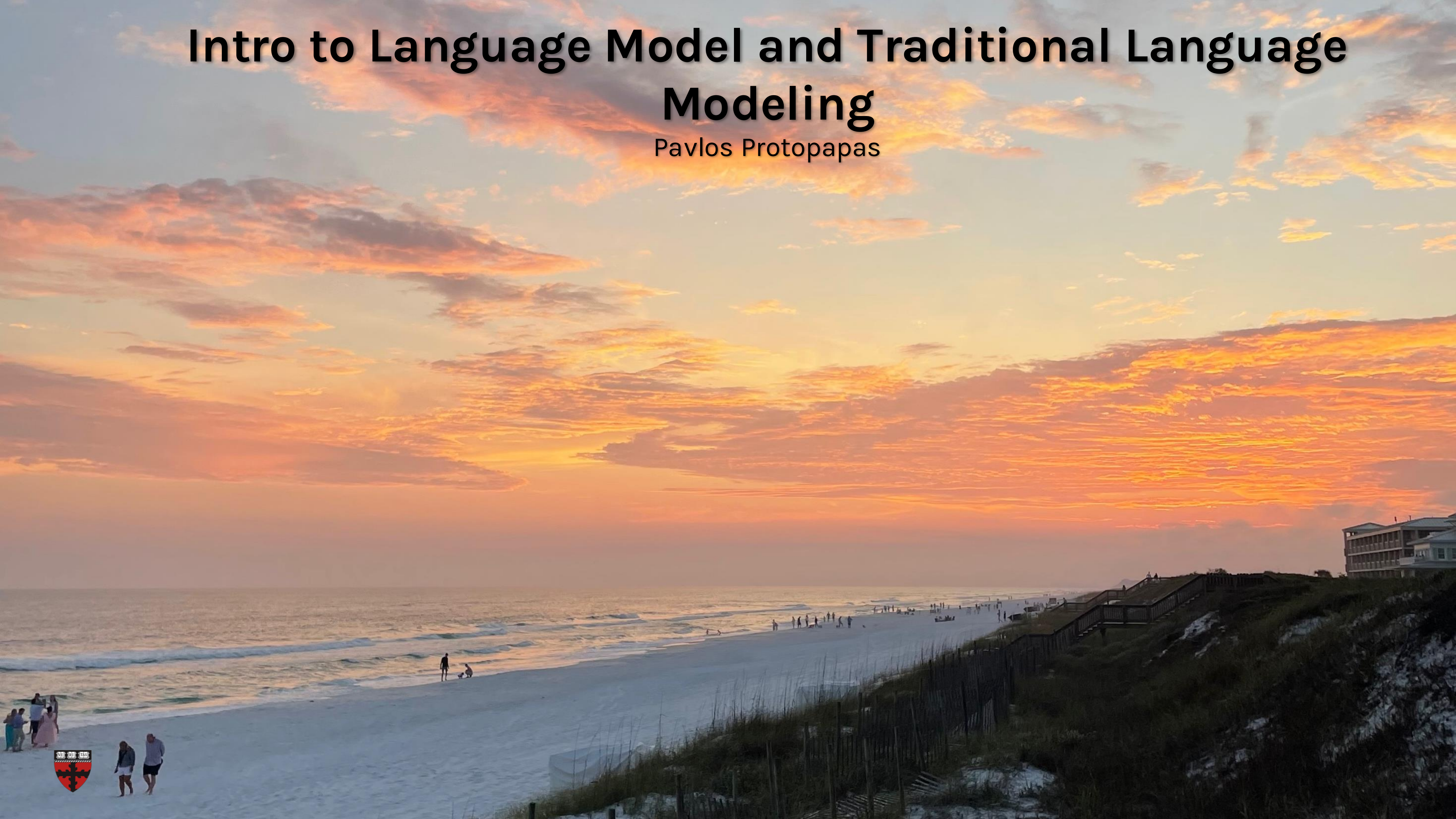
# Intro to Language Model and Traditional Language Modeling

Pavlos Protopapas

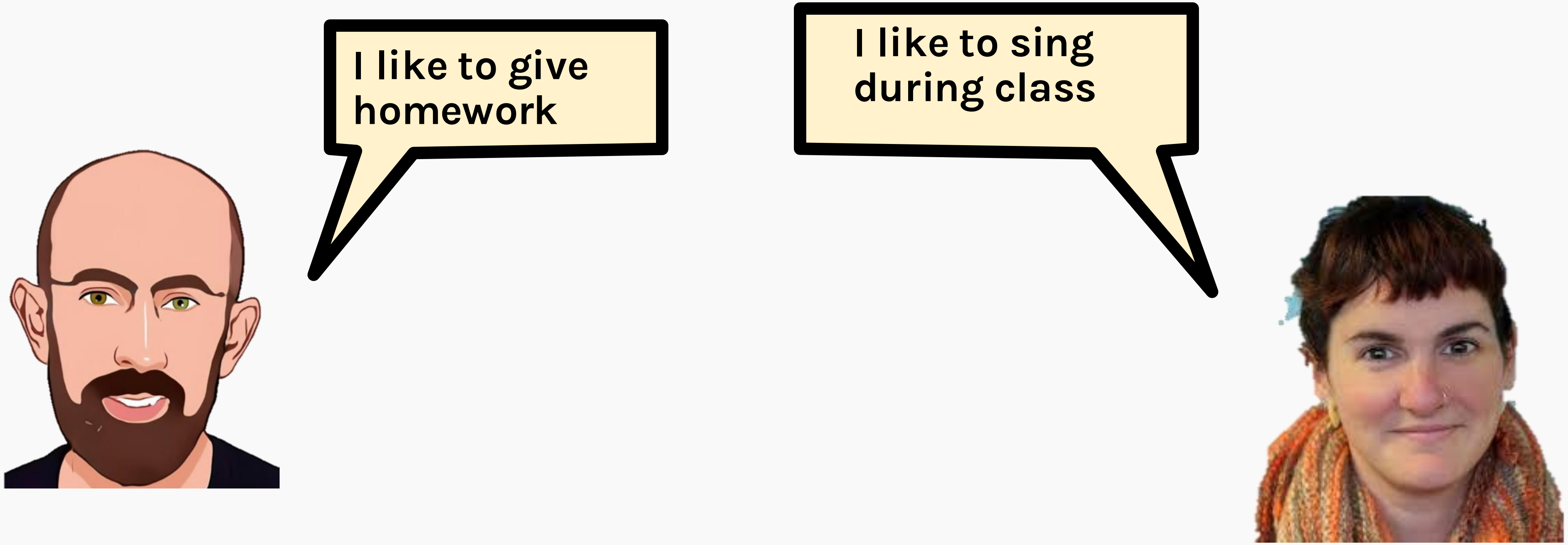


# Outline

- Natural Language Processing
- Text Preprocessing
- Language Modeling
  - Unigrams
  - Bigrams
  - Neural Networks for Language Modeling

# Natural Language Processing

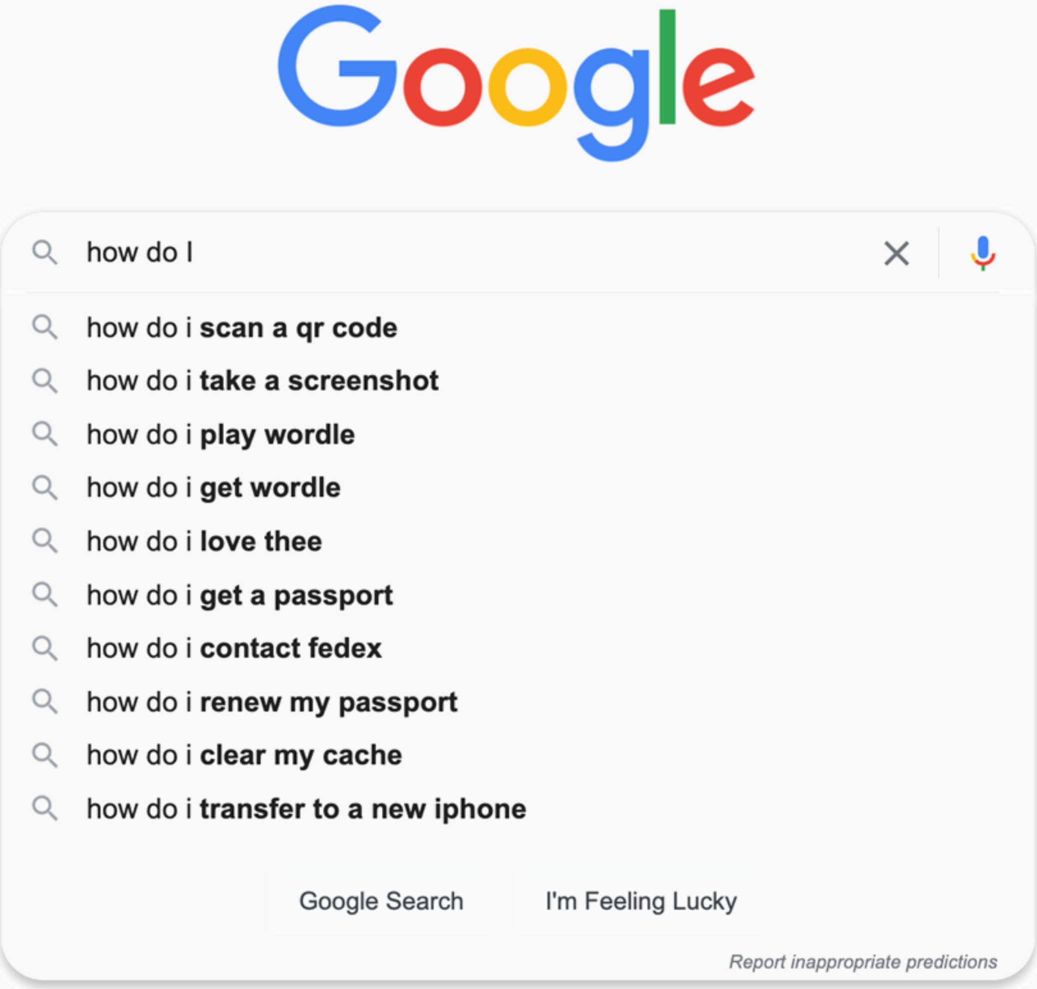
Natural Language Processing (NLP) is the field of study that focuses on the interaction between computers and humans through natural language, aiming to enable machines to understand, interpret, and respond to human language in a meaningful way.



## Text recognition

# Natural Language Processing: **Applications**

## Sentence prediction



# Natural Language Processing: **Applications**

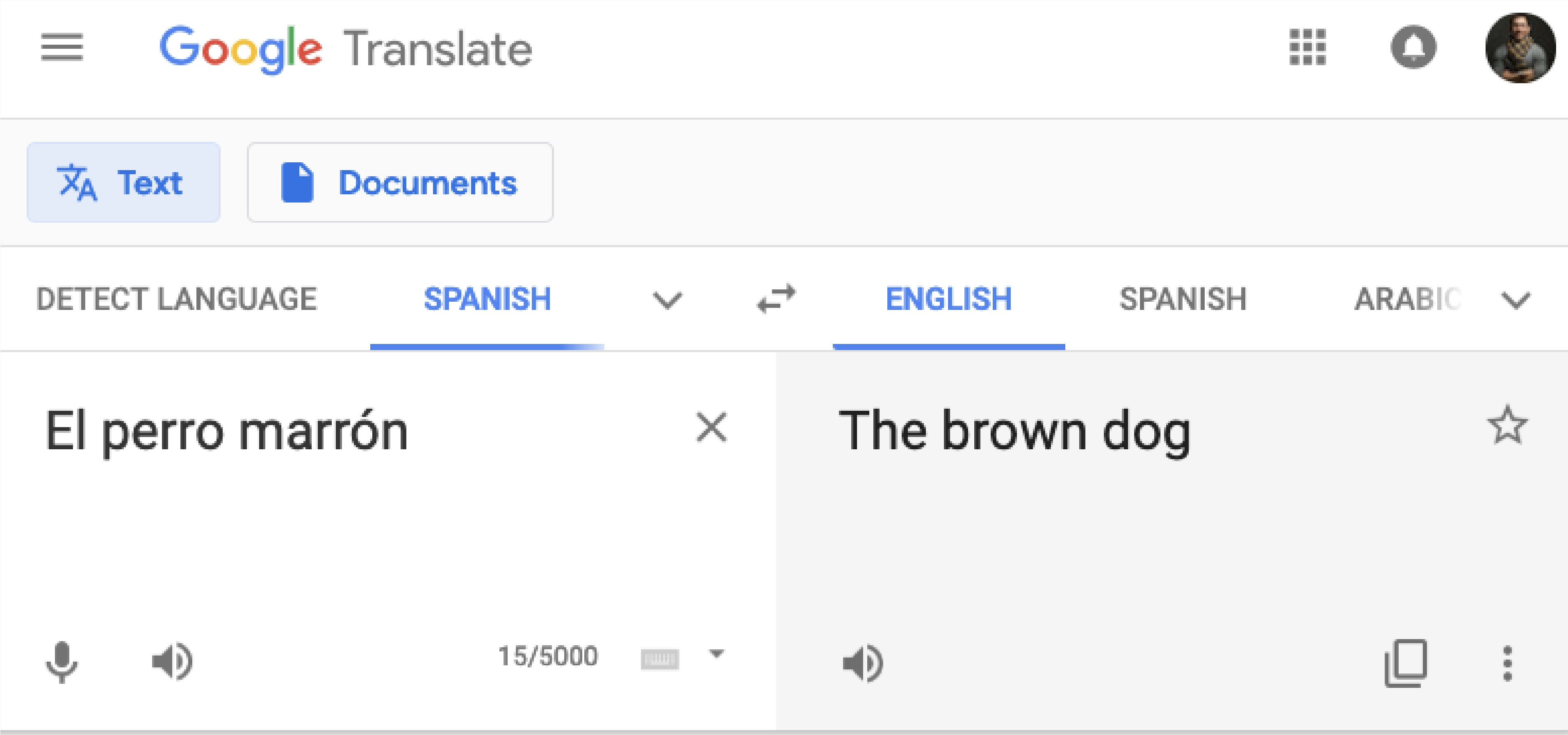
## Sentence prediction

# Named Entity Recognition

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

# Natural Language Processing: **Applications**

## Translation



# Natural Language Processing: **Applications**

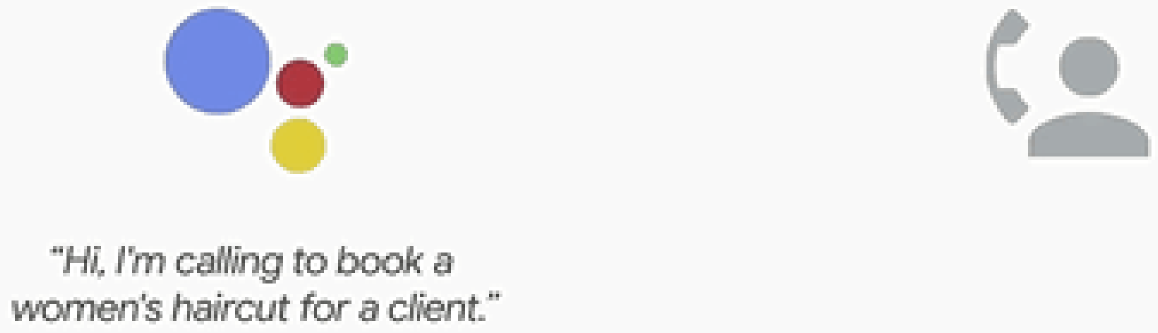
## ChatGPT

| Examples | Capabilities | Limitations |
|---|---|---|
| "Explain quantum computing in simple terms" | Remembers what user said earlier in the conversation | May occasionally generate incorrect information |
| "Got any creative ideas for a 10 year old's birthday?" | Allows user to provide follow-up corrections | May occasionally produce harmful instructions or biased content |
| "How do I make an HTTP request in Javascript?" | Trained to decline inappropriate requests | Limited knowledge of world and events after 2021 |

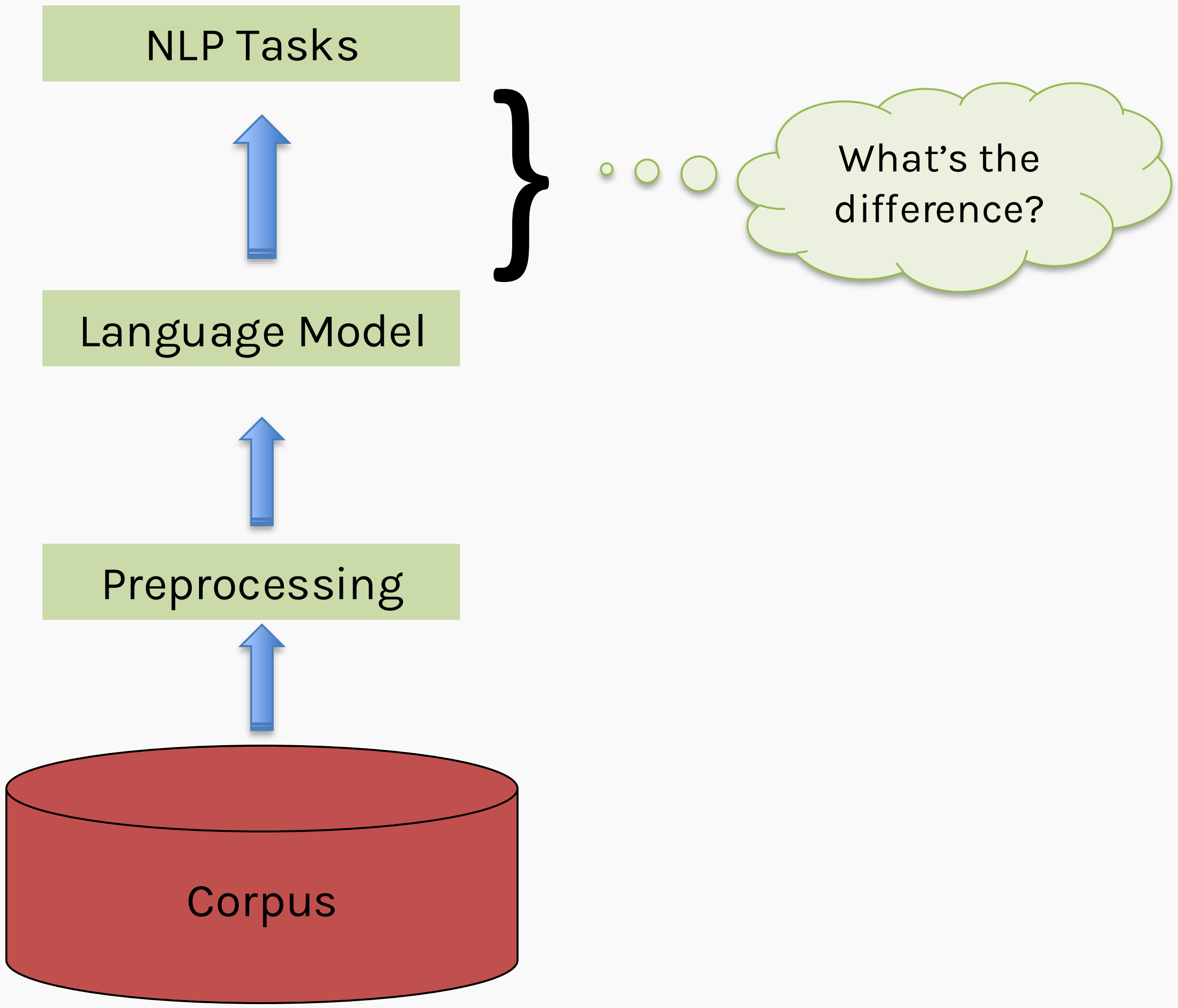
# Natural Language Processing: **Applications**

## Chat Bots

## AI answering calls



# Pipeline



NLP Tasks
What's the difference?
Language Model
Preprocessing
Corpus

# Game Time

How do we teach a toddler to answer a question? Choose the correct order.

I. Teach the toddler to answer questions

II. Teach the toddler language

A. I -> II

B. II -> I

C. I don’t want the toddler to answer questions

D. None of the above

# Game Time

How do we teach a toddler to answer a question? Choose the correct order.

I. Teach the toddler to answer questions

II. Teach the toddler language

A. I -> II

B. II -> I

C. I don’t want the toddler to answer questions

D. None of the above

# NLP Task v/s Language Model

- Similarly, we would want to teach our model to understand language before any downstream tasks like NER, question answering etc.

But can't the model learn that while we're training on a downstream task dataset?

In principle yes, but labeled data are limited.

# NLP Task v/s Language Model

- Similarly, we would want to teach our model to understand language before any downstream tasks like NER, question answering etc.

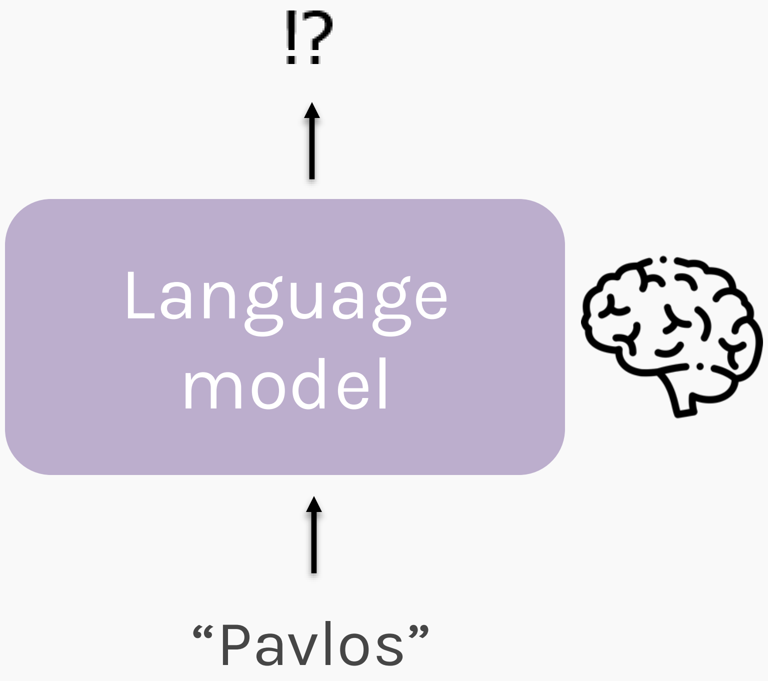
Then what can we do about it?

We can tap into the vast ocean of text from the internet.

!!

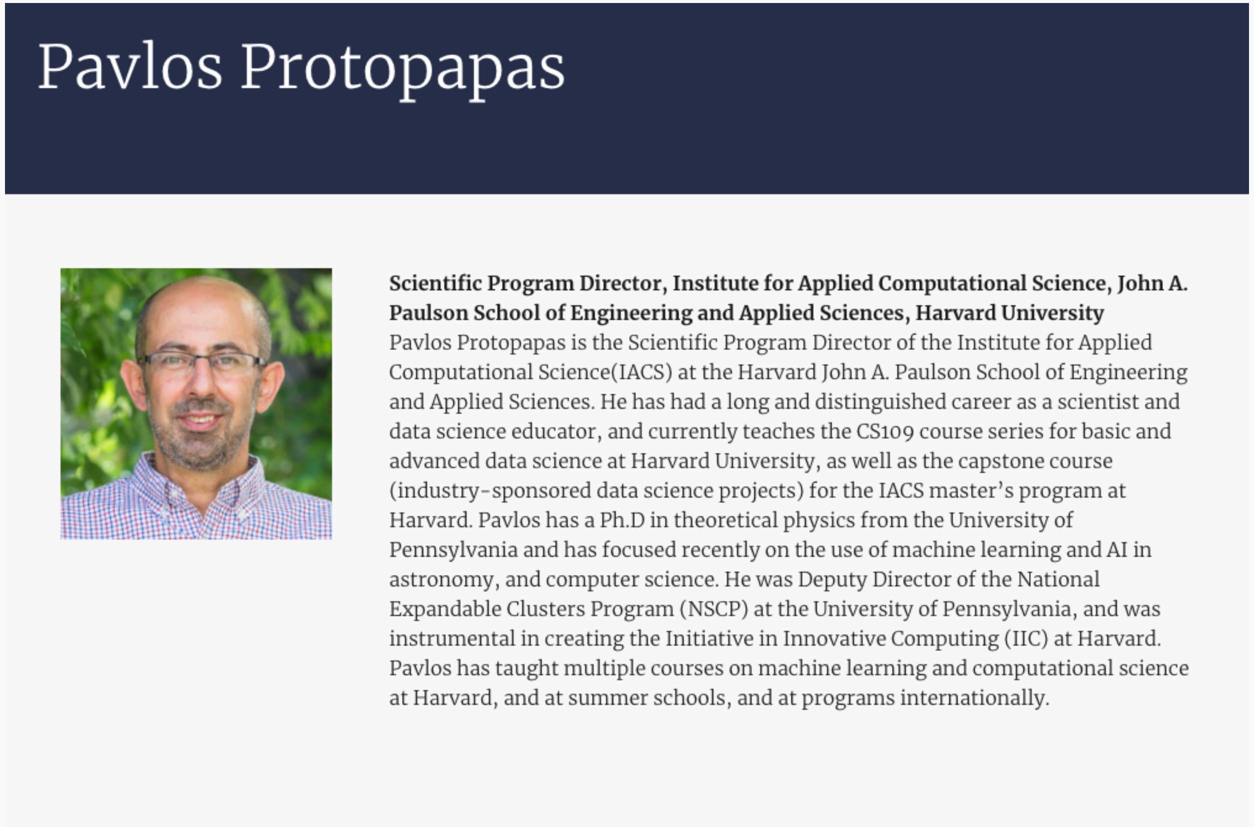
# NLP Task v/s Language Model

1. We start with an untrained language model.



# NLP Task v/s Language Model

2. We get data from the internet.



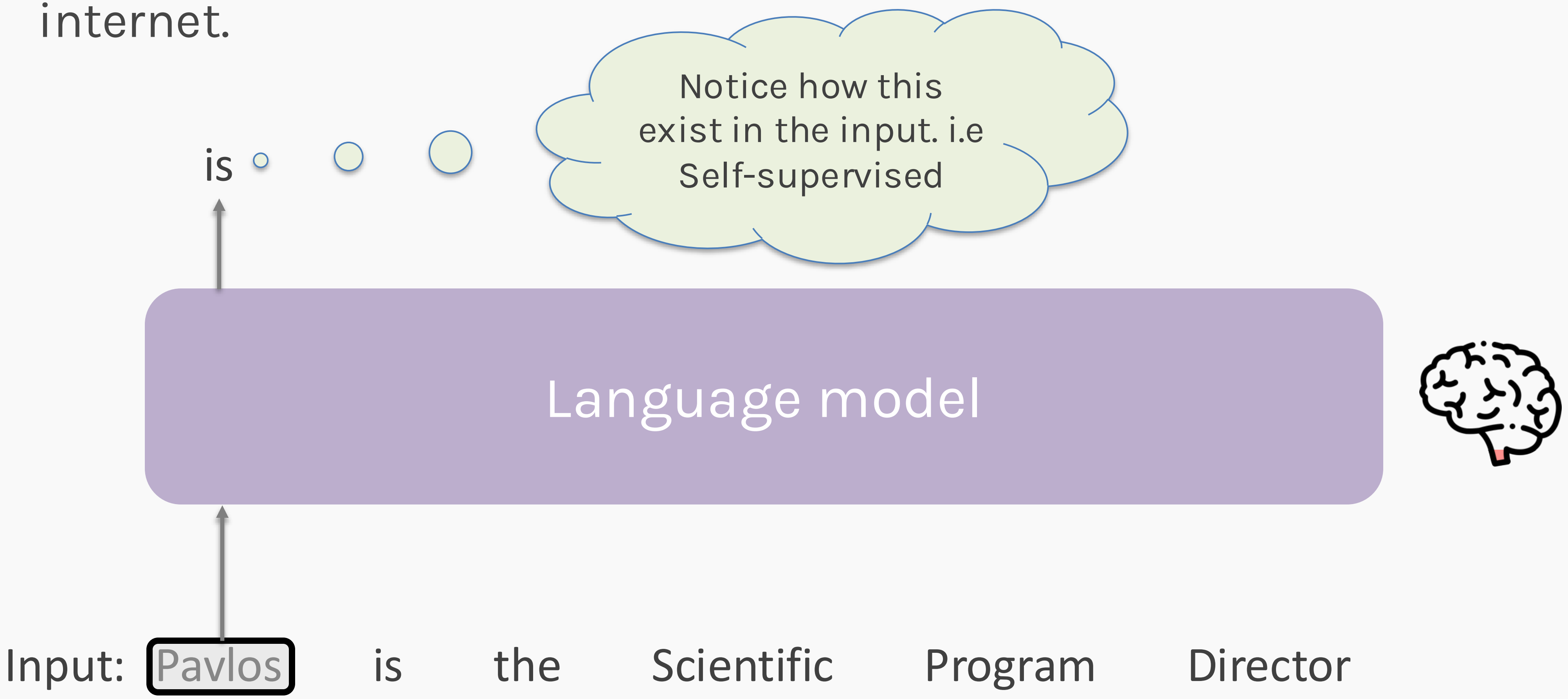
Pavlos Protopapas

**Scientific Program Director, Institute for Applied Computational Science, John A. Paulson School of Engineering and Applied Sciences, Harvard University**
Pavlos Protopapas is the Scientific Program Director of the Institute for Applied Computational Science(IACS) at the Harvard John A. Paulson School of Engineering and Applied Sciences. He has had a long and distinguished career as a scientist and data science educator, and currently teaches the CS109 course series for basic and advanced data science at Harvard University, as well as the capstone course (industry-sponsored data science projects) for the IACS master's program at Harvard. Pavlos has a Ph.D in theoretical physics from the University of Pennsylvania and has focused recently on the use of machine learning and AI in astronomy, and computer science. He was Deputy Director of the National Expandable Clusters Program (NSCP) at the University of Pennsylvania, and was instrumental in creating the Initiative in Innovative Computing (IIC) at Harvard. Pavlos has taught multiple courses on machine learning and computational science at Harvard, and at summer schools, and at programs internationally.

Language model

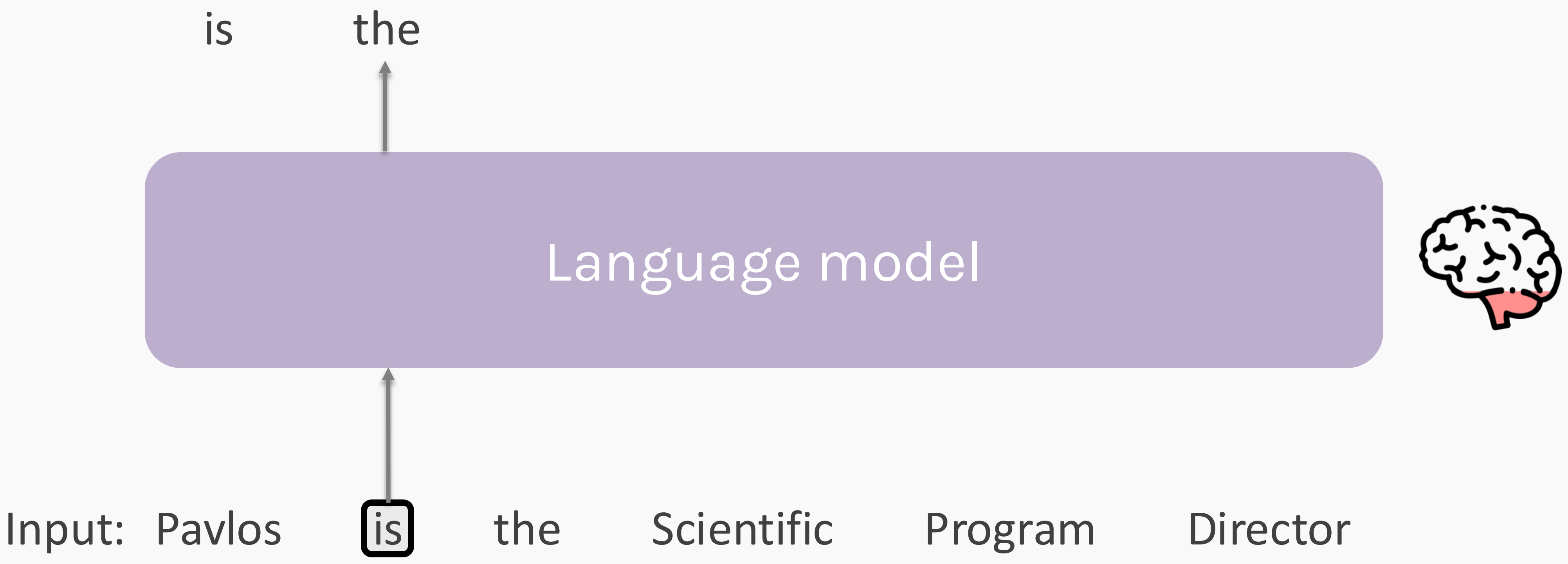
# NLP Task v/s Language Model

3. We train the language model using the data that we got from the internet.



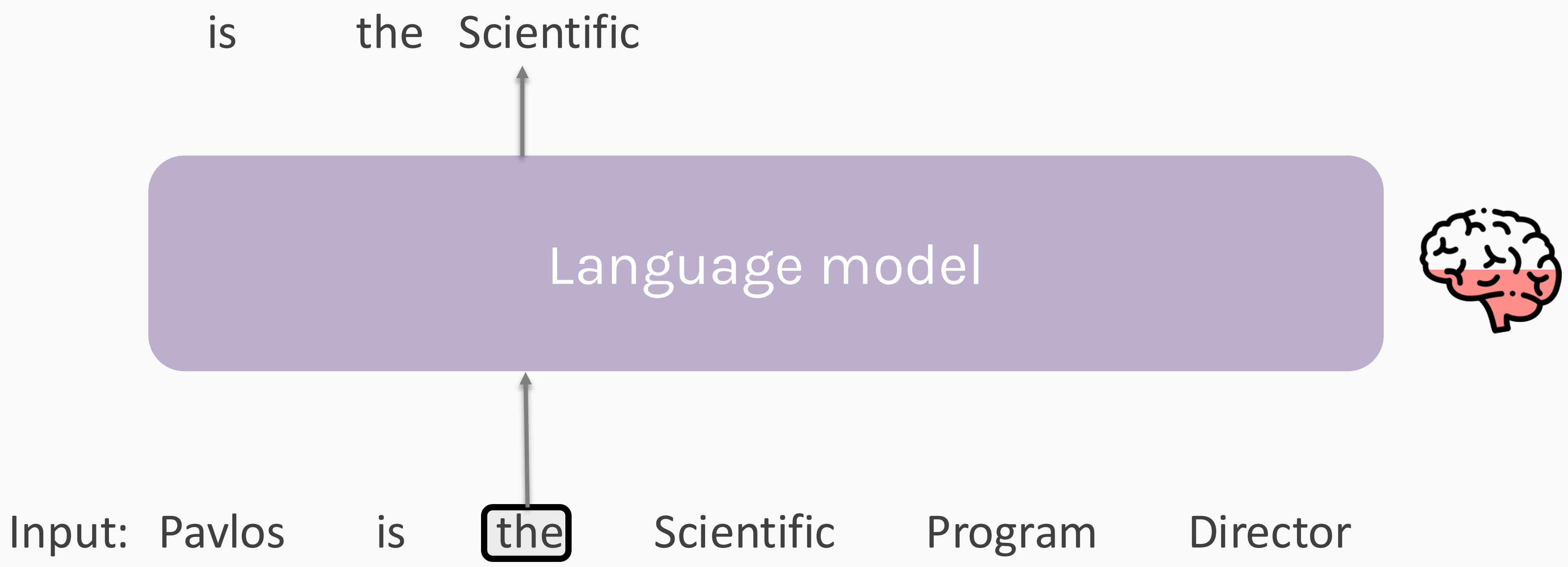
# NLP Task v/s Language Model

3. We train the language model using the data that we got from the internet.



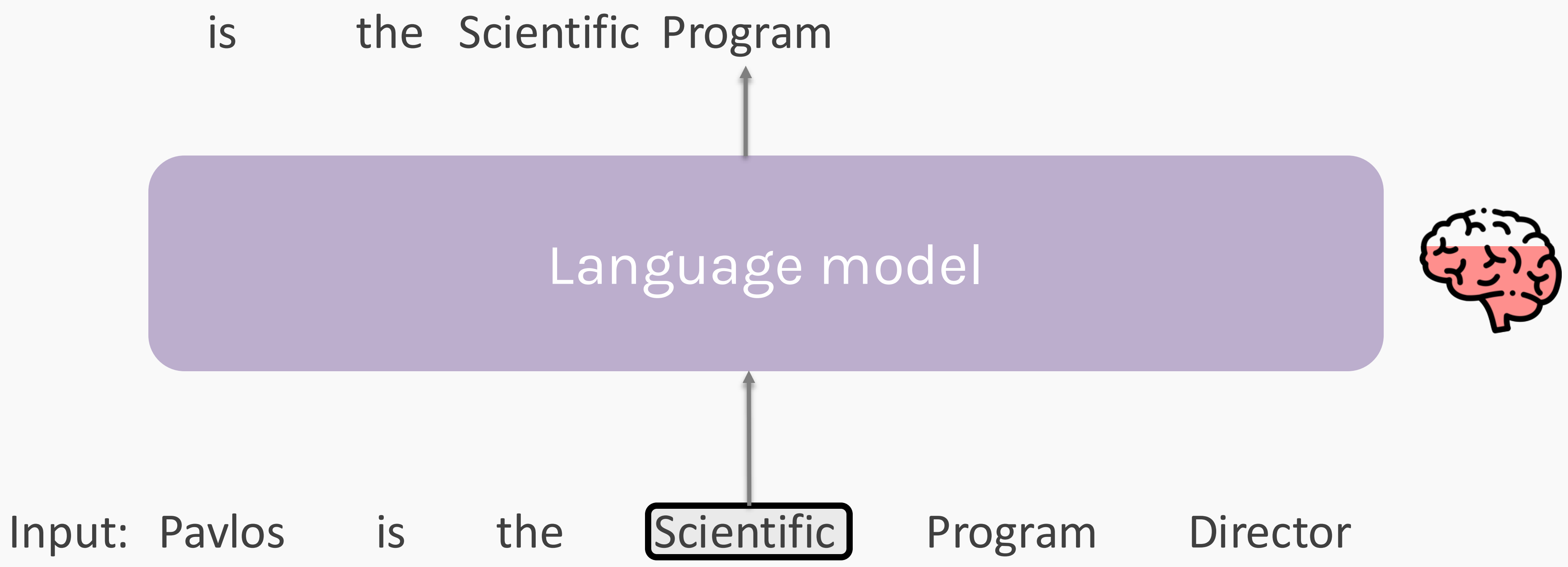
# NLP Task v/s Language Model

3. We train the language model using the data that we got from the internet.



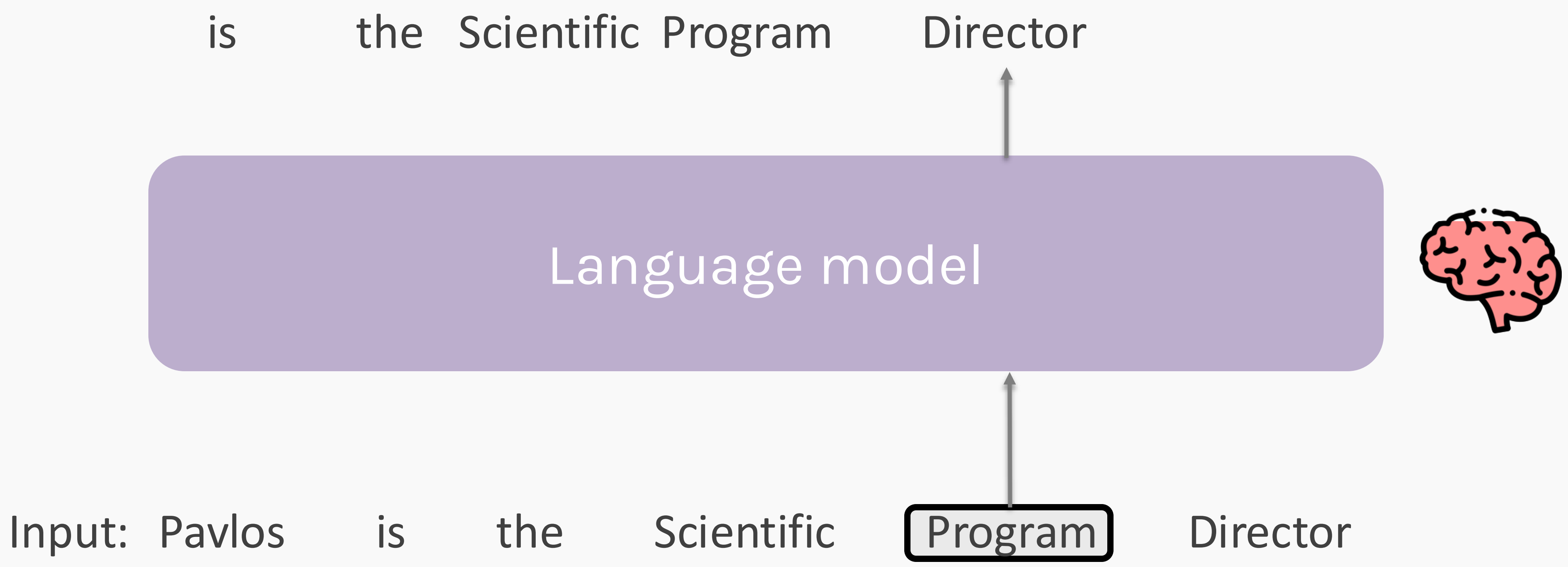
# NLP Task v/s Language Model

3. We train the language model using the data that we got from the internet.



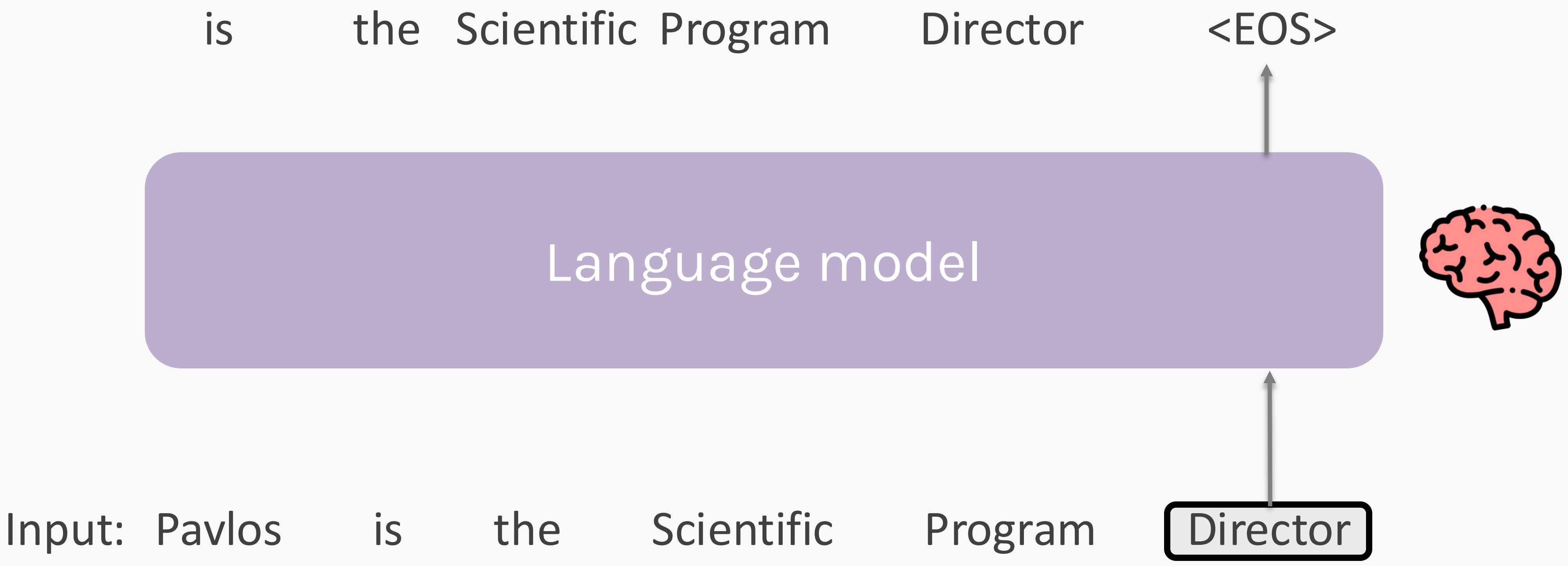
# NLP Task v/s Language Model

3. We train the language model using the data that we got from the internet.



# NLP Task v/s Language Model

3. We train the language model using the data that we got from the internet.



# NLP Task v/s Language Model

4. Finally, we finetune our trained language model to do an NLP task. For eg. Parts of speech tagging.

Notice how this has a label (supervised learning)

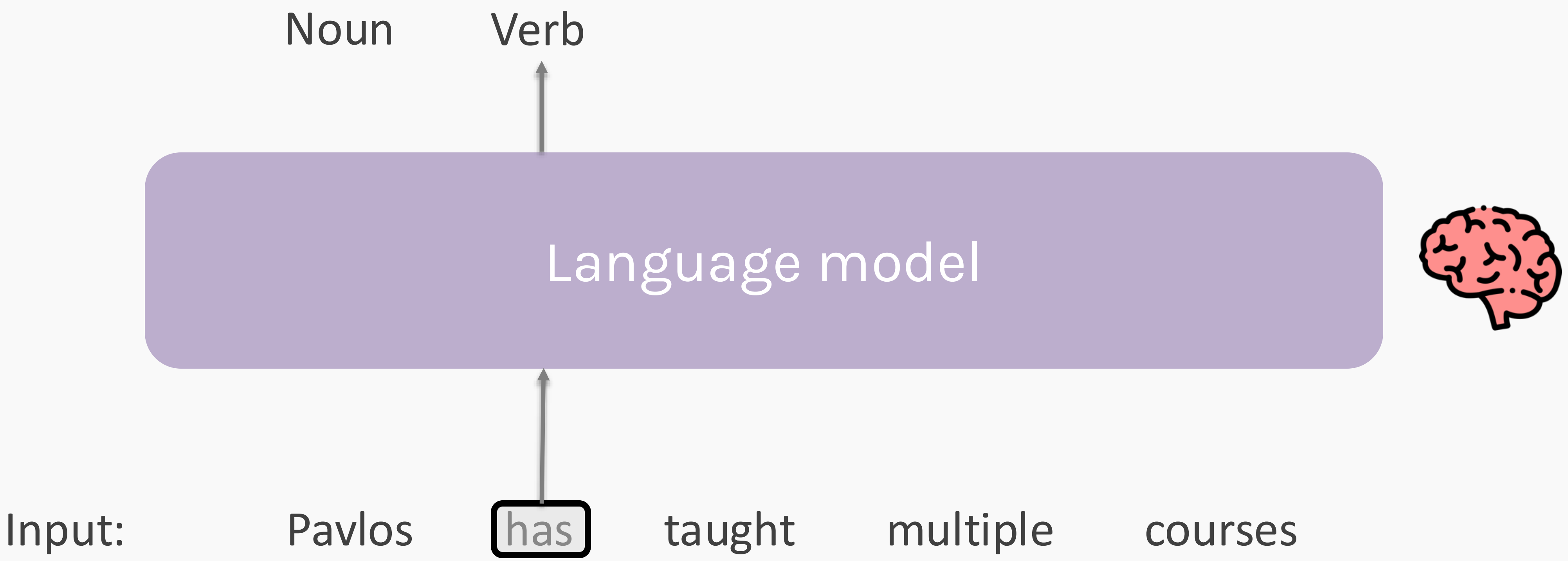
Noun

Language model

Input: Pavlos has taught multiple courses

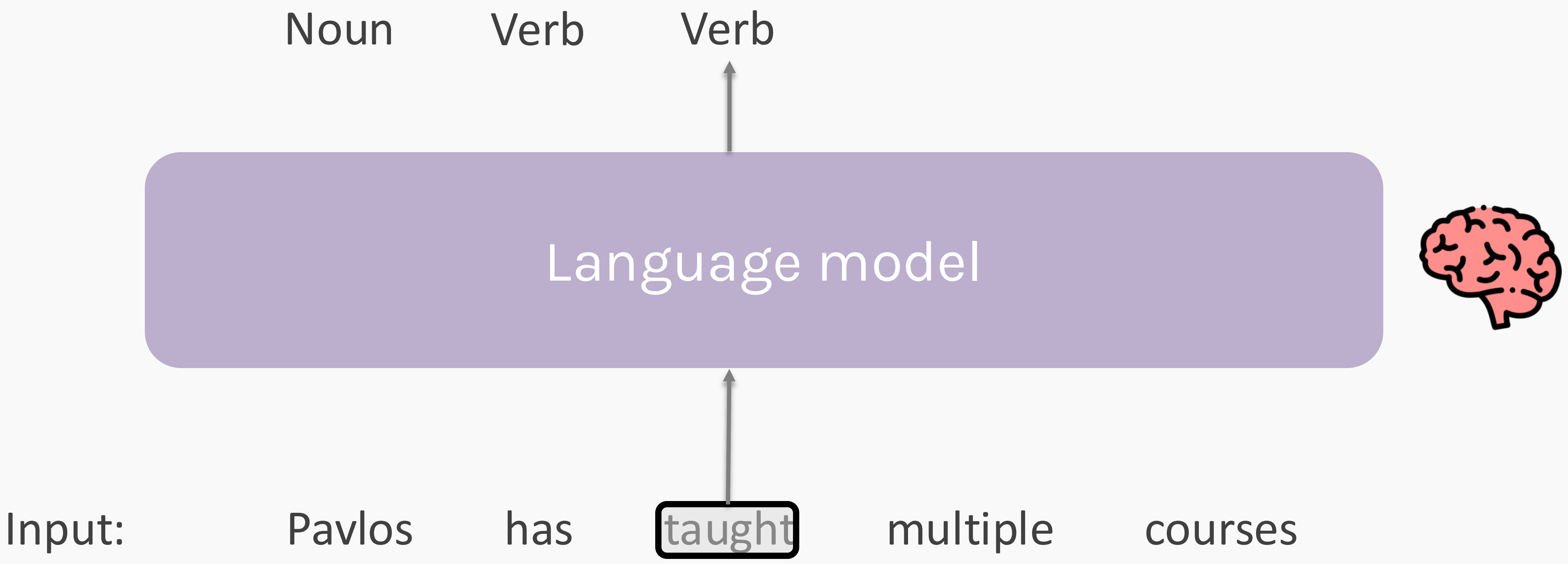
# NLP Task v/s Language Model

4. Finally, we finetune our trained language model to do an NLP task. For eg. Parts of speech tagging.



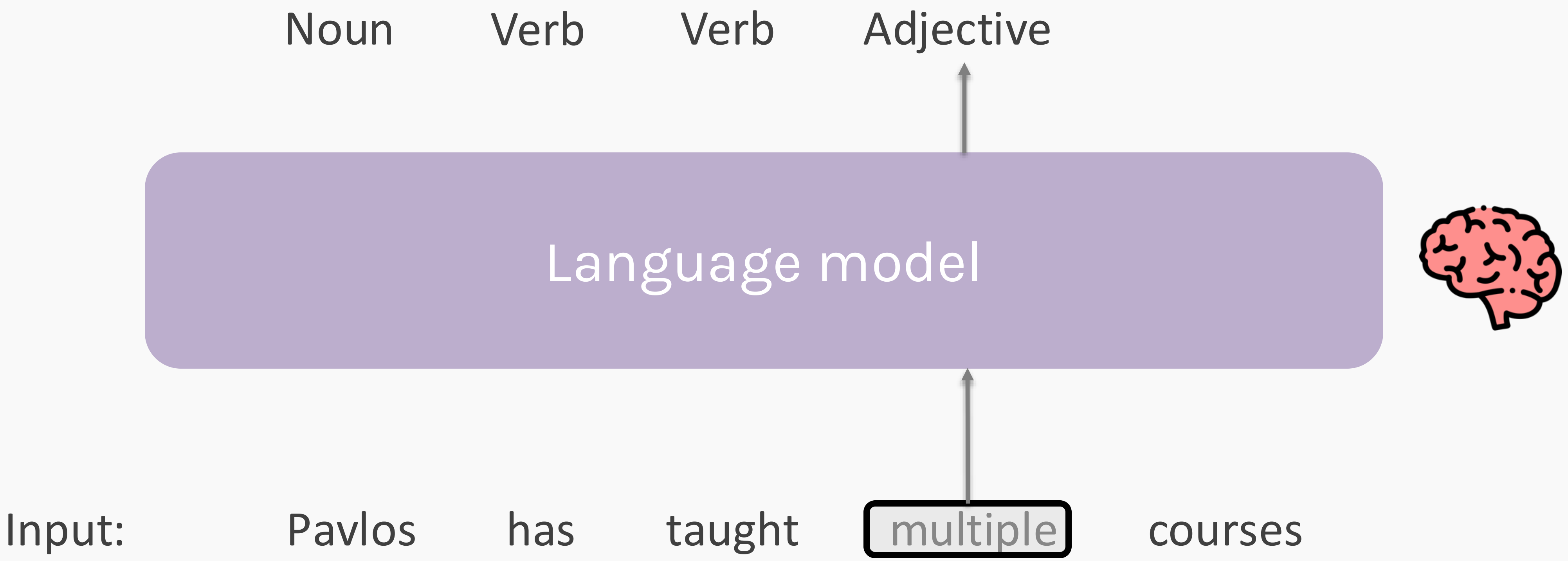
# NLP Task v/s Language Model

4. Finally, we finetune our trained language model to do an NLP task. For eg. Parts of speech tagging.



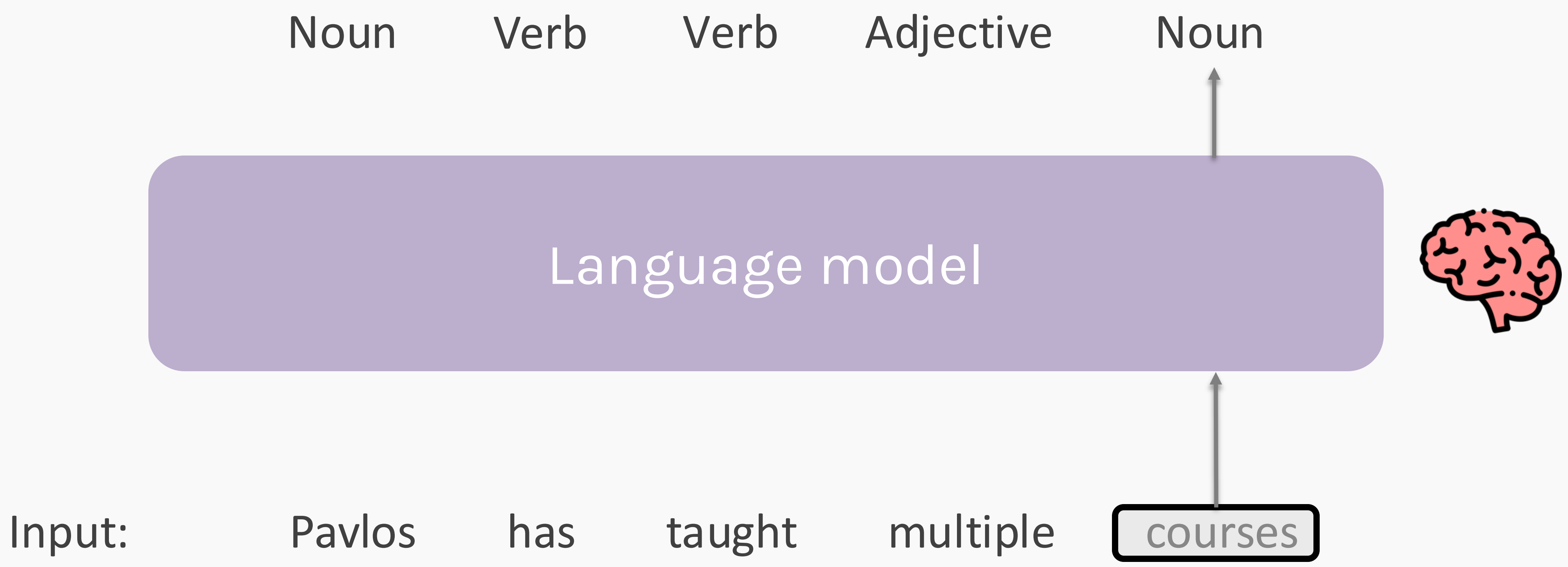
# NLP Task v/s Language Model

4. Finally, we finetune our trained language model to do an NLP task. For eg. Parts of speech tagging.



# NLP Task v/s Language Model

4. Finally, we finetune our trained language model to do an NLP task. For eg. Parts of speech tagging.



# NLP Task v/s Language Model

- Putting it all together:

1. Untrained Language Model

2. Training using a self supervised task

For e.g. - next word prediction

3. Adapt the model to downstream tasks

For e.g. – NER, classification, sentiment analysis, etc.

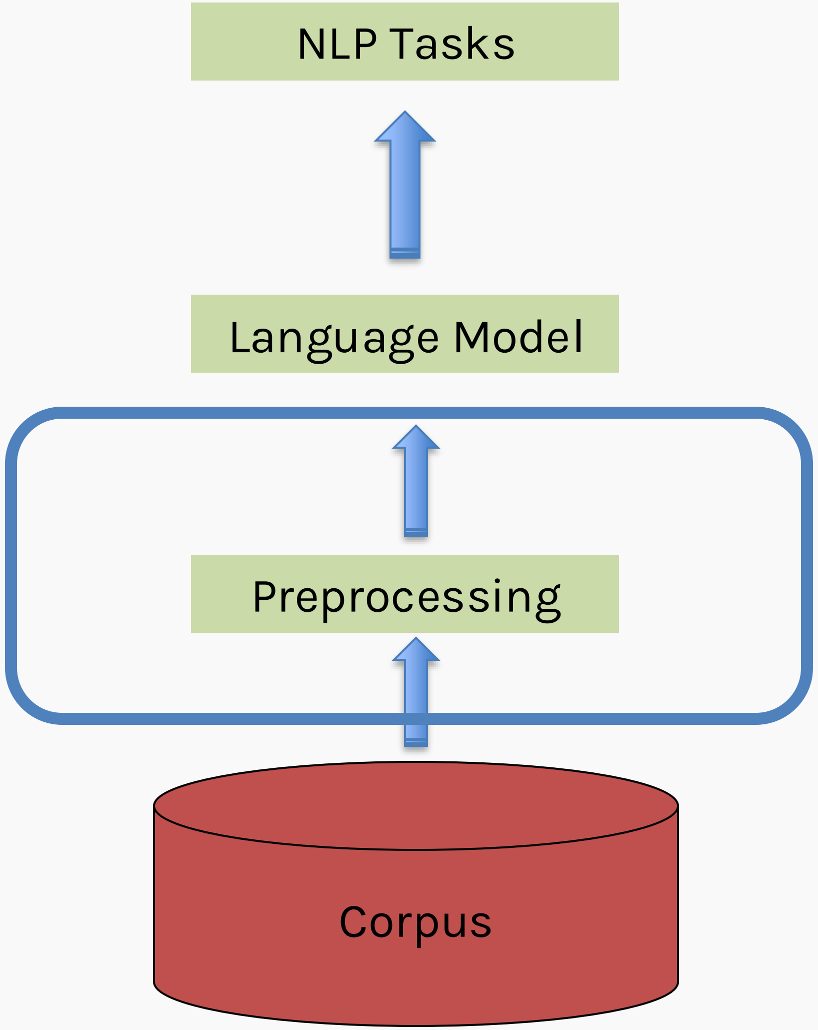
Representation learning!

**Output:** Trained Language Model

NLP Task

# Pipeline

- Now before we look further into language modelling and NLP tasks, let's see how we process our data for the model to be able to understand it.



# Outline

Natural Language Processing

**Text Preprocessing**

Language Modeling

- Unigrams
- Bigrams
- Neural Networks for Language Modeling

# Language Modelling: starting point

- Text is composed of a series of words that carry meaning. To process it, we break down the text into smaller units known as tokens.
- So, a sentence transforms into a sequence of these tokens.
- All the unique tokens from a dataset make up the vocabulary.
- The bigger the vocabulary the bigger the training set required.

# Stemming & Lemmatization

- The goal of both stemming and lemmatization is to reduce inflectional.
- It changes the different forms of a word to a common word.

run

running

runs

run

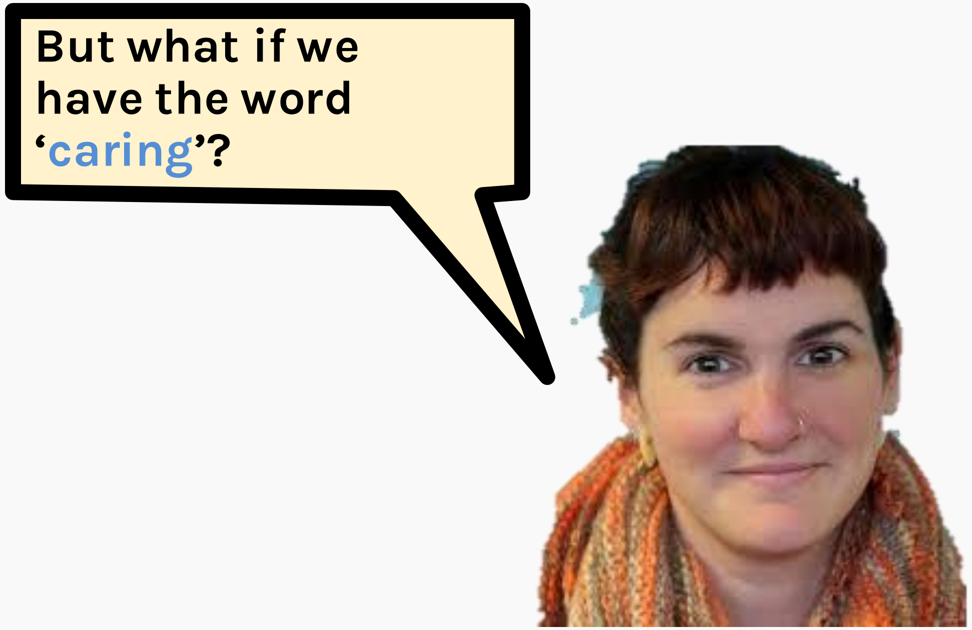
# Stemming

- It is the process of reducing the words to their **roots.**
- **Usually removes the prefixes and suffixes to reduce the word into simpler forms.**

**For eg.**

jumping, jumps, jumped → jump

car, cars, car's, cars' → car



# Lemmatization

- It is similar to stemming, but is context-aware.
- Preserves the semantic meaning of the word.

For eg.

car, car's, cars, cars' $\rightarrow$ car

care, caring, cared $\rightarrow$ care

⚠ It is more **complex** as compared to stemming, and hence **slower**.

# Tokenization

We need to define the basic unit (token) of a sentence.

**First Approach:** whitespace

Split the words on whitespaces only.

# Tokenization

We need to define the basic unit (token) of a sentence.

**First Approach:** whitespace

Split the words on whitespaces only.

“the award-winning actor arrived today.”

# Tokenization

We need to define the basic unit (token) of a sentence.

**First Approach:** whitespace

Split the words on whitespaces only.

| “the | award-winning | actor | arrived | today” |
|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ |

Usually, punctuations are removed before this stage and capitalization is set to lower case

# Tokenization

**First Approach:** whitespace

Split the words on whitespaces only.

| “the | award-winning | actor | arrived | today” |
|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ |

Hyphenated phrases like “*award-winning*” are not split.

| “I | haven’t | seen | her | since | yesterday” |
|---|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ |

Conjunctions such as *haven’t* are not split.

# Tokenization

**Second Approach:** sub-word tokenization

Split the words on statistically significant fragments.

The fragment appears enough times in the corpus

# Tokenization

**Second Approach:** sub-word tokenization

Split the words on statistically significant fragments.

The token loses their direct interpretability but makes a more flexible approach.

# Tokenization

**Second Approach:** sub-word tokenization

Split the words on statistically significant fragments.

# Tokenization

**Second Approach:** sub-word tokenization

Split the words on statistically significant fragments.

How can we define a token?

# Tokenization

**Second Approach:** sub-word tokenization

Split the words on statistically significant fragments.

How can we define a token?

| | | | | | | |
|---|---|---|---|---|---|---|
| “the | award | - | winning | actor | arrived | today” |
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ |

# Tokenization

**Second Approach:** sub-word tokenization

Split the words on statistically significant fragments.

How can we define a token?

| “the | award | - | winning | actor | arrived | today” |
| --- | --- | --- | --- | --- | --- | --- |
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ |

“turn the handle counterclockwise”

# Tokenization

**Second Approach:** sub-word tokenization

Split the words on statistically significant fragments.

How can we define a token?

| “the | award | - | winning | actor | arrived | today” |
|---|---|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ |

| “turn | the | handle | counter | c | lock | wise” |
|---|---|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ |

Tokenization is an area of research in itself.

# Outline

Natural Language Processing

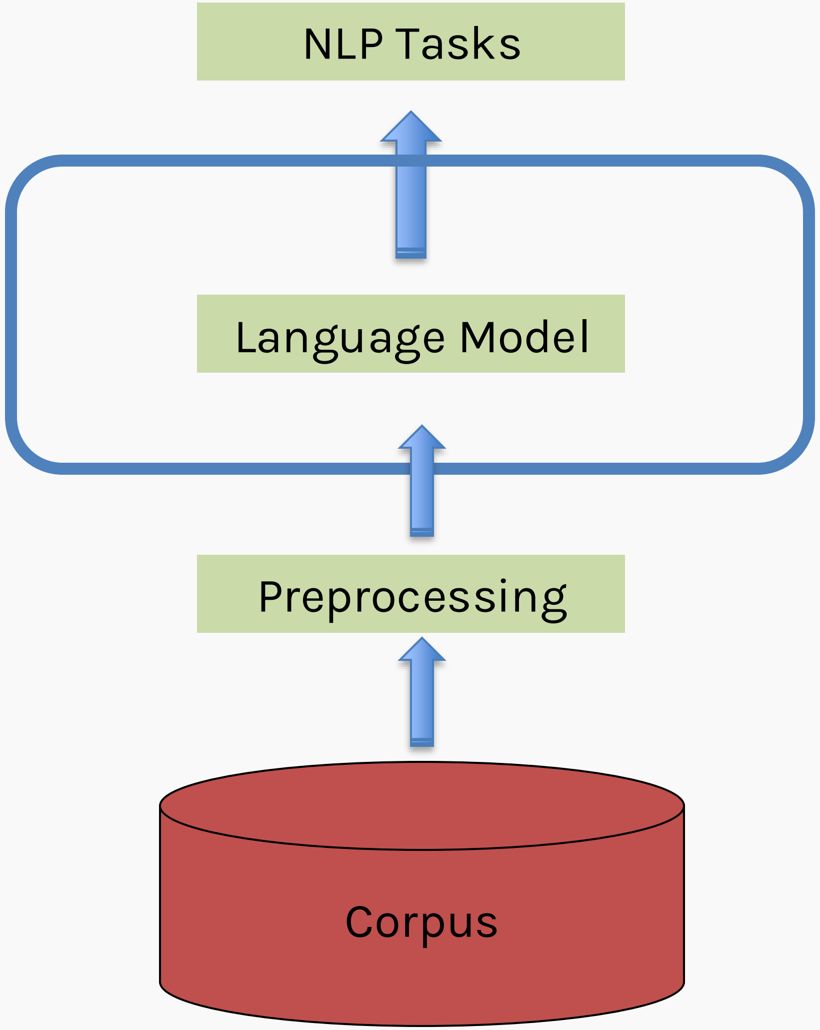
Preprocessing

**Language Modeling**

- Unigrams
- Bigrams
- Neural Networks for Language Modeling

# Pipeline

- Next step in the language modelling



# Language Modelling: Formal Definition

We model any sequential data as follows:

This compounds for all subsequent events, too

$$P(x_1, \dots, x_T) = \prod_{t=2}^{T} p(x_t | x_{t-1}, \dots, x_1)$$

Joint distribution of all measurements

Conditional probability of an event, depends on all of the events that occurred before it.

# Example

If we want to know the probability of the the next on-screen Sesame Street character:

**Scene 1**

**Scene 2**

**Scene 3**

# Example

Remember that, when we evaluate a distribution, we mean

$P($  $,$  $) = P(S_1 =$  $, S_2 =$  $)$

$P($  $,$  $) = P($  $)\,P($  $|$  $)$

# Example

The probability of the the next on-screen Sesame Street character can be computed as

**Scene 1**

**Scene 2**

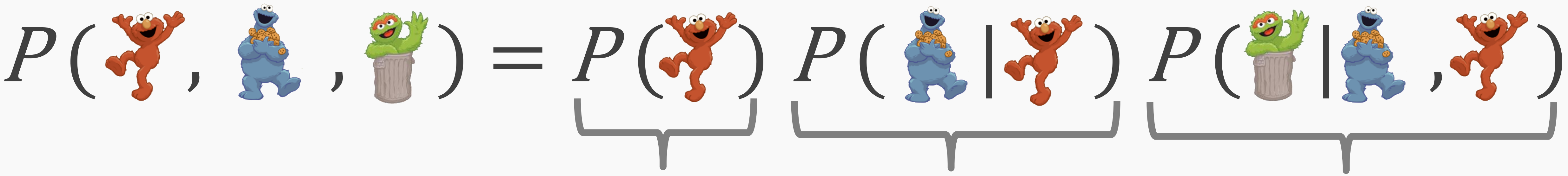
**Scene 3**

$P($  ,  ,  $) =$

# Example

The probability of the the next on-screen Sesame Street character can be computed as

Scene 1 Scene 2 Scene 3



Scene 1 Scene 2 Scene 3

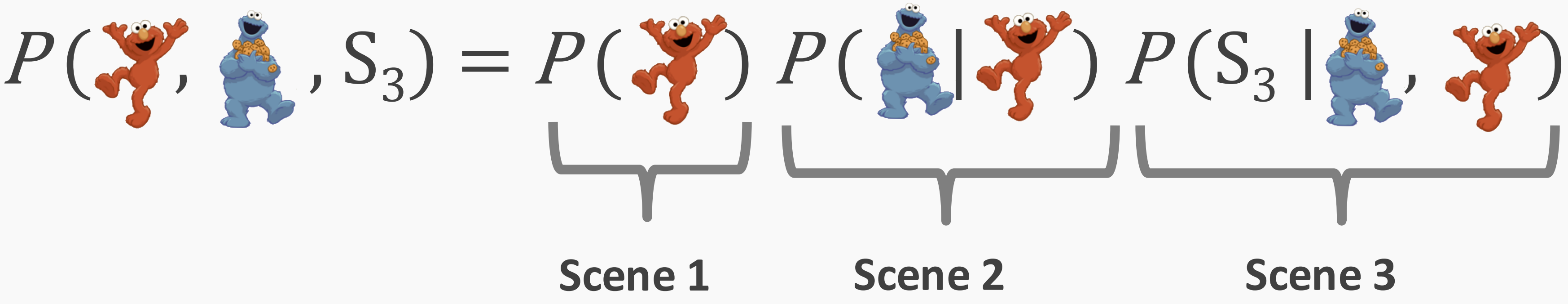
# Example

Why is it useful to accurately estimate the joint probability of any given sequence of length $N$?

# Example

Having learned a Language Model means that we know the behavior of the sequences.

If we have a sequence of length $N$, we can determine the most likely next event (i.e., sequence of length $N + 1$)



# Language Modeling as a sequence of events

A Language Model estimates the probability of any sequence of words

Let $\boldsymbol{X}$ = "Fed was late for class"

$w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5$

$P(\boldsymbol{X})$ = $P$("Fed was late for class")

# Outline

Natural Language Processing

Preprocessing

Language Modeling

- **Unigrams**
- Bigrams
- Neural Networks for Language Modeling

# Language Modelling: Unigrams

$$P(x_1, \dots, x_T) = \prod_{t=1}^{T} p(x_t)$$

How can we build a language model?

**Naive Approach: Unigram model**

Assume each word is independent of all others

Count how often each word occurs (in the training data).

Let $\boldsymbol{X}$ = "Pavlos loves giving surprise quizzes"

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

# Language Modelling: Unigrams

Consider our corpus $d$ has 100,000 words.

| Word | Number of occurrences |
|---|---|
| Pavlos | 15 |
| loves | 1,000 |
| giving | 400 |
| surprise | 3,000 |
| quizzes | 350 |

$n_{w_i}$ = number of times a word $w_i$ appears in the corpus

$|W|$ = Total Number of words, 100000 (corpus size)

Let $\boldsymbol{X}$ = "Pavlos loves giving surprise quizzes"

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

$$P(w_i) = \frac{n_{w_i}(d)}{|W|}$$

$$P(w_1) = P(Pavlos) = \frac{n_{w_1}(d)}{|W|} = \frac{15}{100{,}000} = 0.00015$$

# Language Modelling: Unigrams

Consider our corpus $d$ has 100,000 words.

| Word | Number of occurrences |
|---|---|
| Pavlos | 15 |
| loves | 1,000 |
| giving | 400 |
| surprise | 3,000 |
| quizzes | 350 |

$n_{w_i}$ = number of times a word $w_i$ appears in the corpus

$|W|$ = Total Number of words, 100000 (corpus size)

Let $\boldsymbol{X}$ = "Pavlos loves giving surprise quizzes"

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

What is the dimension of $P(w_i)$?

$$P(w_i) = \frac{n_{w_i}(d)}{|W|}$$

$$P(w_2) = P(loves) = \frac{n_{w_2}(d)}{|W|} = \frac{1{,}000}{100{,}000} = 0.01$$

# Language Modelling: Unigrams

How can we build a language model?

**Naive Approach: Unigram model**

Assume each word is independent of all others

You calculate and store each of these probabilities from the training corpus

Let $\boldsymbol{X}$ = "Pavlos loves giving surprise quizzes"

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

$\mathrm{P}(\boldsymbol{X}) = P(\text{Pavlos})\, P(\text{loves})\, P(\text{giving})\, P(\text{surprise})\, P(\text{quizzes})$

$= 0.00015 * 0.01 * 0.004 * 0.03 * 0.0035$

$= 6.3\text{x}10^{-13}$

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t)$$

# Language Modelling: Unigram issues

**Context doesn't play a role at all**

$$P(\text{"Pavlos loves giving surprise quizzes"}) = P(\text{"quizzes loves giving surprise Pavlos"})$$

**Sequence generation: What's the most likely next word?**

Pavlos loves giving surprise quizzes all ___

Pavlos loves giving surprise quizzes all the

Pavlos loves giving surprise quizzes all the the

**Out of vocabulary words**

$$P(\text{"hungry"}) = 0$$

# Language Modelling: Unigram issues

## Out of vocabulary words

$$P(\text{“hungry”}) = 0$$

**Solution: Additive** Smoothing

$$P(\mathrm{w}) = \frac{n_w(d)}{|W|} = \frac{n_w(d) + \alpha}{|W| + \alpha\,|V|}$$

$\boldsymbol{\alpha}$ values are usually small: 0.5 – 0.2

|V| is the number of unique words in the training corpus – vocabulary size - including an additional token for unknown words

Whenever a word $w$ is not found in the vocabulary it is replaced with a token <UNK> representing unknown

# Language Modelling: Unigram issues

## Before Smoothing :

$$P(\text{“hungry”}) = 0$$

## After Smoothing :

$$P(\text{"UNK"}) = \frac{\alpha}{|W| + \alpha\, |V|} > 0$$

Smoothing allows probability of “UNK” token to be non-zero enabling the model to predict Out of Vocabulary words.

# Language Modelling: Unigram issues

## Context doesn't play a role at all

$$P(\text{“Pavlos loves giving surprise quizzes”}) = P(\text{“quizzes loves giving surprise Pavlos”})$$

## Sequence generation: What's the most likely next word?

Pavlos loves giving surprise quizzes all ___

Pavlos loves giving surprise quizzes all <u>the</u>

Pavlos loves giving surprise quizzes all the <u>the</u>

## Out of vocabulary words

$$P(\text{“hungry”}) = 0$$

# Outline

Natural Language Processing

Preprocessing

Language Modeling

- Unigrams
- **Bigrams**
- Neural Networks for Language Modeling

# Language Modelling: Bigrams

How can we build a language model that uses context?

**Easiest Approach: bigram model**

Look at *pairs* of consecutive words

$$P(x_1, \ldots, x_T) = p(x_1)\prod_{t=2}^{T} p(x_t|x_{t-1})$$

| probability |
|---|
| " Pavlos loves |

Let $\boldsymbol{X}$ = " Pavlos loves giving surprise quizzes "

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

$\mathrm{P}(\boldsymbol{X}) = P(\text{loves}|\text{Pavlos})$

# Language Modelling: Bigrams

How can we build a language model that uses context?

**Easiest Approach: bigram model**

Look at *pairs* of consecutive words

$$P(x_1, \dots, x_T) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1})$$

probability

Let $\boldsymbol{X}$ = " Pavlos loves giving surprise quizzes "

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

$\mathrm{P}(\boldsymbol{X})$ = $P$(loves | Pavlos) $P$(giving | loves)

# Language Modelling: Bigrams

How can we build a language model that uses context?

**Easiest Approach: bigram model**

Look at *pairs* of consecutive words

$$P(x_1, \ldots, x_T) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1})$$

**probability**

Let $\boldsymbol{X}$ = " Pavlos loves giving surprise quizzes "

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

P($\boldsymbol{X}$) = $P$(loves | Pavlos) $P$(giving | loves) $P$(surprise | giving)

# Language Modelling: Bigrams

How can we build a language model that uses context?

**Easiest Approach:** **bigram model**

Look at *pairs* of consecutive words

$$P(x_1, \dots, x_T) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1})$$

**probability**

Let $\boldsymbol{X}$ = " Pavlos loves giving surprise quizzes "

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

P($\boldsymbol{X}$) = $P$(loves | Pavlos) $P$(giving | loves) $P$(surprise | giving) $P$(quizzes | surprise)

# Language Modelling: Bigrams

You calculate each of these probabilities by simply counting the occurrences:

$$P(quizzes|surprise) = \frac{C(W_1W_2)}{C(W_1w)}$$

| probability |
|---|
| |

Let $\boldsymbol{X}$ = " Pavlos loves giving surprise quizzes "

$w_1$ $w_2$ $w_3$ $w_4$ $w_5$

P($\boldsymbol{X}$) = $P$(loves | Pavlos) $P$(giving | loves) $P$(surprise | giving) $P$(quizzes | surprise)

# Language Modelling: Bigram issues

- When a word is out-of-vocabulary, it's given a probability of 0, causing the whole sentence or sequence to also have a probability of 0
- More context (like trigrams, 4-grams) is often desired, but sparsity is a challenge due to infrequent subsequences.
- As we expand the window size for context, storage issues arise.
- Raw counts fail to convey deep semantic relationships, such as the similarity between 'vehicle' and 'car'.

# Outline

Natural Language Processing

Preprocessing

Language Modeling

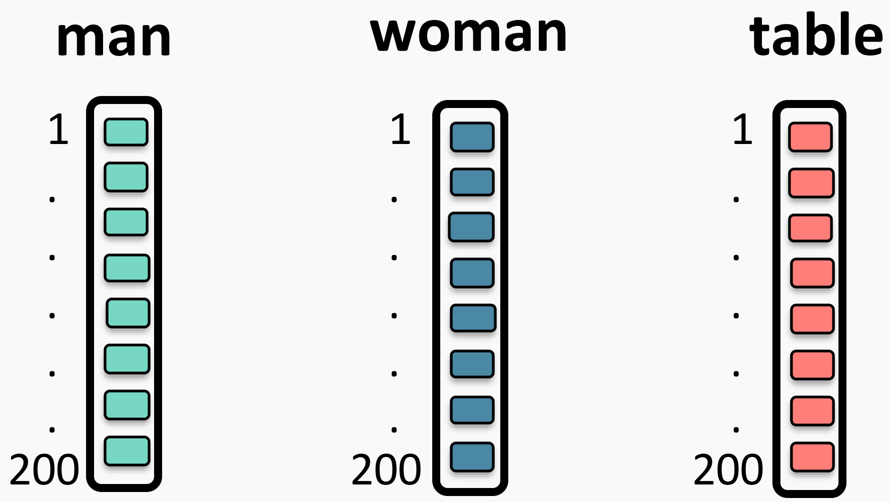
- Unigrams
- Bigrams
- **Neural Networks for Language Modeling**

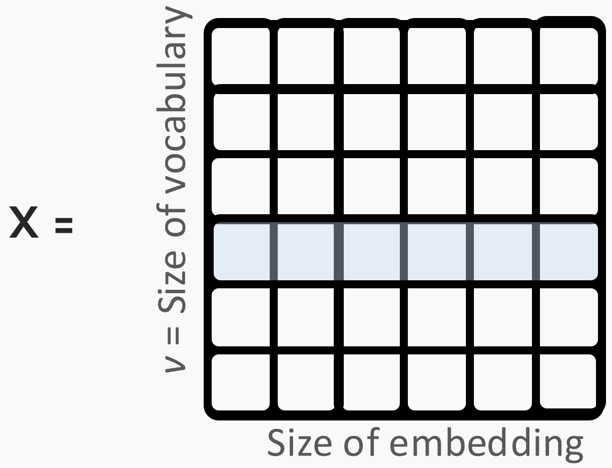
# Language Modelling: Neural Networks

**IDEA:** Let's use a **neural network**!

First, each word is represented by a word embedding (e.g., vector of length 200)

Embedding matrix

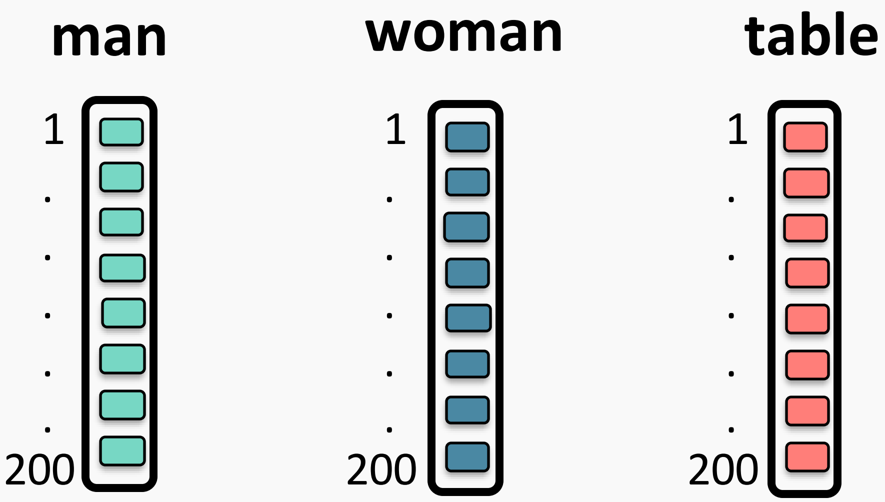


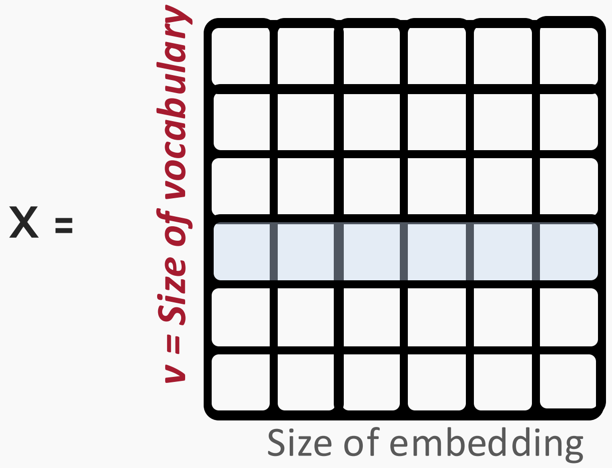
# Language Modelling: **Neural Networks**

**IDEA:** Let's use a **neural network**!

First, each word is represented by a word embedding (e.g., vector of length 200)



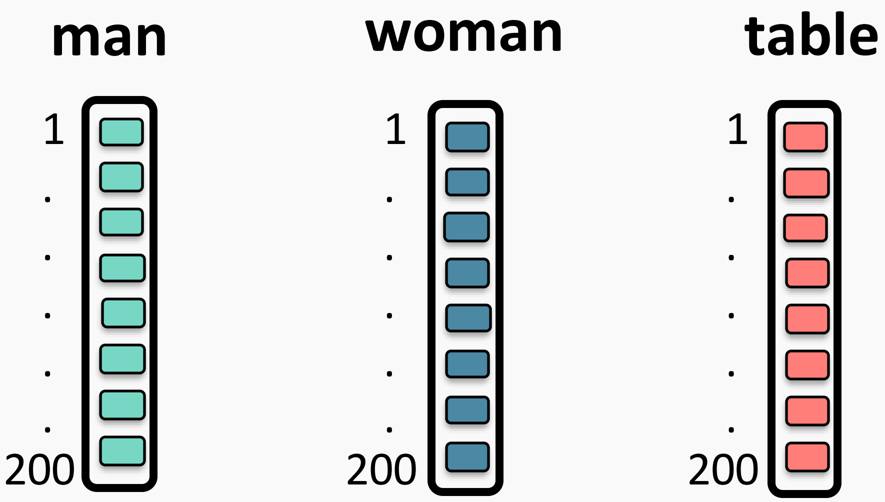

Embedding matrix



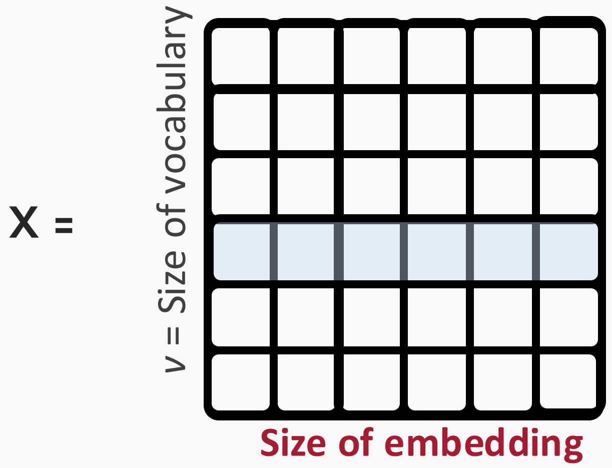
# Language Modelling: **Neural Networks**

**IDEA:** Let's use a **neural network**!

First, each word is represented by a word embedding (e.g., vector of length 200)



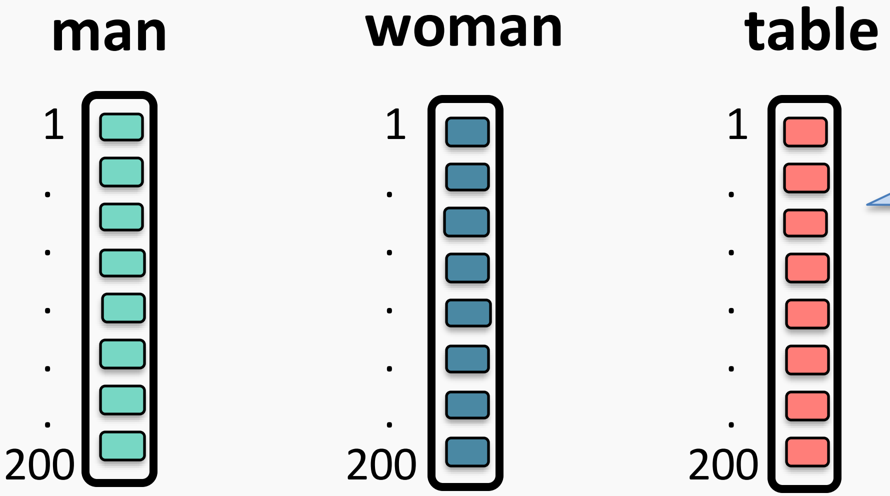

Embedding matrix



# Language Modelling: **Neural Networks**

**IDEA:** Let's use a **neural network**!

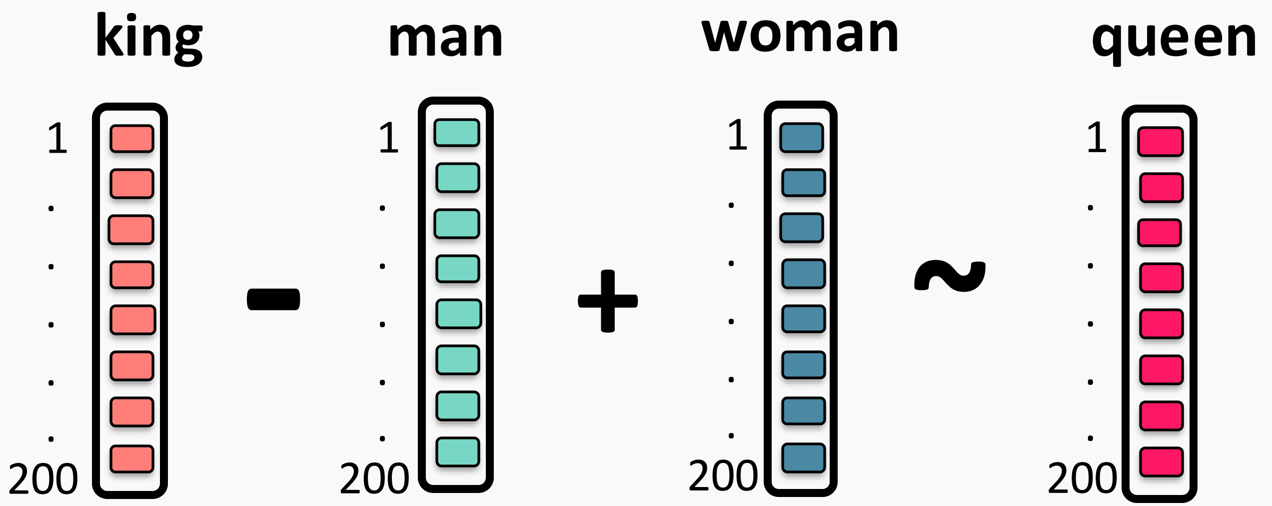
First, each word is represented by a word embedding (e.g., vector of length 200)




- Each rectangle is a *floating-point* scalar
- Words that are more *semantically similar* to one another will have **embeddings** that are also proportionally similar
- We can *use pre-existing* word embeddings that have been trained on gigantic corpora

# Word Embeddings

These word embeddings are so rich that you get nice properties:

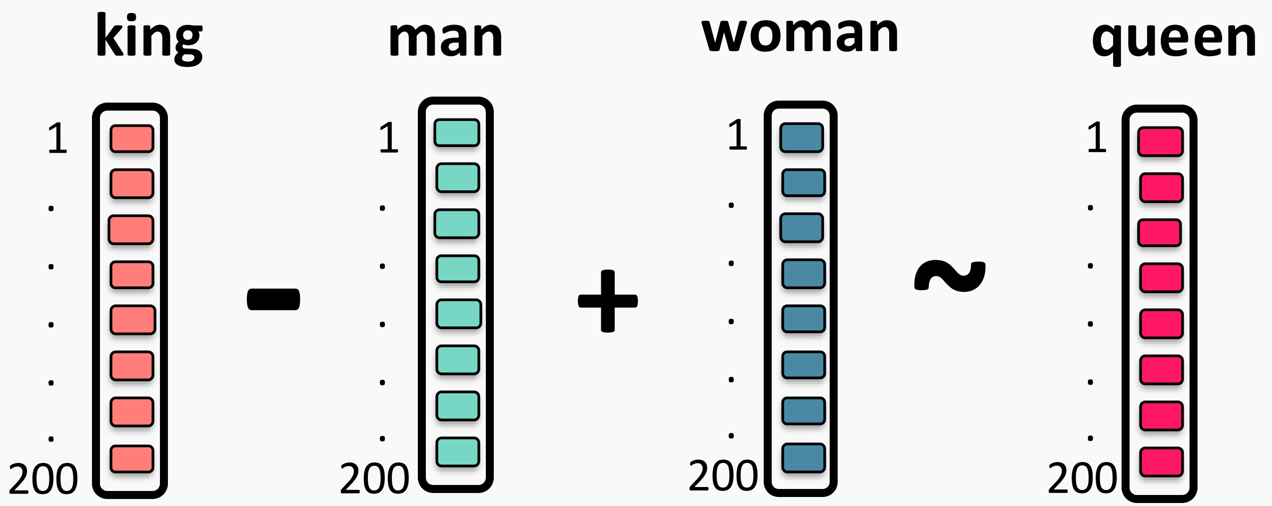





Word2vec: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
GloVe: https://www.aclweb.org/anthology/D14-1162.pdf

# Word Embeddings

These word embeddings are so rich that you get nice properties:




Word2Vec a popular word embedding

Word2vec: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
GloVe: https://www.aclweb.org/anthology/D14-1162.pdf

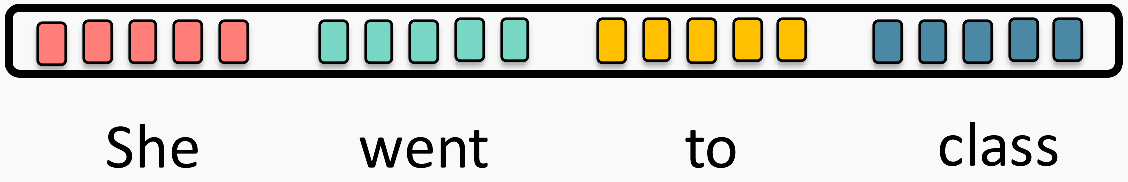
# Language Modelling: Neural Networks

How can we use these embeddings to build a Language Model?

Remember, we only need a system that can estimate:

$$P(\underbrace{x_{t+1}}_{\text{next word}} \mid \underbrace{x_t, x_{t-1}, \dots, x_1}_{\text{previous words}})$$

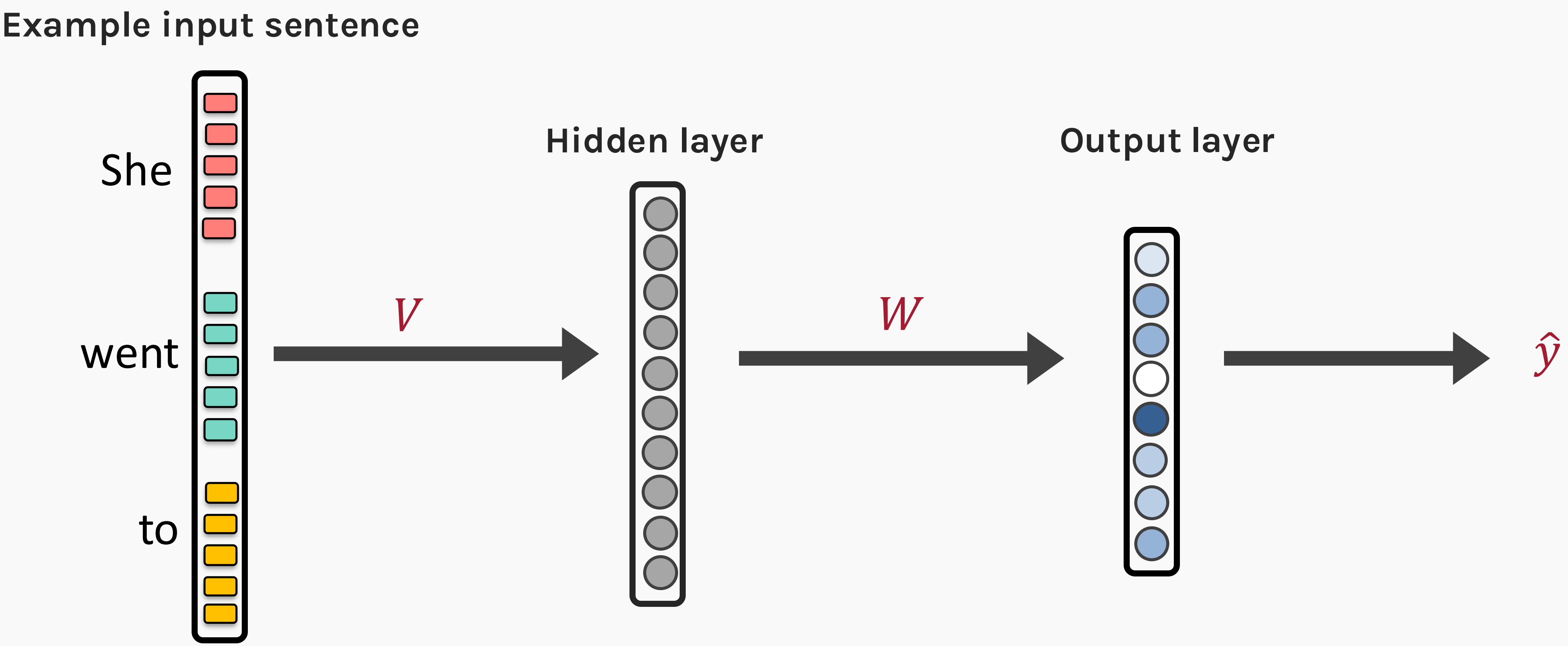
**Example input sentence**



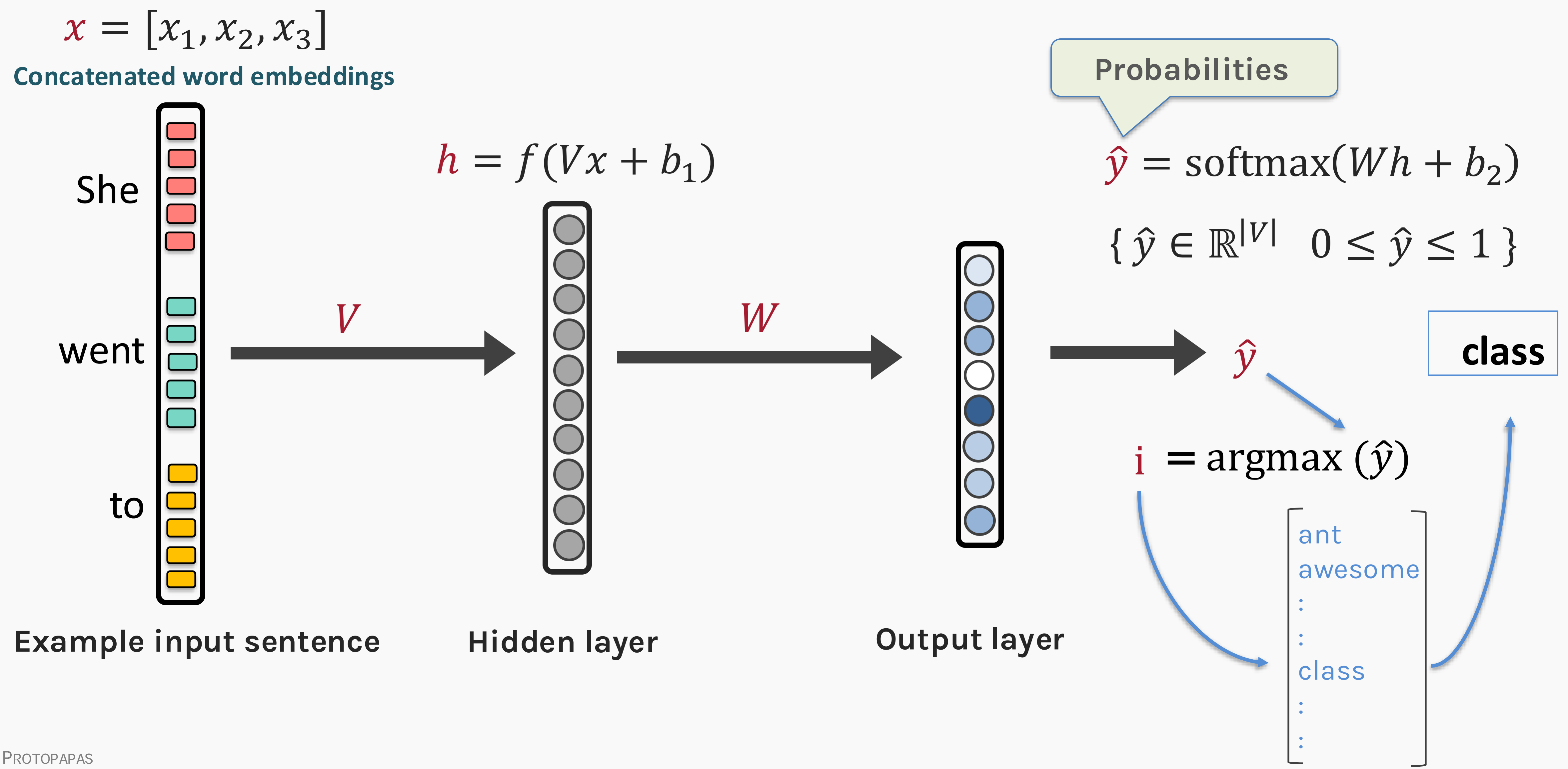
# Language Modelling: Feed-Forward Neural Network

**Neural Approach #1:** Feed-forward neural net

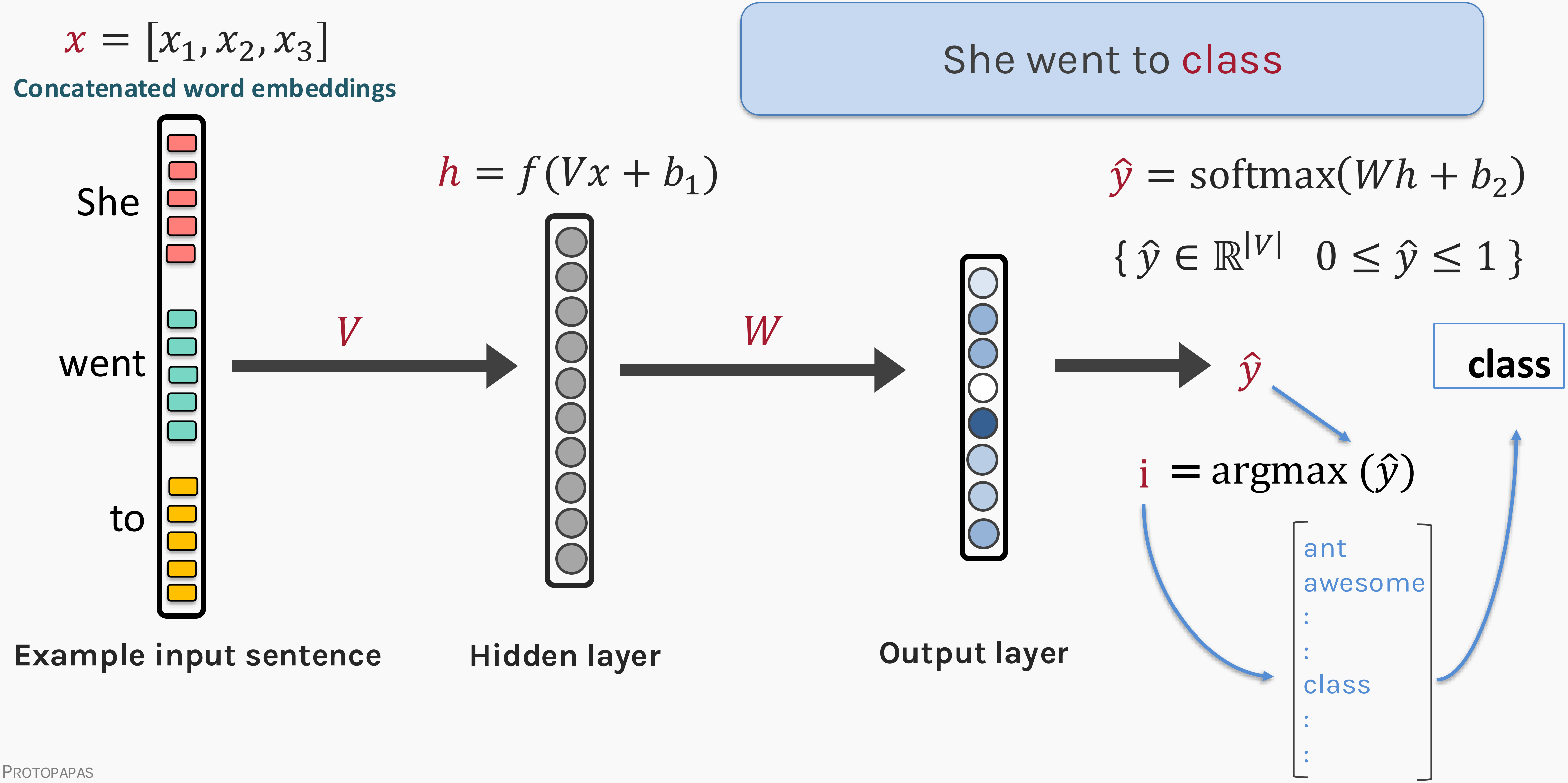
**General Idea:** using *windows* of words, predict the next word



# Language Modelling: Feed-Forward Neural Network



# Language Modelling: Feed-Forward Neural Network



# Language Modelling: Feed-Forward Neural Network

$$x = [x_2, x_3, x_4]$$

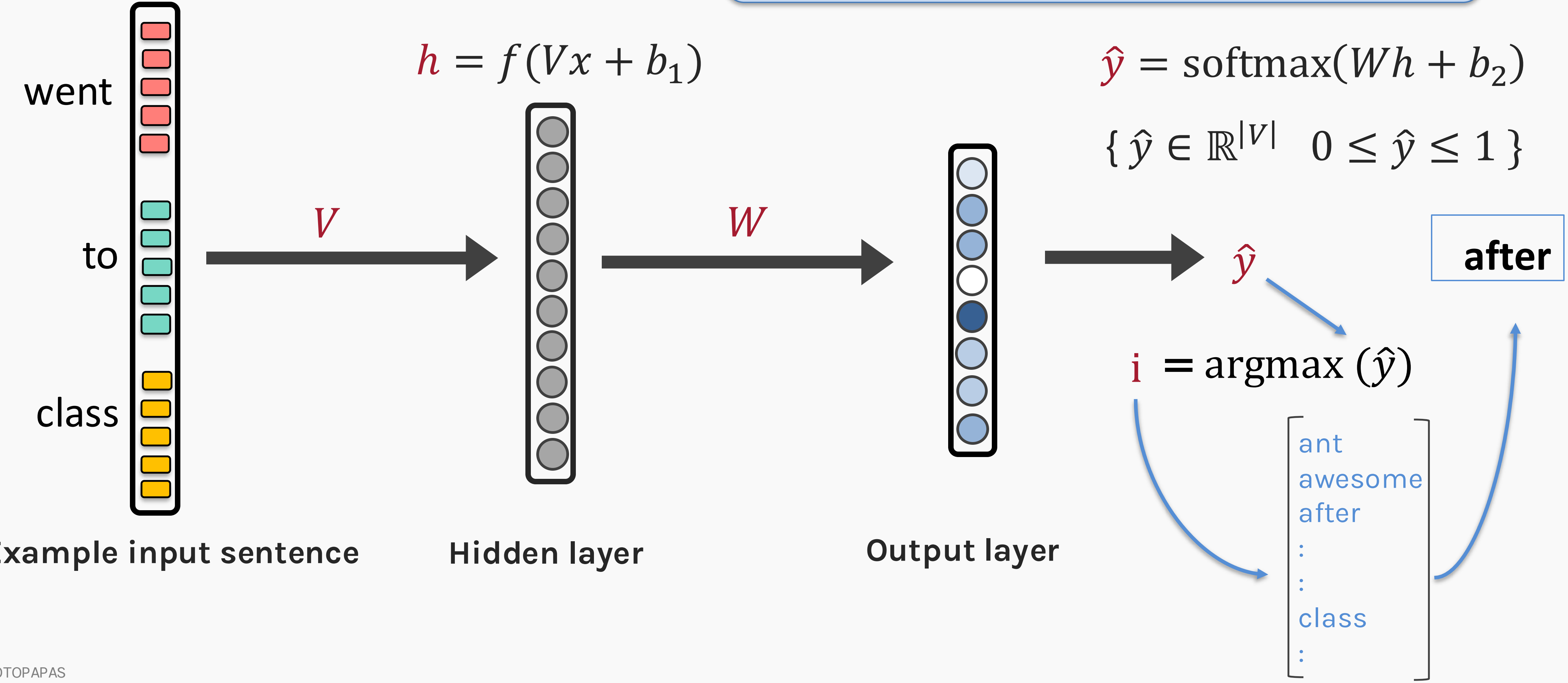
**Concatenated word embeddings**

She went to class after

$$h = f(Vx + b_1)$$

$$\hat{y} = \text{softmax}(Wh + b_2)$$

$$\{\, \hat{y} \in \mathbb{R}^{|V|} \quad 0 \le \hat{y} \le 1 \,\}$$




**Example input sentence** **Hidden layer** **Output layer**

# Language Modelling: Feed-Forward Neural Network

$x = [x_3, x_4, x_5]$

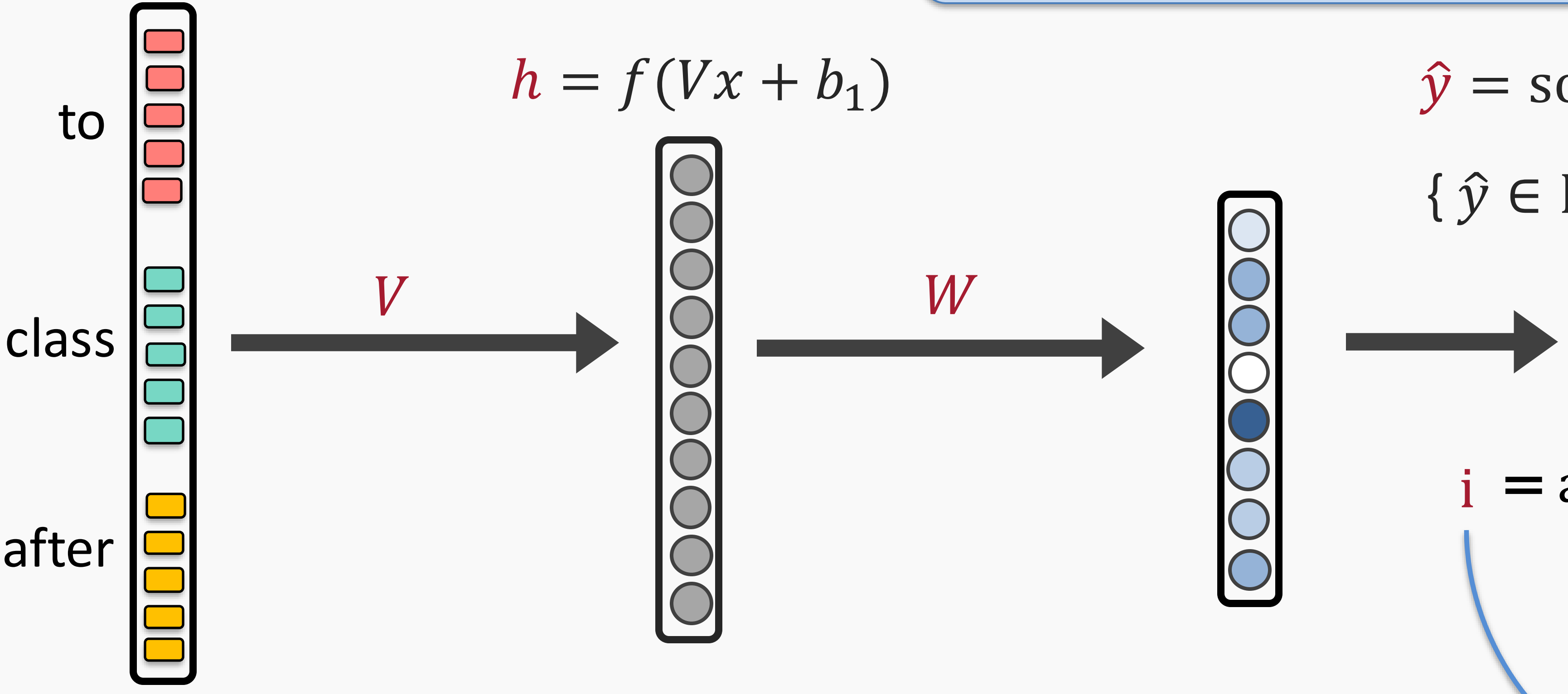
**Concatenated word embeddings**

She went to class after visiting

$h = f(Vx + b_1)$

$\hat{y} = \text{softmax}(Wh + b_2)$

$\{\, \hat{y} \in \mathbb{R}^{|V|} \quad 0 \leq \hat{y} \leq 1 \,\}$




$\hat{y}$

**visiting**

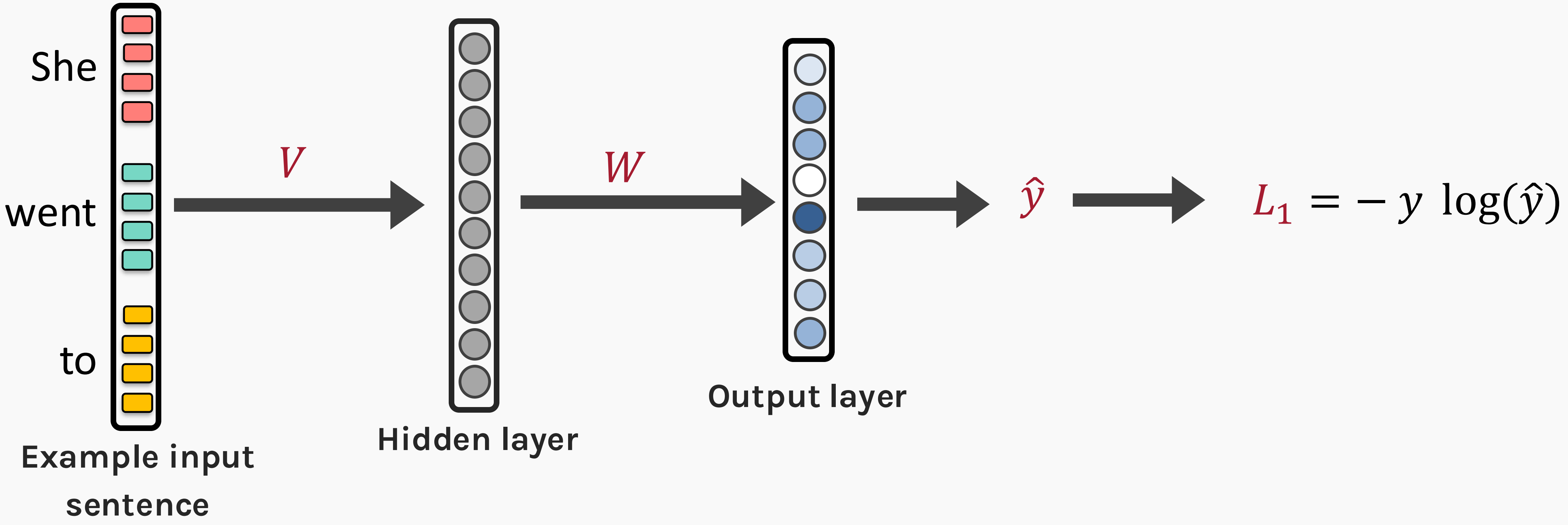
$i = \text{argmax}\,(\hat{y})$

$$\begin{bmatrix} \text{ant} \\ \text{awesome} \\ \vdots \\ \vdots \\ \vdots \\ \text{class} \\ \text{visiting} \end{bmatrix}$$

**Example input sentence** **Hidden layer** **Output layer**

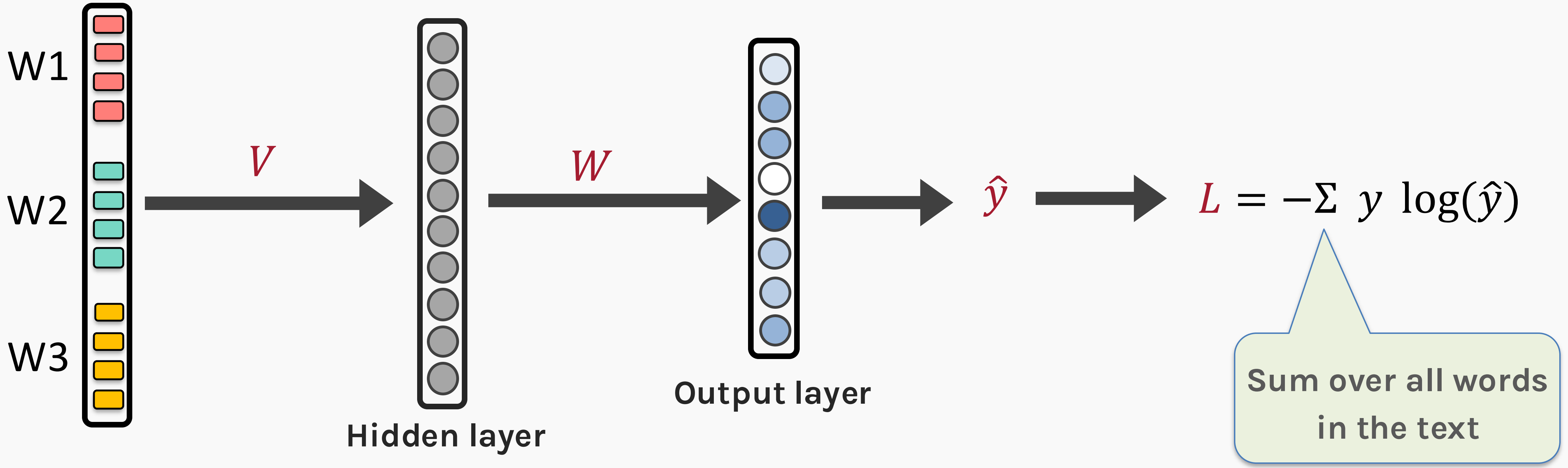
# Language Modelling: Feed-Forward Neural Network Training

$x = [x_1, x_2, x_3]$ $\quad h = f(Vx + b_1)$ $\quad \hat{y} = \text{softmax}(Wh + b_2) \in \mathbb{R}^{|V|}$



# Language Modelling: Feed-Forward Neural Network Training

$x = [x_i, x_{i+1}, x_{i+2}]$ $\quad h = f(Vx + b_1)$ $\quad \hat{y} = \text{softmax}(Wh + b_2) \in \mathbb{R}^{|V|}$




## Back Propagation

$$V^* = V - \eta \nabla_V L$$

$$W^* = W - \eta \nabla_W L$$

# Language Modelling : Feed-Forward Neural Network

## FFNN Strength

- No sparsity issues (it's okay if we've never seen a word)
- No storage issues (we never store counts)

# Language Modelling : Feed-Forward Neural Network

## FFNN Strength

- No sparsity issues (it's okay if we've never seen a word)
- No storage issues (we never store counts)

## FFNN Issues

- Fixed-window size can never be big enough. Need more context
  - Requires inputting entire context just to predict one word
  - Increasing window size adds many more weights
- The weights awkwardly handle word position
- No concept of time

# Language Modelling

**We especially need a system that:**

- Has a concept of an “infinite” past, not just a fixed window
- For each new input, output the most likely next event (e.g., word)

THANK YOU