neural networks

Inteligencia artificial y control

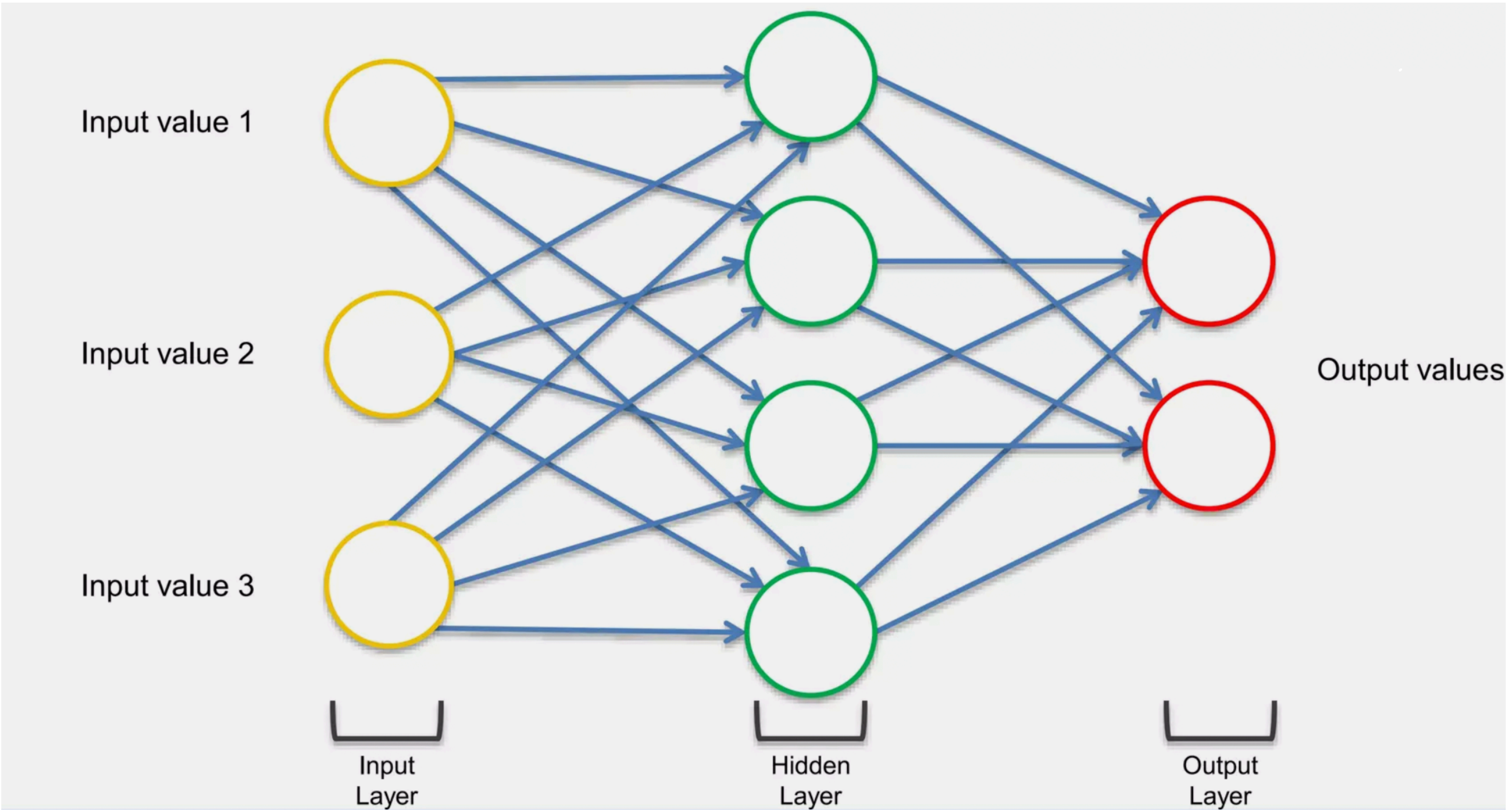
Gerardo Marx Chávez-Campos

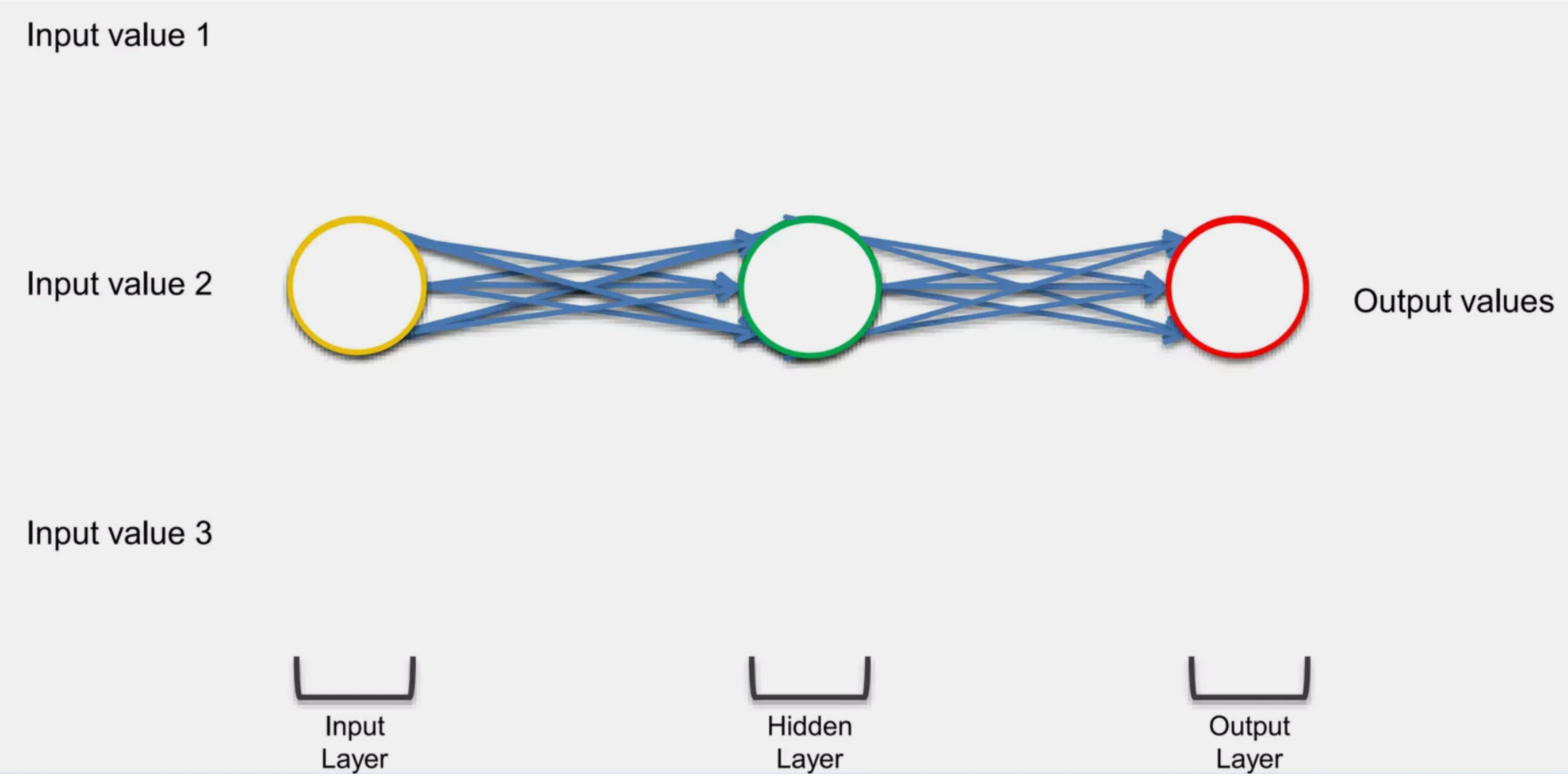
Part 2
Recurrent Neural Networks

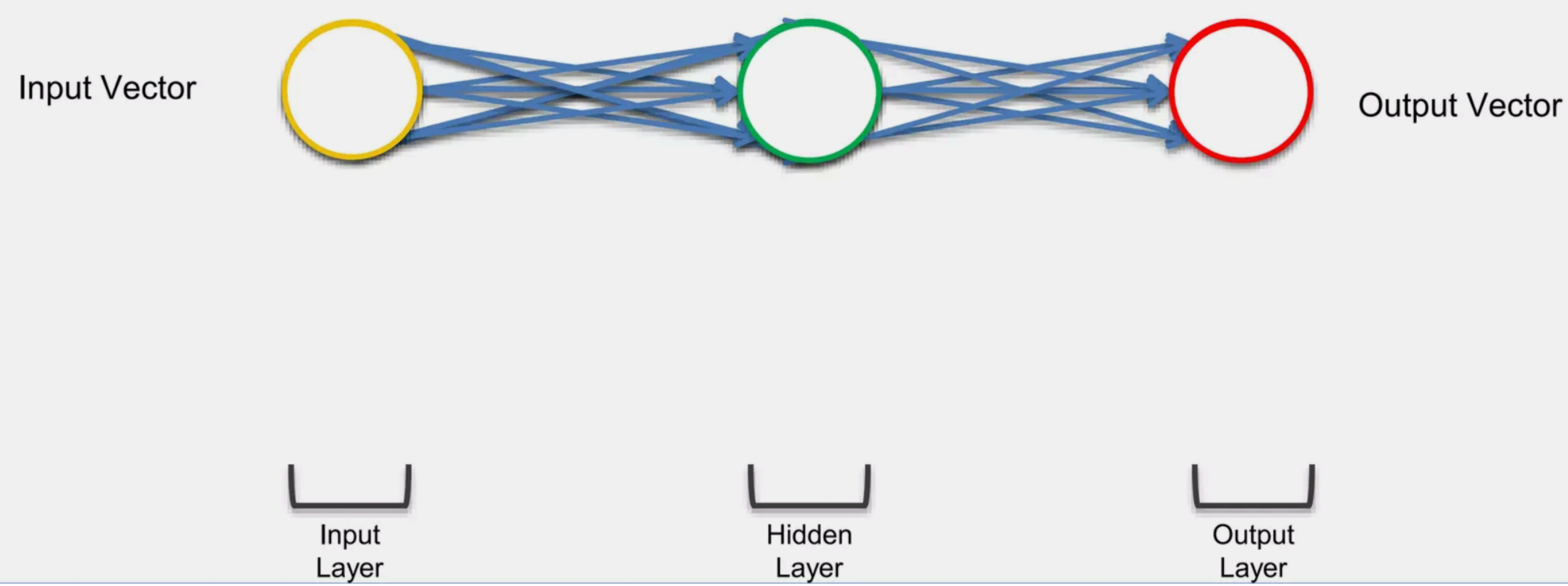
Recurrent Neural Networks (RNN)

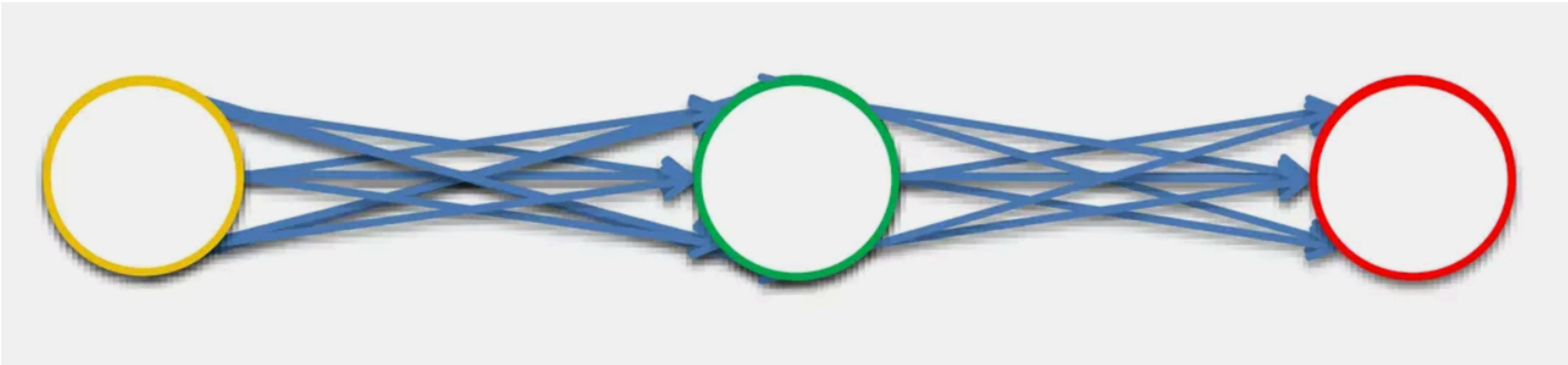
Supervised

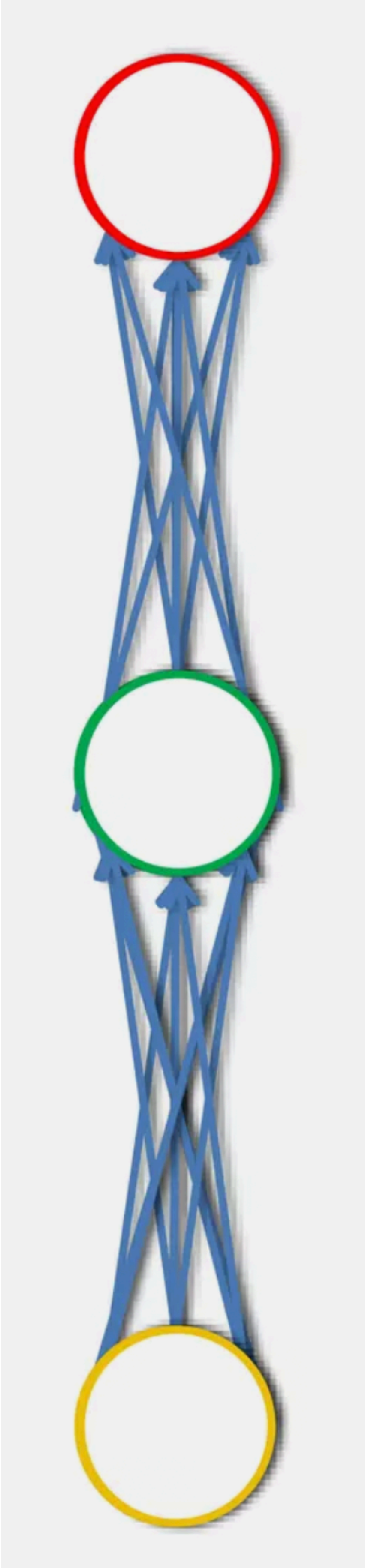
Artificial Neural Networks	Regression and classification
Convolutional Neural Networks	Computer Visión
Recurrent Neural Networks	Time Series Analysis

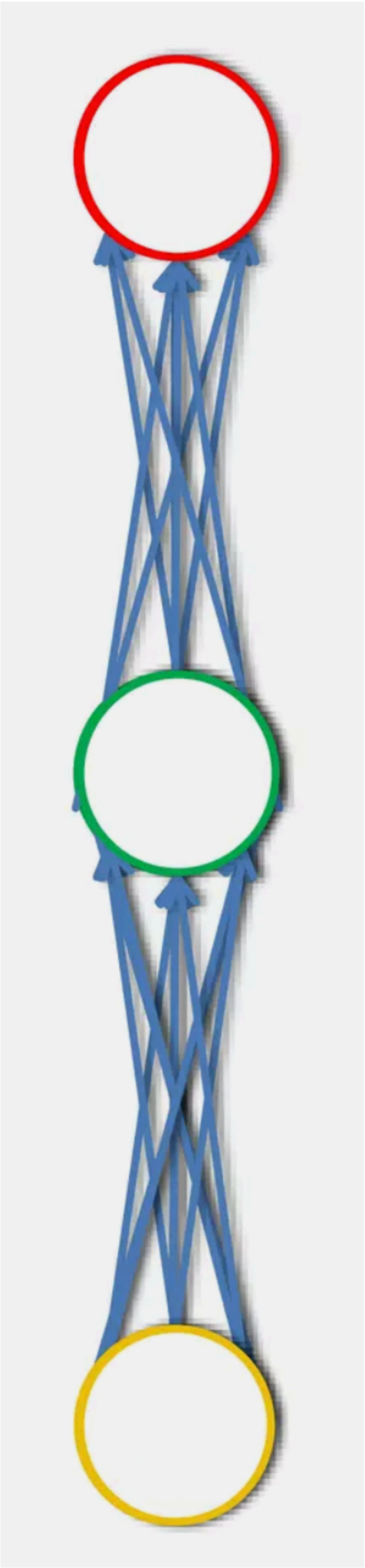

Input value 1
Input value 2
Input value 3
Output values
Input Layer
Hidden Layer
Output Layer

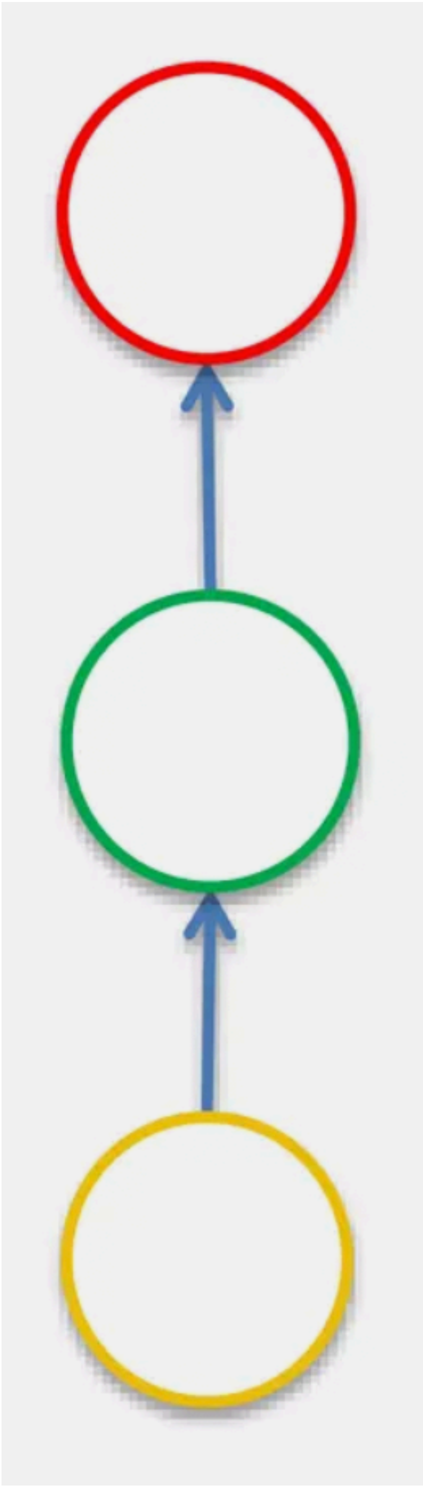

Input value 1
Input value 2
Input value 3
Output values
Input Layer
Hidden Layer
Output Layer

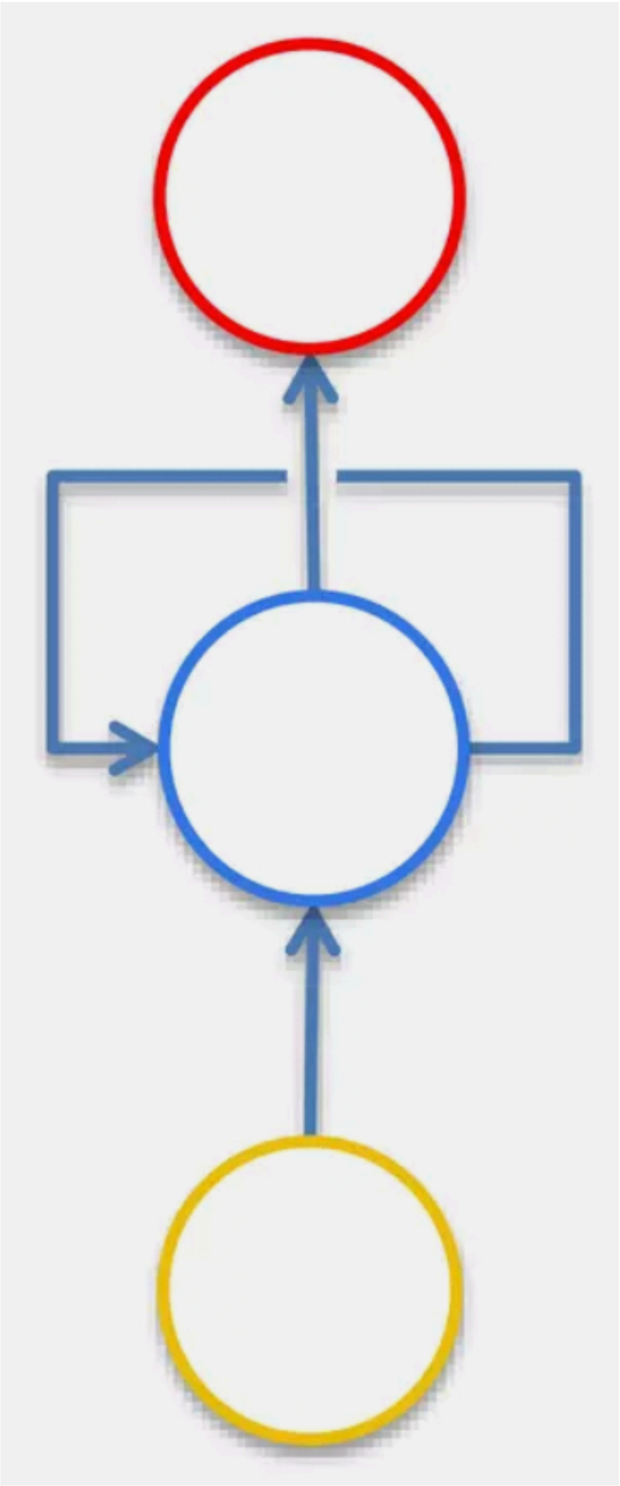

Input Vector
Output Vector
Input
Layer
Hidden
Layer
Output
Layer

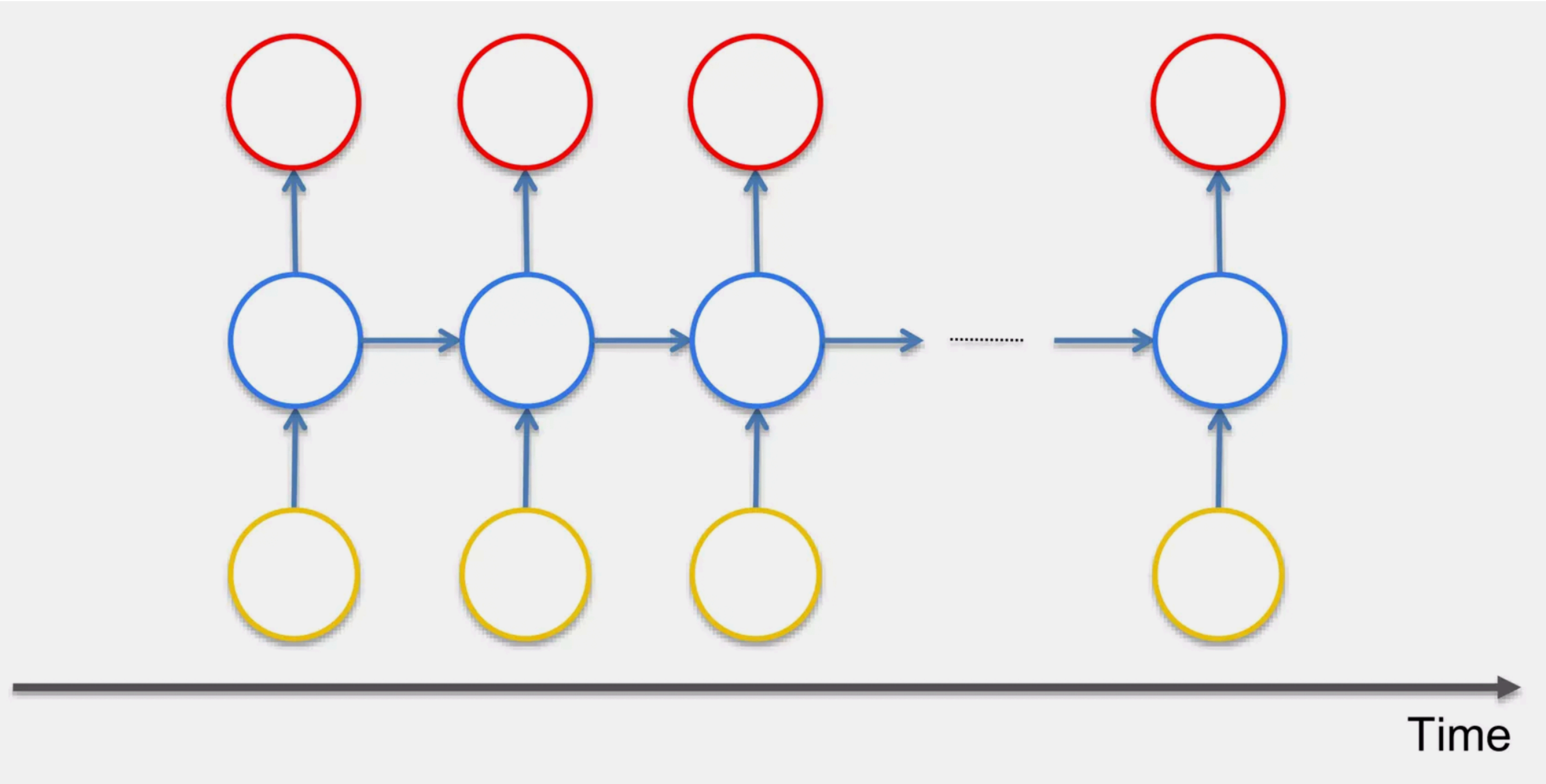


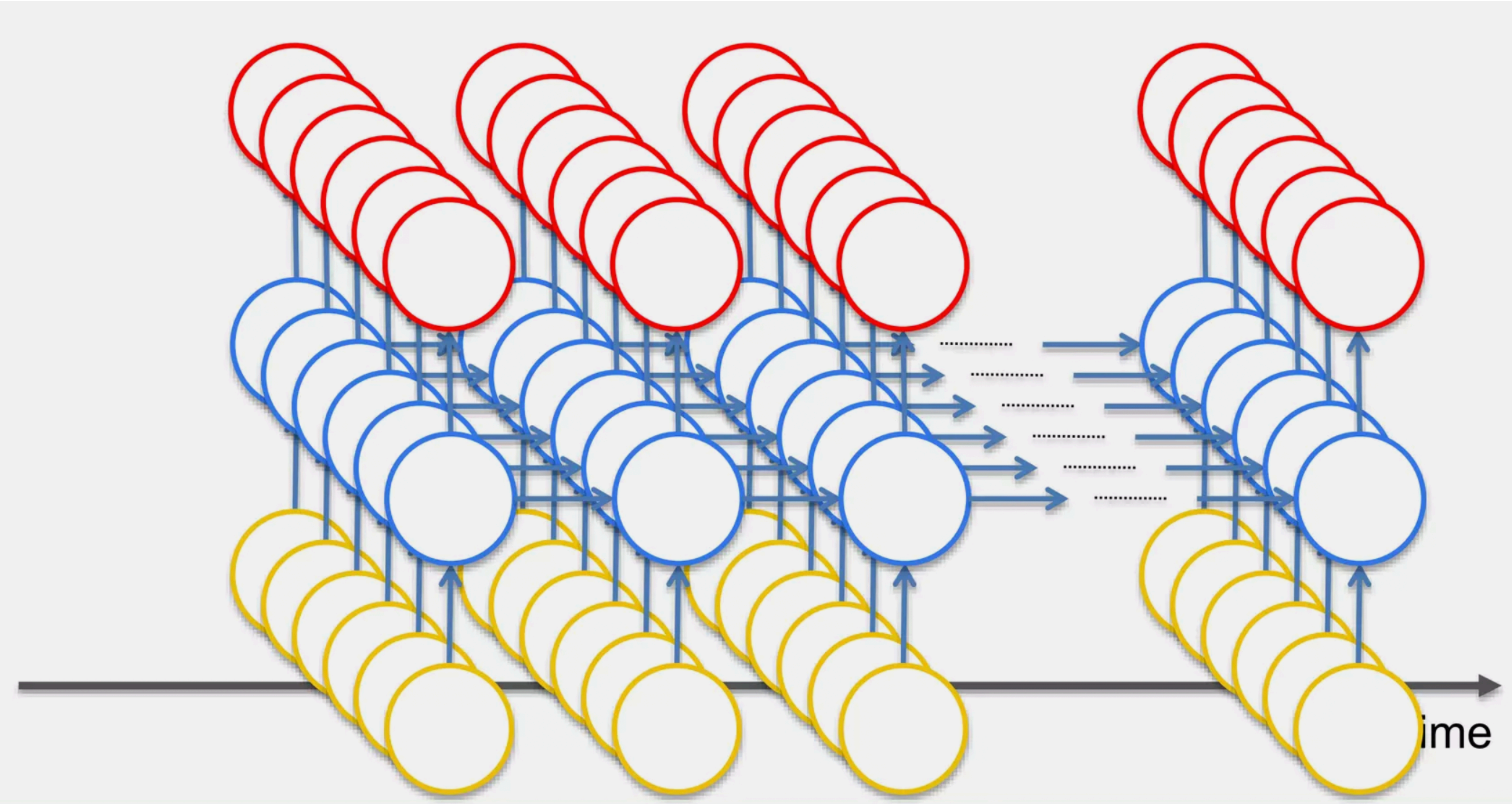


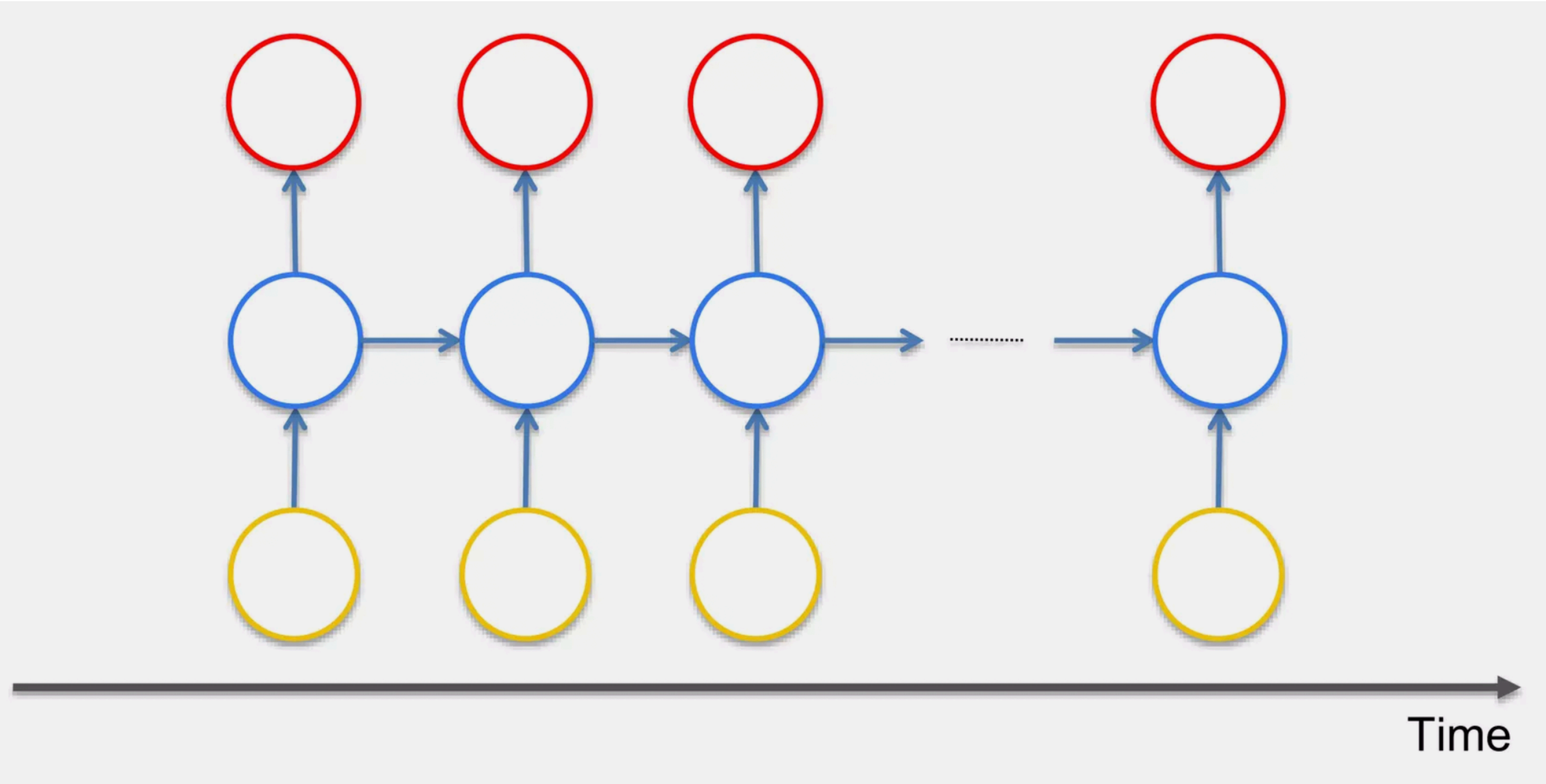







Time

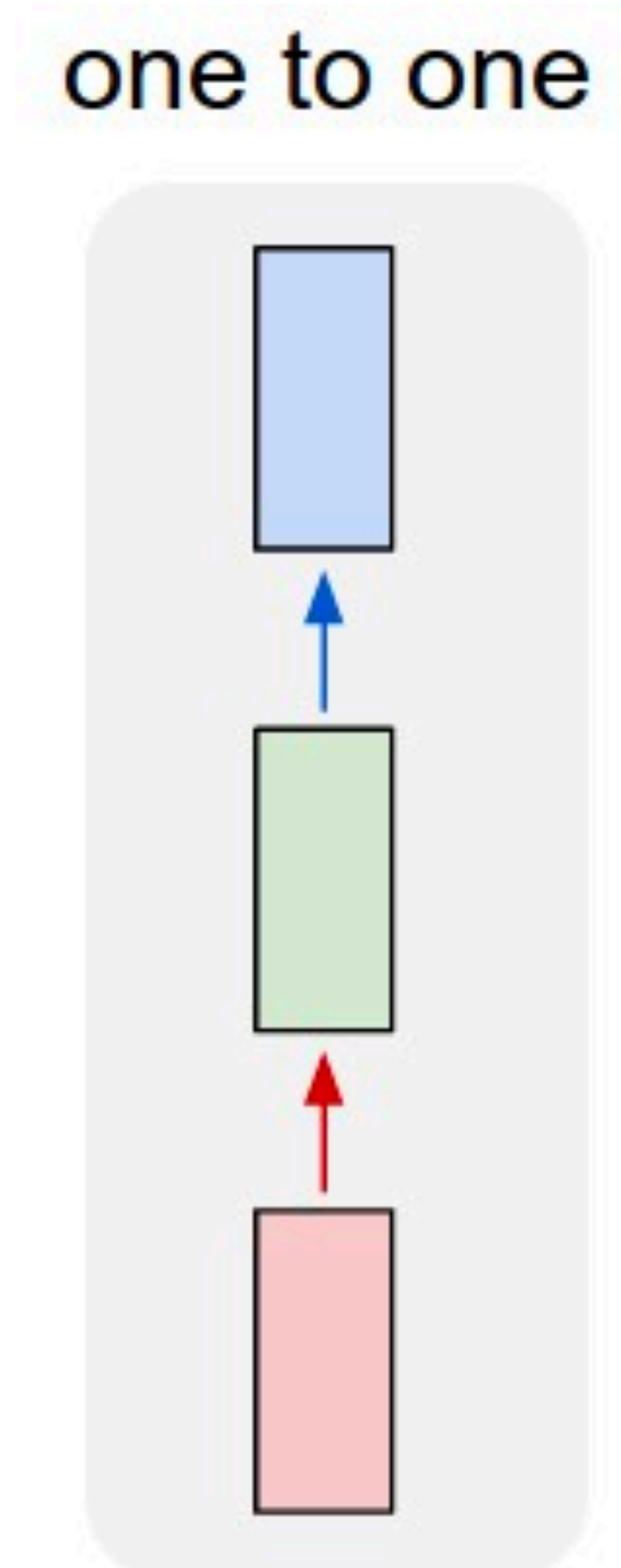

ime


Time

RRN’s Aplications

One to one

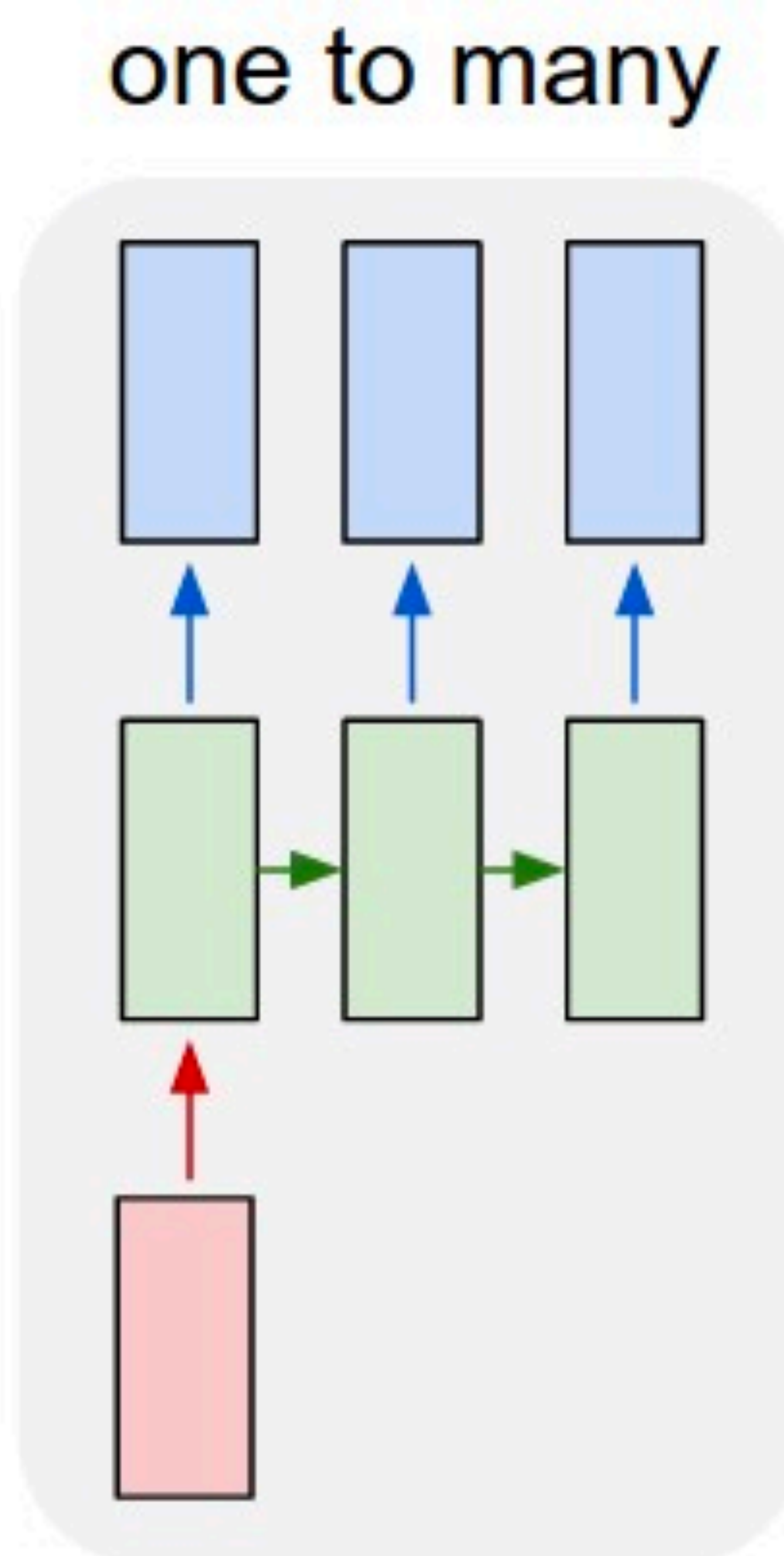
- Vanilla type without RNN
- fixed-sized input
- fixed-sized output
- Image classification



From http://karpathy.github.io/2015/05/21/rnn-effectiveness

One to many

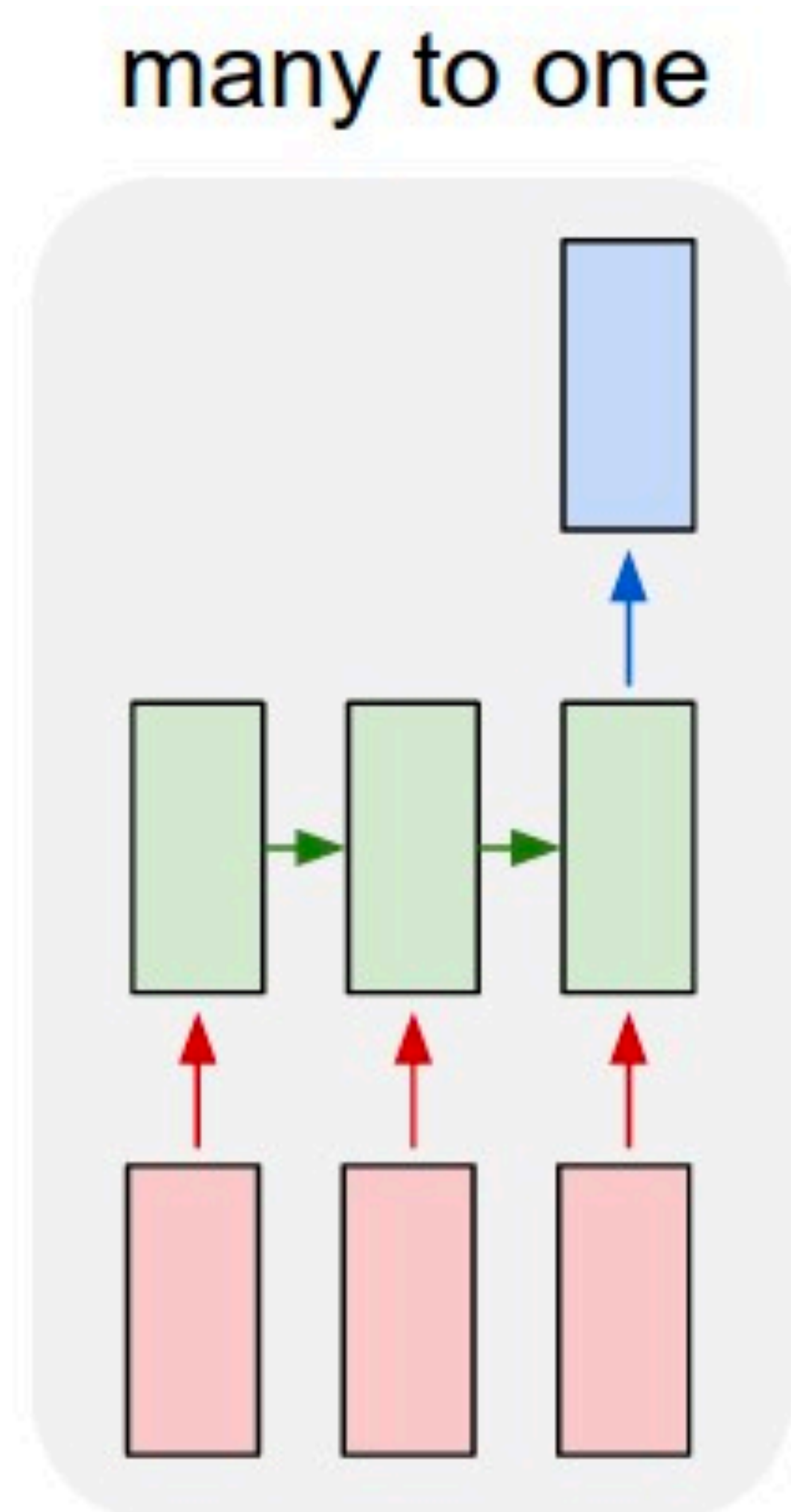
- Sequence output
- image captioning takes an image and outputs a sentence of words
- OCR (Optical Character Recognition)



From http://karpathy.github.io/2015/05/21/rnn-effectiveness

Many to one

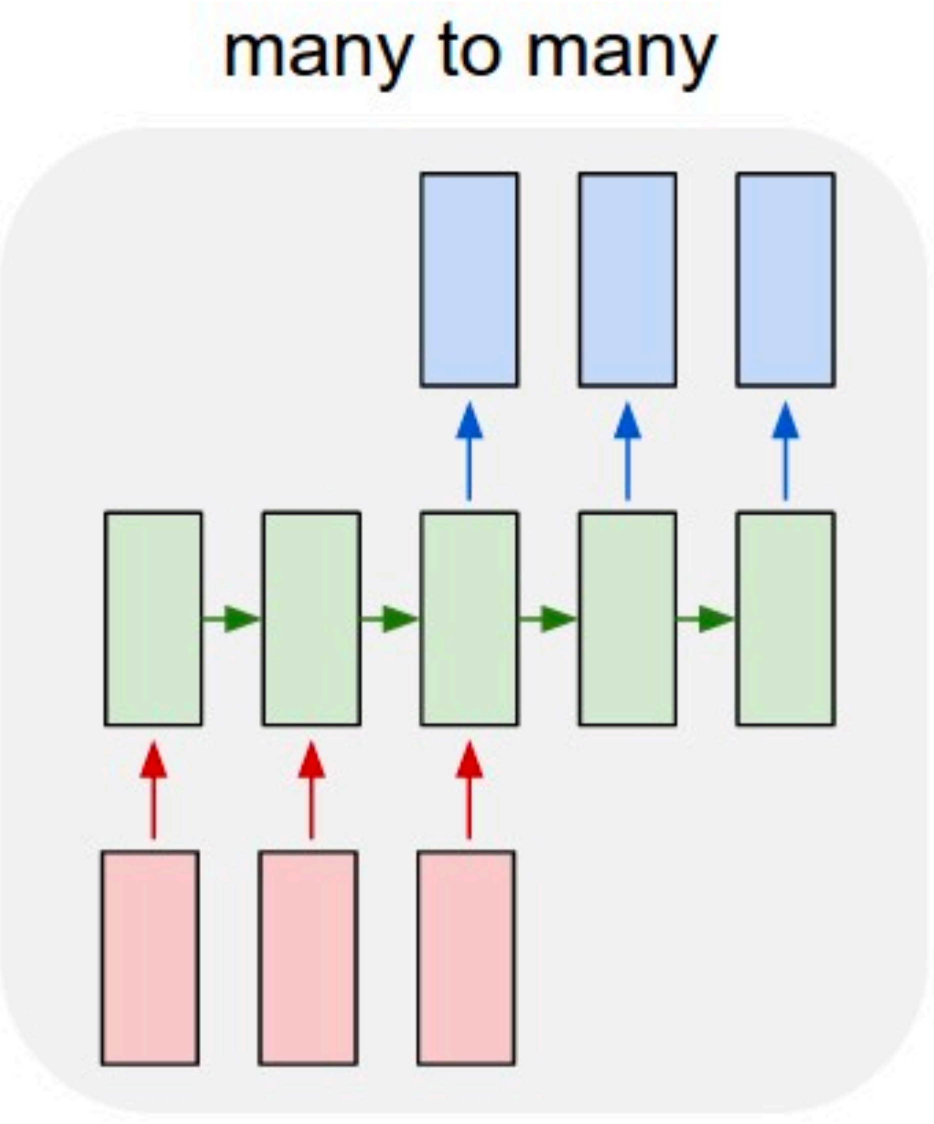
- Sequence input
- sentiment analysis where a given sentence is classified as expressing positive or negative sentiment
- Grammarly intent checker



From http://karpathy.github.io/2015/05/21/rnn-effectiveness

Many to many

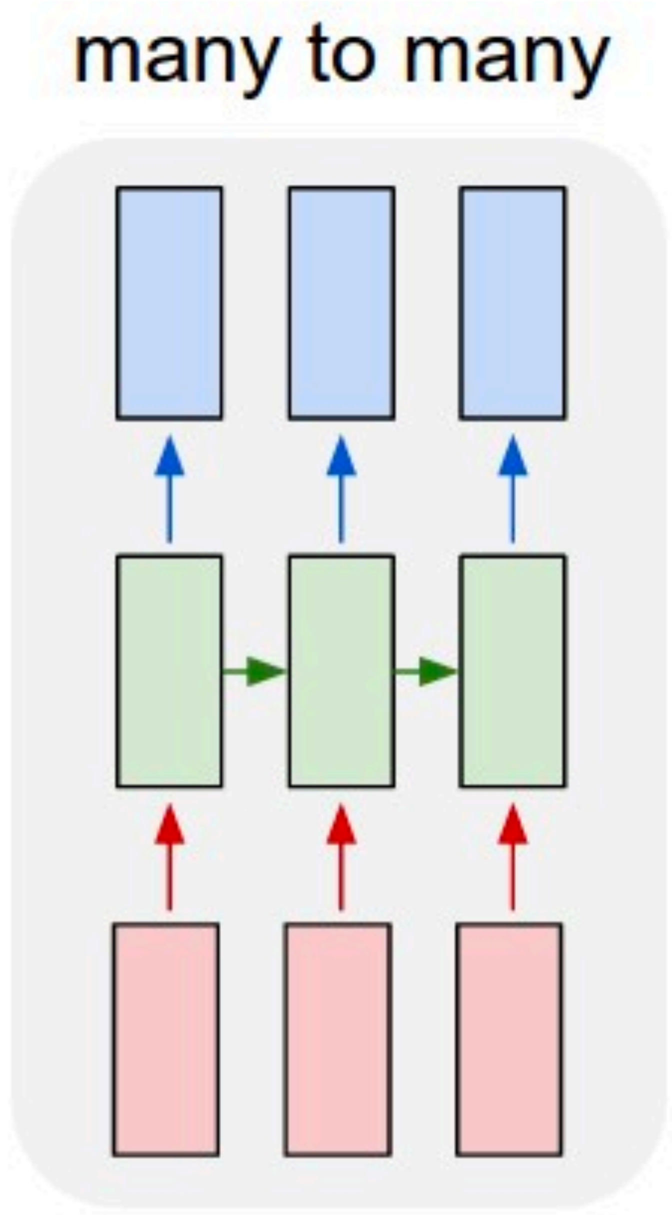
- Sequence input and sequence output
- Machine Translation: an RNN reads a sentence in English and then outputs a sentence in French
- Grammarly spelling checker



From http://karpathy.github.io/2015/05/21/rnn-effectiveness

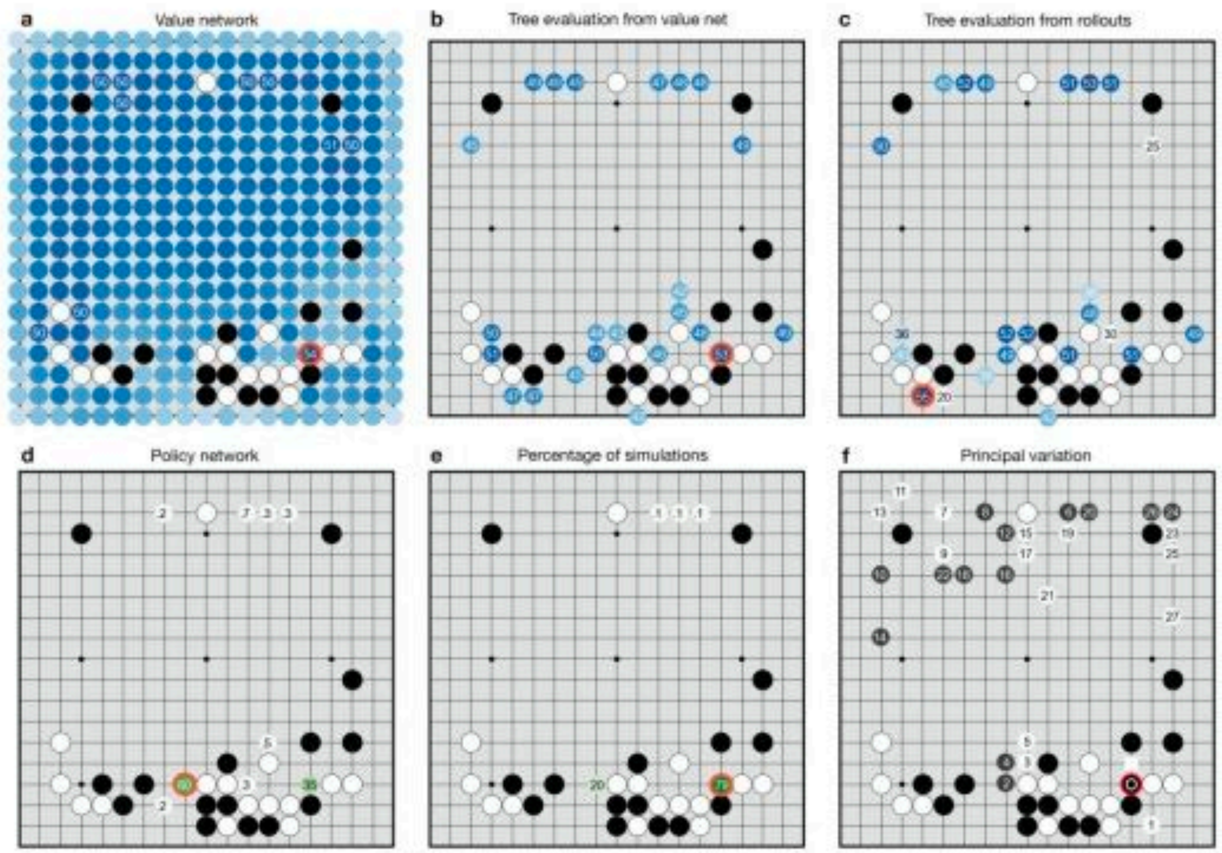
Many to many

- Synced sequence input and output
- video classification where we wish to label each frame of the video



From http://karpathy.github.io/2015/05/21/rnn-effectiveness

Vanishing Gradient

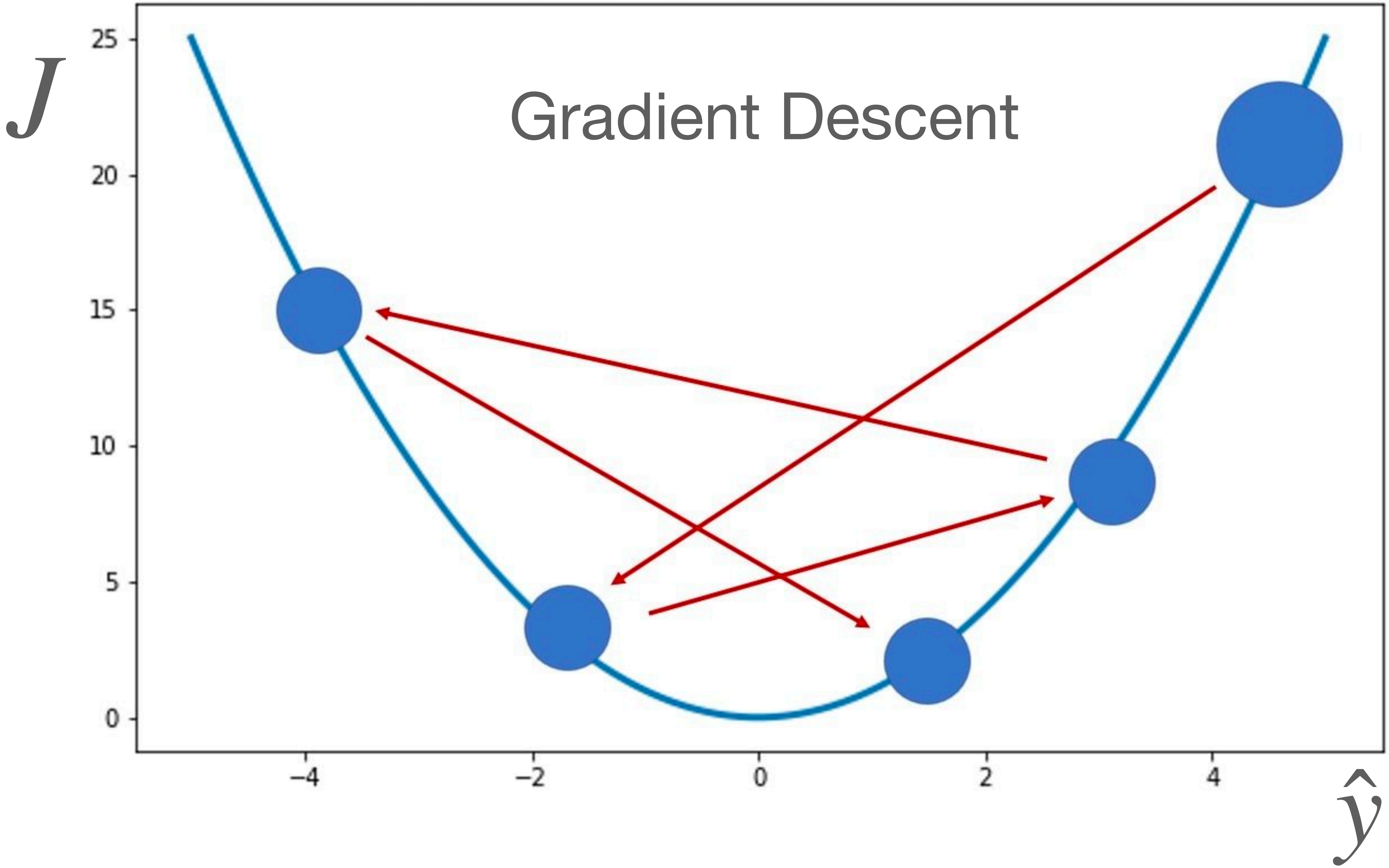


Sepp Hochreiter

AlphaGo

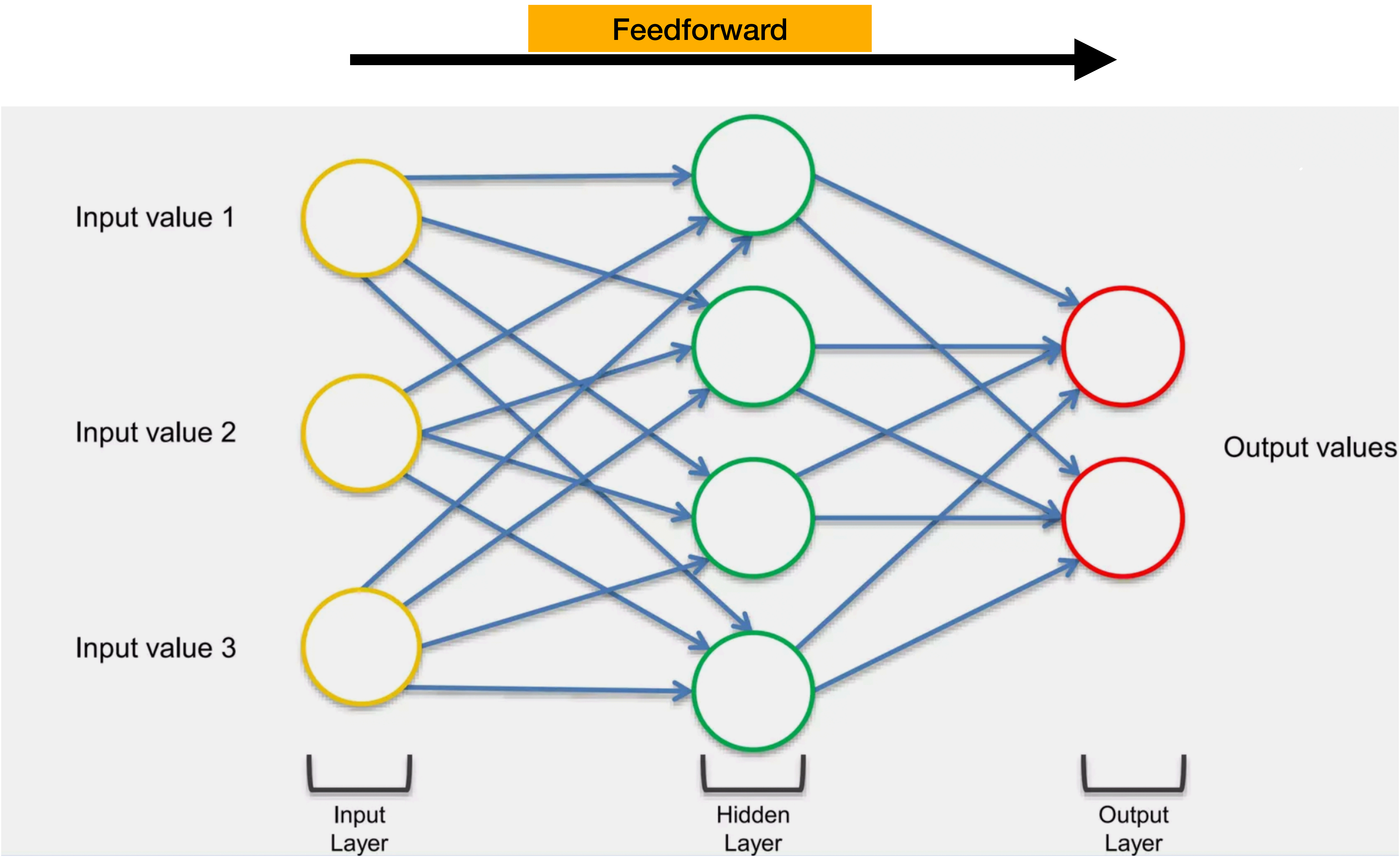
IARAI

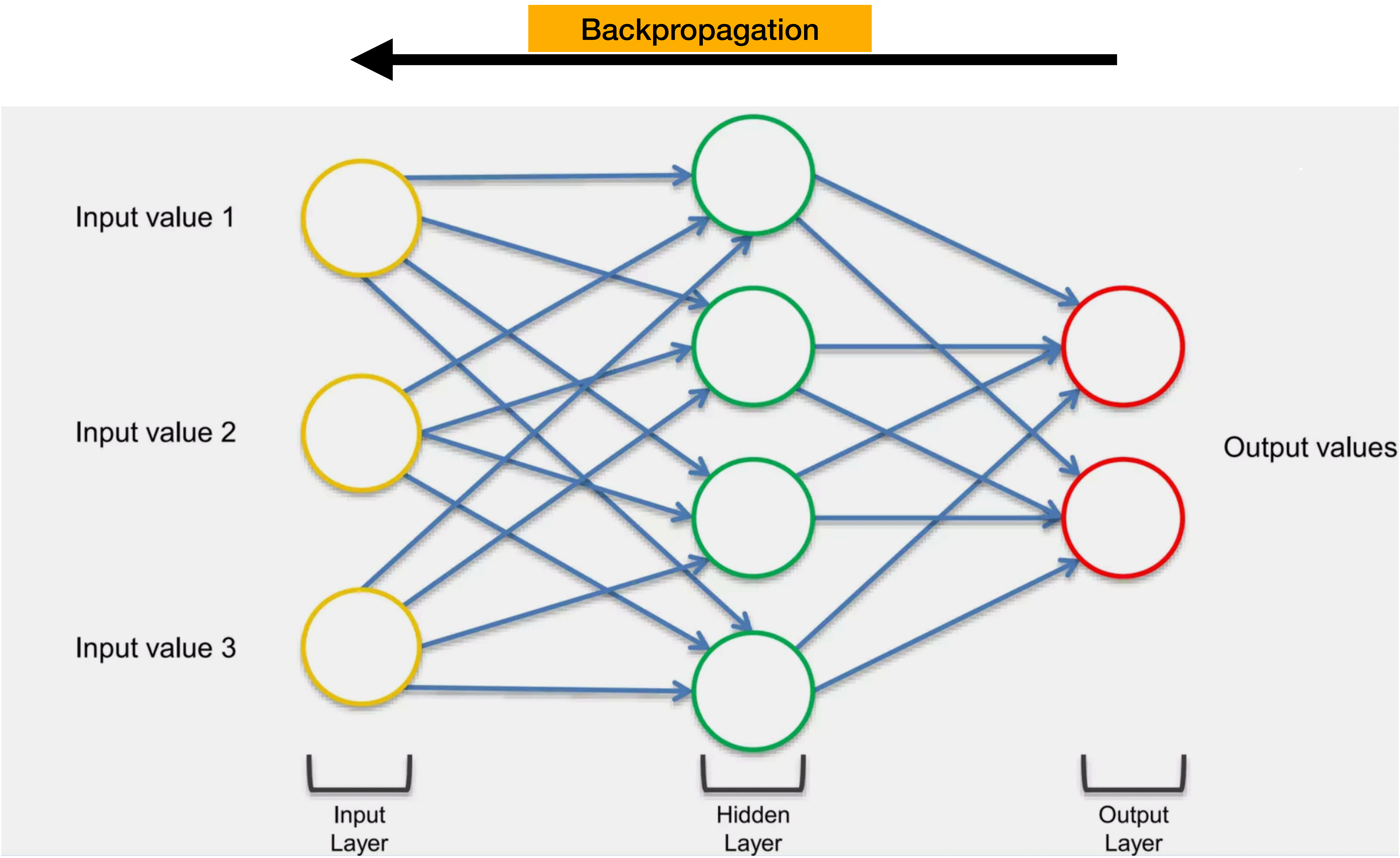
institute of advanced research in artificial intelligence

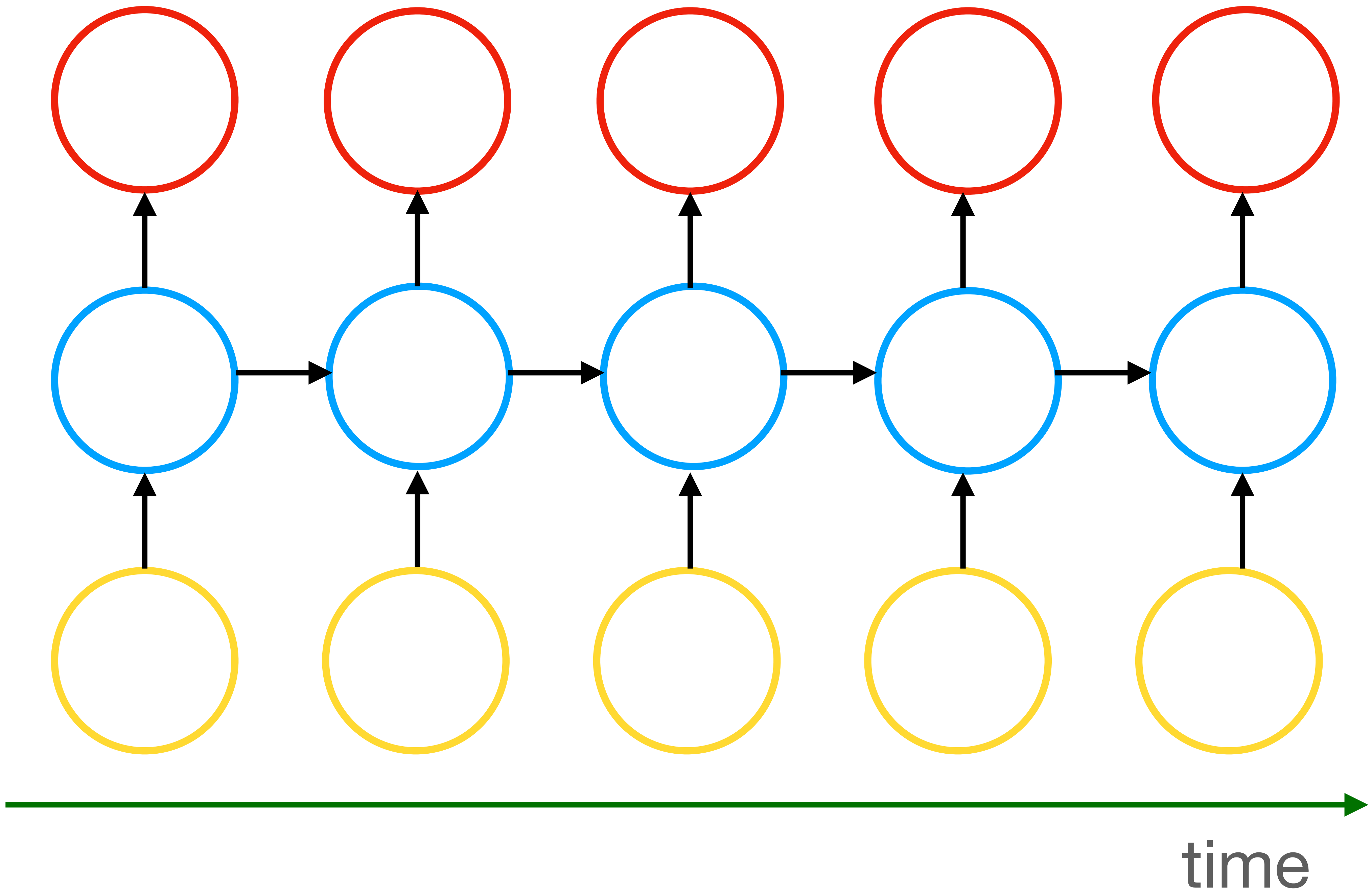


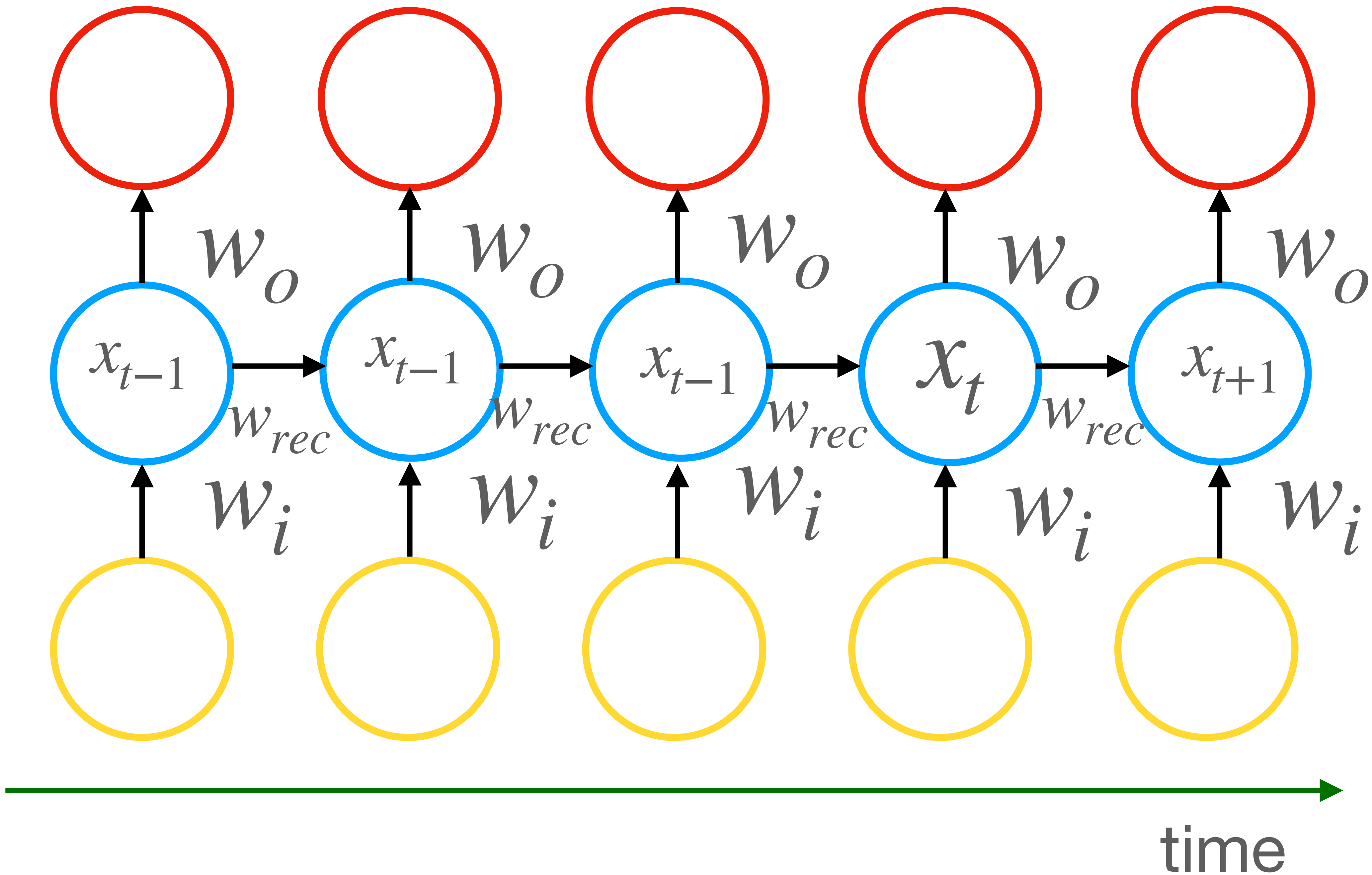
$$J = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y} - y)^2$$

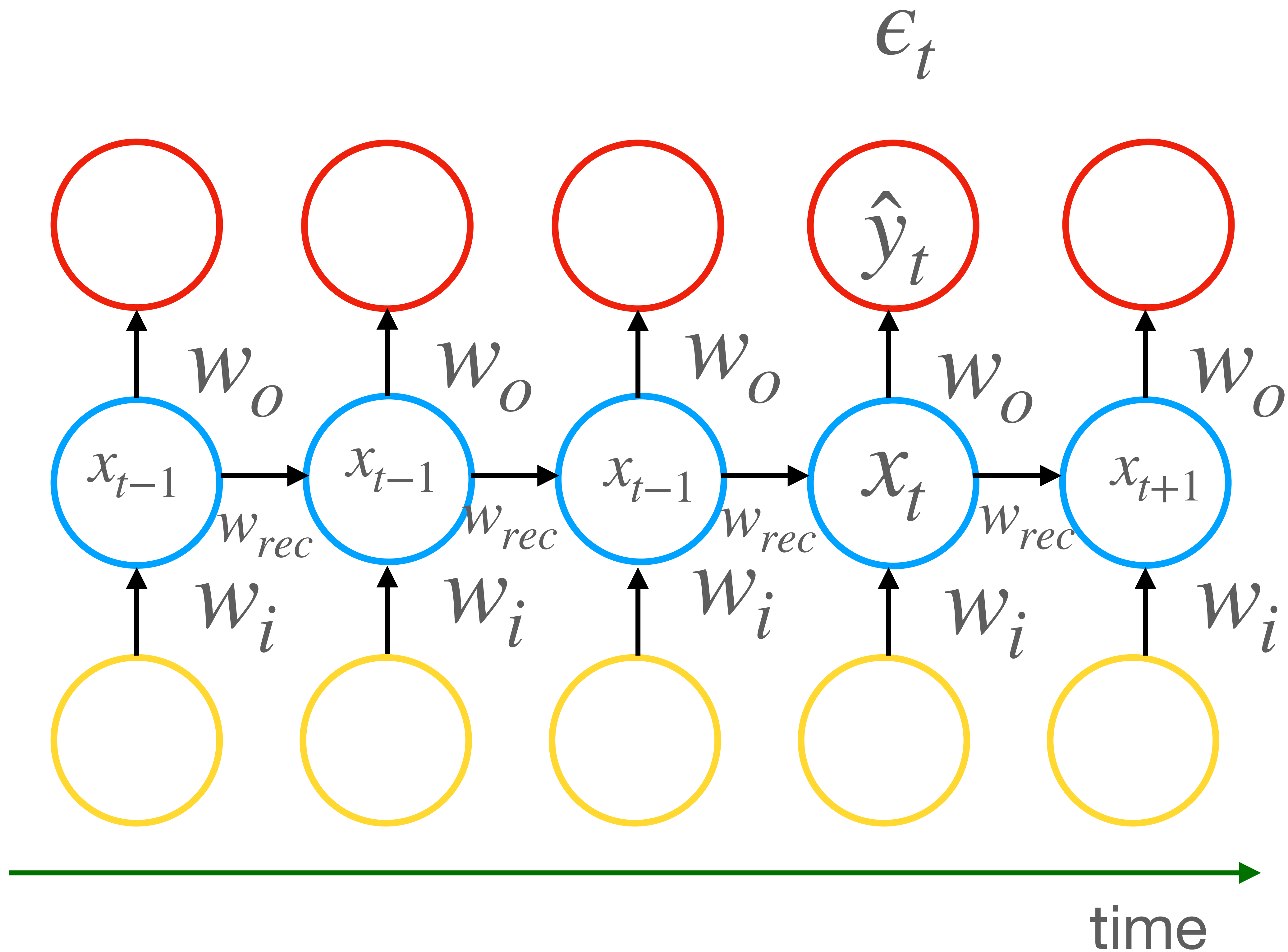
$$J = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(X) - y)^2$$


Feedforward
Input value 1
Input value 2
Input value 3
Output values
Input Layer
Hidden Layer
Output Layer

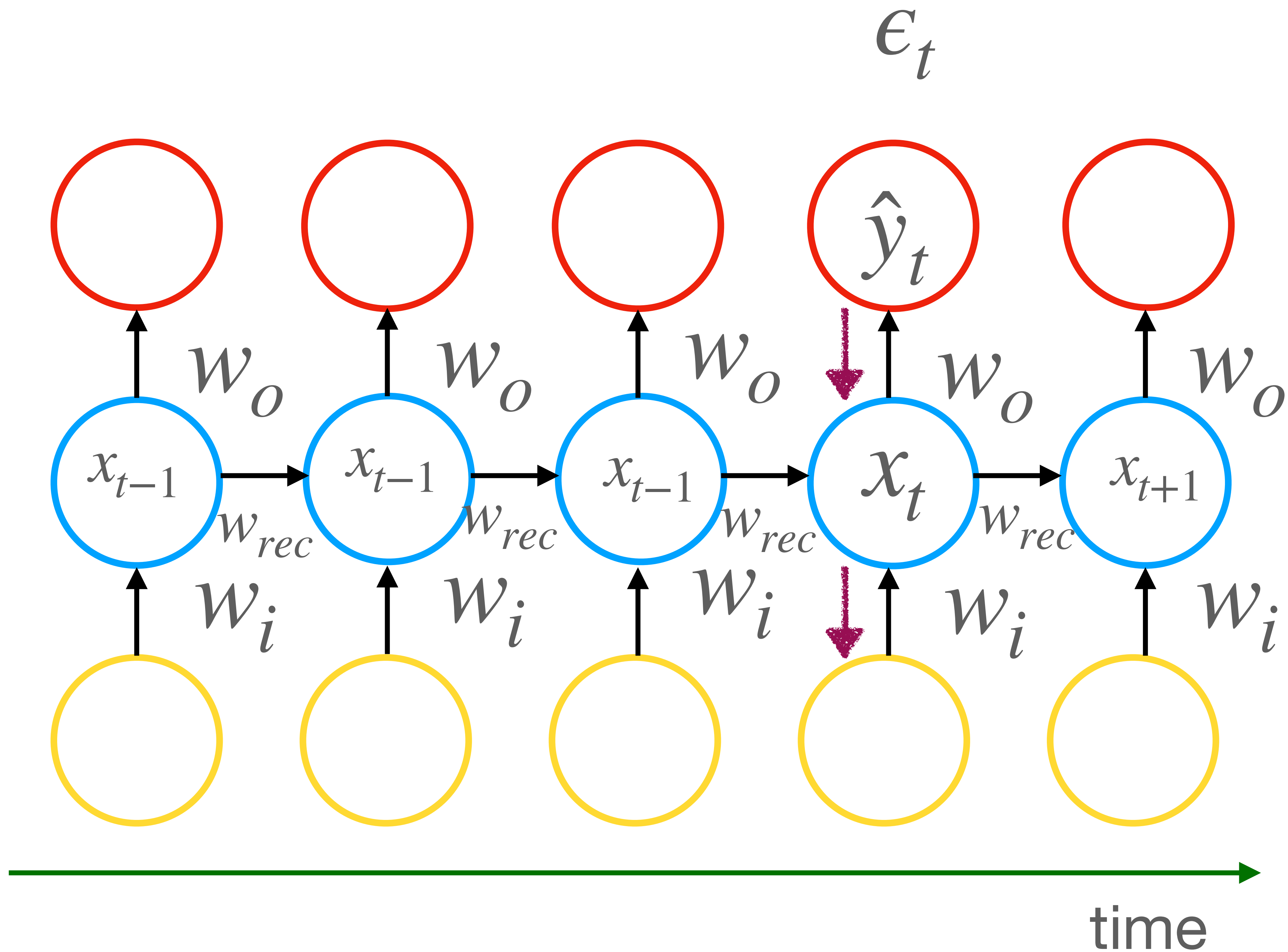

Backpropagation
Input value 1
Input value 2
Input value 3
Output values
Input
Layer
Hidden
Layer
Output
Layer


time

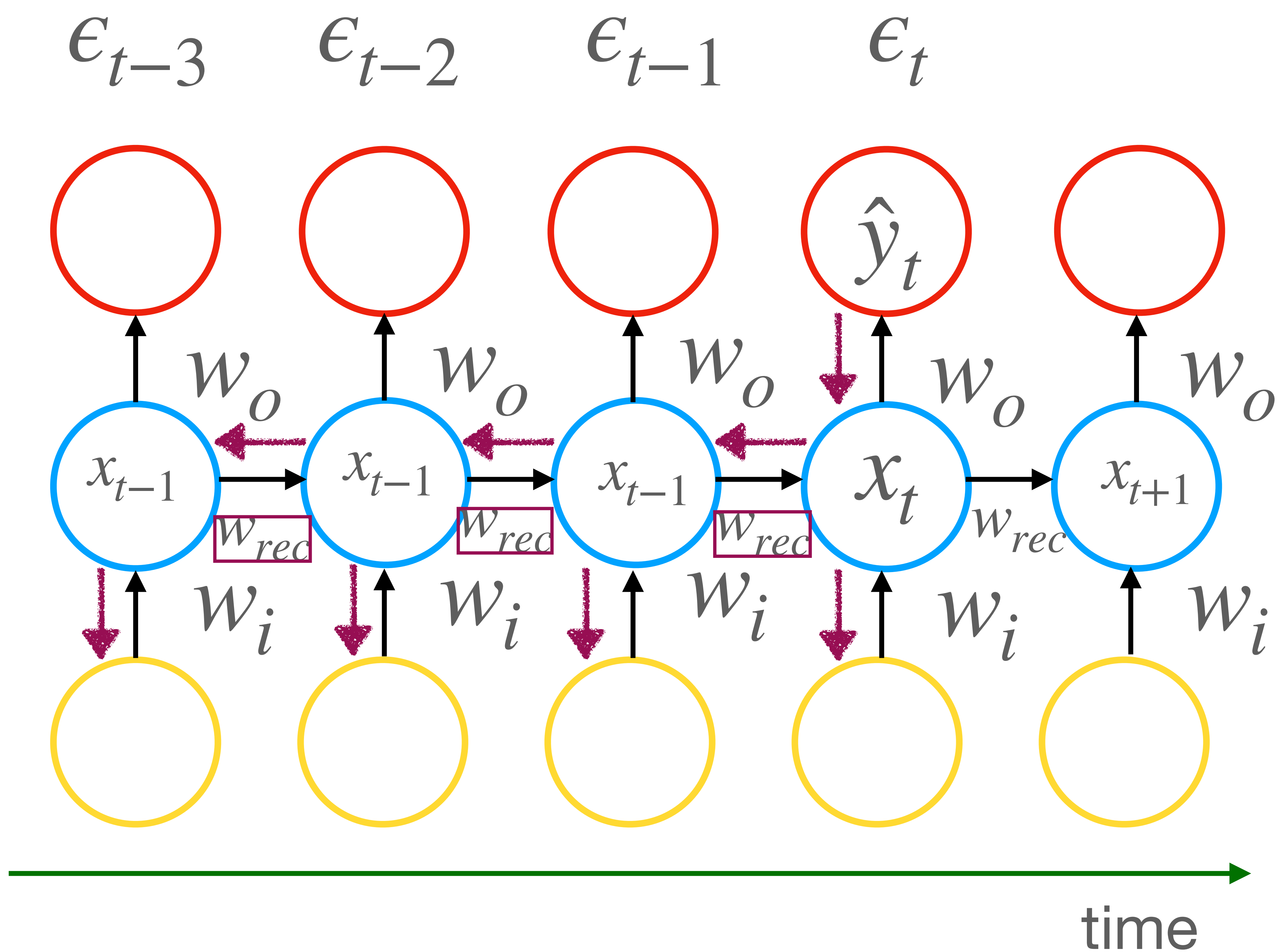

x_{t-1}
x_{t-1}
x_{t-1}
x_t
x_{t+1}
w_o
w_i
w_{rec}
time



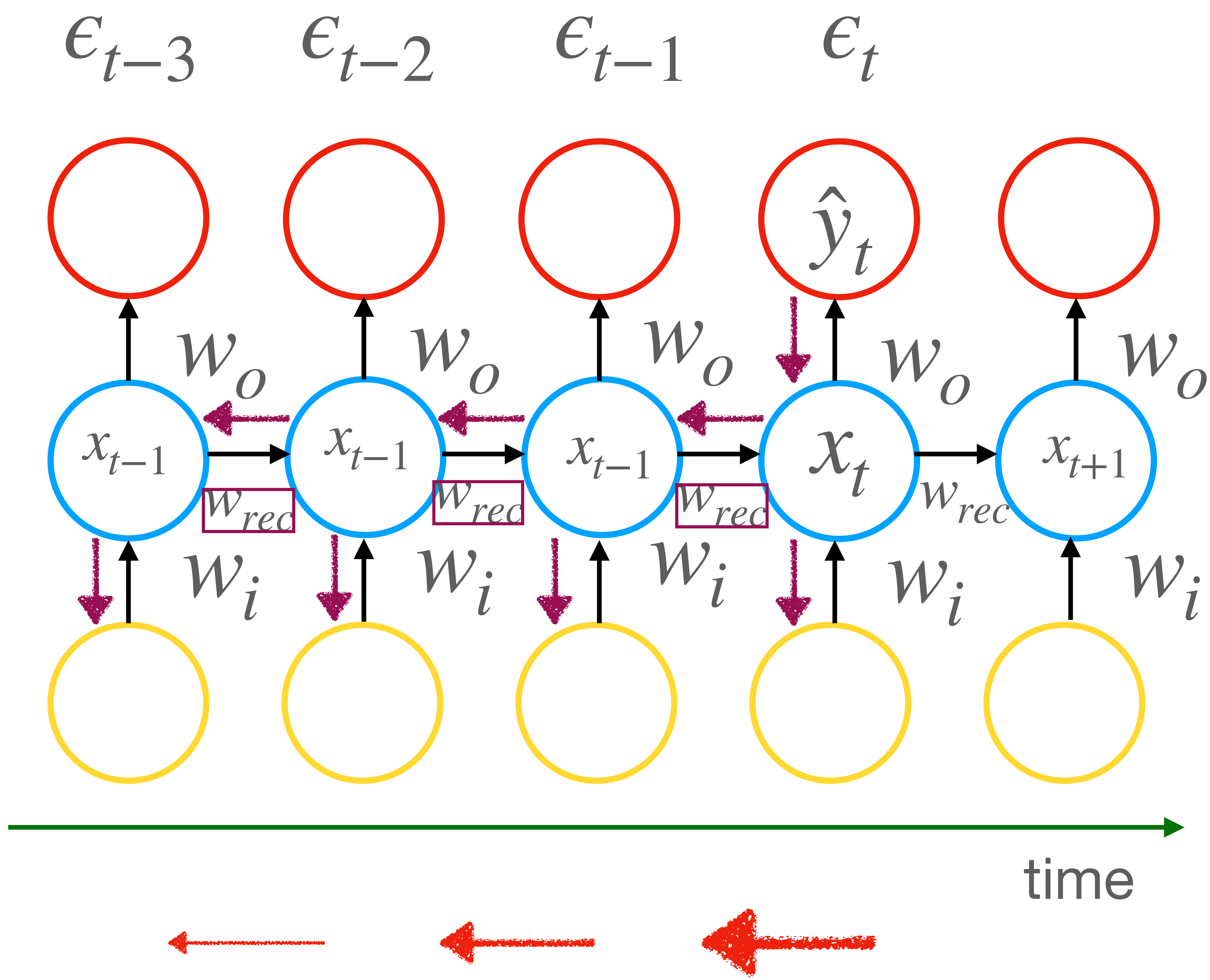
ϵ_t
$\hat{y}_t$
w_o
x_{t-1}
x_{t-1}
x_{t-1}
x_t
x_{t+1}
w_{rec}
w_i
time



ϵ_t
$\hat{y}_t$
w_o
x_{t-1}
x_{t-1}
x_{t-1}
x_t
x_{t+1}
w_{rec}
w_i
time



ϵ_{t-3}
ϵ_{t-2}
ϵ_{t-1}
ϵ_t
$\hat{y}_t$
w_o
w_o
w_o
w_o
w_o
x_{t-1}
x_{t-1}
x_{t-1}
x_t
x_{t+1}
w_{rec}
w_{rec}
w_{rec}
w_{rec}
w_i
w_i
w_i
w_i
w_i
time



ϵ_{t-3}
ϵ_{t-2}
ϵ_{t-1}
ϵ_t
$\hat{y}_t$
w_o
w_o
w_o
w_o
w_o
x_{t-1}
x_{t-1}
x_{t-1}
x_t
x_{t+1}
w_{rec}
w_{rec}
w_{rec}
w_{rec}
w_i
w_i
w_i
w_i
w_i
time

$$\frac{\partial \epsilon}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\partial \epsilon_t}{\partial \theta}$$

$$\frac{\partial \epsilon_t}{\partial \theta} = \sum_{1 \leq t \leq T} \left(\frac{\partial \epsilon_t}{\partial X_t} \frac{\partial X_t}{\partial X_k} \frac{\partial X_k}{\partial \theta} \right)$$

$$\frac{\partial X_t}{\partial X_k} = \prod_{1 \geq i \geq k} \frac{\partial X_i}{\partial X_{i-1}} = \prod_{1 \geq i \geq k} W_{rec}^T diag(\sigma'(x_{i-1}))$$

$$\frac{\partial X_t}{\partial X_k} = \prod_{1 \geq i \geq k} \frac{\partial X_i}{\partial X_{i-1}} = \prod_{1 \geq i \geq k} W_{rec}^T diag(\sigma'(x_{i-1}))$$

$$W_{rec}$$

$$W_{rec}$$

The vanishing gradient problem

Solutions

- Exploding Gradient
  - Truncated Backpropagation
  - Penalties
  - Gradient clipping
- Vanishing Gradient
  - Weight initialization
  - Echo state networks
  - Long Short-Term Memory Networks