## RNN Basic Theory

### What are examples of sequence data set?

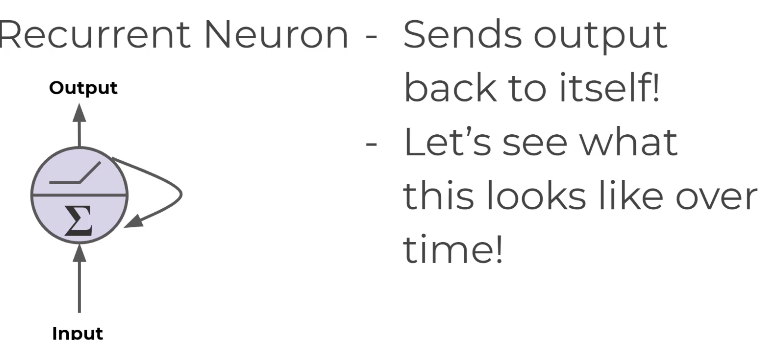
Sine waves, forecasting share prices. Audio, music..

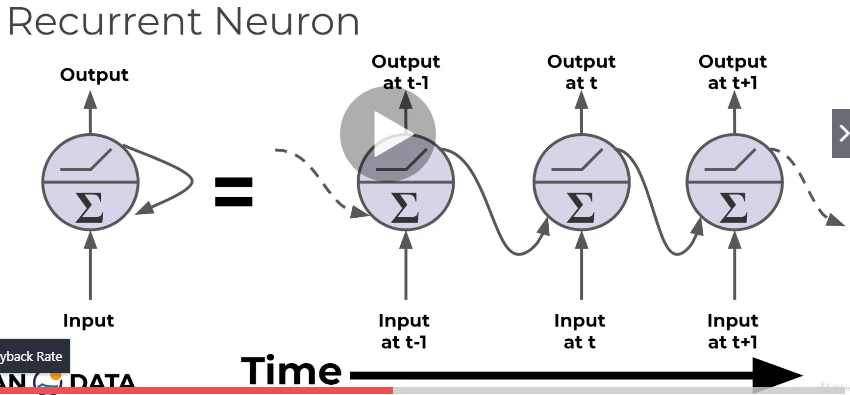
Basically, any data where the ordering of the data is important i.e. sequenced data.

### What is the basic mechanism of how a Recurrent Neural Network?

We need to let the neuron know about its previous history of outputs. We do this by simply feeing its output back into itself as an input.

### How does a recurrent neuron work?

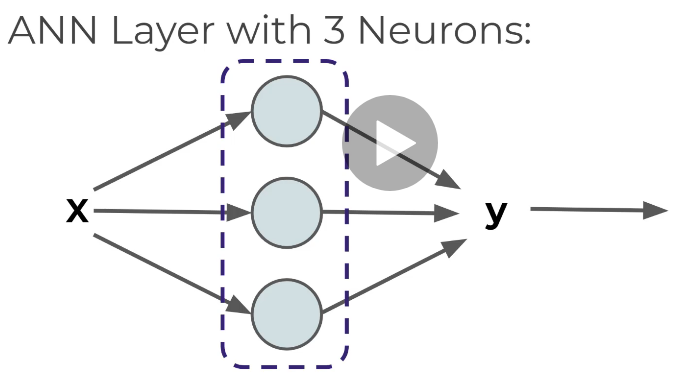


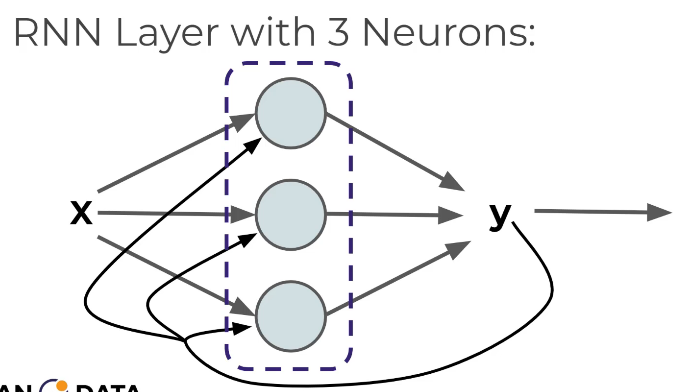


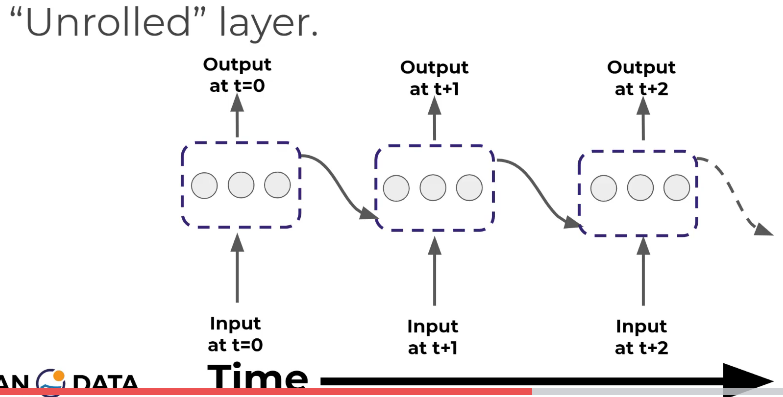
This is the recurrent neuron unrolled throughout time.

### What is the name of cells that have used historical data as inputs?

Memory cells



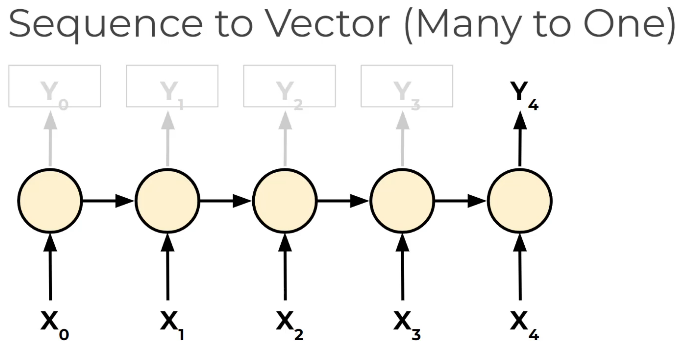




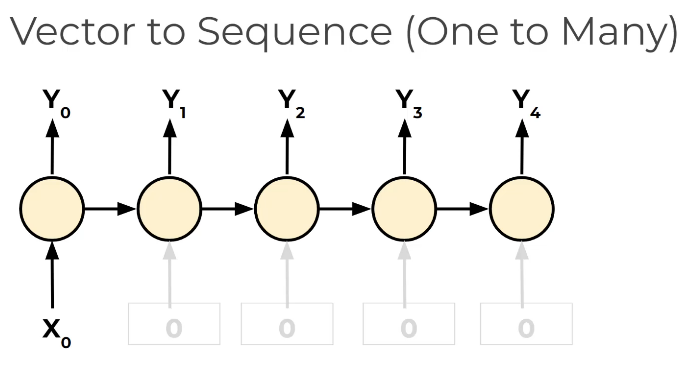
### What the different architectures of inputs and outputs in a RNN?

### 

Given 5 words, output 5 words (same number of input and output)



Given 5 words, output one word



Given one word, predict a sequence of 5 words.

### What is a major disadvantage of a basic RNN?

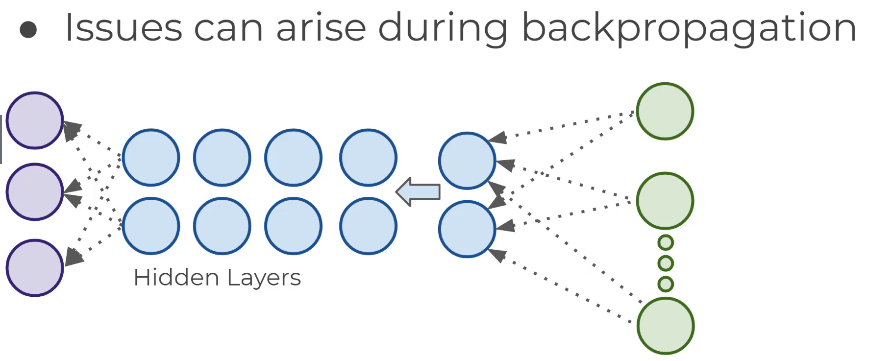
RNN has a short term memory because only the most recent history is input into a neuron.

Another issue is the vanishing gradient.

## Vanishing Gradients

### What are 2 issues that arise when our network gets more complex and deeper?

Exploding gradients and vanishing gradients.

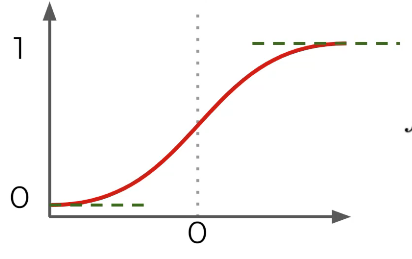


### What are vanishing/exploding gradients?

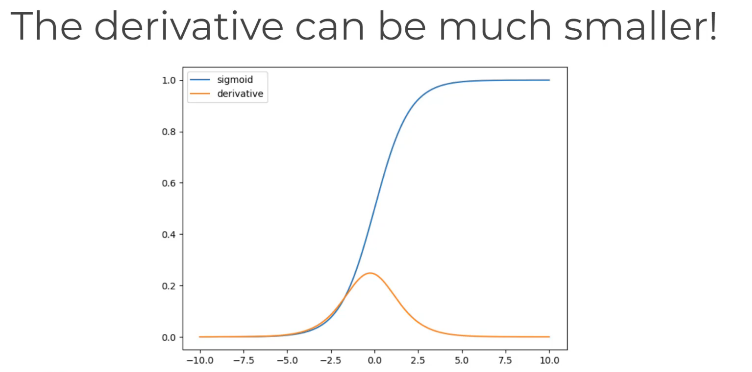
For vanishing gradients, as you go back to the “lower” layers, gradients often get smaller, eventually causing weights to never change at lower levels.

For exploding gradients, the opposite happens

### Why does this occur?



For the sigmoid function, if the values are close to 0 or 1, the derivative of the function is close to 0, as shown below.



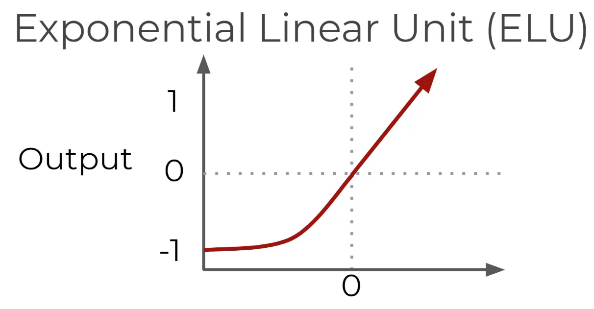
The derivatives can be very small as it moves away from 0. Then this doesn’t affect the weights and biases in the initial layers and the model backpropagates.

The weights and biases of the initial layers is very important in getting the model to work well.

### How can we fix vanishing gradients?

We can use the Relu function because it doesn’t saturate positive values.

Or use the exponential linear unit (ELU)



Or use batch normalization.

Or choose different initialization of weights, such as Xavier initialization.

### How can we fix exploding gradients?

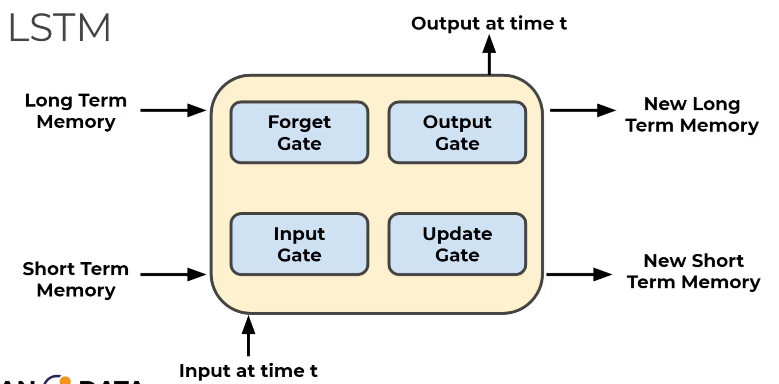
A quick and dirty trick is known as gradient clipping where gradients are cut off before reaching a predetermined gradient.

## LSTMs GRUs

### What does LSTM stand for?

Long Short-Term Memory cell

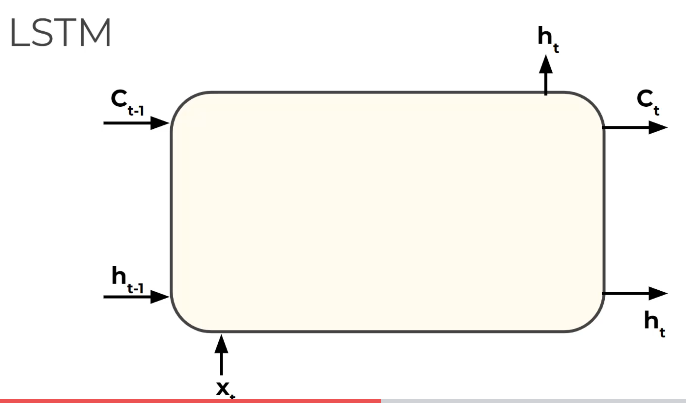
### What are the differences of a LSTM vs normal?



### What is a gate?

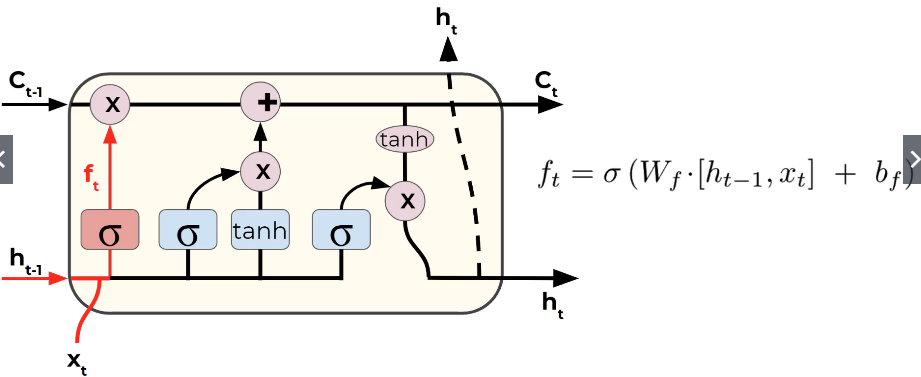
Lets some information through, based on a function, usually sigmoid.

### What are the symbols for all the inputs and outputs of a LSTM?



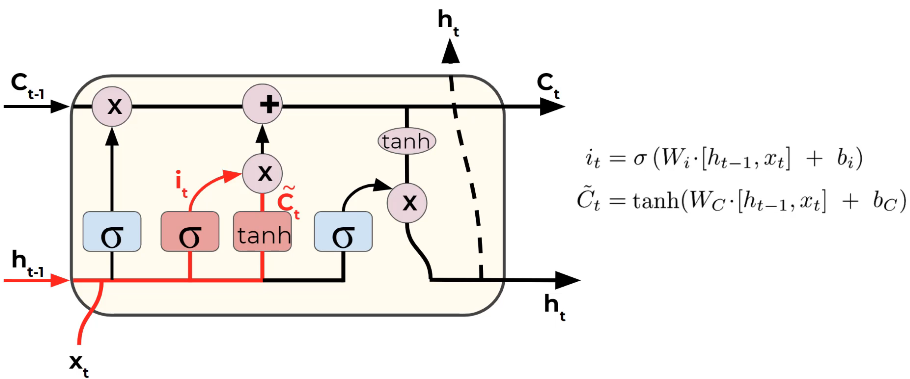
### What is the first step of the LSTM?

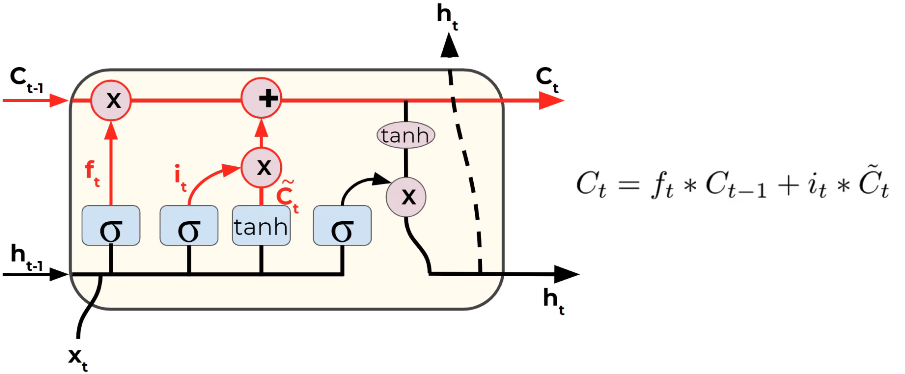
What are we going to forget?

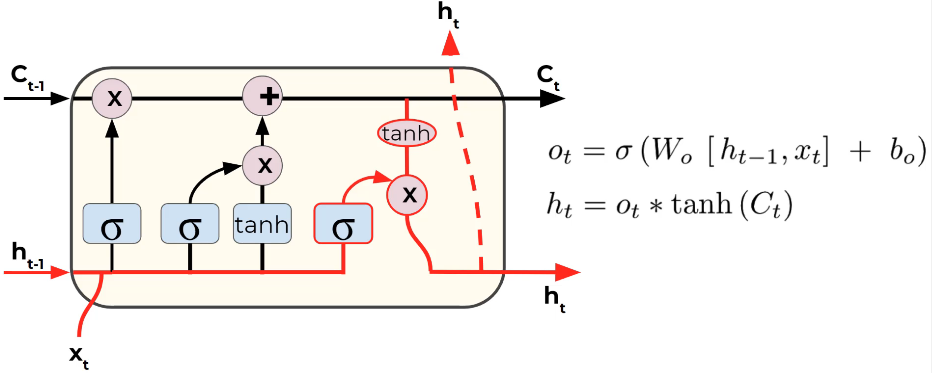


### What is next?

What is the new information we’re going to store in the cell state?



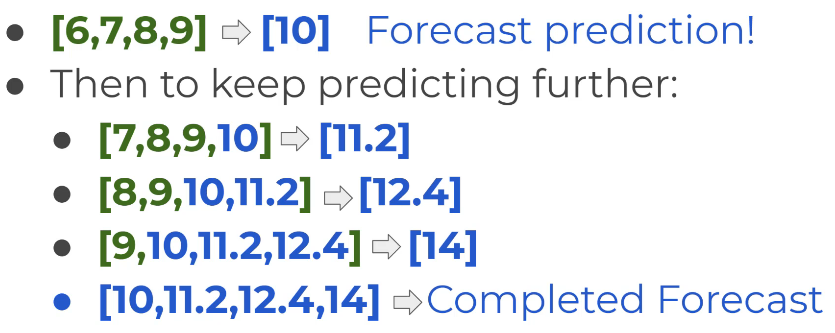




### What are the other variations of the normal LSTM?

Peephole, Gated Recurrent Unit (GRU)

## RNN Batches



### Does the data for RNN need to be scaled?

Yes.

### What does TimeseriesGenerator do?

Defines the length of the data to use for forecasting, and the length of data to be forecasted.

The variables are length and batch size.

### What does the ‘patience’ parameter in EarlyStopping do?

How many epochs are you going to keep waiting until the early stopping takes into effect.

### When forecasting from the entire dataset, why can’t you do early stop call back?

Recall that early stopping works by assessing the validation loss. This validation loss is calculated by a set of data that is set aside just for this purpose. But if you’re using the whole data set to train the model, then there isn’t anything left to validate, otherwise that would be cheating by using some data for validation and training.

### So then how do you choose the number of epochs when training with the full data set and you can’t implement early stopping?

The number of epochs can be chosen from the loss value when using the validation data set that also included early stopping. By looking at the graph, you can estimate the number of epochs needed to get to the lowest validation loss.

## Multivariate time-series

### What are some other simpler models that can deal with time-series data?

SARIMAX and VARMAX

### What changes are required for multivariate time series?

2 things need to be changed.

1. Change input shape in LSTM layer to reflect a 2D structure
2. Final dense layer should have a neuron per feature/variable