# RNN & Transformers

## NEU365

Jianing Mu 04/24/2025

# Time series prediction

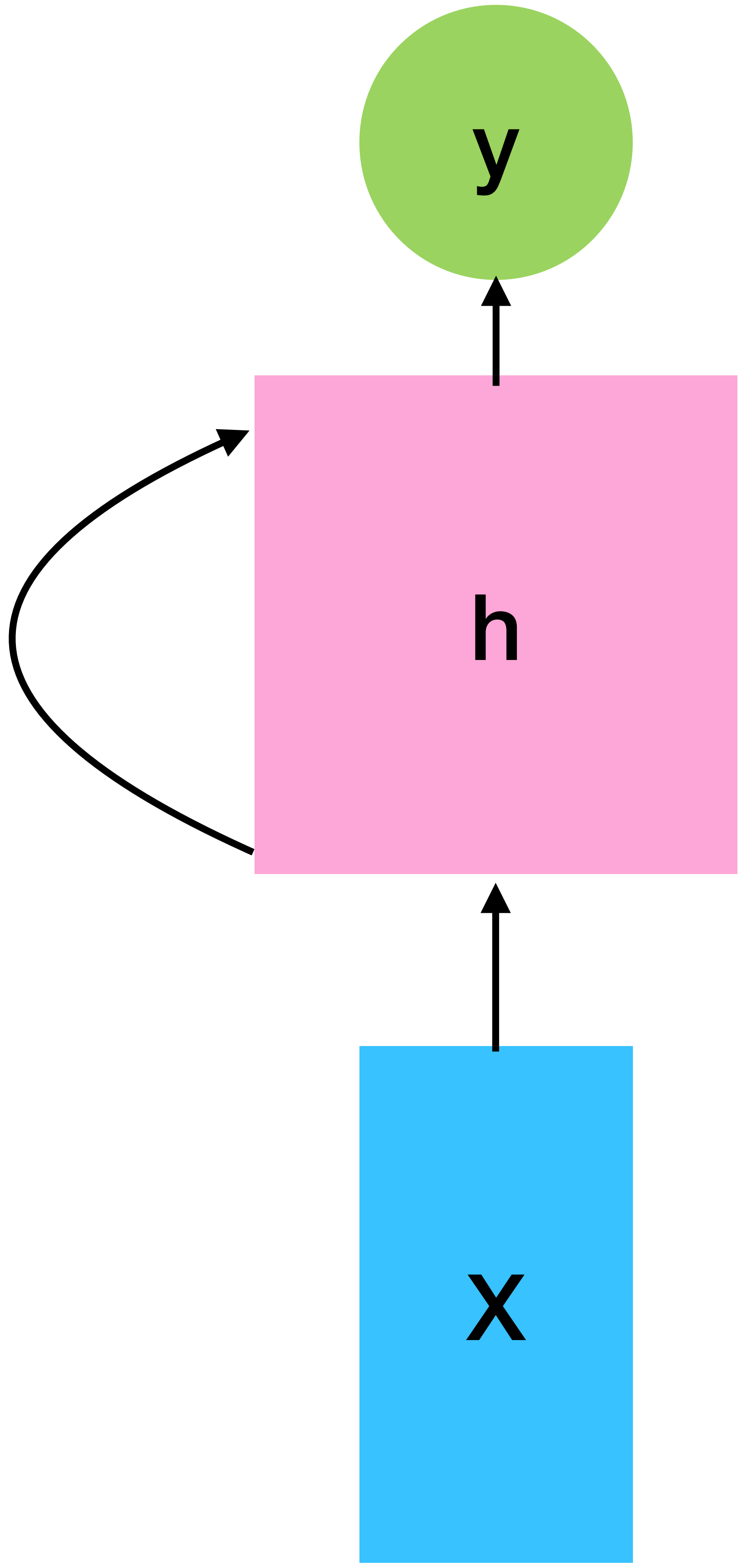
- Suppose you're predicting the **stock price (y)** of NVDA tomorrow (t+1), given input features at that time (general economy status, news on NVDIA GPU demand etc.)

# Time series prediction

- Suppose you pack all input features at time t into a vector **x(t)**
- But temporal dependency can go far back in time - what if yesteday's price is useful to predict today's price? What if last week's price is useful to predict today's price?
- We don't know - we use a neural network to learn this dependency

$$y_t = f(x_t, x_{t-1}, x_{t-2}, \ldots)$$

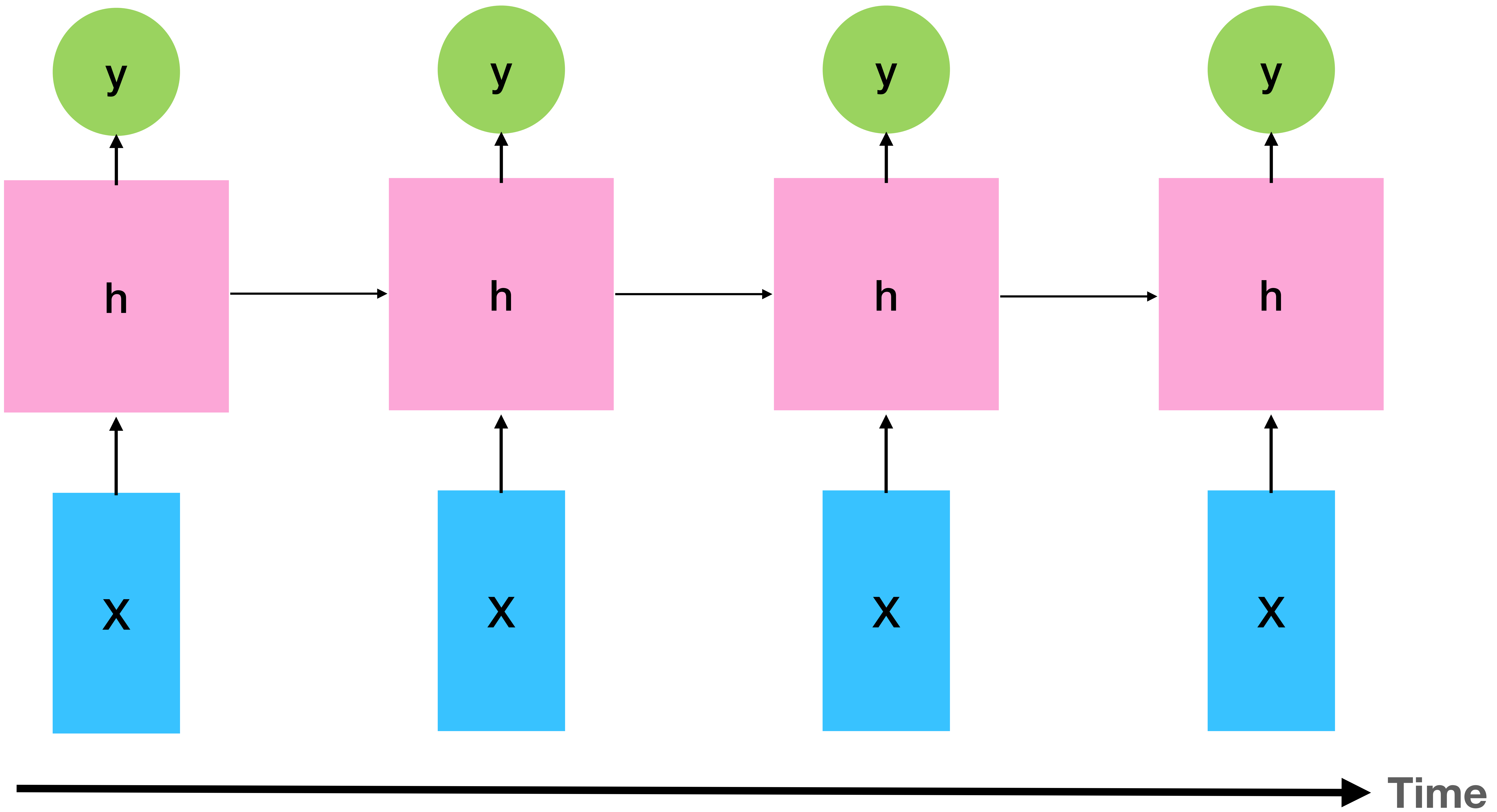
# Recurrent neural network (RNN)




- Output **y** is a function of hidden state **h**
- $y_t = f(h_t)$
- Hidden state h is a function of the current input **x** and hidden state from the last time point
- $h_t = g(x_t, h_{t-1})$

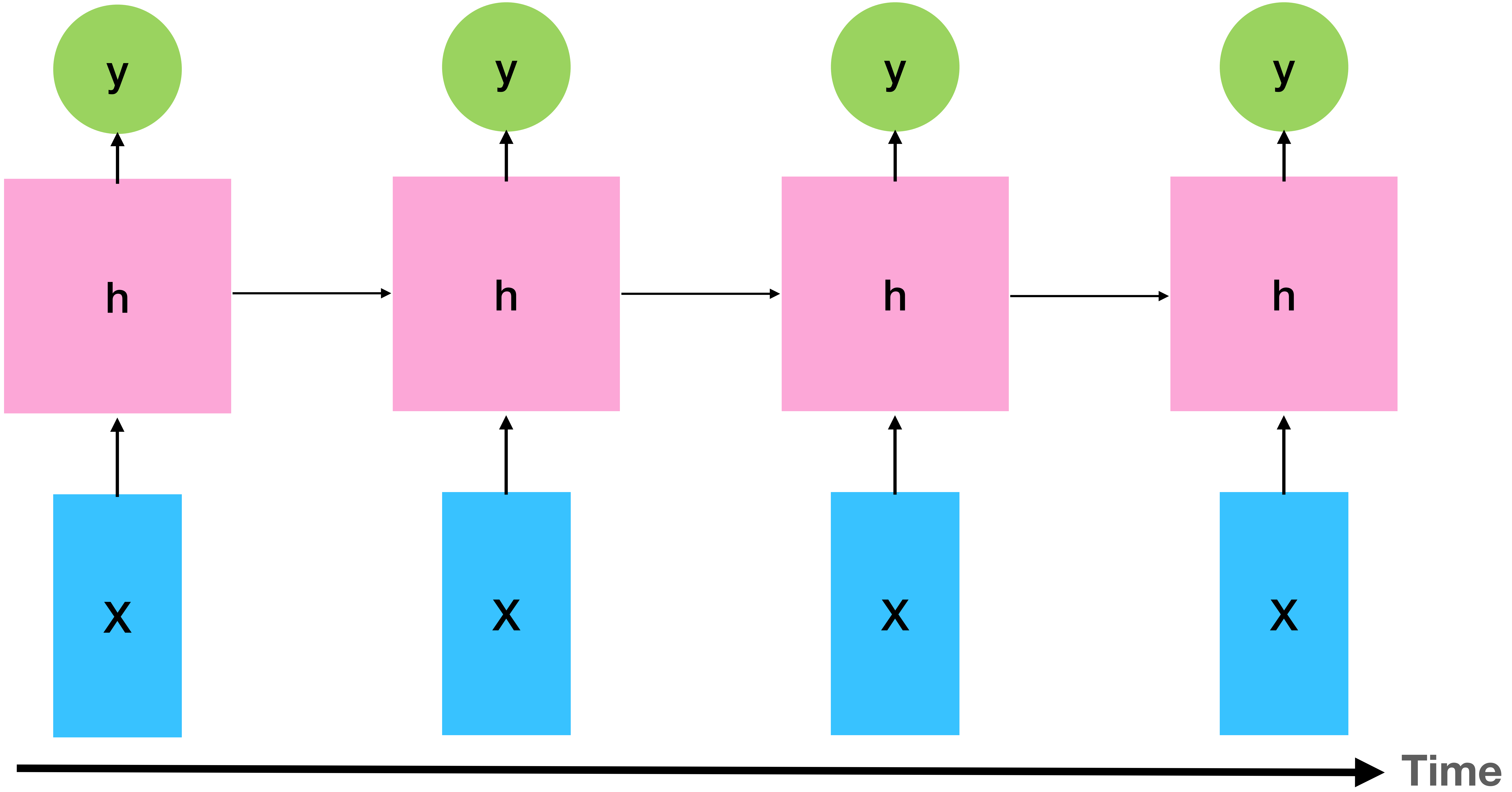
# RNN

## How does information flow in time? Unroll an RNN in time



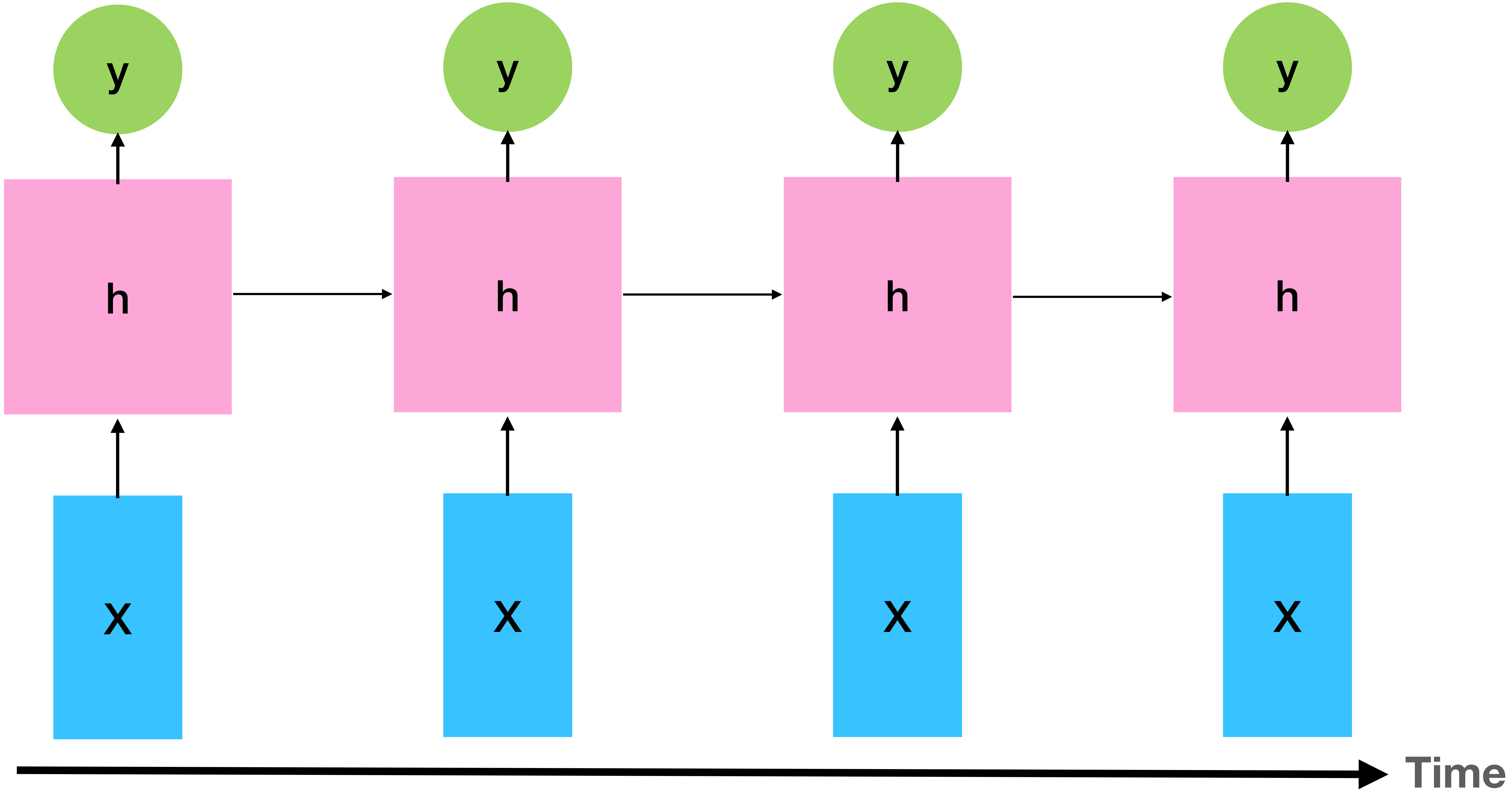
# Unroll an RNN

- This looks like the feedforward neural network we've seen, but each unrolled "layer" have the same weights

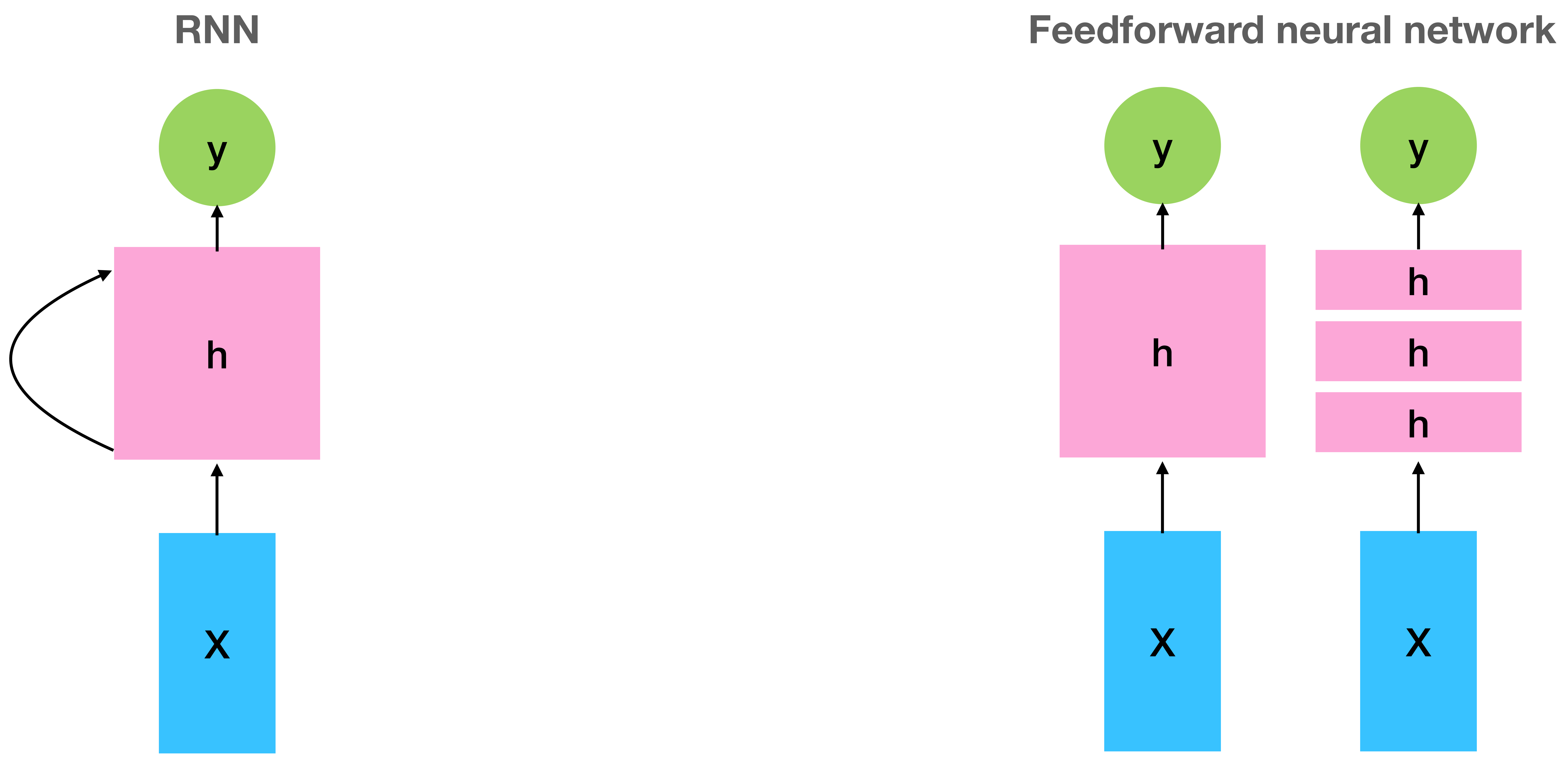


# Unroll an RNN

- RNNs only process one input **X** at a time (unlike CNN/feedforward NNs)
- But they have memory of the previous inputs stored in the hidden state

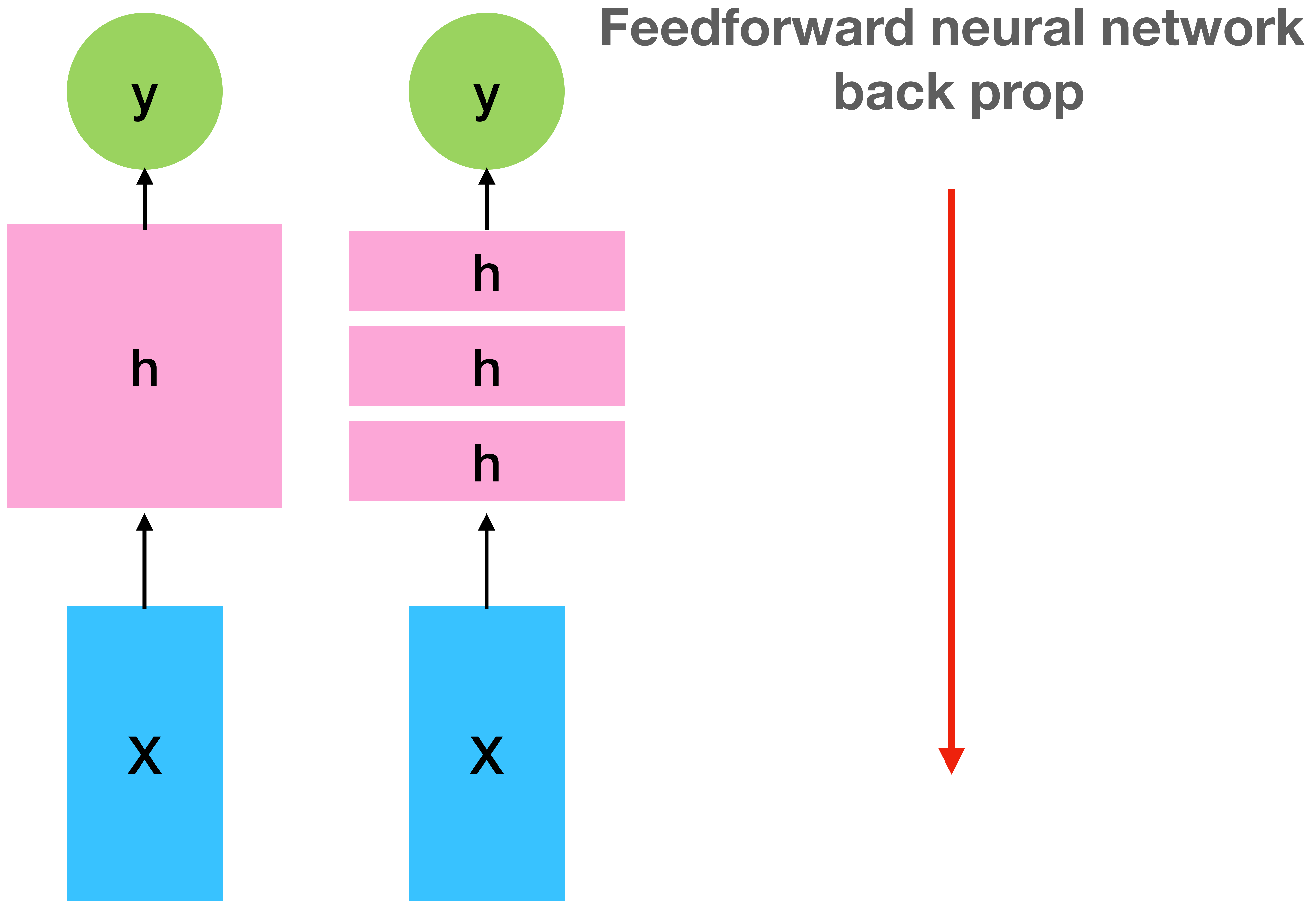


# RNN vs. feedforward NN



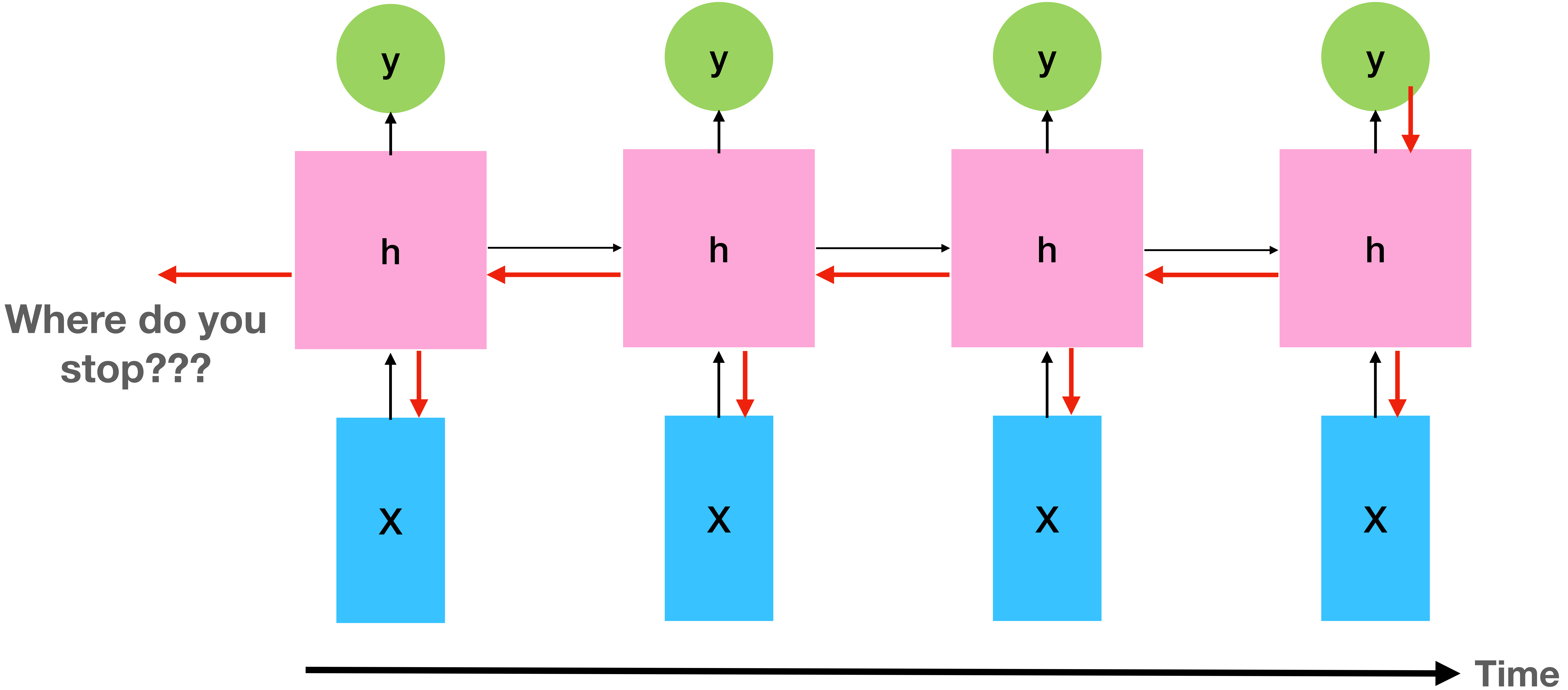
# RNN training

- Review: a feedforward neural network trains via back propoagation through the network



# RNN training: BPTT

## Back propagation through time

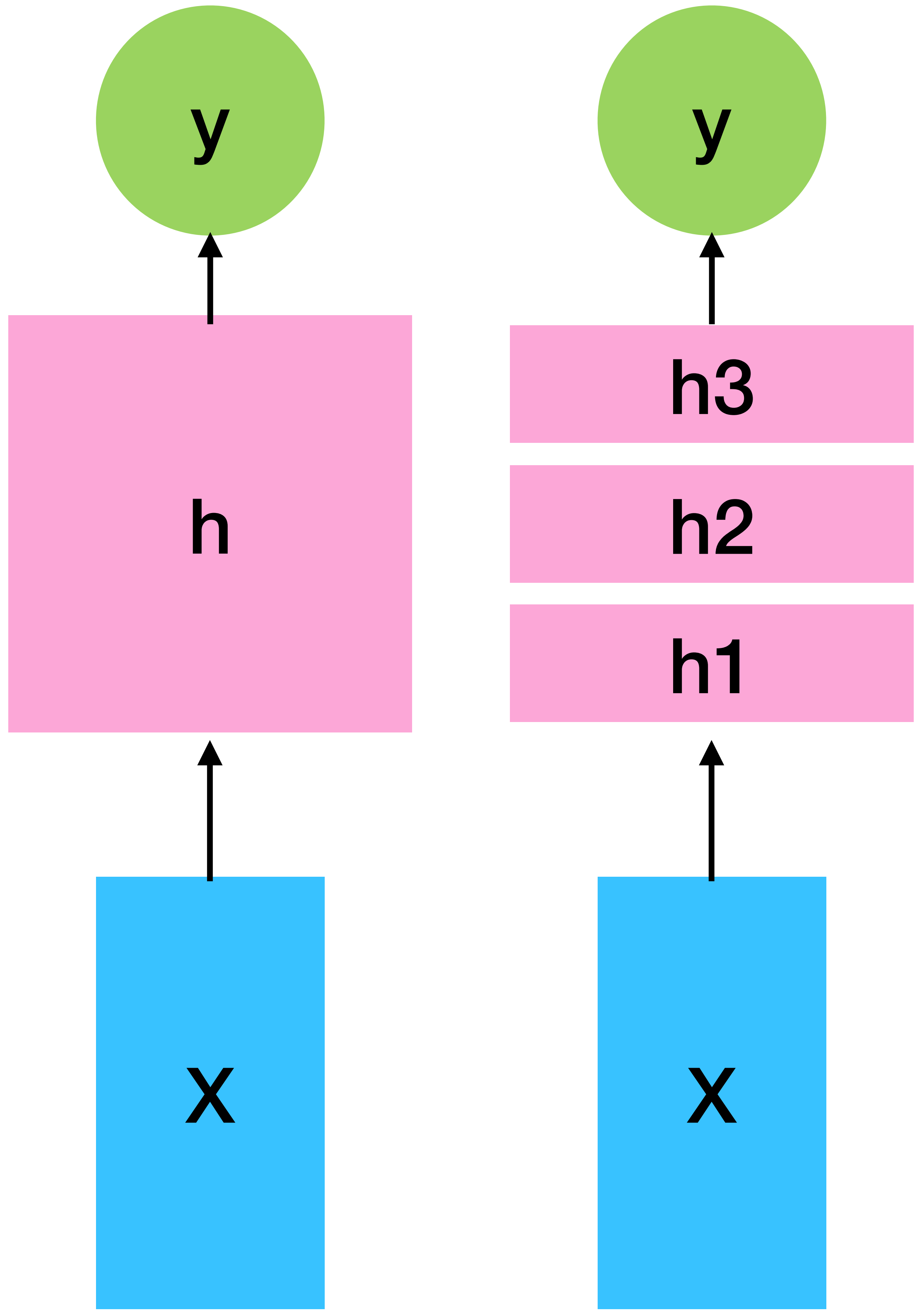


# RNN training: truncated BPTT

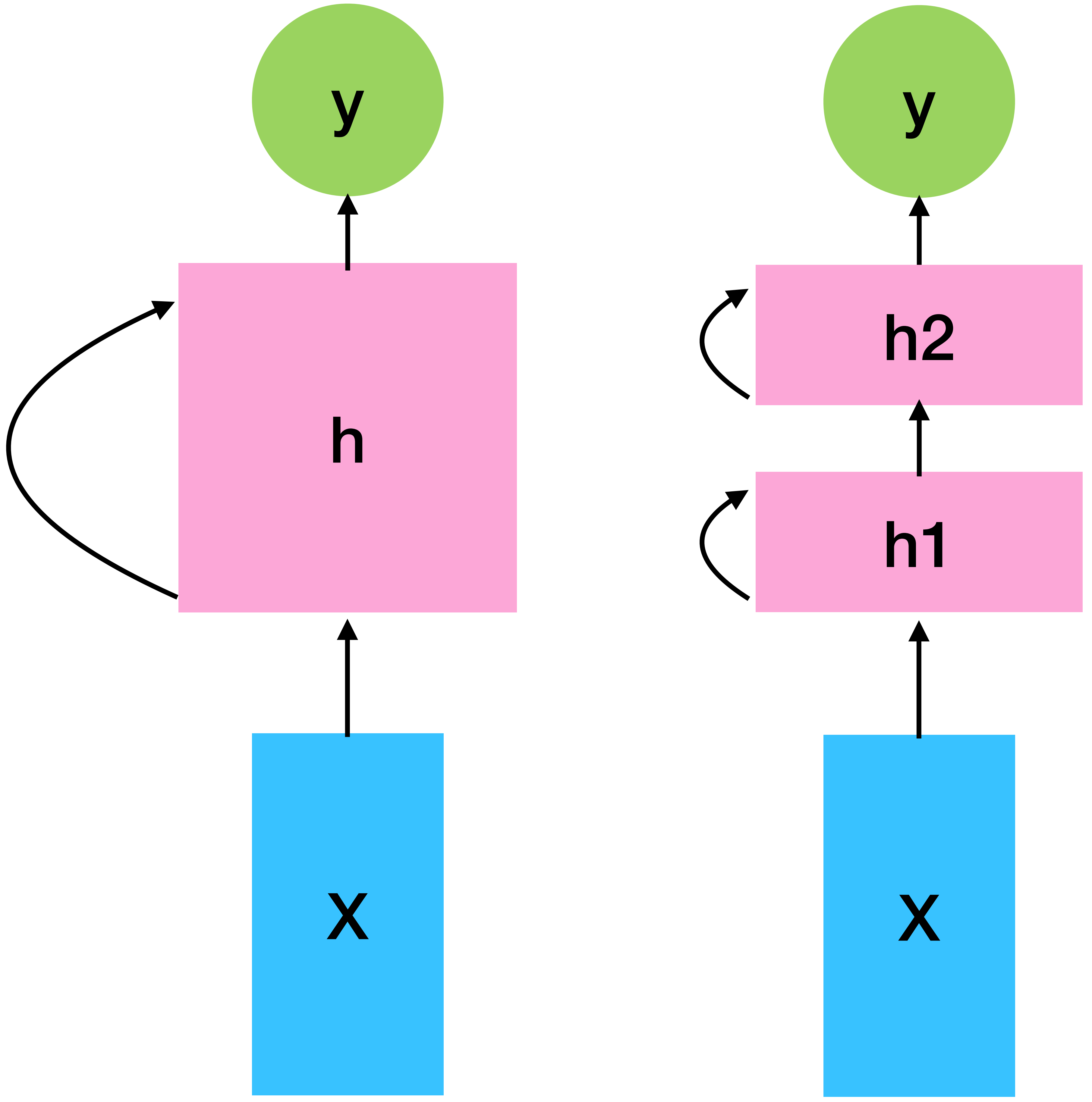
- Instead of backpropagating all the way back in time, RNN is usually trained with backpropagating k timesteps only

# Stacked RNNs

## Feedforward neural network




## RNN



# Issues with RNNs

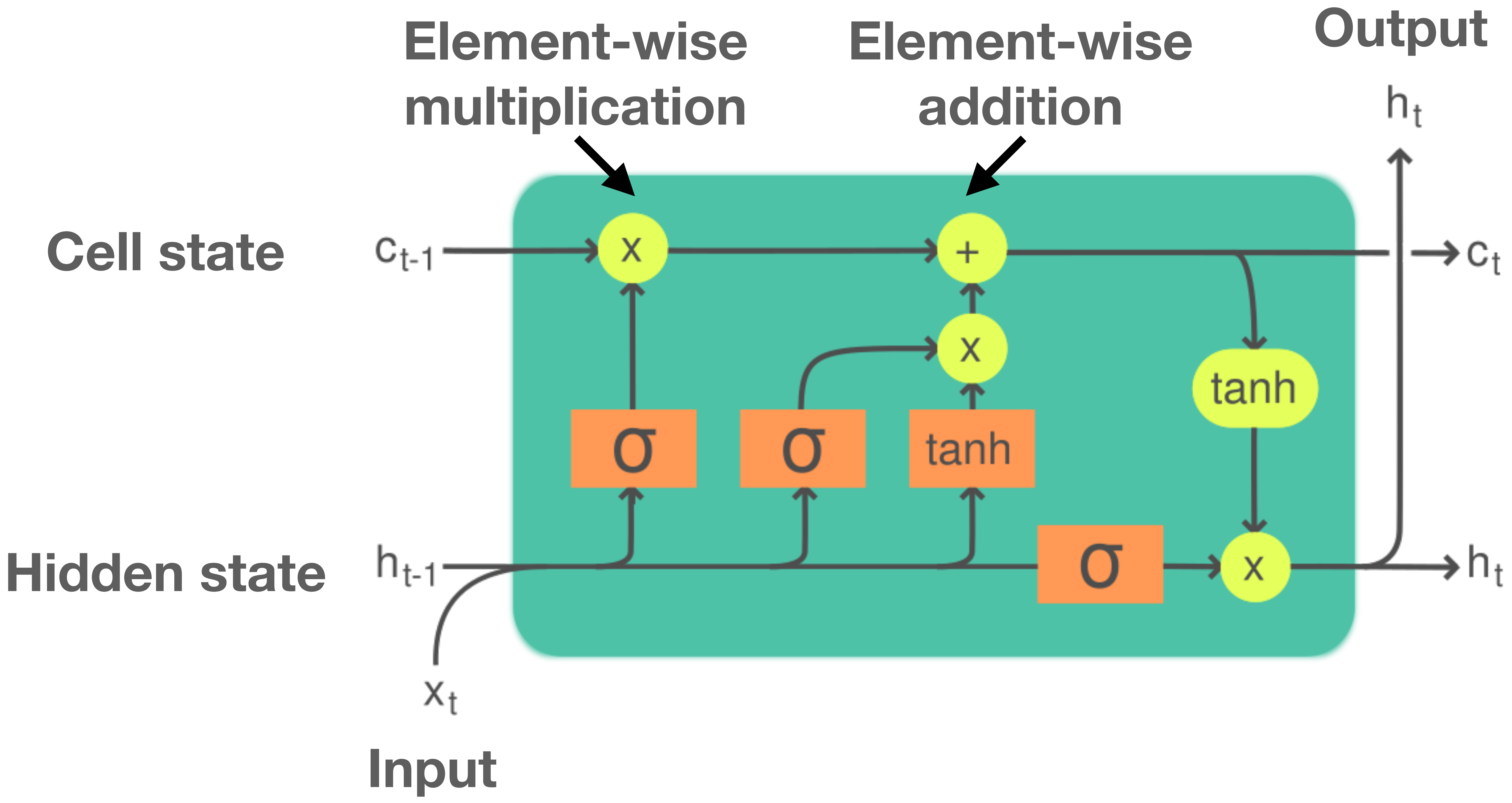
- Vanishing and exploding gradients
  - Vanishing gradients: gradient updates are close to 0, not learning much
  - Exploding gradients: gradient updates are HUGE, unstable training
- Both issues get worse when backpropagating over long sequence
- Limiting the temporal dependence that can be learned
- Introducing…

# Gated Recurrent Neural Network

## Long short-term memory network (LSTM)

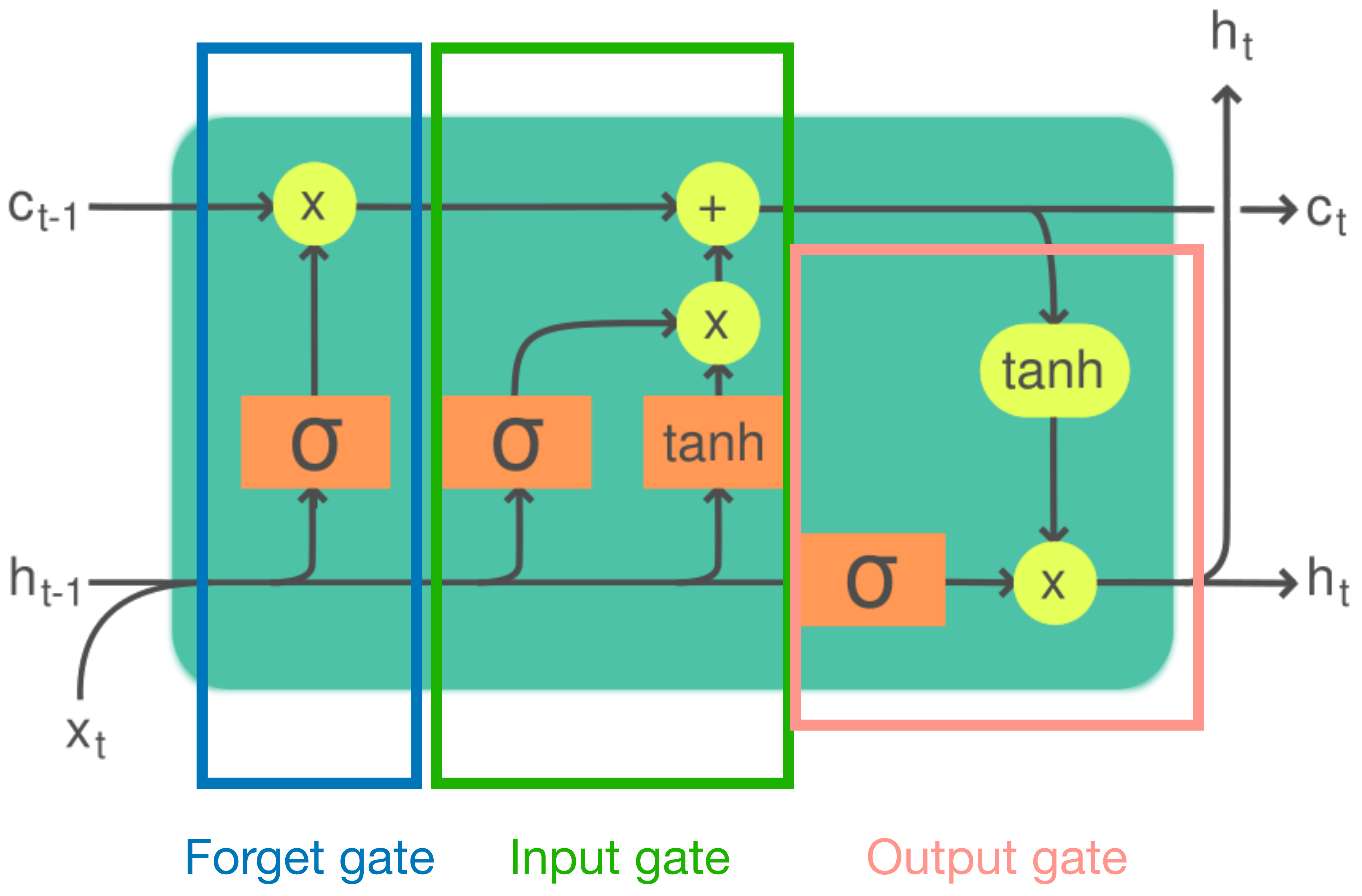
- Alleviates the issues of vanishing and exploding gradients, allows learning of long-term dependences
- Besides a hidden state, LSTM contains *a cell state* that does not pass through non-linearities
  - Allows past information to be preserved
- LSTM contains a set of gates, controlling how information is added/removed from previous states
  - Each gate is its own neural network

# LSTM


Element-wise multiplication
Element-wise addition
Output
$h_t$
Cell state
$c_{t-1}$
x
+
x
$c_t$
tanh
σ
σ
tanh
σ
x
Hidden state
$h_{t-1}$
$h_t$
$x_t$
Input

# LSTM: gates

- Each gate is a neural network, learning a function on $x_t$ and $h_{t-1}$, and applying sigmoid
- Forget gate controls how information is removed from the previous cell state $c_{t-1}$
- Input gate controls how information is added to the cell state
- Output gate controls how information from the cell state becomes the output



# RNN summary

- Learns temporal dependency
- Not great at long term dependency: vanishing and exploding gradients
- LSTM alleviates this issue but not entirely
- Questions?

# Language modeling

## Next-word prediction

“Good Morning!” said Bilbo, and he meant it. The sun was shining, and the grass was very green. But Gandalf looked at him from under long bushy eyebrows that stuck out further than the brim of his shady hat. “What do you mean?” he said. “Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?” “All of them at once,” said ___.

**How would an LSTM solve this problem?**

- Remember RNNs only process one input (word) at a time
- Store “Bilbo” in its cell state/ hidden state
- Learn to keep the information for 90 steps
- Decode “Bilbo” from cell state/ hidden state

Is that efficient? Is that how you as a human would solve this problem?

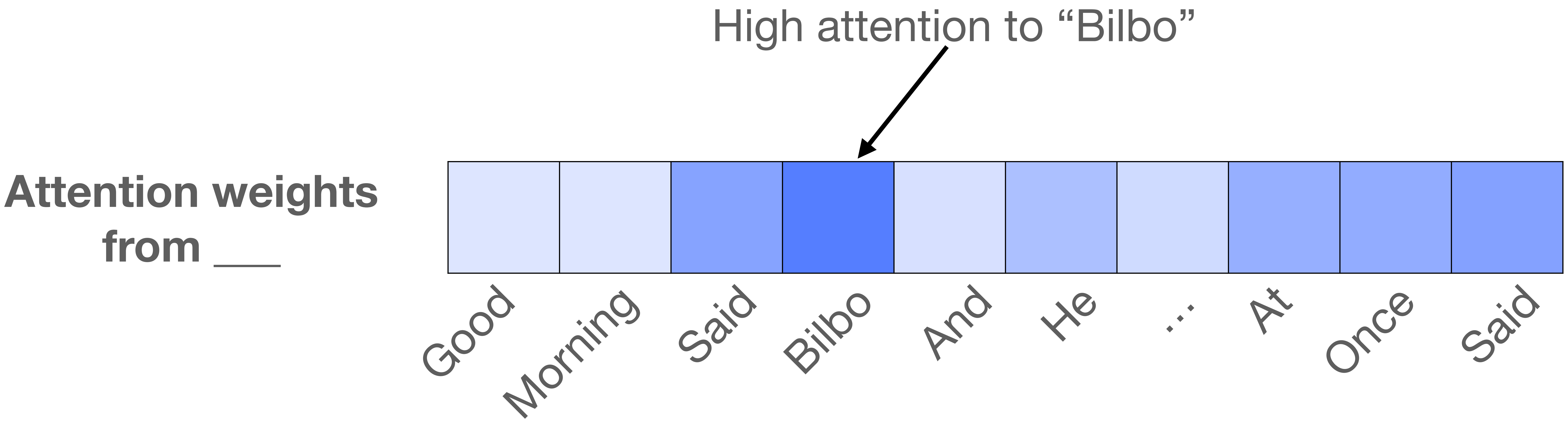
# Attention

- A human would know that you're supposed to fill in a name, check what's the name of the person at the start of the paragraph, and fill it in the blank.
- We need a model that can select the right things at the right time!

"Good Morning!" said Bilbo, and he meant it. The sun was shining, and the grass was very green. But Gandalf looked at him from under long bushy eyebrows that stuck out further than the brim of his shady hat. "What do you mean?" he said. "Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?" "All of them at once," said ___.

# Attention

## Vaswani et al. 2017: *Attention is all you need*



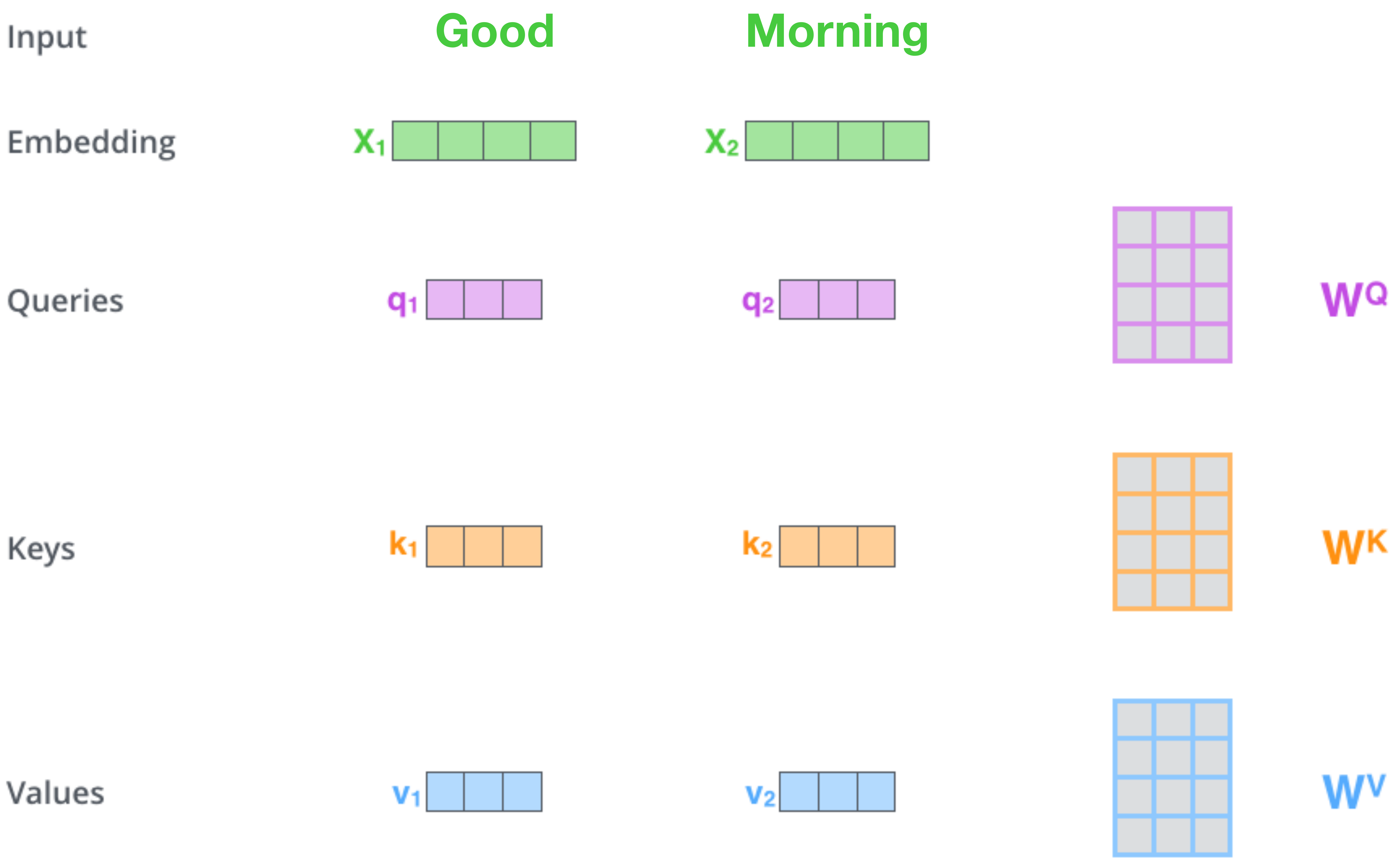

Attention weights to all words sum up to 1

# Attention

## How are attention weights computed?

- Each word is represented by a high-dimensional embedding vector (GPT-2 embeddings have 768 dimensions)
- Each embedding vector is then converted to a **query vector**, a **key vector**, and a **value vector**
- Given a **query**, select the **key** that's most relevant, and fetch its **value**

# Attention

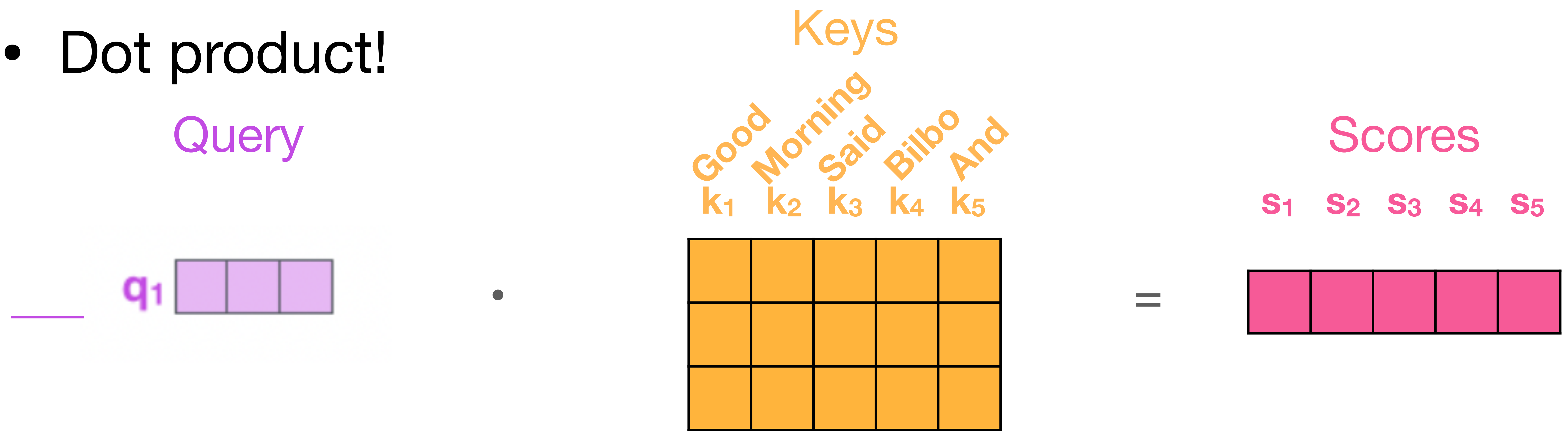



The Illustrated Transformer, 2018

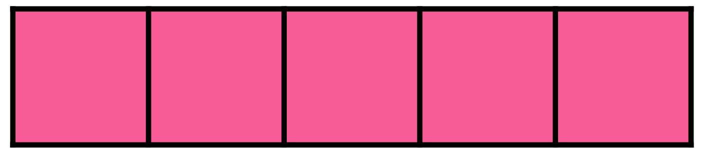
# Attention

Attentions = $\mathrm{softmax}(\frac{qK^\top}{\sqrt{d_k}})$

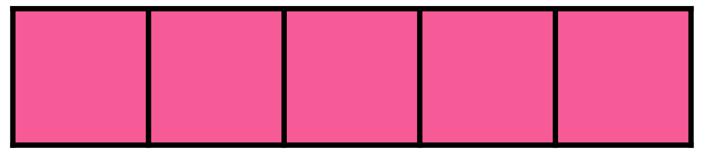
- Given a query, how do you find the most relevant keys?
- Dot product!




- Apply softmax to convert the Scores to the range of 0 and 1

Attentions = Softmax(  )

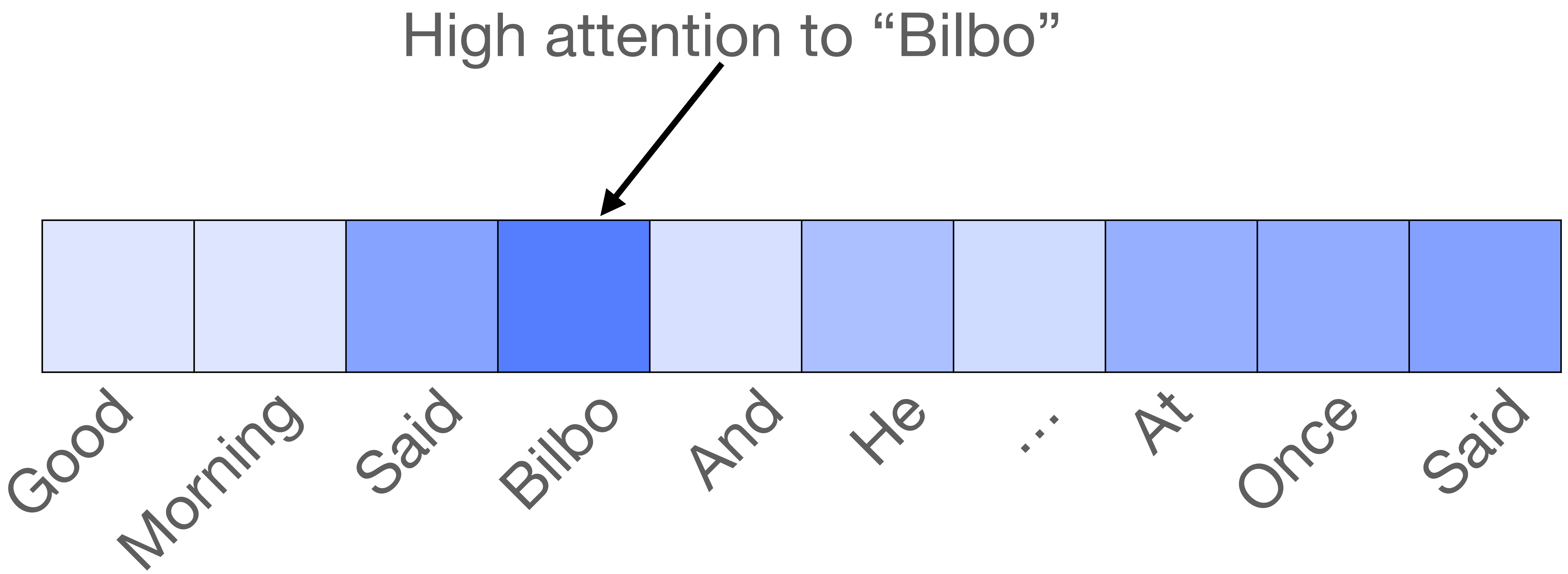
- This tells you how likely each key is matched to the query!
- Attentions are scaled by sqrt(# dimensions in keys), giving **scaled dot product attention**:

Softmax(  / $\sqrt{d_k}$ )

# Attention

## Vaswani et al. 2017: *Attention is all you need*

**Attention weights from ___**

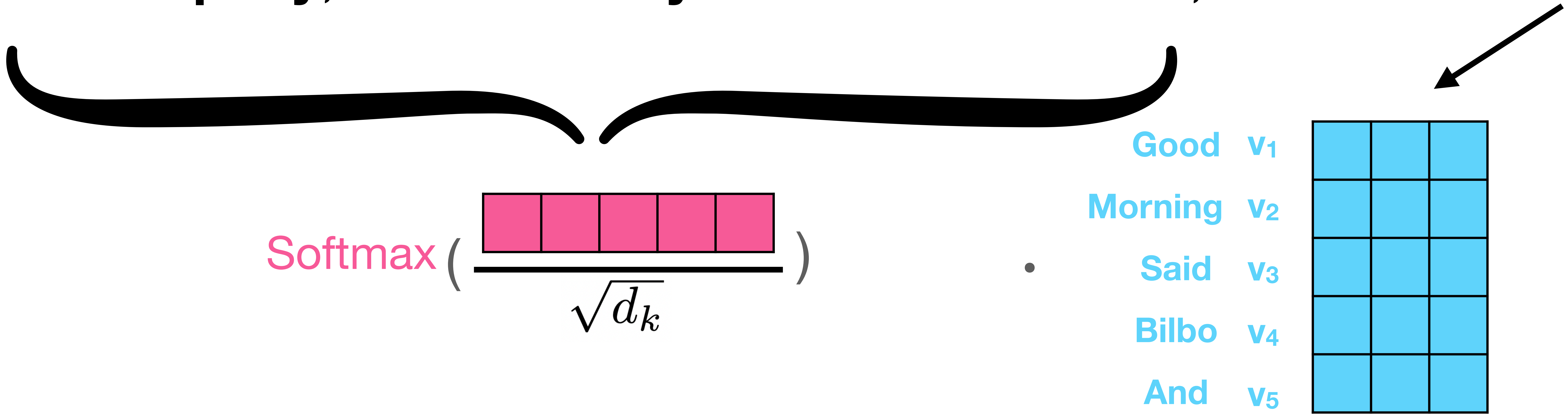



Attention weights to all words sum up to 1

Attentions = $\mathrm{softmax}(\frac{qK^\top}{\sqrt{d_k}})$

# Attention

- Given a **query**, select the **key** that's most relevant, and fetch its **value**

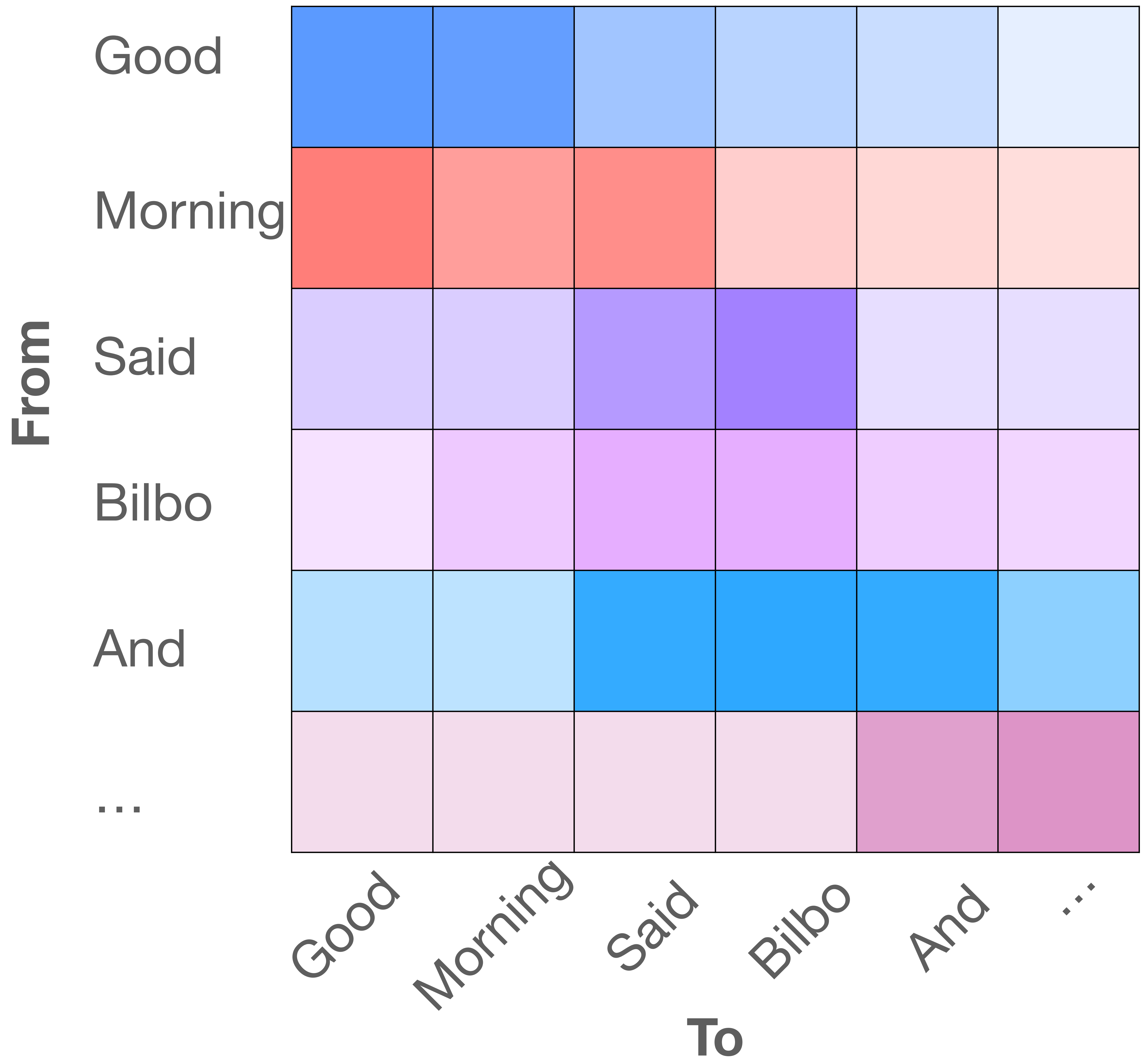



- i.e. Use the attention weights to take a weighted sum of the values

$$\text{Attention}(q, K, V) = \text{softmax}(\frac{qK^\top}{\sqrt{d_k}})V$$

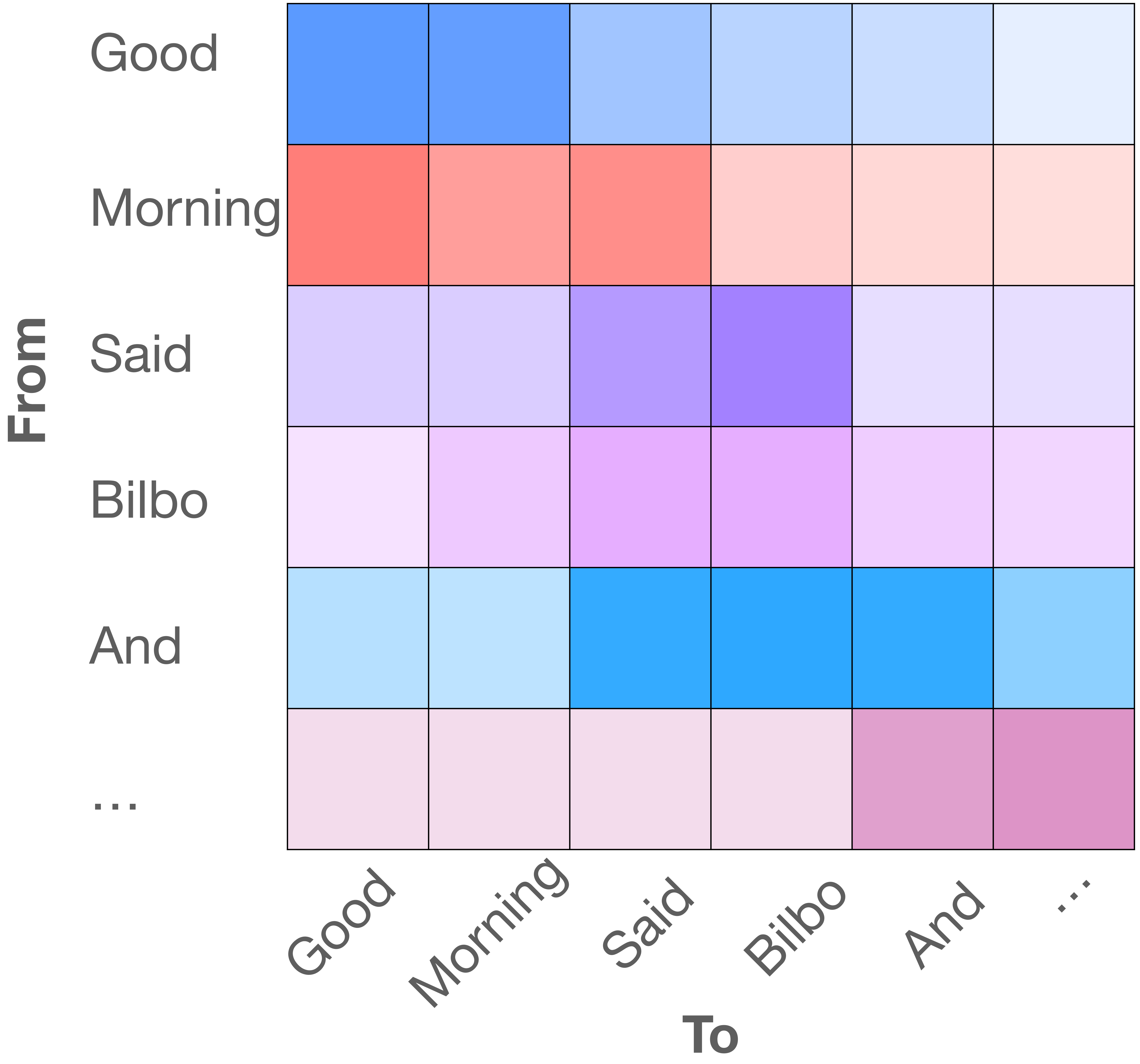
# Self-attention

- Where do the queries come from?
- Self attention: Each word is a query! Each word attends to every other word
- Each row sums to 1



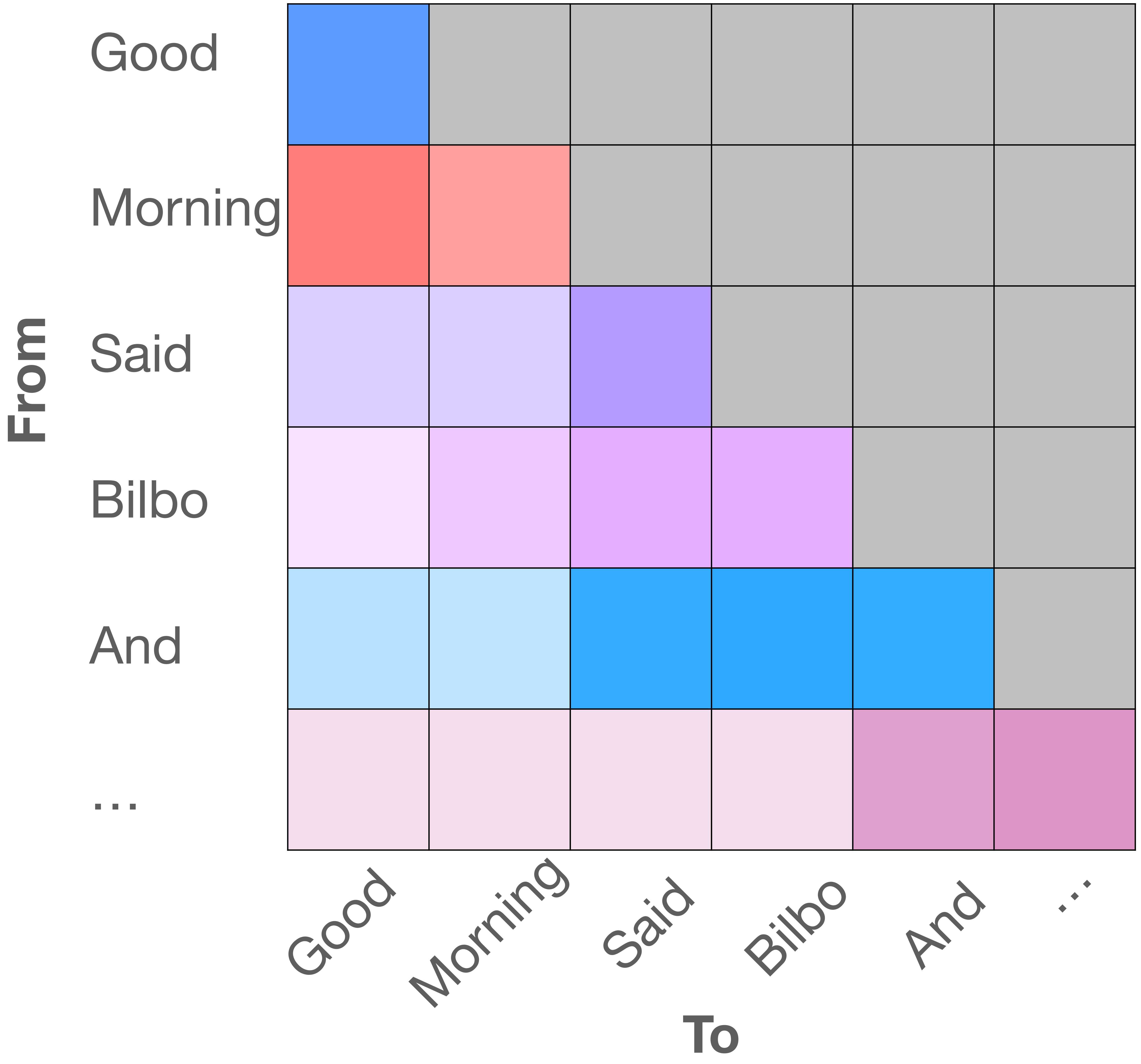
# Causal self-attention

- With self-attention, every word attends to every other word
- But when training language models to predict the next word, that's cheating!
- To prevent cheating, mask out attention to future words

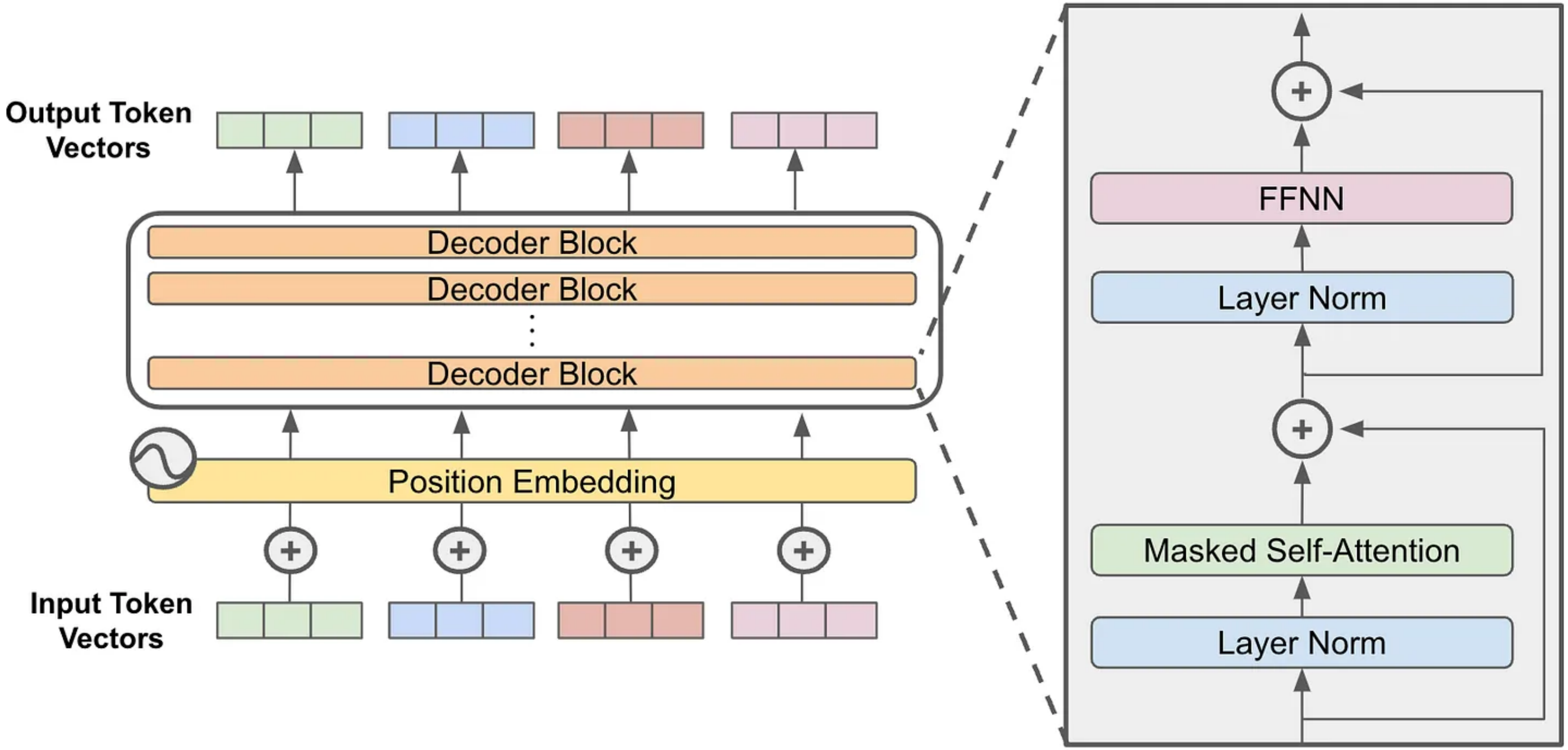


# Causal self-attention

- With self-attention, every word attends to every other word
- But when training language models to predict the next word, that's cheating!
- To prevent cheating, mask out attention to future words



# Transformer architecture




https://cameronrwolfe.substack.com/p/decoder-only-transformers-the-workhorse

# Language modeling, formally

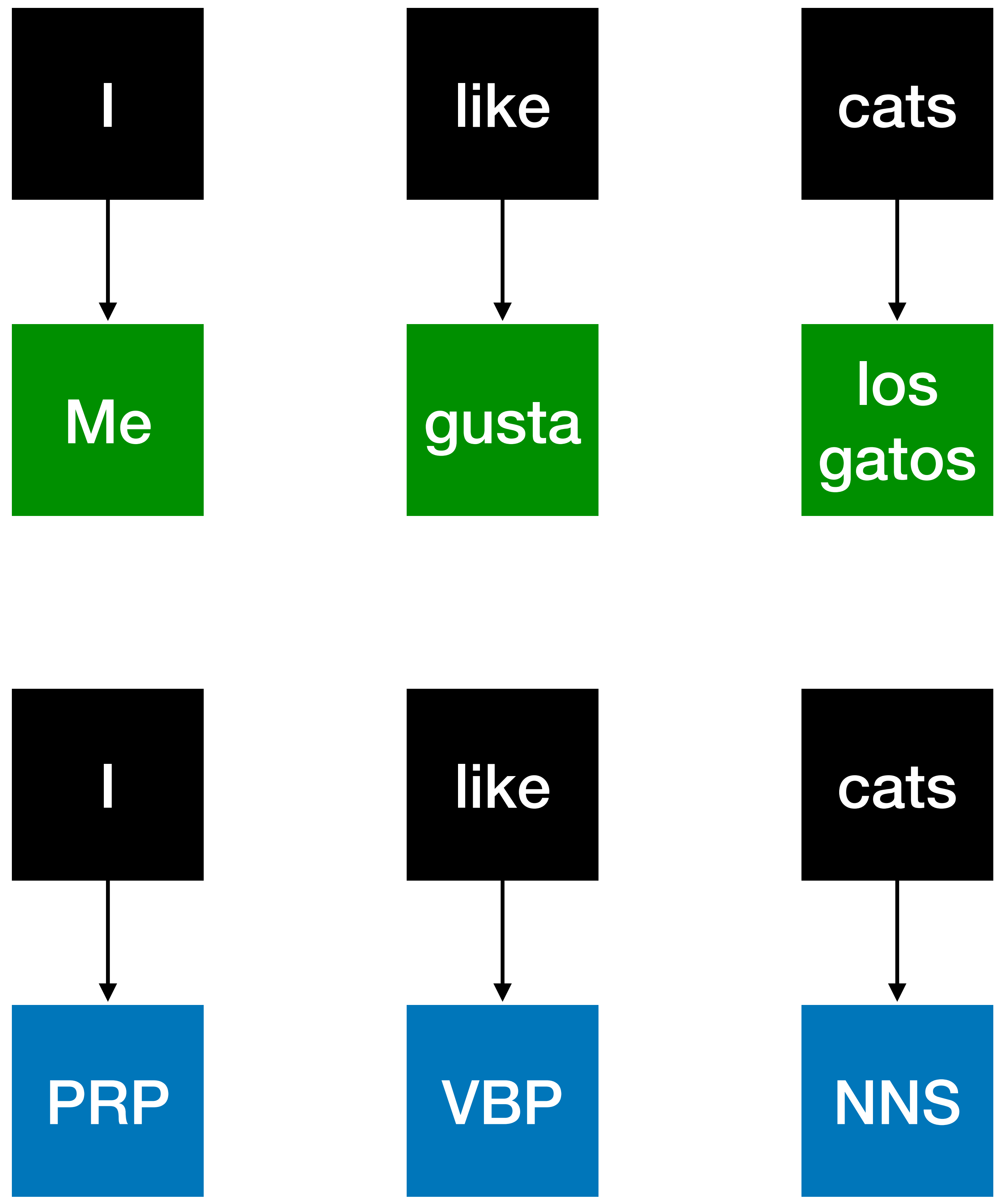
- At each position, language models generate a probability distribution of P(next word | previous words)
- A multi-class classification problem!
- What kind of loss do we use?
- Cross-entropy loss: how much does the predicted probability deviates from the correct word
- Generation: sample from the probability distribution

"Good Morning!" said Bilbo, and he meant it. The sun was shining, and the grass was very green. But Gandalf looked at him from under long bushy eyebrows that stuck out further than the brim of his shady hat. "What do you mean?" he said. "Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?" "All of them at once," said ___.

# But language models are good at many things!

- Translation
- Part of speech tagging
- Summarization
- Creative writing
- …

**Why does training models to predict the next word make them good at many other tasks?**



# Language modeling

- Excelling at the language modeling objective requires the model to learn the **statistics of language**
- Translation, POS tagging, generation, all requires knowledge of the statistics of language
- With minimal fine-tuning, the model can transfer their knowledge of language modeling to specialized tasks: a.k.a **transfer learning**

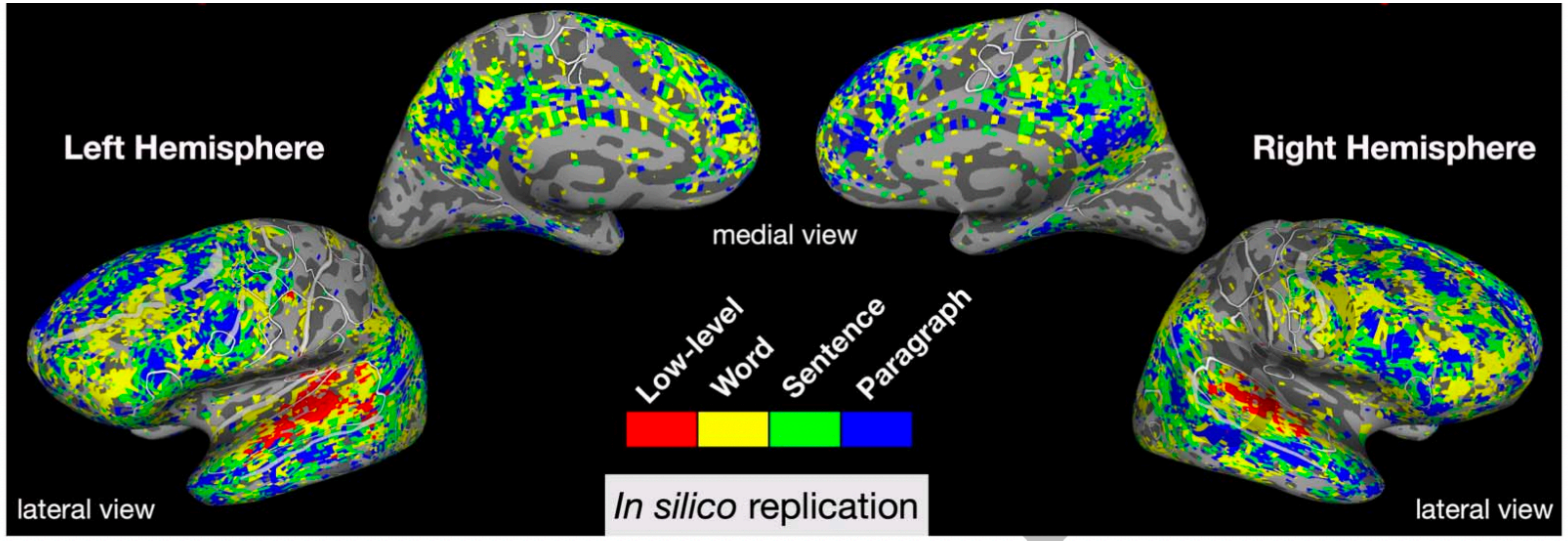
# Challenges of LLMs

## i.e. Active areas of research

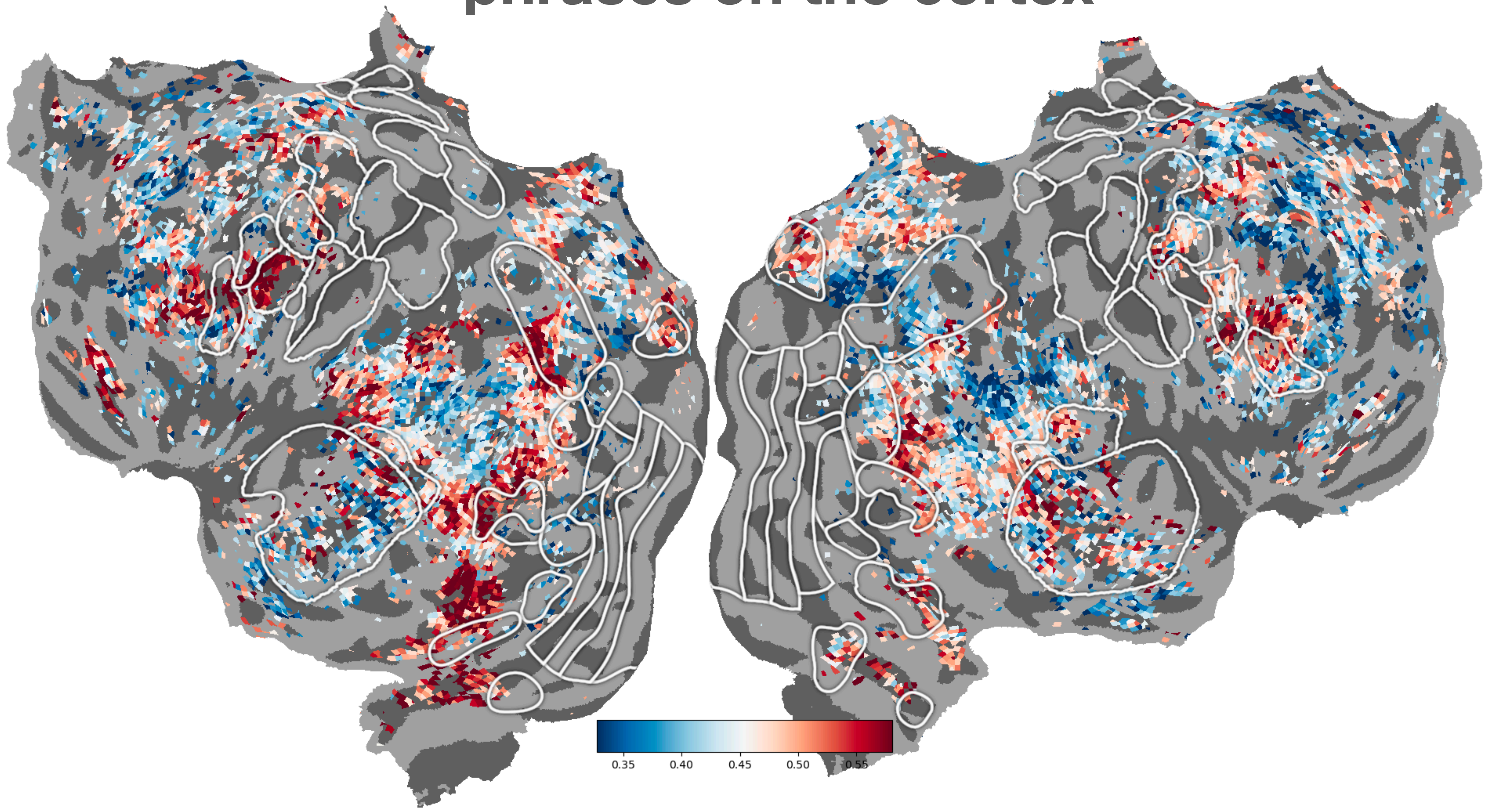
- Factual error and hallucination
- Limited context-window length
- Retrieving precise information from long-context
- Reasoning
- Injecting new knowledge to keep LLMs up to date
- Reducing training and inference cost (memory requirements, GPU usage etc.)
- …

# Using language models to study the brain

Temporal receptive windows of the brain (Jain et al. 2023)




Selectivity of concrete & abstract phrases on the cortex



# Using language models to study the brain

## Decoding speech from fMRI (Tang et al. 2023)

**c**

| Actual stimulus | Decoded stimulus |
| --- | --- |
| *i got up from the air mattress and pressed my face against the glass of the bedroom window expecting to see eyes staring back at me but instead finding only darkness* | i just continued to walk up to the window and open the glass i stood on my toes and peered out i didn’t see anything and looked up again i saw nothing |
| *i didn't know whether to scream cry or run away instead i said leave me alone i don't need your help adam disappeared and i cleaned up alone crying* | started to scream and cry and then she just said i told you to leave me alone you can't hurt me anymore i'm sorry and then he stormed off i thought he had left i started to cry |
| *that night i went upstairs to what had been our bedroom and not knowing what else to do i turned out the lights and lay down on the floor* | we got back to my dorm room i had no idea where my bed was i just assumed i would sleep on it but instead i lay down on the floor |
| *i don't have my driver's license yet and i just jumped out right when i needed to and she says well why don't you come back to my house and i'll give you a ride i say ok* | she is not ready she has not even started to learn to drive yet i had to push her out of the car i said we will take her home now and she agreed |

Exact

Gist

Error

# That’s all for today!