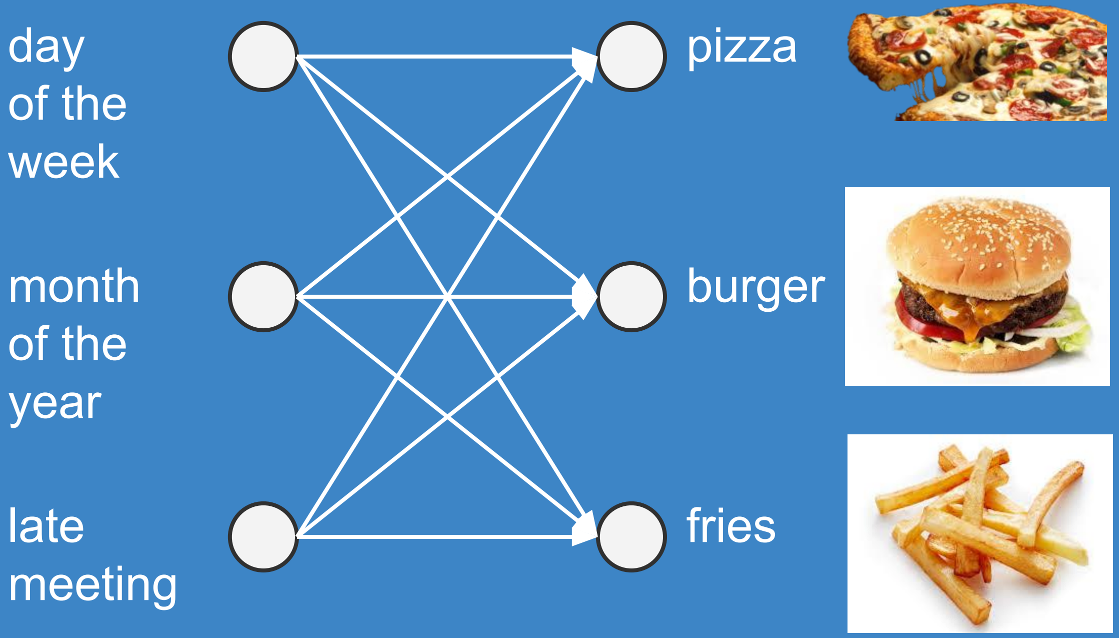
IDL

INTRODUCTION TO DEEP LEARNING

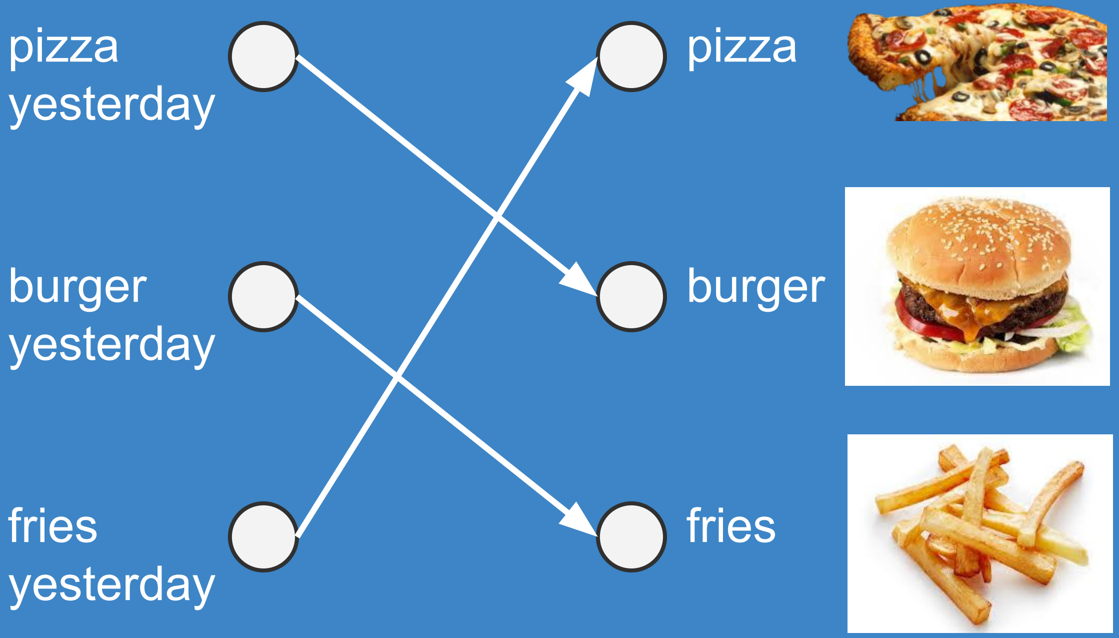
Recurrent Neural Networks (RNNs)

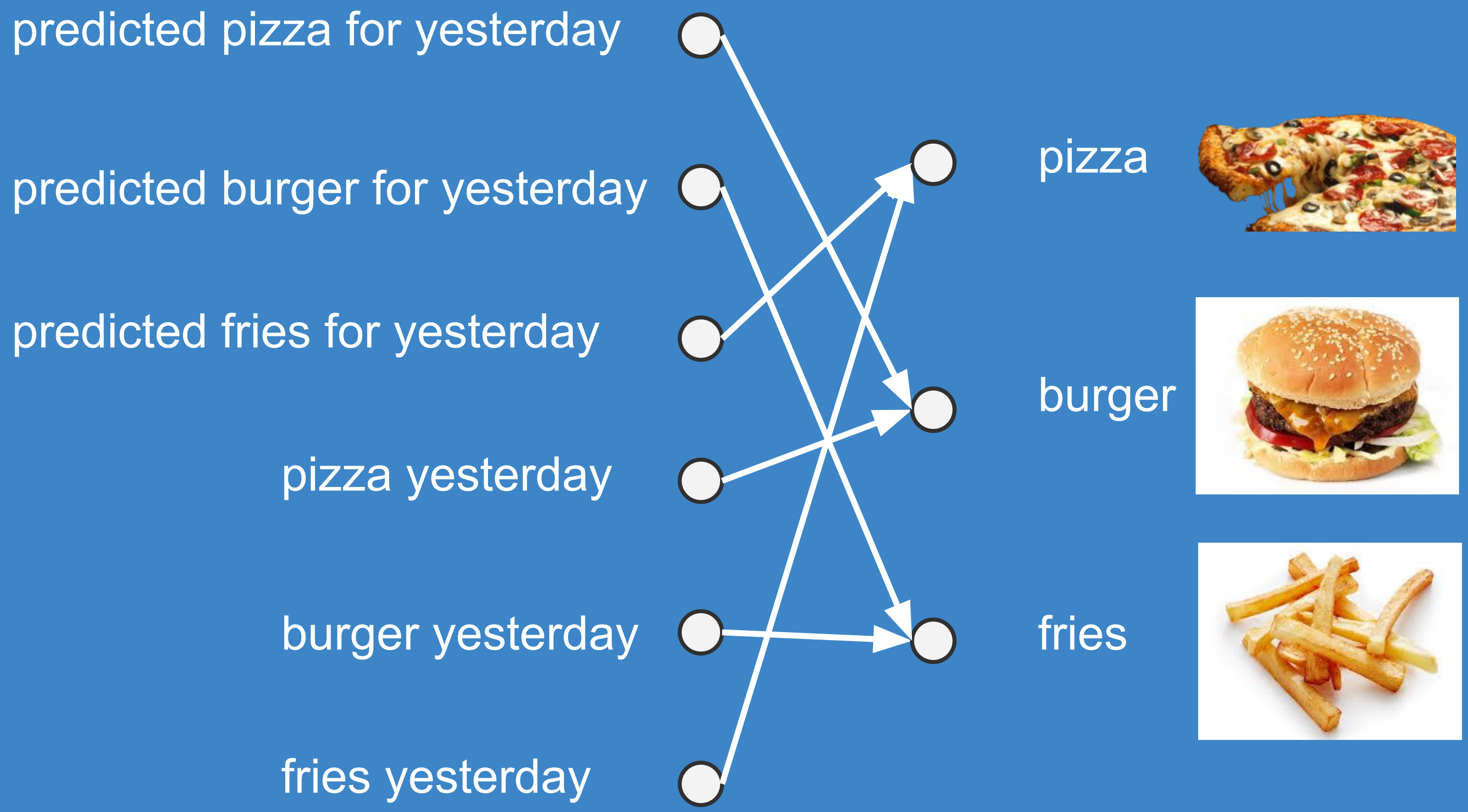
Long Short-Term Memory (LSTM)

What's for dinner?



What’s for dinner?





predicted pizza for yesterday
predicted burger for yesterday
predicted fries for yesterday
pizza yesterday
burger yesterday
fries yesterday
pizza
burger
fries

A vector is a list of values

"High is 67 F.
Low is 43 F.
Wind is 13 mph.
.25 inches of rain.
Relative humidity
is 83%."

=

High temperature	67
Low temperature	43
Wind speed	13
Precipitation	.25
Humidity	.83

=

Weather vector
67
43
13
.25
.83

A vector is a list of values

“It’s Tuesday” =

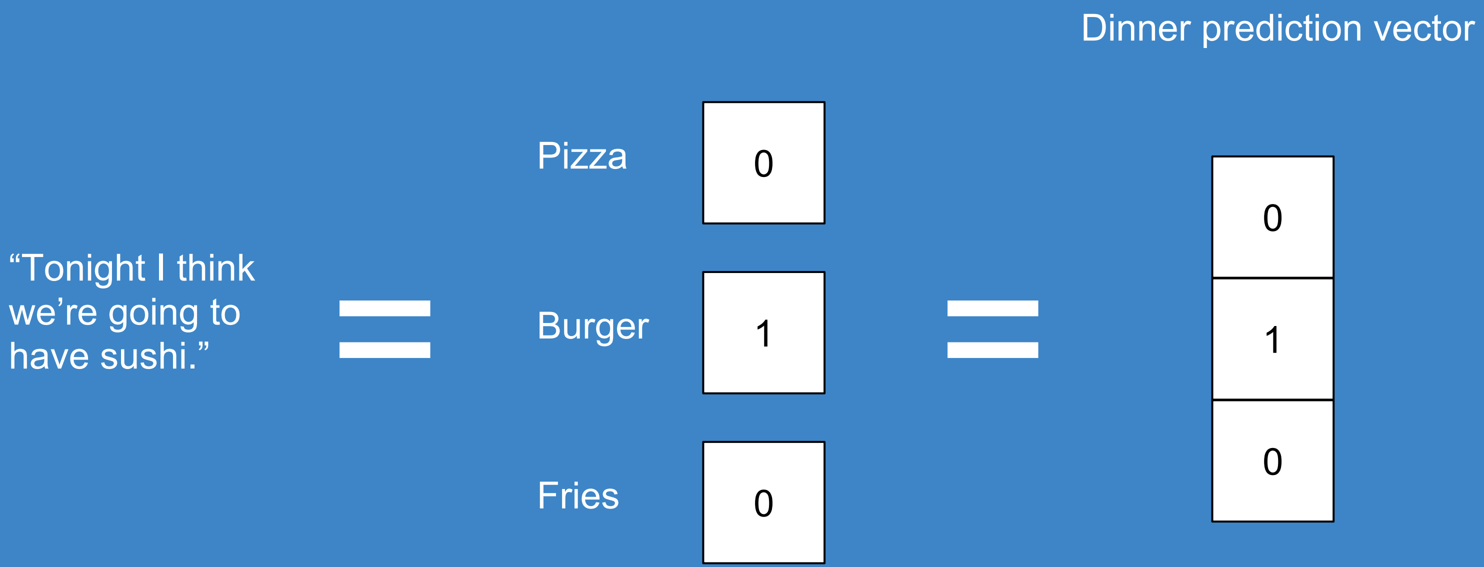
Sunday	0
Monday	0
Tuesday	1
Wednesday	0
Thursday	0
Friday	0
Saturday	0

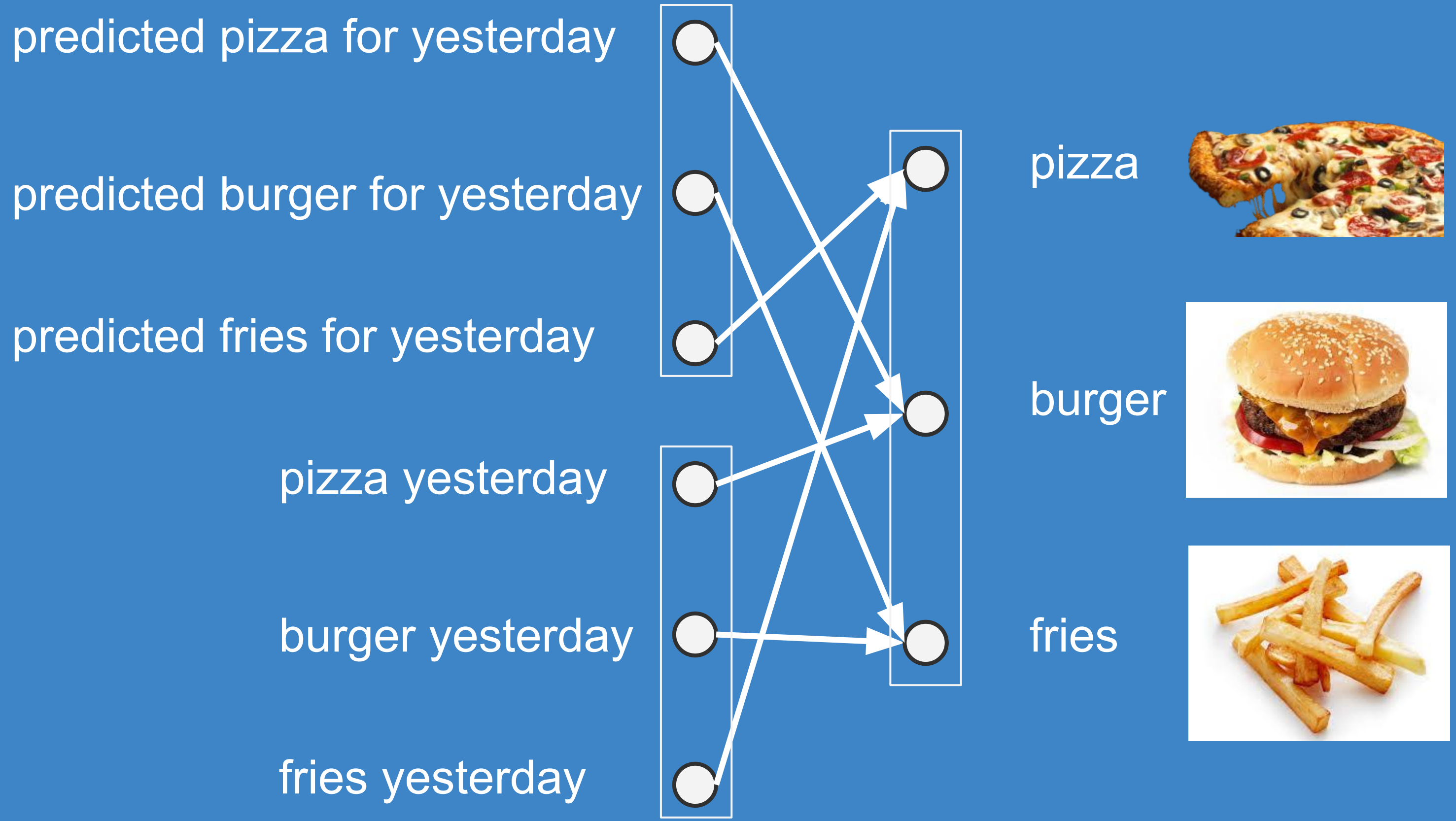
=

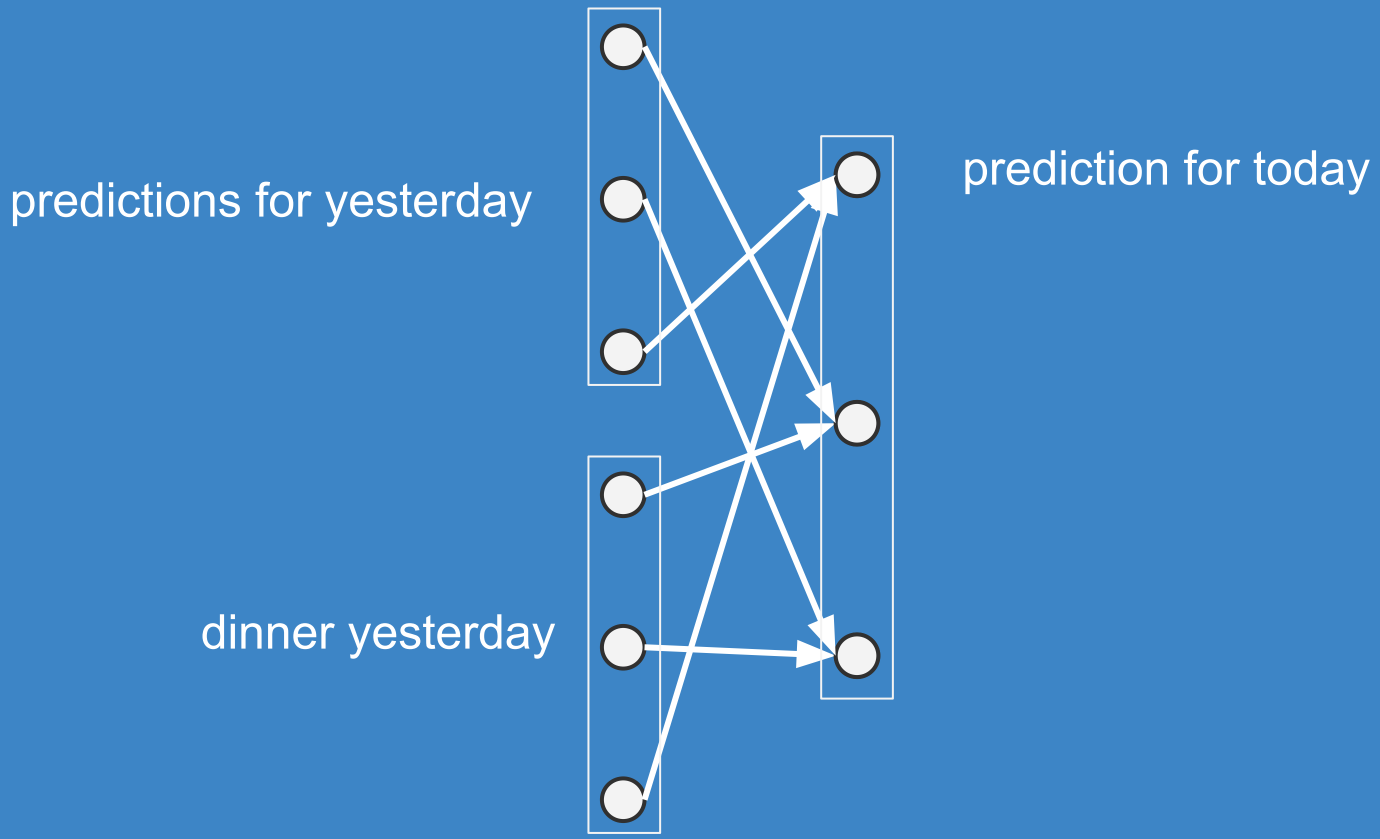
Day of week vector

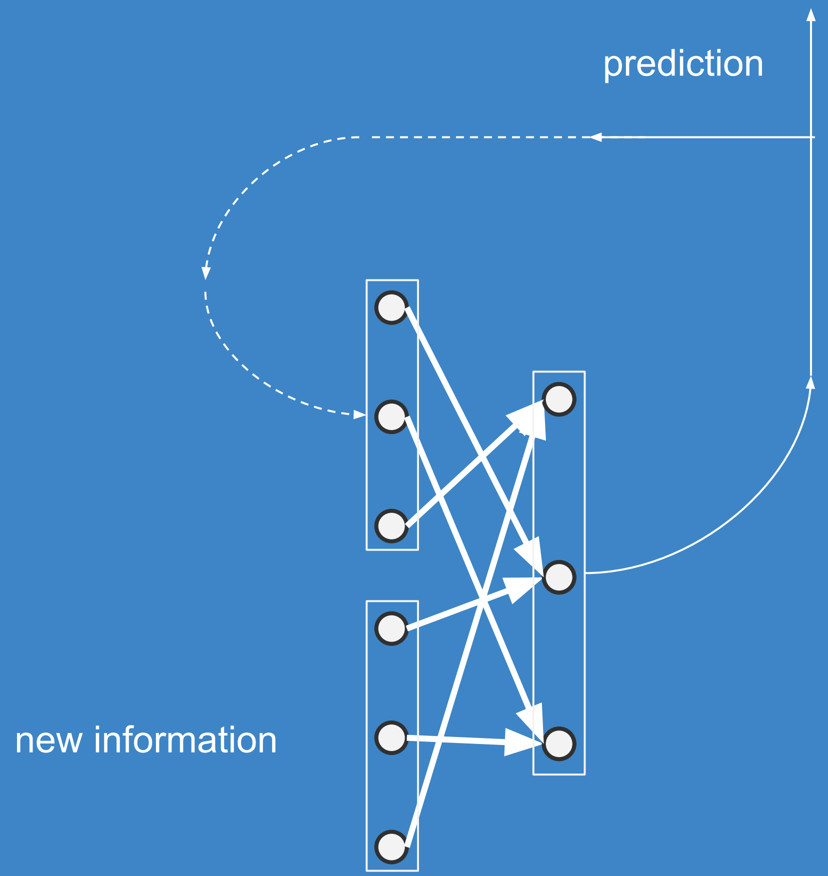
$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

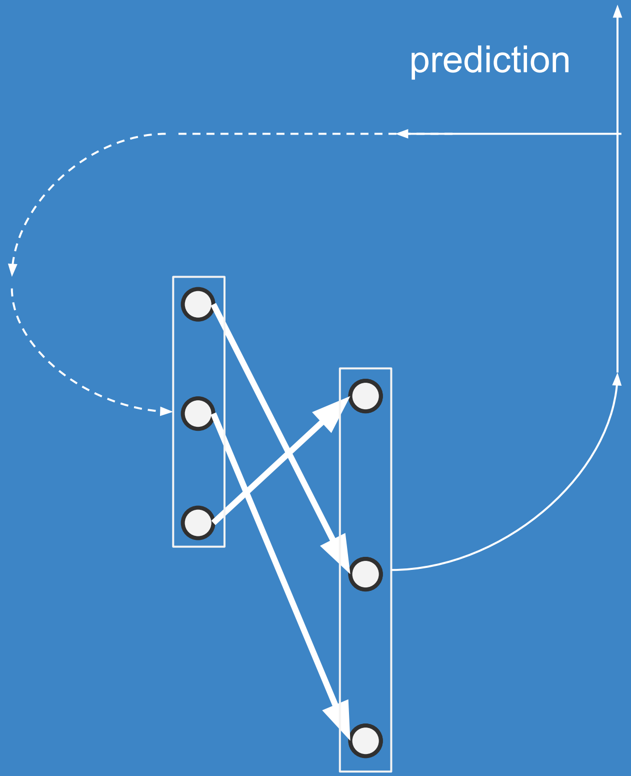
A vector is a list of values



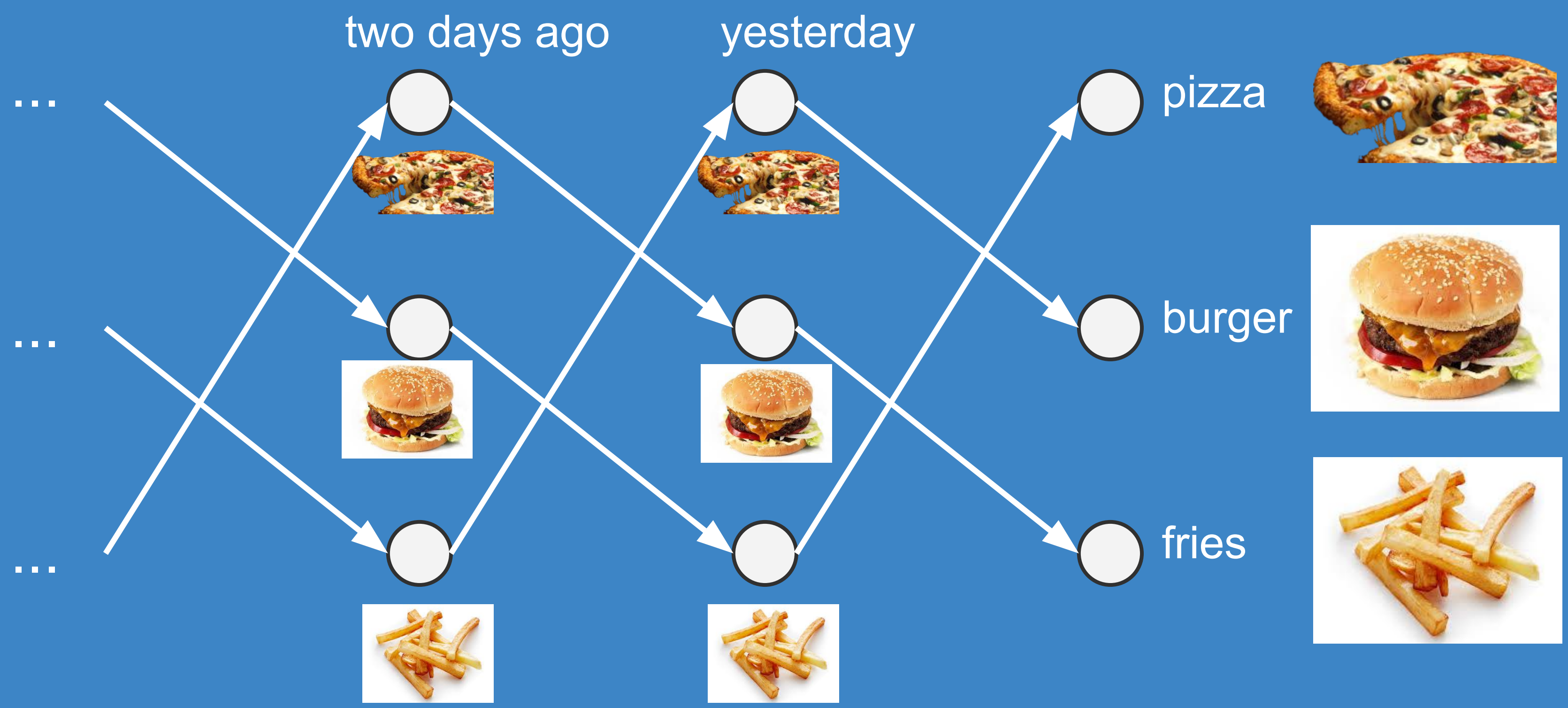

predicted pizza for yesterday
predicted burger for yesterday
predicted fries for yesterday
pizza yesterday
burger yesterday
fries yesterday
pizza
burger
fries


predictions for yesterday
dinner yesterday
prediction for today


prediction
new information


prediction

Unrolled predictions



Write a children's book

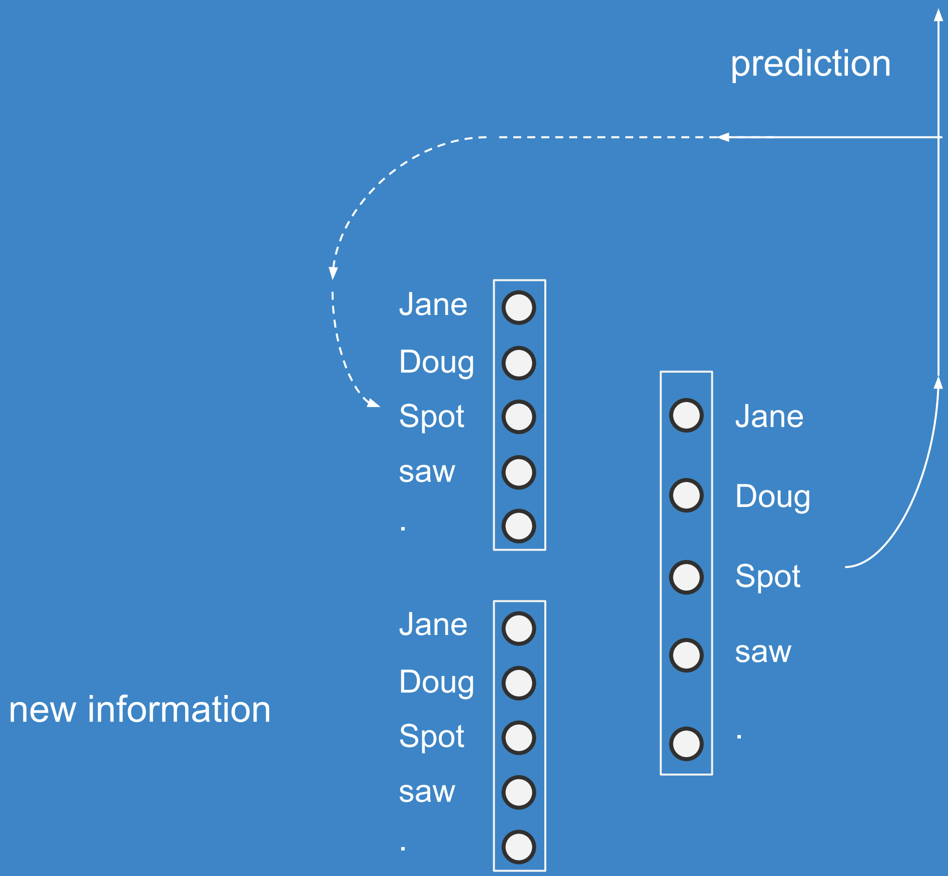
Doug saw Jane.

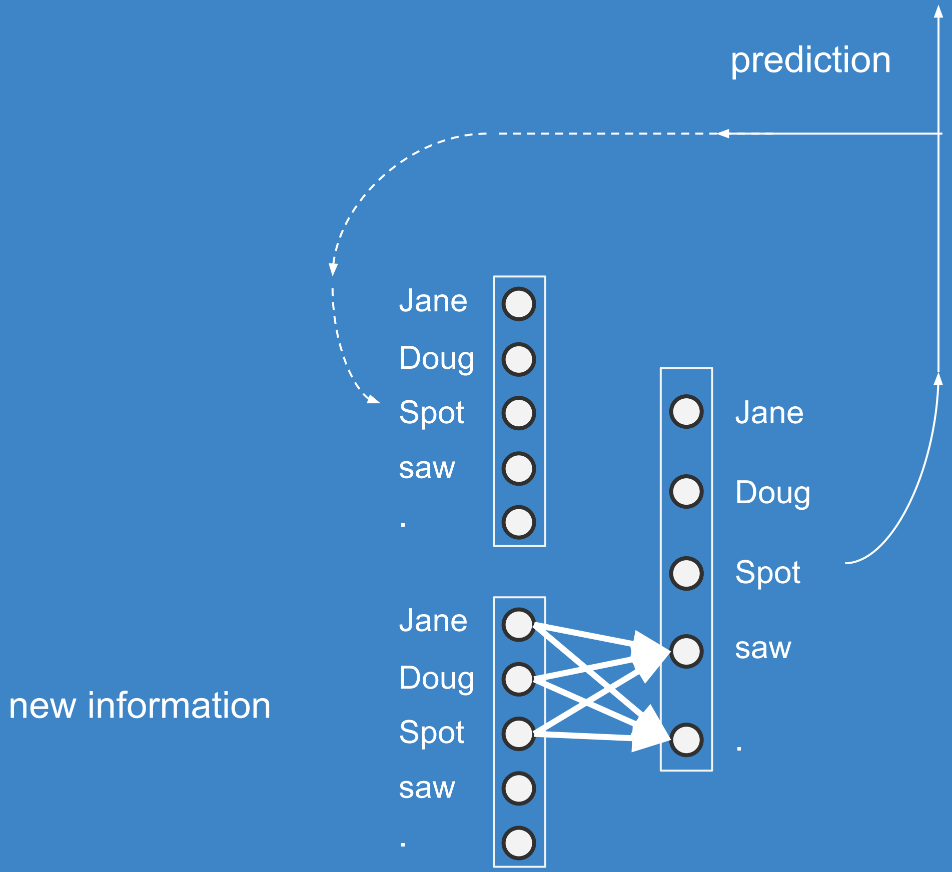
Jane saw Spot.

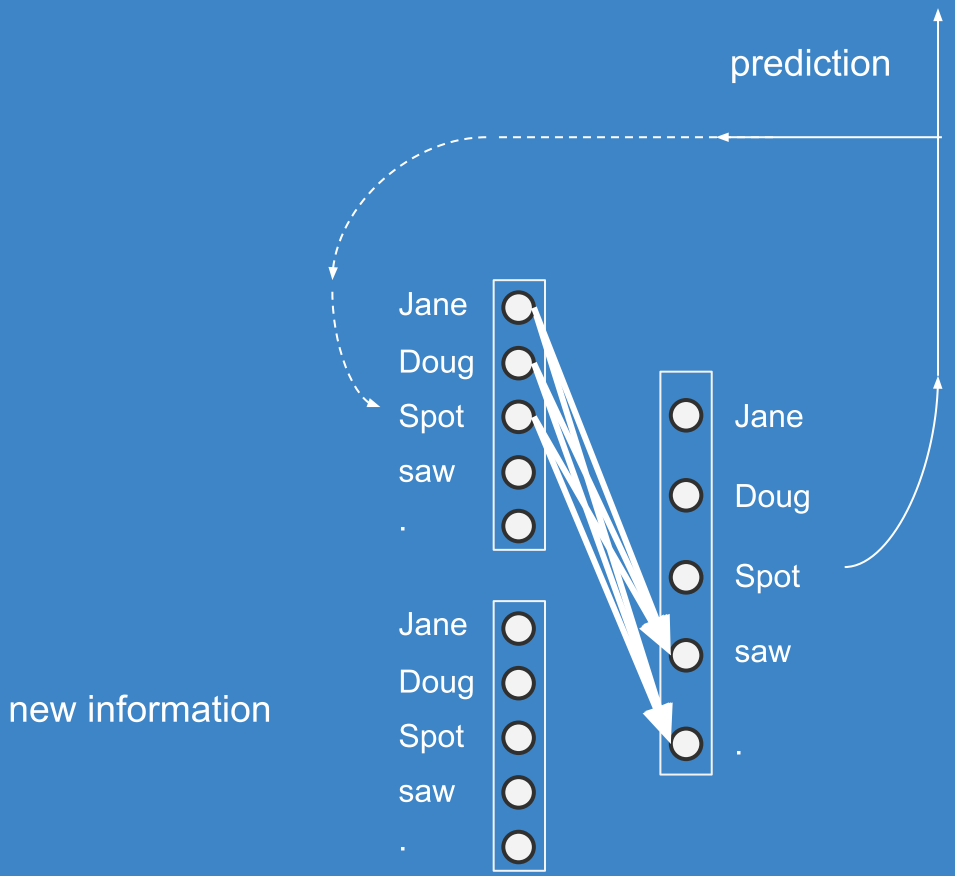
Spot saw Doug.

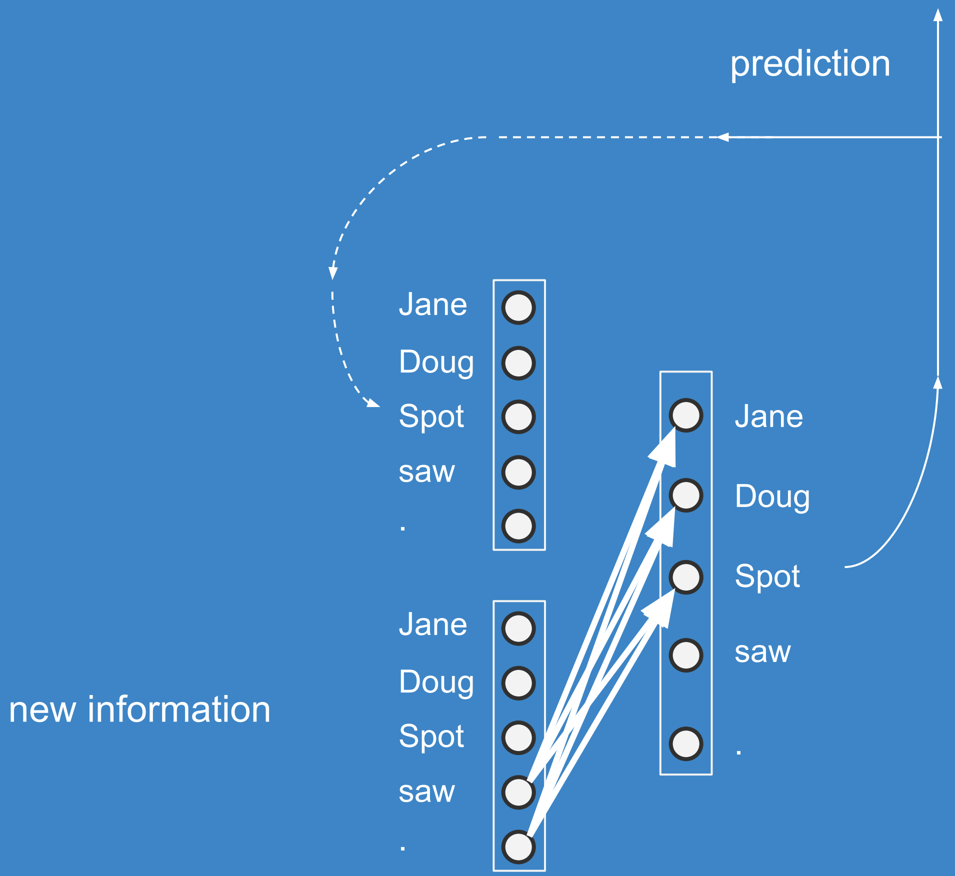
…

Your dictionary is small: {Doug, Jane, Spot, saw, .}


prediction
Jane
Doug
Spot
saw
.
Jane
Doug
Spot
saw
.
Jane
Doug
Spot
saw
.
new information

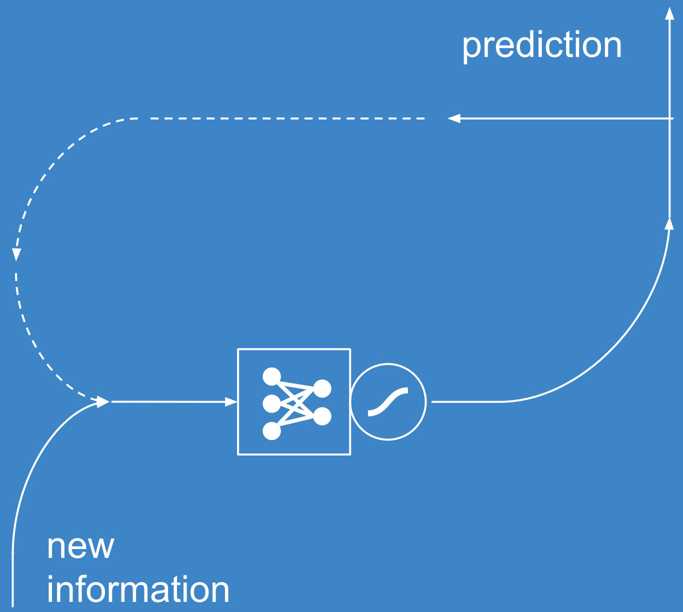

prediction
Jane
Doug
Spot
saw
.
Jane
Doug
Spot
saw
.
new information
Jane
Doug
Spot
saw
.


prediction
Jane
Doug
Spot
saw
.
Jane
Doug
Spot
saw
.
Jane
Doug
Spot
saw
.
new information

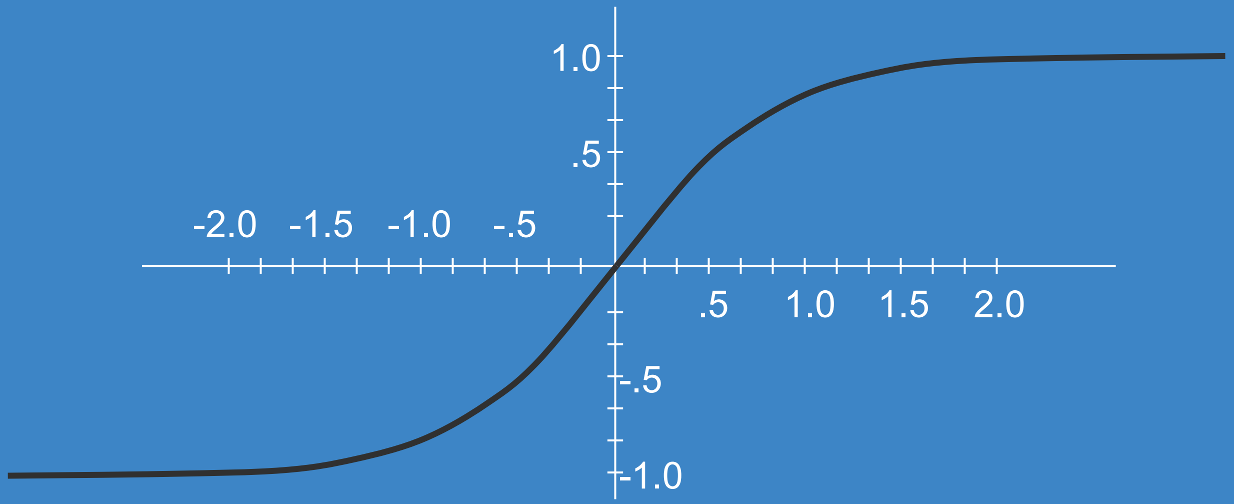


prediction
Jane
Doug
Spot
saw
.
Jane
Doug
Spot
saw
.
new information
Jane
Doug
Spot
saw
.

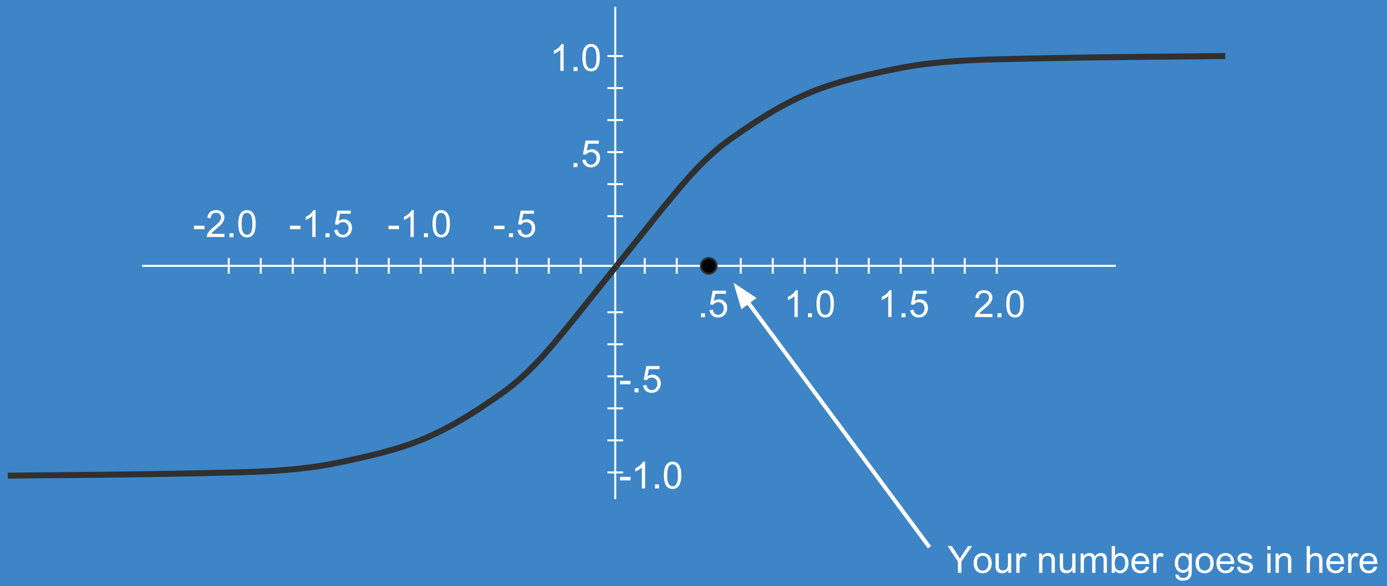
recurrent neural network



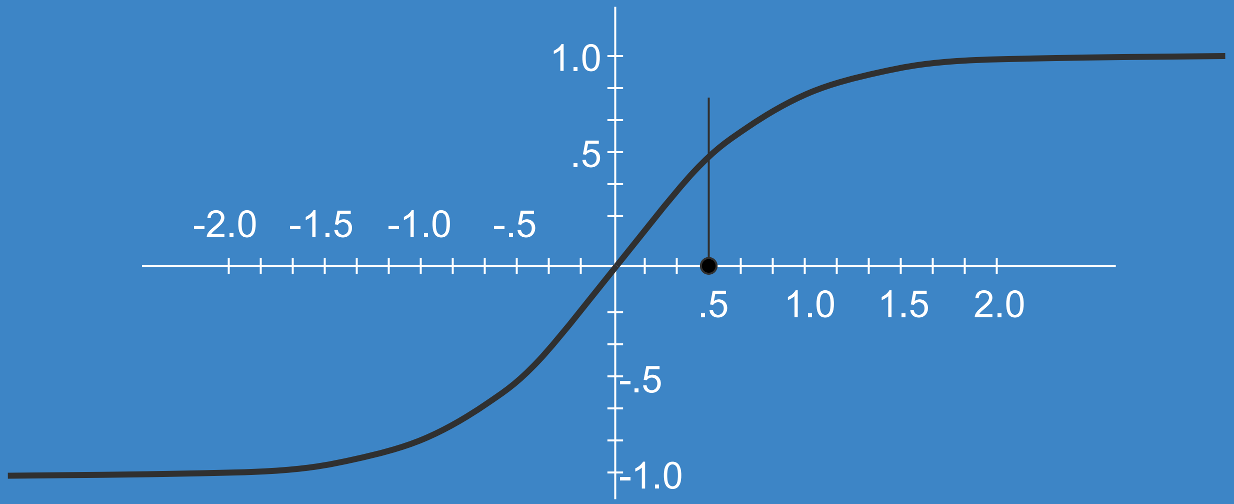
Hyperbolic tangent (tanh) squashing function



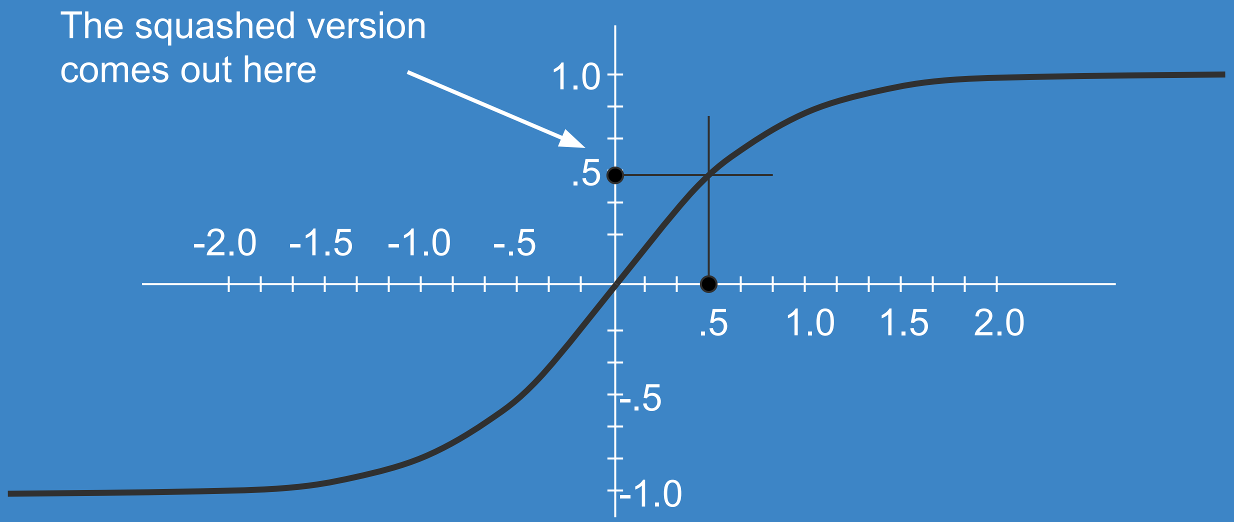
tanh squashing function



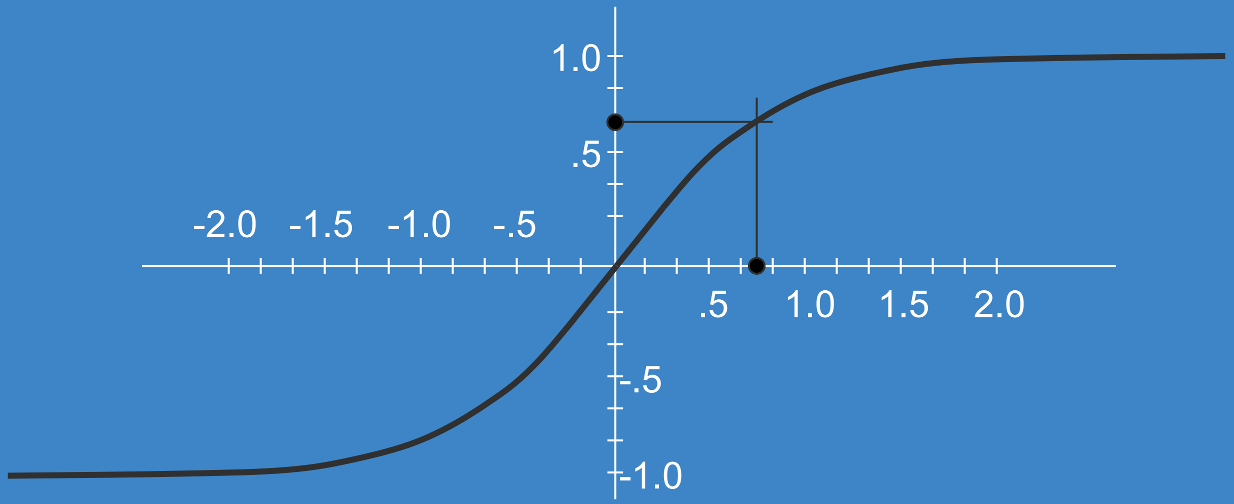
tanh squashing function



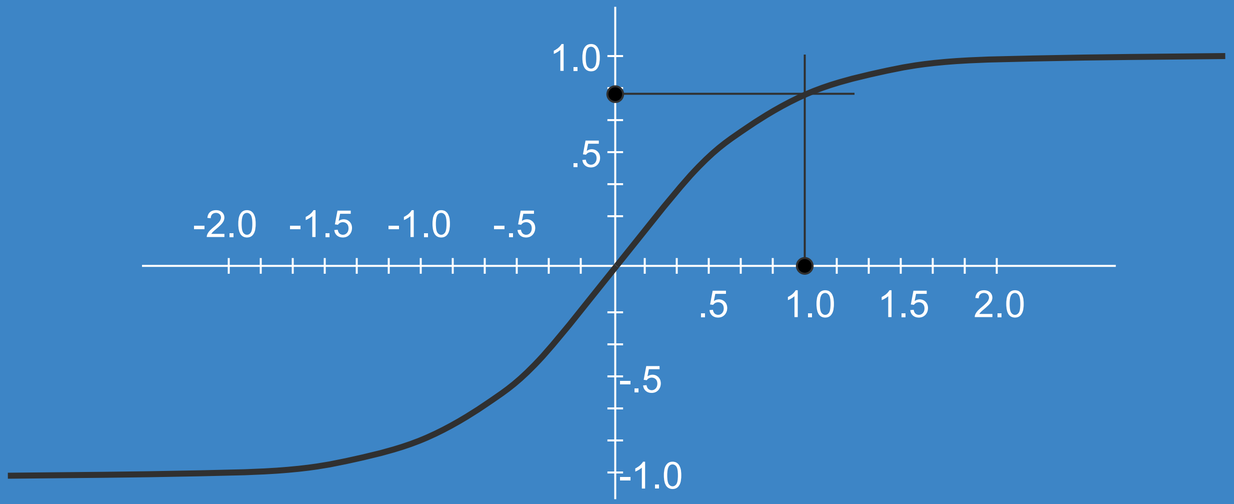
tanh squashing function



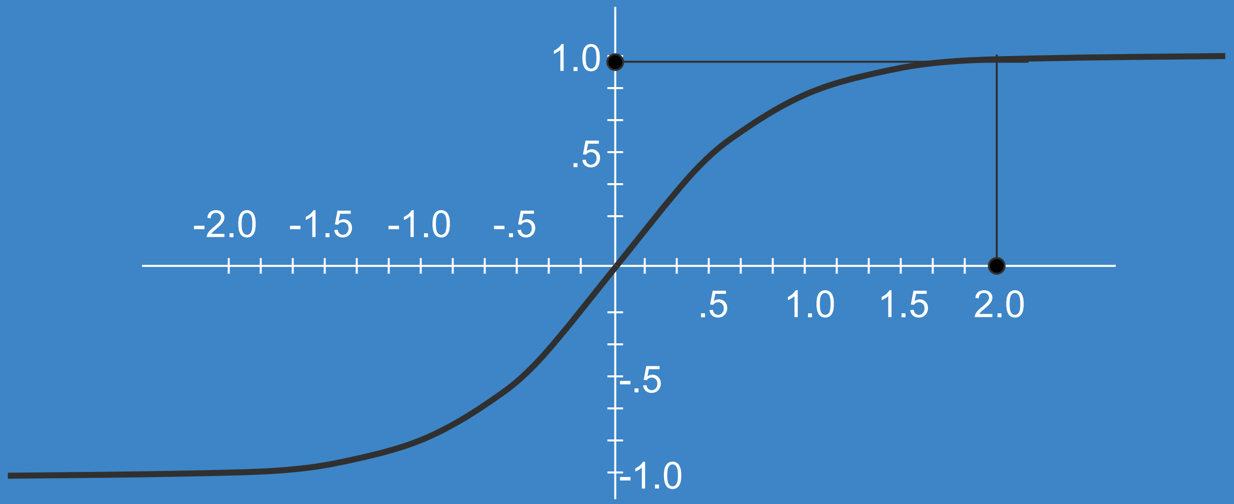
tanh squashing function



tanh squashing function

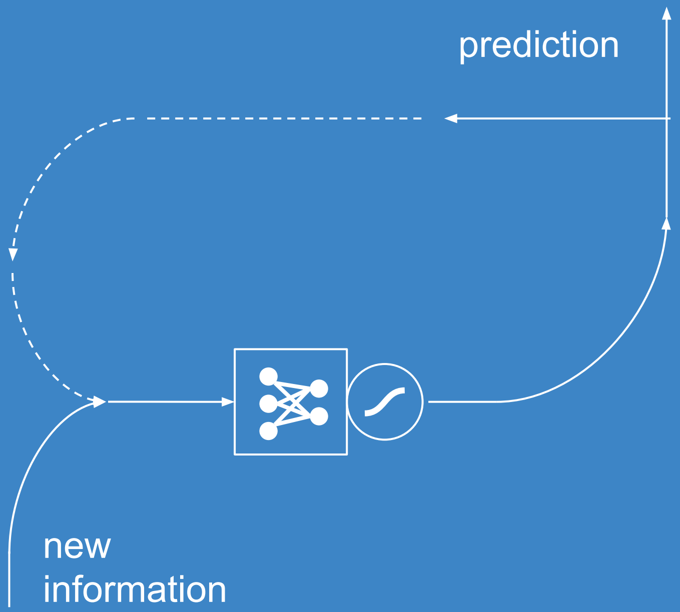


tanh squashing function



No matter what you start with, the answer stays between -1 and 1.

recurrent neural network



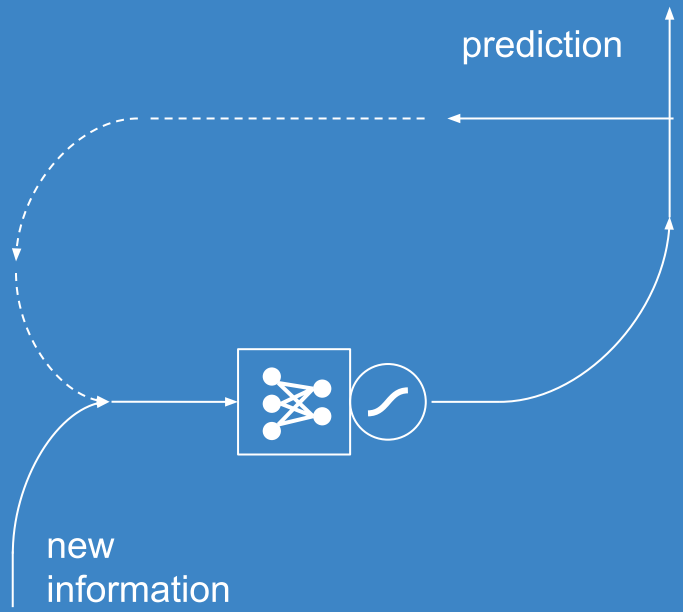
Mistakes an RNN can make

Doug saw Doug.

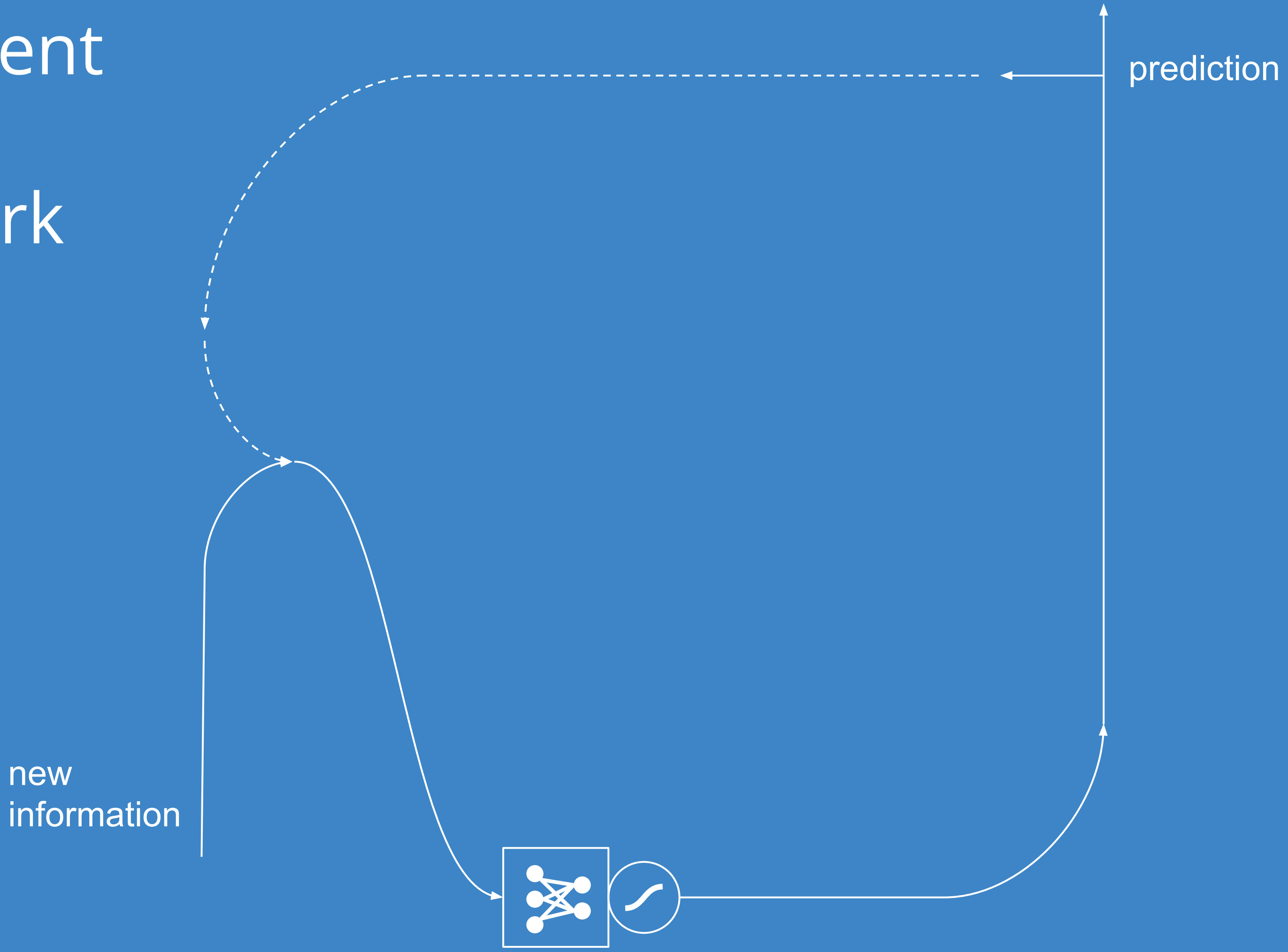
Jane saw Spot saw Doug saw ...

Spot. Doug. Jane.

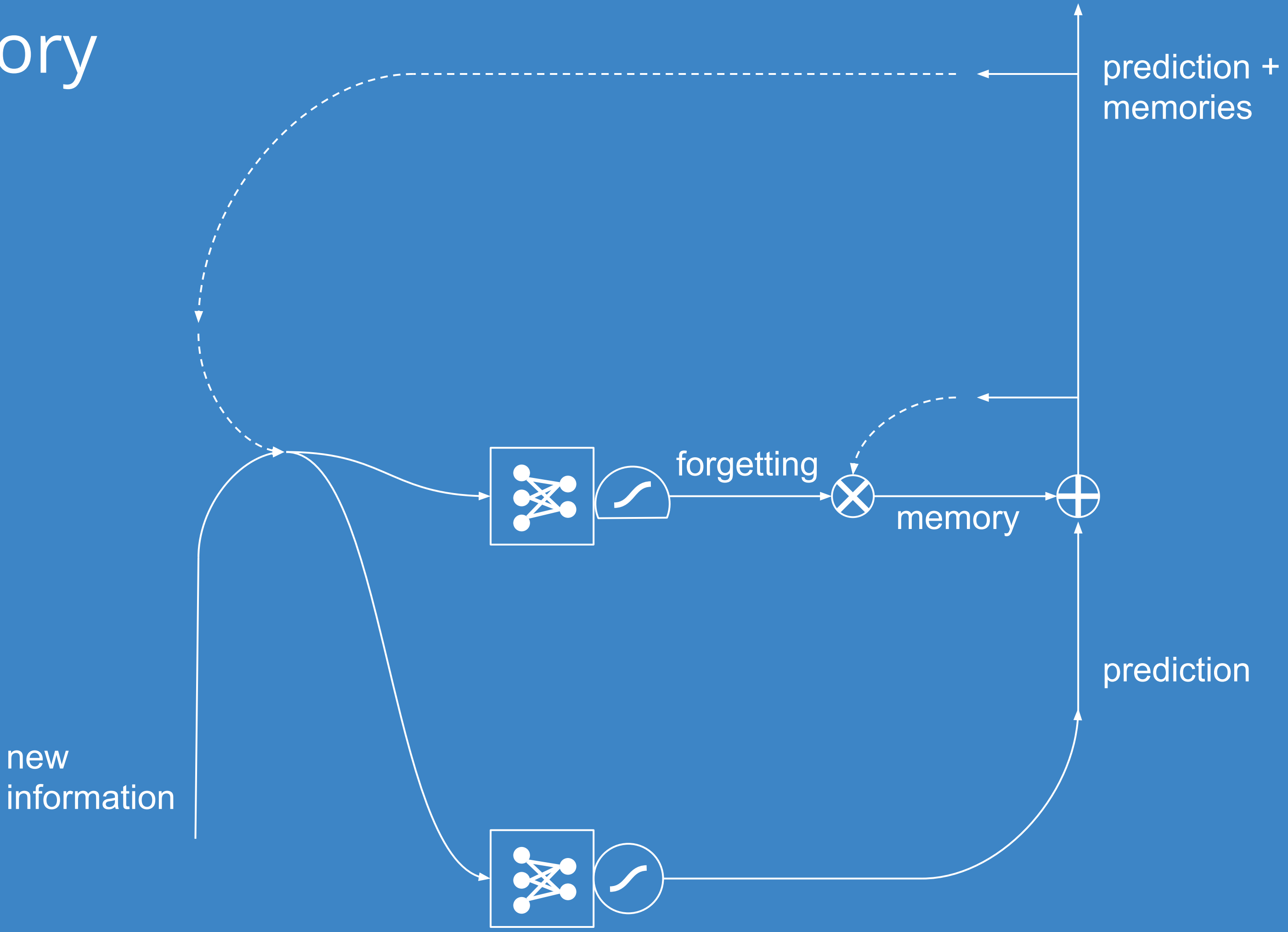
recurrent neural network



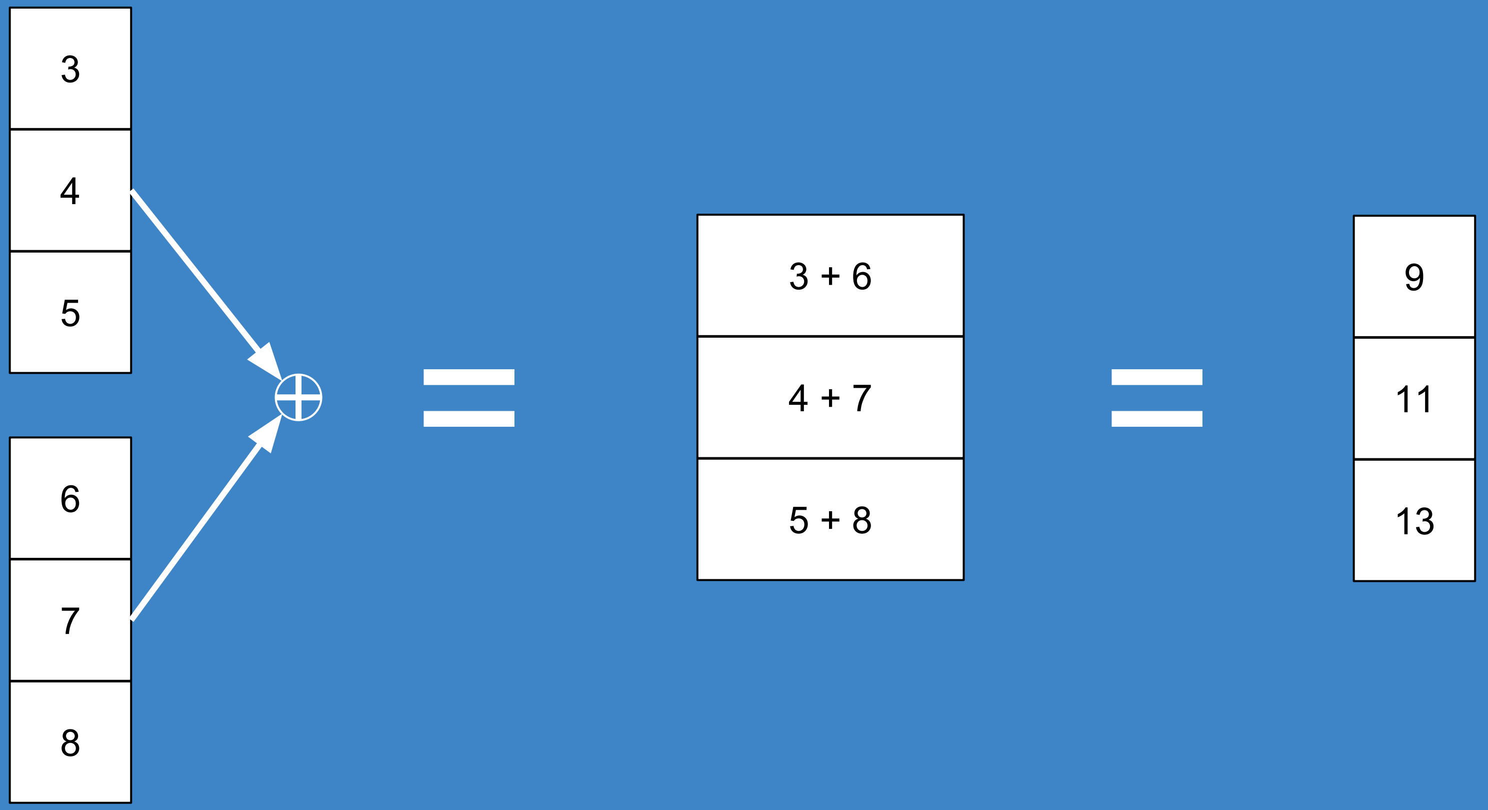
recurrent neural network



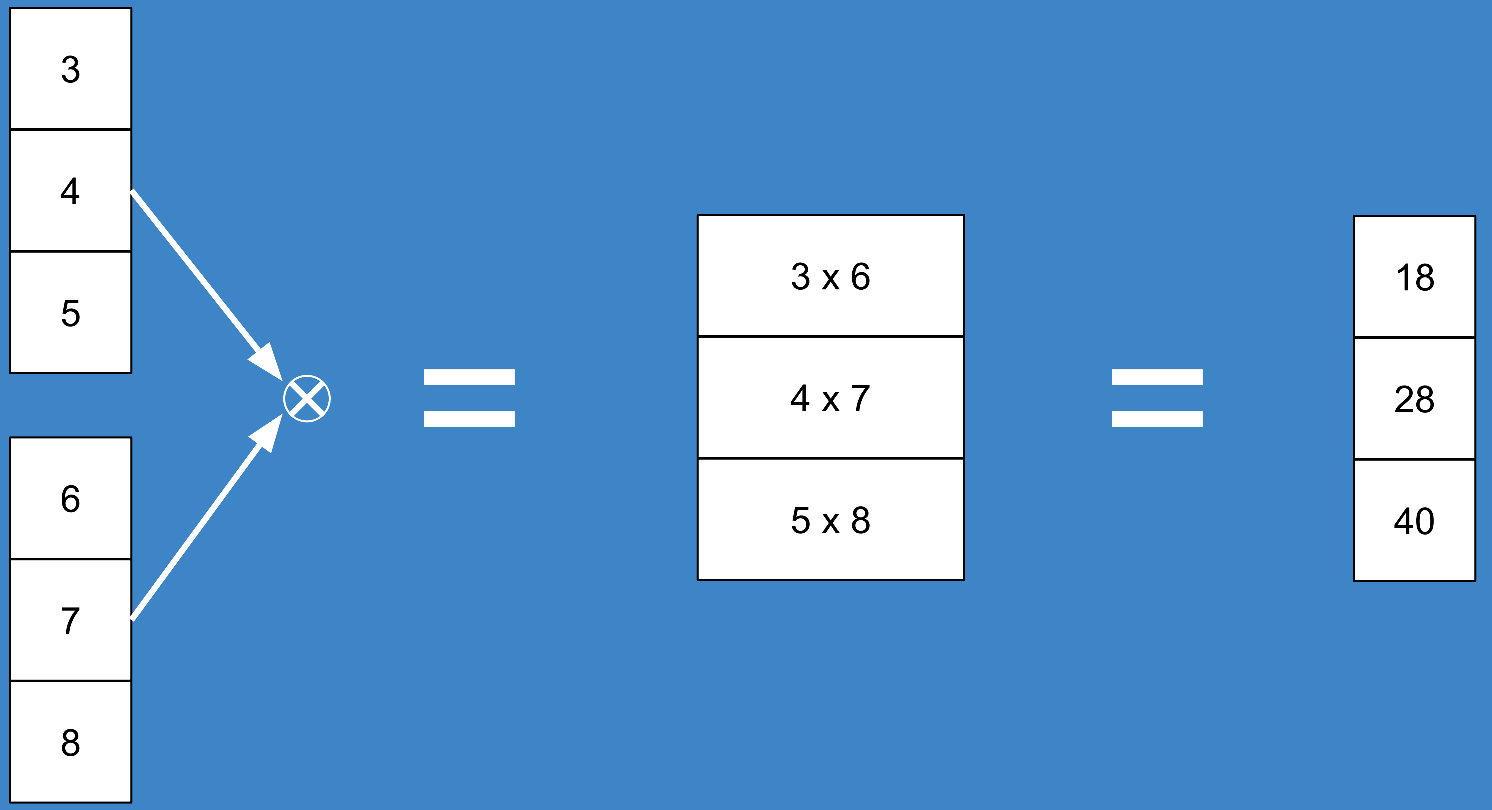
memory


prediction +
memories
forgetting
memory
prediction
new
information

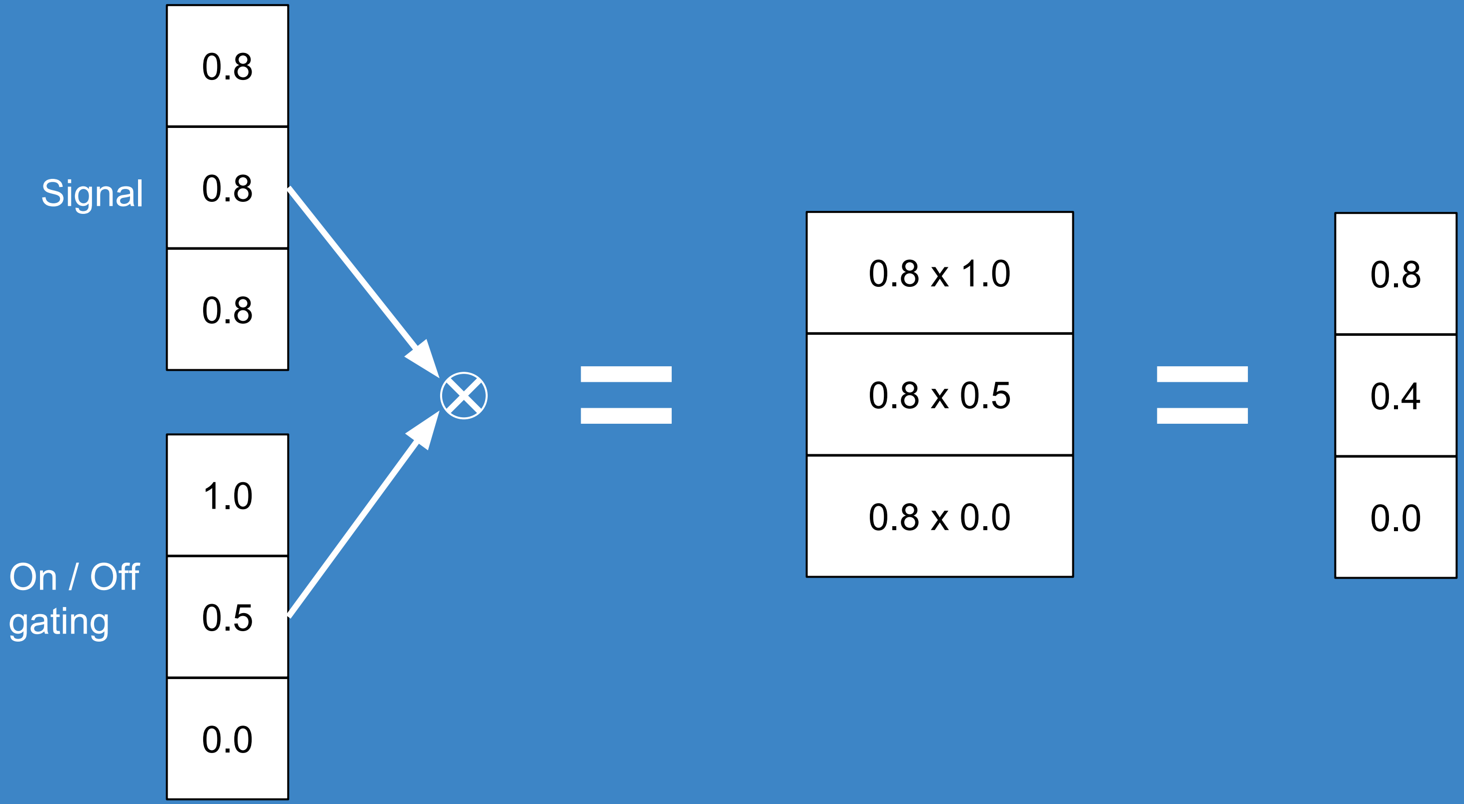
Plus junction: element-by-element addition $\oplus$



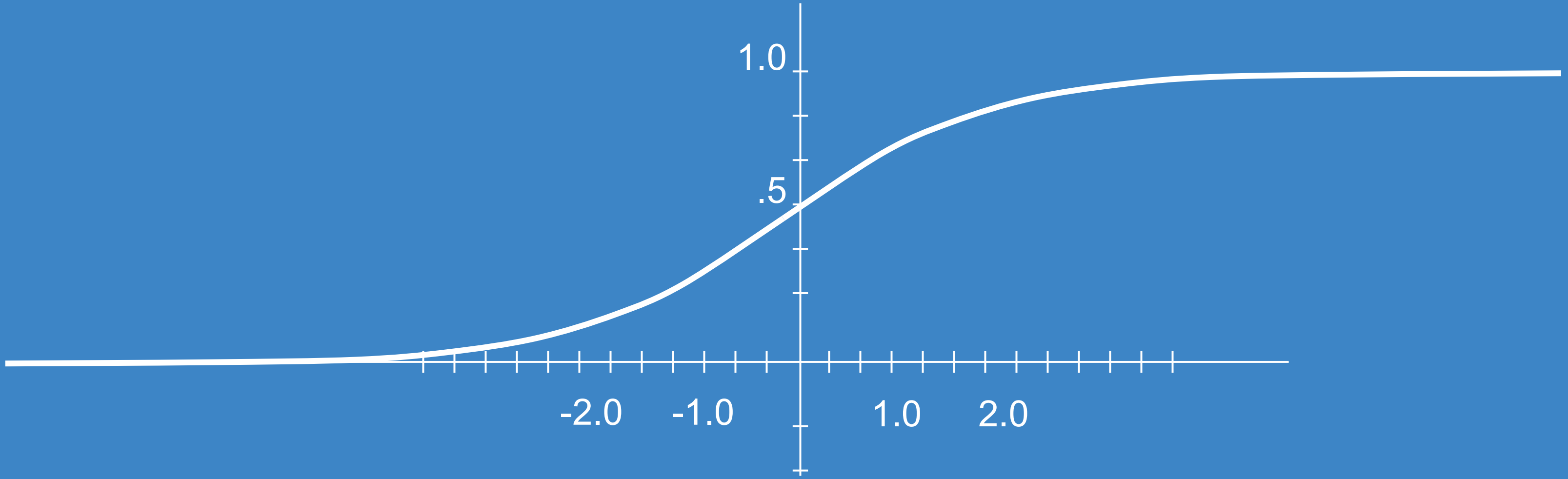
Times junction: element-by-element multiplication



Gating

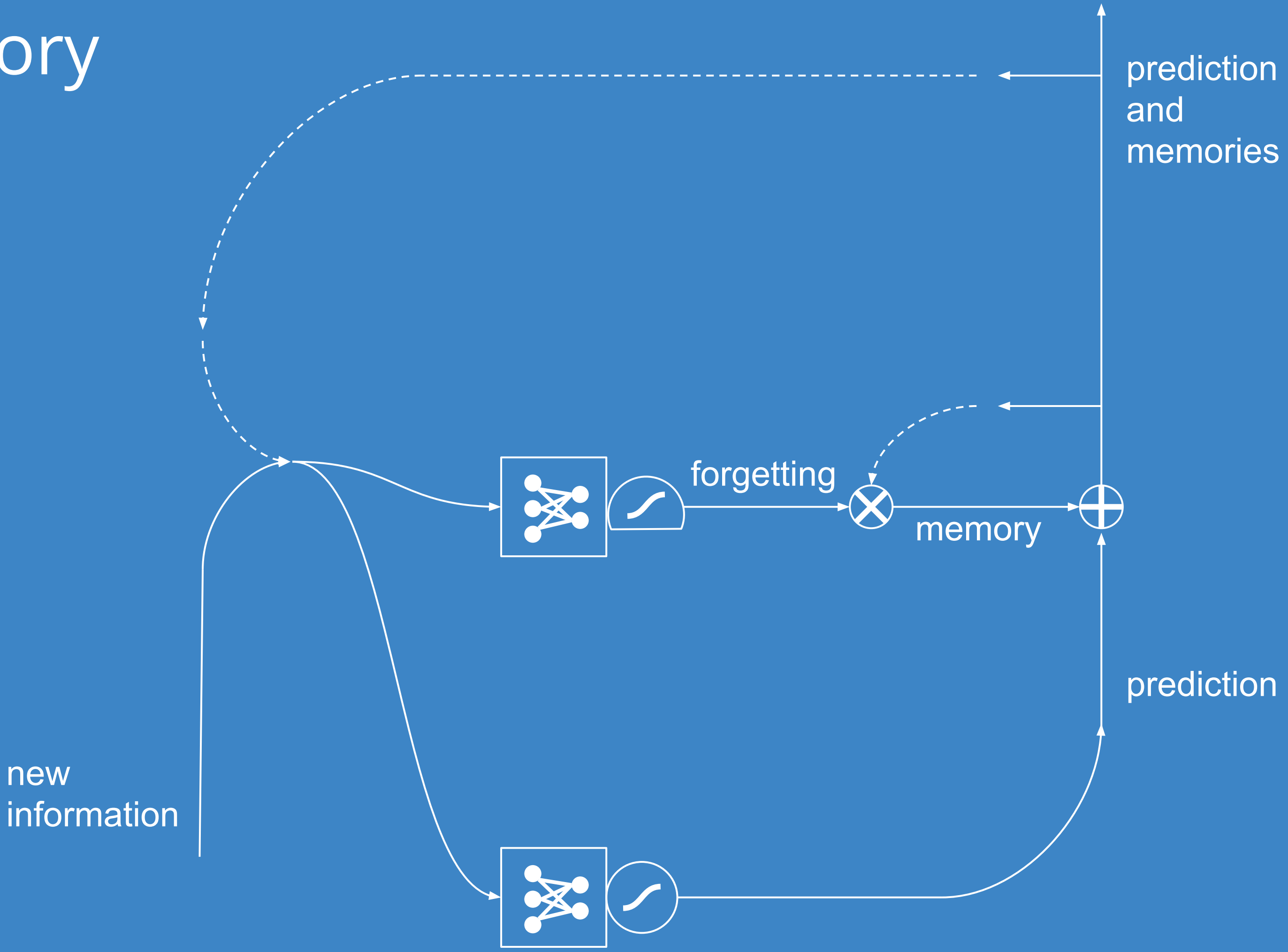

Signal
0.8
0.8
0.8
On / Off
gating
1.0
0.5
0.0
=
0.8 x 1.0
0.8 x 0.5
0.8 x 0.0
=
0.8
0.4
0.0

Logistic (sigmoid) squashing function

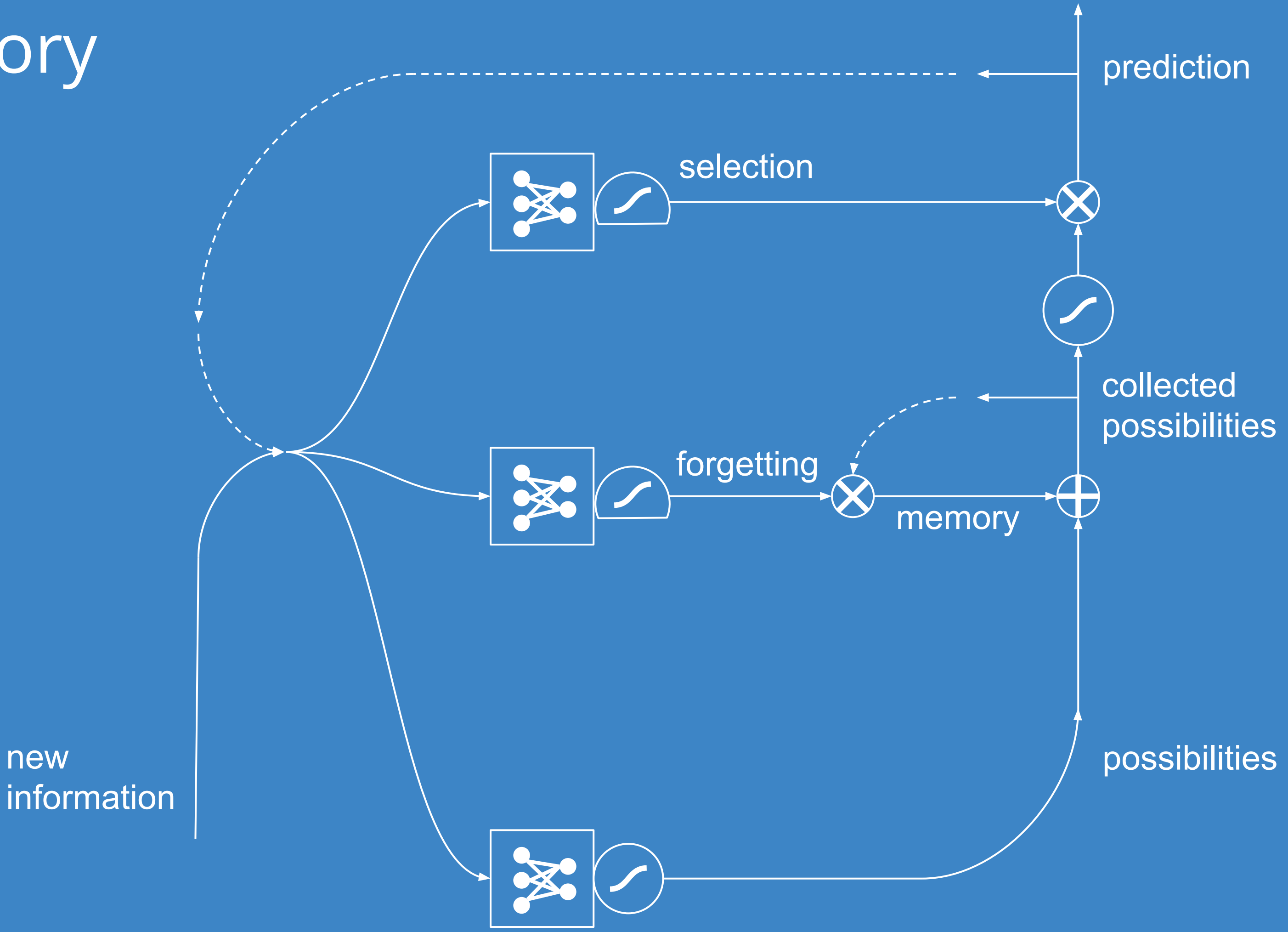


No matter what you start with, the answer stays between 0 and 1.

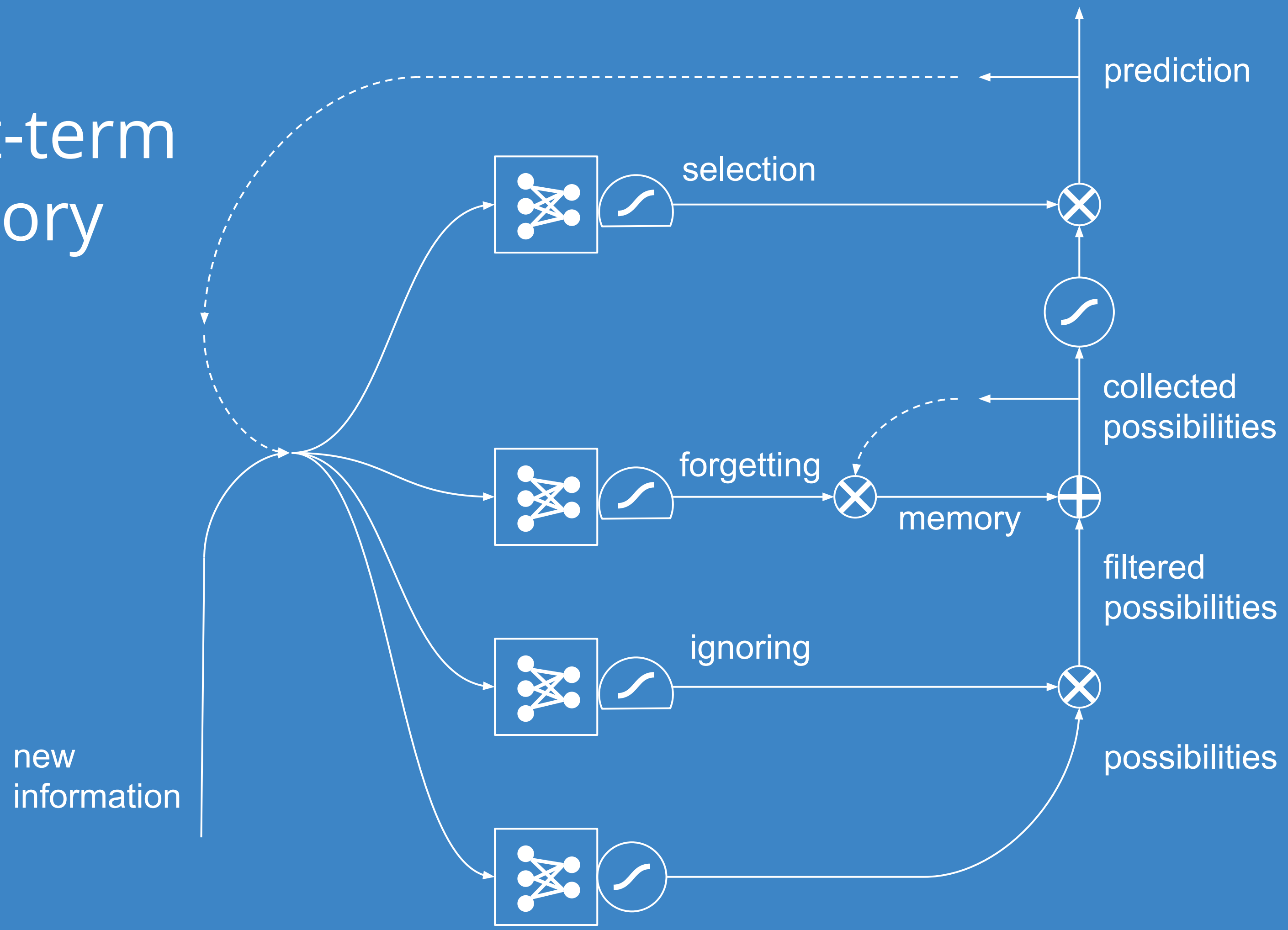
memory


prediction
and
memories
forgetting
memory
prediction
new
information

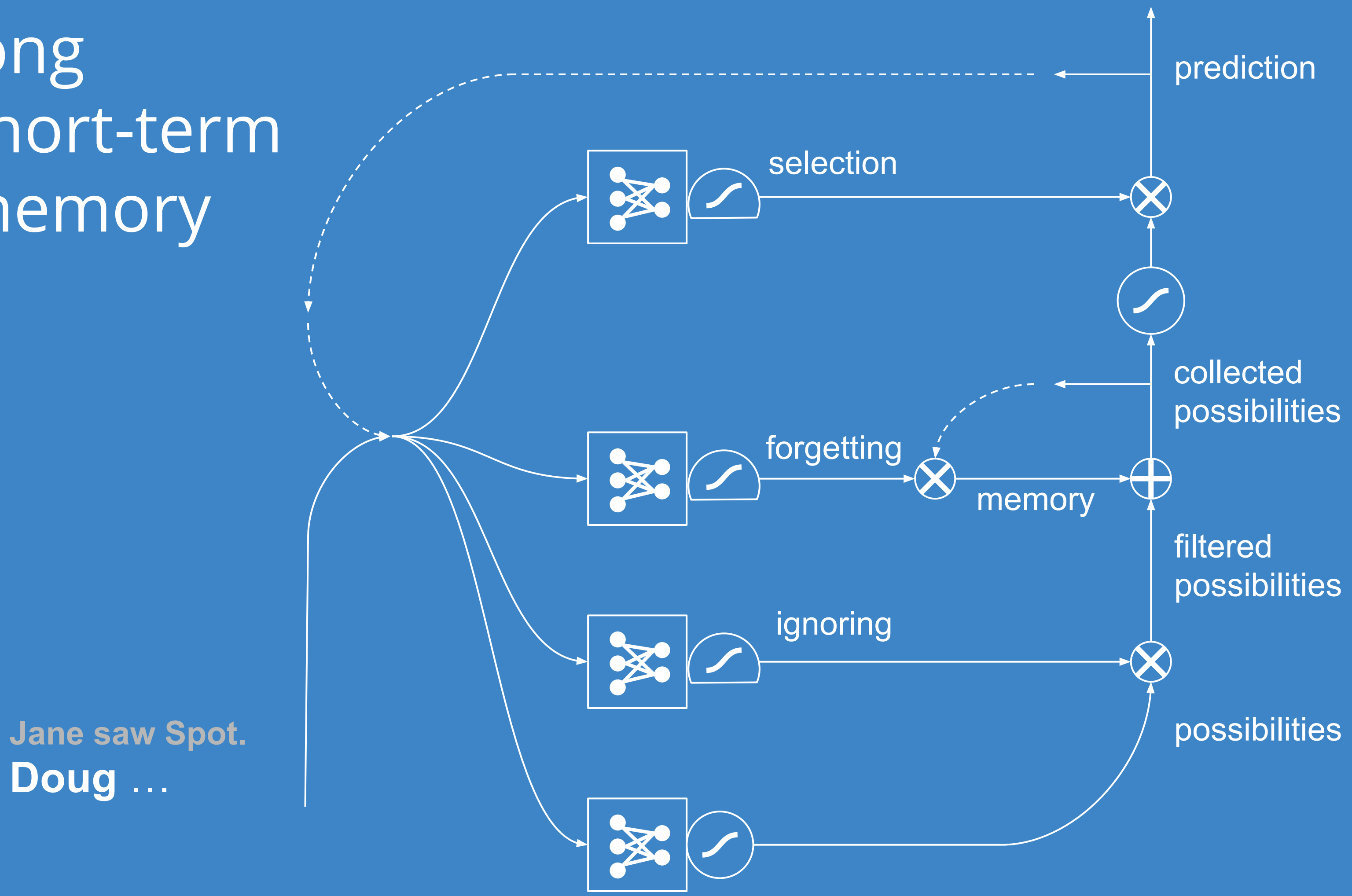
memory


prediction
selection
collected
possibilities
forgetting
memory
new
information
possibilities

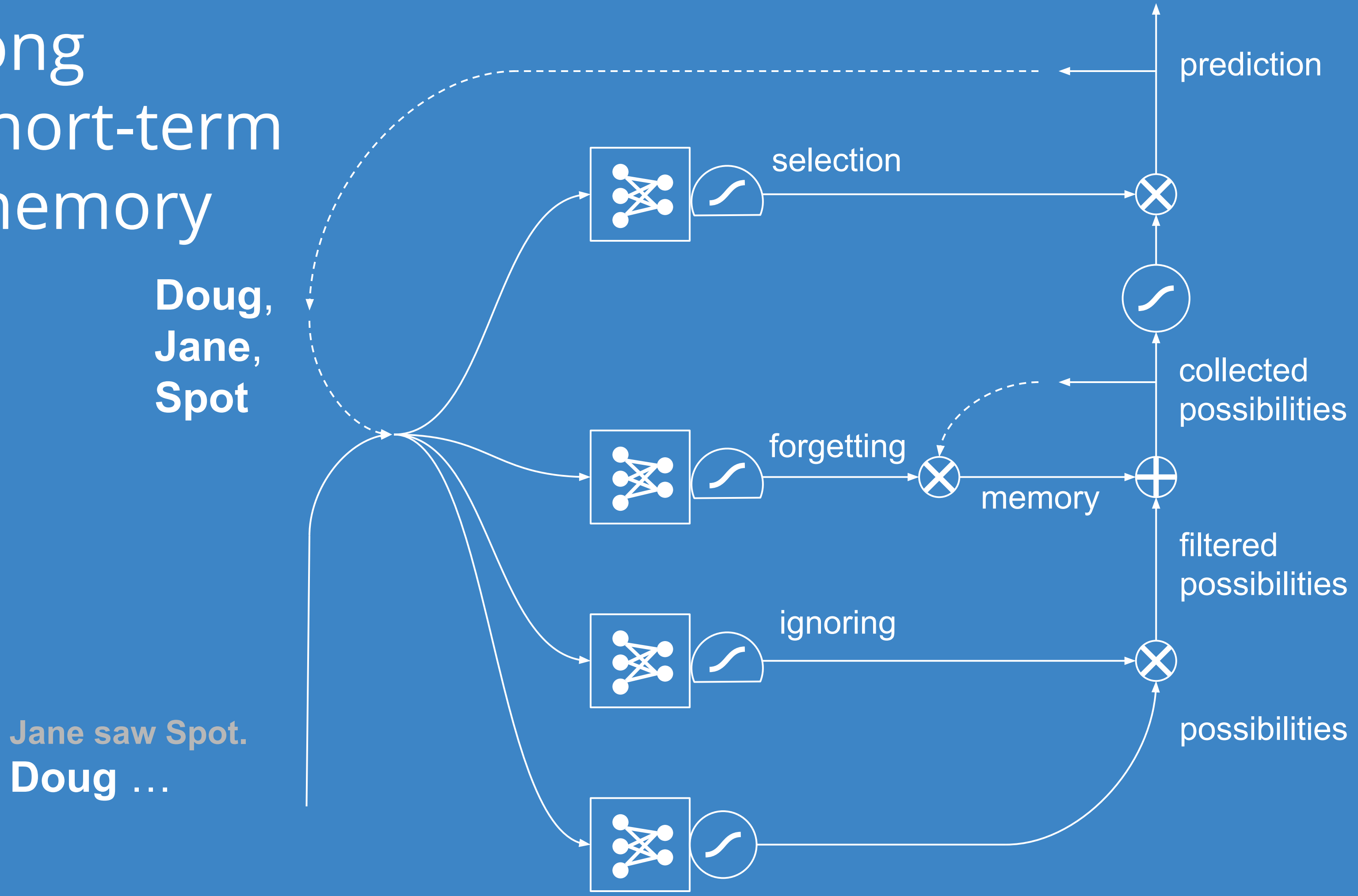
long short-term memory



long short-term memory

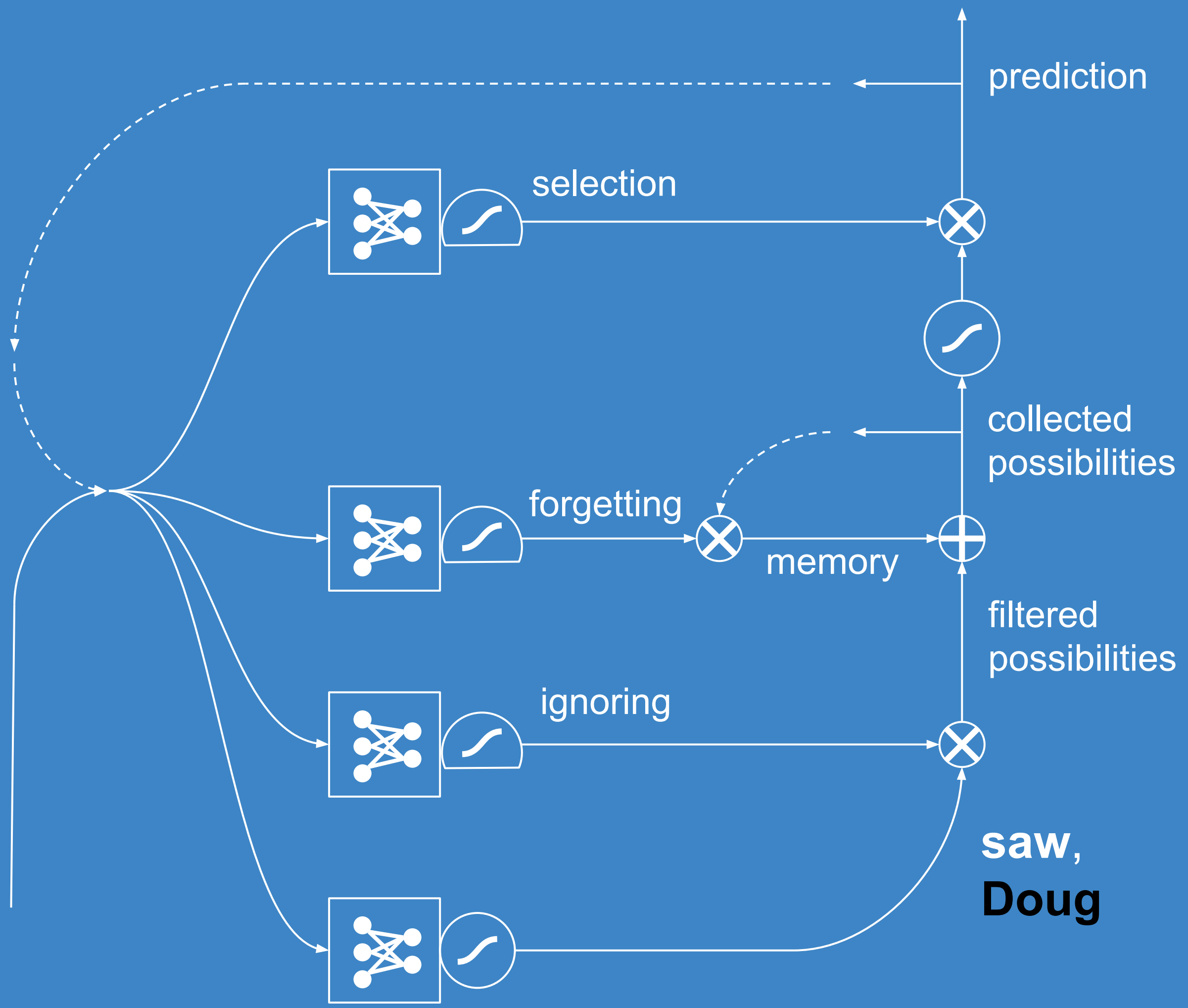


long short-term memory



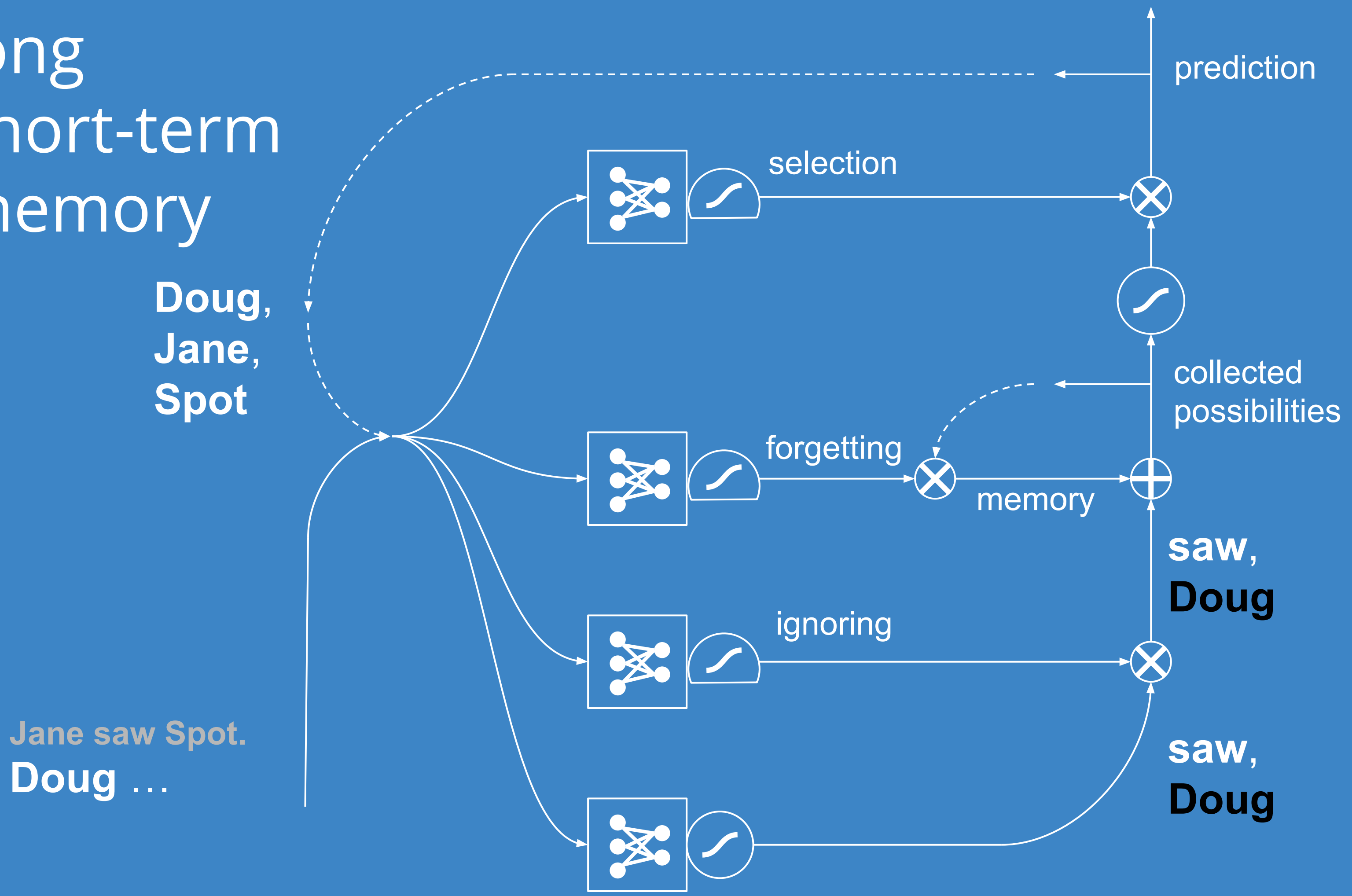
long short-term memory

**Doug,
Jane,
Spot**

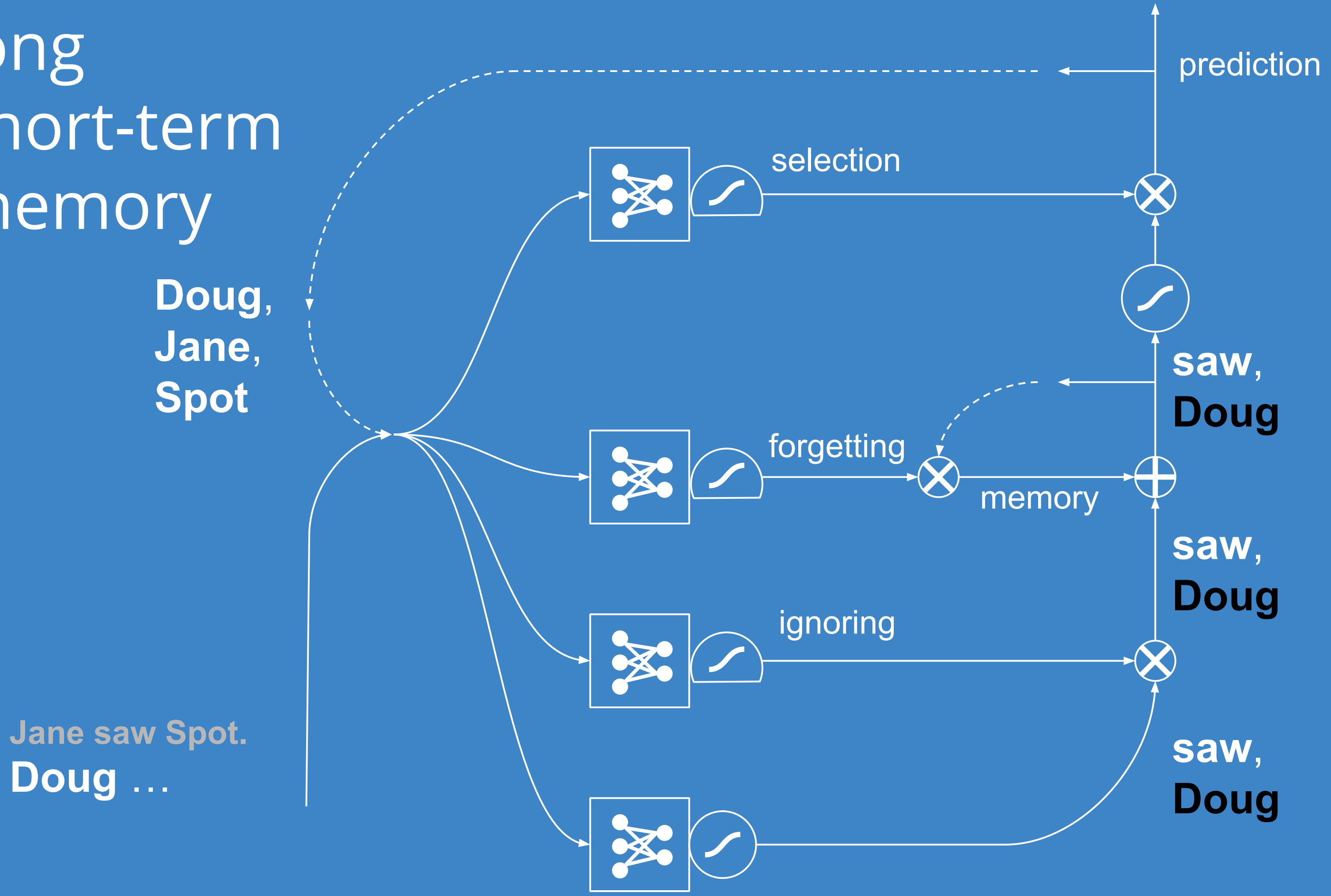


Jane saw Spot.
Doug …

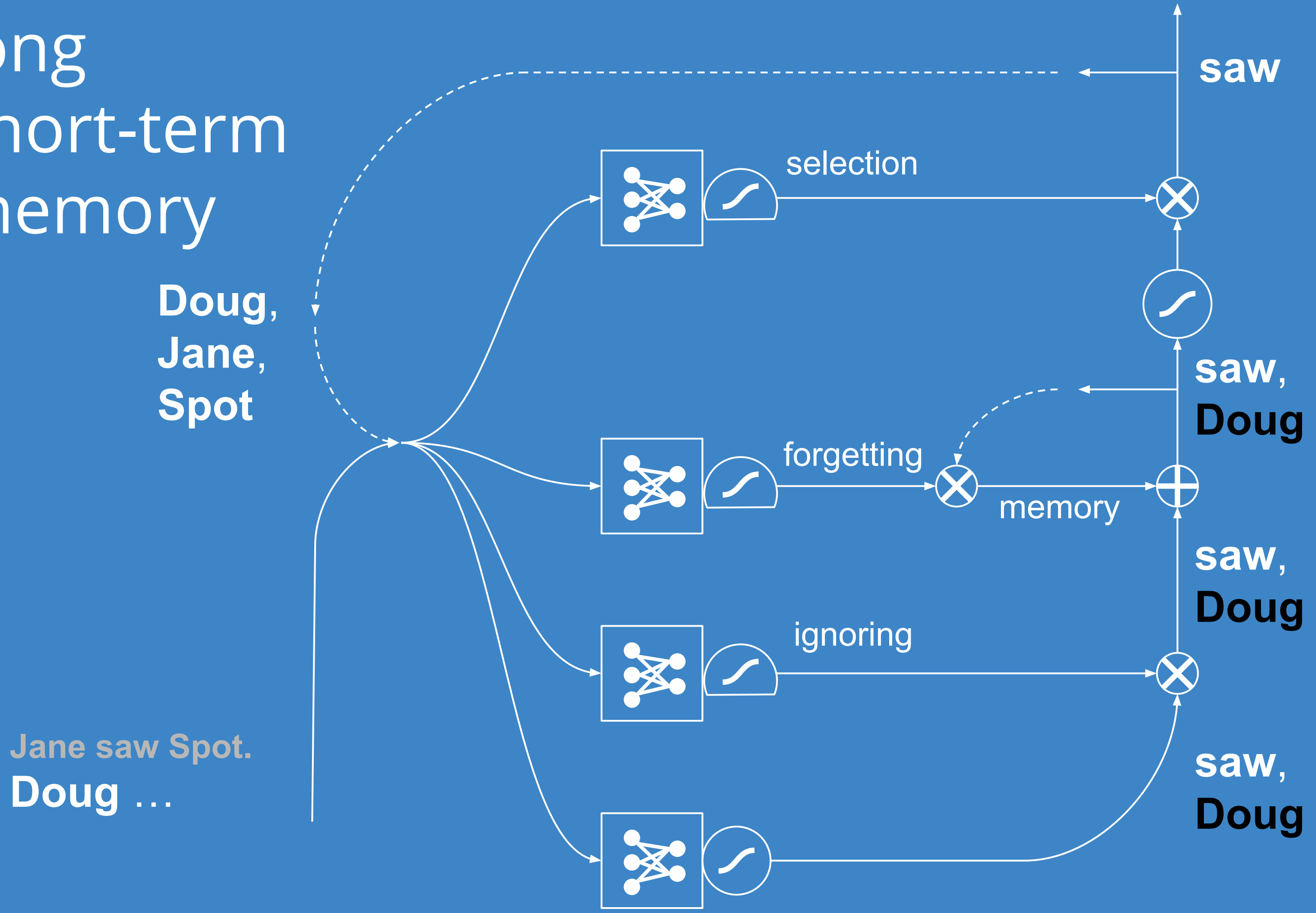
long short-term memory



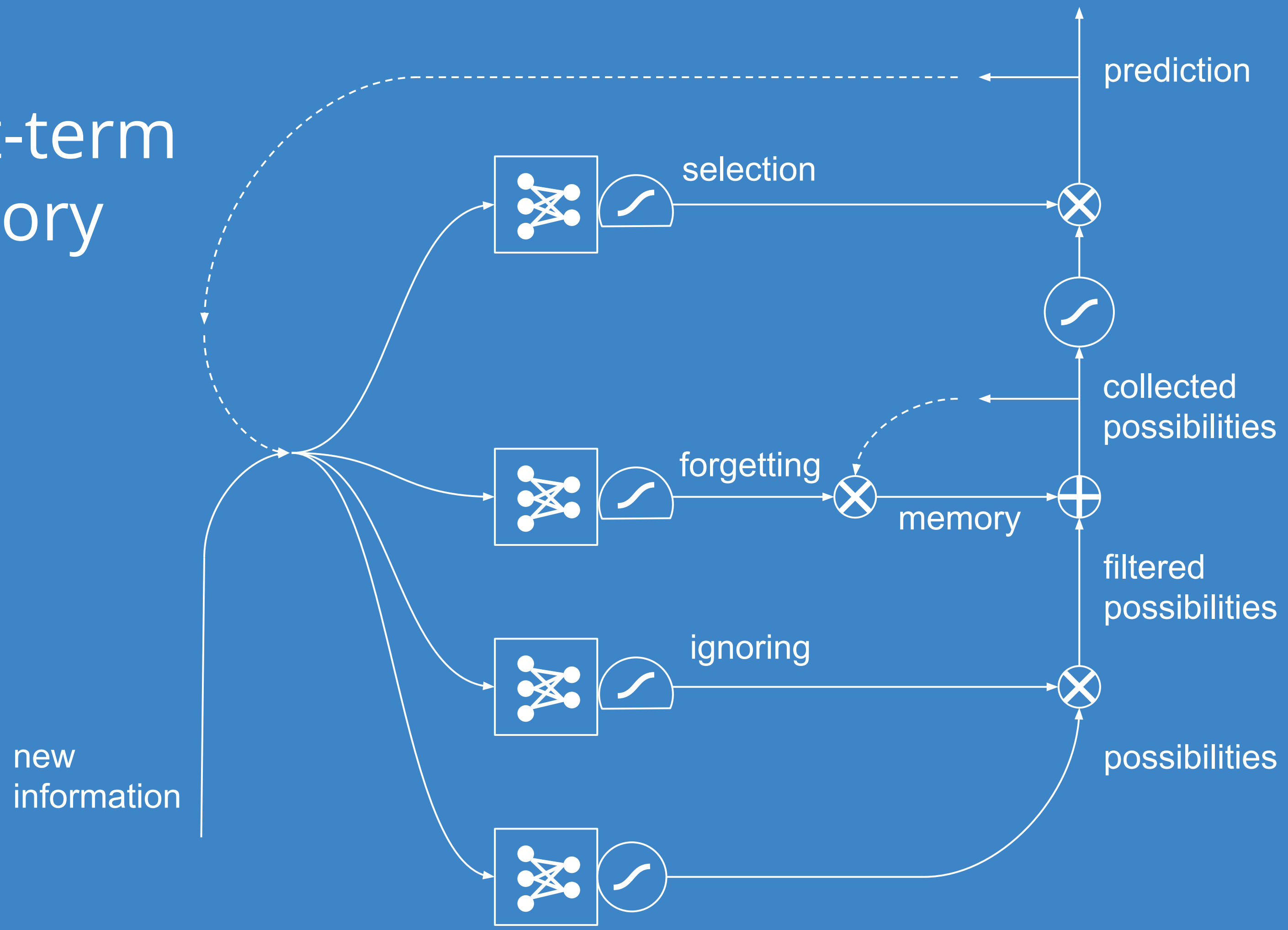
long short-term memory



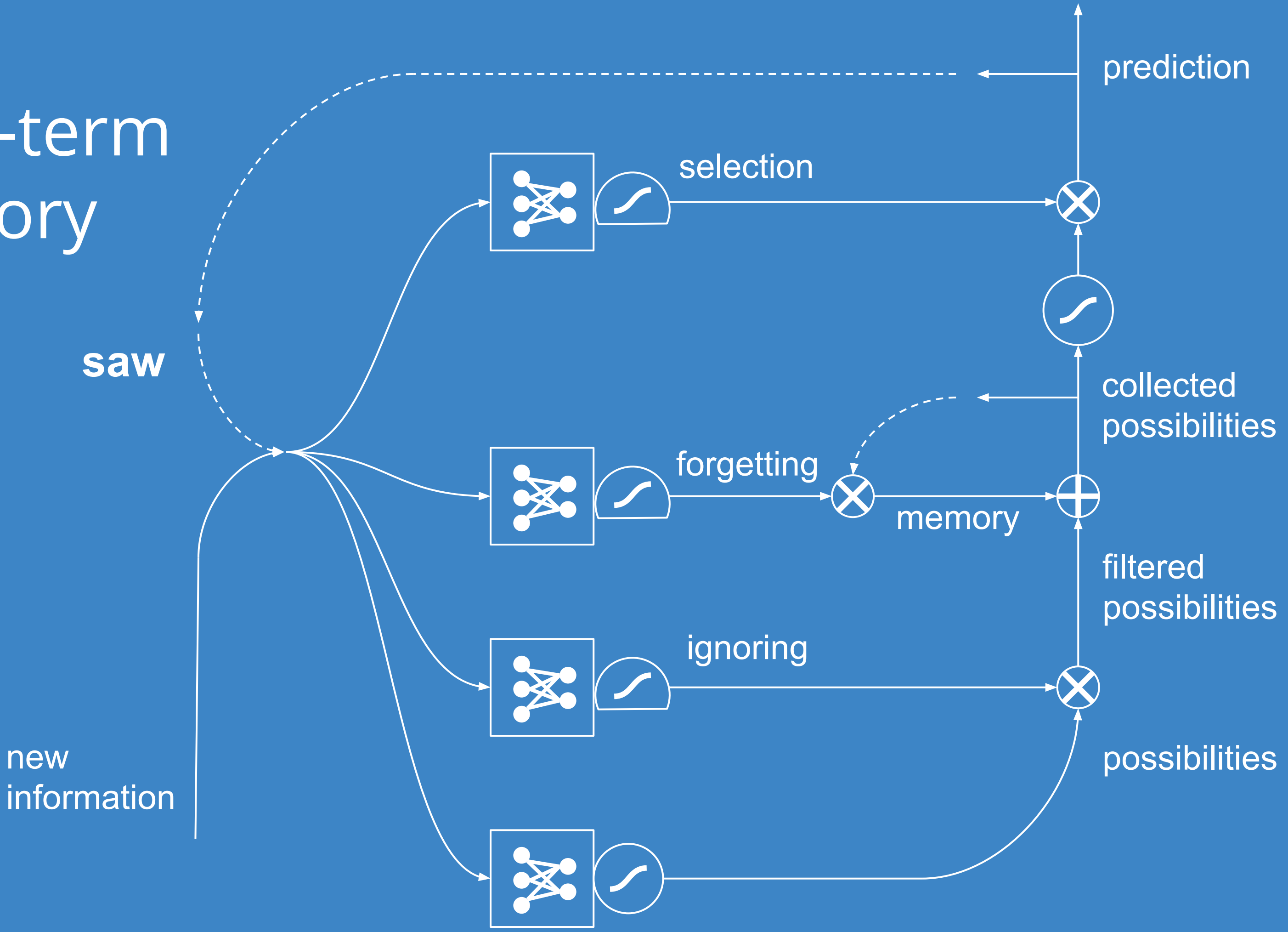
long short-term memory



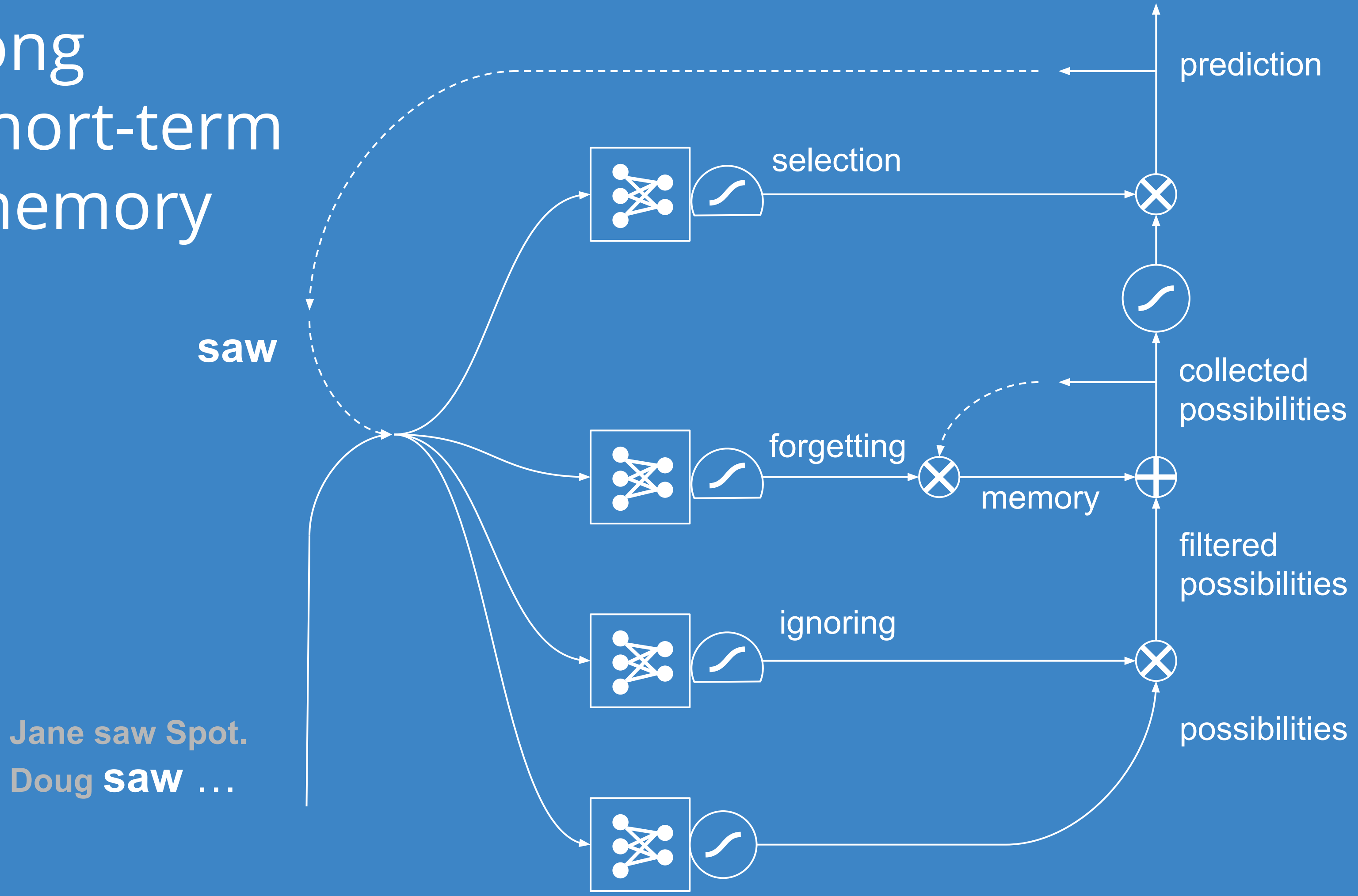
long short-term memory



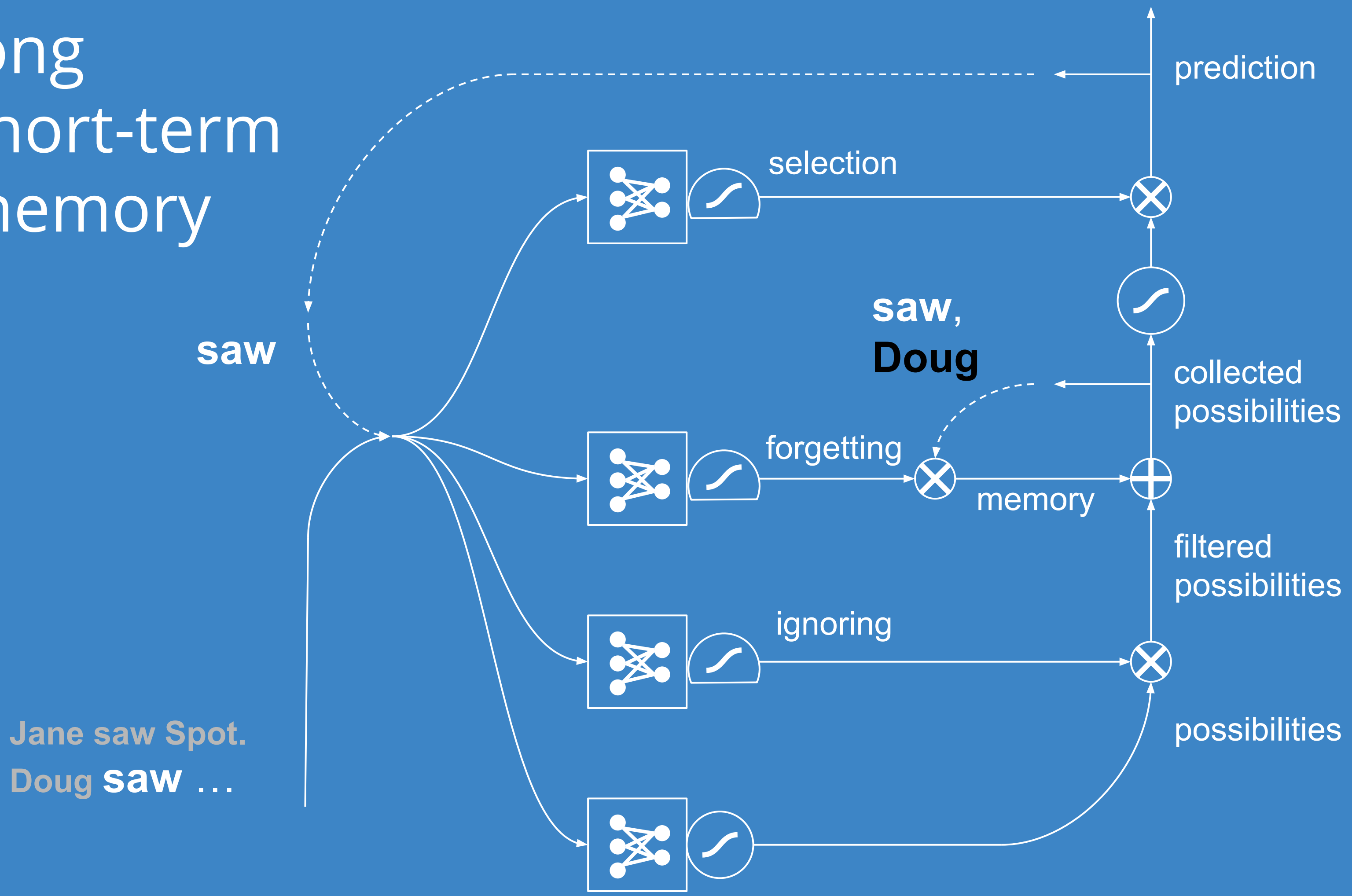
long short-term memory



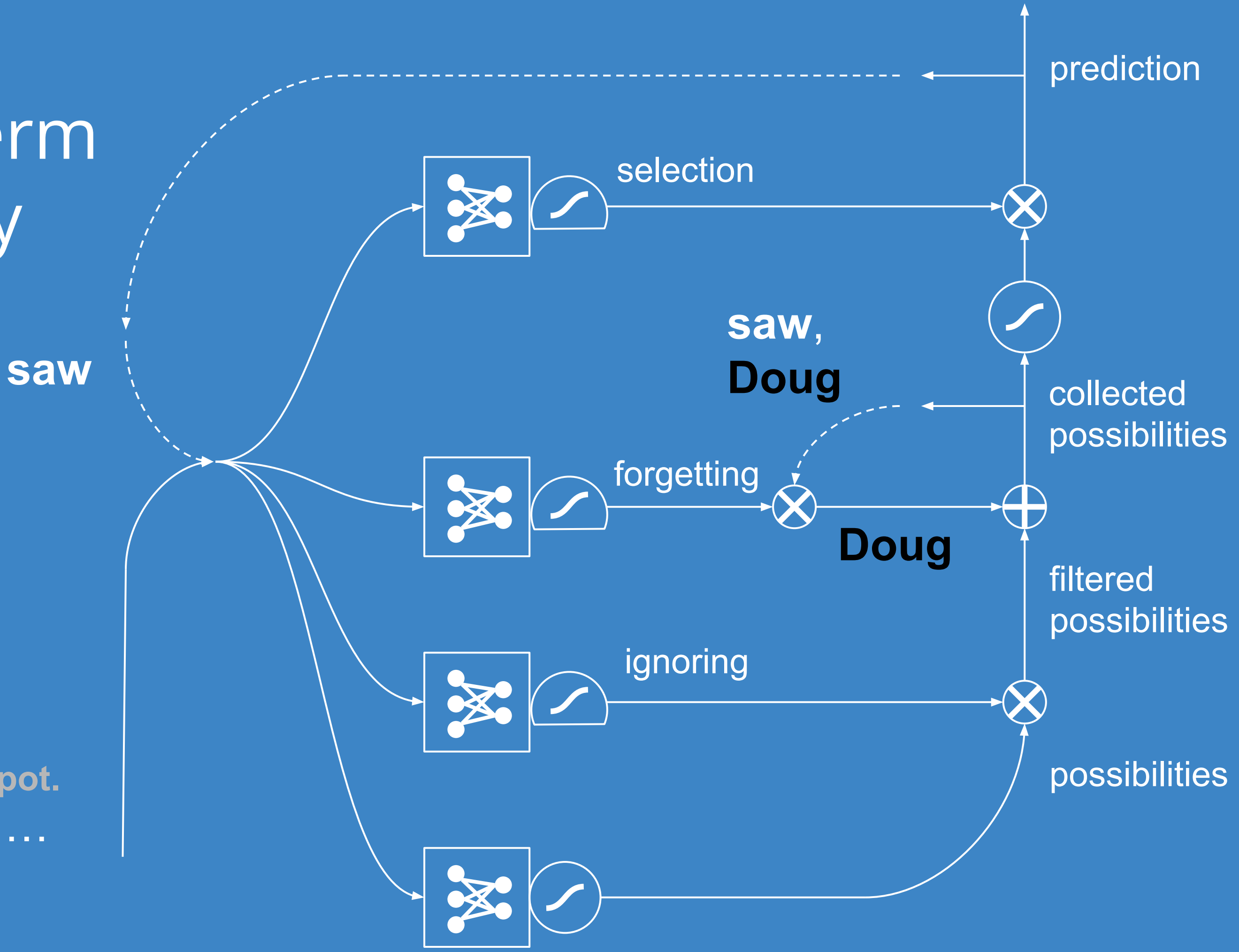
long short-term memory



long short-term memory

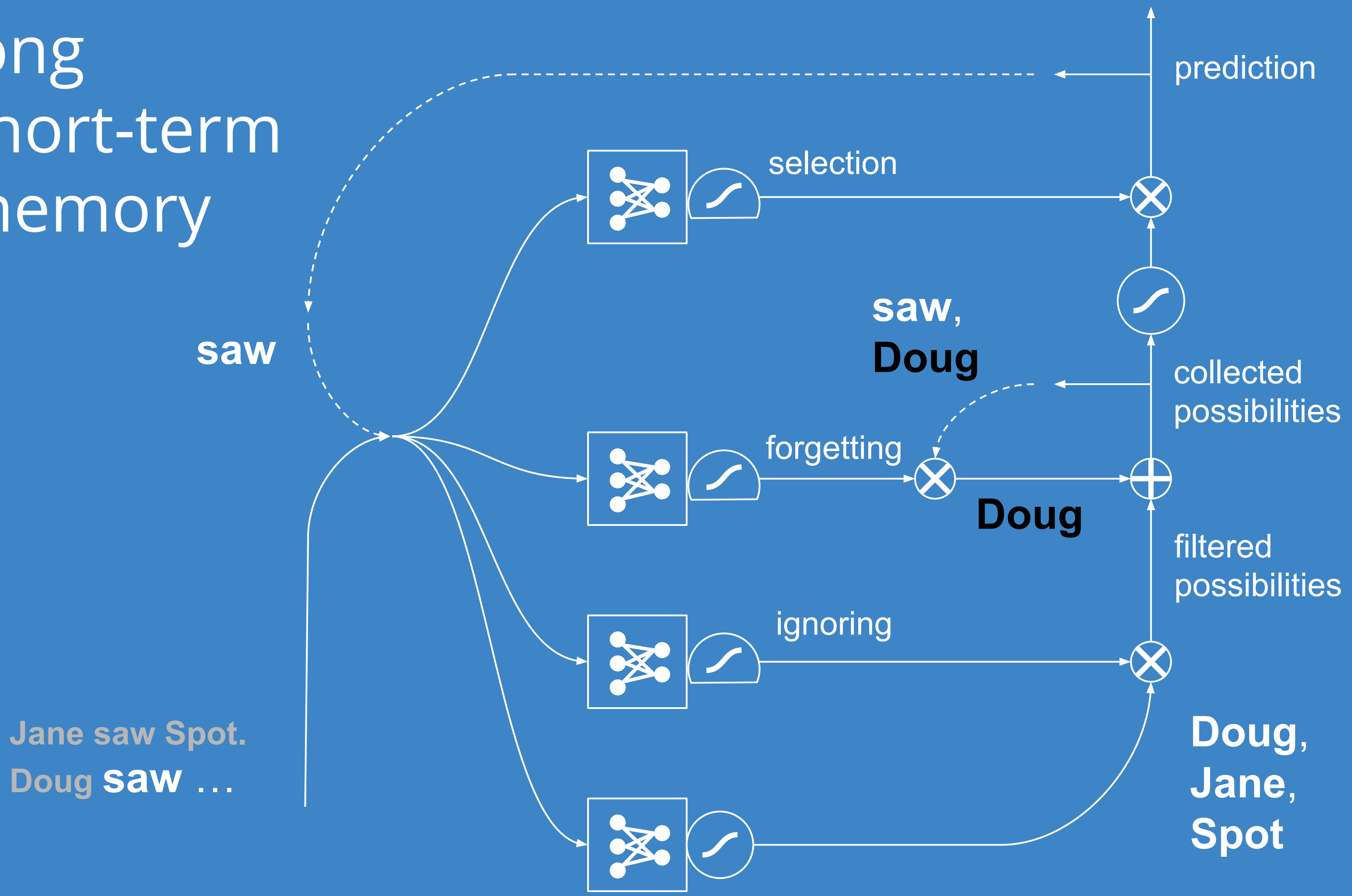


long short-term memory

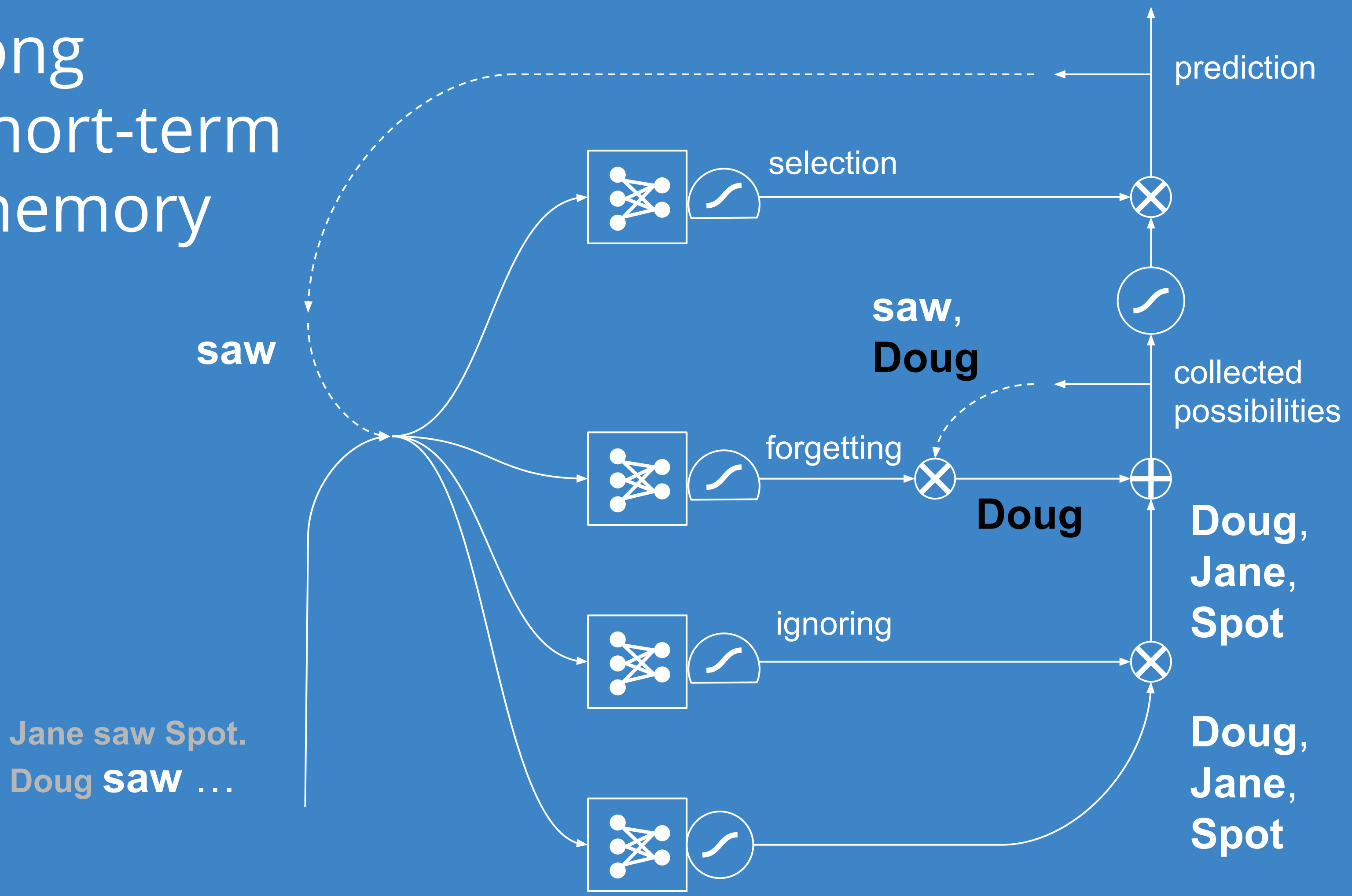


Jane saw Spot.
Doug **saw** …

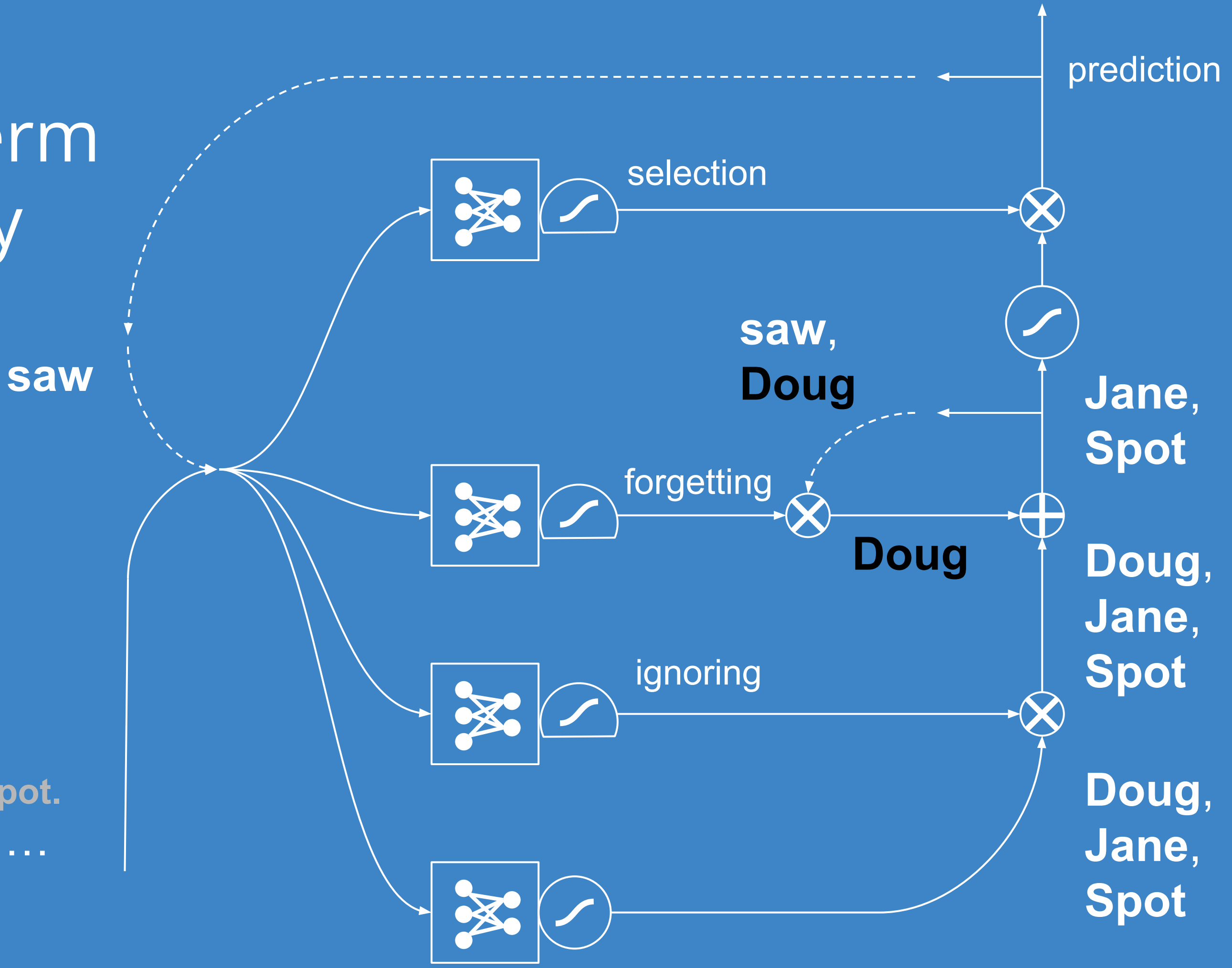
long short-term memory



long short-term memory

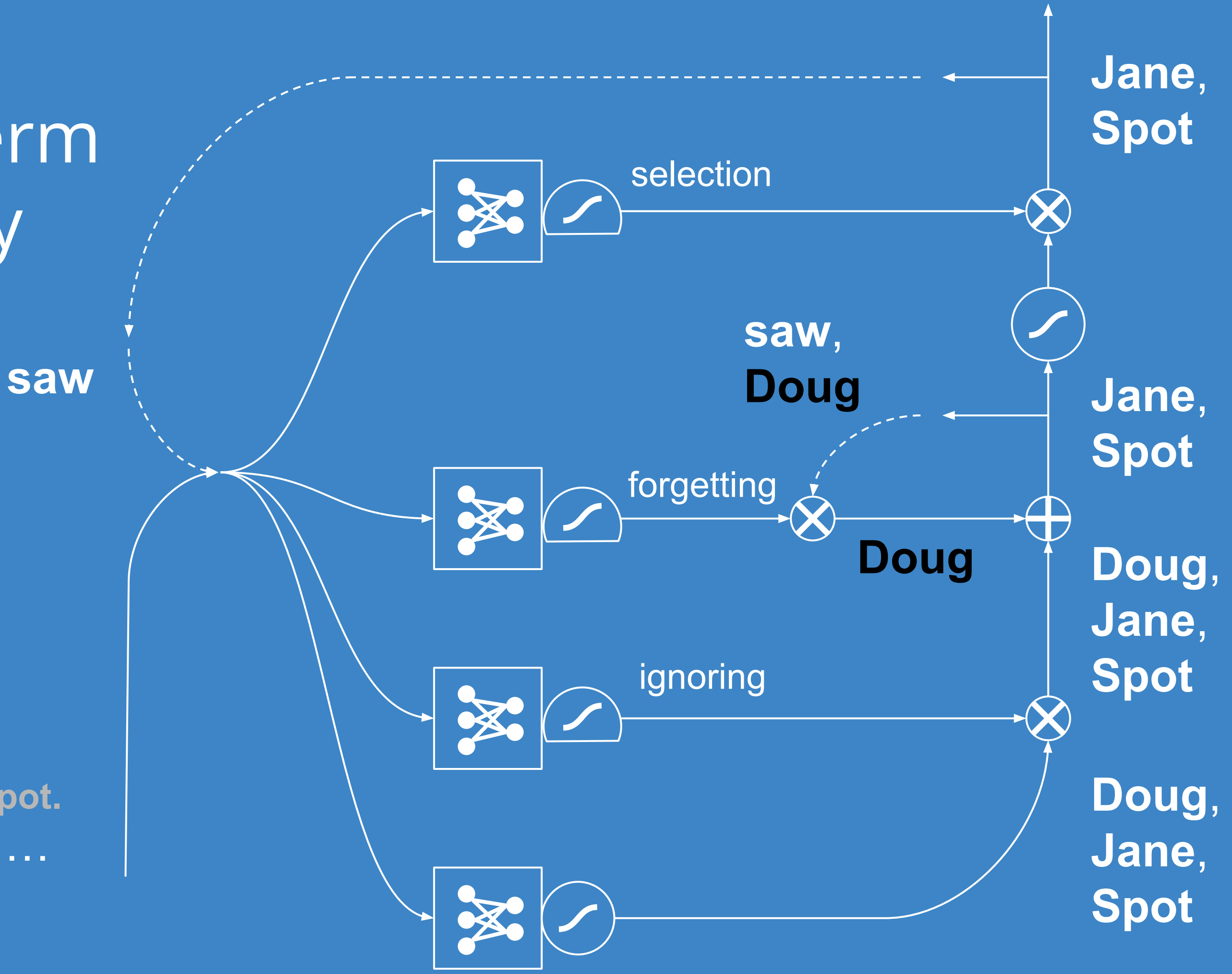


long short-term memory



Jane saw Spot.
Doug **saw** …

long short-term memory



Jane saw Spot.
Doug **saw** …

Sequential patterns

Text

Speech

Audio

Video

Physical processes

Anything embedded in time (almost everything)

Traditional LSTM with forget gates.[2][3]

Initial values: $c_0 = 0$ and $h_0 = 0$. The operator $\circ$ denotes the Hadamard product (entry-wise product).

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

Variables

- x_t : input vector
- h_t : output vector
- c_t : cell state vector
- W, U and b: parameter matrices and vector
- f_t, i_t and o_t : gate vectors
  - f_t : Forget gate vector. Weight of remembering old information.
  - i_t : Input gate vector. Weight of acquiring new information.
  - o_t : Output gate vector. Output candidate.

Activation functions

- σ_g : The original is a sigmoid function.
- σ_c : The original is a hyperbolic tangent.
- σ_h : The original is a hyperbolic tangent, but the peephole LSTM paper suggests $\sigma_h(x) = x$.[18][19]