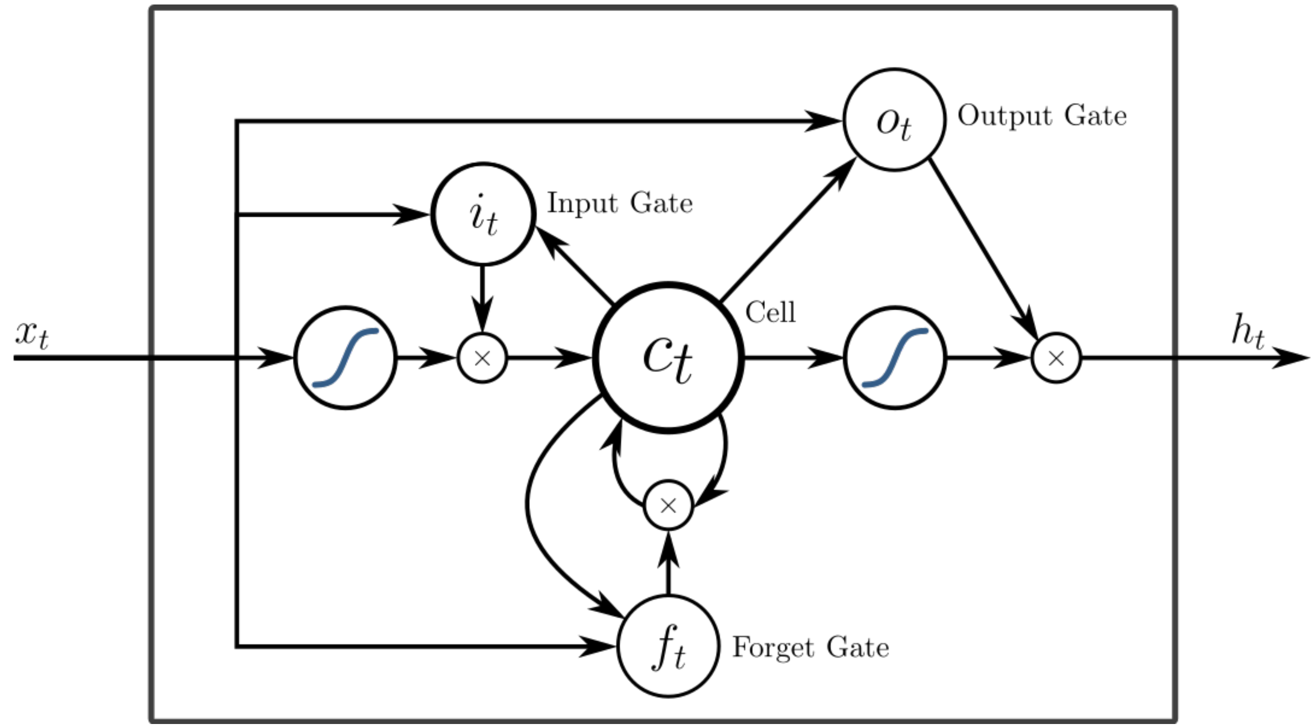
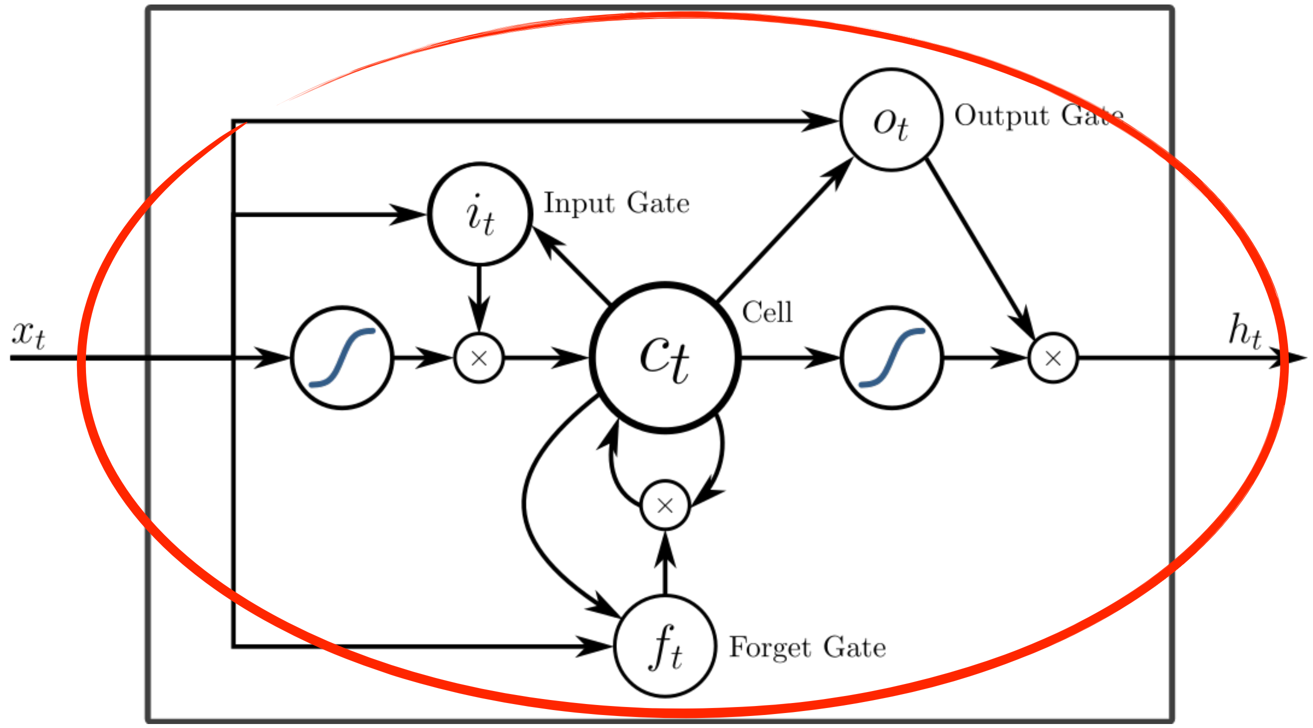
LSTM

LSTMs are so powerful that we dedicate an entire lecture on how they are working.

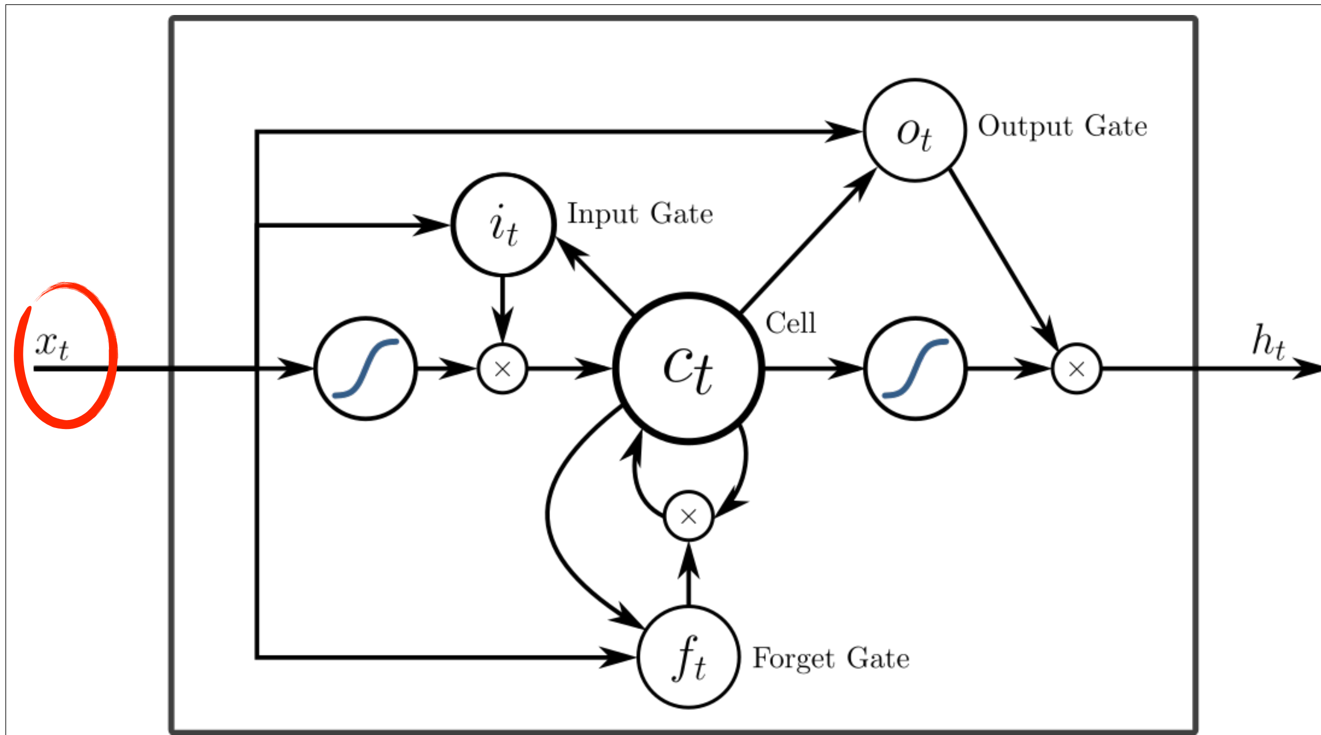


You could take an entire course on LSTMs and if you are planning to do so, please check out the description of this video. But we try to give you a more intuitive view on LSTMs.

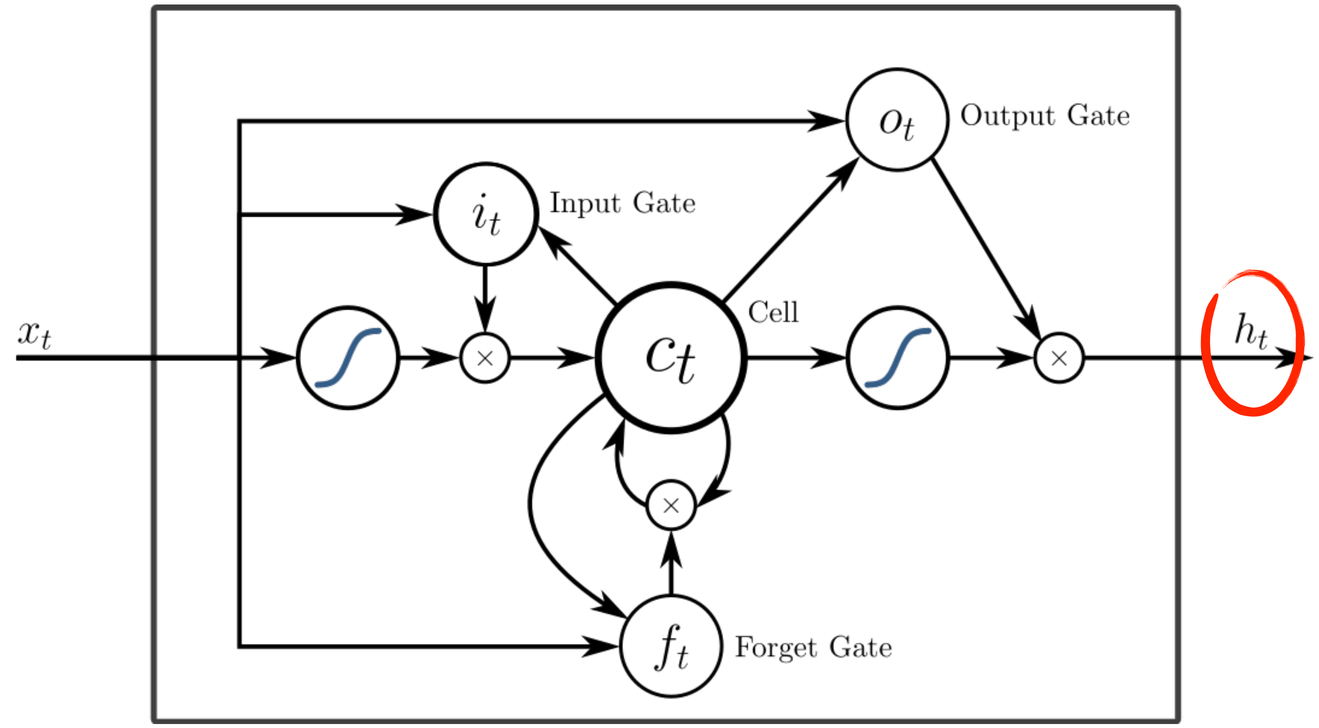


By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

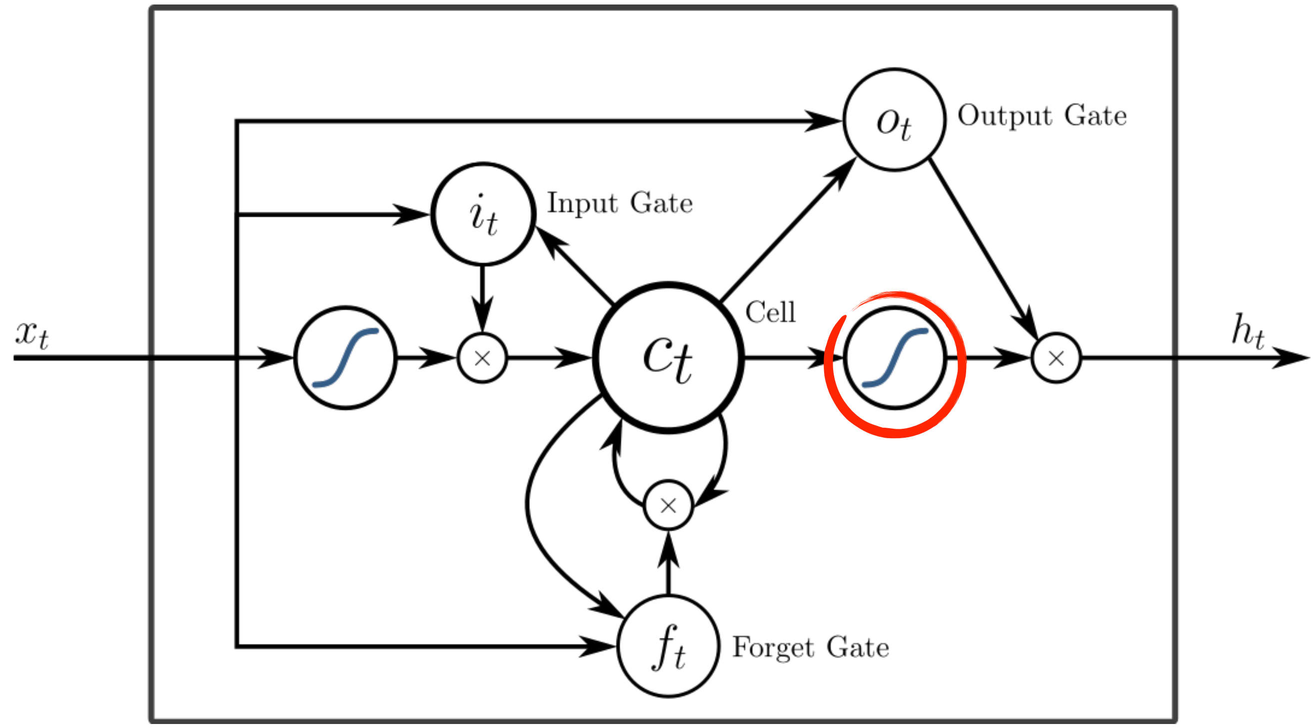
So this is a single neuron.



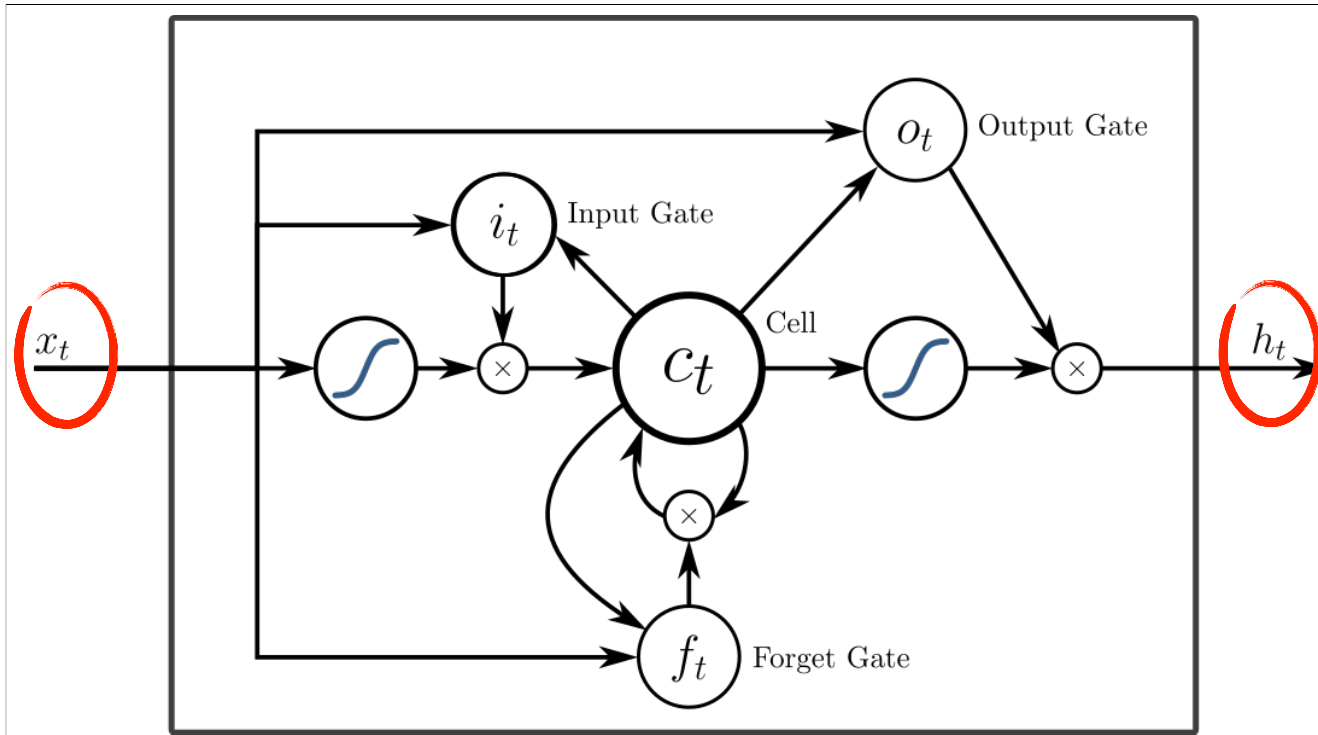
And as in a feed forward network it maps an input vector x t



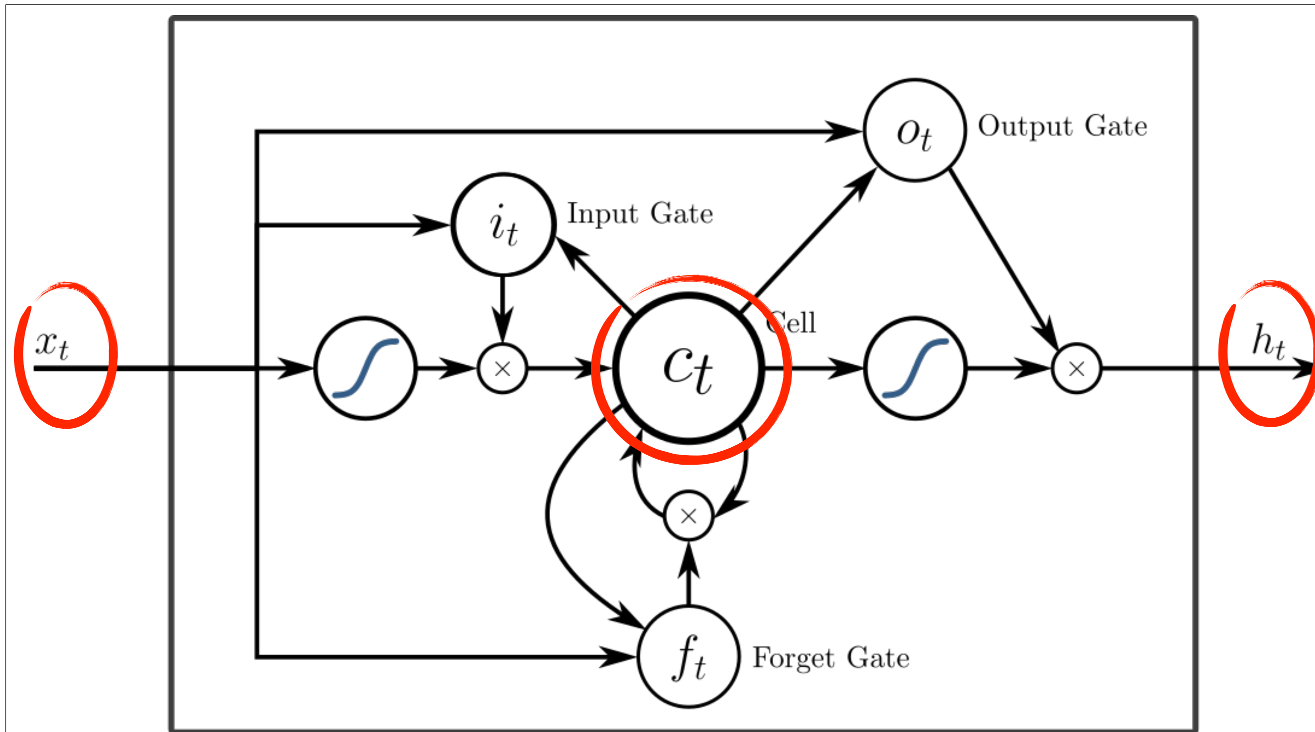
to an output vector h t by using weights and an



to an output vector h t by using weights and an activation function. Note that the same hold independently of wether we use scalars or vectors as input and output. But we also see a lot of additional components.

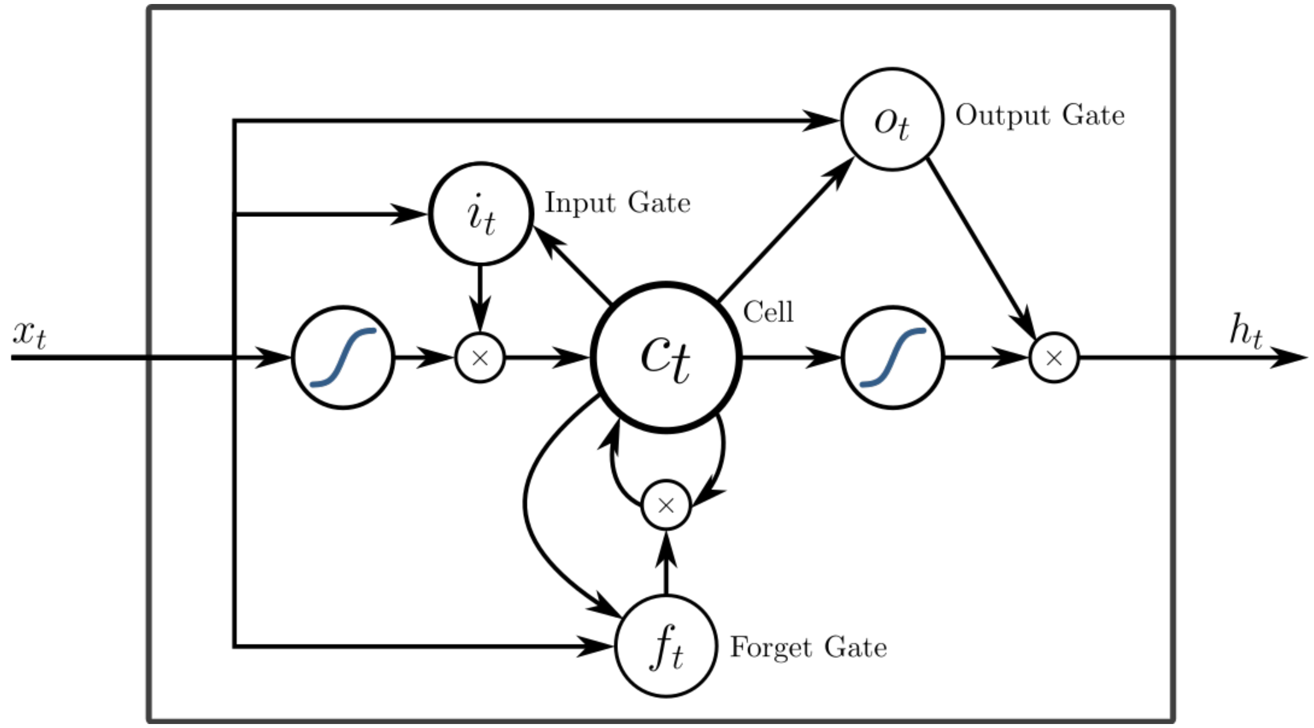


So the first thing we notice is that there is no direct connection between x t and h t.



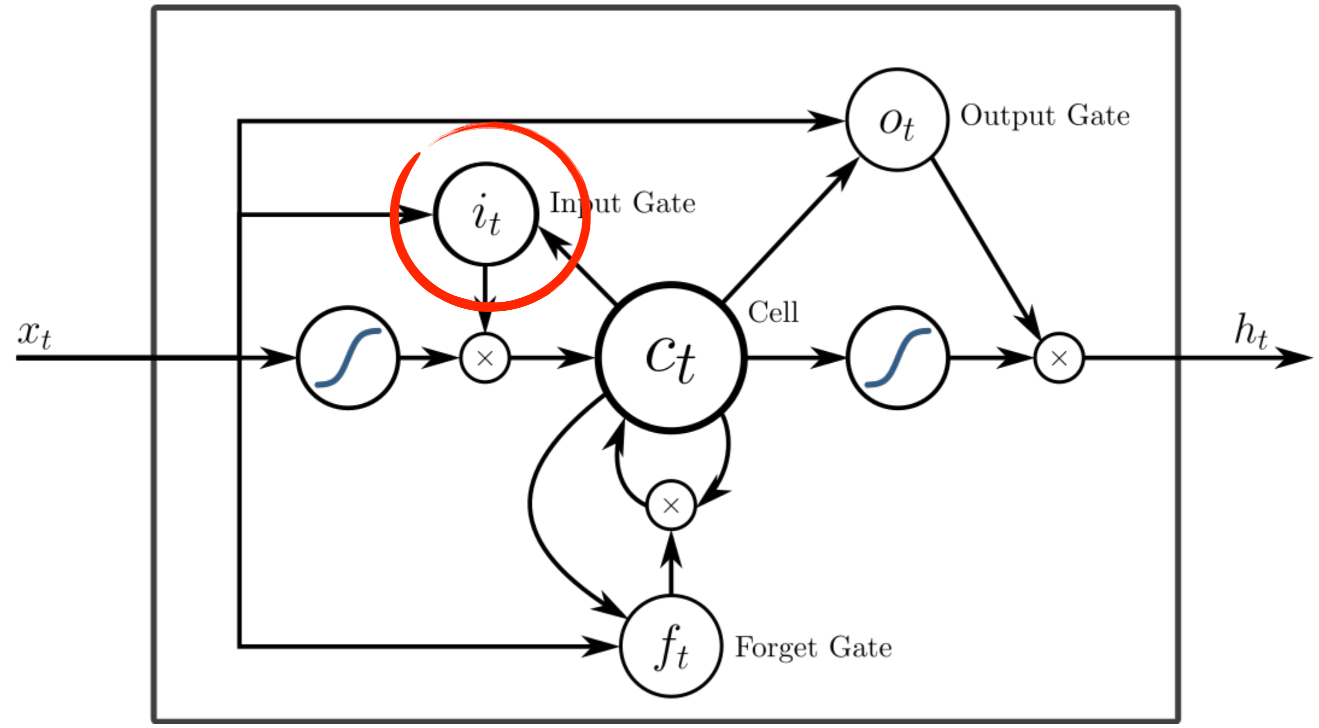
By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

All data flows though c t which is the so called cell state. Cell state is the actual memory of the LSTM neuron

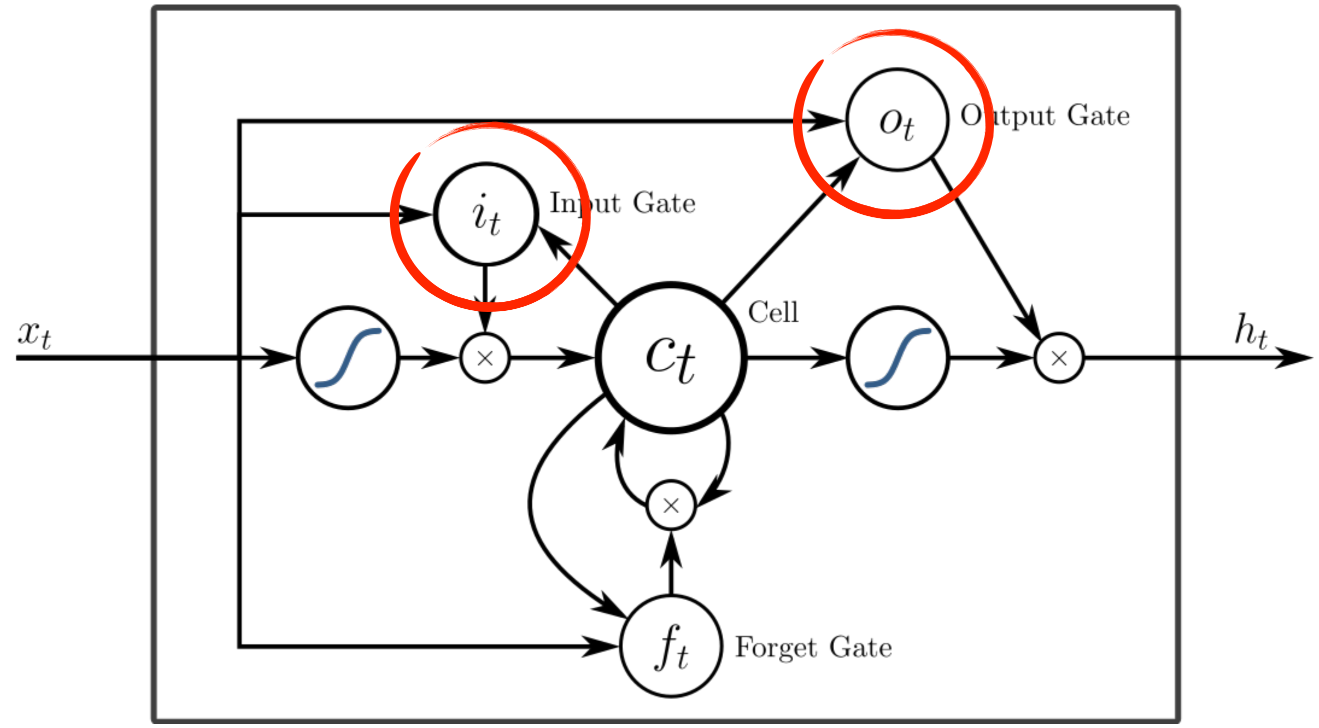


By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

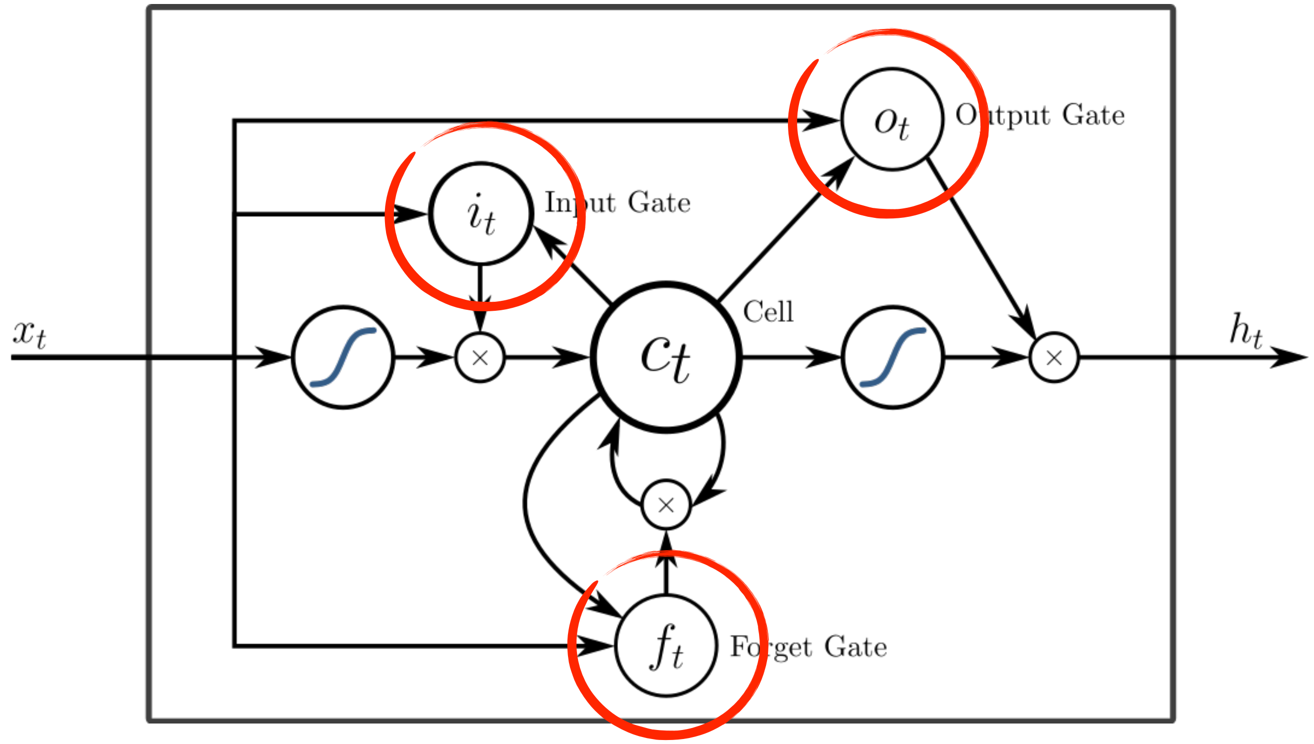
Please notice that there are three additional units present in the LSTM neutron.



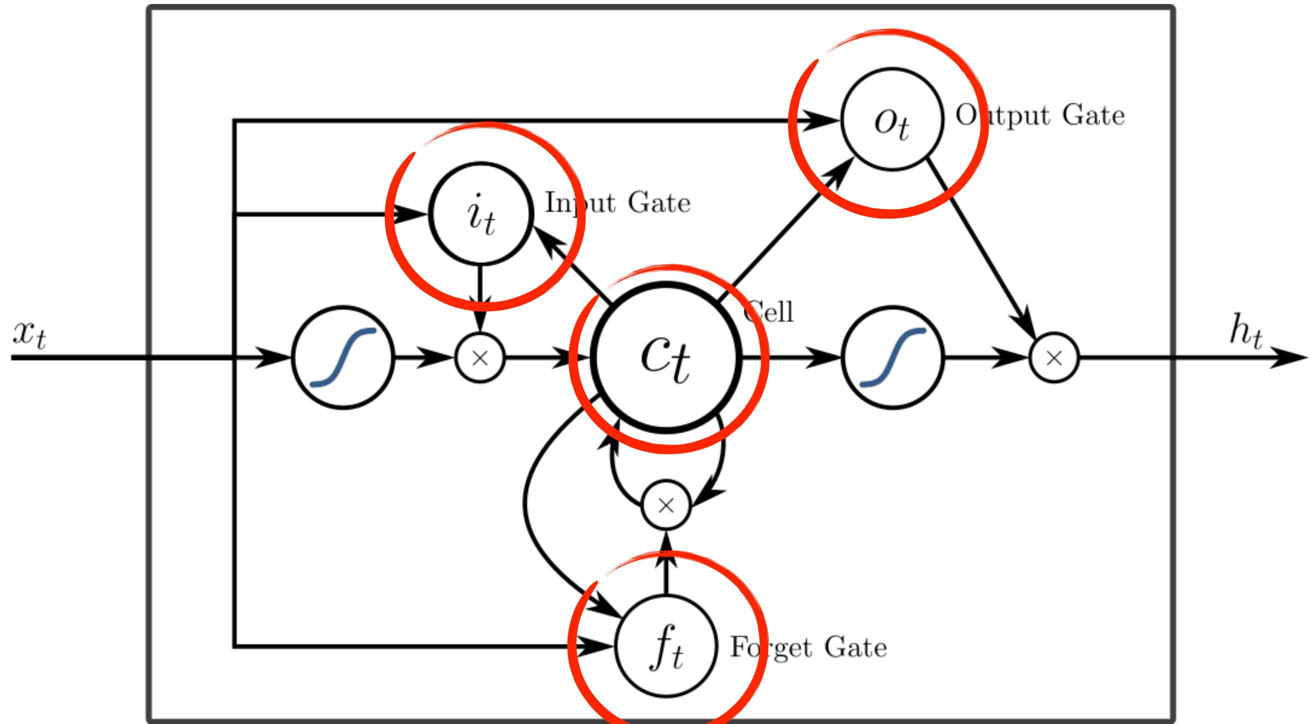
An input gate…



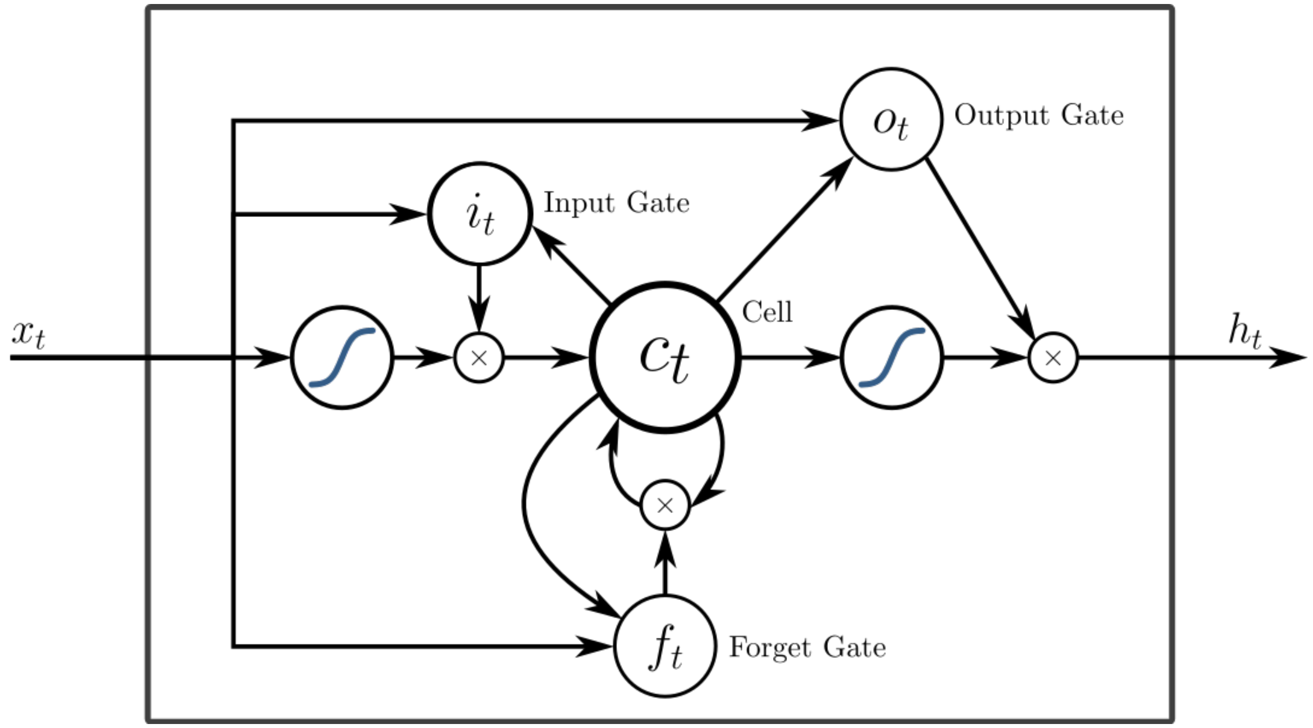
…an output gate…



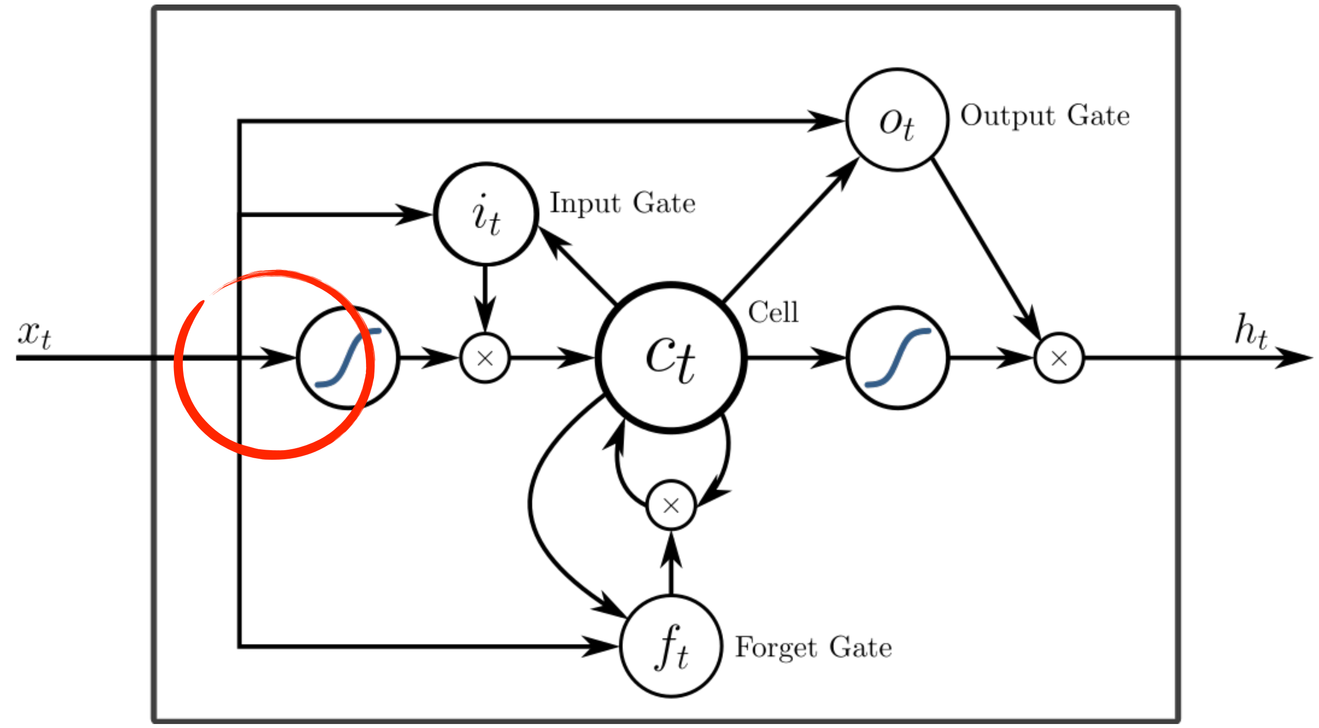
…and a forget gate.



Those three gates are controlling the state of c t.

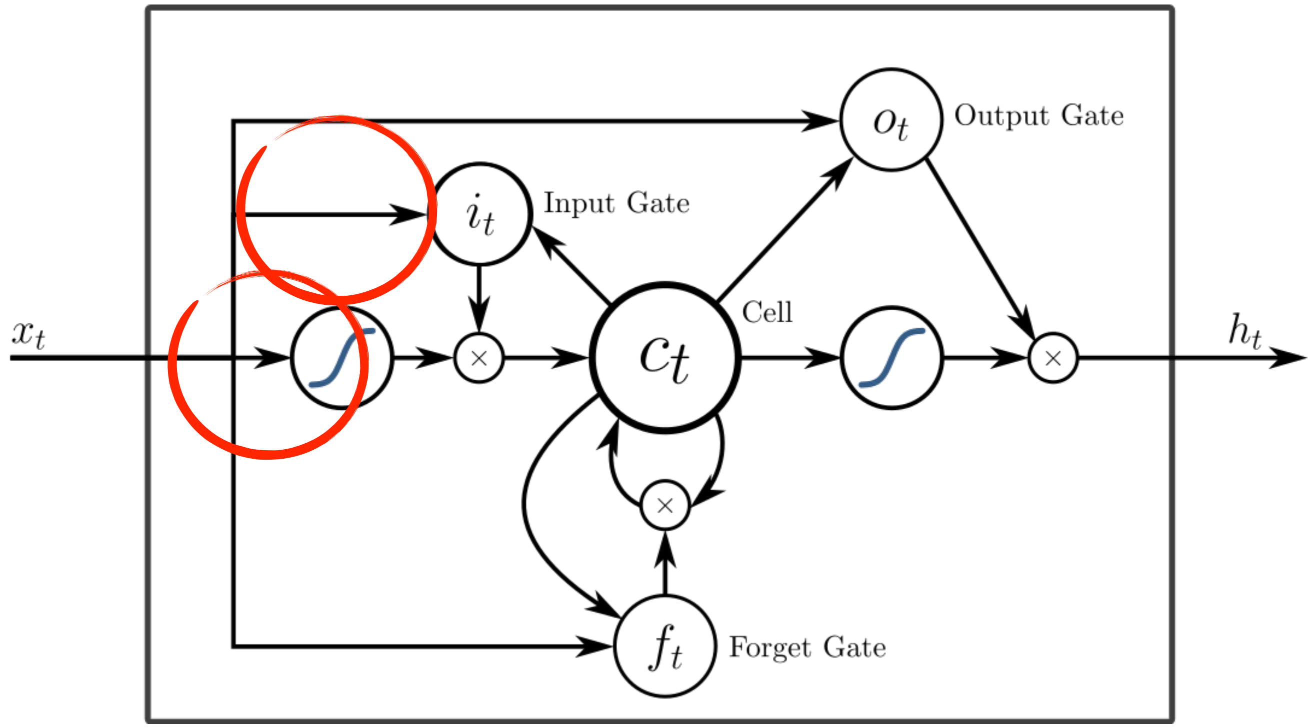


The way how this is controlled is as follows. So look at the input first.

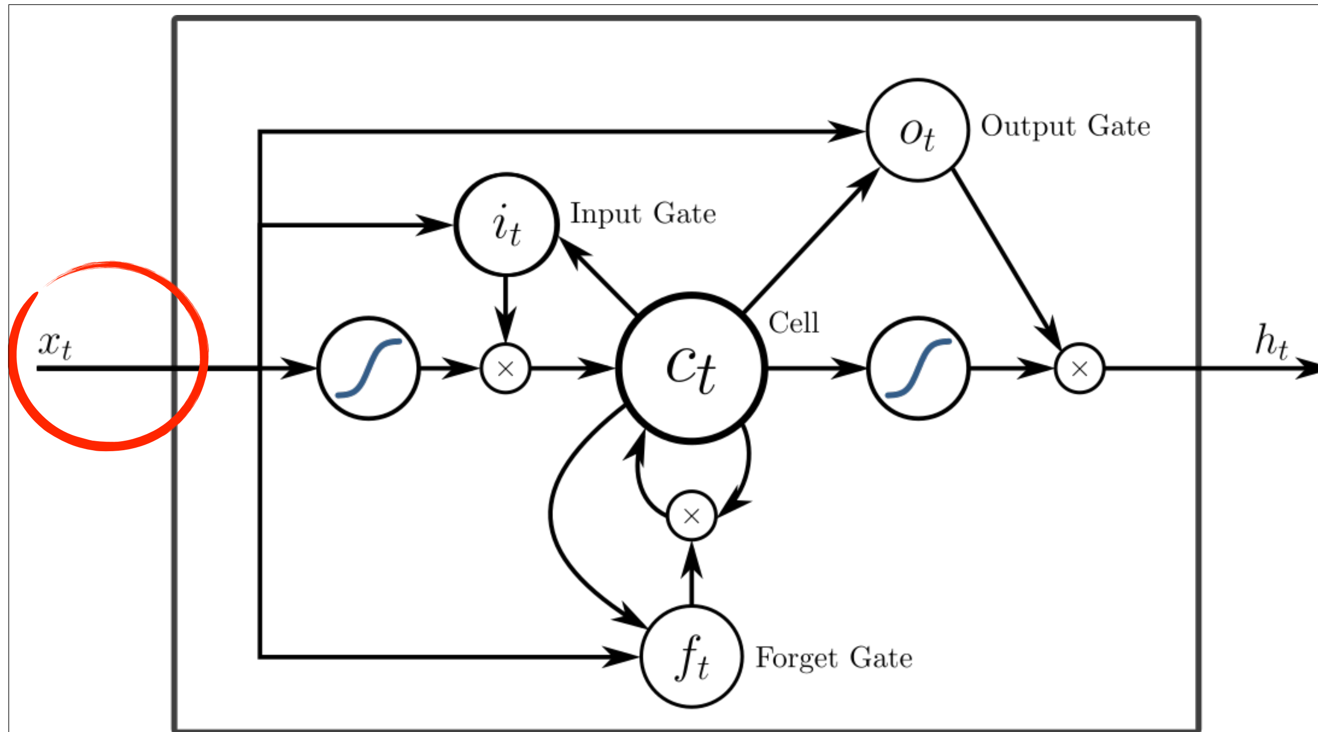


By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

The first thing we notice is that x t is not only used as input of the neuron but also…

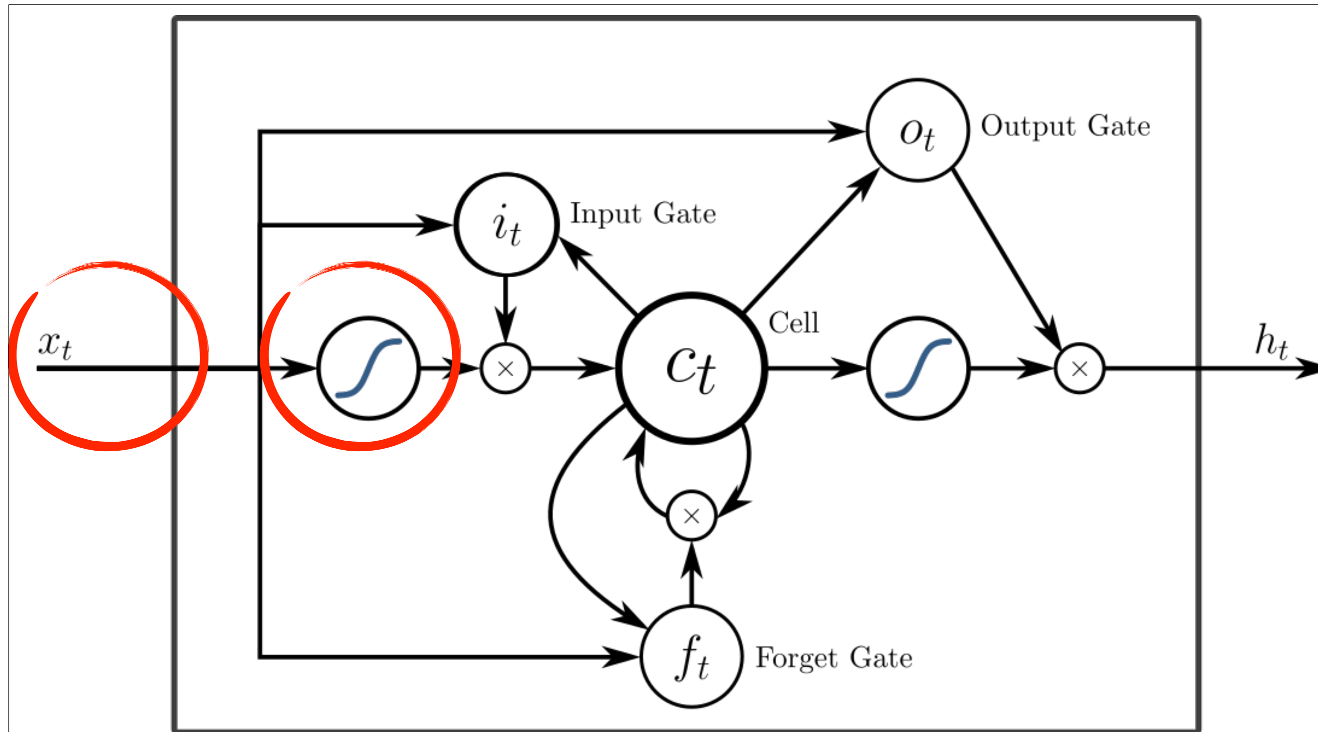


..as input to the input gate. So the input gate, as the other gates, has a separate weight vector which is trained from the input data and learns to control the influx of information into cell state c t. This is done by a vector dot product between the

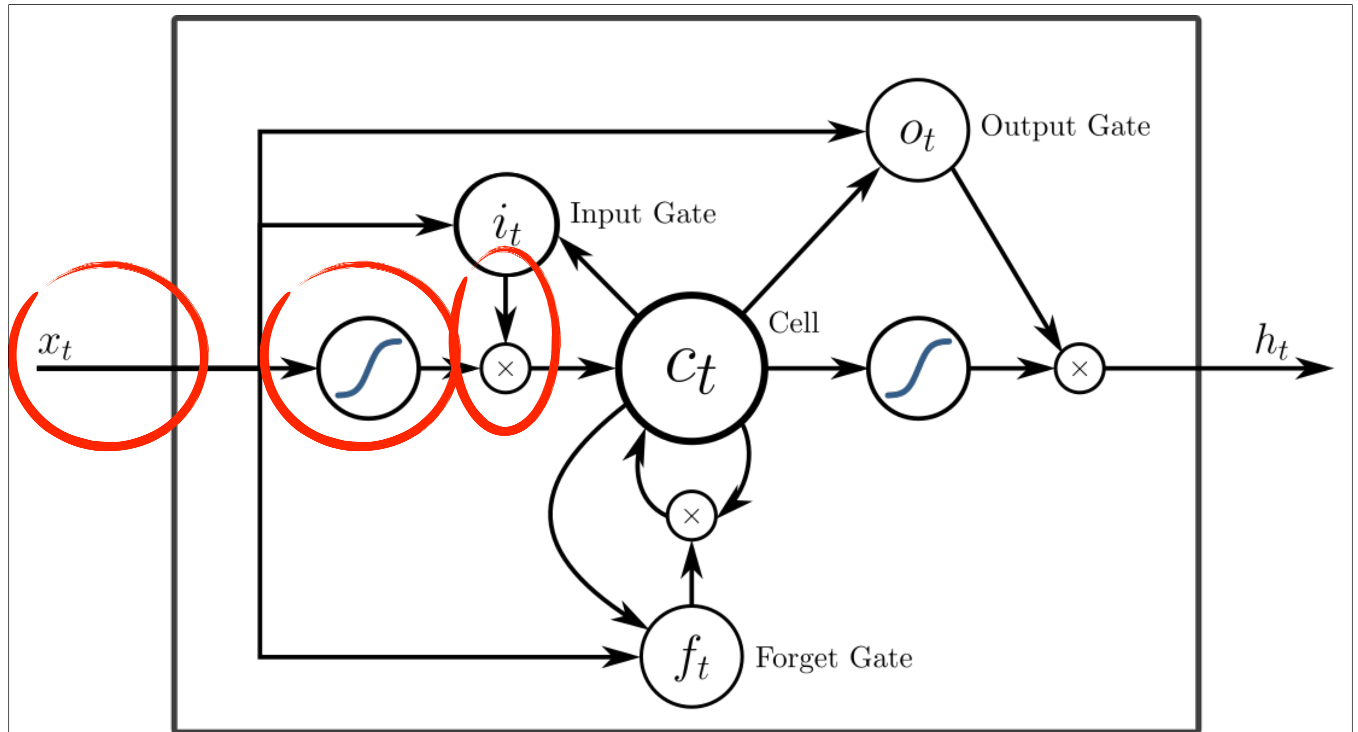


By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

.. input x t after it has been squashed by the activation function

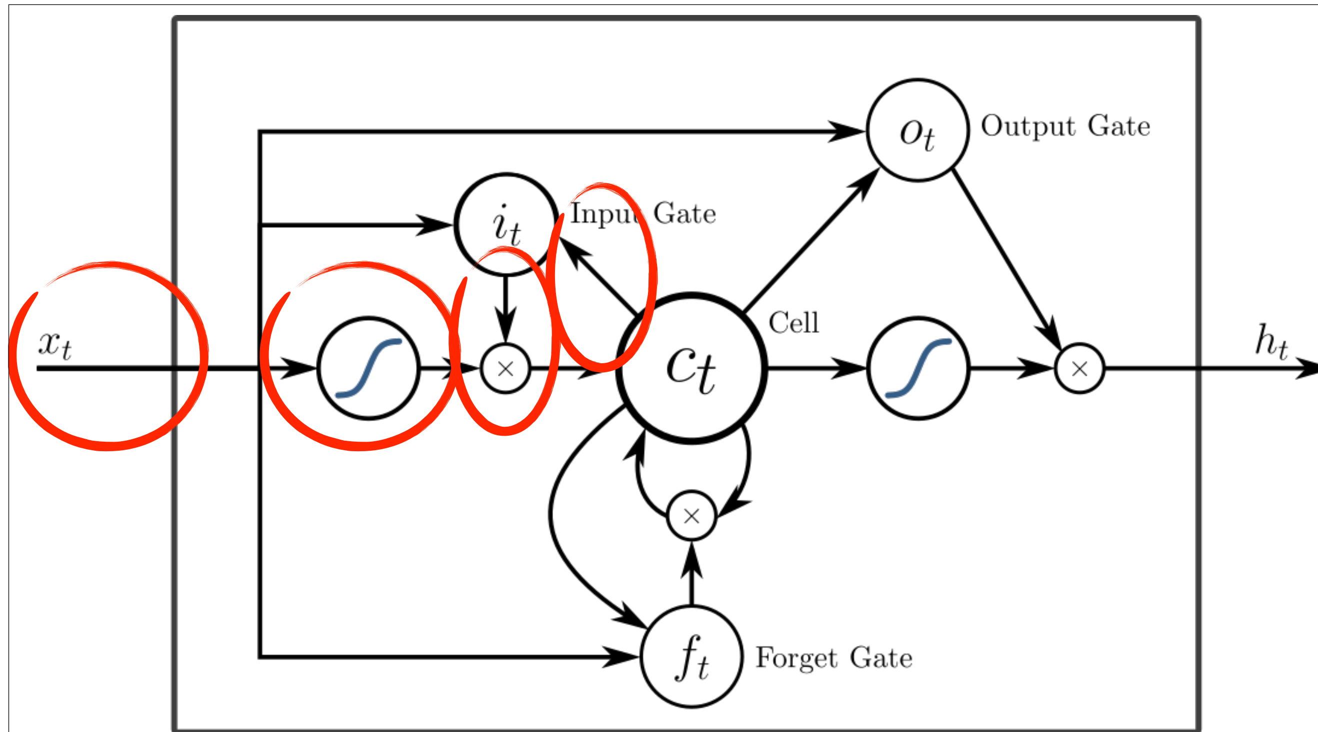


and the output of the input gate

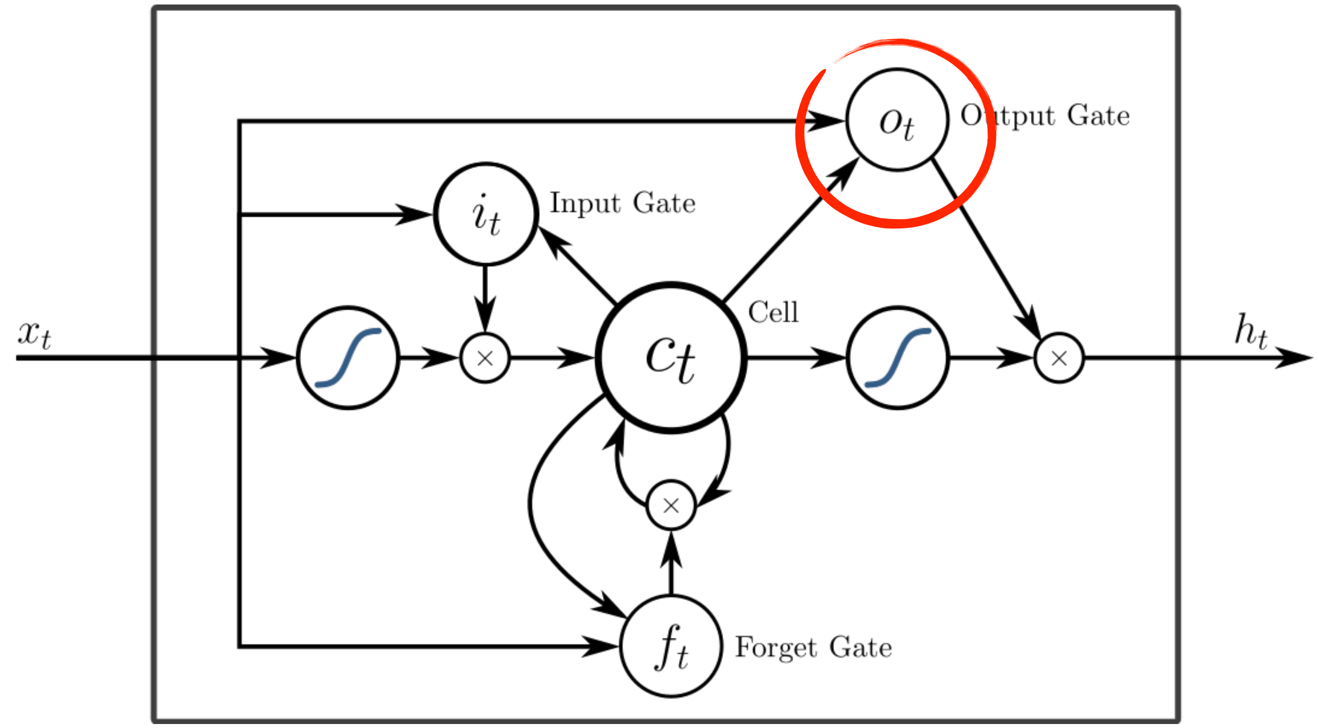


By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

and the output of the input gate. In other words, through the weight vector of the input gate the neuron can learn from the training data when it is a good idea to open the gate and have the input stored in the cell or when it is a bad idea to remember things and close the influx information into cell state c t. Note that this is a continuous value, so you can see it as a valve which can be partially opened and closed.

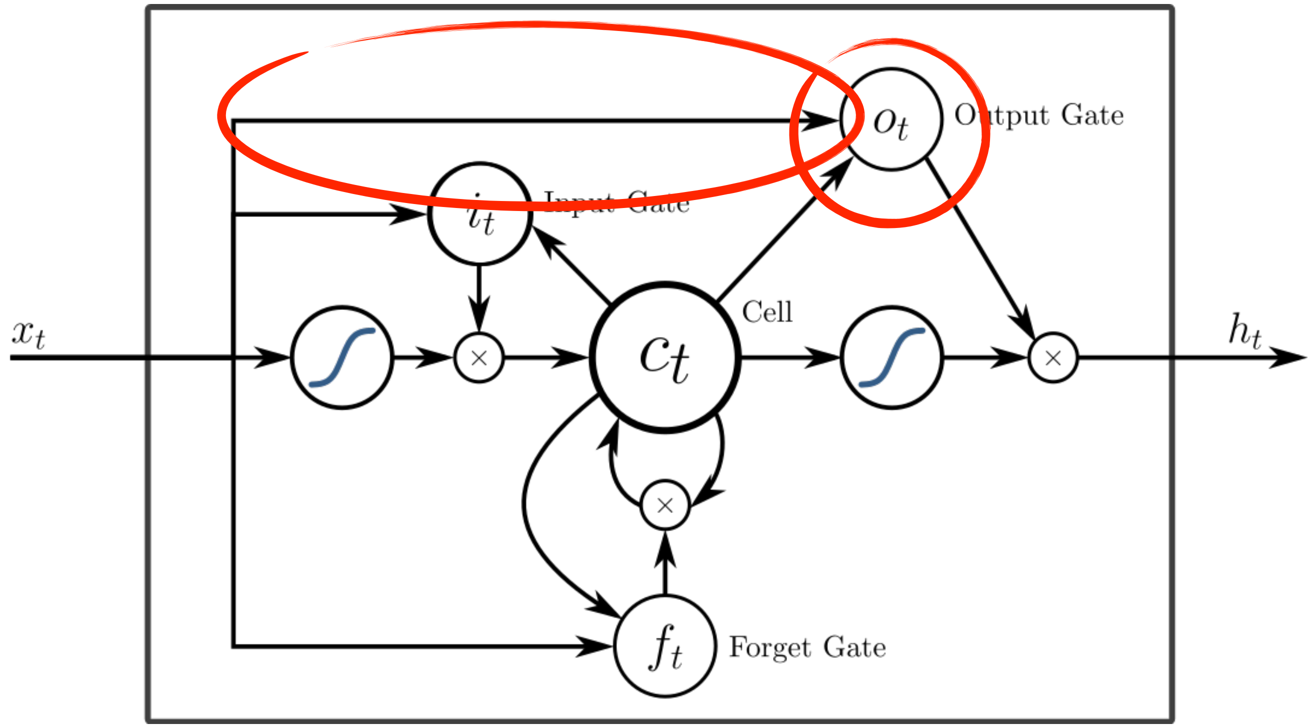


Finally it is important to notice that all the cell state has an influence on the gate. This is again accomplished through a separate weight vector, so that the actual input gate is controlled by the historic cell state as well as by the actual value x t.



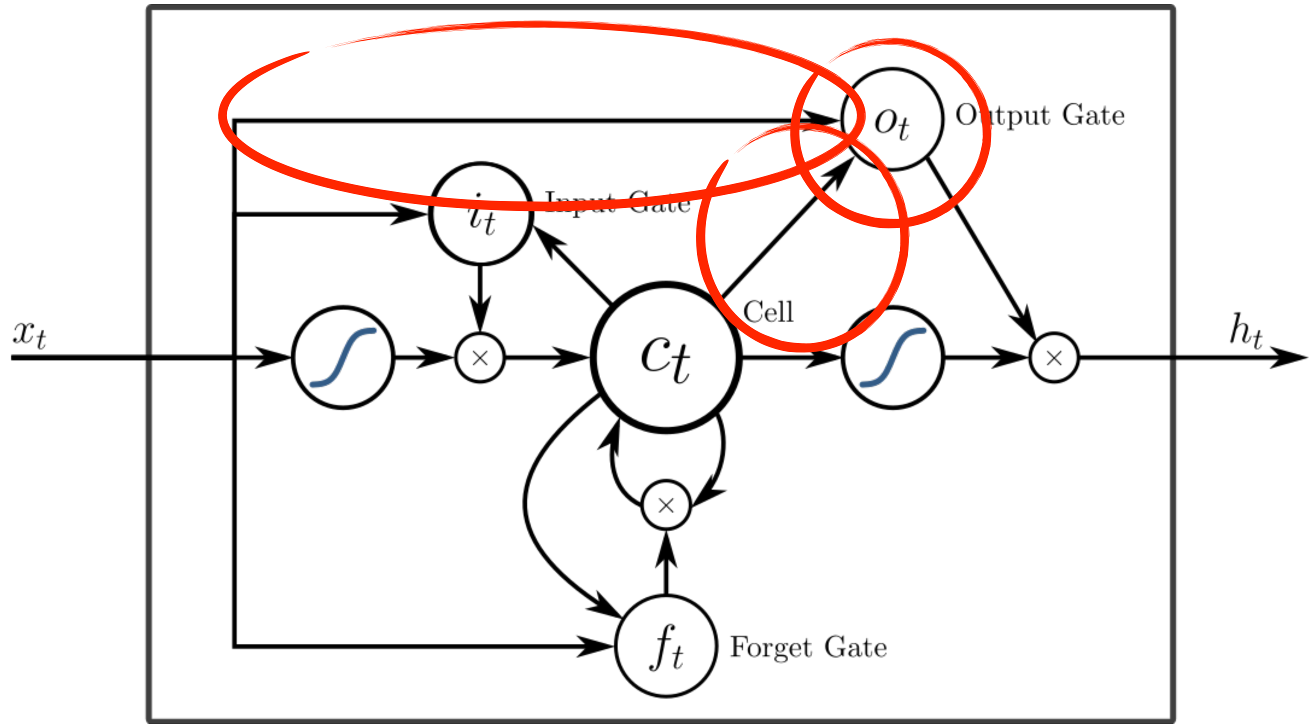
By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

So now let’s have a look at the output gate.

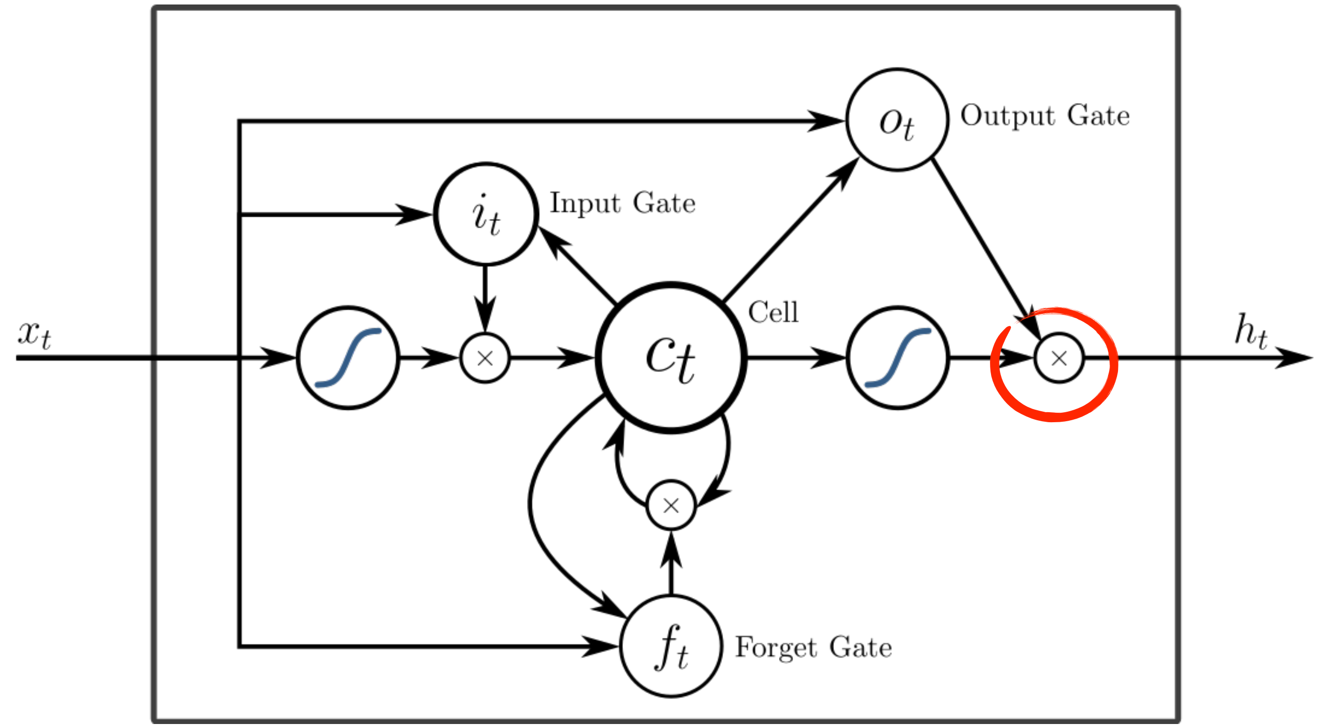


By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

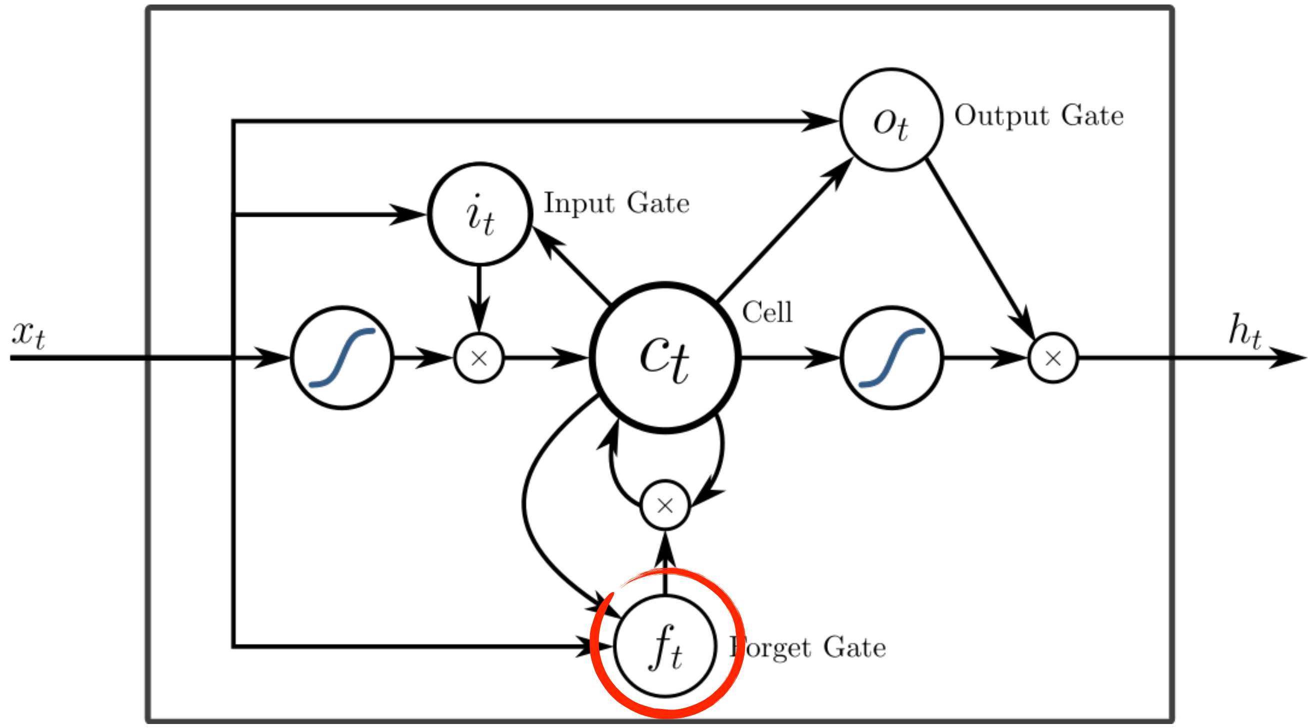
Again it is controlled by the actual input value x t



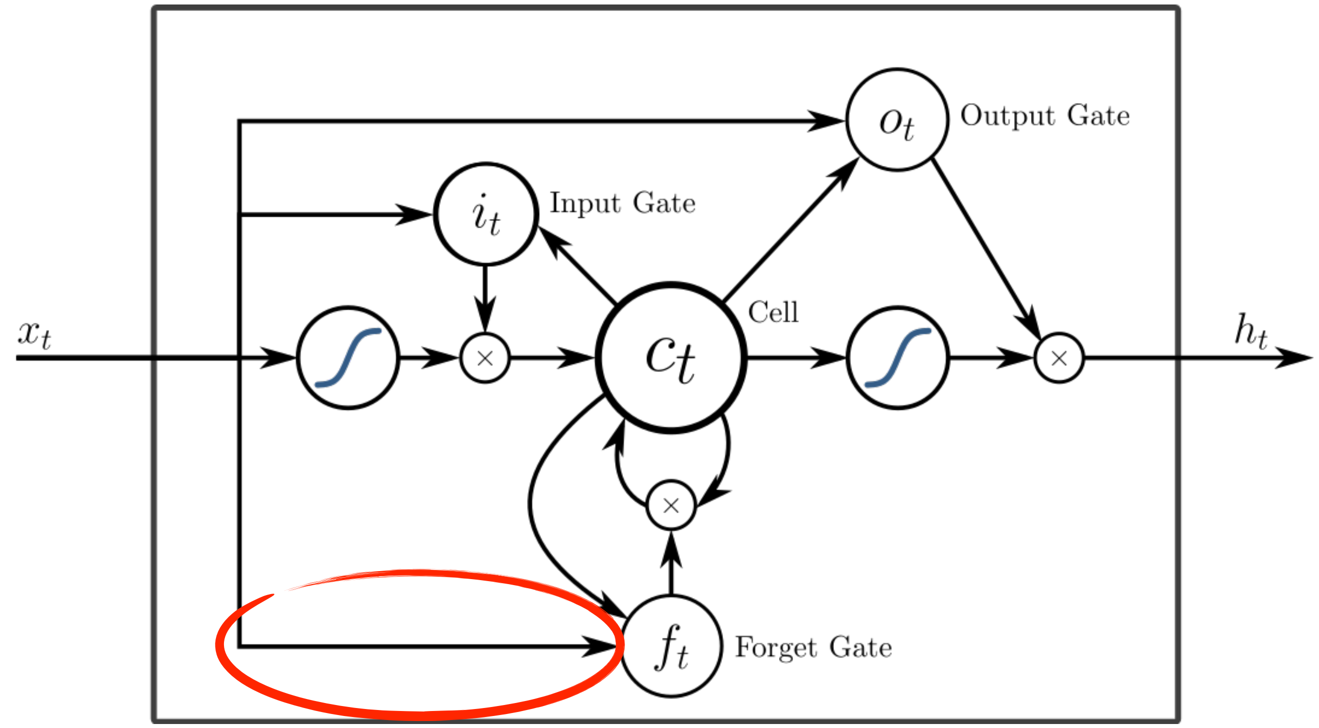
and by the actual cell state c t.



Here the output gate controls how much of cell state c t get’s output to downstream neurons connected to h t. So this topology is the initial LSTM proposed by Sepp Hochreiter and Juergen Schmidhuber in 1997.

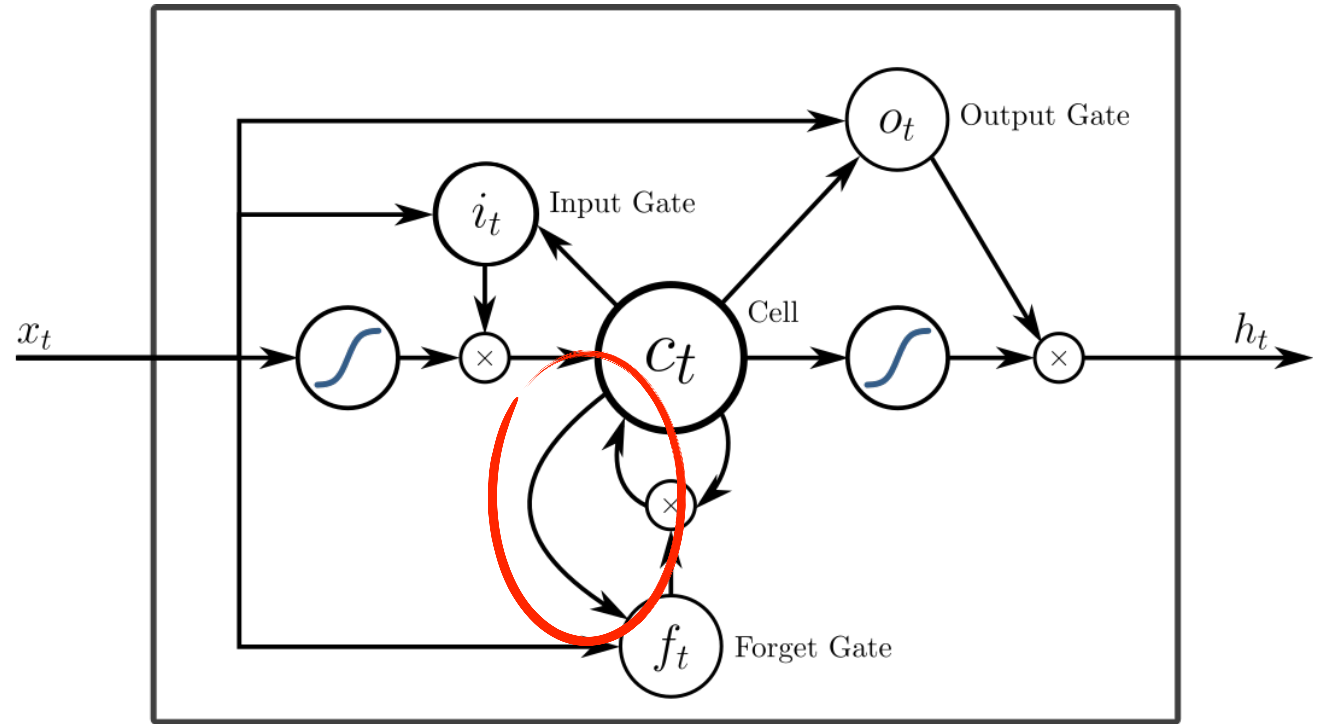


In 1999 Felix Gers , Jürgen Schmidhuber and Fred Cummins added an additional component. The forget gate. They discovered that without the capability of forgetting the cell state c t may grow indefinitely and eventually cause the network to break down.

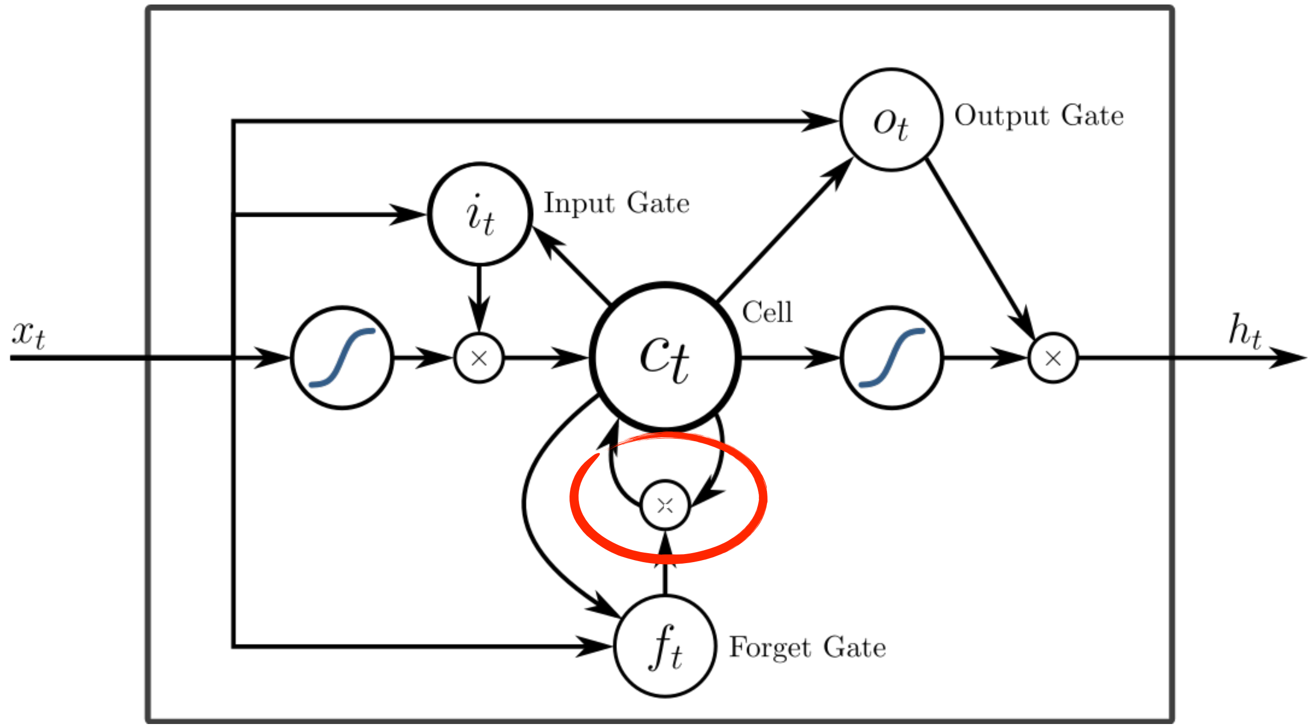


By https://commons.wikimedia.org/wiki/User:BiObserve (Raster version previously uploaded to Wikimedia) Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton (original) Eddie Antonio Santos (SVG version with TeX math) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

Again the forget gate is controlled by the actual input x t …



…and the current cell state c t.



…and again trough calculation of a dot product between the output of the forget gate and the previous cell state c t it controls how much of the actual cell state c t is preserved.

Autoencoder

Another exotic but totally exiting neural network topology is an autoencoder, so let’s learn about it in the next lecture.