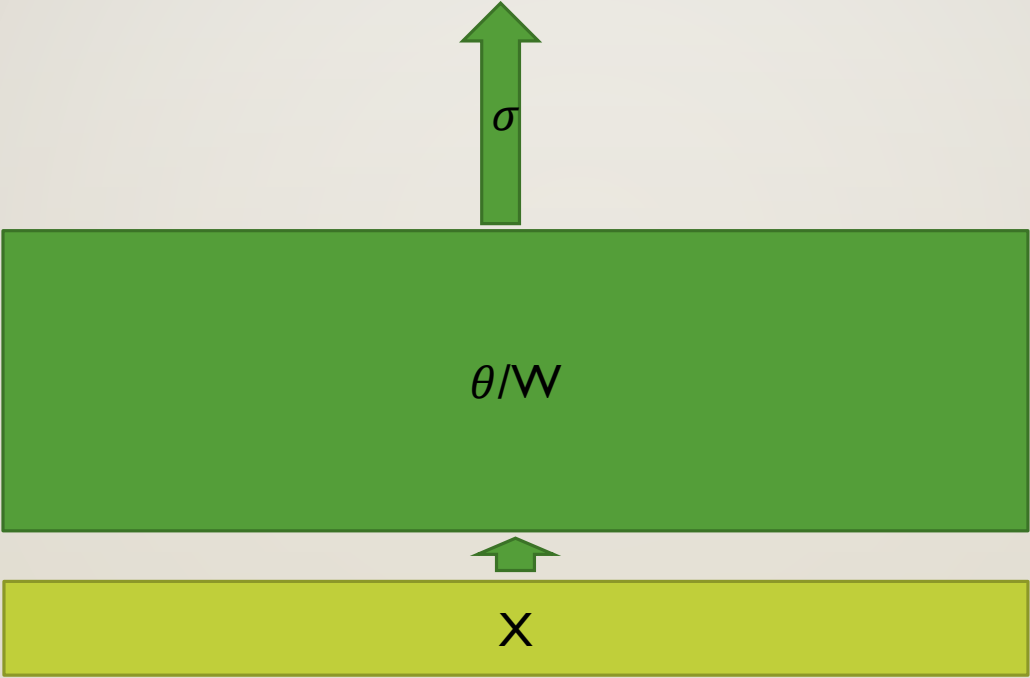
---

$\text{relu}(x) = 0; \text{ if } x < 0$   
 $\text{relu}(x) = x; \text{ if } x \geq 0$



# TENSORS!

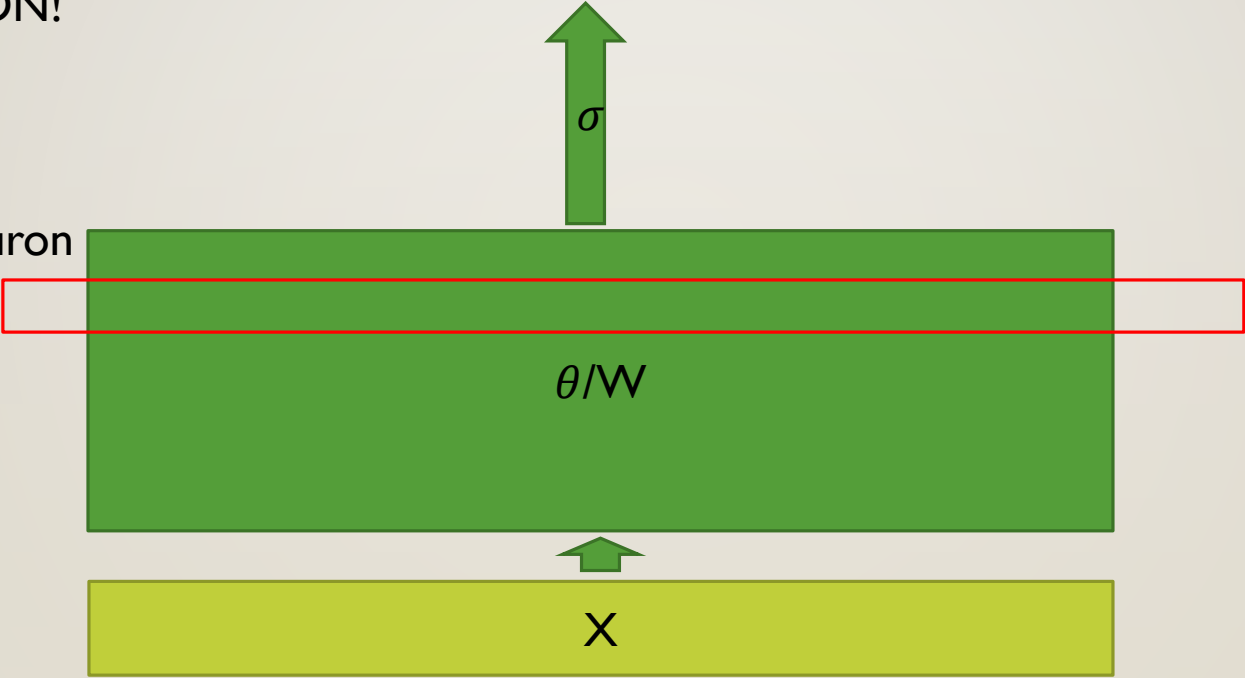
---



# TENSORS!

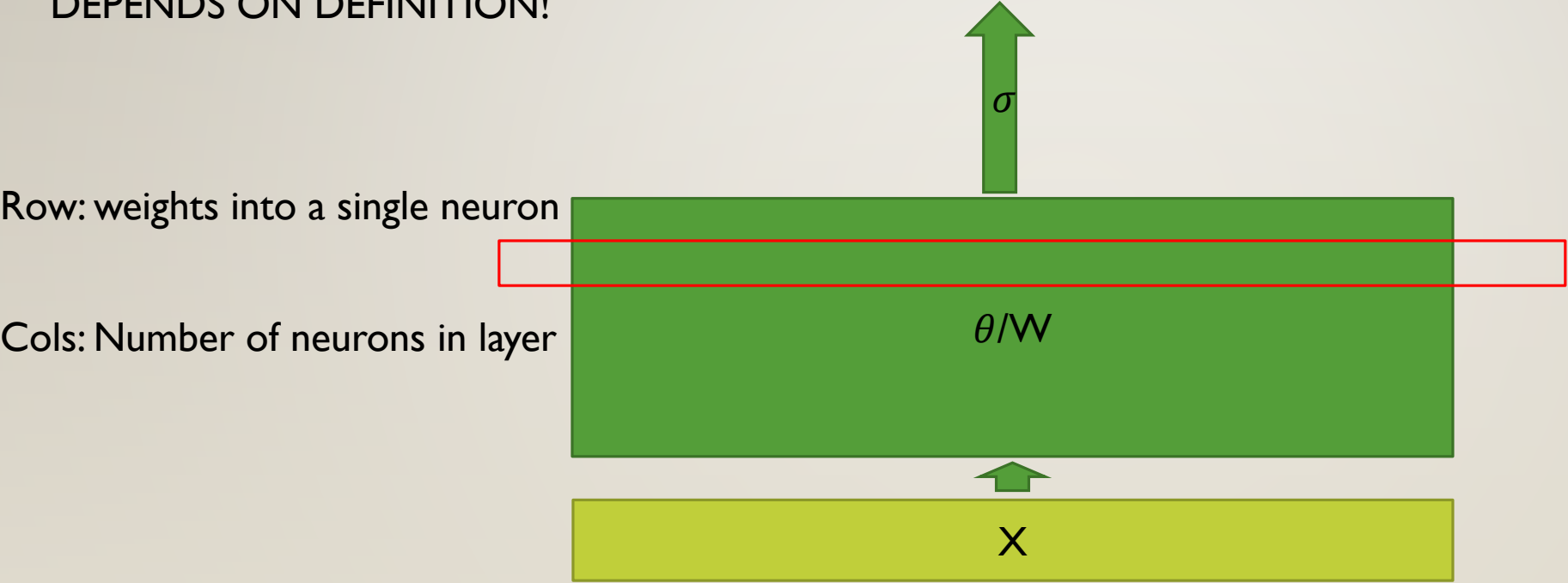
DEPENDS ON DEFINITION!

Row: weights into a single neuron



# TENSORS!

DEPENDS ON DEFINITION!





# WHY NOW? (LAST 10 YEARS)

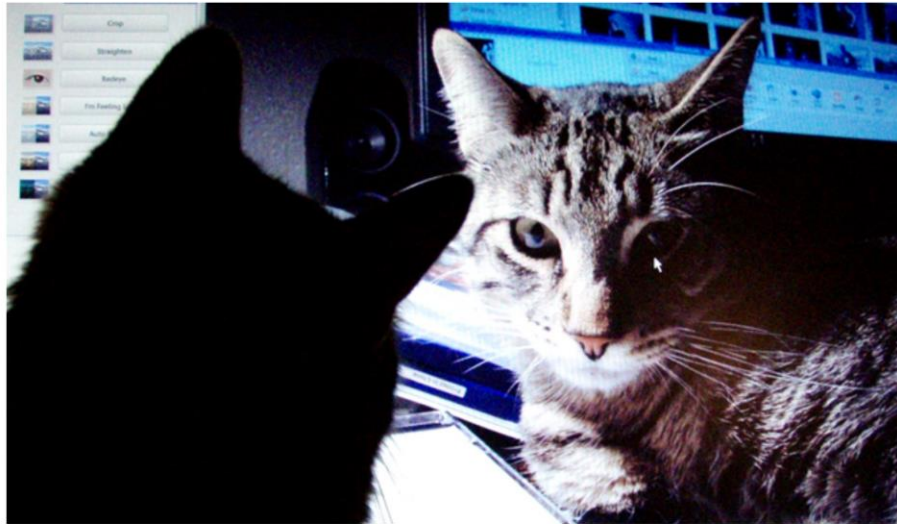
---



# CATS!

## Google's Artificial Brain Learns to Find Cat Videos

When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.



WATCH



Neuroscientist Explains One Concept in 5 Levels of Difficulty


# CATS!

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE


CORONAVIRUS HOW TO GET A COVID VACCINE BEST FACE MASKS COVID-19 FAQ NEWSLETTER LATEST NEWS

## Google's Artificial Brain Learns to Find Cat Videos

When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.



WATCH



makes the human race strong.

Neuroscientist Explains One Concept in 5 Levels of Difficulty

Get WIRED

# CATS!

# GPUs!


The image shows a screenshot of a Wired article. At the top, the Wired logo is on the left, and navigation links for 'BACKCHANNEL', 'BUSINESS', 'CULTURE', 'GEAR', 'IDEAS', 'SCIENCE', and 'SECURITY' are in the center. On the right, there are 'SIGN IN' and 'SUBSCRIBE' buttons. Below this is a secondary navigation bar with 'CORONAVIRUS' (with a red virus icon), 'HOW TO GET A COVID VACCINE', 'BEST FACE MASKS', 'COVID-19 FAQ', 'NEWSLETTER', and 'LATEST NEWS'. The main article title is 'Google's Artificial Brain Learns to Find Cat Videos'. The sub-headline reads: 'When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.' The number '16,000' is circled in red. Below the text is a large image of a black cat and a tabby cat looking at a computer screen. To the right of the main image is a 'WATCH' section featuring a video thumbnail of a man and a woman on a couch. The video title is 'Neuroscientist Explains One Concept in 5 Levels of Difficulty'. At the bottom right, there is a red banner with the text 'Get WIRED'.

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE


CORONAVIRUS HOW TO GET A COVID VACCINE BEST FACE MASKS COVID-19 FAQ NEWSLETTER LATEST NEWS

## Google's Artificial Brain Learns to Find Cat Videos

When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.



WATCH



Neuroscientist Explains One Concept in 5 Levels of Difficulty

Get WIRED

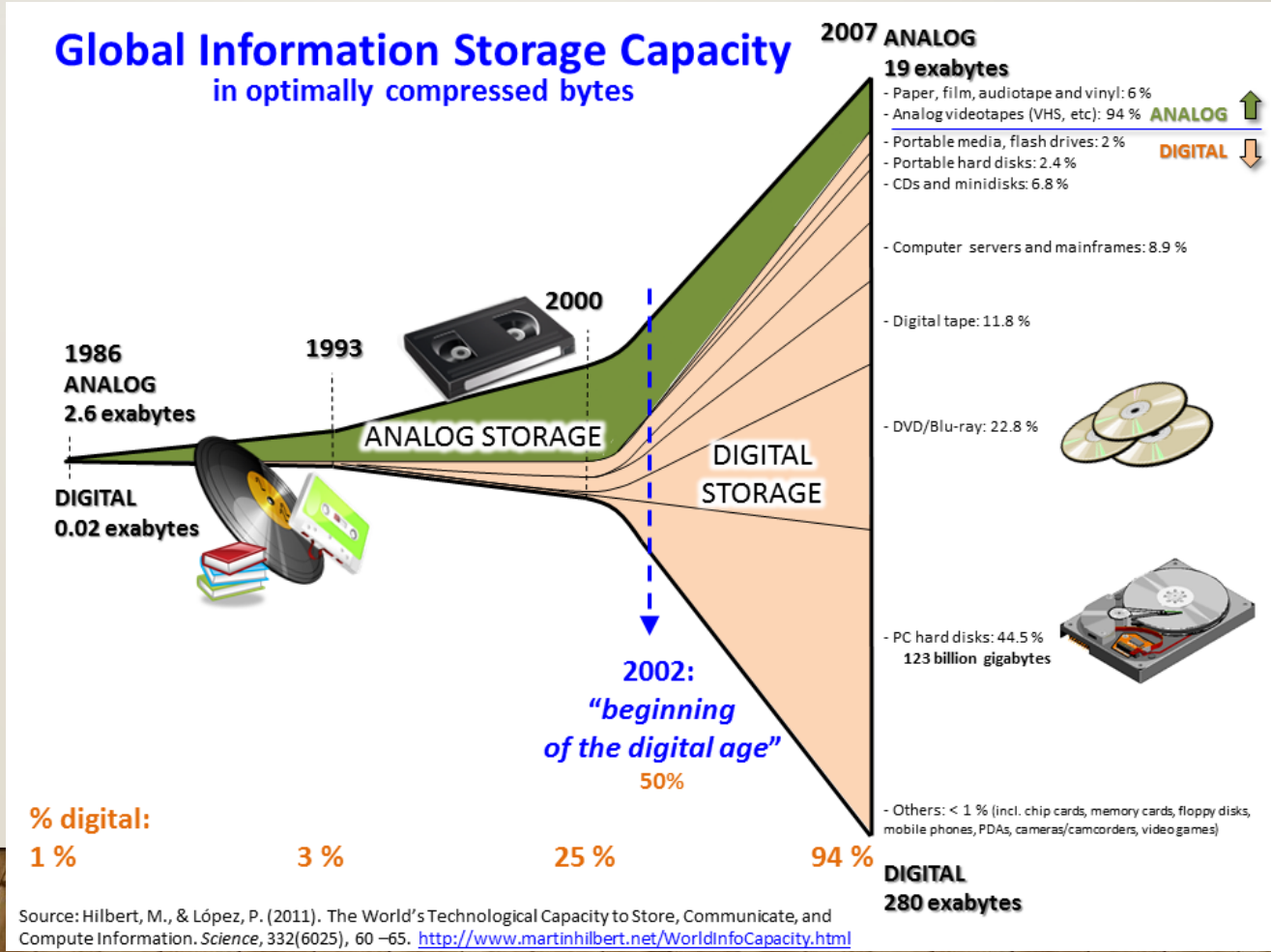
# WHY NOW? (LAST 10 YEARS)

---

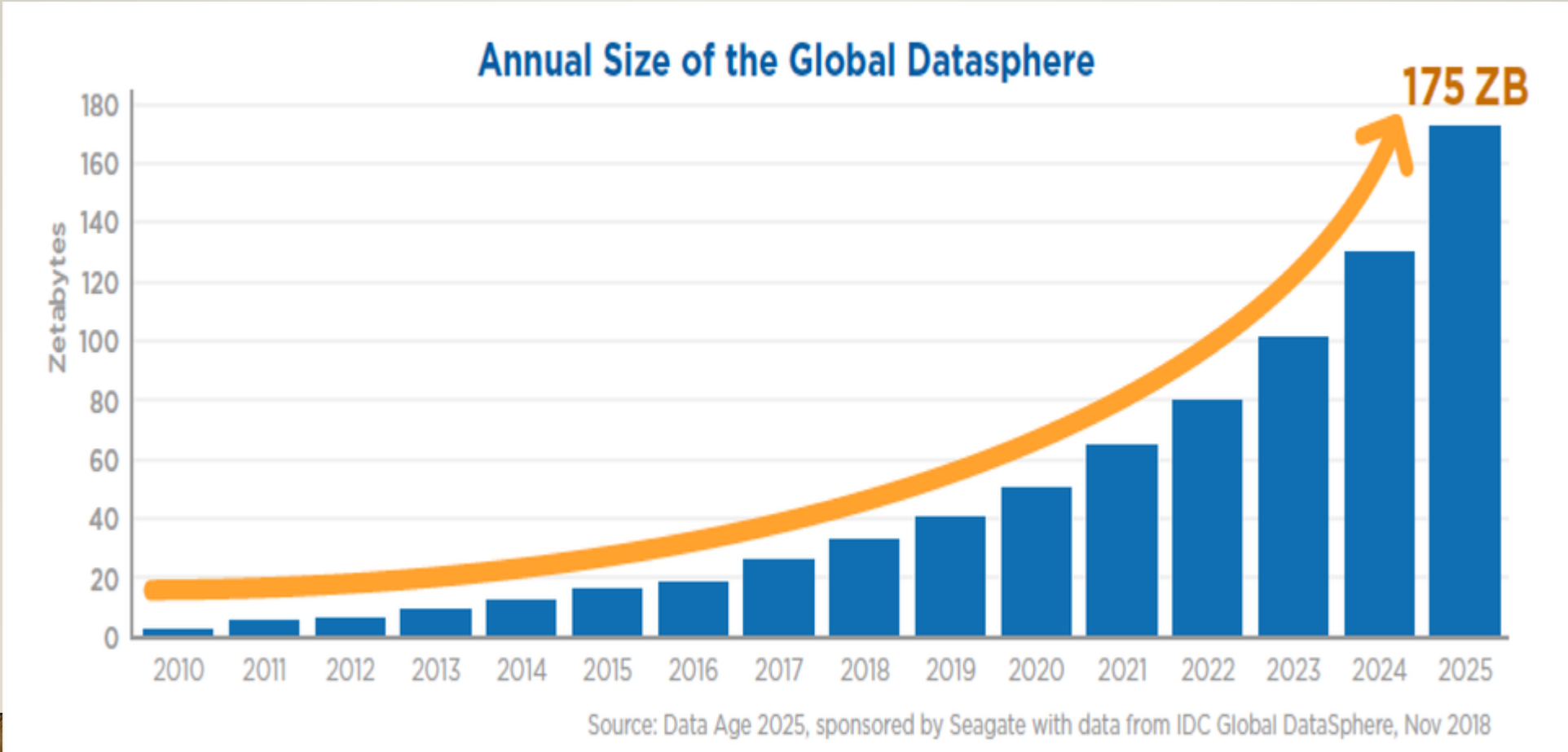
- Murray and Chiang, 2015
- 5-gram Language Model ~28 Days CPU (Grid restrictions)
- Today ~Couple of hours GPU

# WHY NOW? (LAST 10 YEARS)

- Data availability



# WHY NOW? (LAST 10 YEARS)



Forbes, 2018

# LANGUAGE MODELS

---

- Throwback to first week
- Predict the next word in a sequence



# LANGUAGE MODELS

---

- Throwback to first week
- Predict the next word in a sequence



# LANGUAGE MODELS

---

- Throwback to first week
- Predict the next word in a sequence

Johns Hopkins University was \_\_\_\_\_

# LANGUAGE MODELS

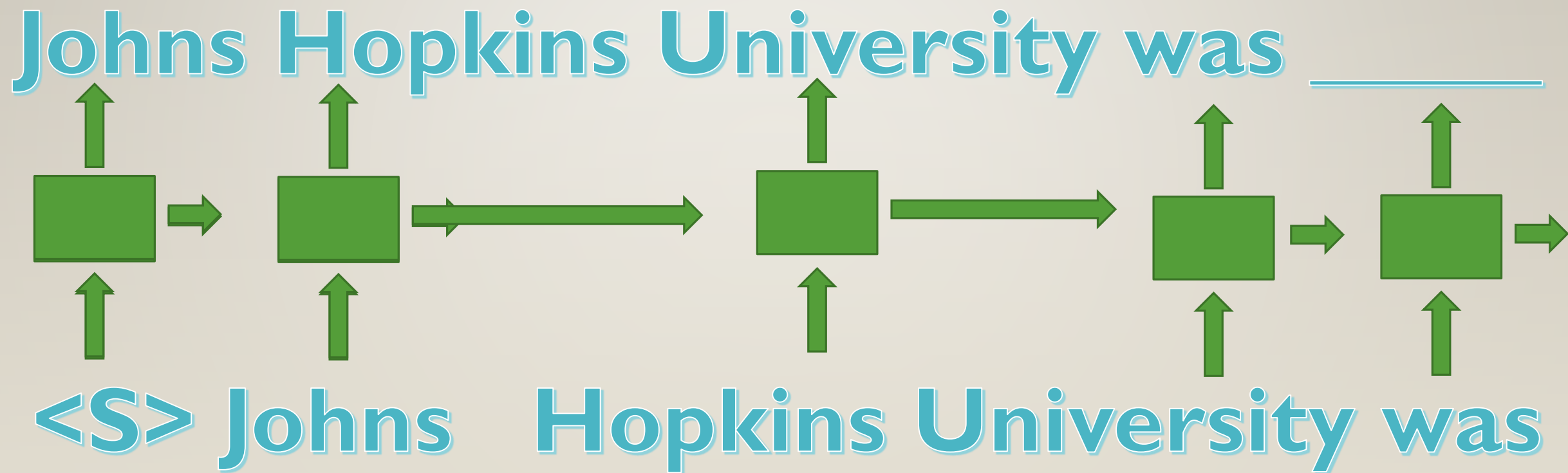
---

- Historically: vocab,  $V$
- n-gram language model was  $|V|^n$
- Data Sparsity issues (5-gram was common)

Johns Hopkins University was \_\_\_\_\_

# RNN LMS

---



# RNN LMS



# BRNN

- Schuster and Paliwal, 1997

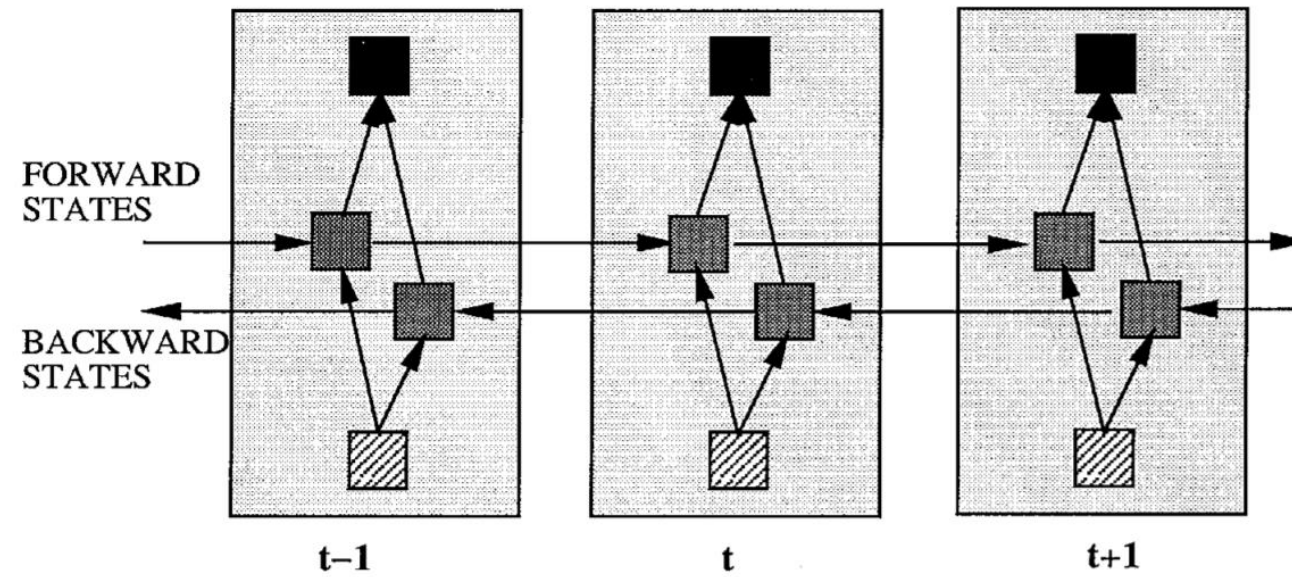


Fig. 3. General structure of the bidirectional recurrent neural network (BRNN) shown unfolded in time for three time steps.

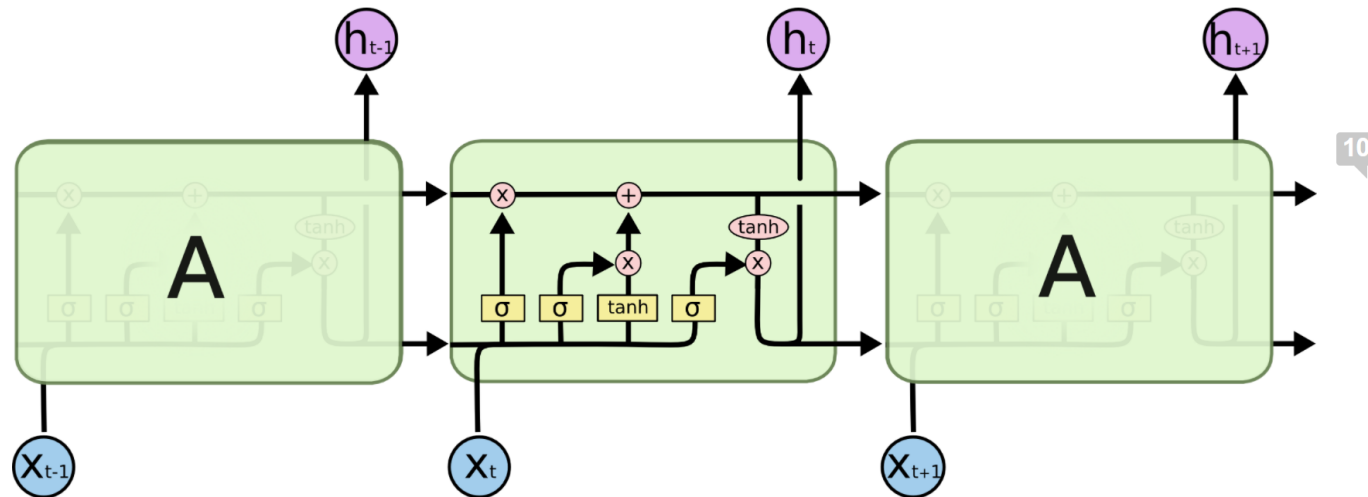
# LSTMS

---

- Long-Short Term Memory
- Cell
- Input Gate
- Output Gate
- Forget Gate
- S. Hochreiter and J. Schmidhuber, 1997

# LSTMS

- <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

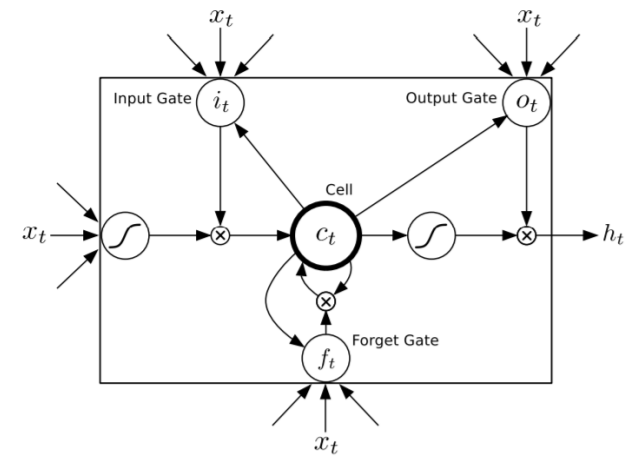


The repeating module in an LSTM contains four interacting layers.

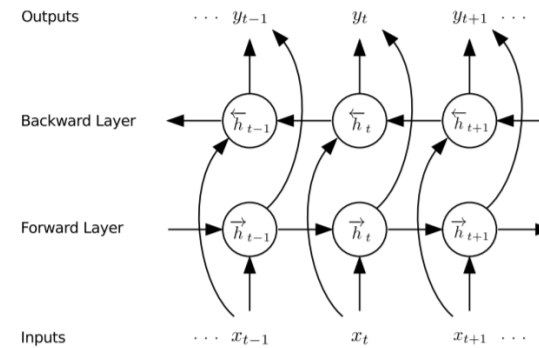


# BI-LSTM

- Graves et al., 2013



**Fig. 1.** Long Short-term Memory Cell



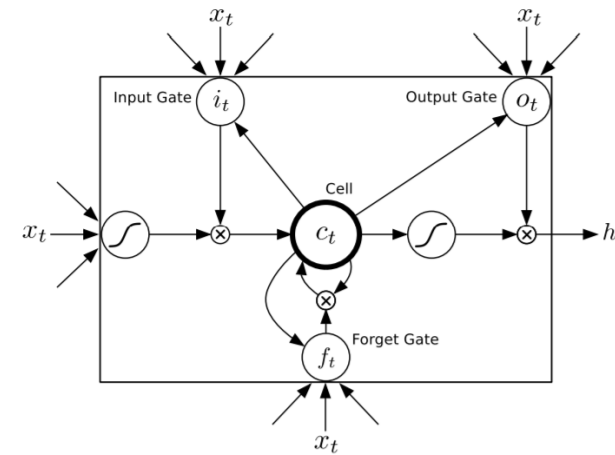
**Fig. 2.** Bidirectional RNN

# BI-LSTM

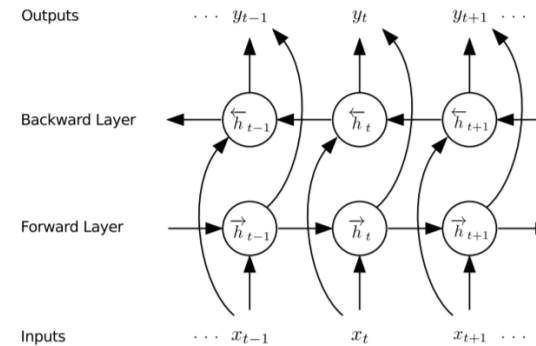
- Graves et al., 2013

**Table 1.** TIMIT Phoneme Recognition Results. ‘Epochs’ is the number of passes through the training set before convergence. ‘PER’ is the phoneme error rate on the core test set.

NETWORK	WEIGHTS	EPOCHS	PER
CTC-3L-500H-TANH	3.7M	107	37.6%
CTC-1L-250H	0.8M	82	23.9%
CTC-1L-622H	3.8M	87	23.0%
CTC-2L-250H	2.3M	55	21.0%
CTC-3L-421H-UNI	3.8M	115	19.6%
CTC-3L-250H	3.8M	124	18.6%
CTC-5L-250H	6.8M	150	18.4%
TRANS-3L-250H	4.3M	112	18.3%
<b>PRETRANS-3L-250H</b>	<b>4.3M</b>	<b>144</b>	<b>17.7%</b>



**Fig. 1.** Long Short-term Memory Cell



**Fig. 2.** Bidirectional RNN

# MACHINE TRANSLATION

---

- Source Sentence,  $f$ : “Yo tengo hambre”
- Target Sentence,  $e$ : “I am hungry”
- $P(e|f)$  ..... Neural Network

# MACHINE TRANSLATION

- Cho et al., 2014

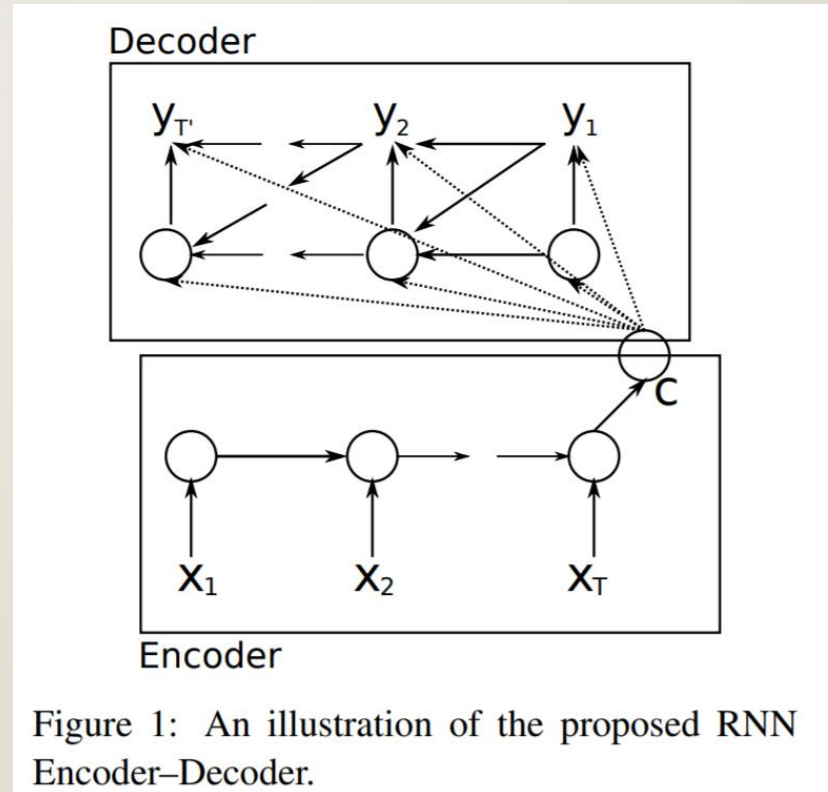


Figure 1: An illustration of the proposed RNN Encoder-Decoder.

# MACHINE TRANSLATION

- Cho et al., 2014
- WMT '14 En-Fr

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64

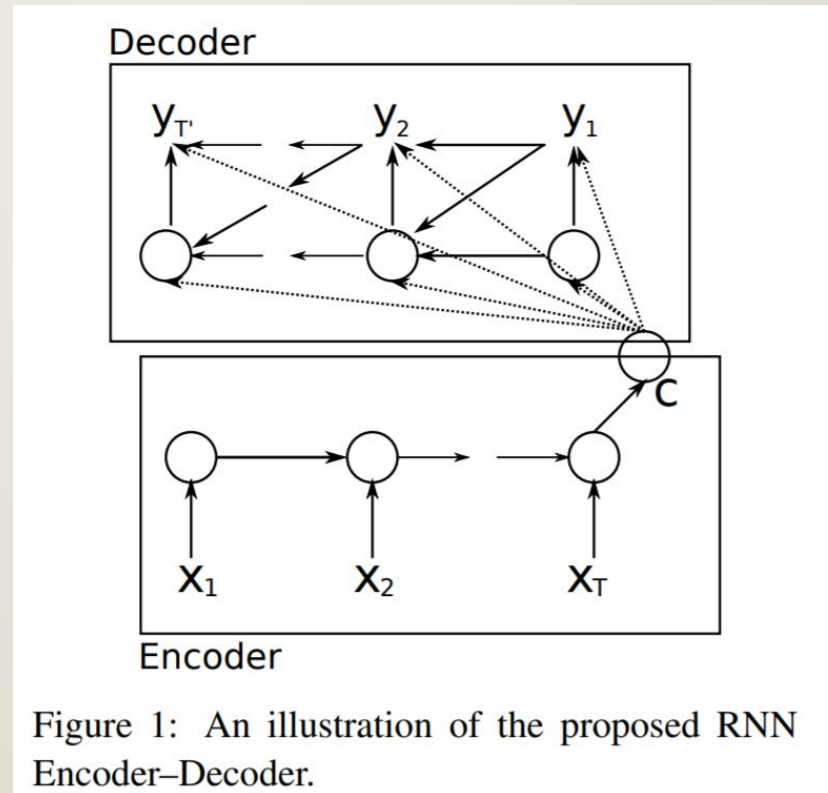


Figure 1: An illustration of the proposed RNN Encoder-Decoder.

# MACHINE TRANSLATION

---

- Sutskever et al., 2014

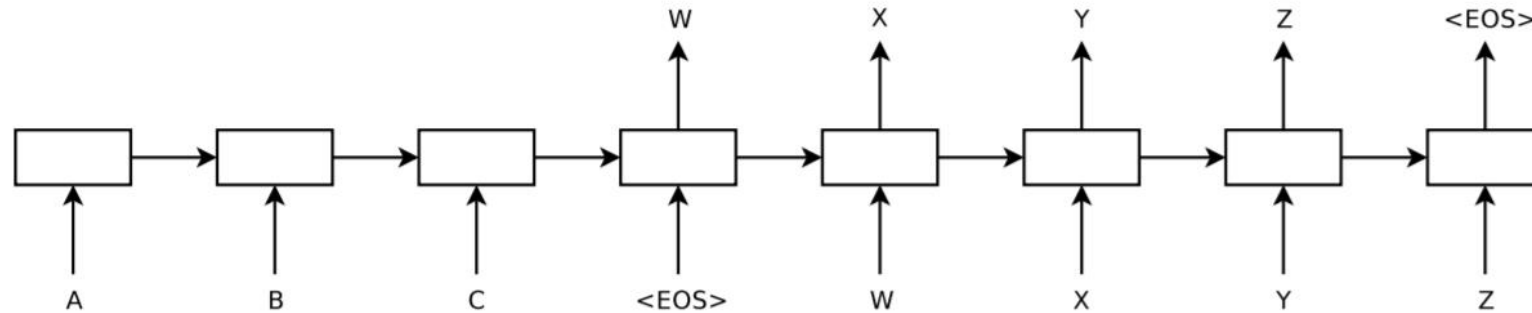
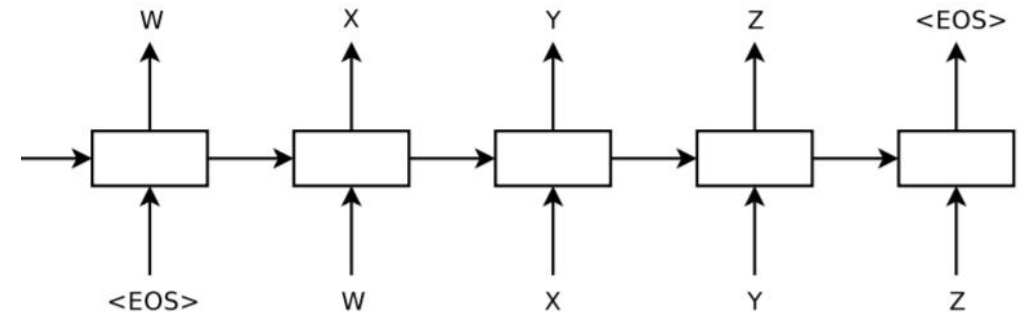


Figure 1: Our model reads an input sentence “ABC” and produces “WXYZ” as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

# MACHINE TRANSLATION

- Sutskever et al., 2014
- WMT '14 En-Fr

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	<b>34.81</b>



entence “ABC” and produces “WXYZ” as the output sentence. The outputting the end-of-sentence token. Note that the LSTM reads the so introduces many short term dependencies in the data that make the

# ATTENTION

- Bahdanau et al., 2015

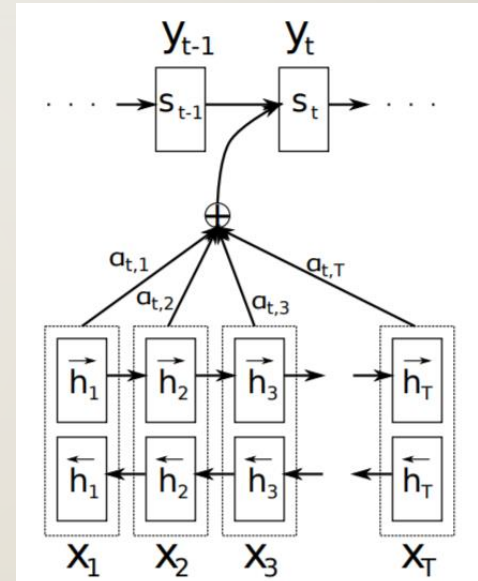


Figure 1: The graphical illustration of the proposed model trying to generate the  $t$ -th target word  $y_t$  given a source sentence  $(x_1, x_2, \dots, x_T)$ .



# ATTENTION

- Bahdanau et al., 2015
- WMT '14 En-Fr

Model	All	No UNK <sup>o</sup>
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63

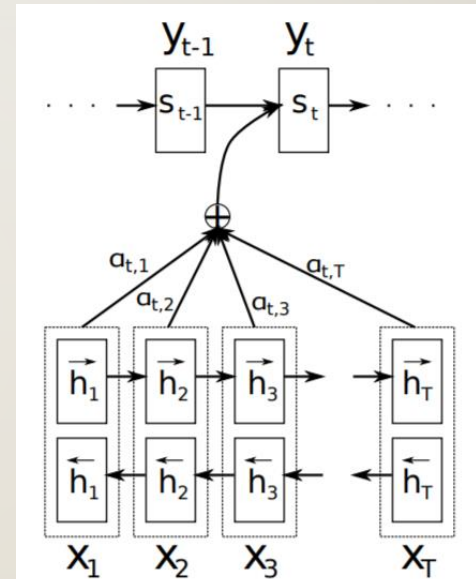


Figure 1: The graphical illustration of the proposed model trying to generate the  $t$ -th target word  $y_t$  given a source sentence  $(x_1, x_2, \dots, x_T)$ .

# ATTENTION

- Bahdanau et al., 2015

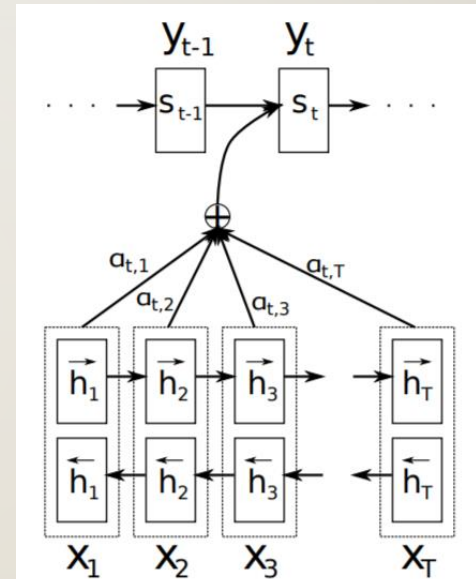
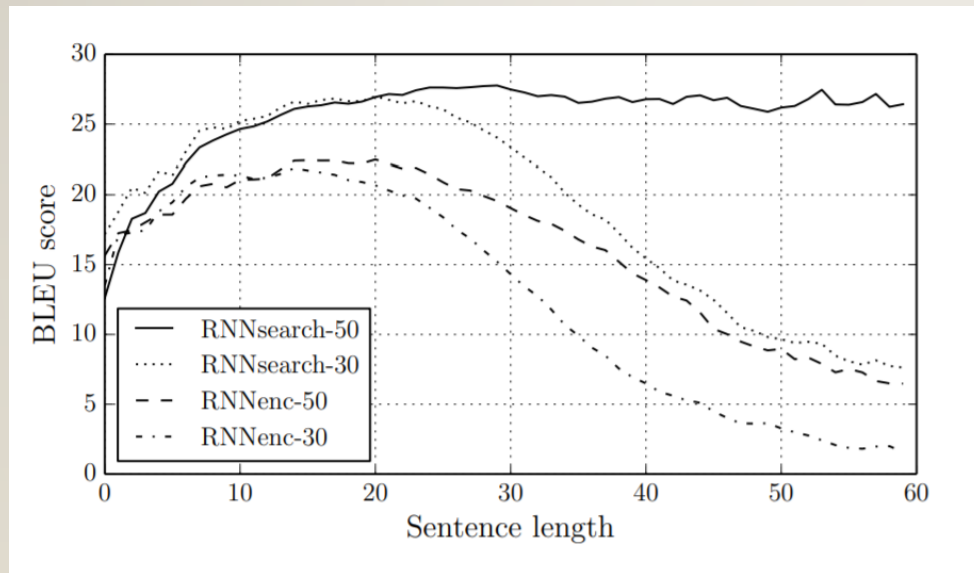


Figure 1: The graphical illustration of the proposed model trying to generate the  $t$ -th target word  $y_t$  given a source sentence  $(x_1, x_2, \dots, x_T)$ .

# ATTENTION IS ALL YOU NEED

- Vaswani et al., 2017
- Transformer

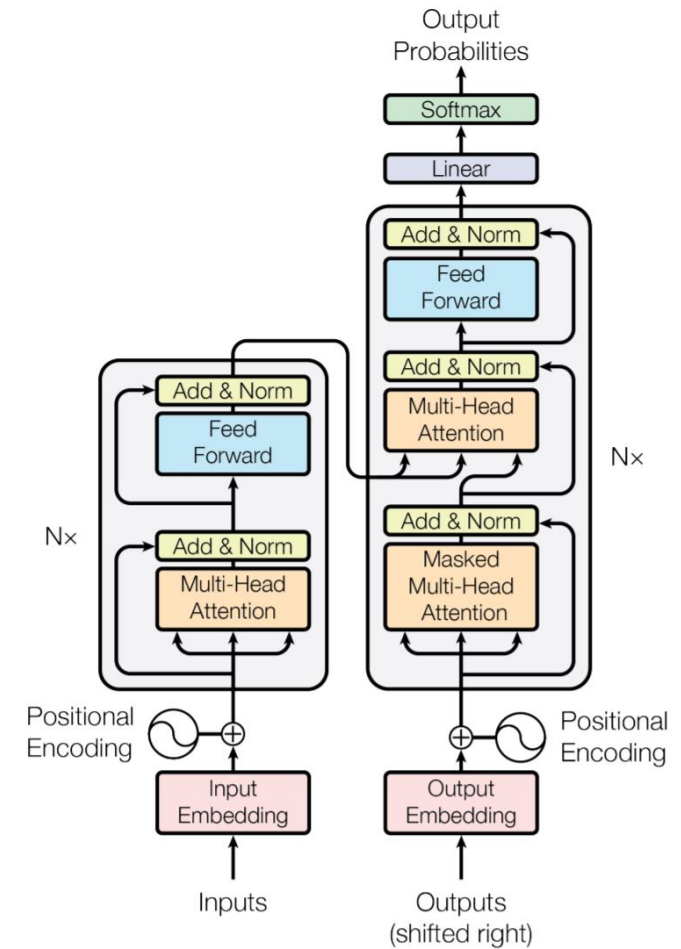


Figure 1: The Transformer - model architecture.

# ATTENTION IS ALL YOU NEED

- Vaswani et al., 2017

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	<b>41.29</b>	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	<b><math>3.3 \cdot 10^{18}</math></b>	
Transformer (big)	<b>28.4</b>	<b>41.8</b>	$2.3 \cdot 10^{19}$	

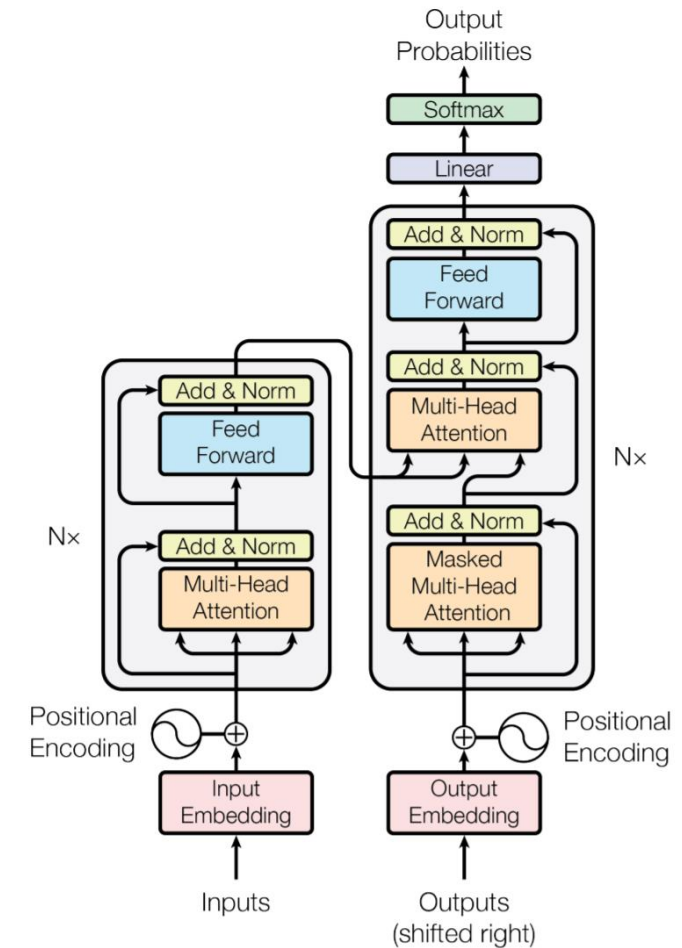


Figure 1: The Transformer - model architecture.

# BERT

---

- Devlin et al., 2018
- Bidirectional Encoder Representations from Transformers
- Masked Language Model



BERT



Johns Hopkins University was \_\_\_\_\_

BERT



Johns Hopkins University was founded in the year eighteen seventy six .

# BERT

Johns Hopkins \_\_\_\_\_ was founded in the year eighteen \_\_\_\_\_ six .



Johns Hopkins [redacted] was founded in the year eighteen [redacted] six .



# BERT

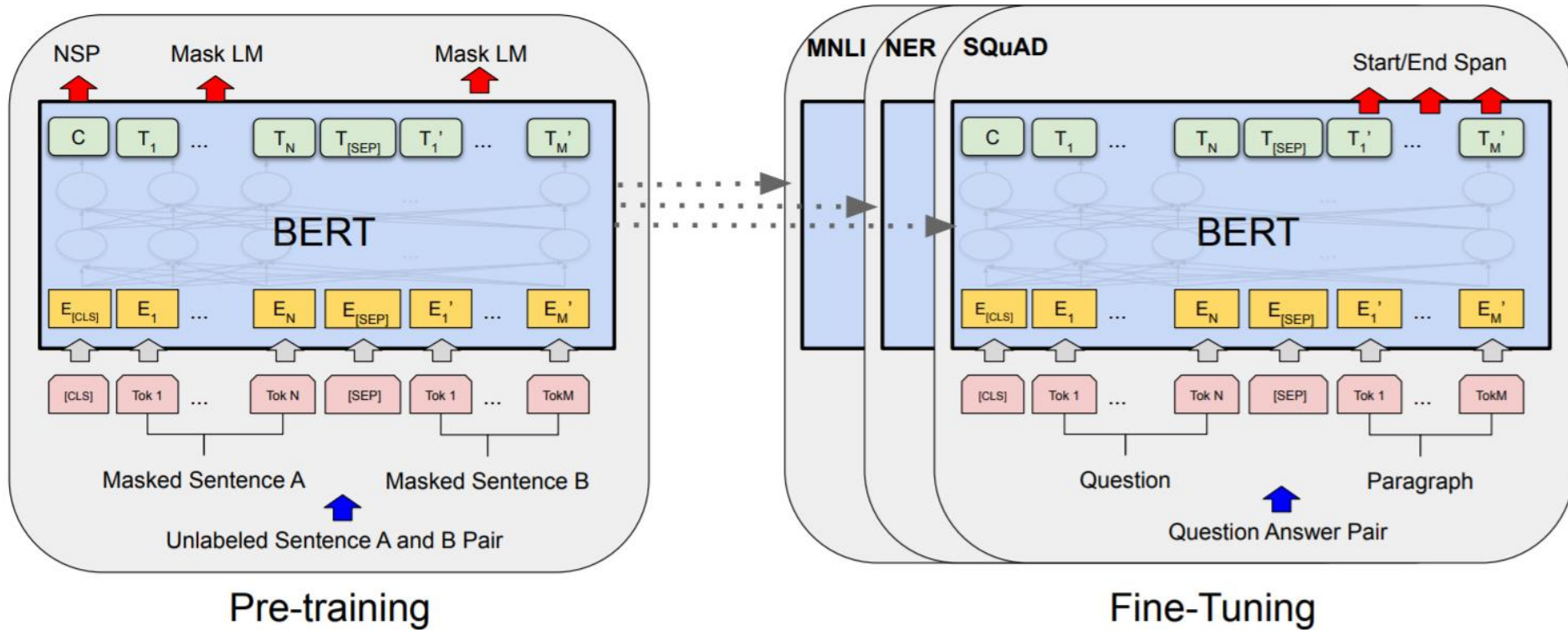


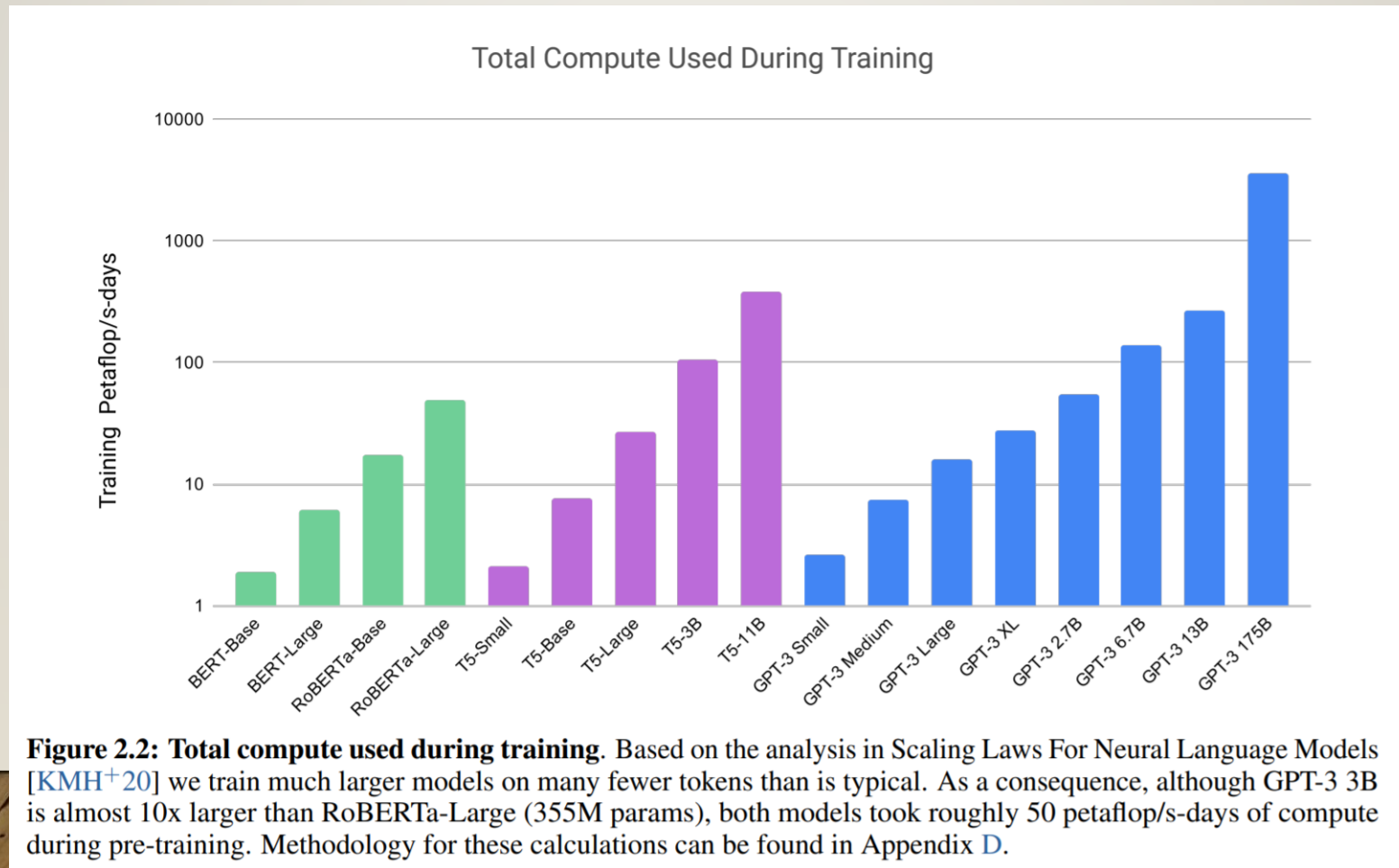
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

# GPT-3

---

- Brown et al., 2020
- Auto-Regressive (Looks to previous context)

# GPT-3



# GPT-3

- WMT '14 MT
- SacreBLEU (proper eval)
- Trained only on English

Setting	En→Fr
SOTA (Supervised)	<b>45.6<sup>a</sup></b>
XLM [LC19]	33.4
MASS [STQ <sup>+</sup> 19]	<u>37.5</u>
mBART [LGG <sup>+</sup> 20]	-
GPT-3 Zero-Shot	25.2
GPT-3 One-Shot	28.3
GPT-3 Few-Shot	32.6