Sequence

is a collection of **ordered** items.

Sequences entail 2 kinds of information:
1. what are the individual items;
2. how those items are arranged.

Time flies like an arrow

An arrow flies like time

An arrow time like flies

To capture the meaning of the sentence you need to find a way to model the dependencies between words.

The goal of sequence modeling is to learn a representation of the sequence, which incorporates the information about the order.

.

How to model a sequence ?

word embeddings
! are INDEPENDENT
Know nothing about words before or after

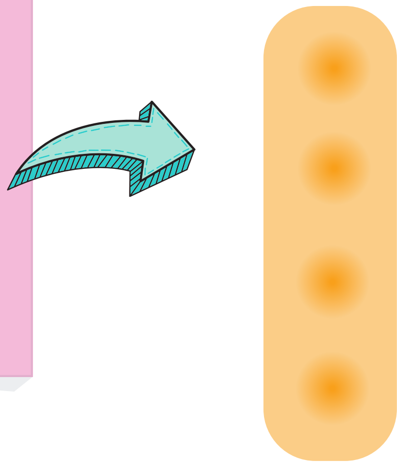
Design a **modeling strategy** to capture information about the dependency between items.

RNN

Modeling straregy

1. encode items sequentially
2. maintain a **memory** of items seen so far.

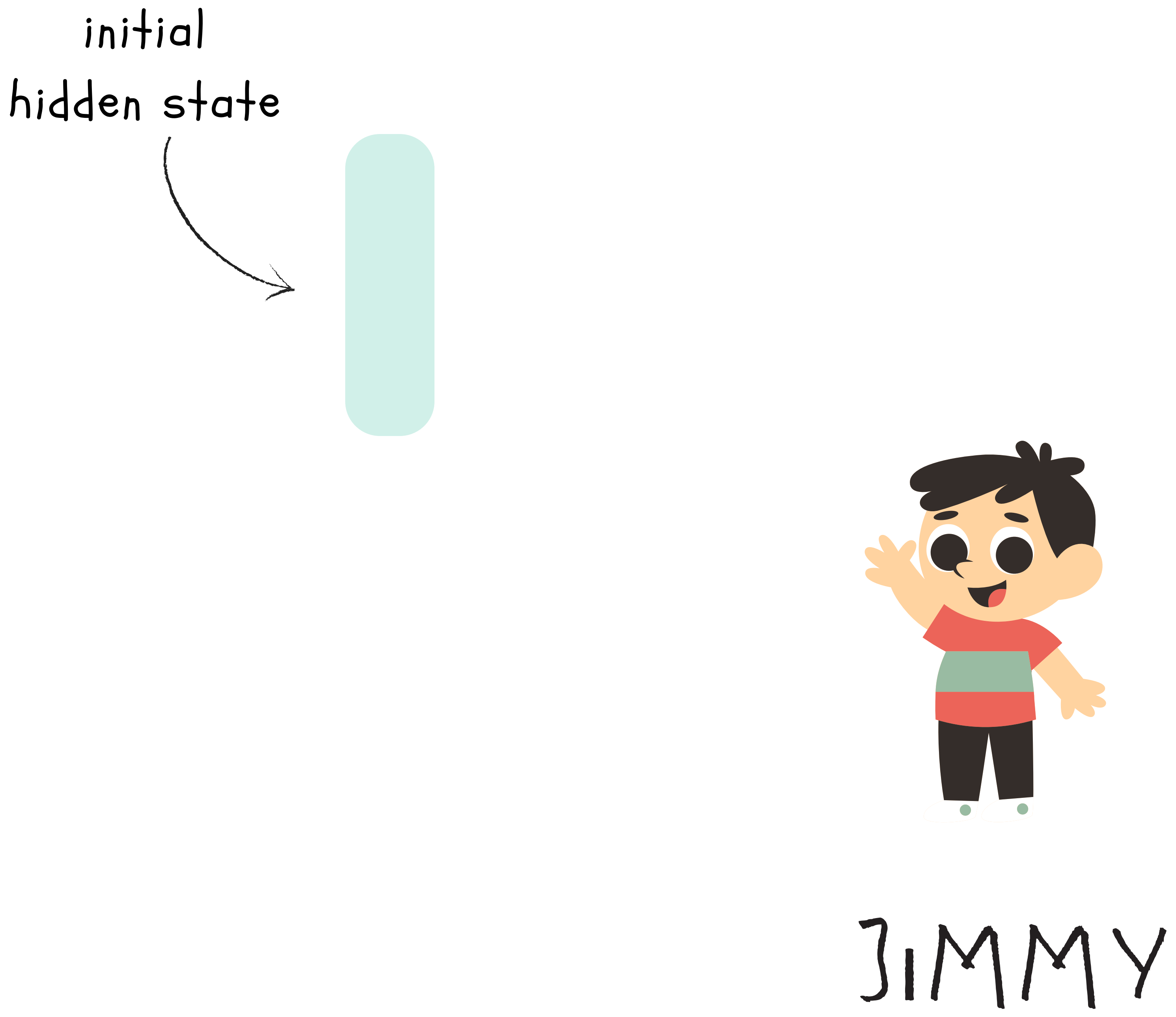
HIDDEN STATE



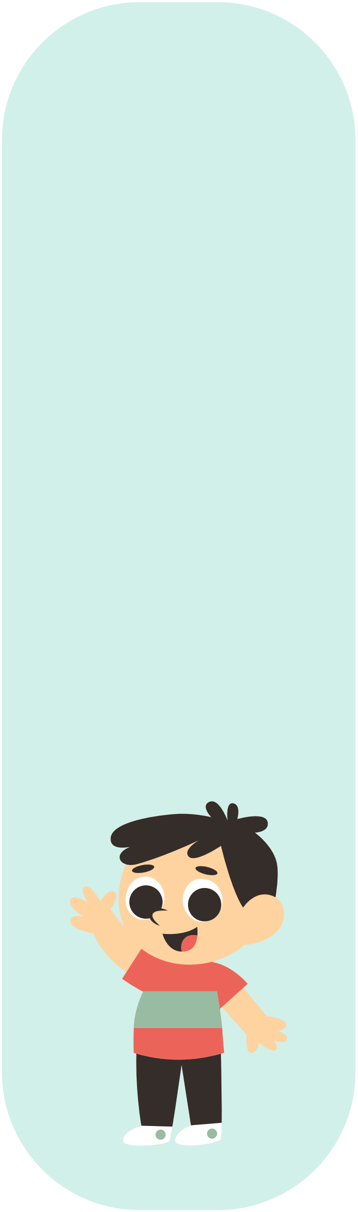
Goal of RNN:
to bake a sequence representation

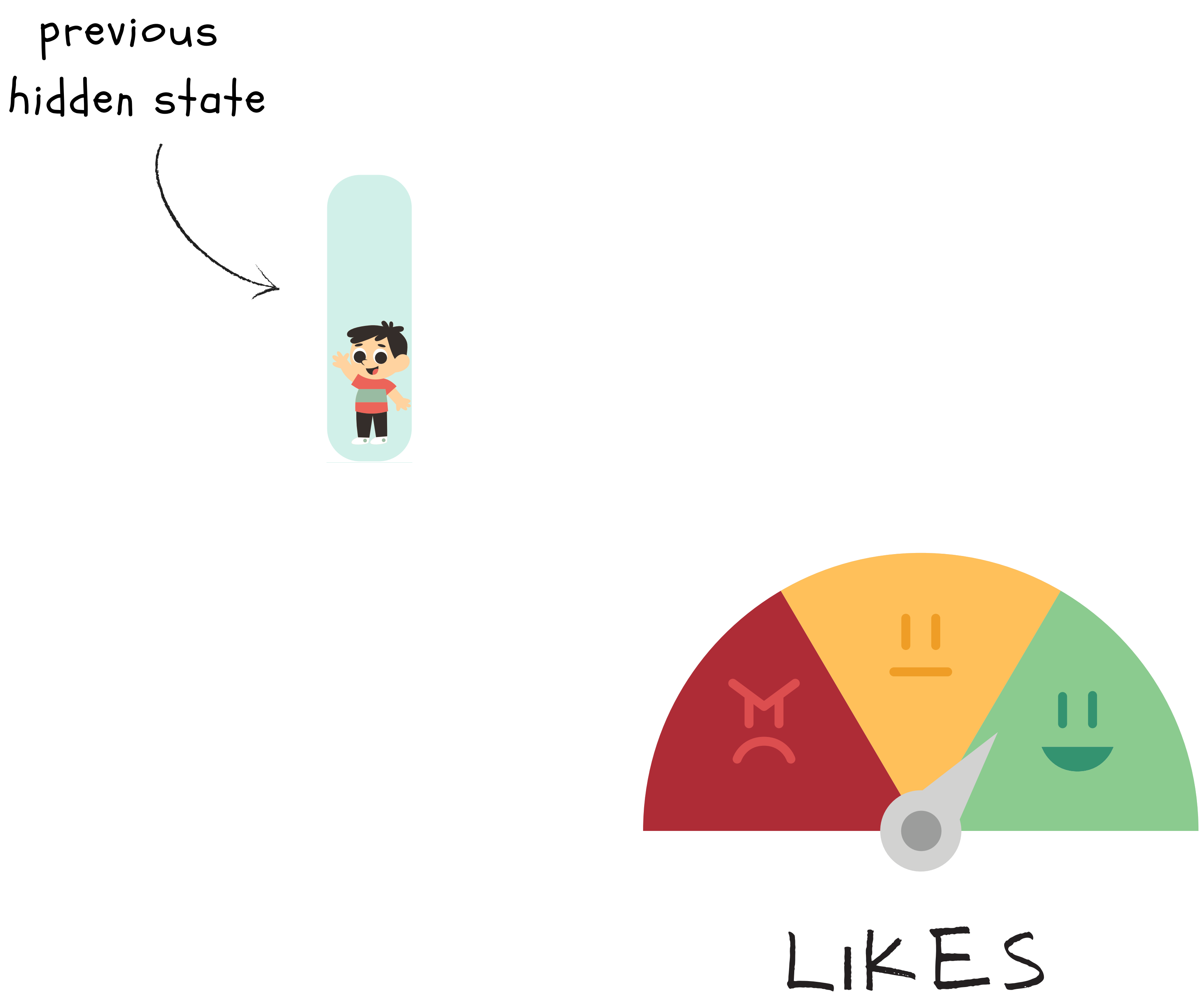
Assume that we want to use a RNN
to build a vector representation of the sentence:

Jimmy likes candies

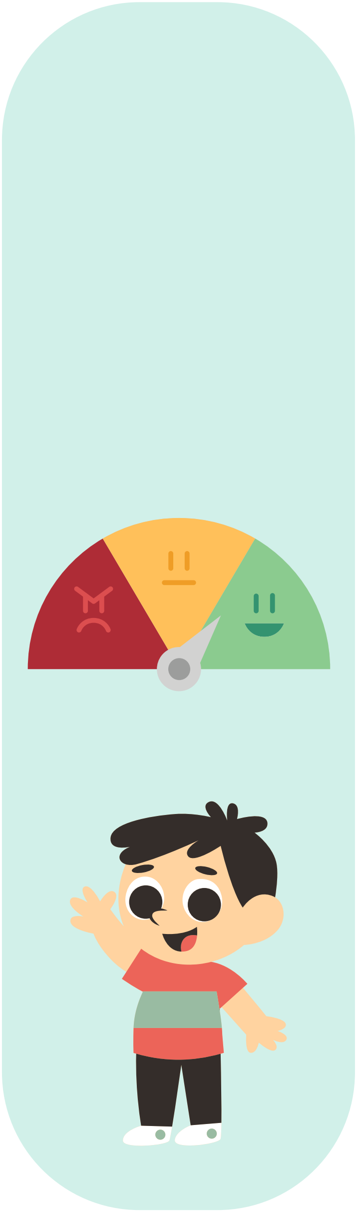

initial
hidden state
JIMMY

updated
hidden state

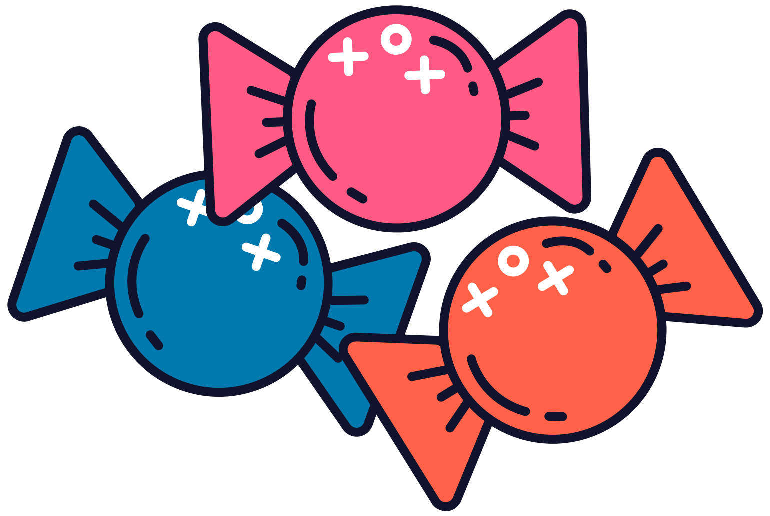



previous hidden state
LIKES

updated
hidden state



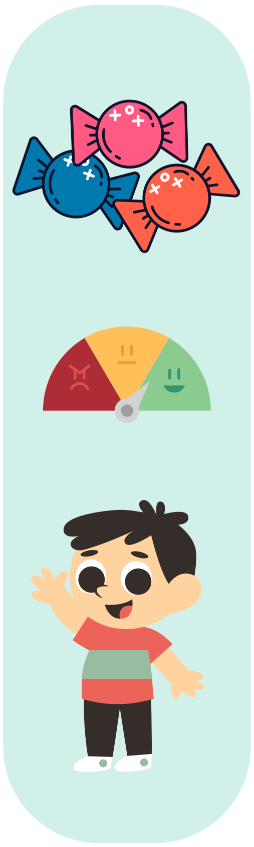
hidden state



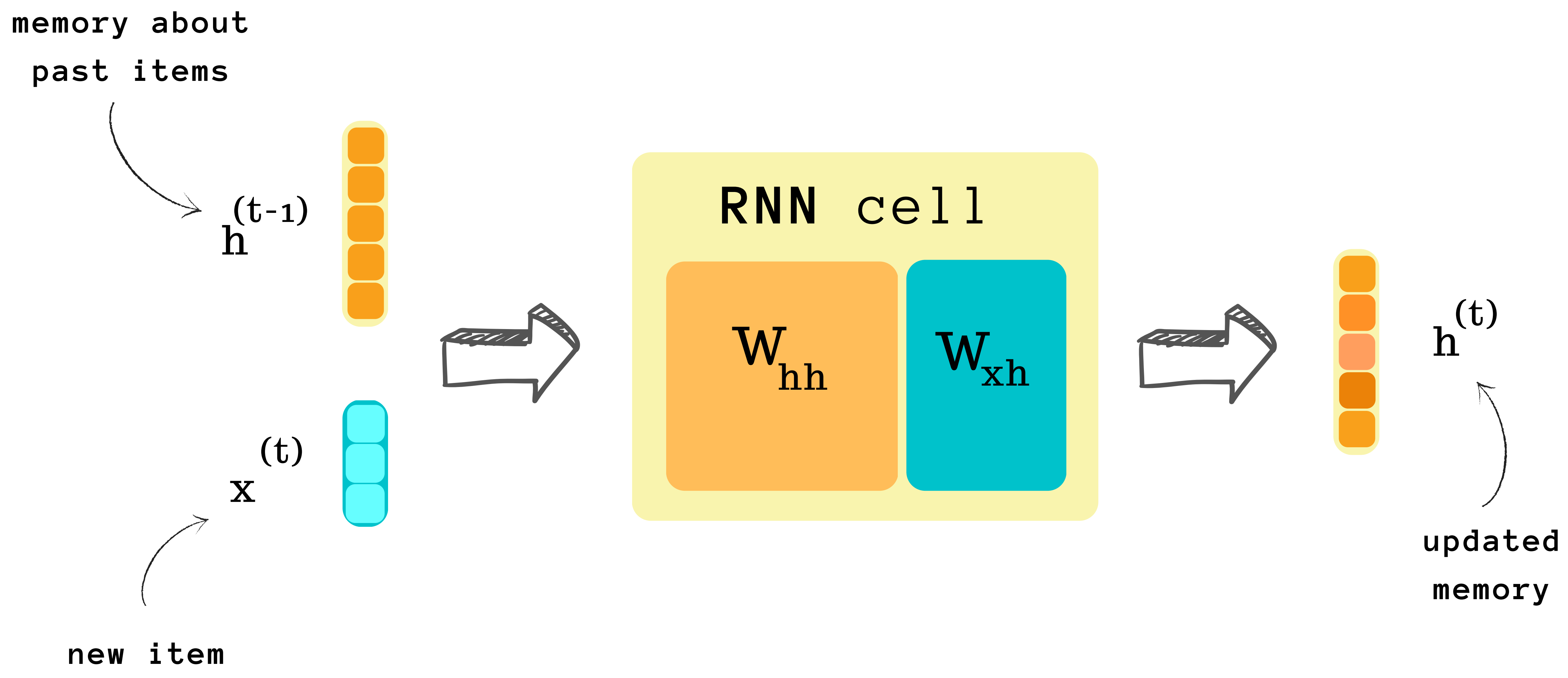
CANDIES

The final hidden state vector provides a summary of the whole sequence.

final
hidden state

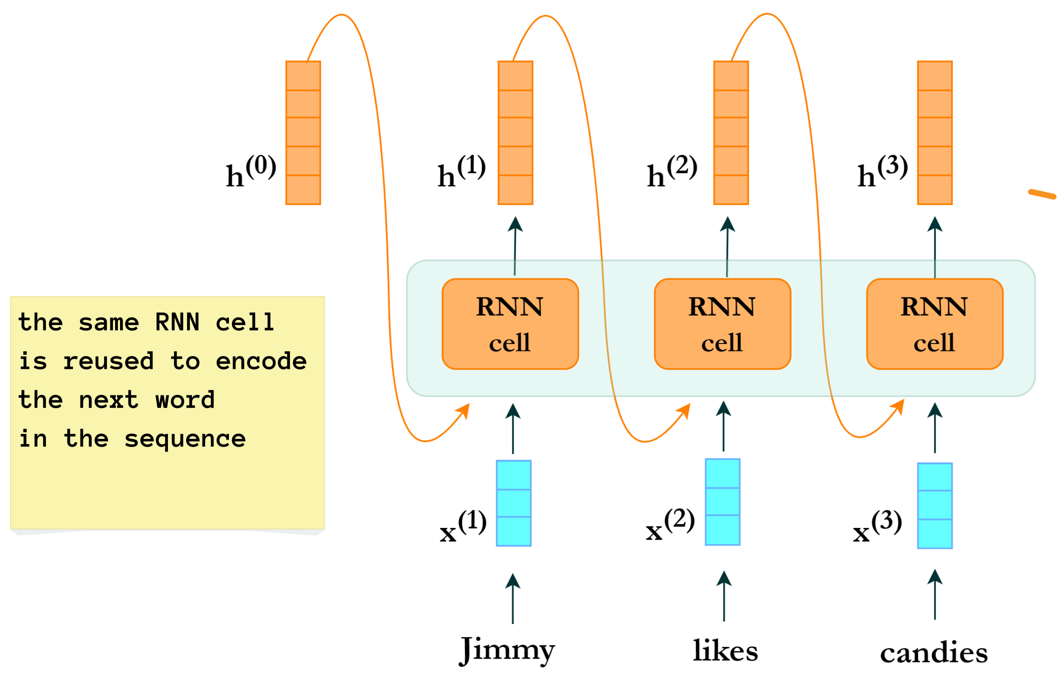


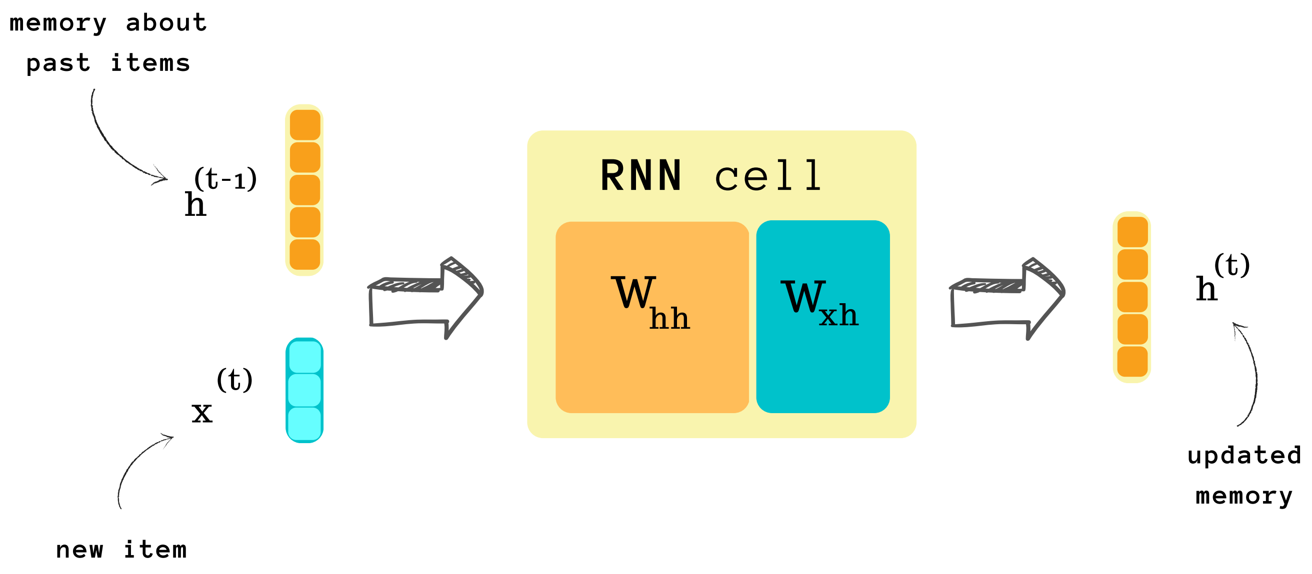
HOW DO WE COMPUTE A NEW HIDDEN STATE?



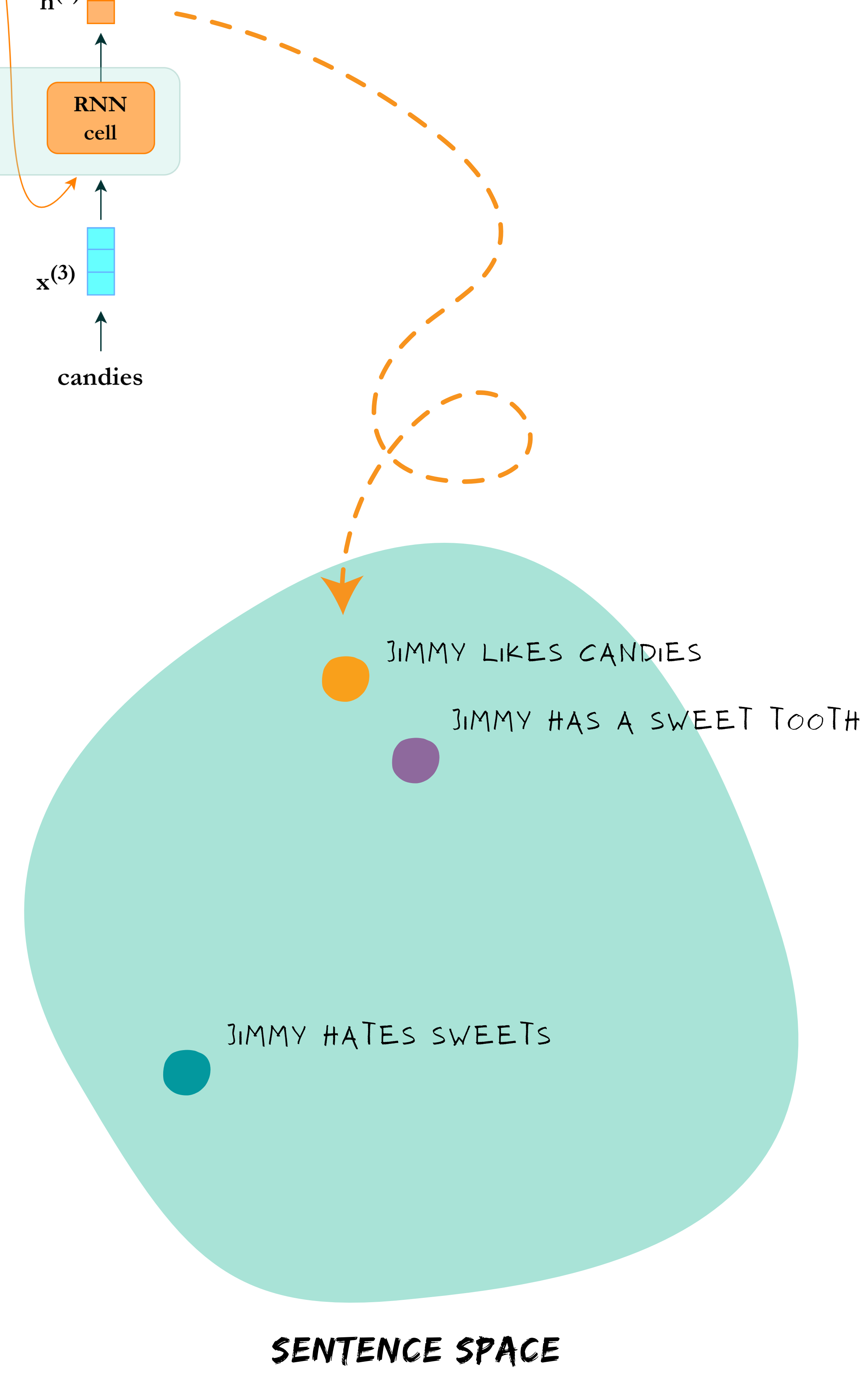
$$h^{(t)} = \tanh\left(W_{hh}h^{(t-1)} + \text{bias} + W_{xh}x^{(t)} + \text{bias}\right)$$

RECURRENT NEURAL NETWORK



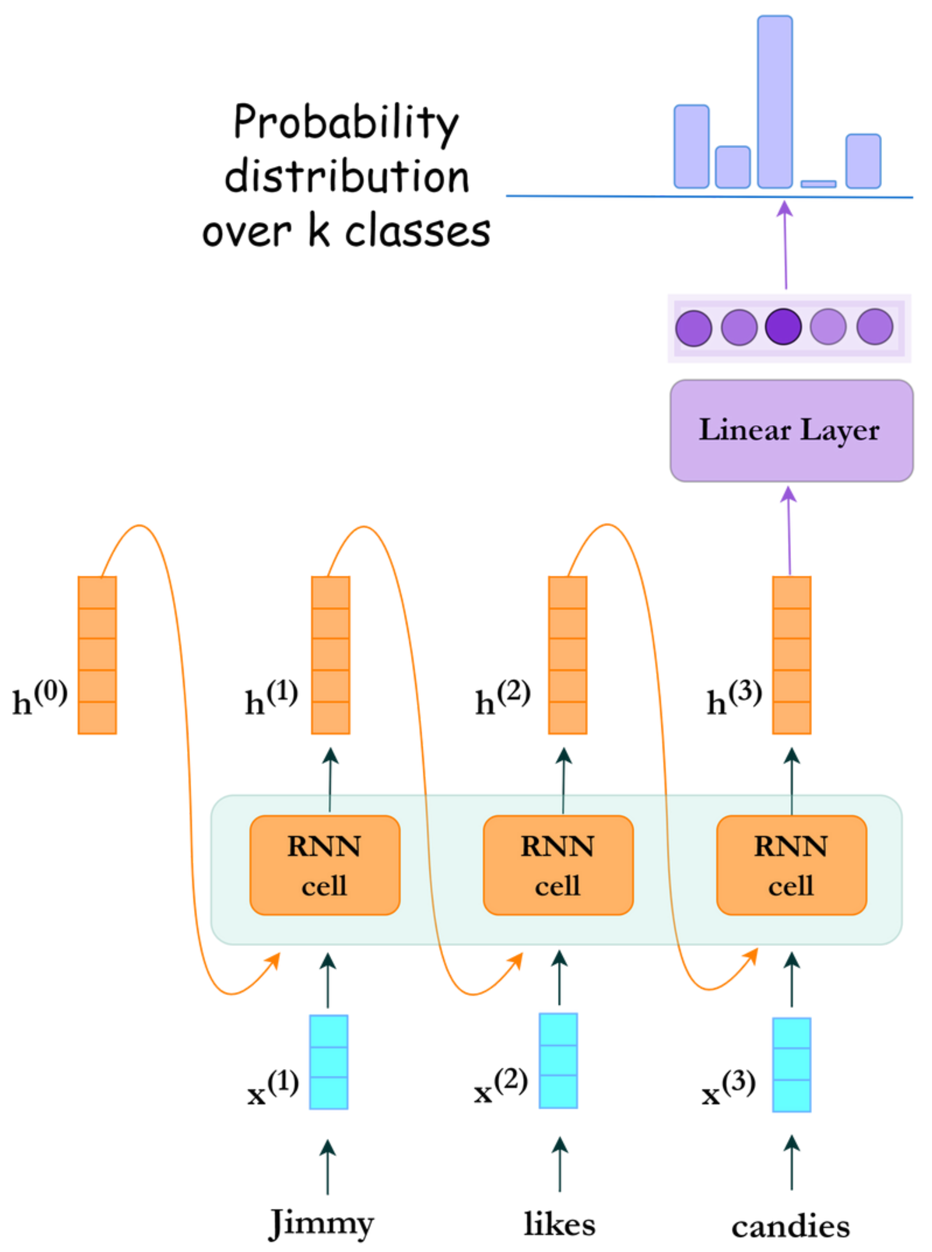


$$h^{(t)} = \tanh\left(W_{hh} h^{(t-1)} + bias + W_{xh} x^{(t)} + bias\right)$$

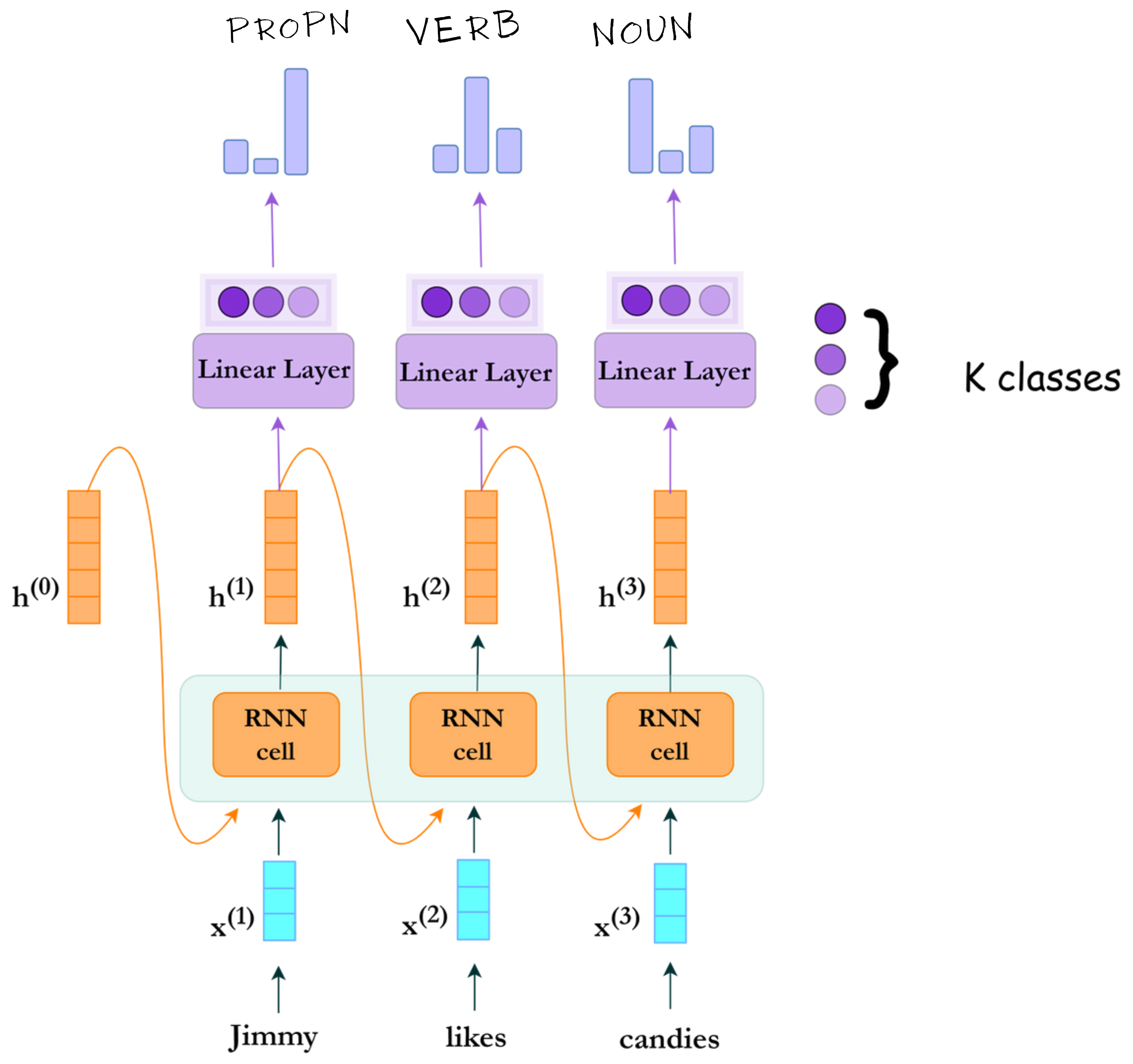


WHEN DO WE NEED RNNS?

SEQUENCE CLASSIFICATION



SEQUENCE PREDICTION



ADVANTAGES OF RNN

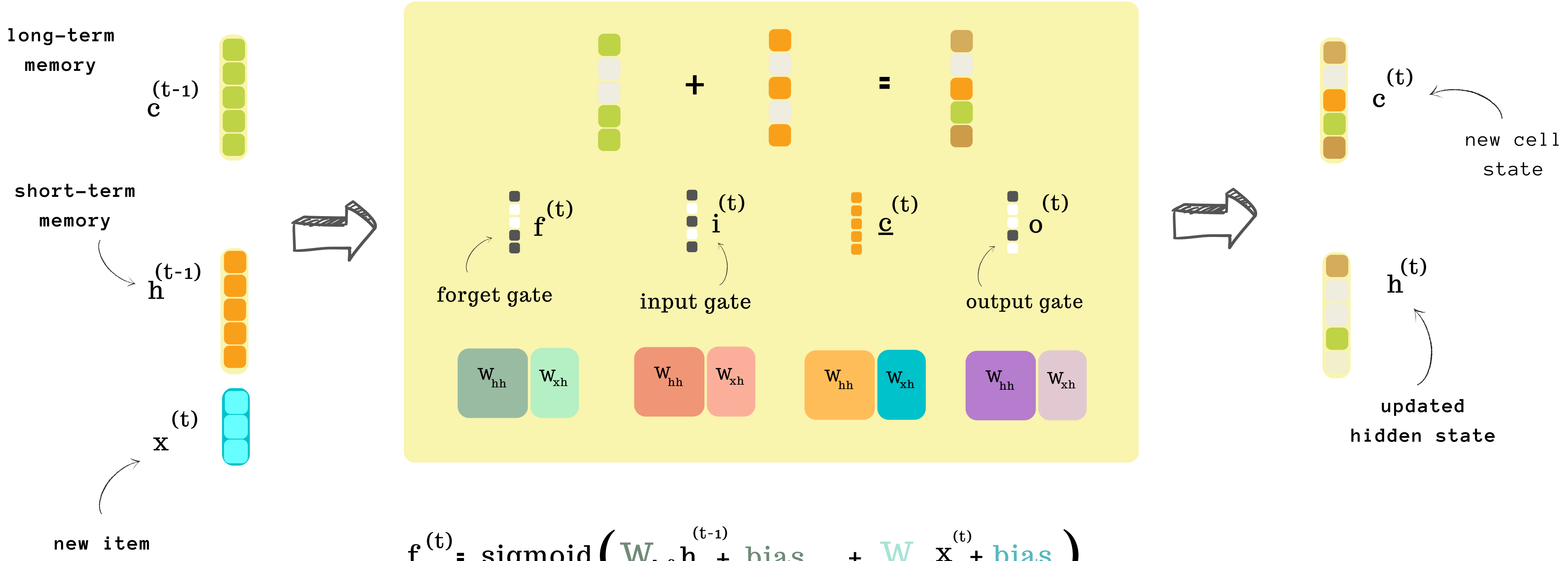
- can process sequences of any length
- can maintain the memory about the past

DISADVANTAGES OF RNN

- Recurrent computation is slow
- Computational instability with gradients spiraling out of control
- Vanilla RNN struggle to maintain long-range information

LONG-SHORT TERM MEMORY NETWORK

LSTM cell



$$f^{(t)} = \text{sigmoid}\left(W_{hf}h^{(t-1)} + \text{bias} + W_{xf}x^{(t)} + \text{bias}\right)$$

$$i^{(t)} = \text{sigmoid}\left(W_{hi}h^{(t-1)} + \text{bias} + W_{xi}x^{(t)} + \text{bias}\right)$$

$$o^{(t)} = \text{sigmoid}\left(W_{ho}h^{(t-1)} + \text{bias} + W_{xo}x^{(t)} + \text{bias}\right)$$

$$\underline{c}^{(t)} = \tanh\left(W_{hc}h^{(t-1)} + \text{bias} + W_{xc}x^{(t)} + \text{bias}\right)$$

$$c^{(t)} = c^{(t-1)} * f^{(t)} + \underline{c}^{(t)} * i^{(t)}$$

$$h^{(t)} = o^{(t)} * \tanh\left(c^{(t)}\right)$$