Towards Understanding Long Short Term Memory Networks

HMI

1/28/2019

Jordan Rodu

Department of Statistics

University of Virginia

Towards Understanding Long Short Term Memory Networks

HMI

1/28/2019

Jordan Rodu

Department of Statistics

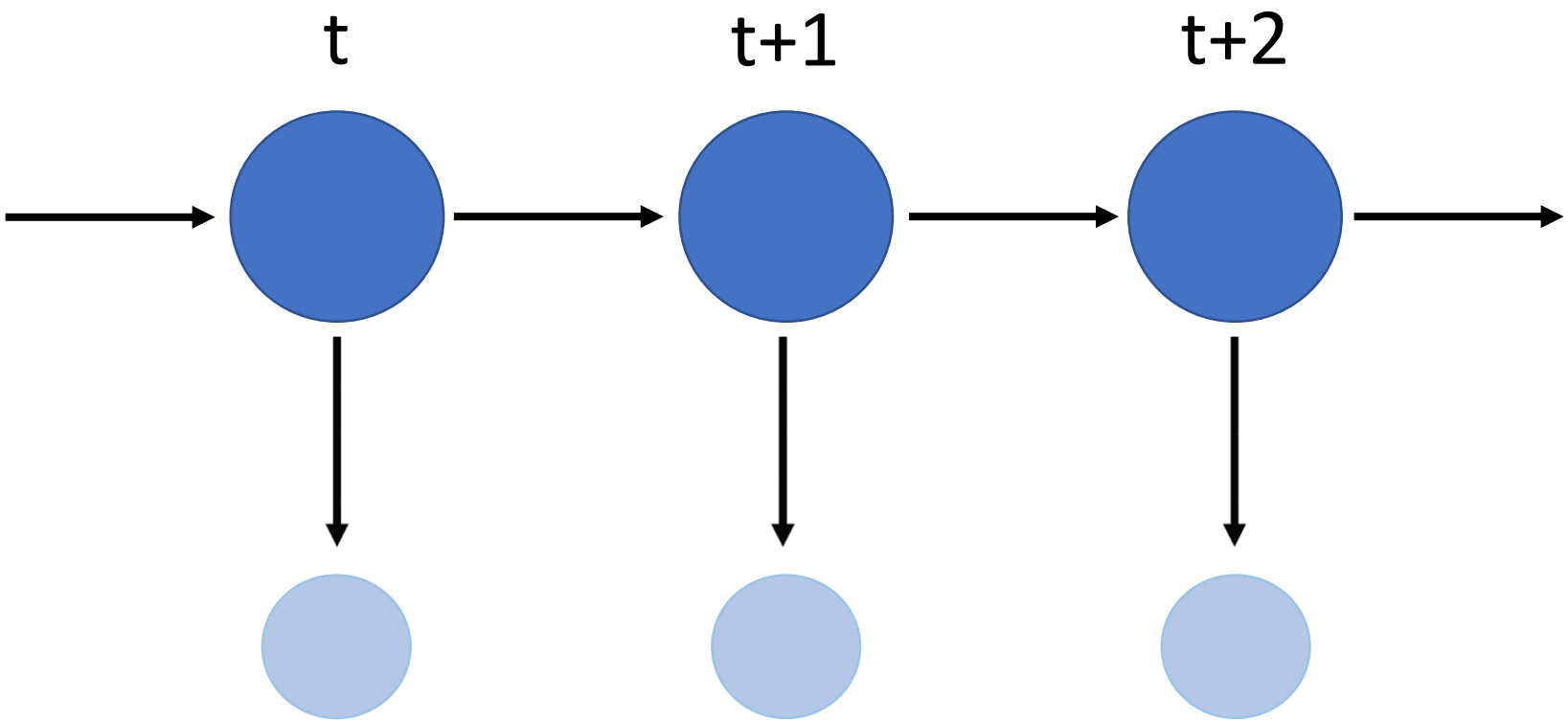
University of Virginia

with Joao Sedoc
University of Pennsylvania
Department of Computer Science

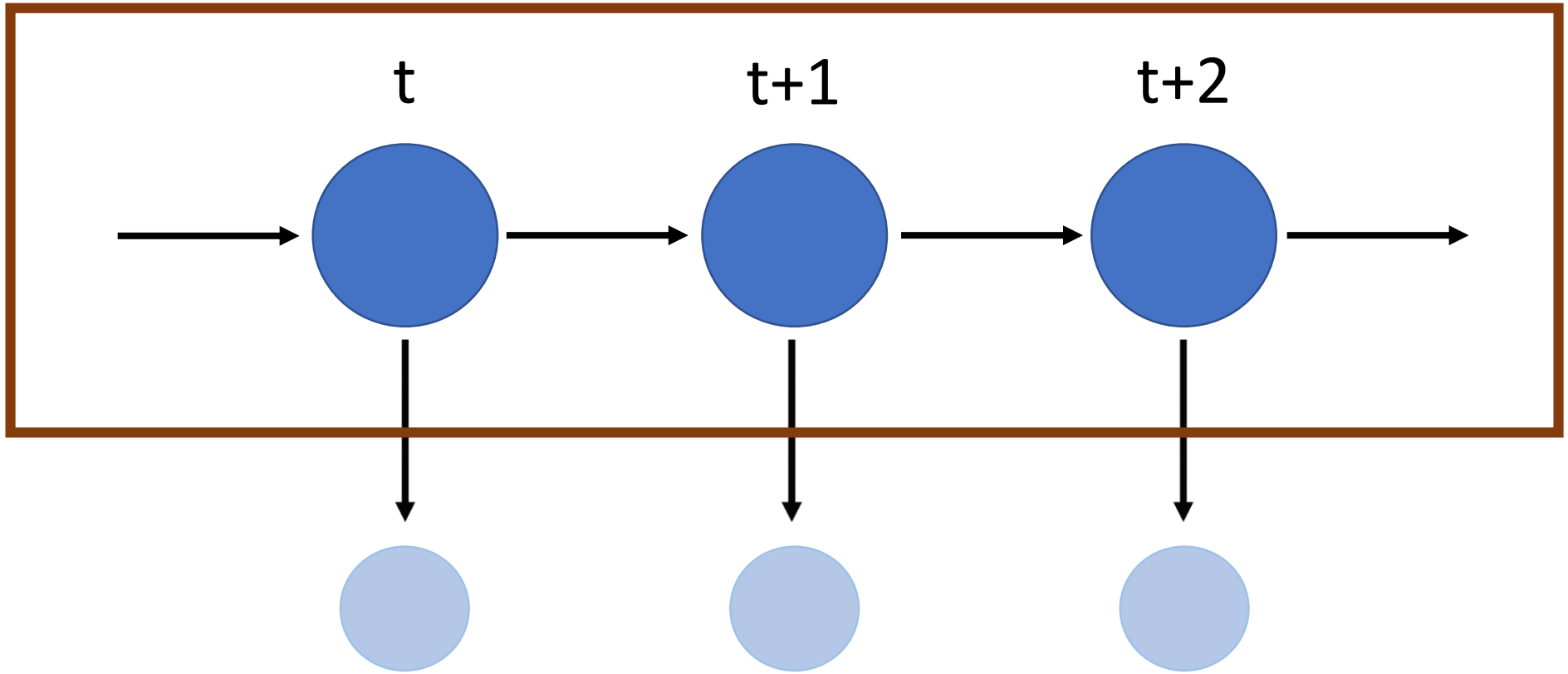
Mapping LSTMs to state space models

- Goal is not to interpret results on specific data
- Rather, map LSTM onto reasonable models- understand the space of sequences captured by LSTMs
- Preliminary work, basic ideas

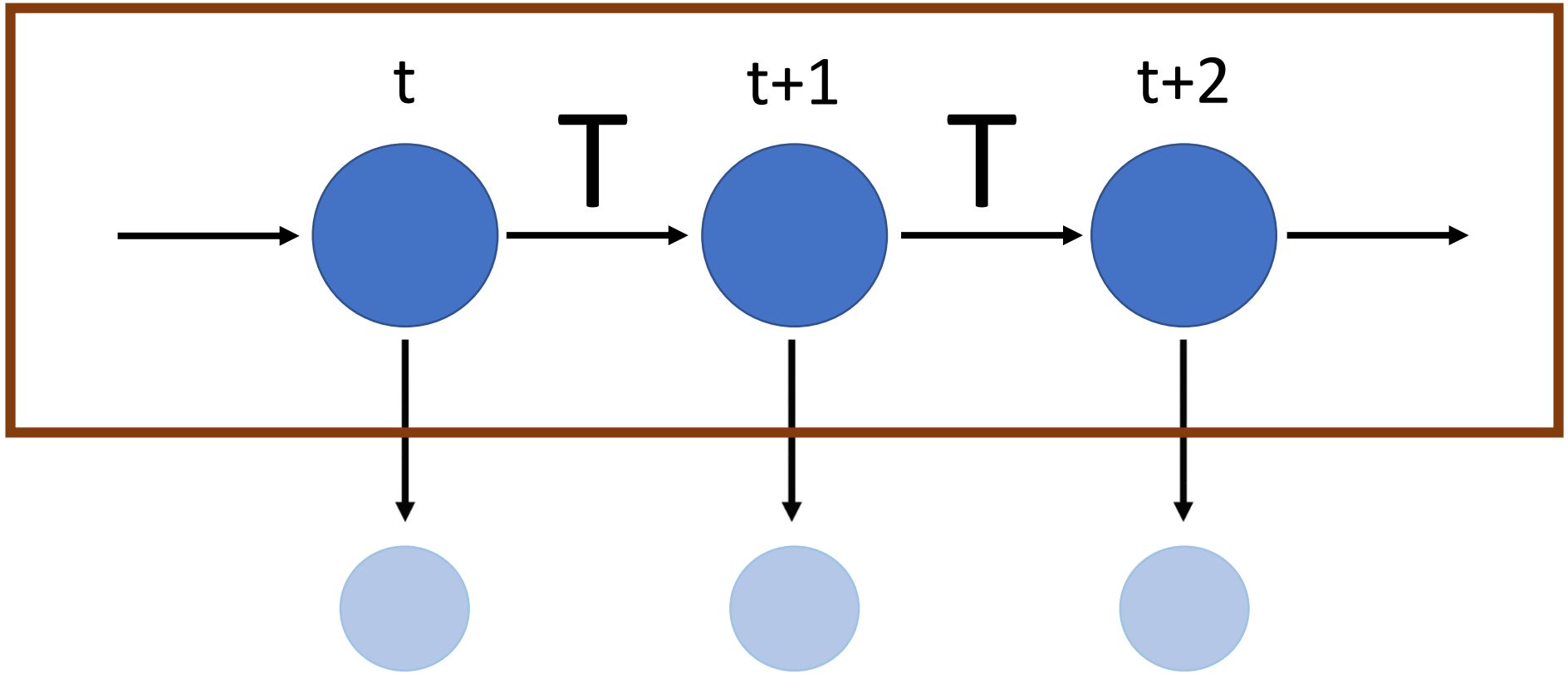
Hidden Markov Models



Hidden Markov Models

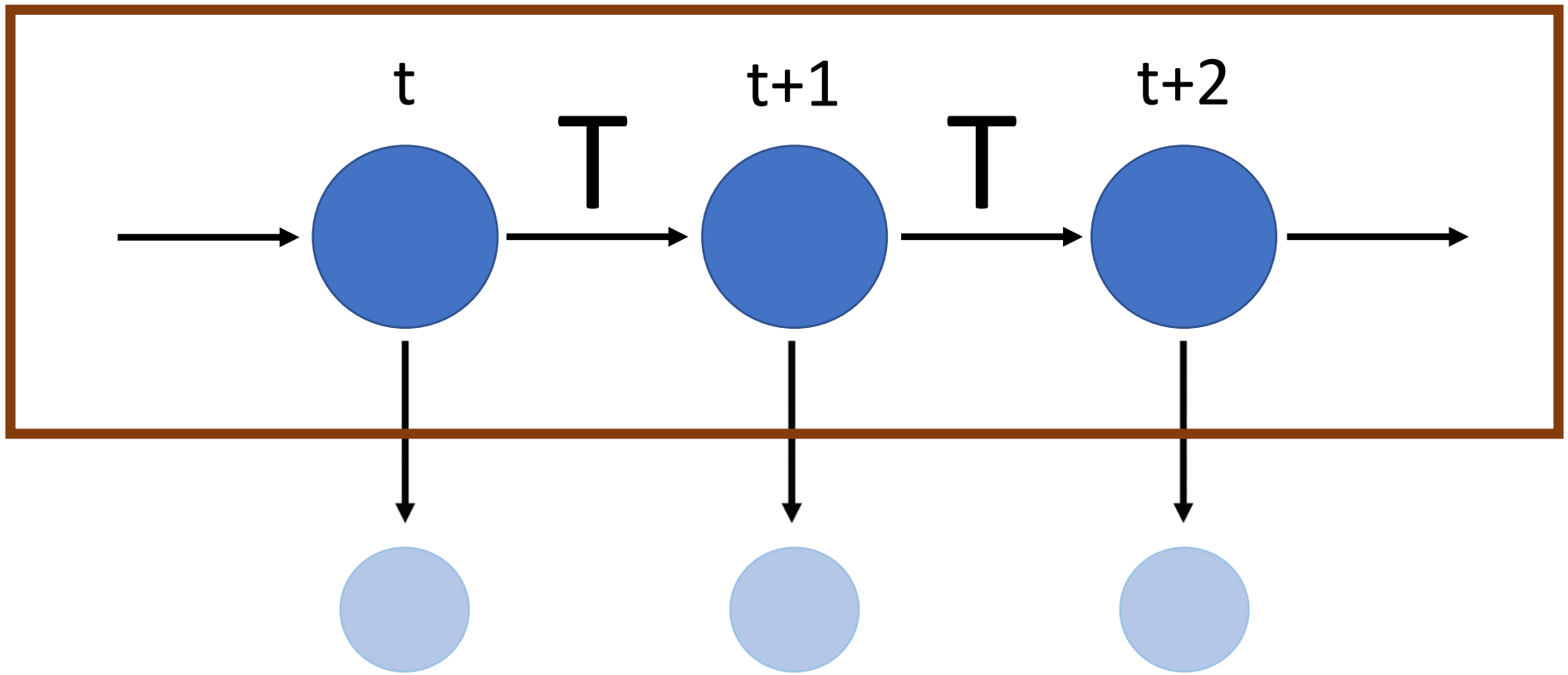


Hidden Markov Models

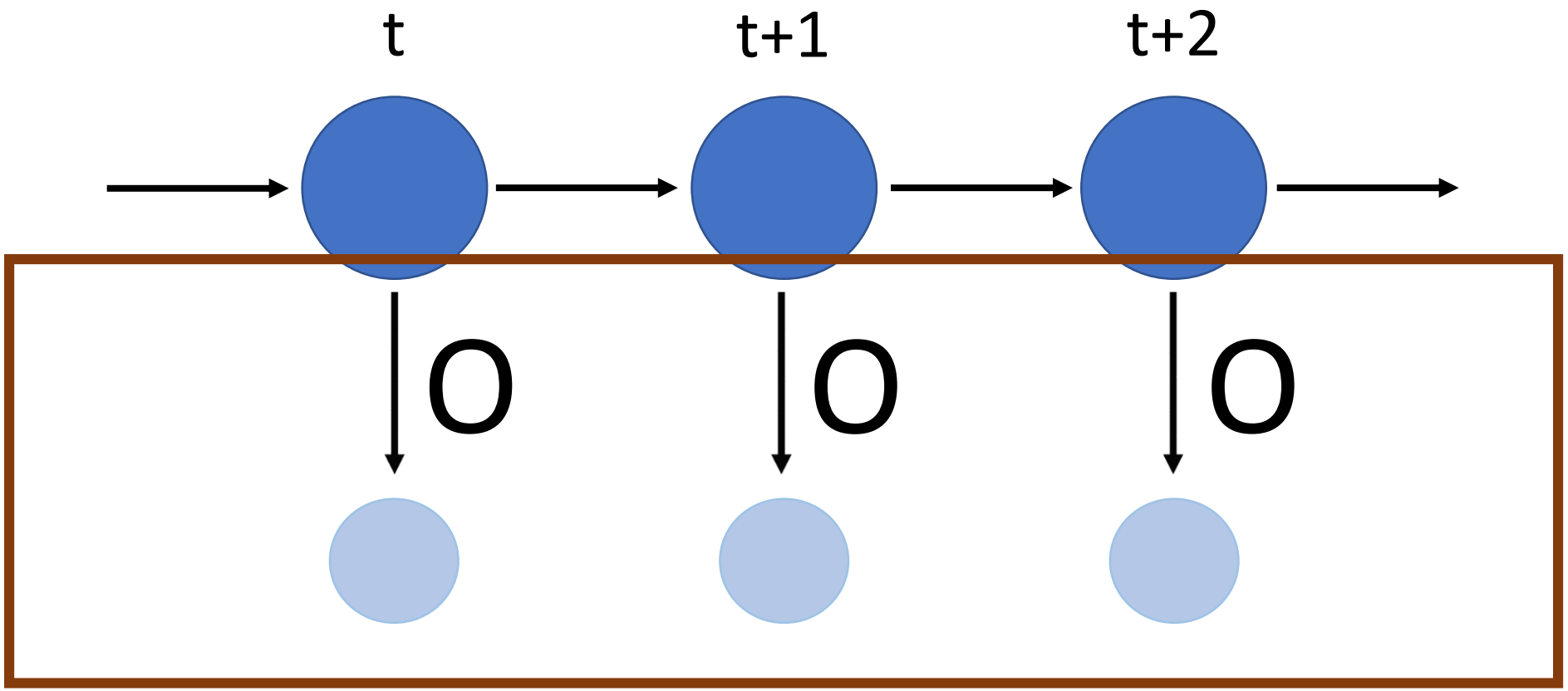


Hidden Markov Models

$$p(x_{t+1}|\, x_{1:t}) = p(x_{t+1}|\, x_t)$$

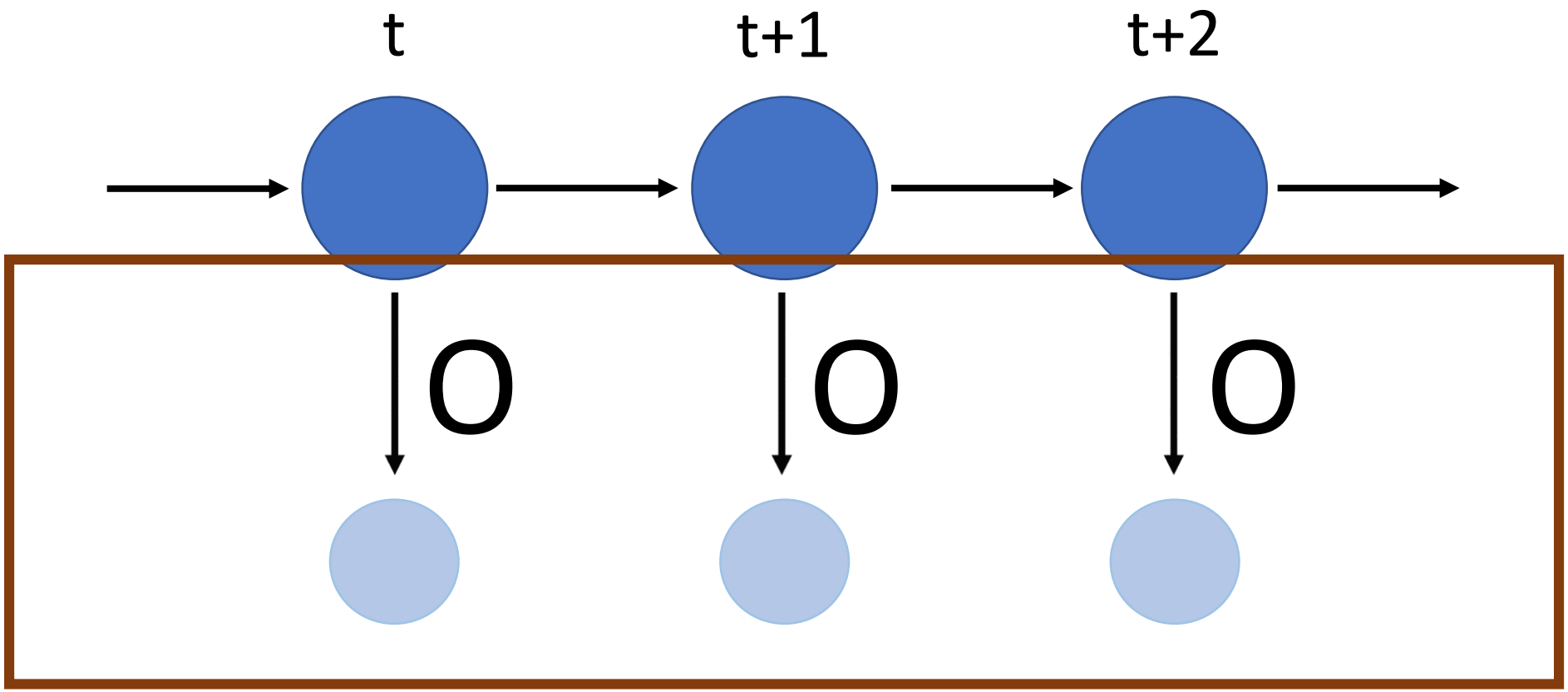


Hidden Markov Models



Hidden Markov Models

$p(y_t | x_t)$



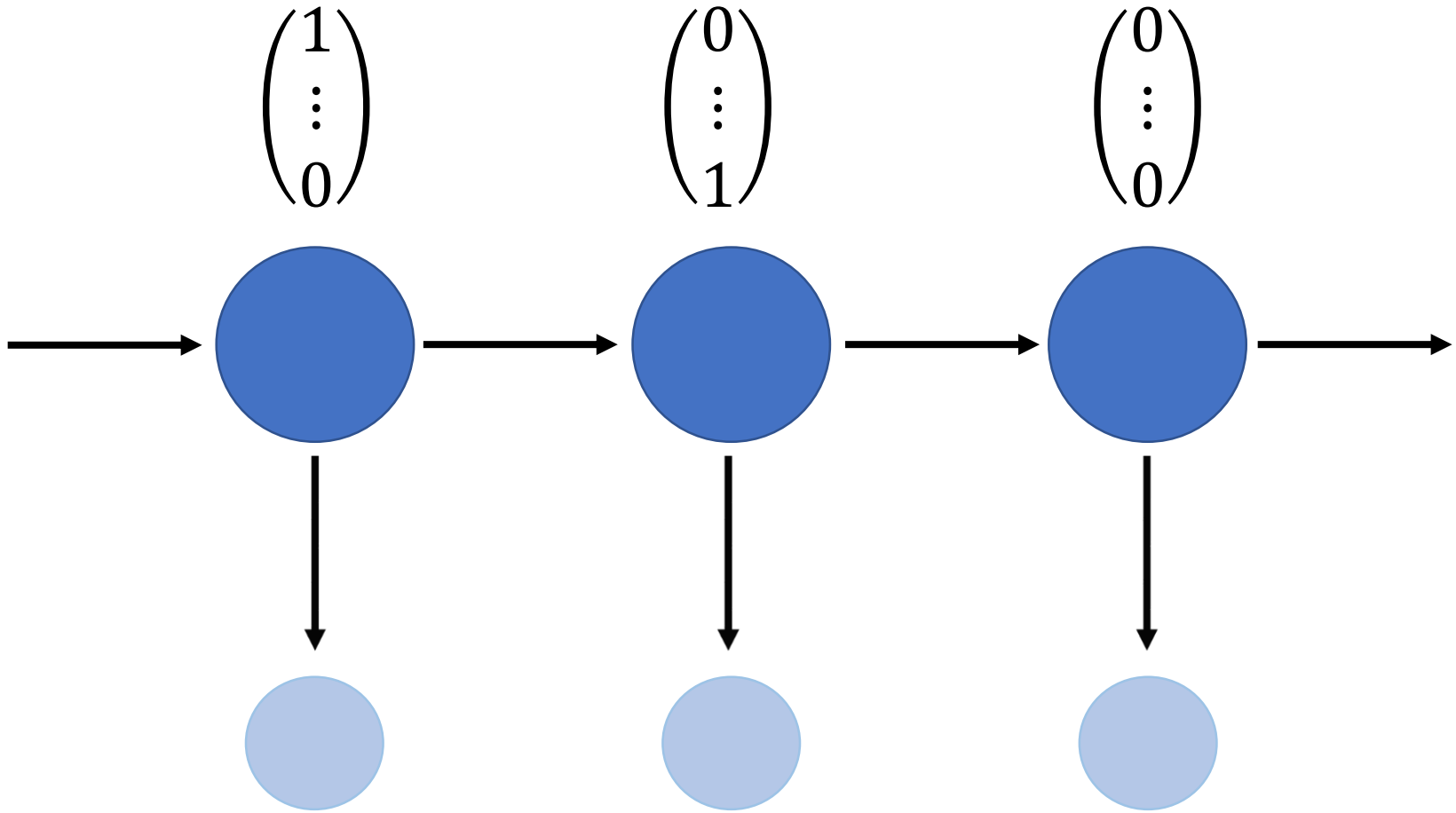
Hidden Markov Models- flavors

- Output
  - Discrete
  - Continuous
  - Low dimensional
  - High dimensional
- States
  - Discrete
  - Continuous
    - Low dimensional
    - High dimensional
- Time
  - Discrete
  - Continuous

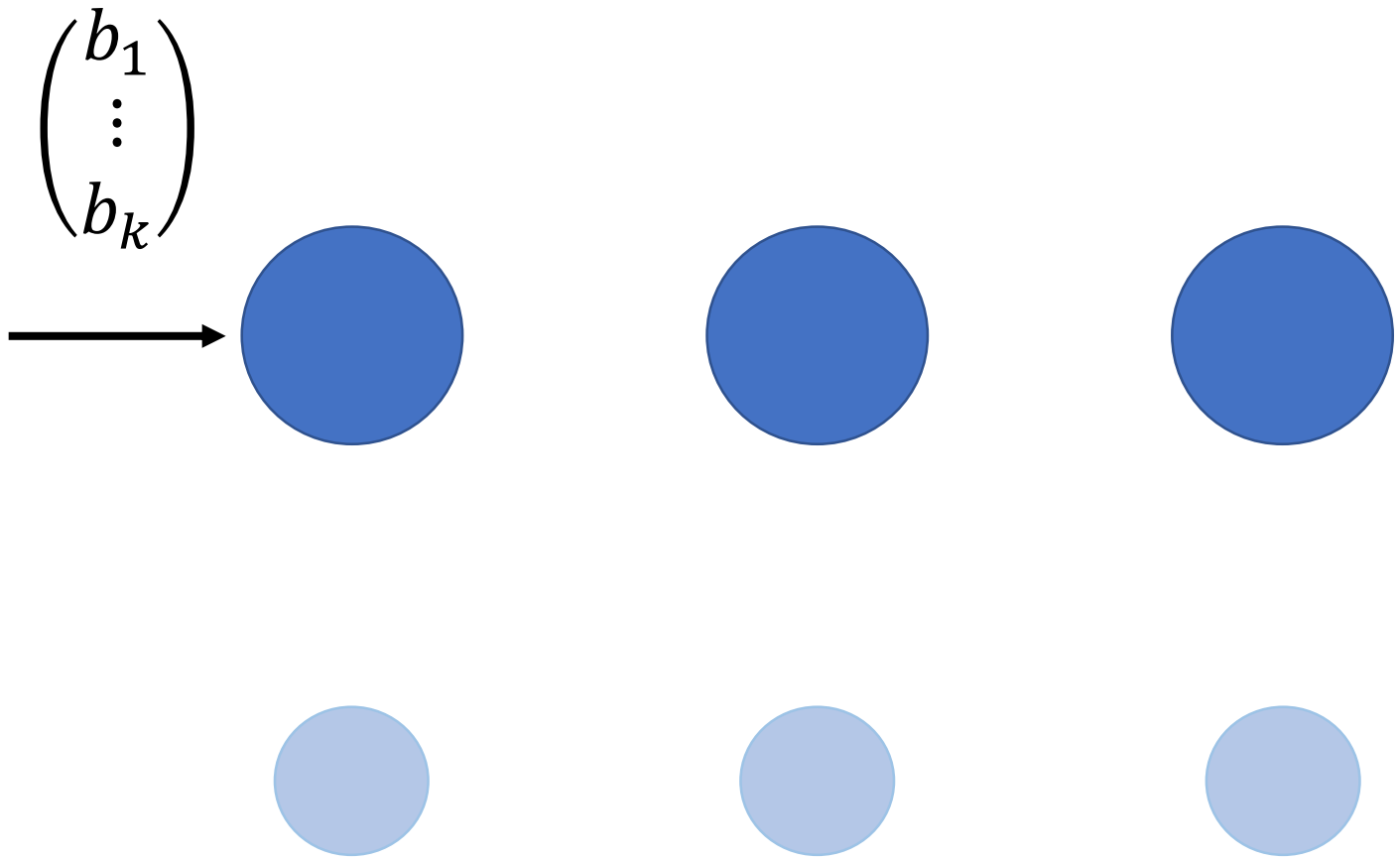
Hidden Markov Models- flavors

- Output
  - Discrete
  - Continuous
  - Low dimensional
  - High dimensional
- States
  - **Discrete**
  - Continuous
    - Low dimensional
    - High dimensional
- Time
  - **Discrete**
  - Continuous

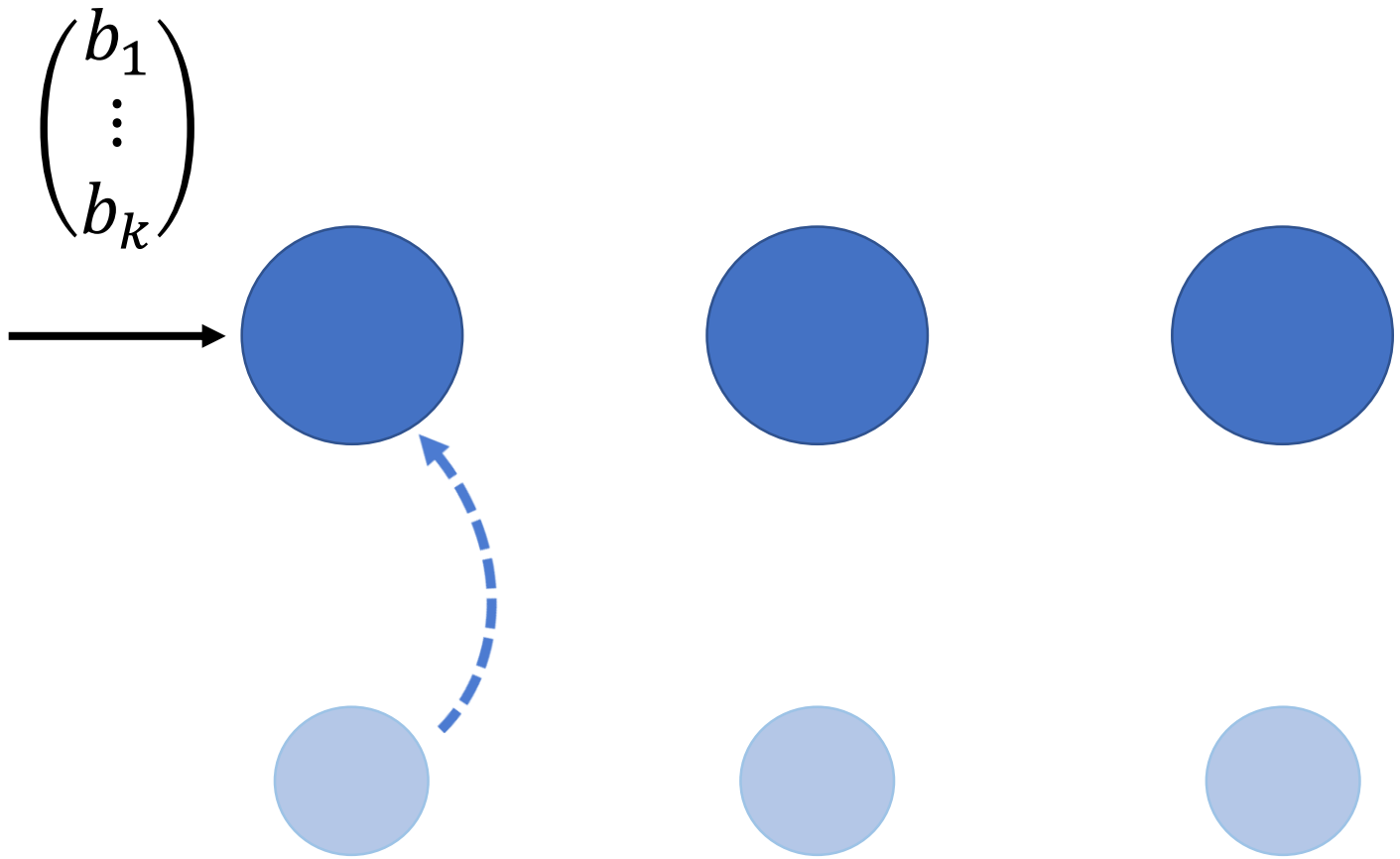
Hidden Markov Models



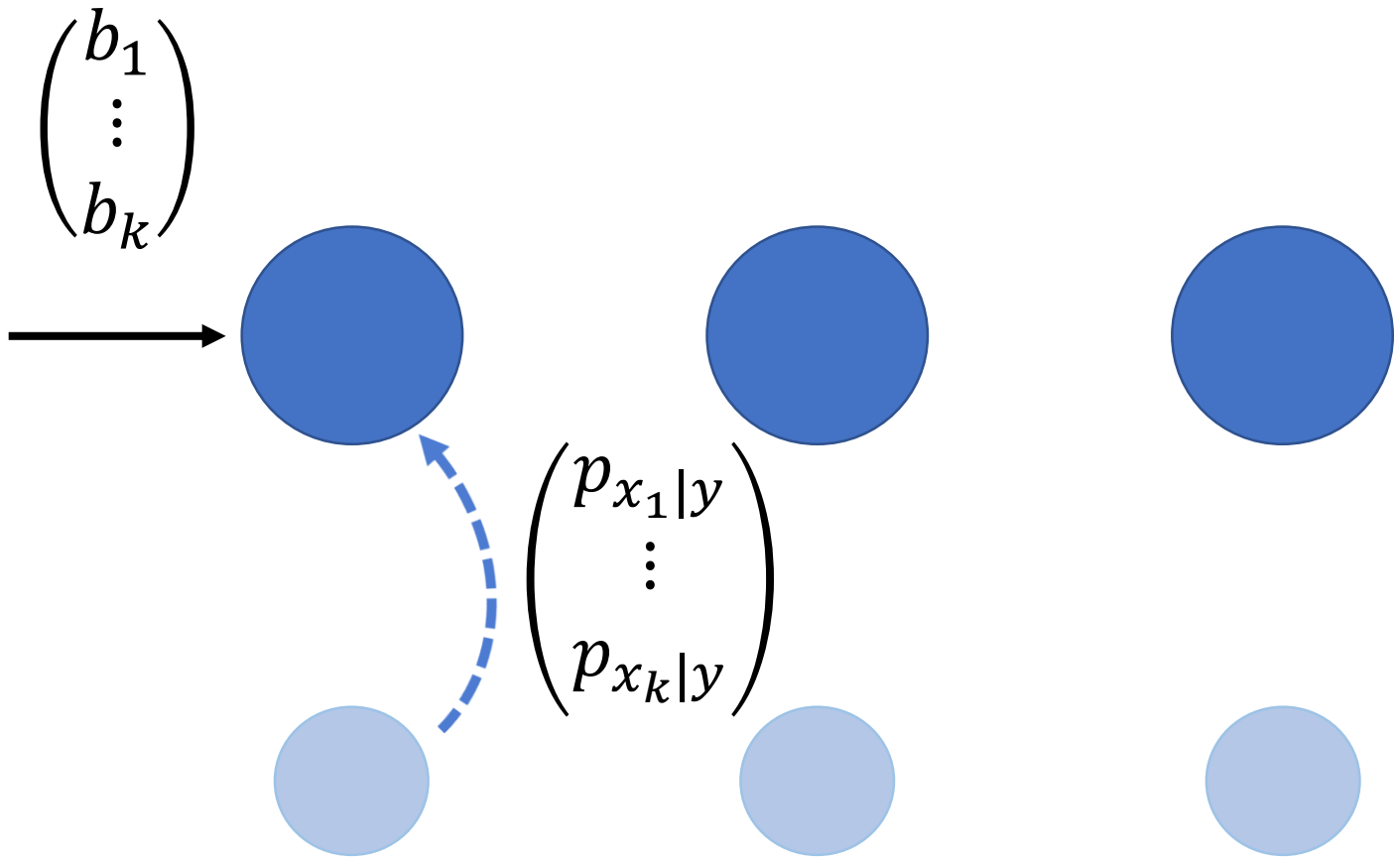
Hidden Markov Models



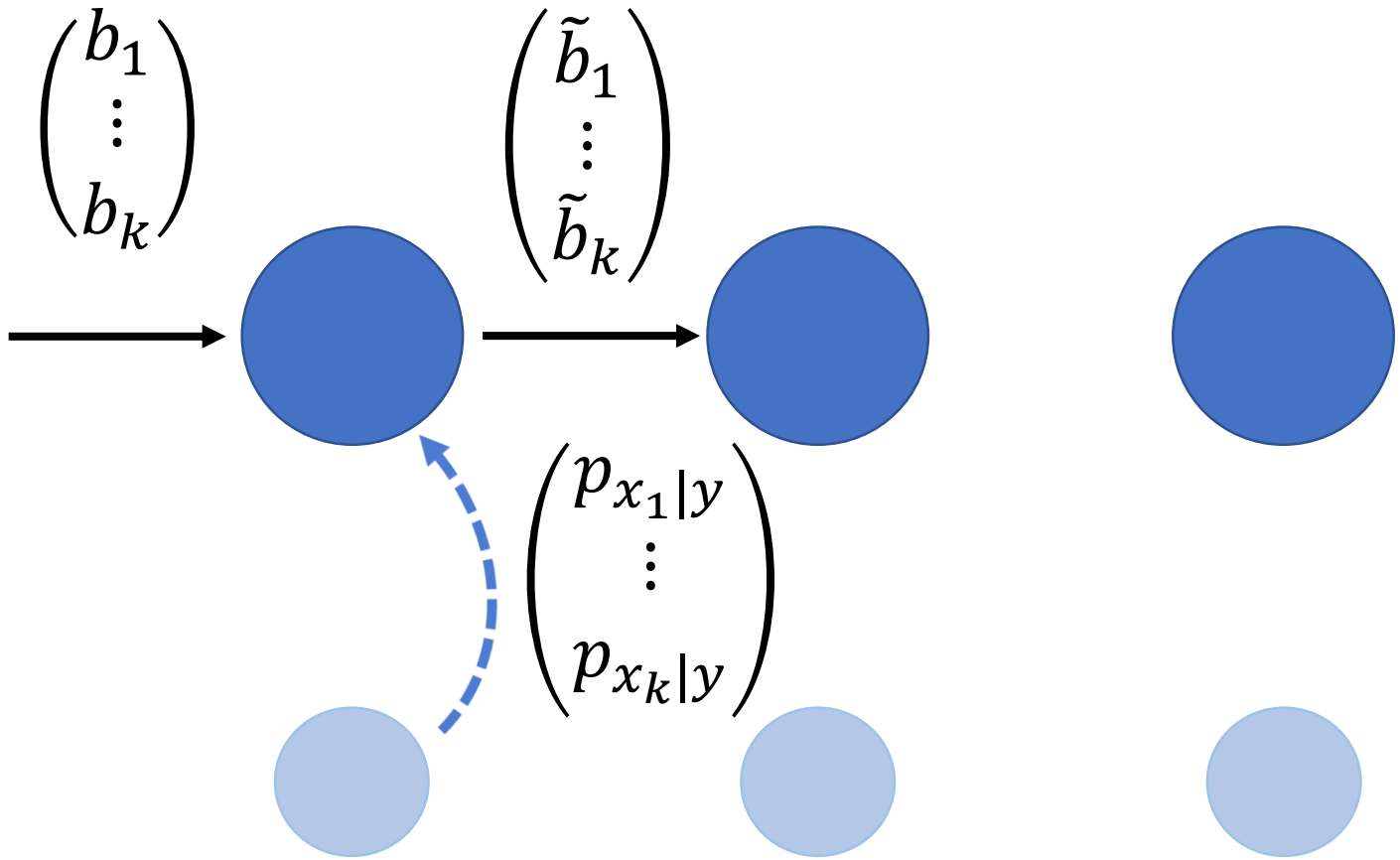
Hidden Markov Models



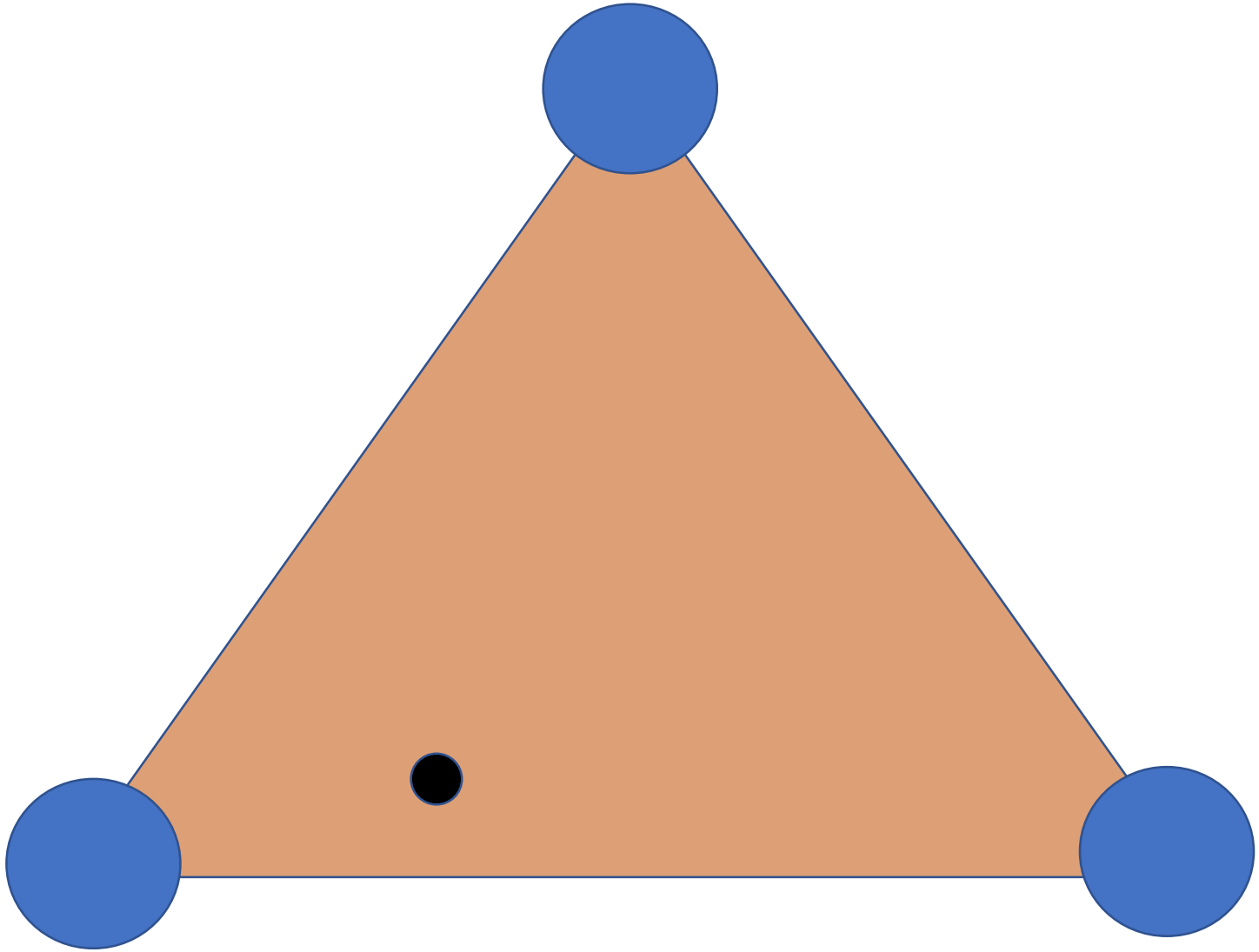
Hidden Markov Models



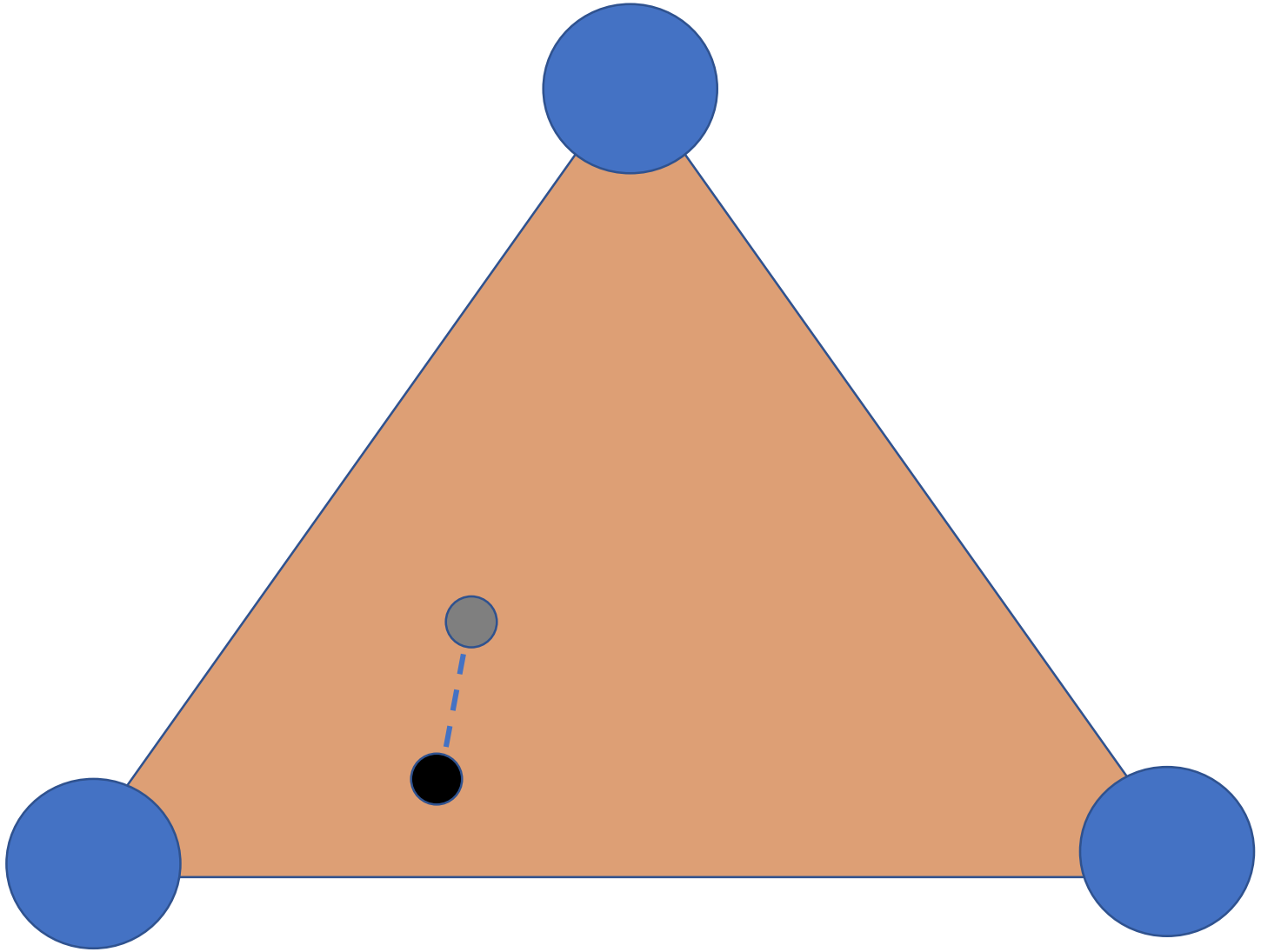
Hidden Markov Models



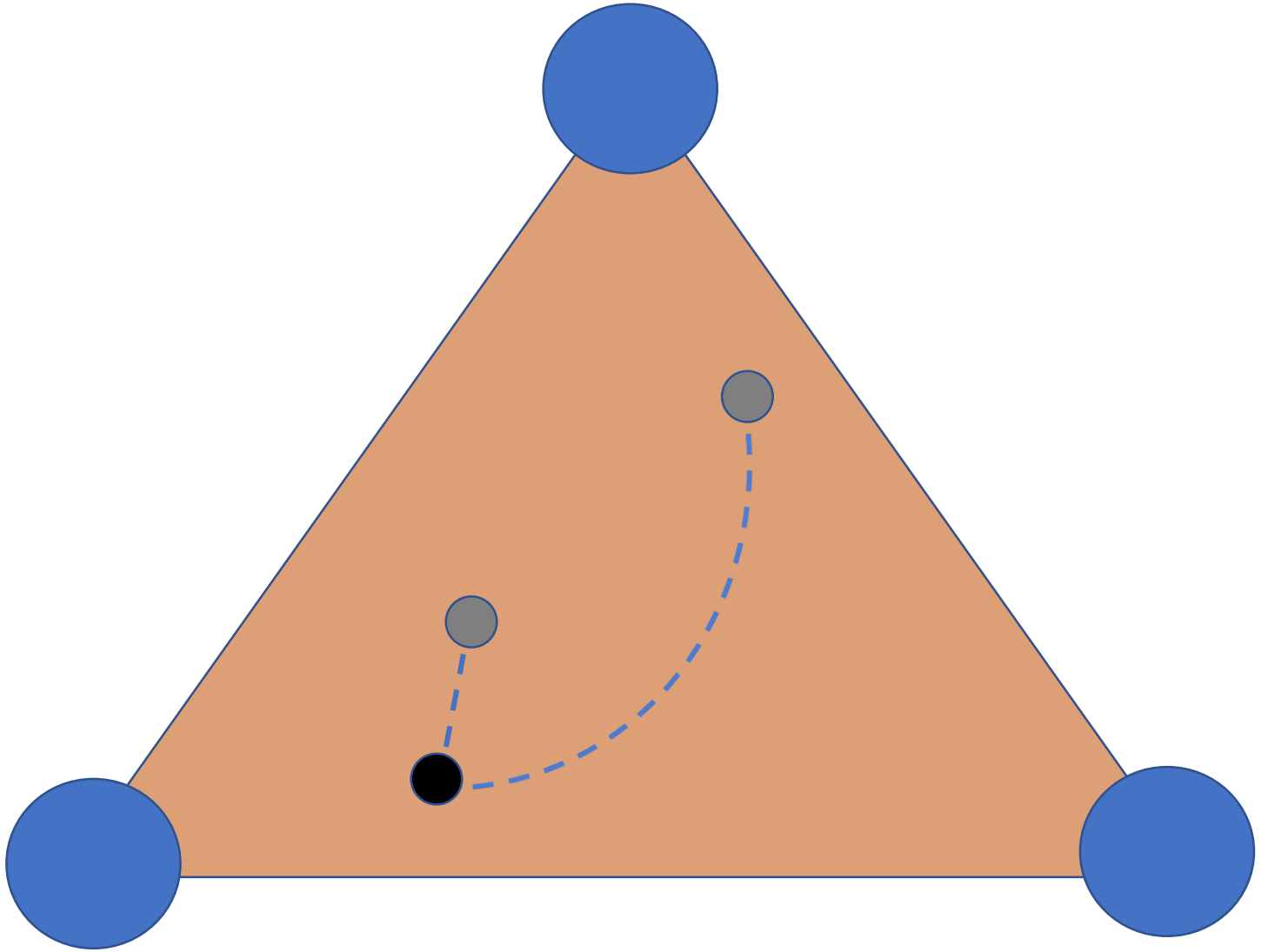
Hidden Markov Model- belief states



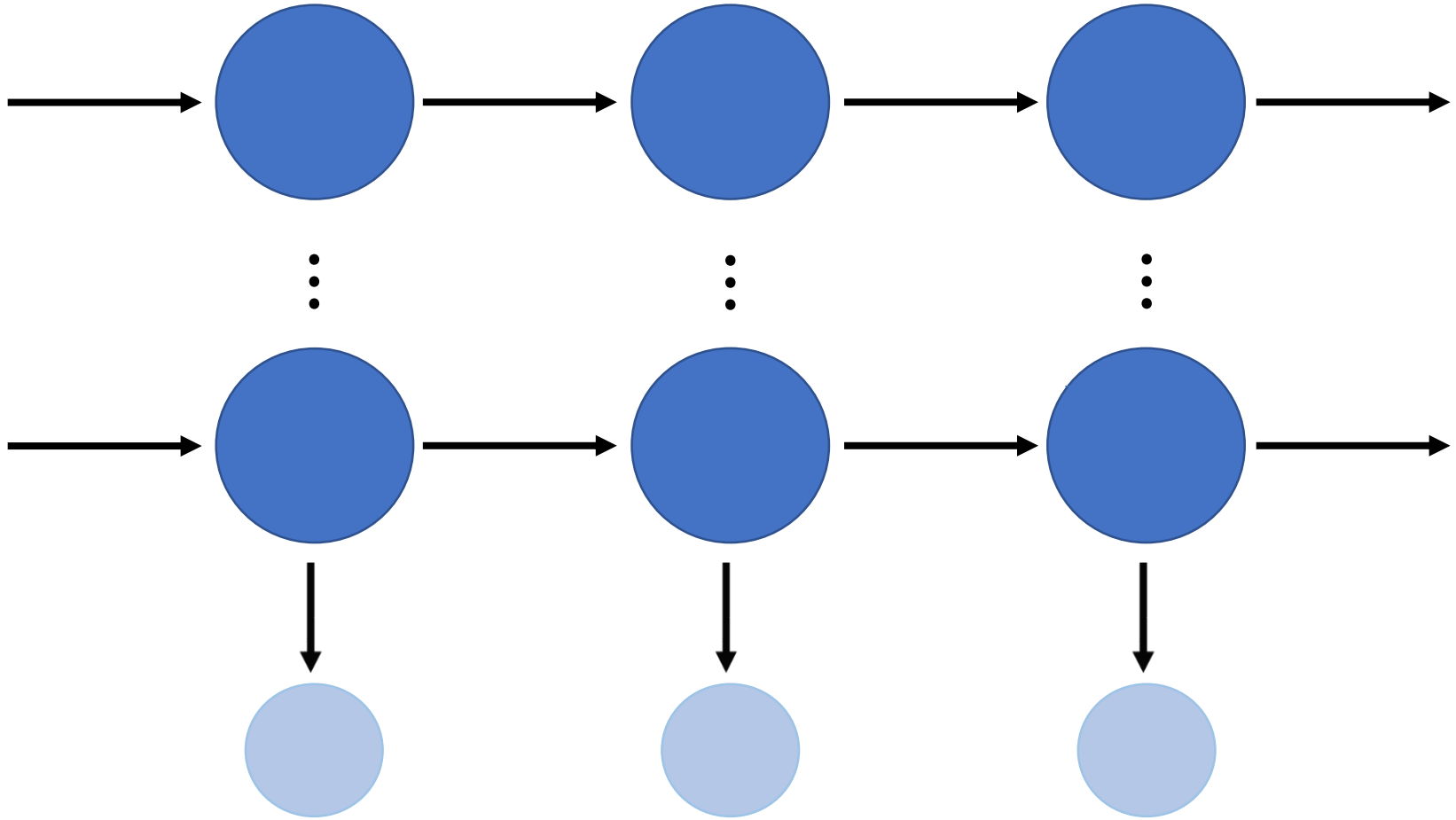
Hidden Markov Model- belief states



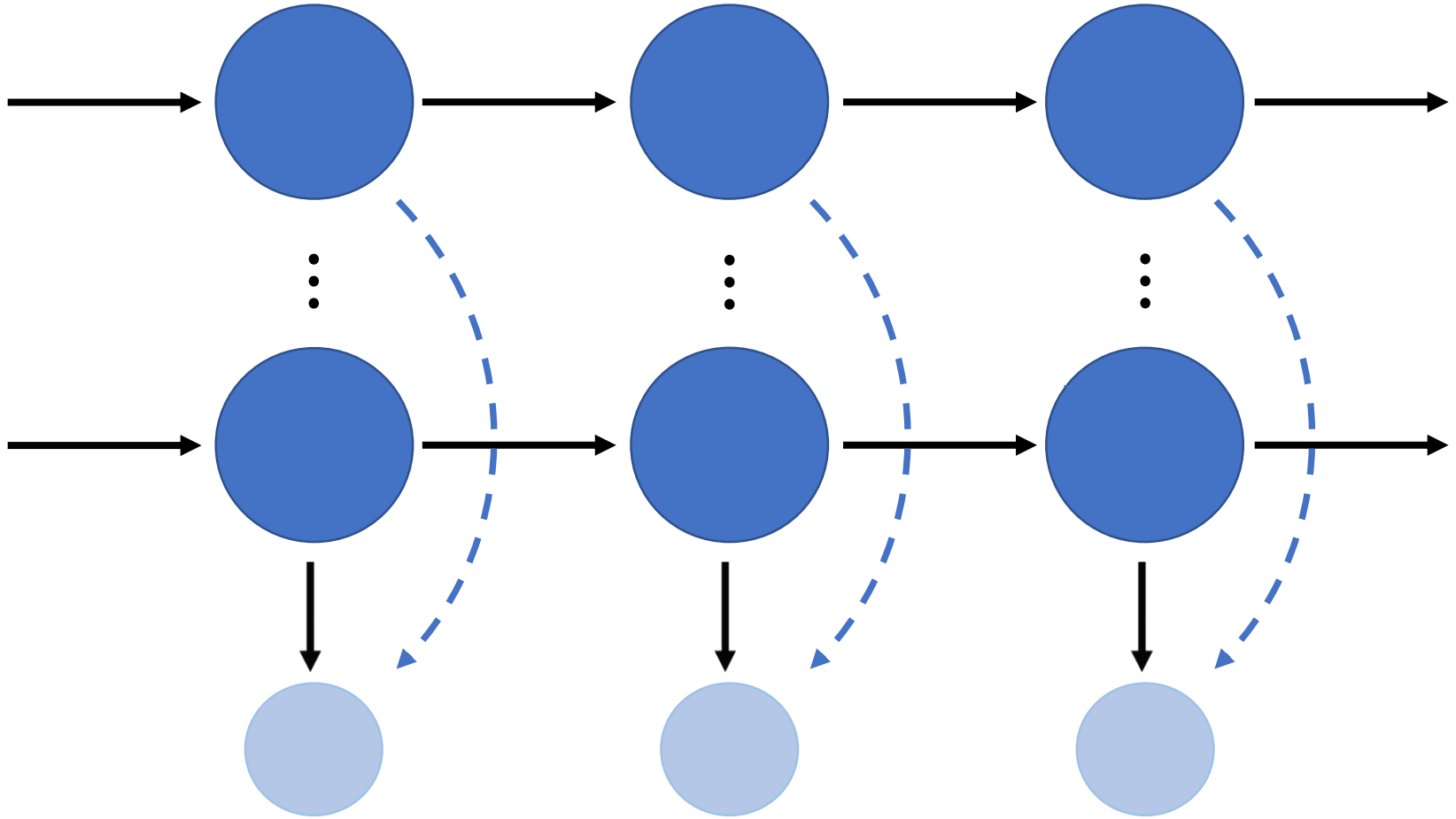
Hidden Markov Model- belief states



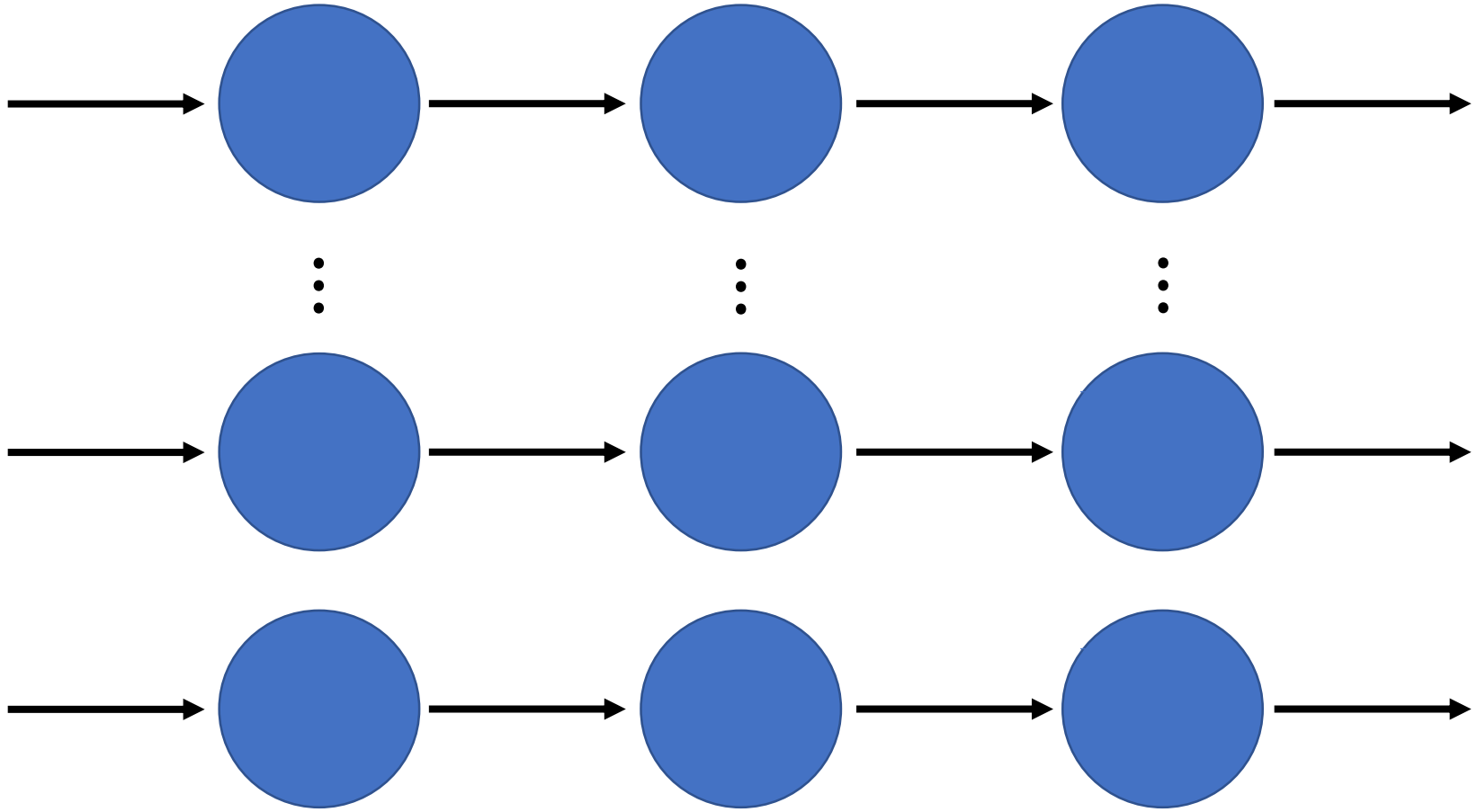
A few related architectures



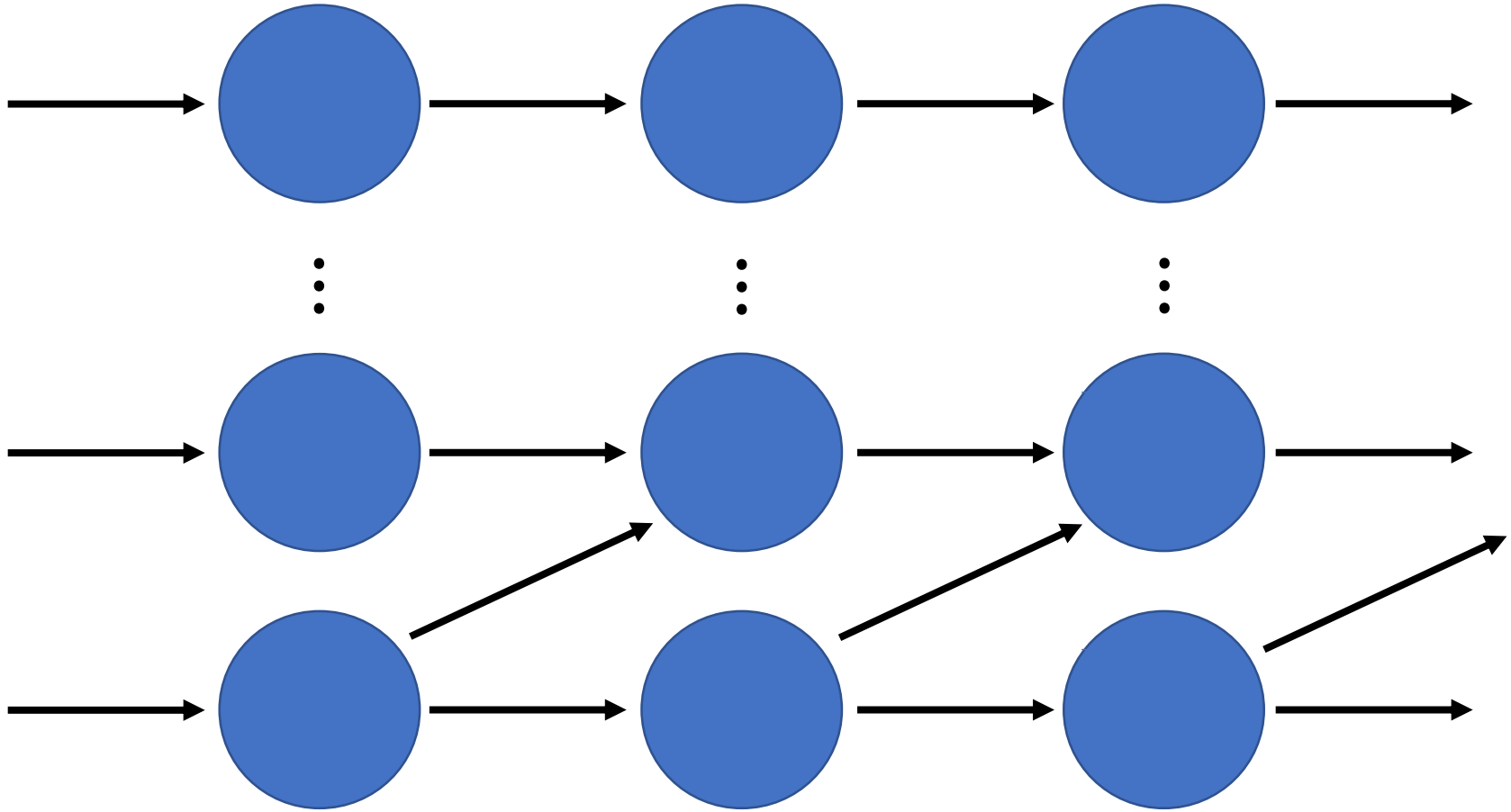
A few related architectures



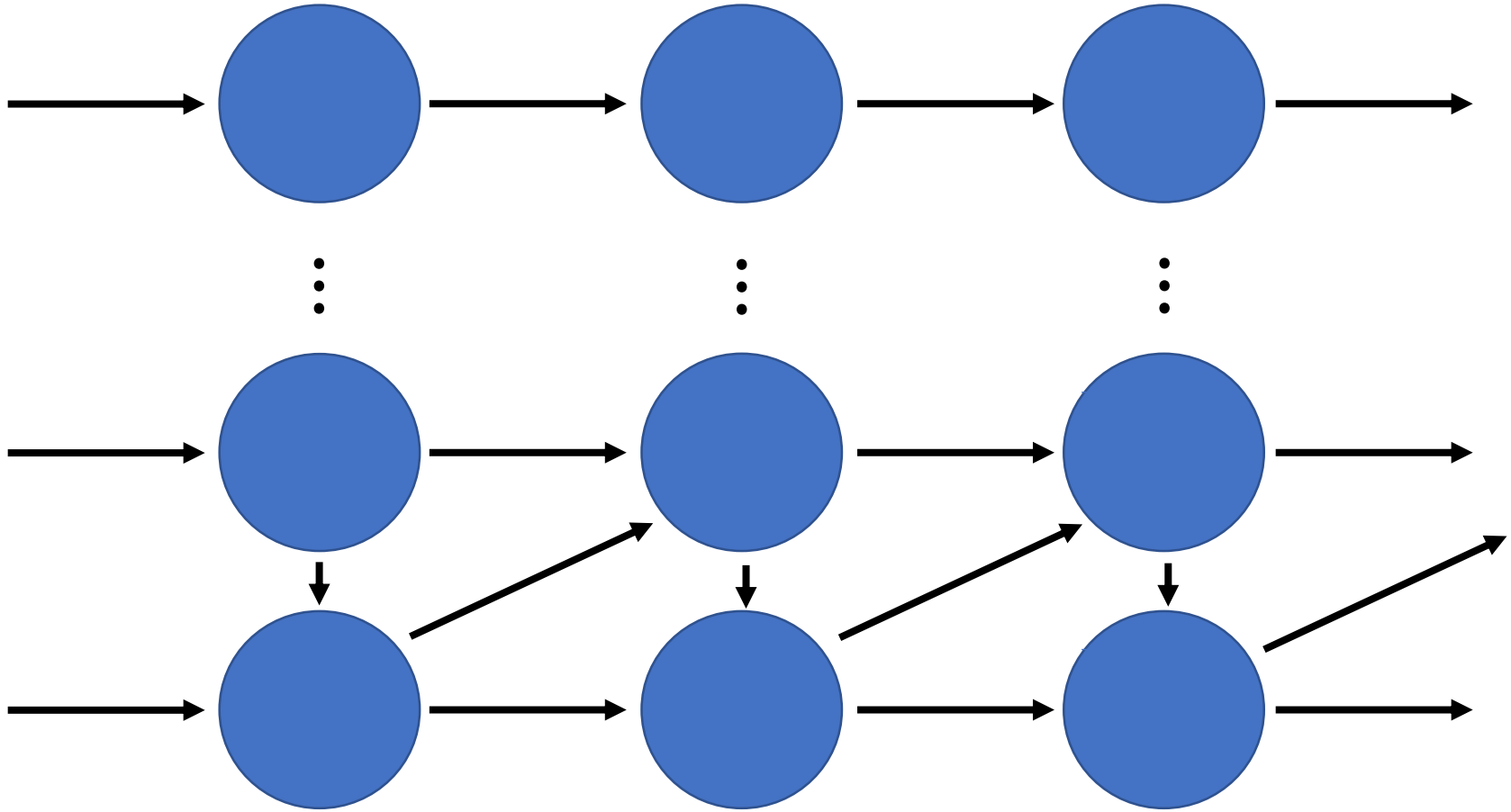
A few related architectures



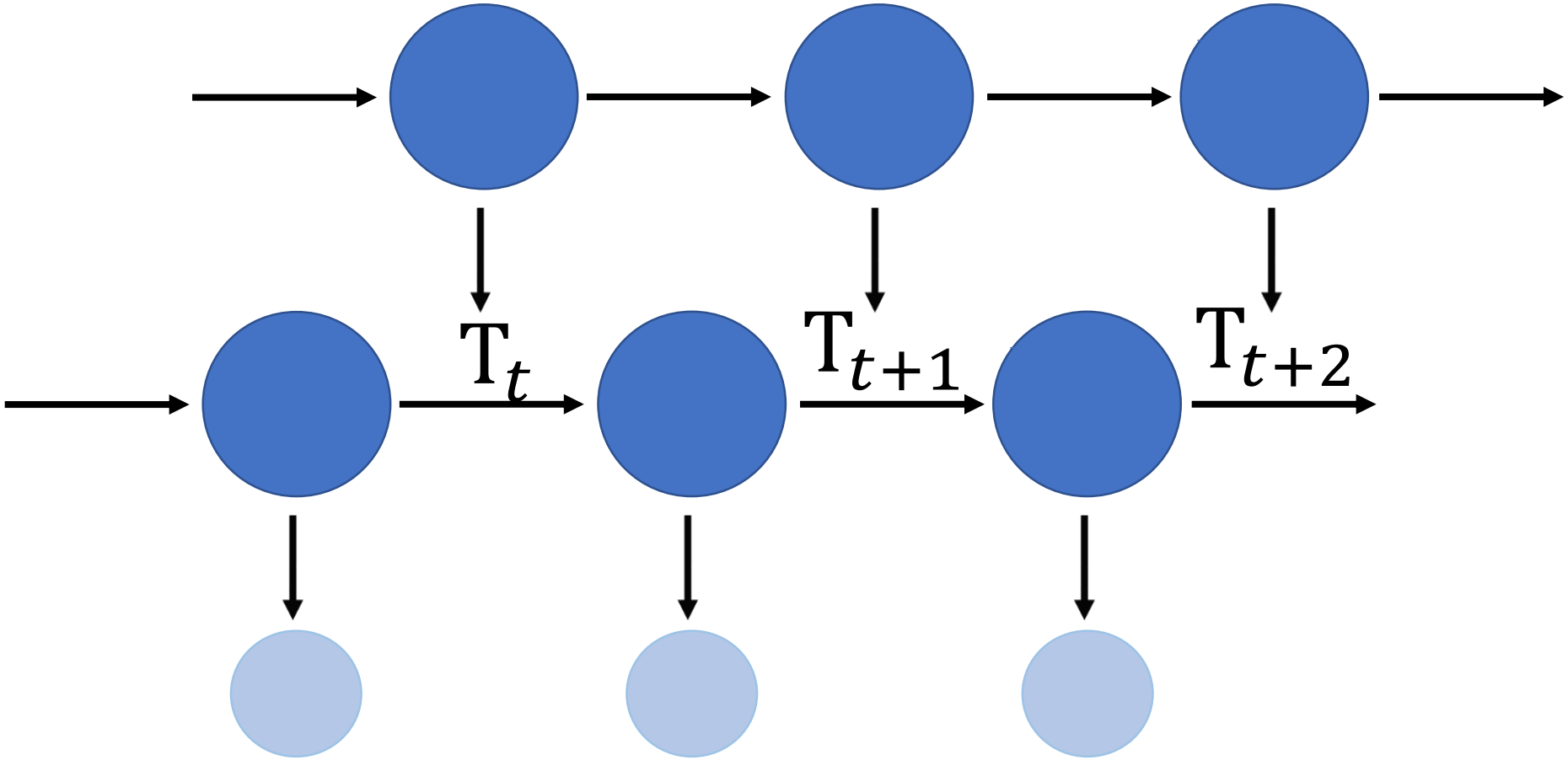
A few related architectures



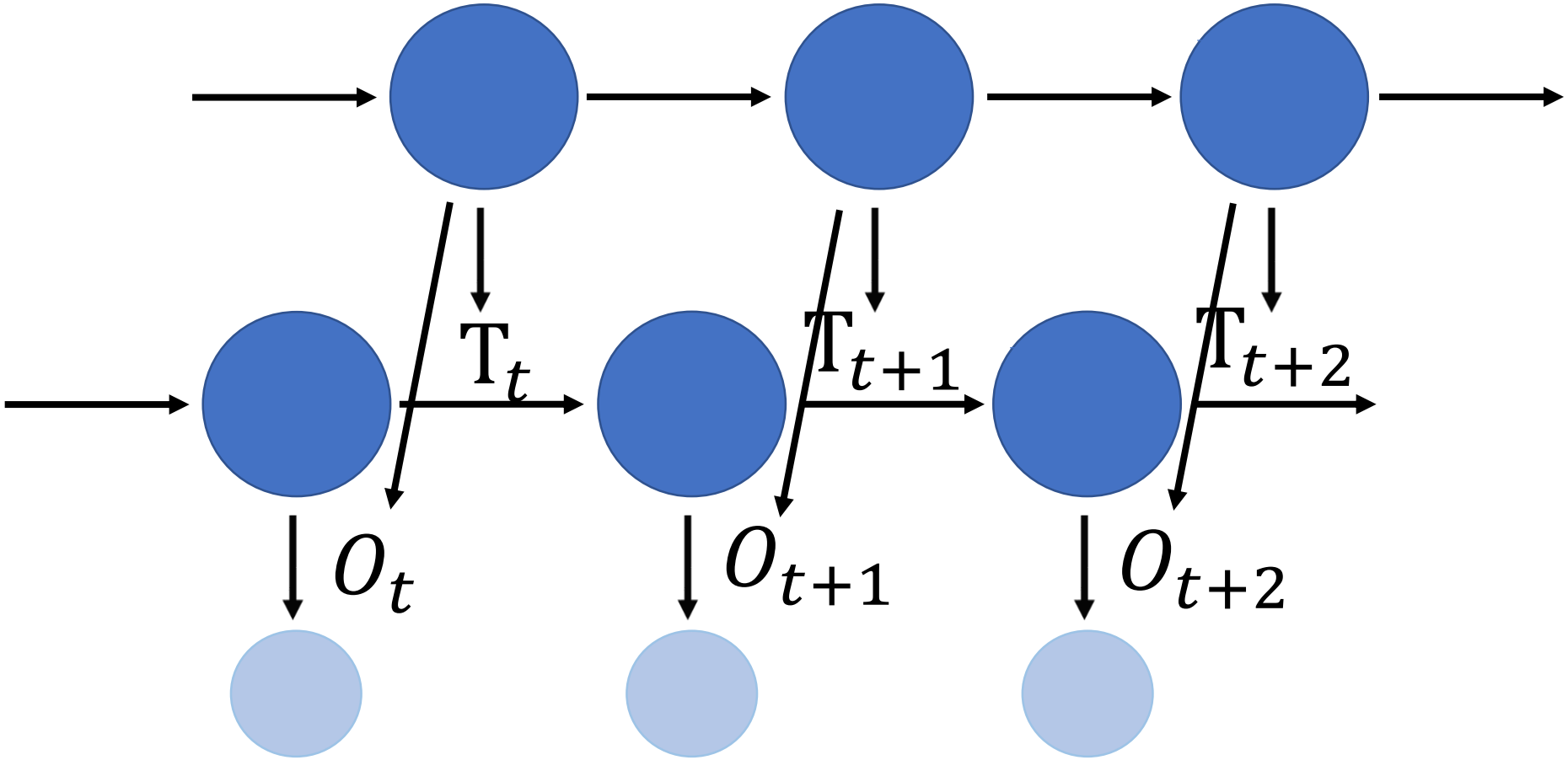
A few related architectures



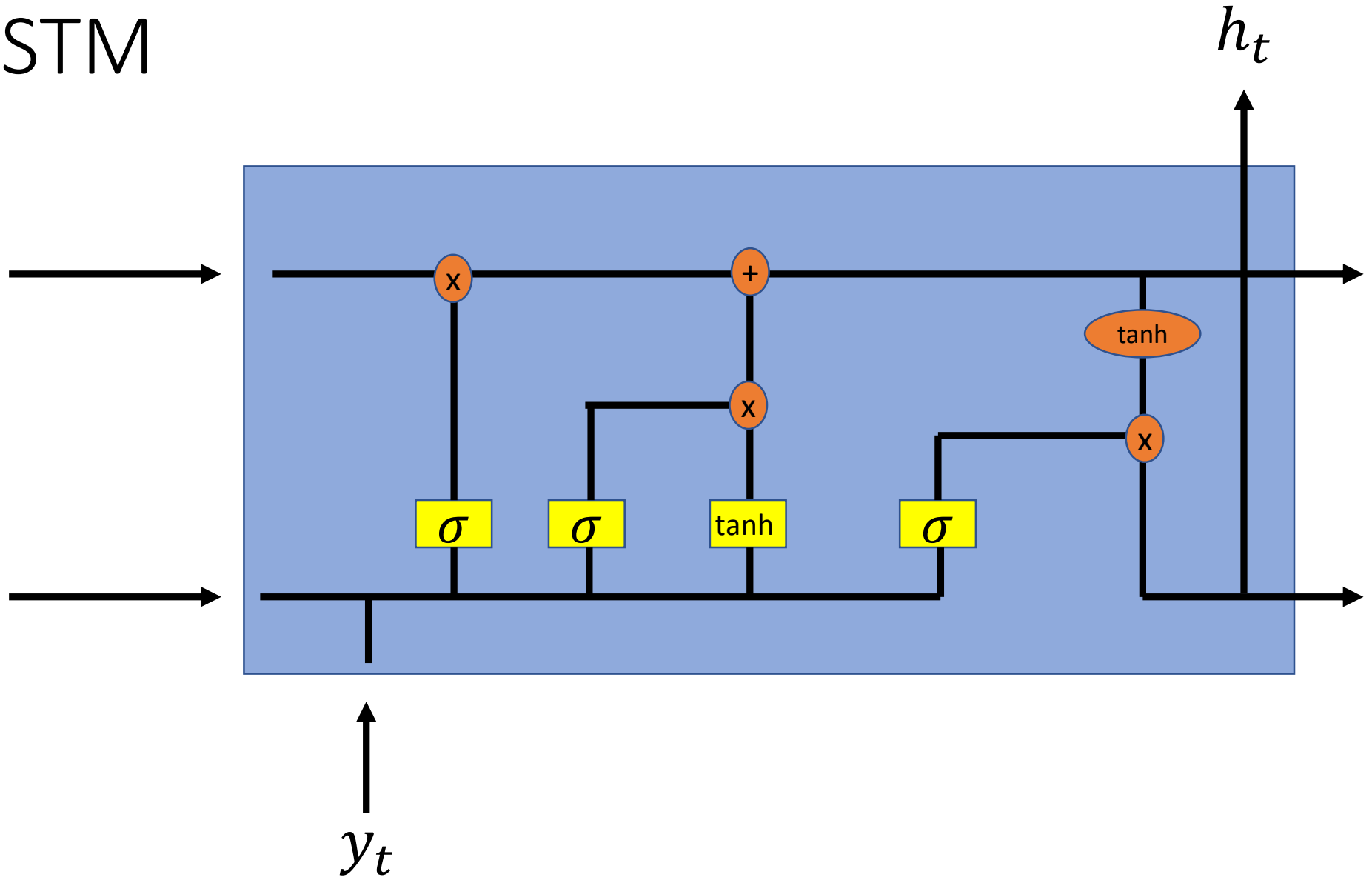
A few related architectures



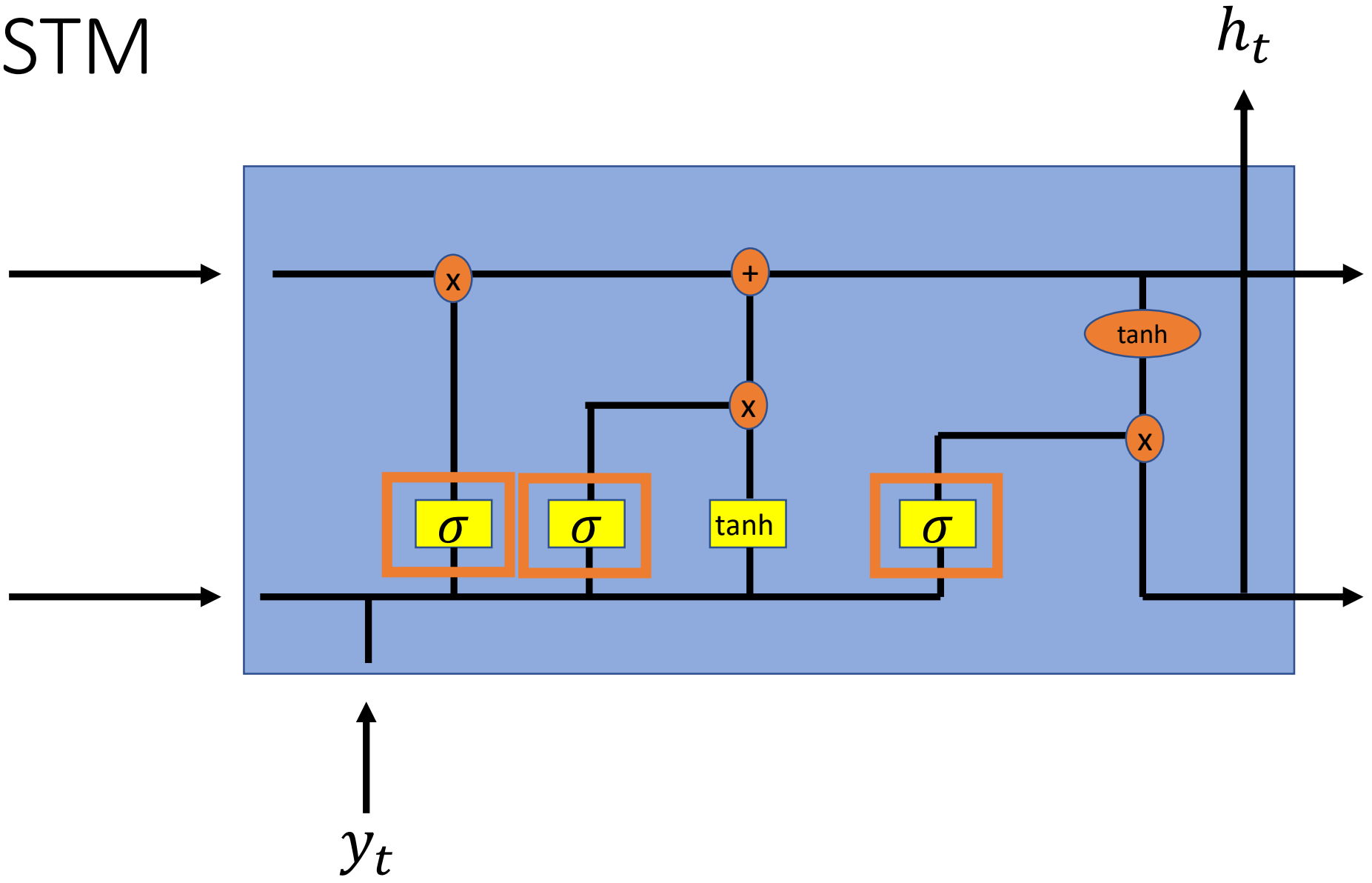
A few related architectures



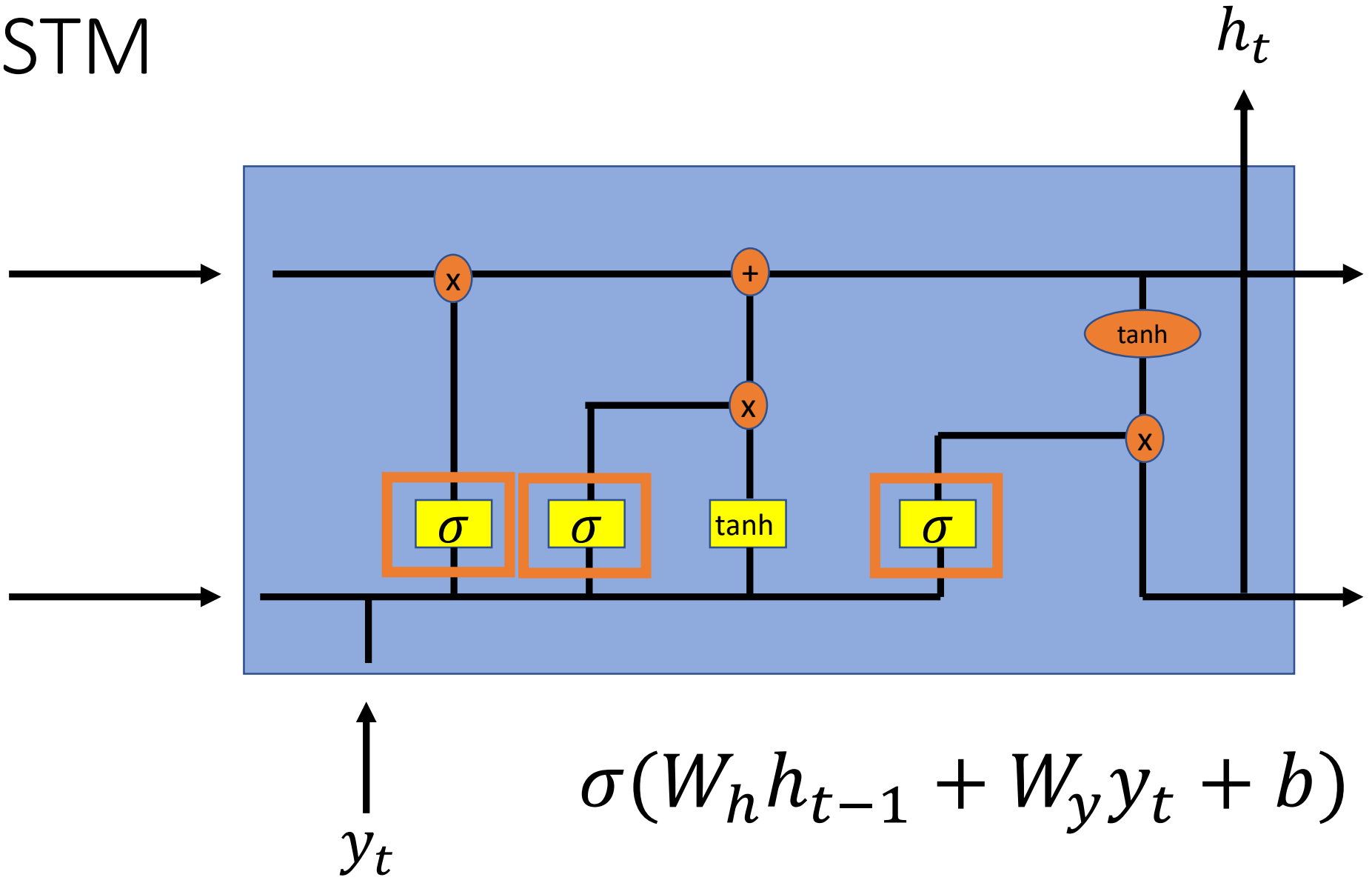
LSTM



LSTM

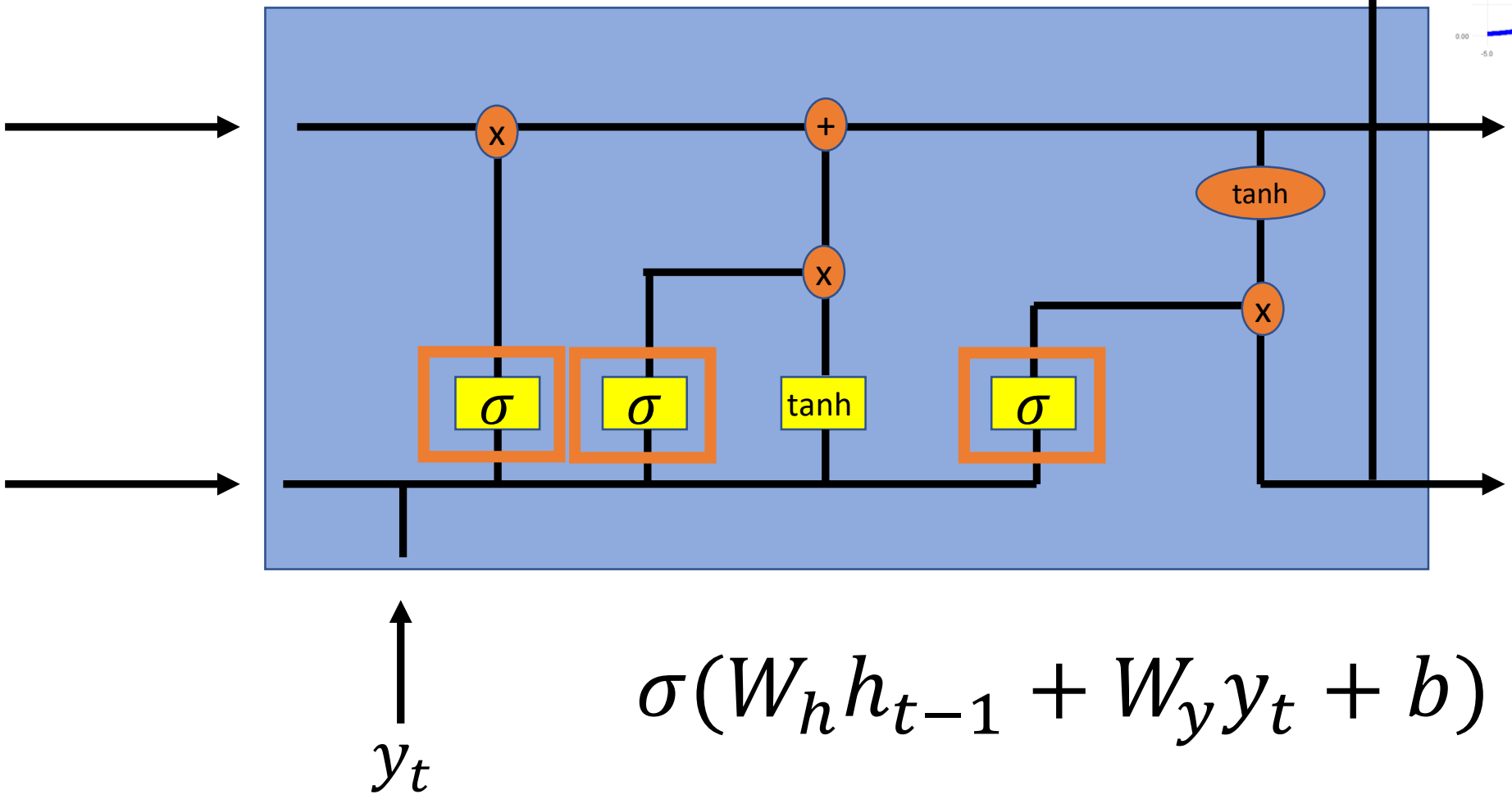


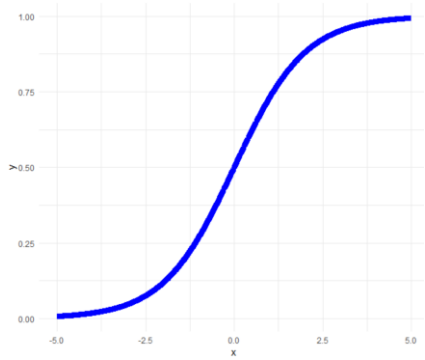
LSTM



$$\sigma(W_h h_{t-1} + W_y y_t + b)$$

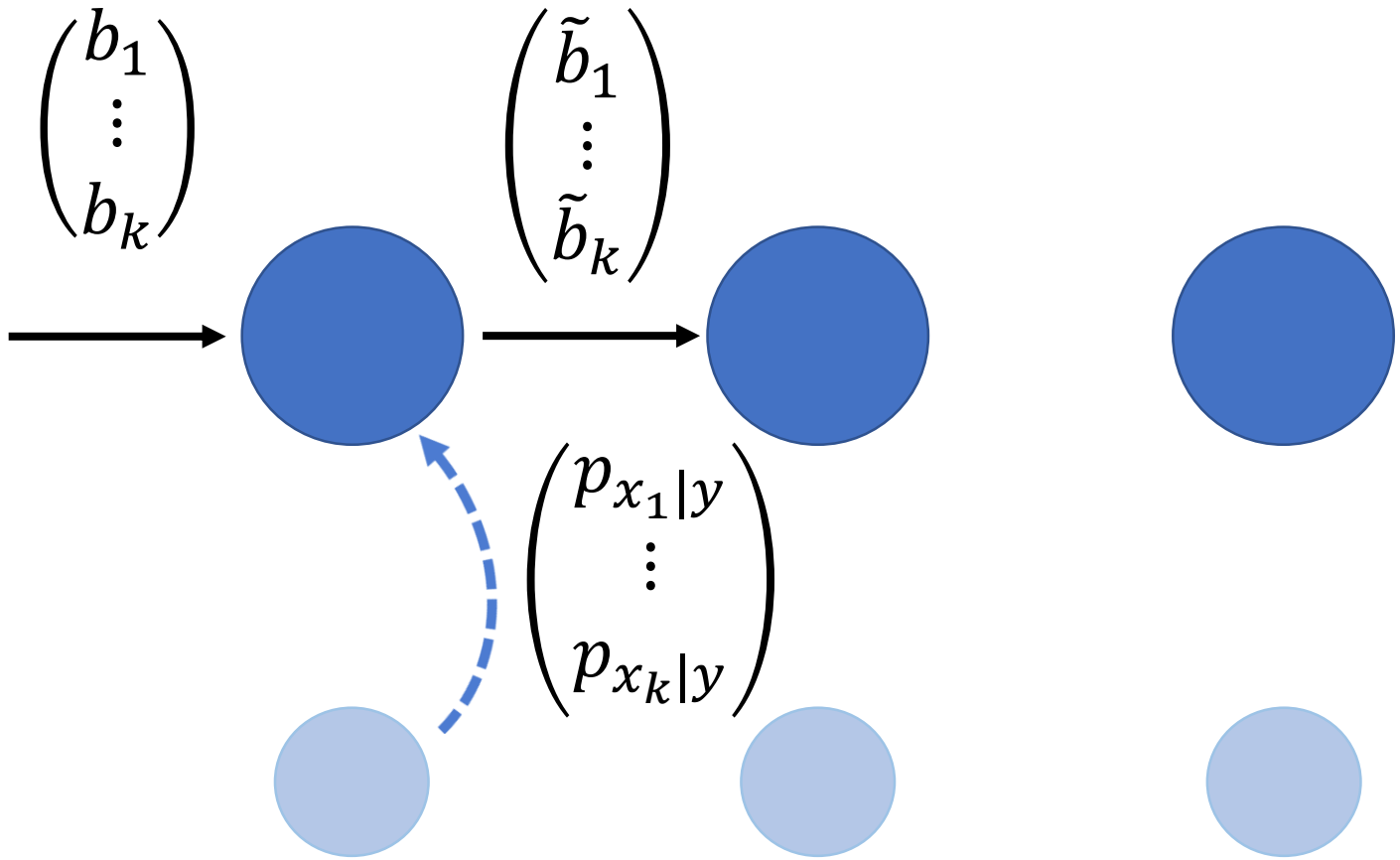
LSTM



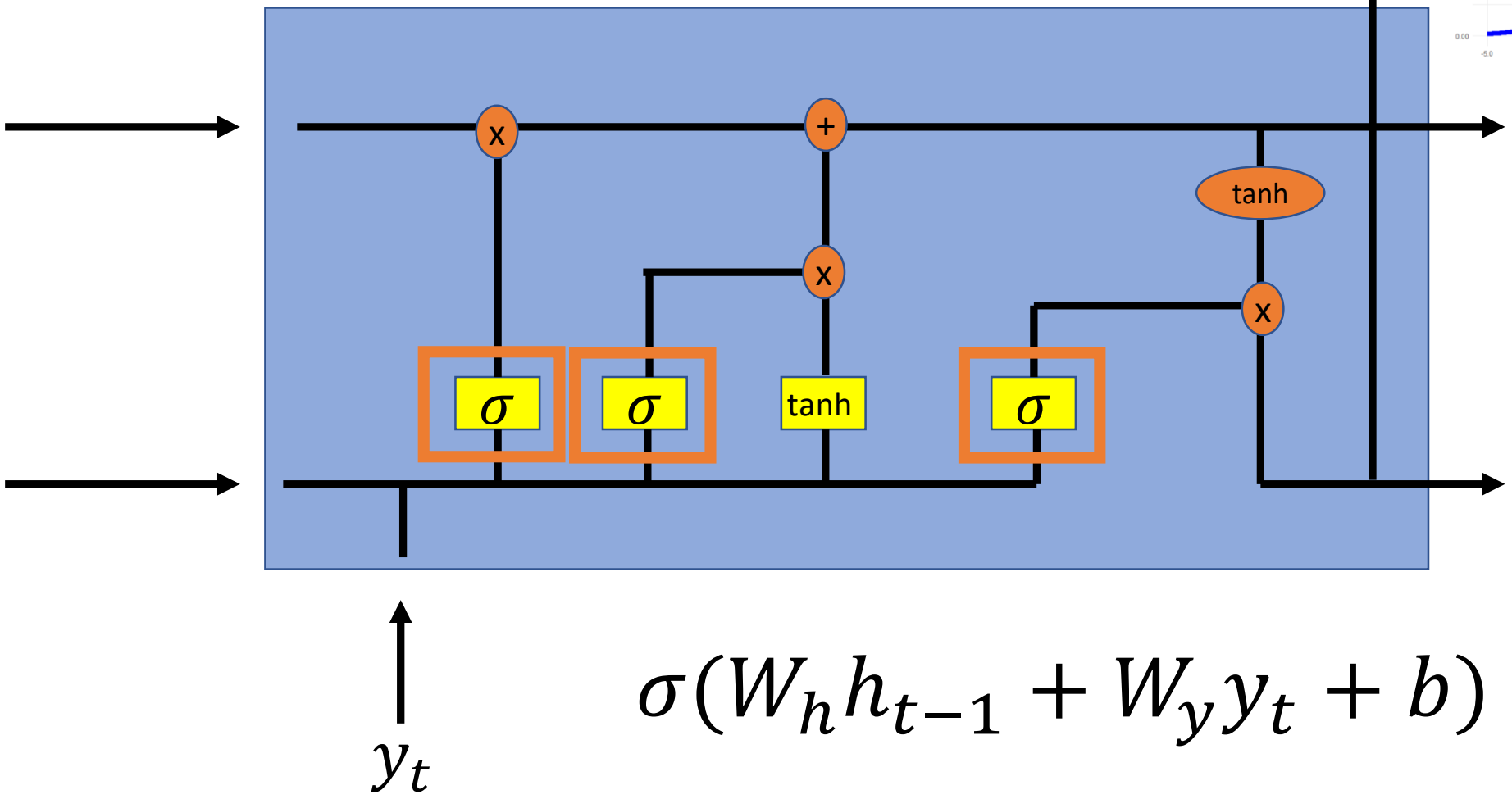


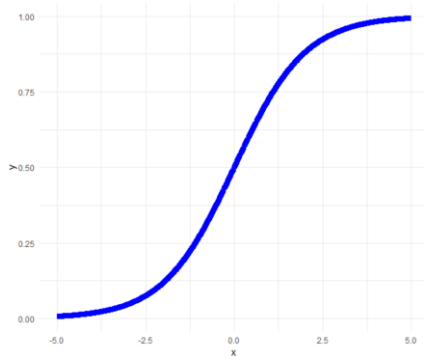
$$\sigma(W_h h_{t-1} + W_y y_t + b)$$

Hidden Markov Models-reminder



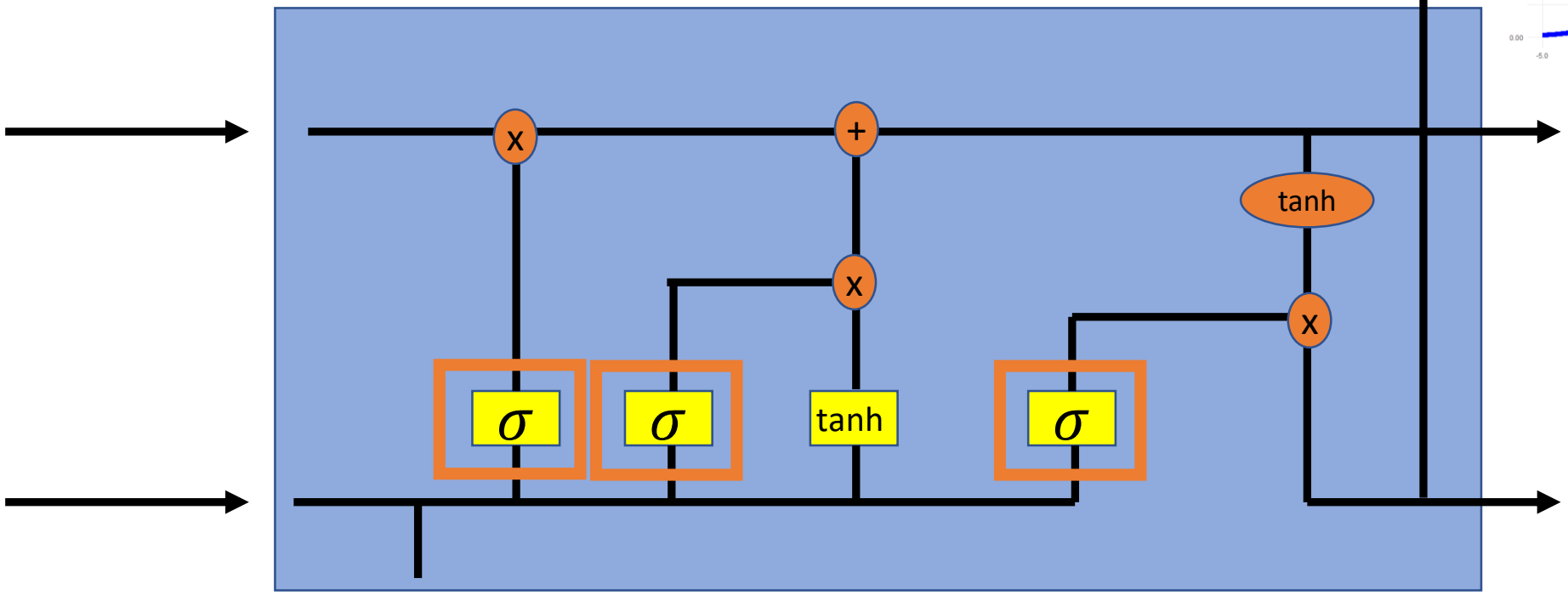
LSTM



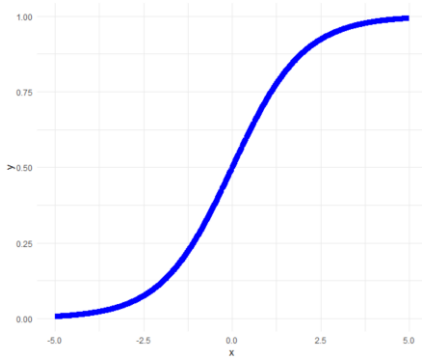


$$\sigma(W_h h_{t-1} + W_y y_t + b)$$

LSTM



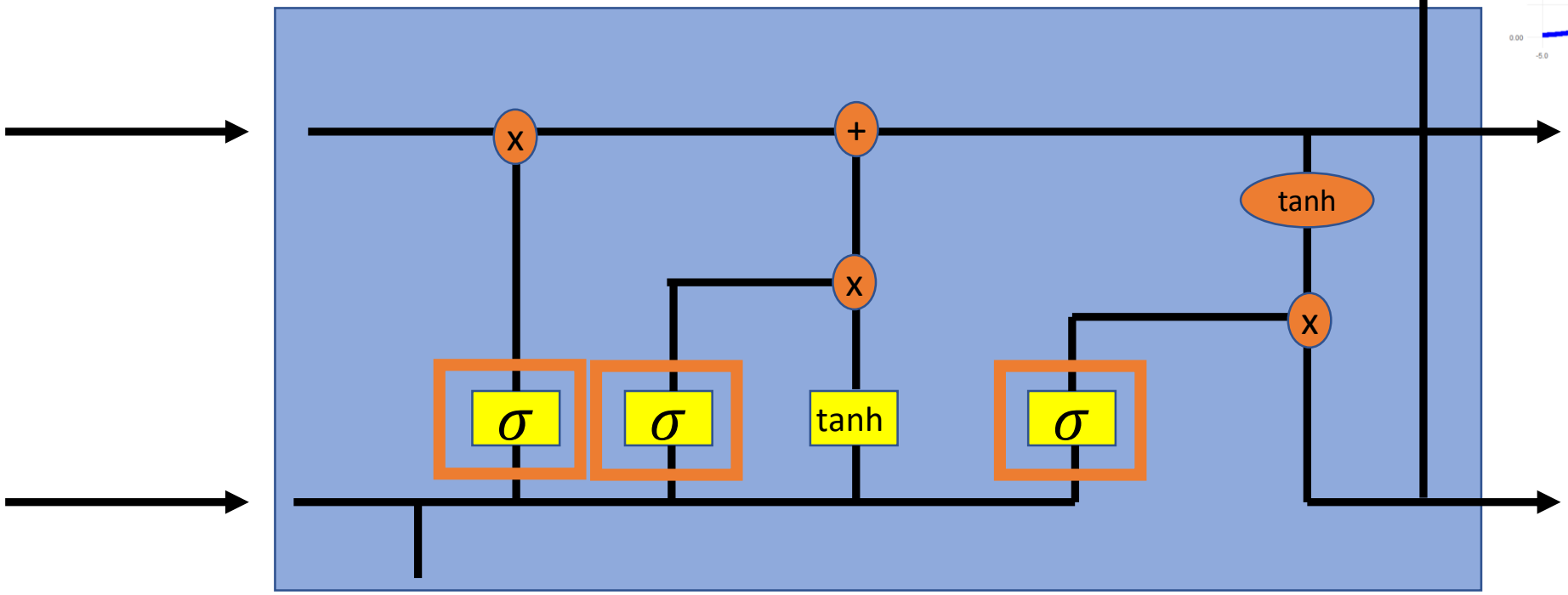
h_t



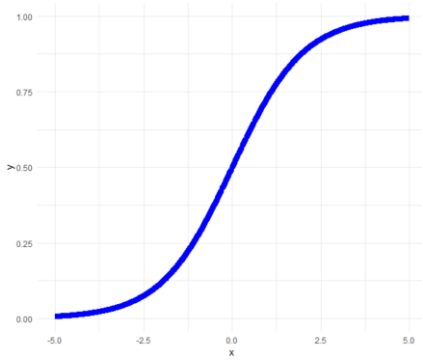
y_t

$$\sigma(\underbrace{W_h h_{t-1}}_{\text{prior (T)}} + W_y y_t + b)$$

LSTM

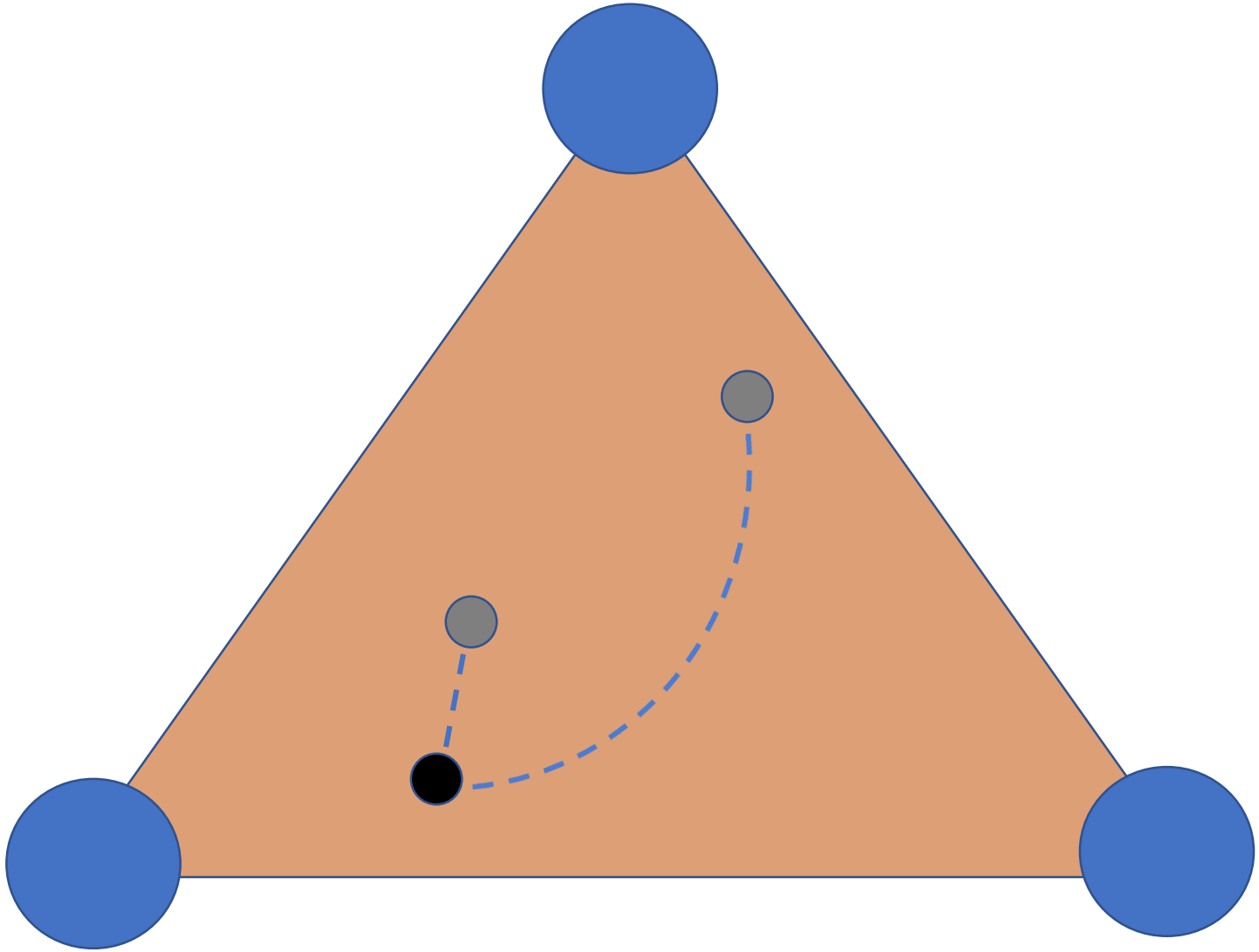


h_t

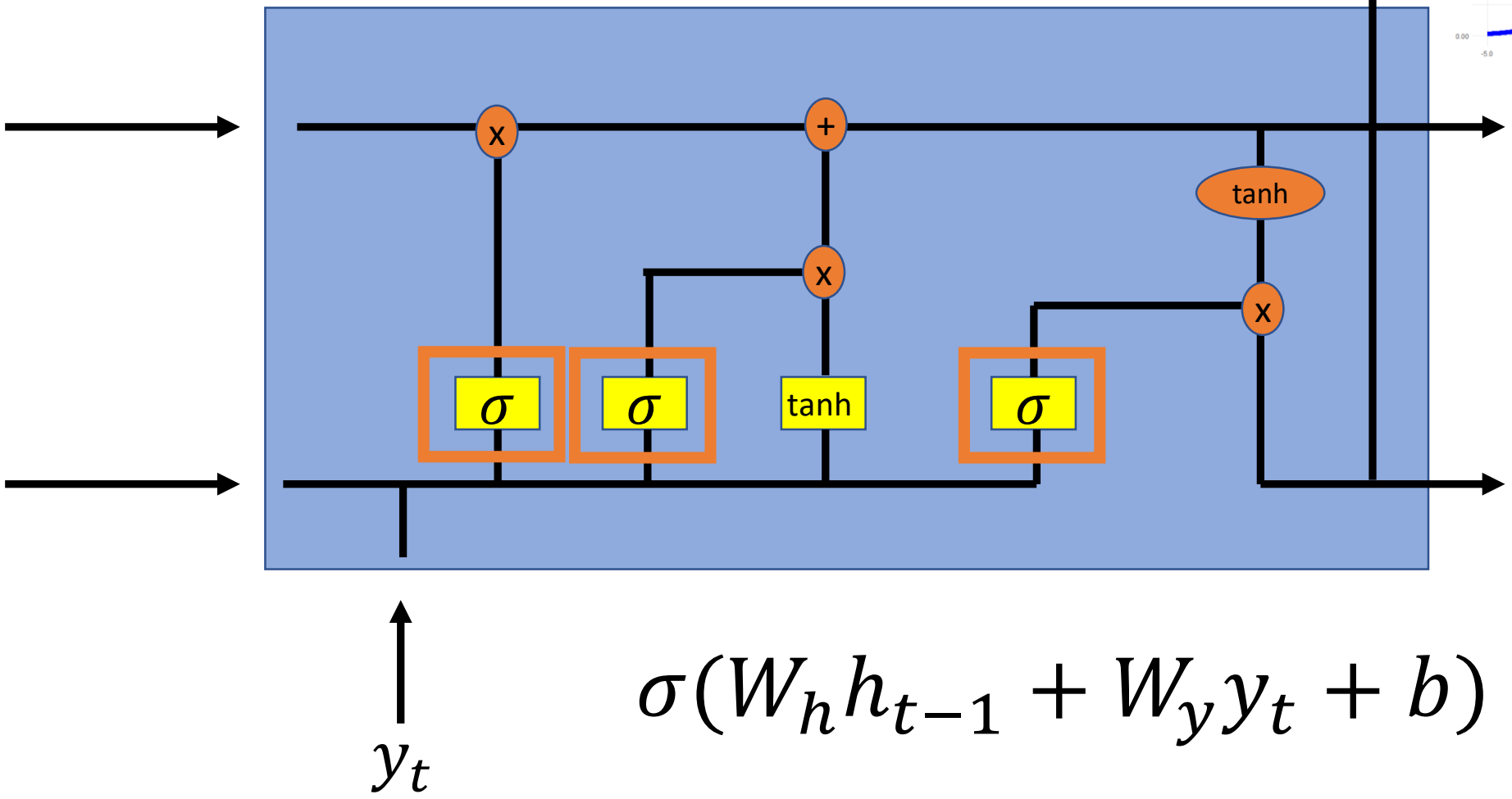


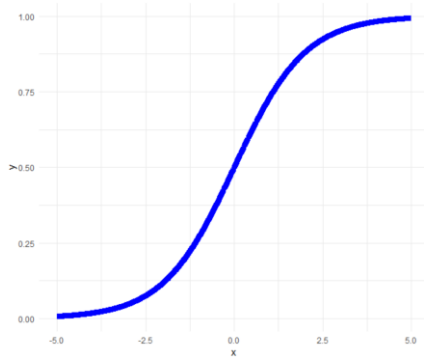
y_t

$$\sigma(\underbrace{W_h h_{t-1} + W_y y_t}_{\text{posterior (T and O)}} + b)$$



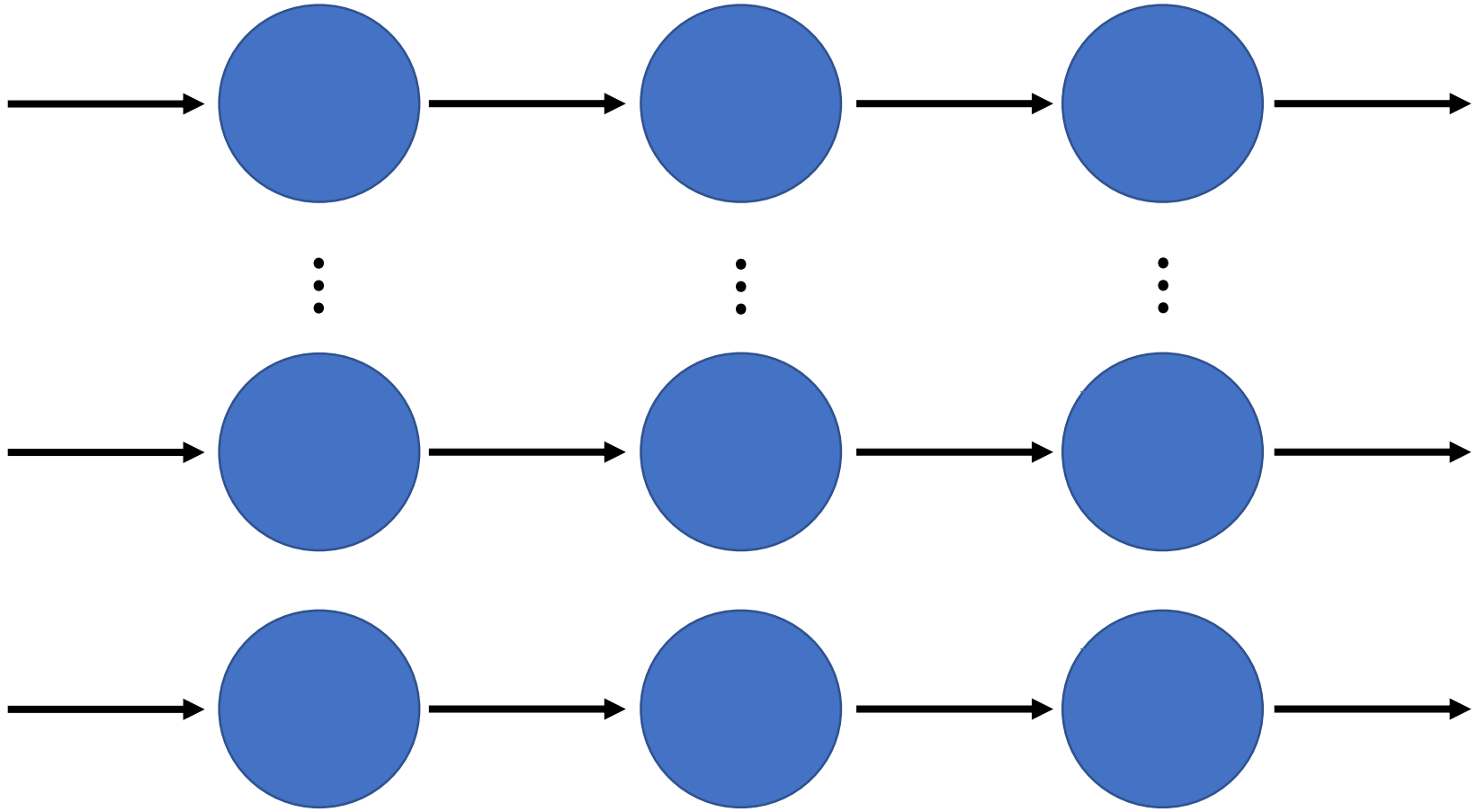
LSTM



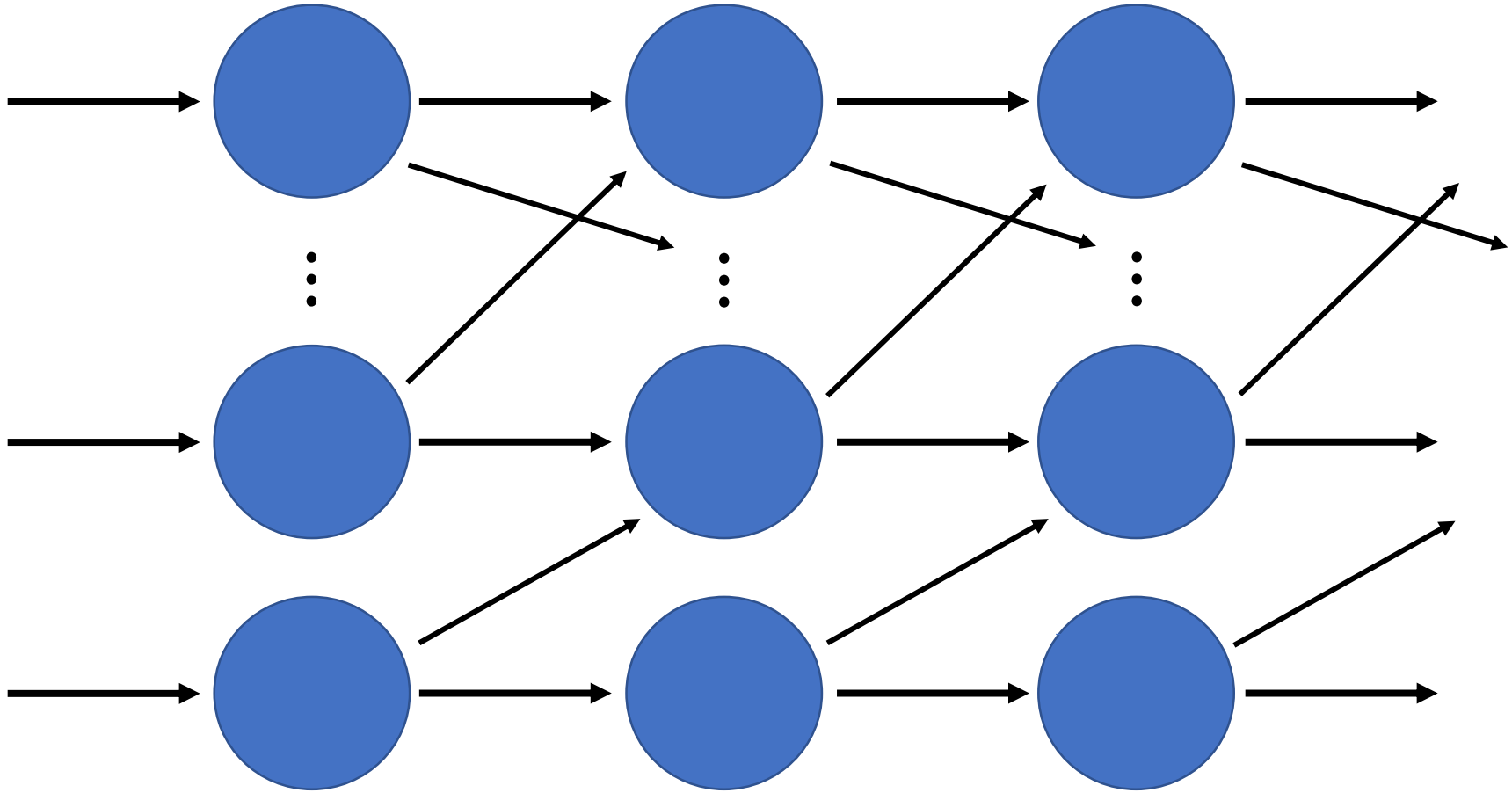


$$\sigma(W_h h_{t-1} + W_y y_t + b)$$

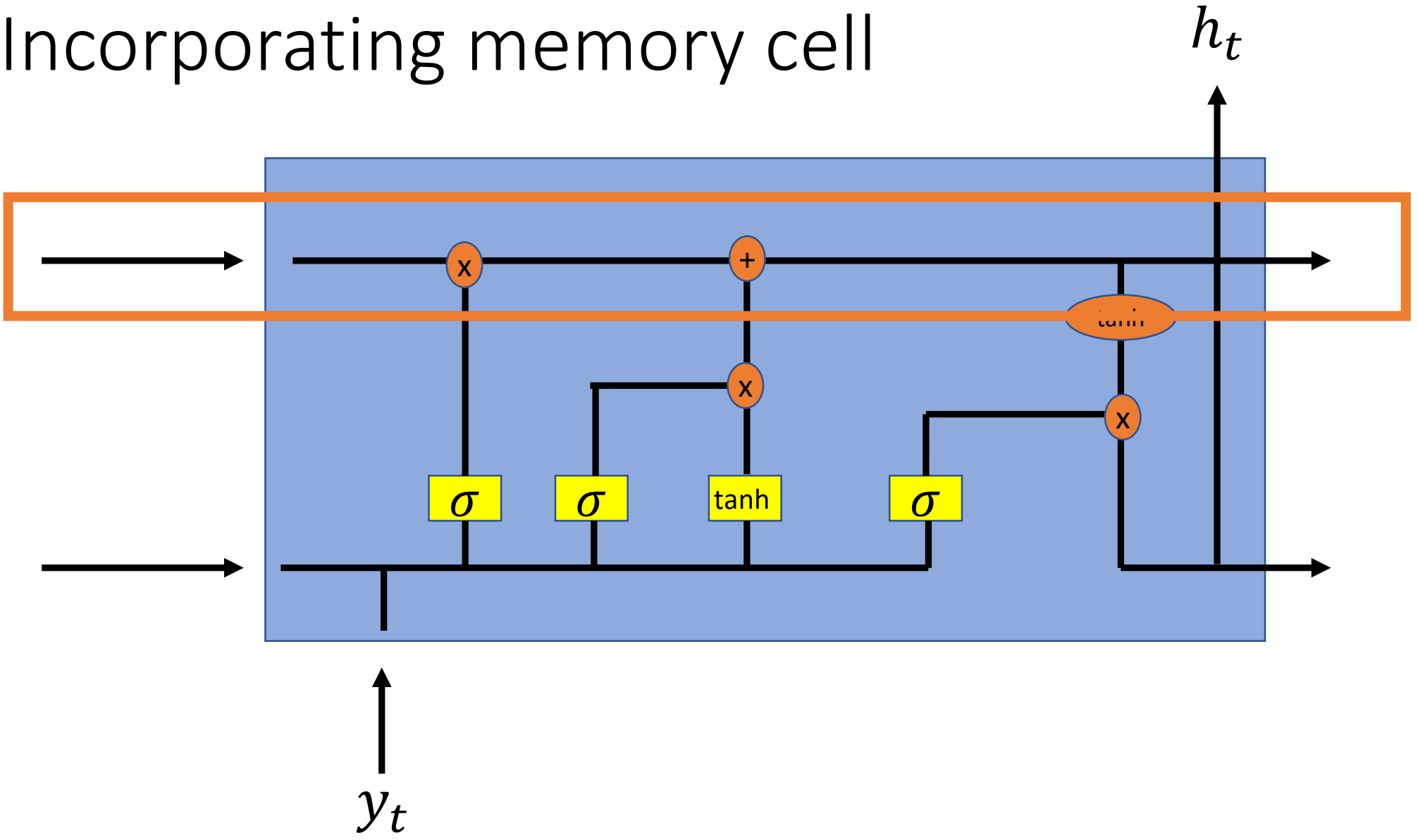
Hidden states, 2^k states for k hidden nodes



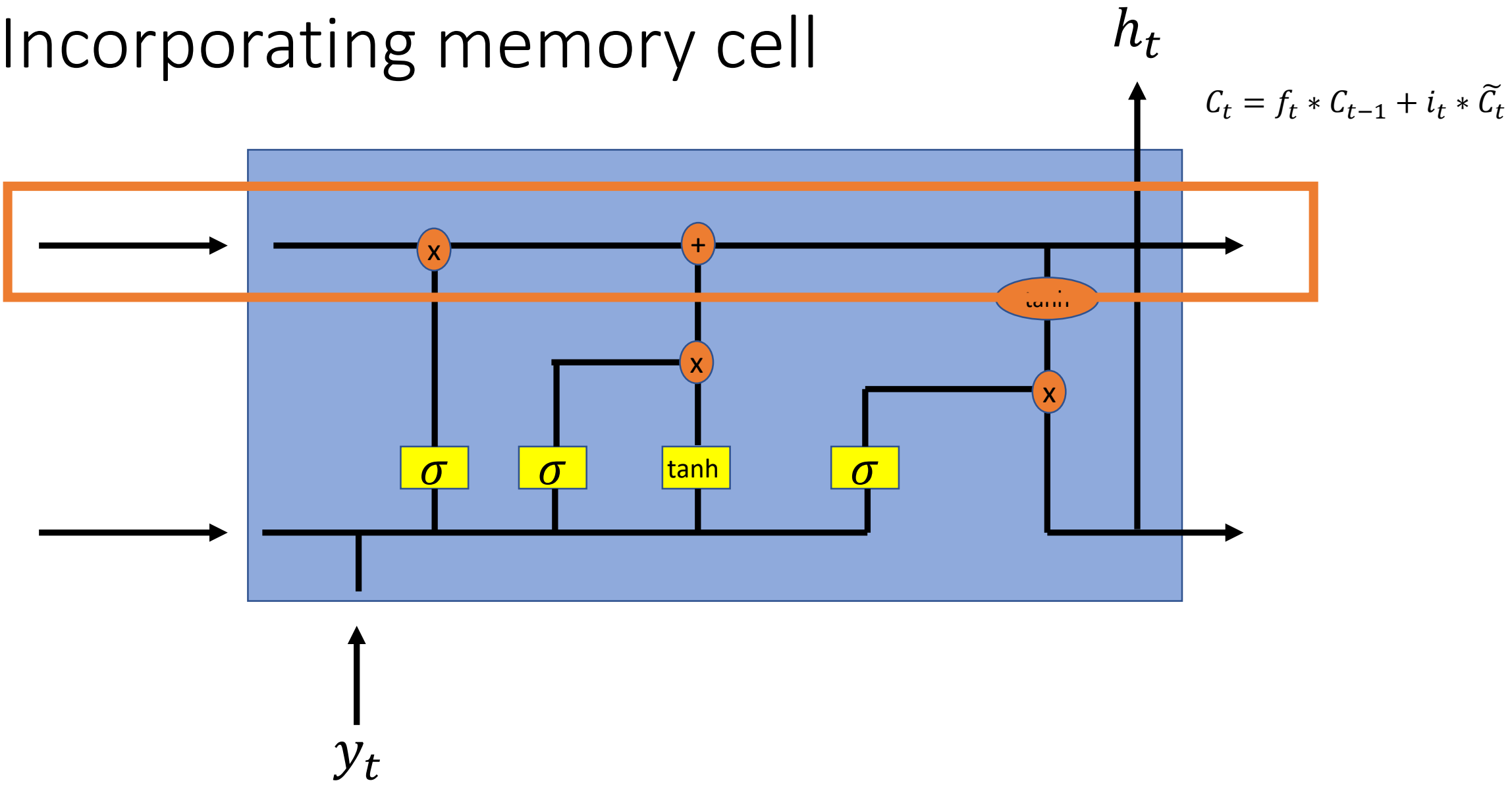
Dependencies specified by weights from W_h



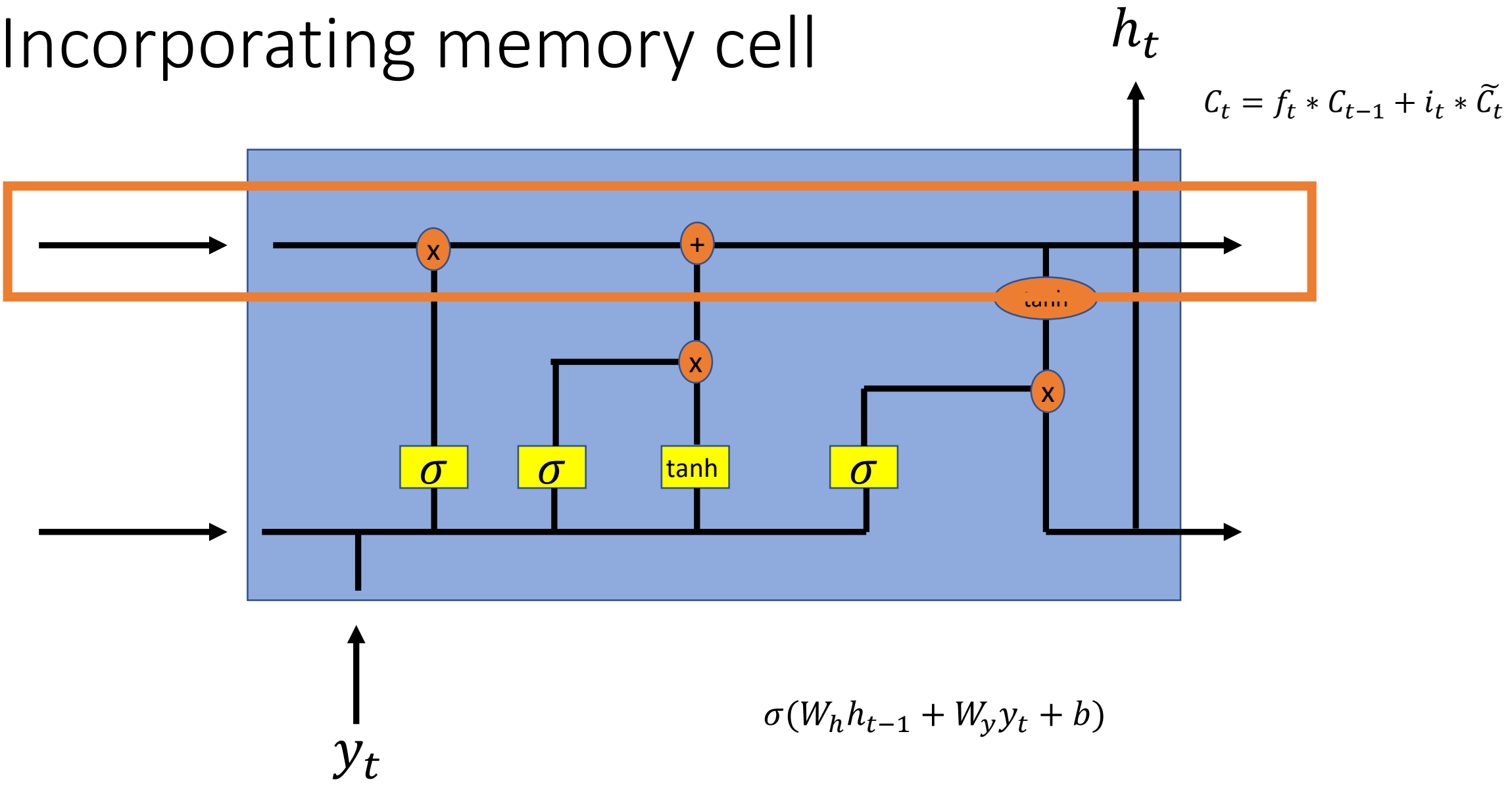
Incorporating memory cell

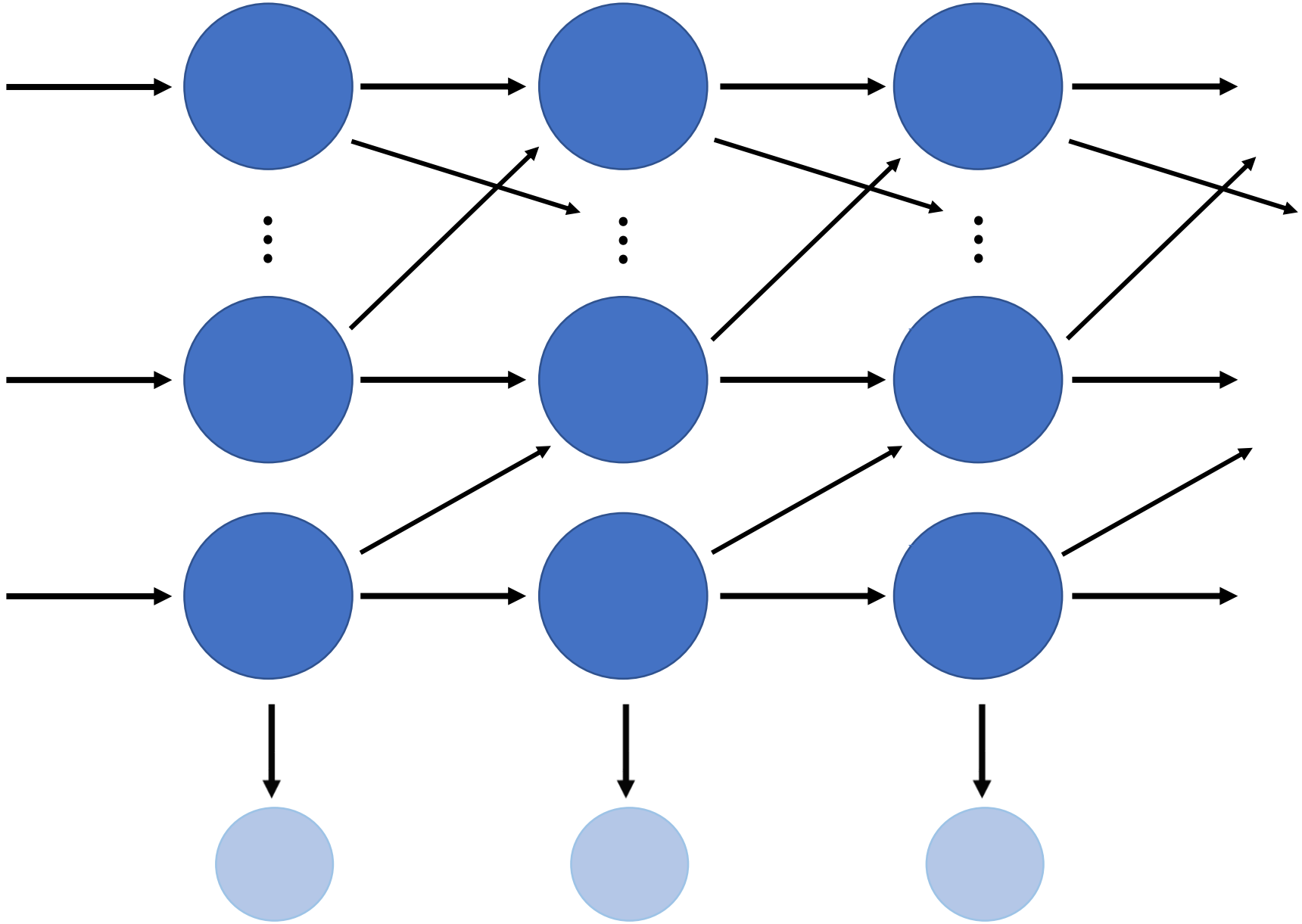


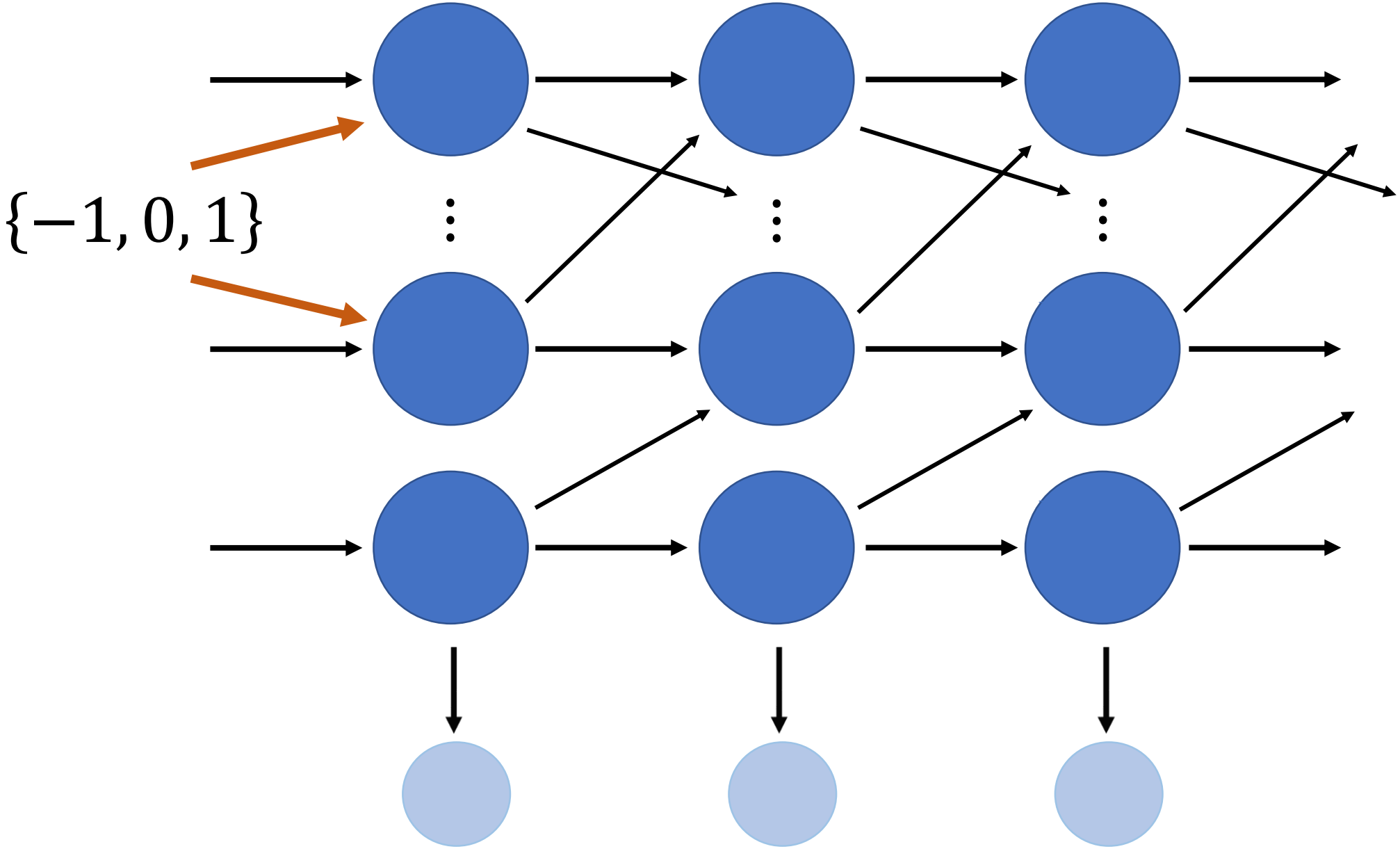
Incorporating memory cell

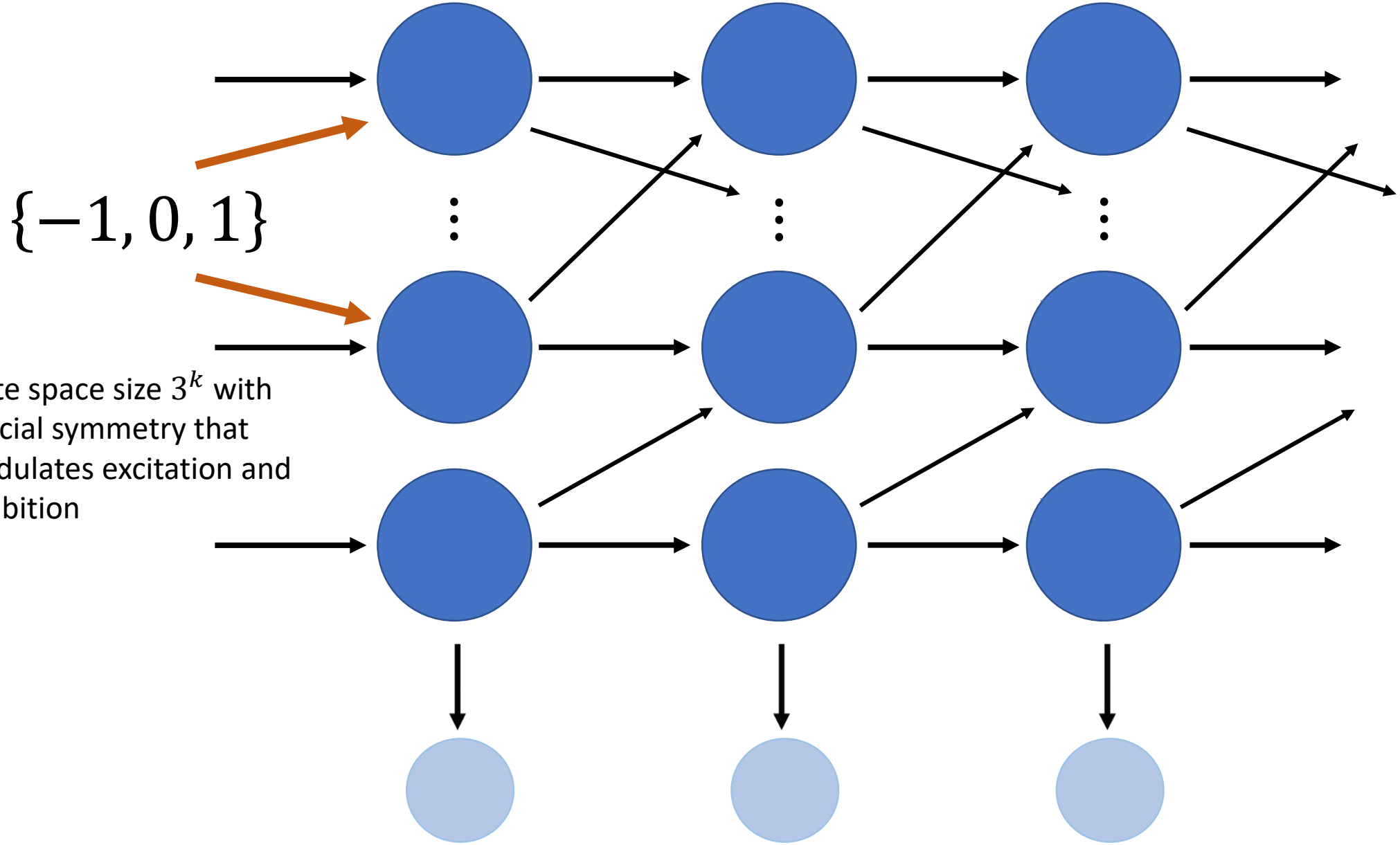


Incorporating memory cell





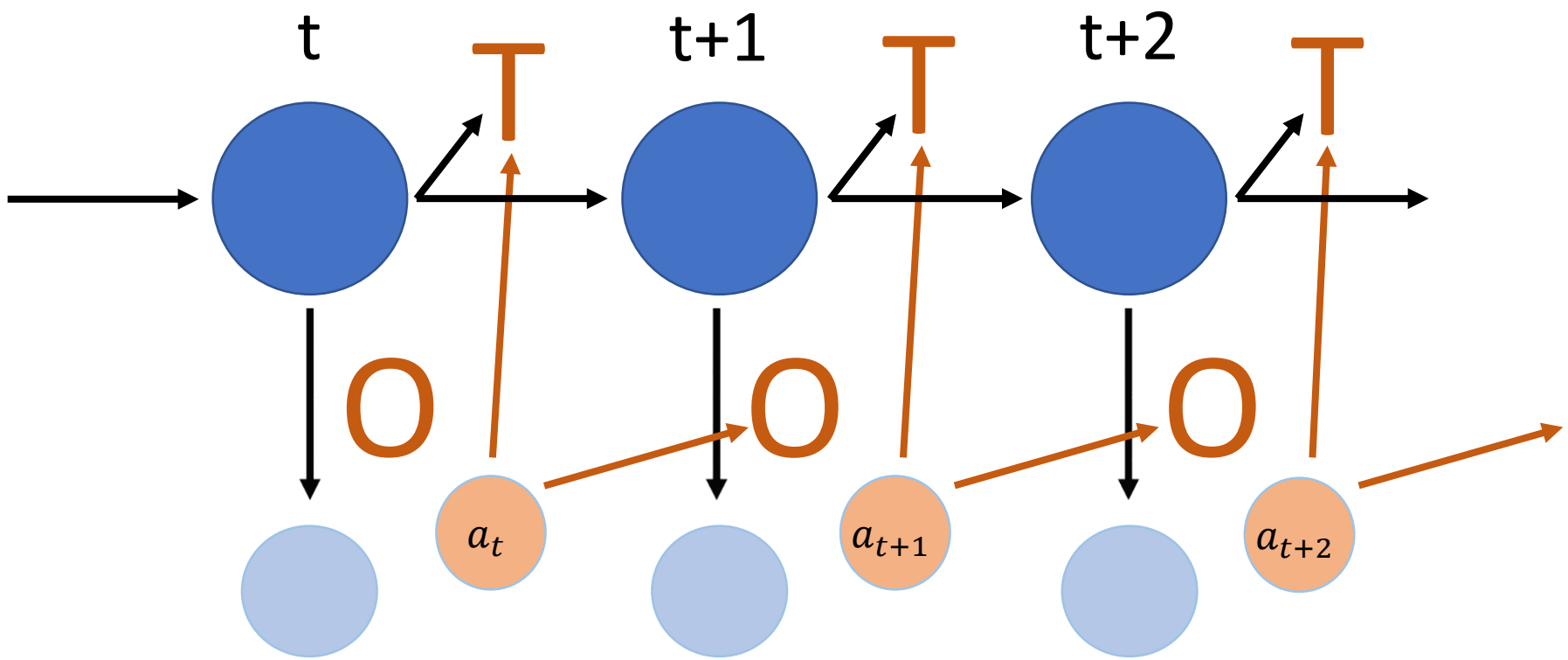

$\{-1, 0, 1\}$



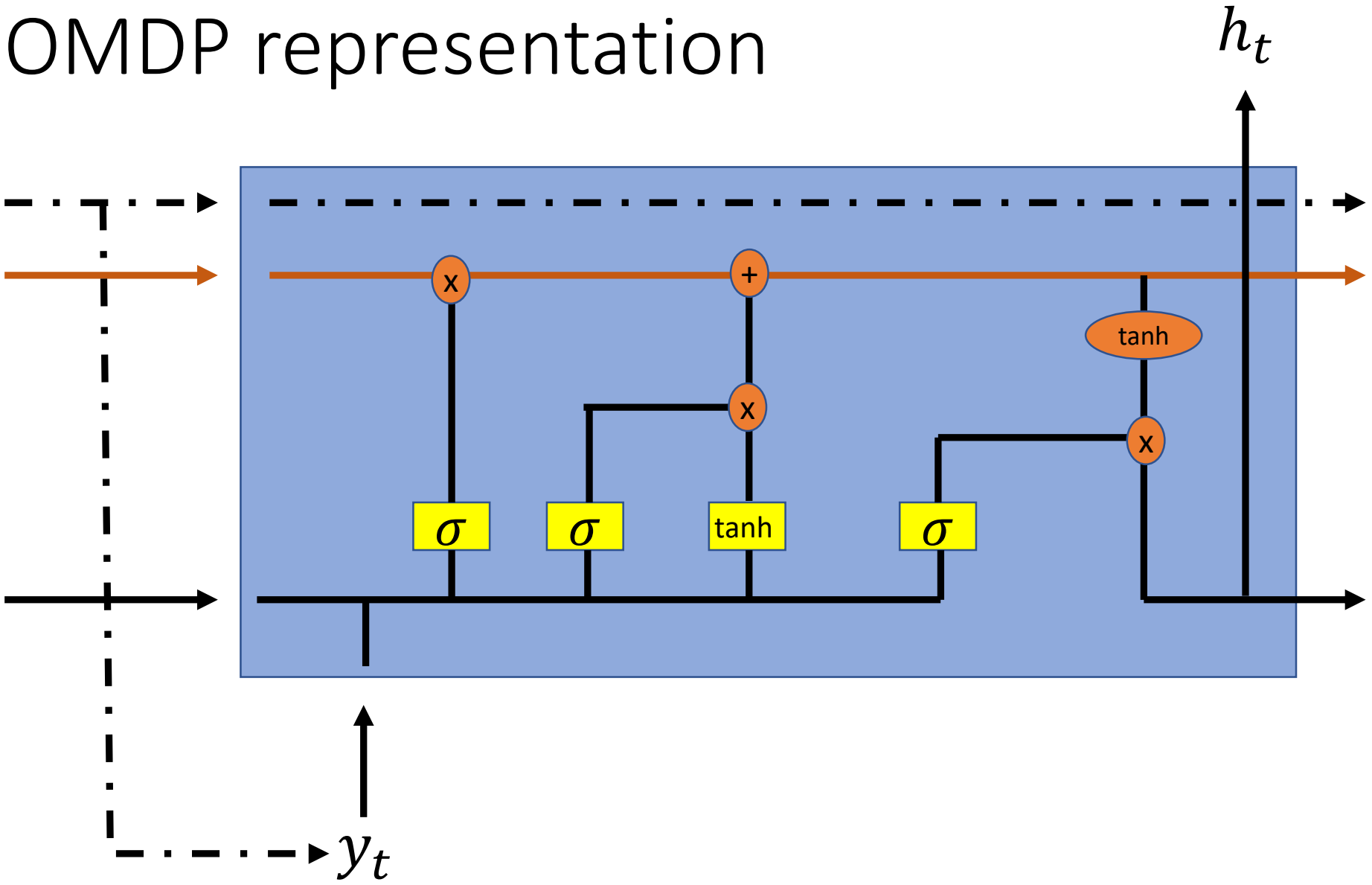
State space size 3^k with special symmetry that modulates excitation and inhibition

Partially Observable Markov Decision Process

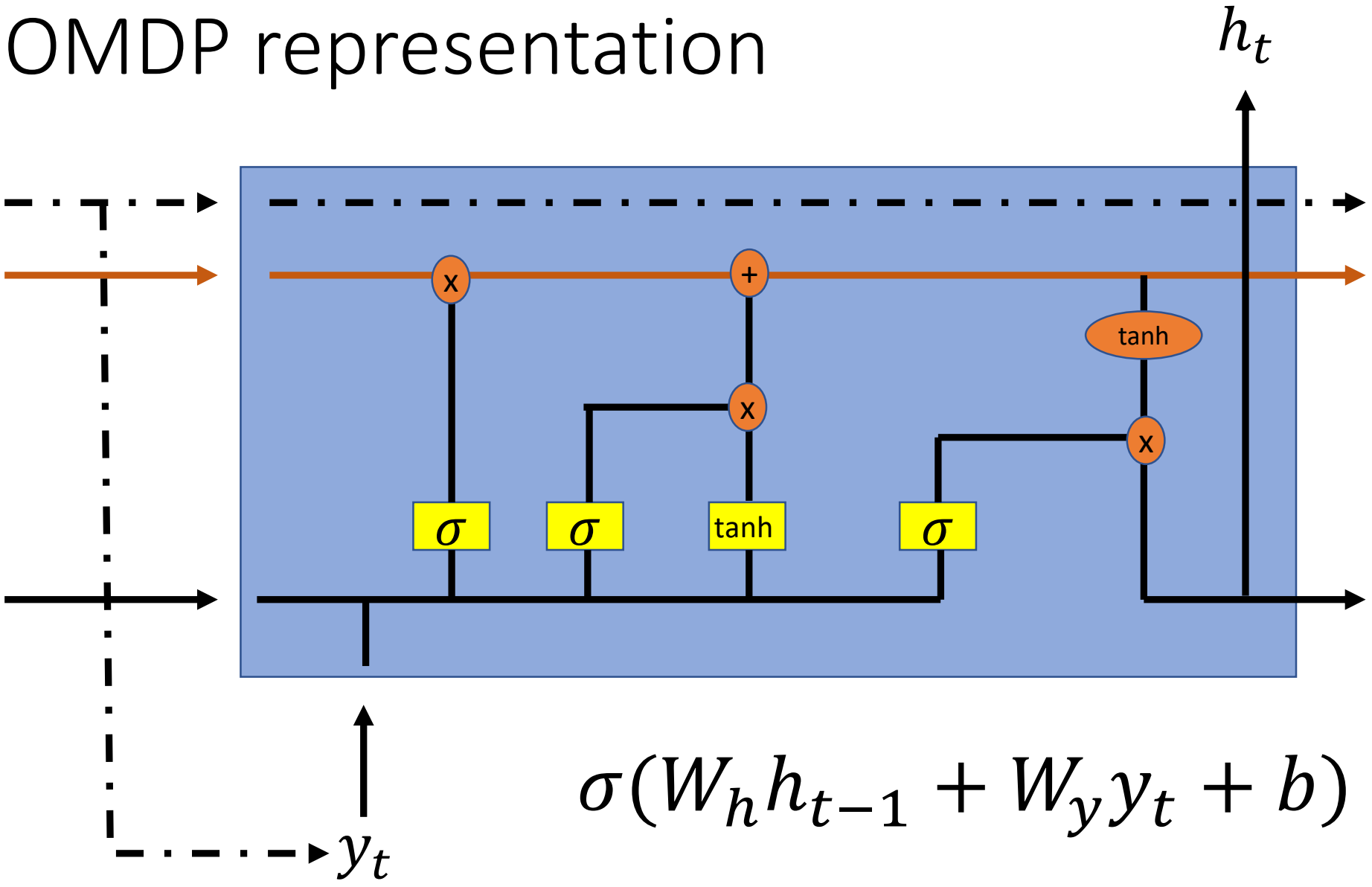
*relaxed visualization



POMDP representation

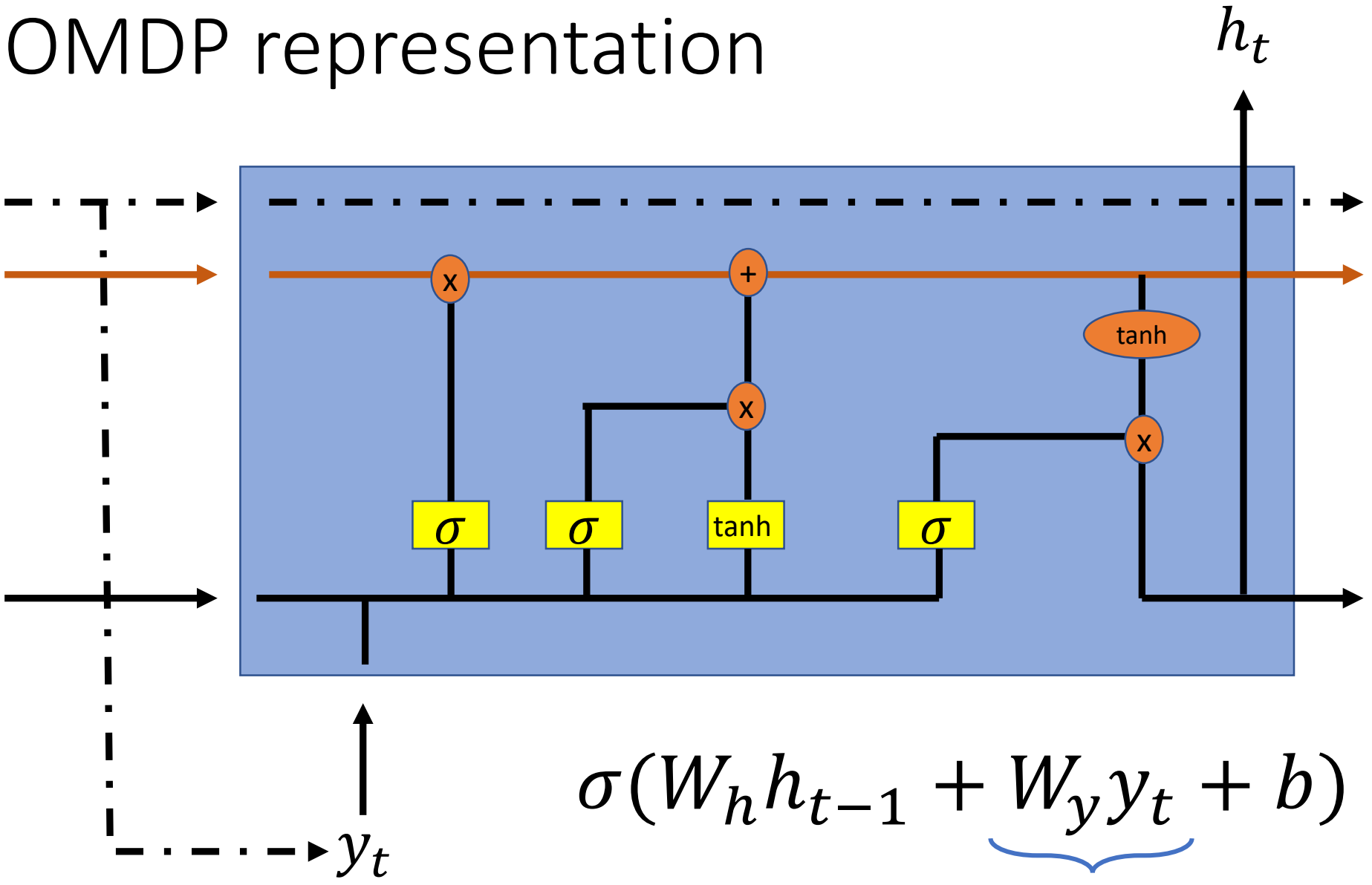


POMDP representation



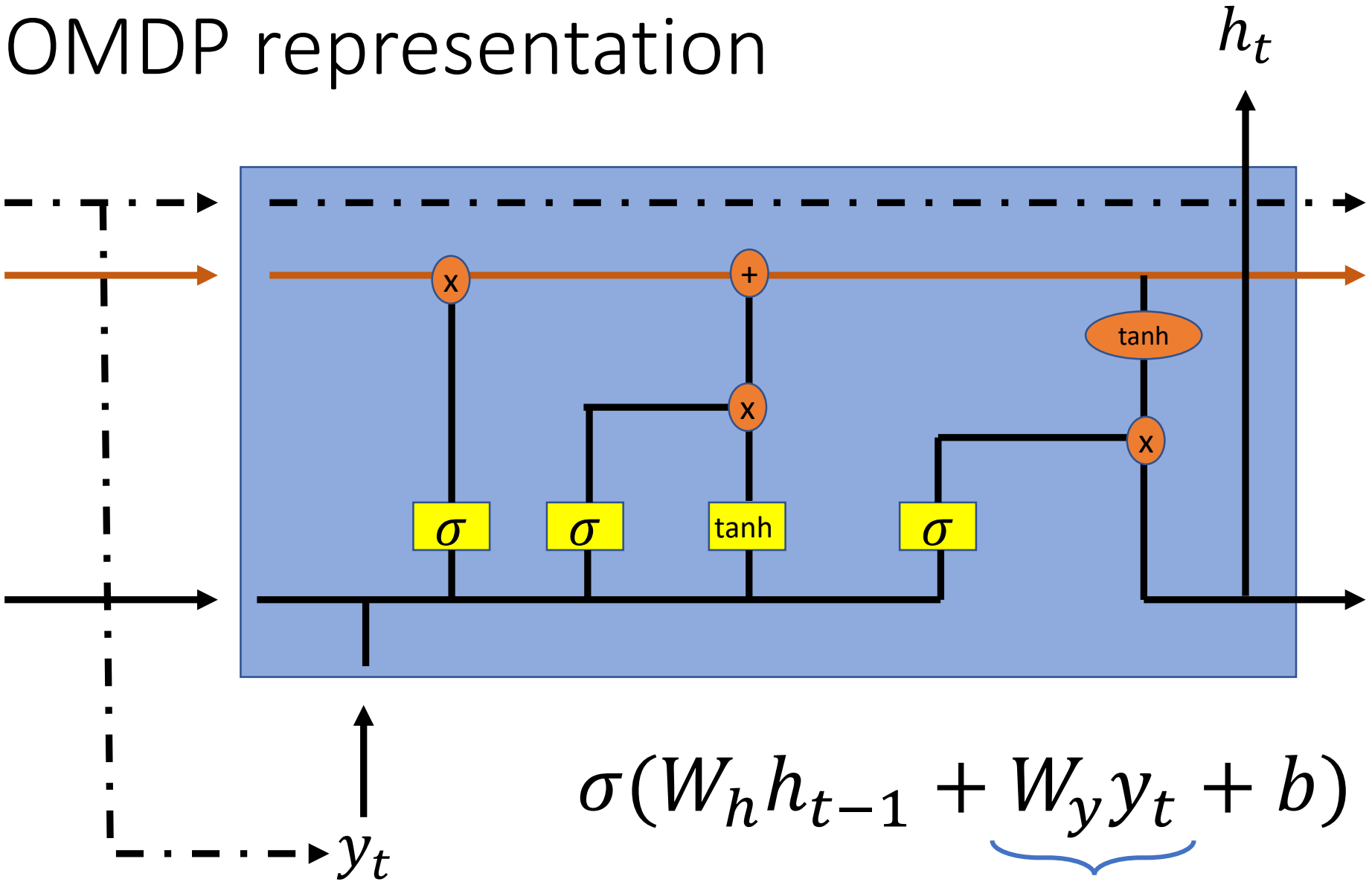
$$\sigma(W_h h_{t-1} + W_y y_t + b)$$

POMDP representation



$$\sigma(W_h h_{t-1} + \underbrace{W_y y_t}_{\text{policy}} + b)$$

POMDP representation



Recall that our goal here is not to learn a POMDP or to approximate a POMDP using LSTMs, rather to wrap LSTMs in models for which we have a more robust understanding.

$$\sigma(W_h h_{t-1} + \underbrace{W_y y_t}_{\text{policy}} + b)$$

Thanks!