Recurrent Neural Networks for Multivariate Time Series with Missing Values

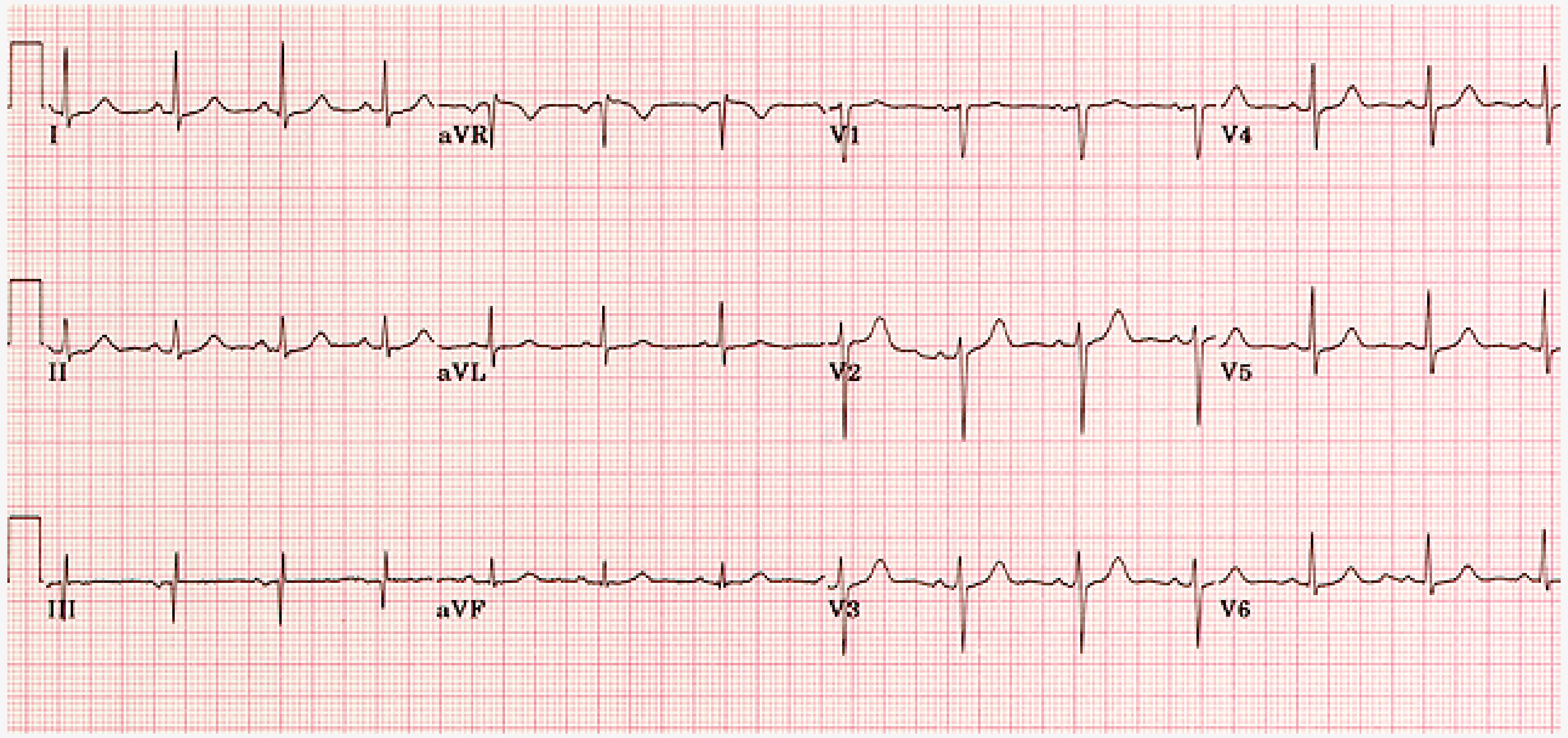
Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag & Yan Liu

Aslesha Pokhrel, Sujay Nagaraj

Why do we care?

Multivariate time-series data is prevalent in a wide variety of fields:

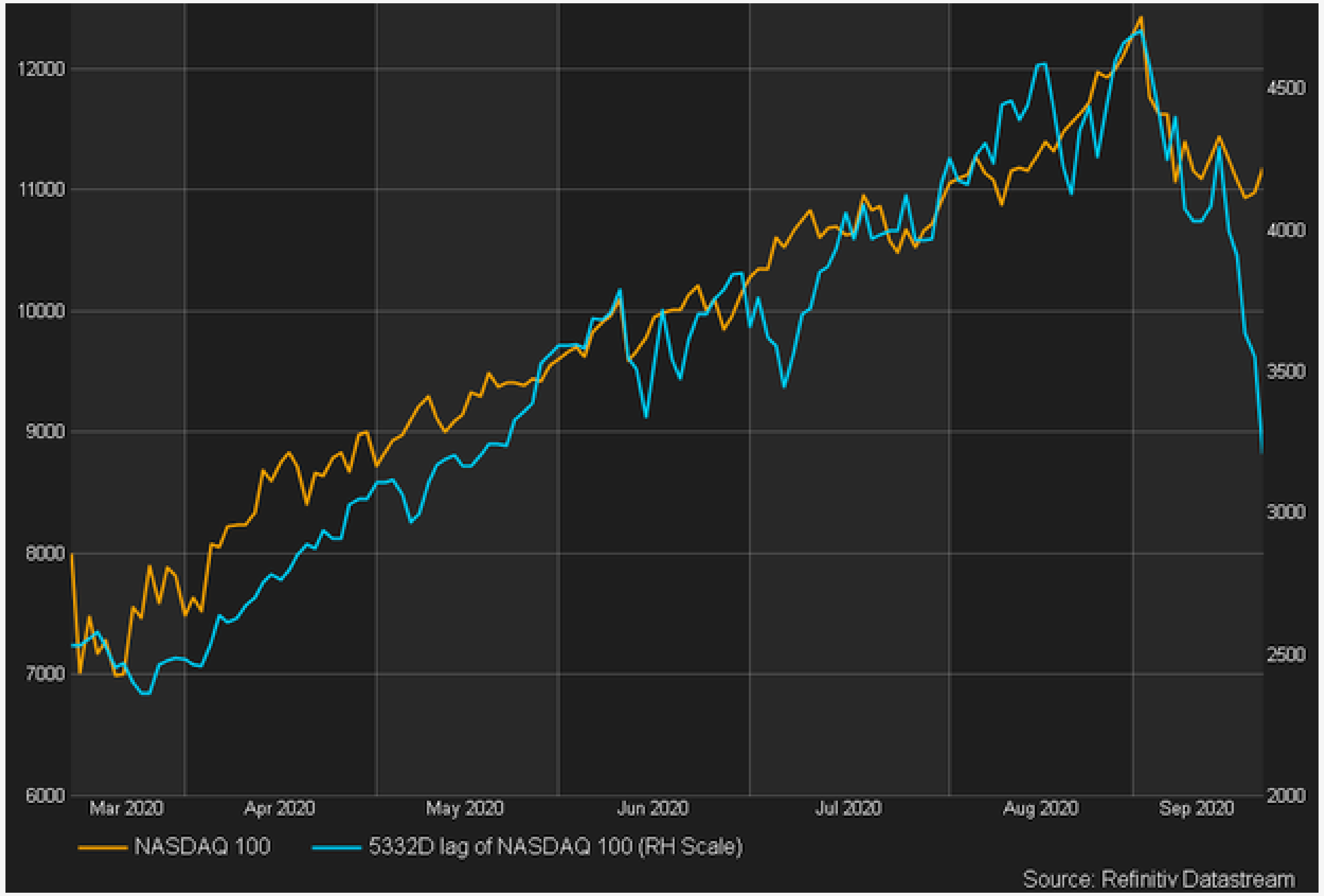
- Health care (ICU, wearables)



Why do we care?

Multivariate time-series data is prevalent in a wide variety of fields:

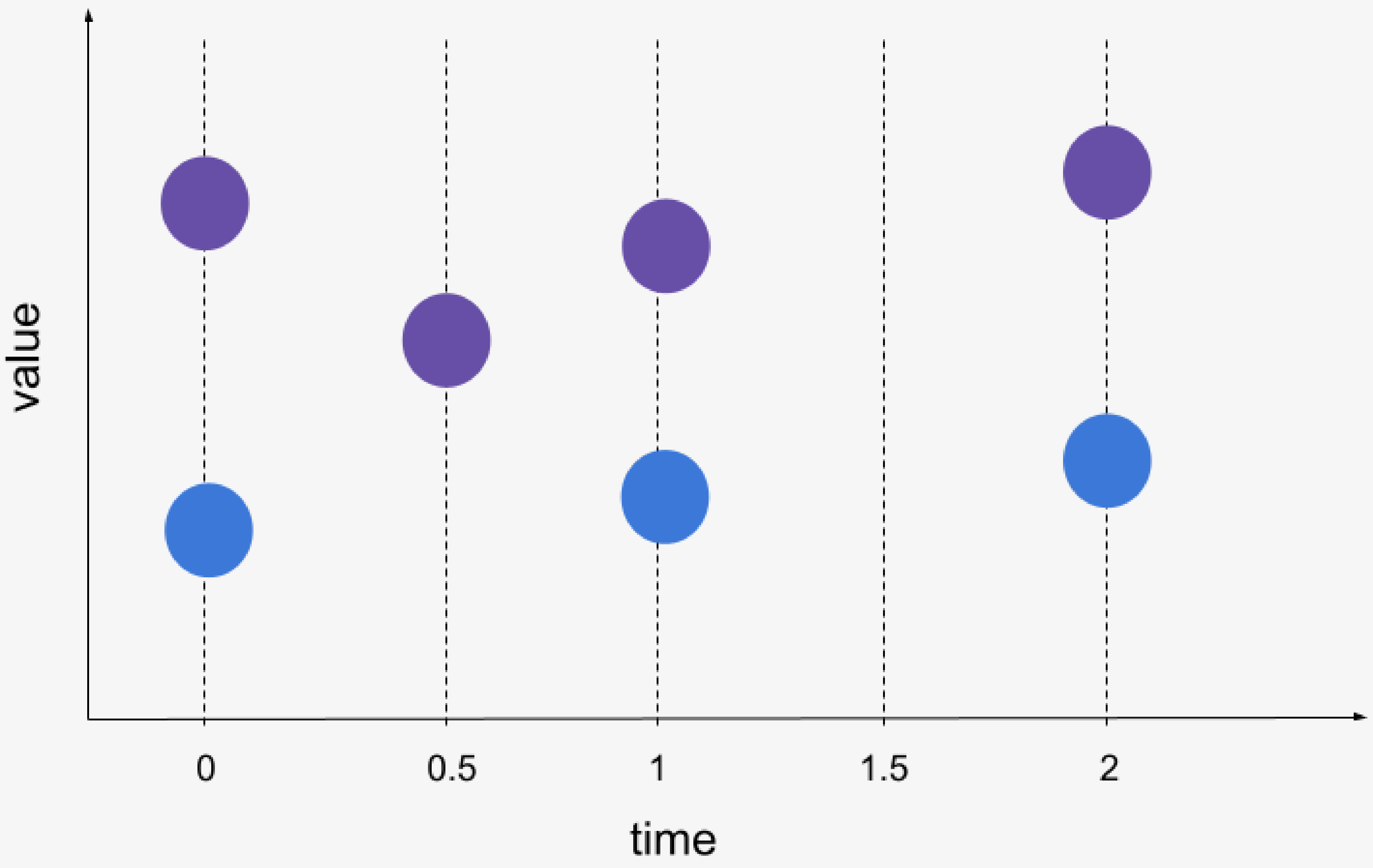
- Health care (ICU, wearables)
- Economics,
- Geoscience



Problem

Missing values

- Different frequencies
- Device malfunction
- Data Corruption
- Human errors
- Intentional



Baseline Approaches

Omit the missing data - Inadequate samples when the missing rate is high.

Data imputation - Methods such as smoothing, interpolation and spline are agnostic to the variable correlation and do not capture complex pattern.

Better methods - Spectral analysis, kernel methods, EM algorithm, matrix completion and matrix factorization

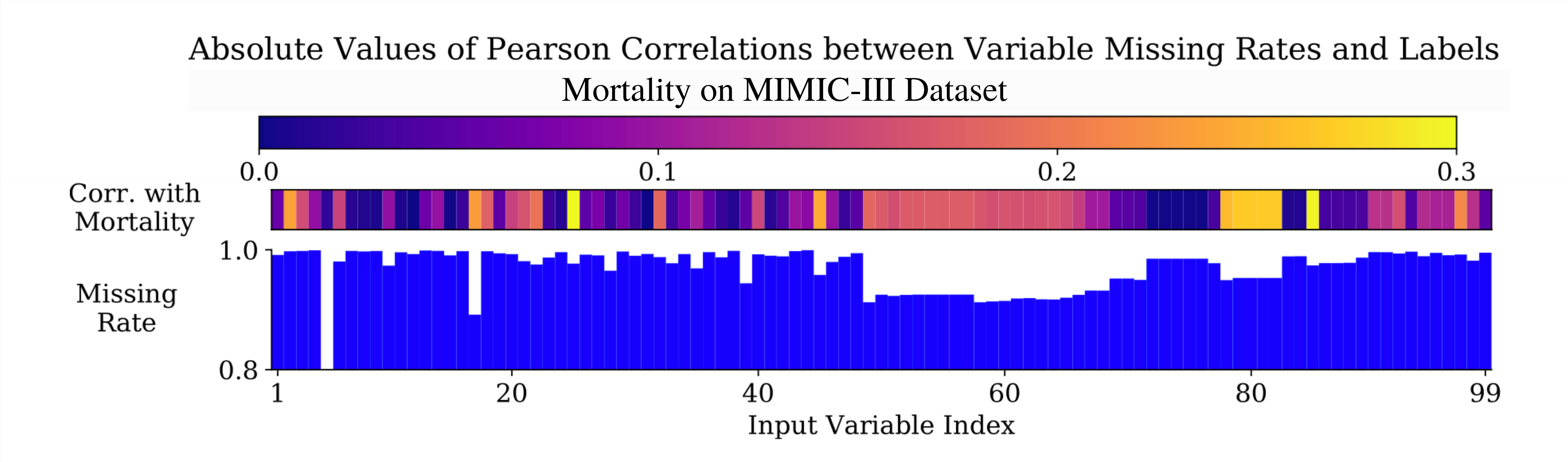
Multiple imputation and averaging data

Problem

- Assumptions of imputation method often not satisfied:
  - Small missing rate
  - Missing at random
- Results in a two-step process where imputation and prediction are separated
  - Missing patterns not effectively explored

Informative Missingness

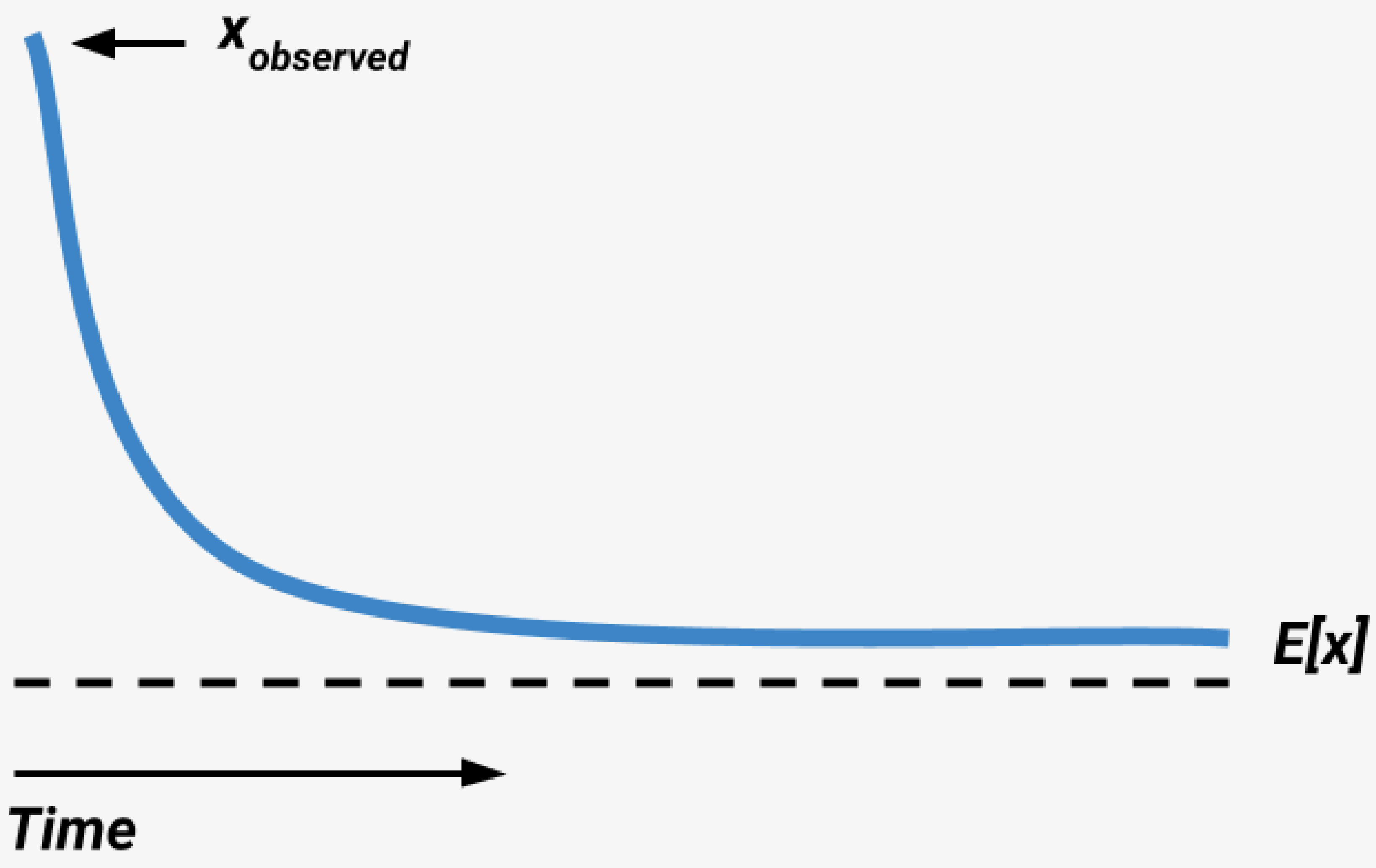
- Missing pattern correlated with the target labels
- Provide information about labels in the supervised learning tasks



Missingness in health data

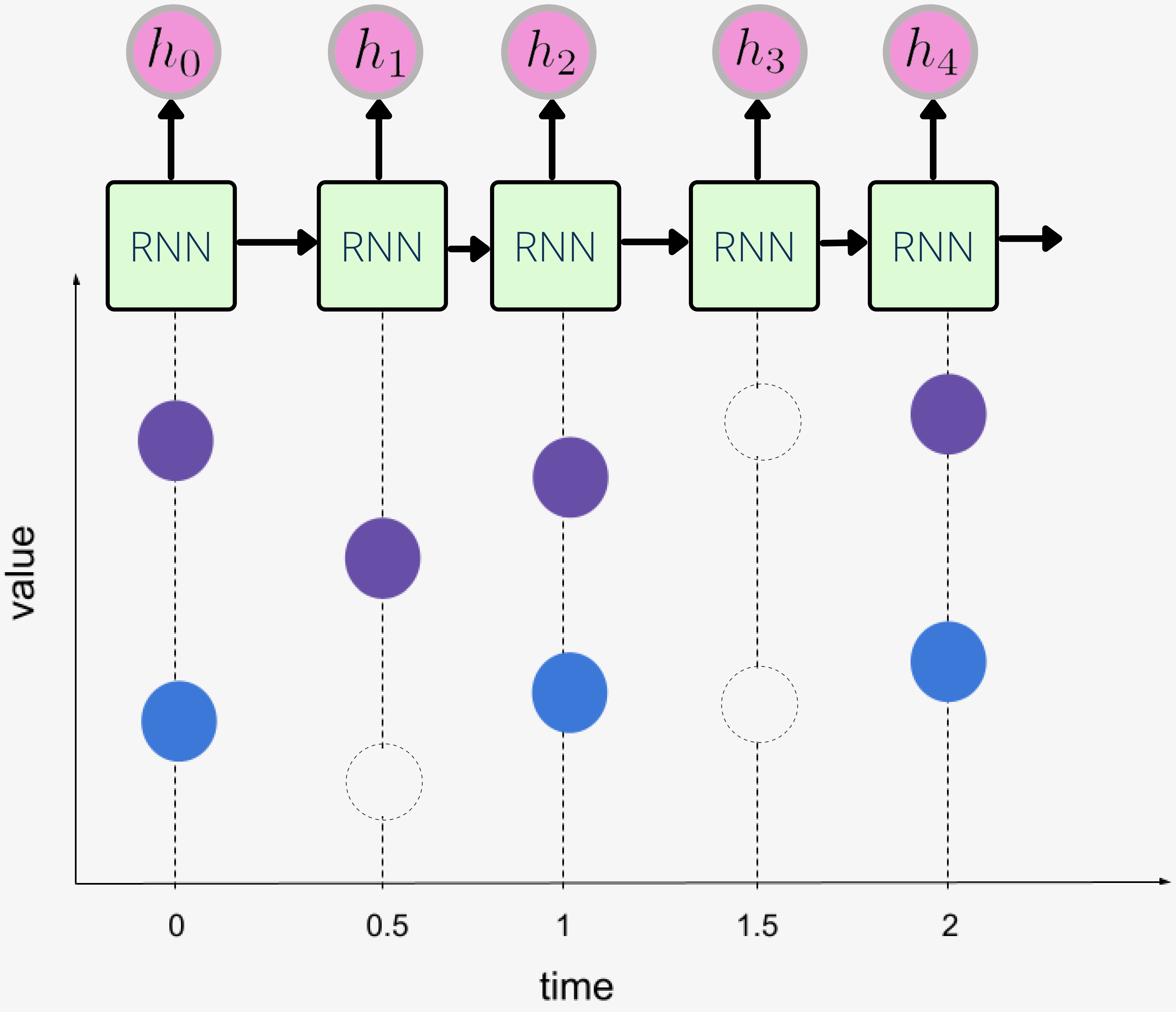
- Missing value of a variable tends to be close to some default value
    - Homeostasis

- Influence of the input variables that has been missing for a while will fade away over time

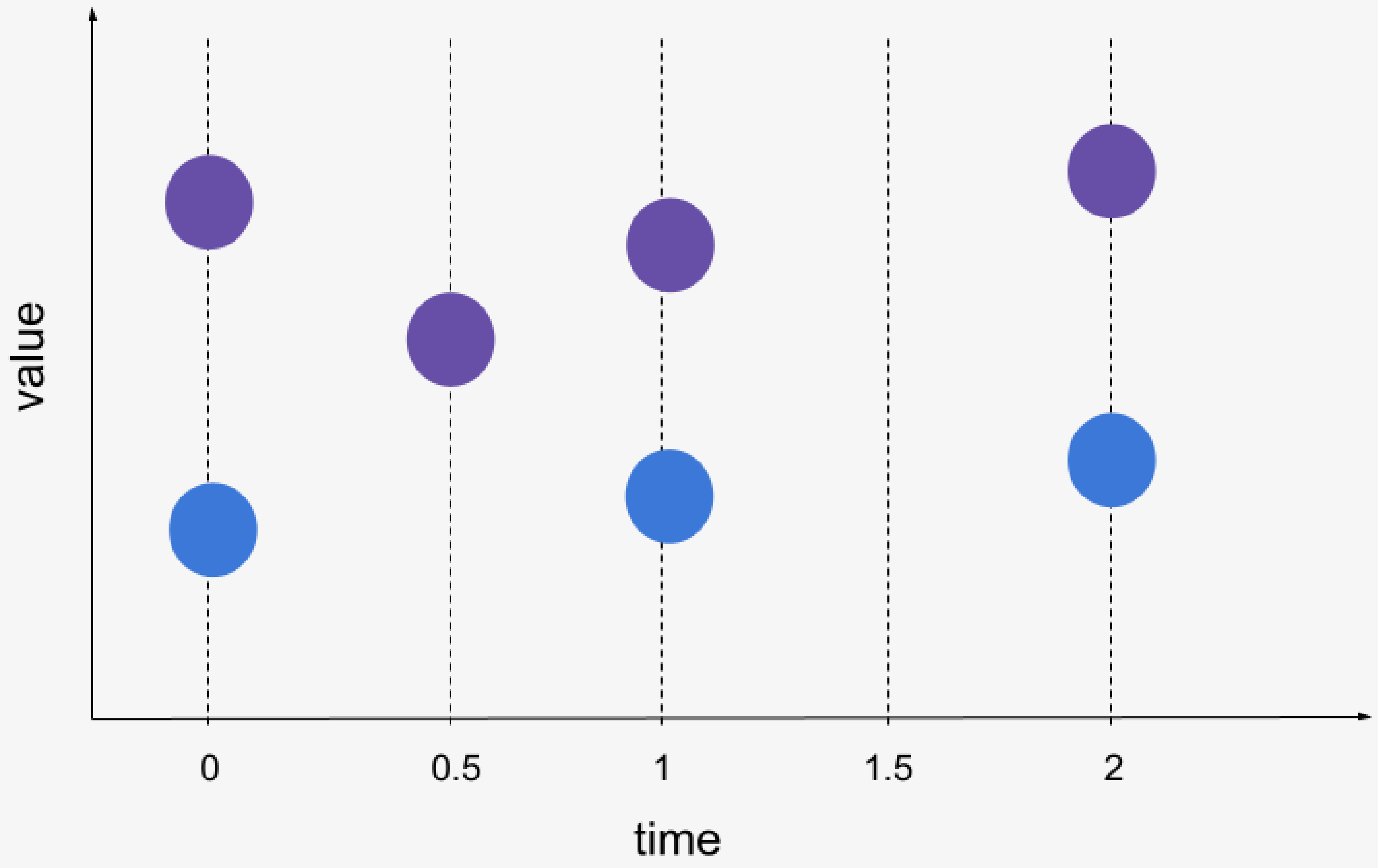


RNN

- Models sequential data
- Auto-regressive
- Vanilla RNN suffer from vanishing gradient problem
  - GRU mitigates this problem



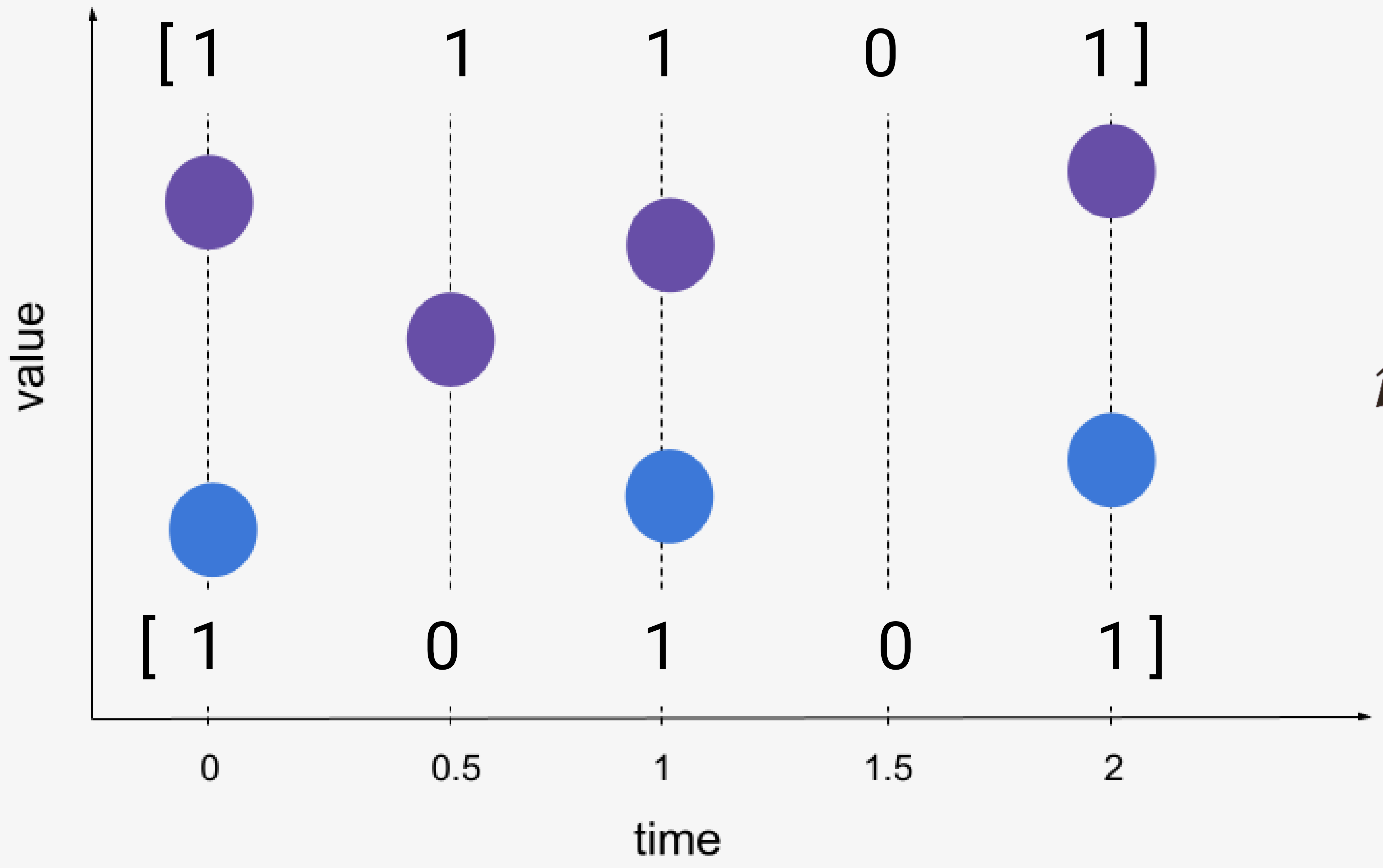
Input



$$\mathbf{X} = \begin{bmatrix} 12 & 10 & 11 & NA & 13 \\ 6 & NA & 7 & NA & 8 \end{bmatrix}$$

$$\mathbf{s} = \begin{bmatrix} 0 & 0.5 & 1.0 & 1.5 & 2.0 \end{bmatrix}$$

Masking

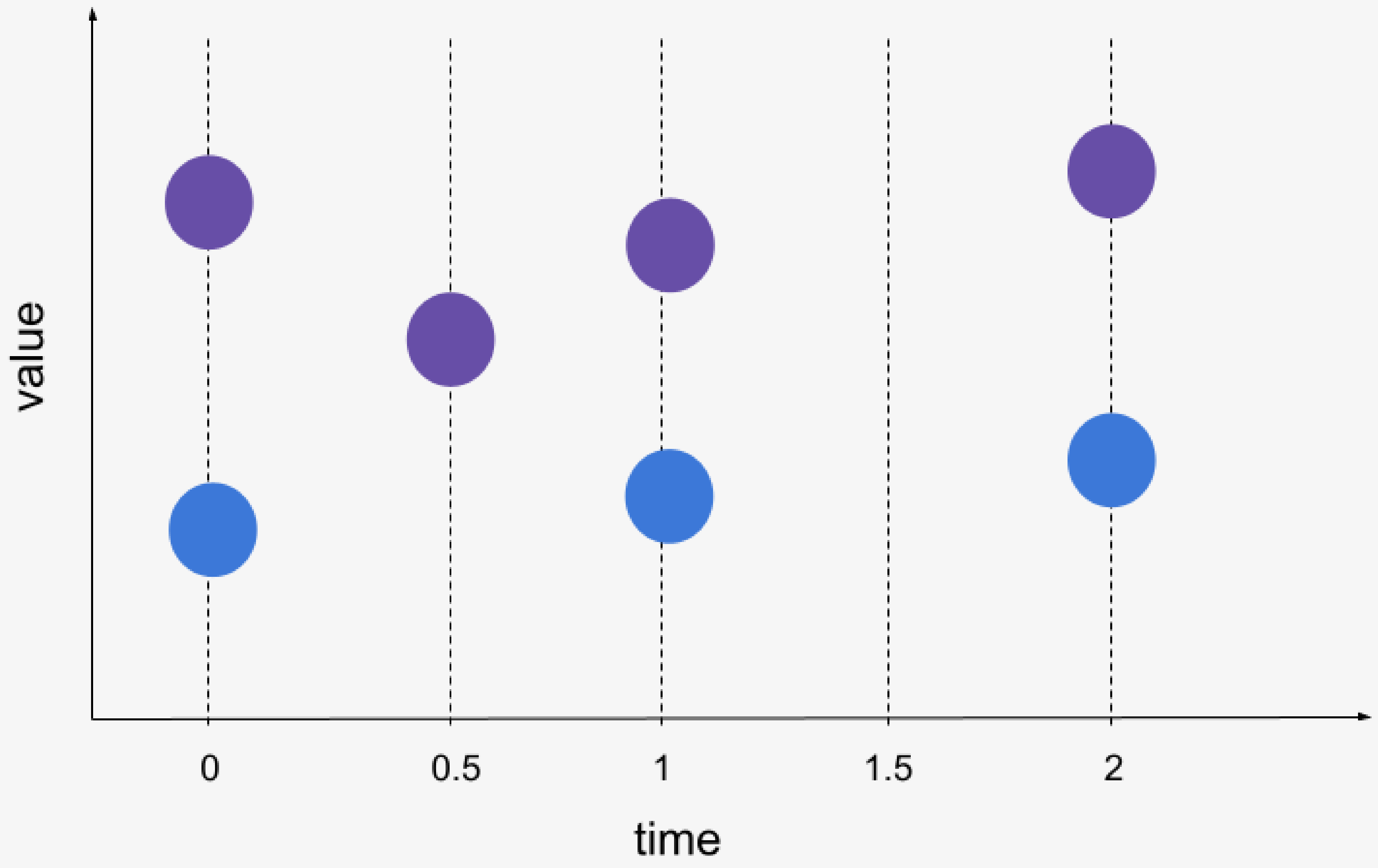


$$\mathbf{X} = \begin{bmatrix} 12 & 10 & 11 & NA & 13 \\ 6 & NA & 7 & NA & 8 \end{bmatrix}$$

$$\mathbf{s} = \begin{bmatrix} 0 & 0.5 & 1.0 & 1.5 & 2.0 \end{bmatrix}$$

$$m_t^d = \begin{cases} 1, & \text{if } x_t^d \text{ is observed} \\ 0, & \text{otherwise} \end{cases}$$

Masking

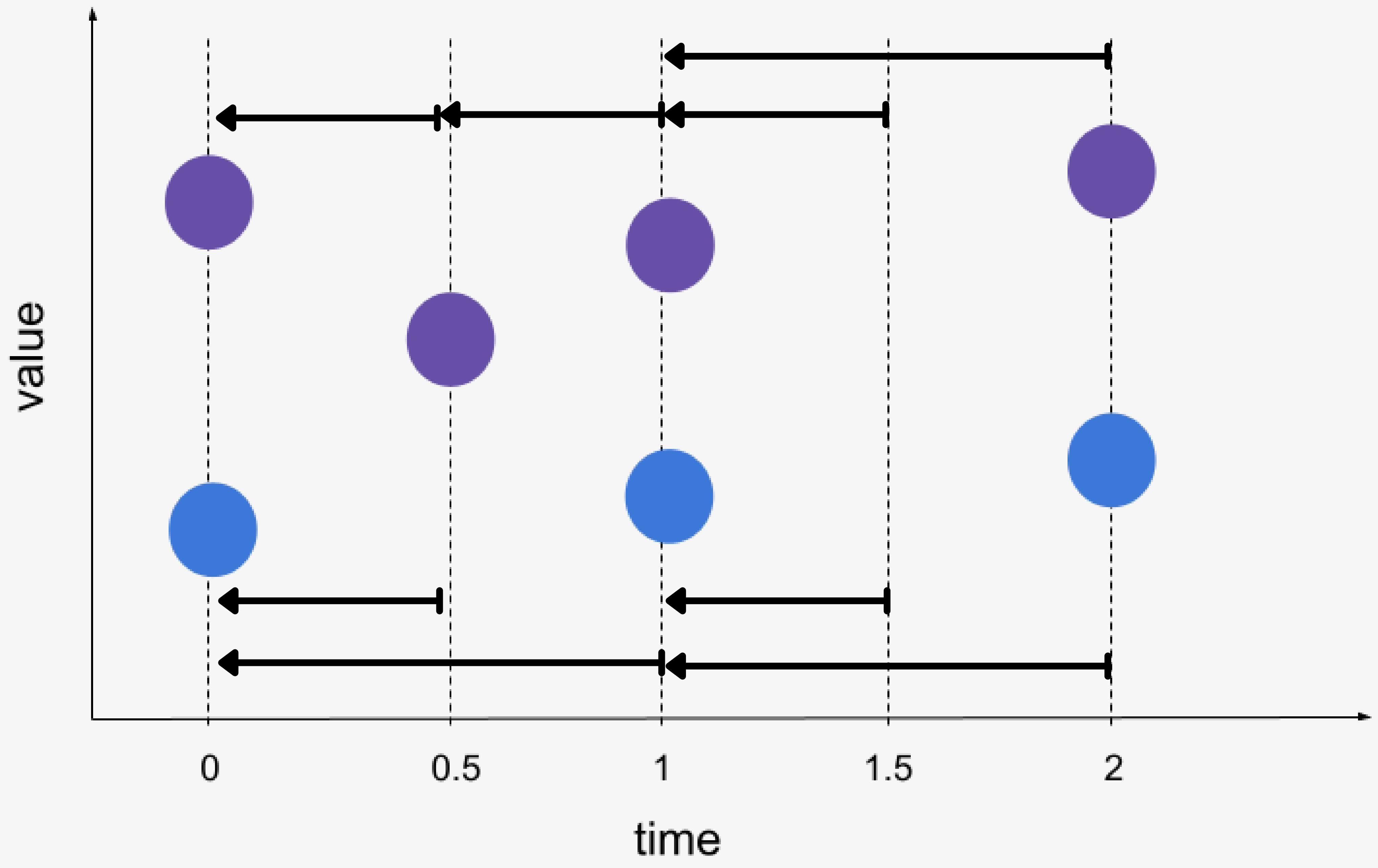


$$\mathbf{X} = \begin{bmatrix} 12 & 10 & 11 & NA & 13 \\ 6 & NA & 7 & NA & 8 \end{bmatrix}$$

$$\mathbf{s} = \begin{bmatrix} 0 & 0.5 & 1.0 & 1.5 & 2.0 \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Time Interval



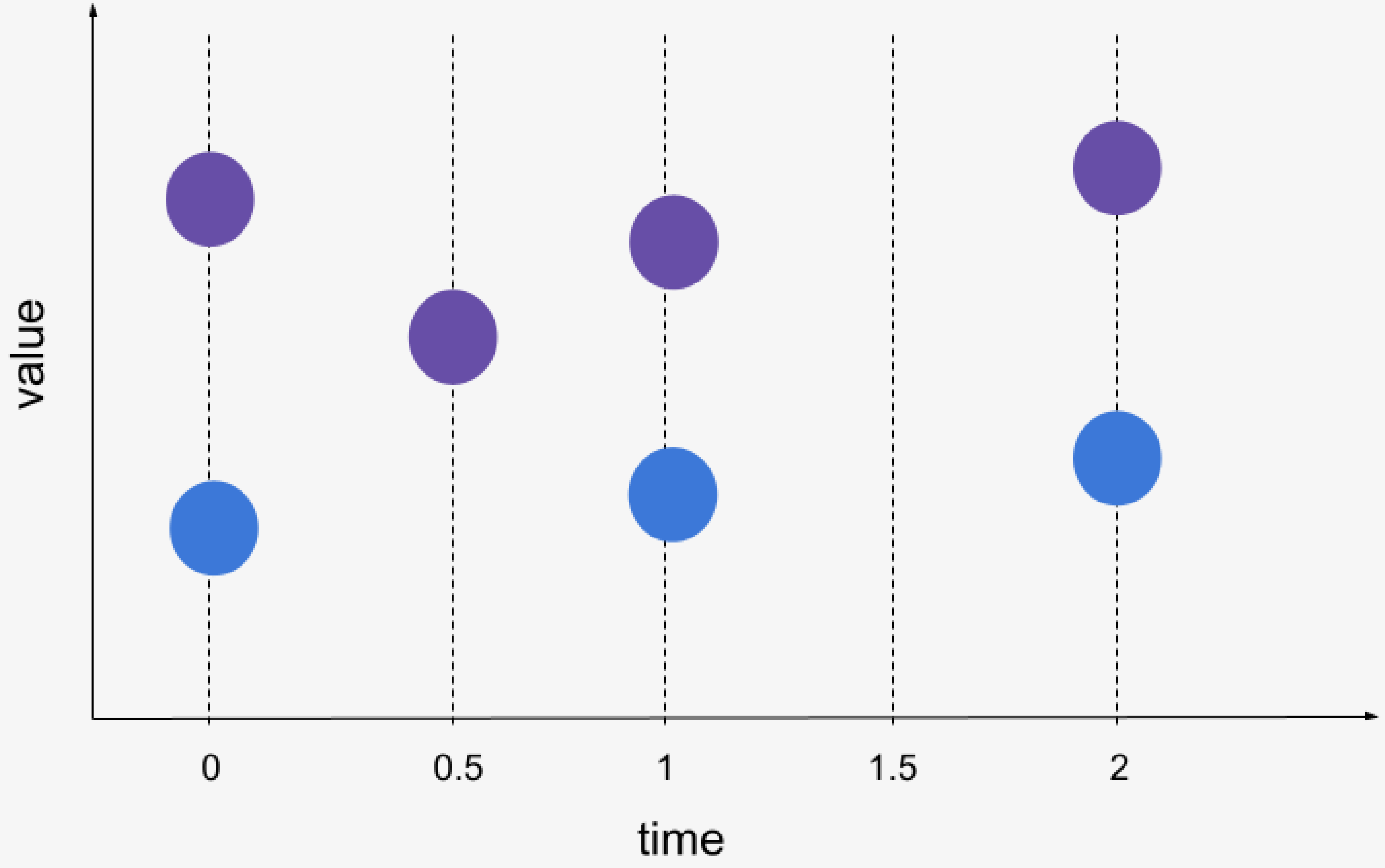
$$\mathbf{X} = \begin{bmatrix} 12 & 10 & 11 & NA & 13 \\ 6 & NA & 7 & NA & 8 \end{bmatrix}$$

$$\mathbf{s} = \begin{bmatrix} 0 & 0.5 & 1.0 & 1.5 & 2.0 \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\delta_t^d = \begin{cases} s_t - s_{t-1} + \delta_{t-1}^d, & t > 1,\ m_{t-1}^d = 0 \\ s_t - s_{t-1}, & t > 1,\ m_{t-1}^d = 1 \\ 0, & t = 1 \end{cases}$$

Time Interval



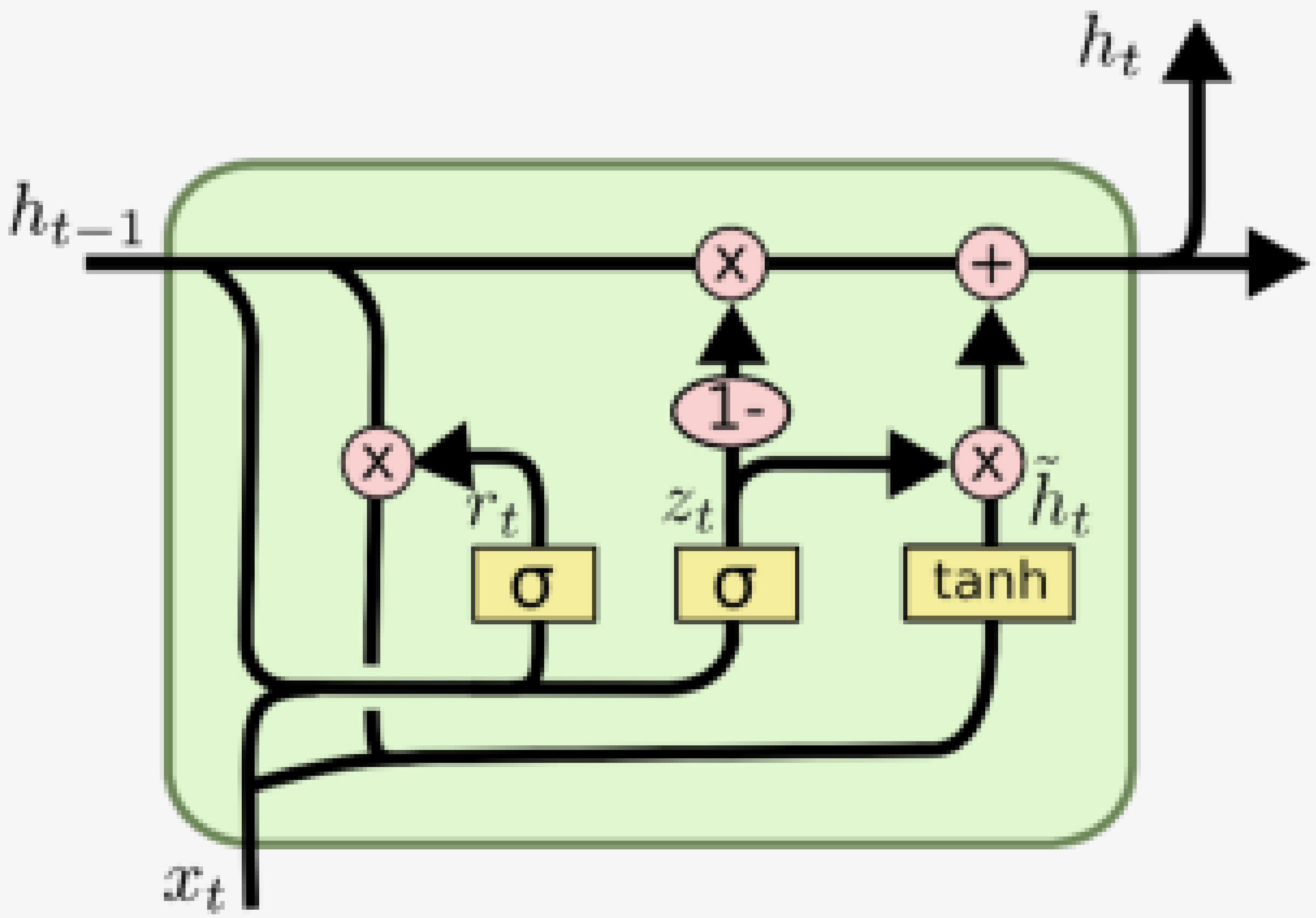
$$\mathbf{X} = \begin{bmatrix} 12 & 10 & 11 & NA & 13 \\ 6 & NA & 7 & NA & 8 \end{bmatrix}$$

$$\mathbf{s} = \begin{bmatrix} 0 & 0.5 & 1.0 & 1.5 & 2.0 \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{\Delta} = \begin{bmatrix} 0.0 & 0.5 & 0.5 & 0.5 & 1.0 \\ 0.0 & 0.5 & 1.0 & 0.5 & 1.0 \end{bmatrix}$$

GRU



$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

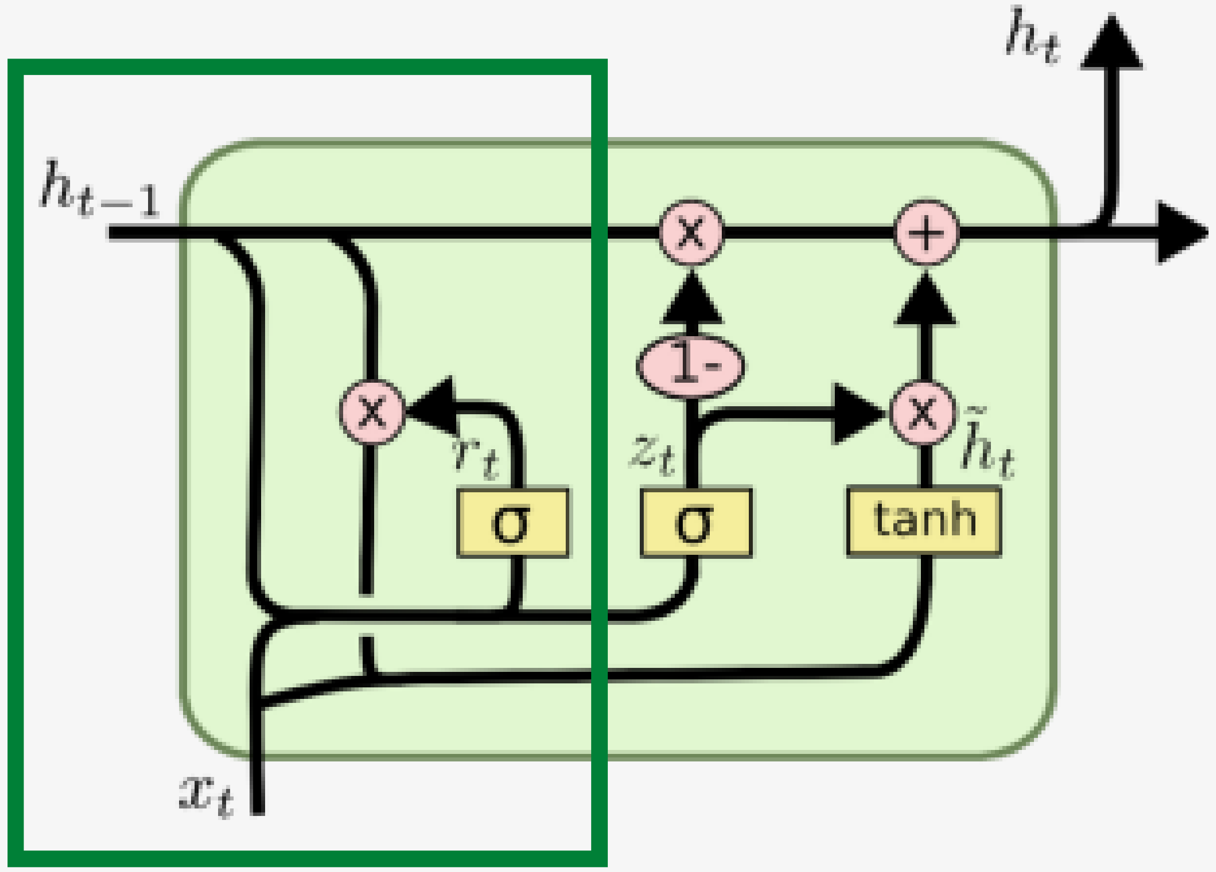
$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

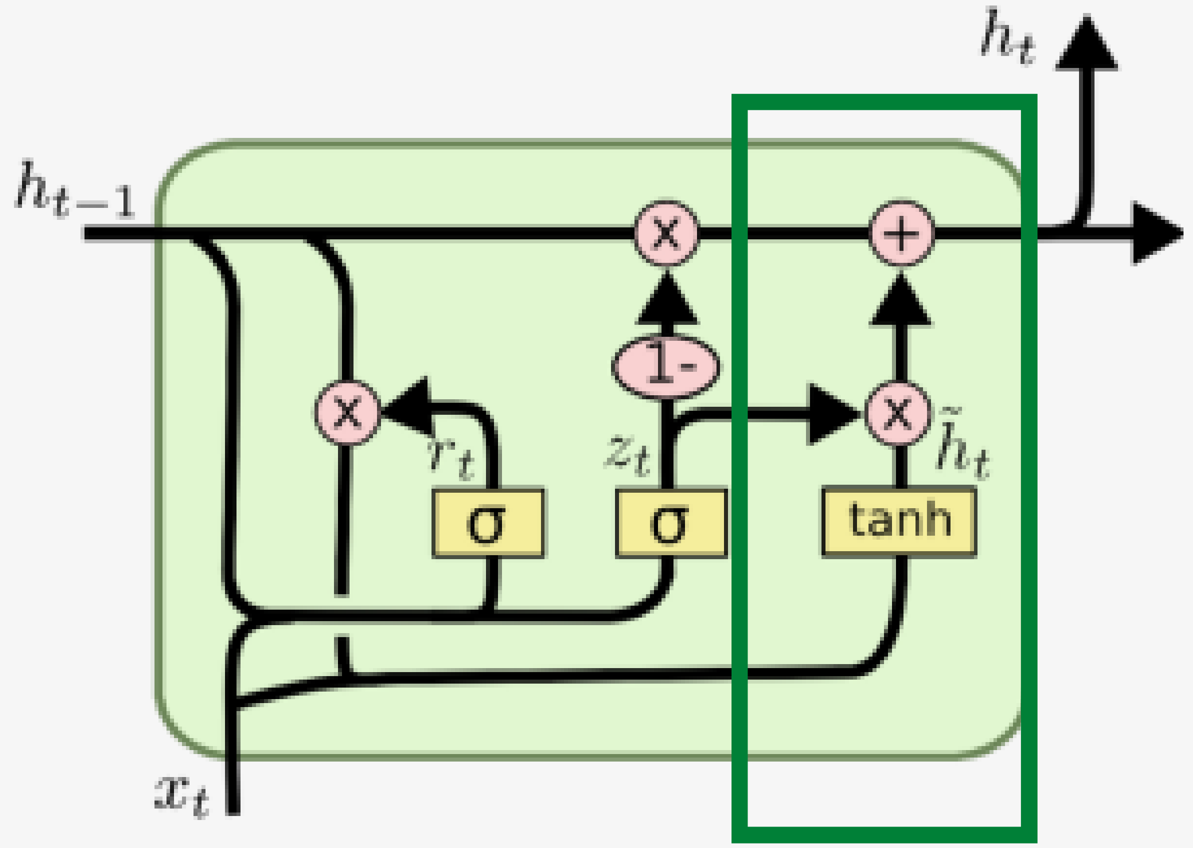
GRU

$$r_t = \sigma(\underline{W_r x_t + U_r h_{t-1} + b_r)}$$ RESET GATE



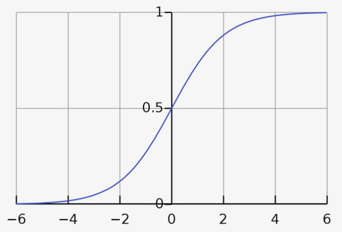
Adapted from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

GRU



$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

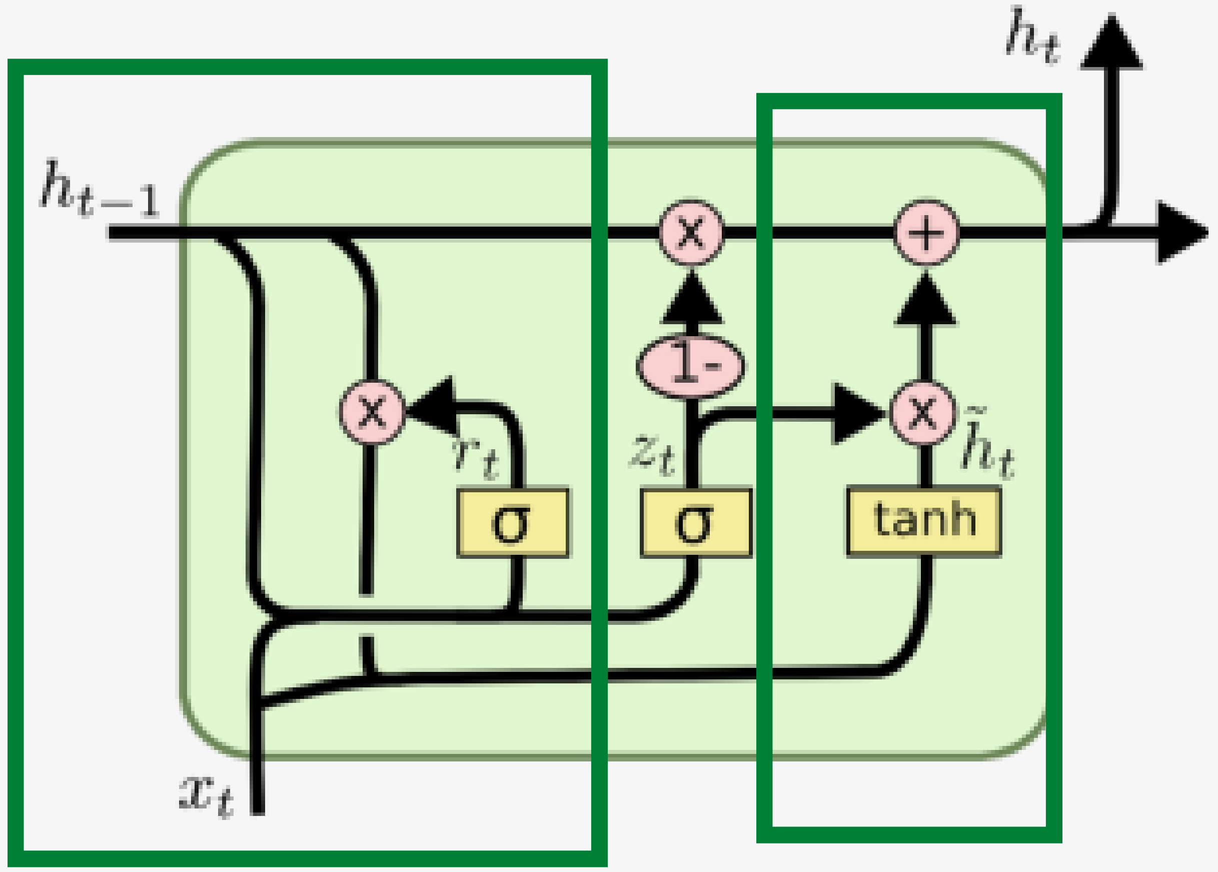
RESET GATE



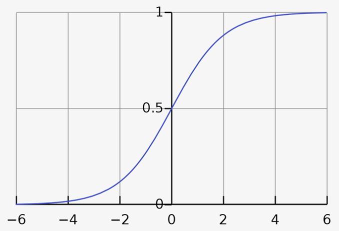
$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b)$$

CANDIDATE HIDDEN STATE

GRU



$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$ RESET GATE

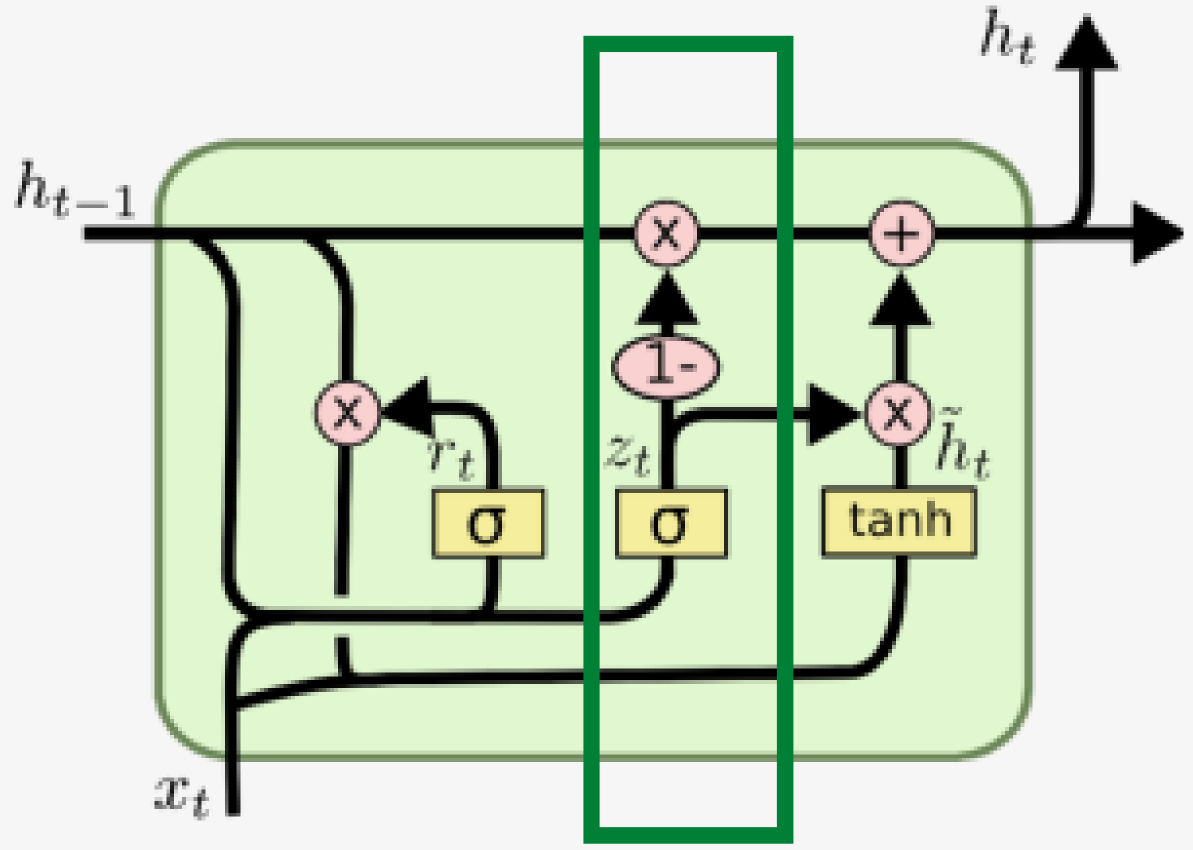


$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b)$$ CANDIDATE HIDDEN STATE

"How much do we want to remember?"

Adapted from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

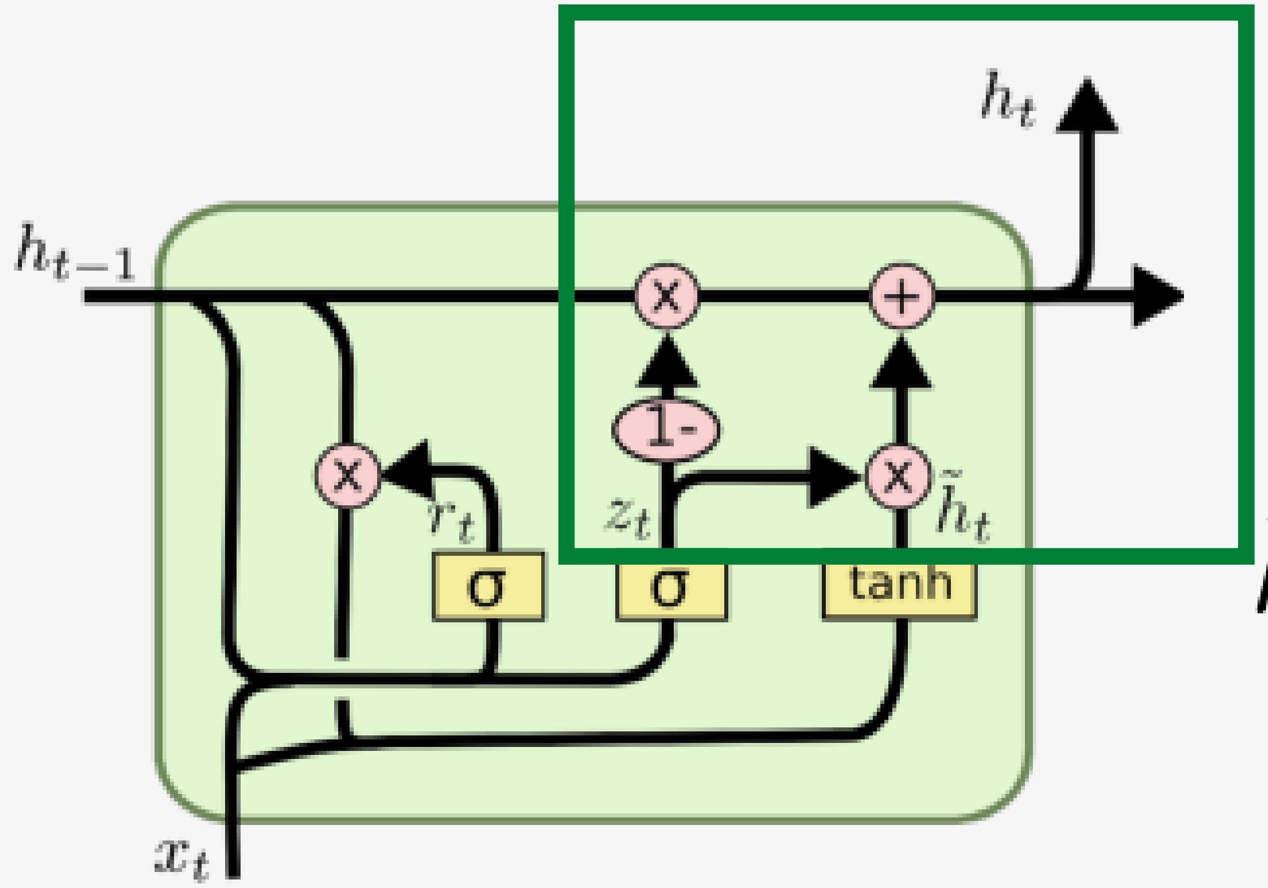
GRU



$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$ UPDATE GATE

Adapted from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

GRU

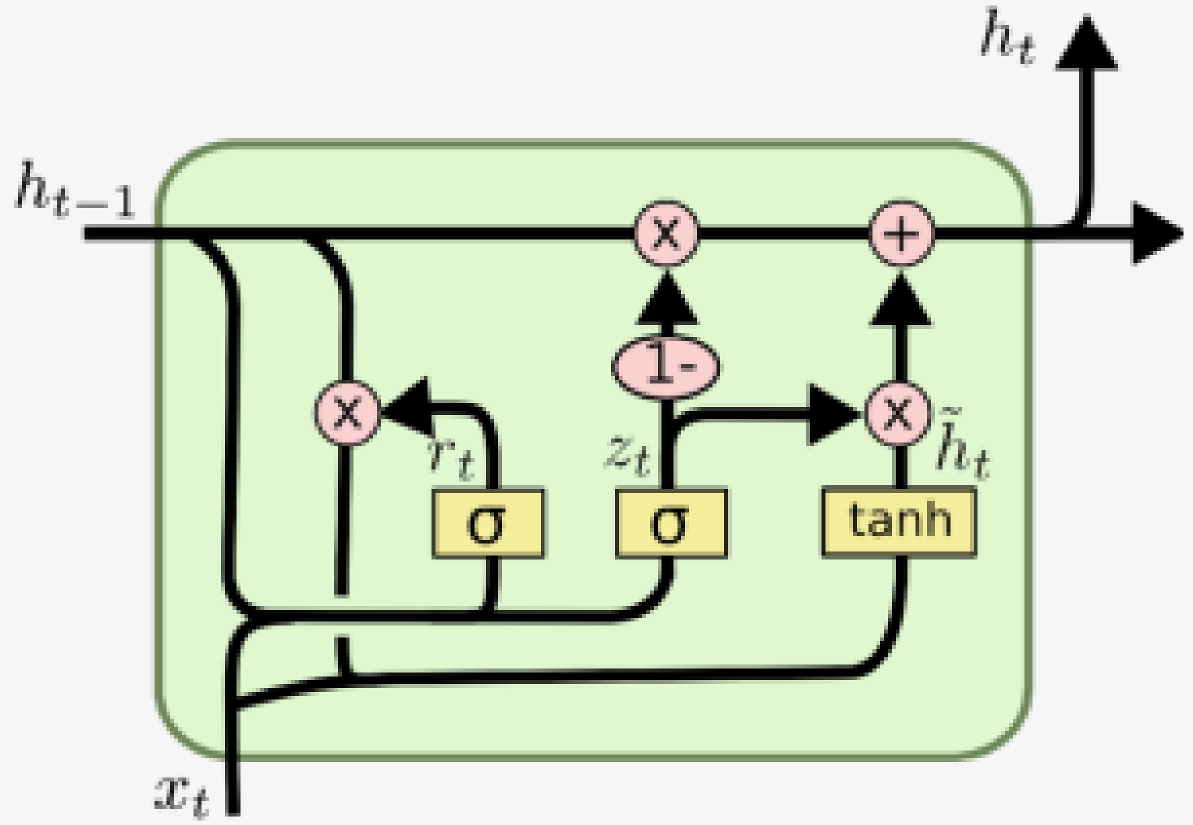


$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$ UPDATE GATE

CANDIDATE HIDDEN STATE

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$ HIDDEN STATE

GRU



$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$ UPDATE GATE

CANDIDATE HIDDEN STATE

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$ HIDDEN STATE

Tradeoff between old hidden state and candidate hidden state

Adapted from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

GRU - D

- Input Decay
- Hidden State Decay
- Mask Vector Inputs

GRU - D

Input Decay

$$\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\}$$

$$\hat{x}_t^d = m_t^d x_t^d + (1 - m_t^d)\left(\gamma_{x_t}^d x_{t'}^d + (1 - \gamma_{x_t}^d)\tilde{x}^d\right)$$

Hidden State Decay

Mask Vector Inputs

GRU - D

- Input Decay
- Hidden State Decay
- Mask Vector Inputs

$$\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\}$$

$$\hat{x}_t^d = m_t^d x_t^d + \boxed{(1 - m_t^d)}(\gamma_{x_t}^d x_{t'}^d + \boxed{(1 - \gamma_{x_t}^d)\tilde{x}^d})$$

GRU - D

Input Decay

$$\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\}$$

$$\hat{x}_t^d = m_t^d x_t^d + (1 - m_t^d)\left(\gamma_{x_t}^d x_{t'}^d + (1 - \gamma_{x_t}^d)\tilde{x}^d\right)$$

Hidden State Decay

$$\hat{h}_{t-1} = \gamma_{h_t} \odot h_{t-1},$$

Mask Vector Inputs

GRU - D

Input Decay

$$\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\}$$

$$\hat{x}_t^d = m_t^d x_t^d + (1 - m_t^d)\left(\gamma_{x_t}^d x_{t'}^d + (1 - \gamma_{x_t}^d)\tilde{x}^d\right)$$

Hidden State Decay

$$\hat{h}_{t-1} = \gamma_{h_t} \odot h_{t-1},$$

Mask Vector Inputs

[0 0 0 1 1 1 0 0 0]

GRU Updates

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

GRU-D Updates

$$r_t = \sigma(W_r \hat{x}_t + U_r \hat{h}_{t-1} + V_r m_t + b_r)$$

$$z_t = \sigma(W_z \hat{x}_t + U_z \hat{h}_{t-1} + V_z m_t + b_z)$$

$$\tilde{h}_t = \tanh(W \hat{x}_t + U(r_t \odot \hat{h}_{t-1}) + V m_t + b)$$

$$h_t = (1 - z_t) \odot \hat{h}_{t-1} + z_t \odot \tilde{h}_t$$

GRU Updates

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

GRU-D Updates

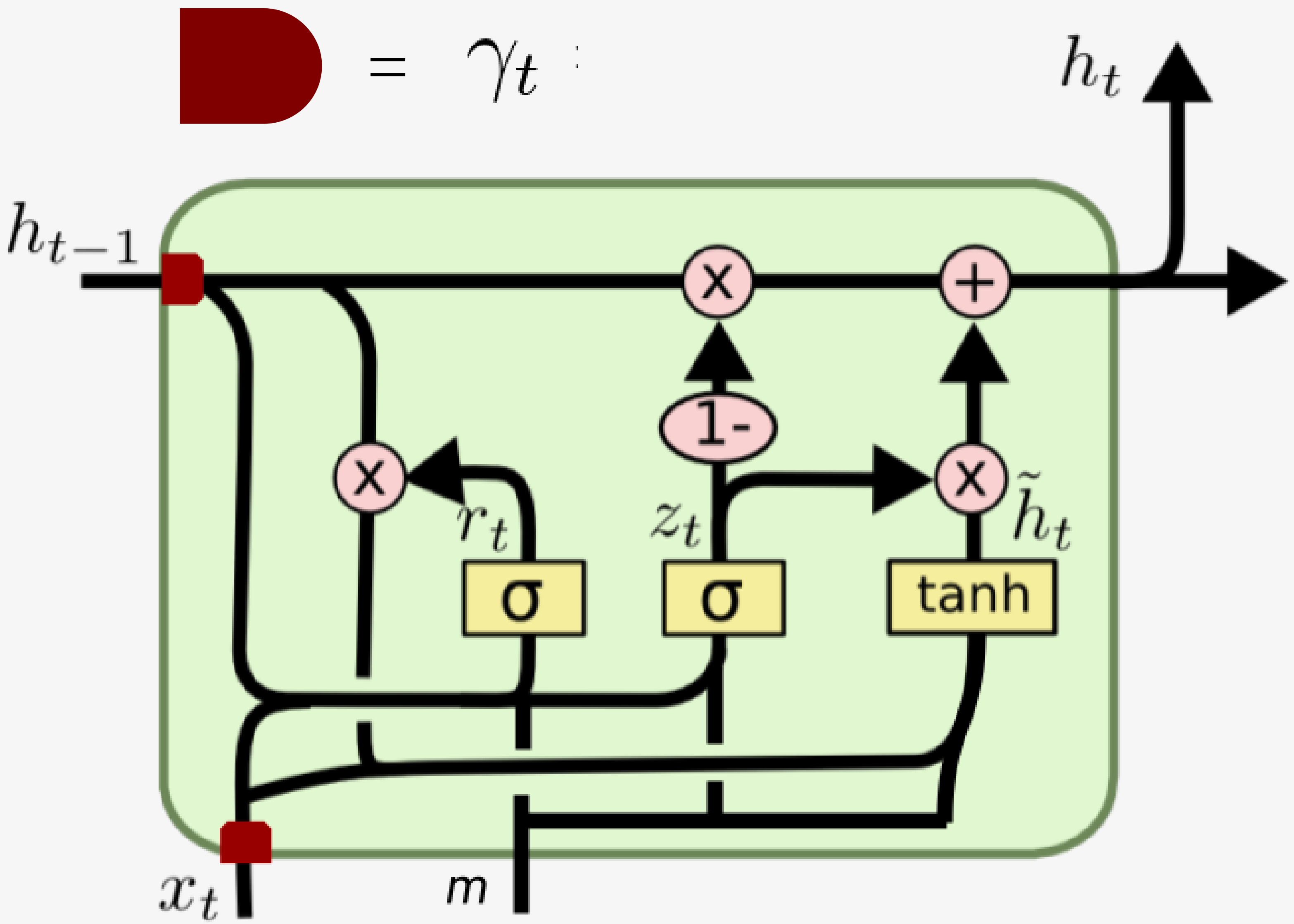
$$r_t = \sigma(W_r \hat{x}_t + U_r \hat{h}_{t-1} + V_r m_t + b_r)$$

$$z_t = \sigma(W_z \hat{x}_t + U_z \hat{h}_{t-1} + V_z m_t + b_z)$$

$$\tilde{h}_t = \tanh(W \hat{x}_t + U(r_t \odot \hat{h}_{t-1}) + V m_t + b)$$

$$h_t = (1 - z_t) \odot \hat{h}_{t-1} + z_t \odot \tilde{h}_t$$

Input Decay **Hidden State Decay** **Mask Vector Inputs**



GRU-D Updates

$$r_t = \sigma(W_r\hat{x}_t + U_r\hat{h}_{t-1} + V_r m_t + b_r)$$

$$z_t = \sigma(W_z\hat{x}_t + U_z\hat{h}_{t-1} + V_z m_t + b_z)$$

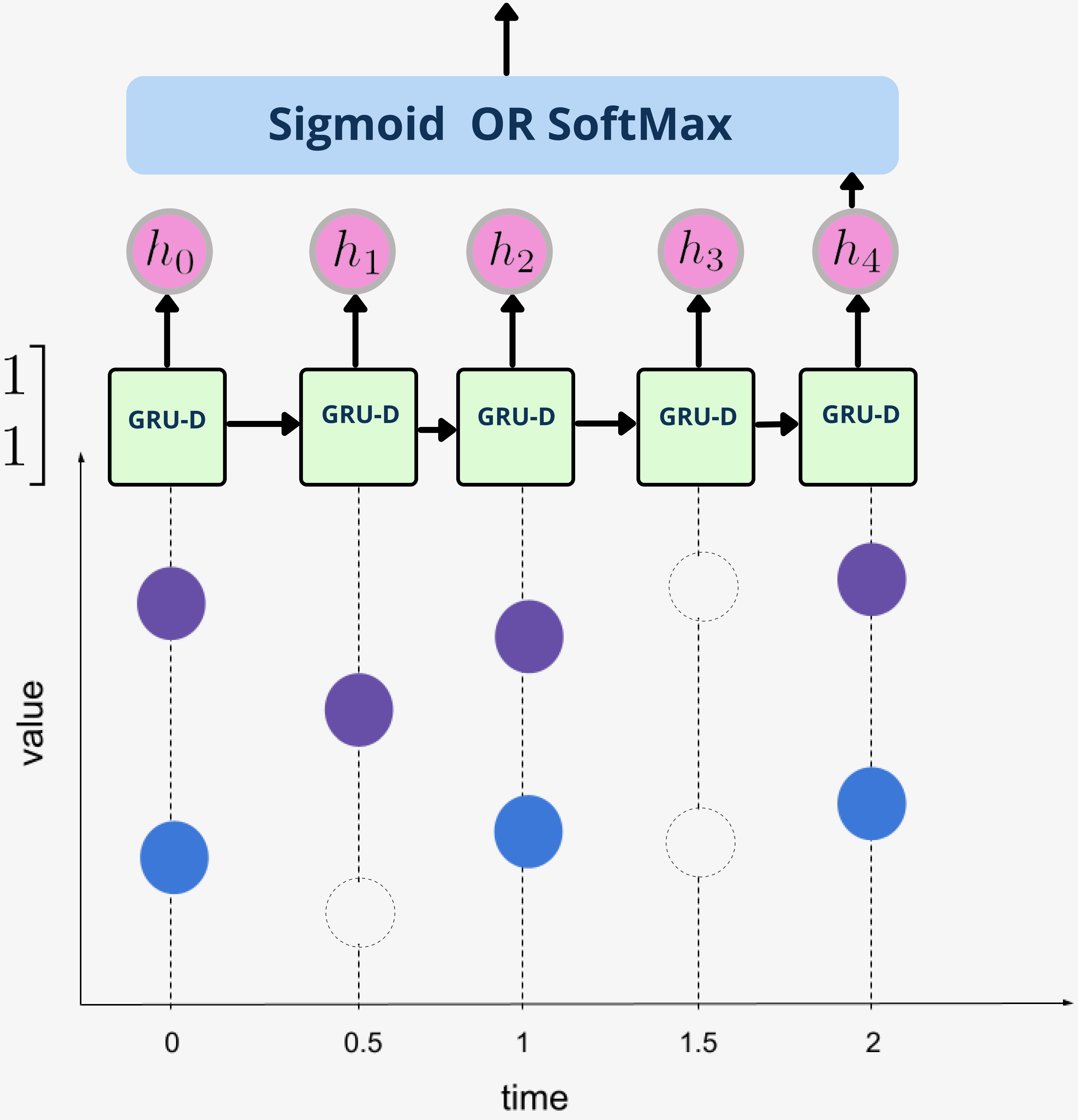
$$\tilde{h}_t = \tanh(W\hat{x}_t + U(r_t \odot \hat{h}_{t-1}) + V m_t + b)$$

$$h_t = (1 - z_t) \odot \hat{h}_{t-1} + z_t \odot \tilde{h}_t$$

Input Decay **Hidden State Decay** **Mask Vector Inputs**

Adapted from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$



Non-RNN Baselines	Logistic Regression, SVM, Random Forest
GRU - Mean	Missing values filled with feature mean
GRU - Forward	Missing values filled with previous observed value
GRU - Simple	Missingness masks and time since last observation passed as feature inputs
GRU - D	Proposed model

Gesture Phase Segmentation

Multivariate time series with synthetic missingness introduced - multiclass classification

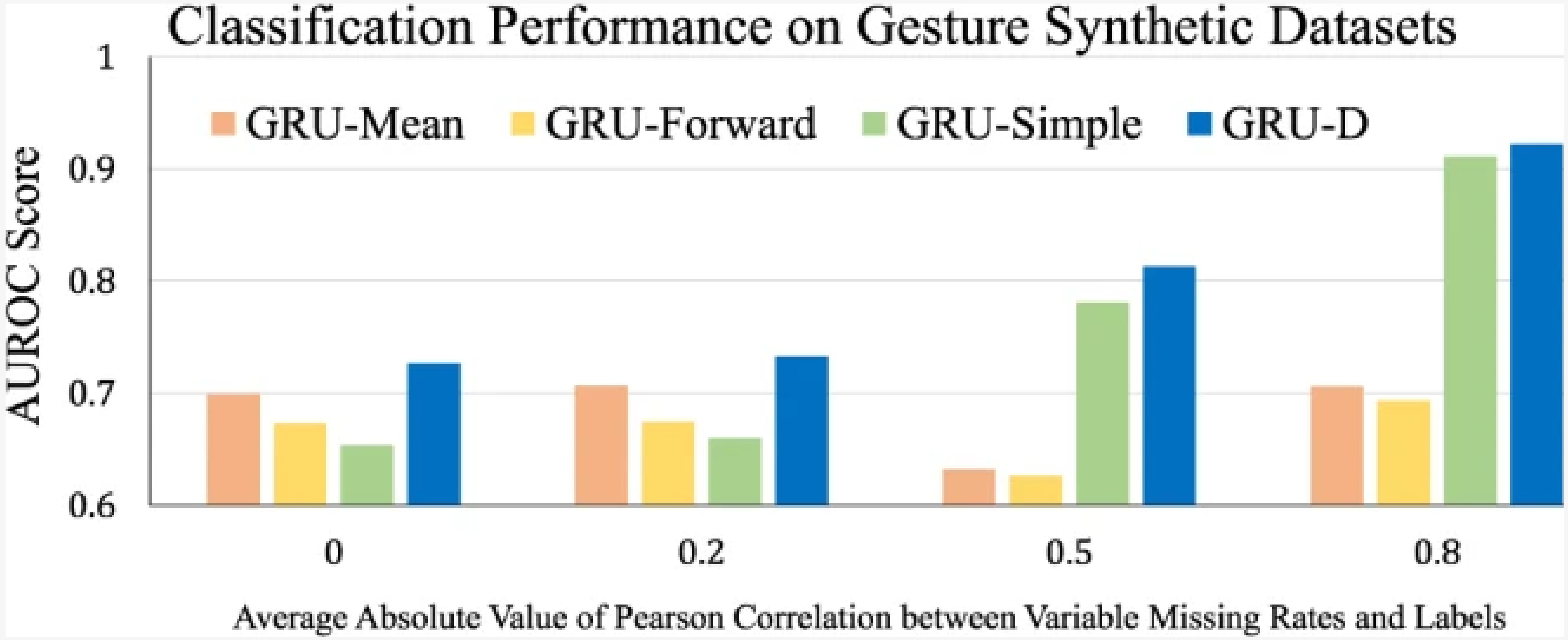
PhysioNet Challenge 2012

Multivariate time series ICU records
*Binary classification (mortality)
*Multiclass classification (4 tasks)

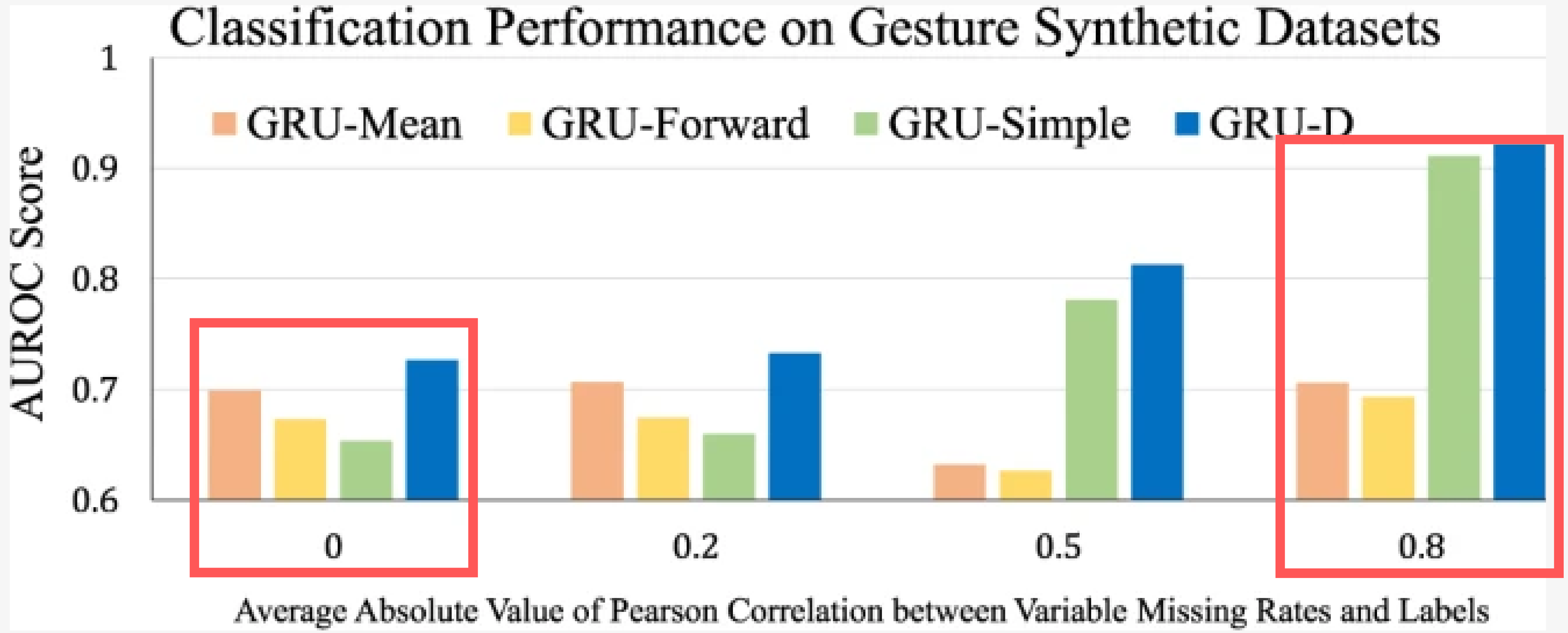
MIMIC III

Multivariate time series ICU records
*Binary classification (mortality)
*Multiclass classification (ICD9 codes, 20 tasks)

Gesture Phase Segmentation



Gesture Phase Segmentation



PhysioNet (Mortality)

MODEL	AUROC
GRU - Mean	0.8252 ± 0.011
GRU - Forward	0.8192 ± 0.013
GRU - Simple	0.8380 ± 0.008
GRU - D	0.8527 ± 0.003

PhysioNet (4 tasks)

MODEL	AUROC
GRU - Mean	0.8099 ± 0.011
GRU - Forward	0.8091 ± 0.008
GRU - Simple	0.8249 ± 0.010
GRU - D	0.8370 ± 0.012

MIMIC (Mortality)

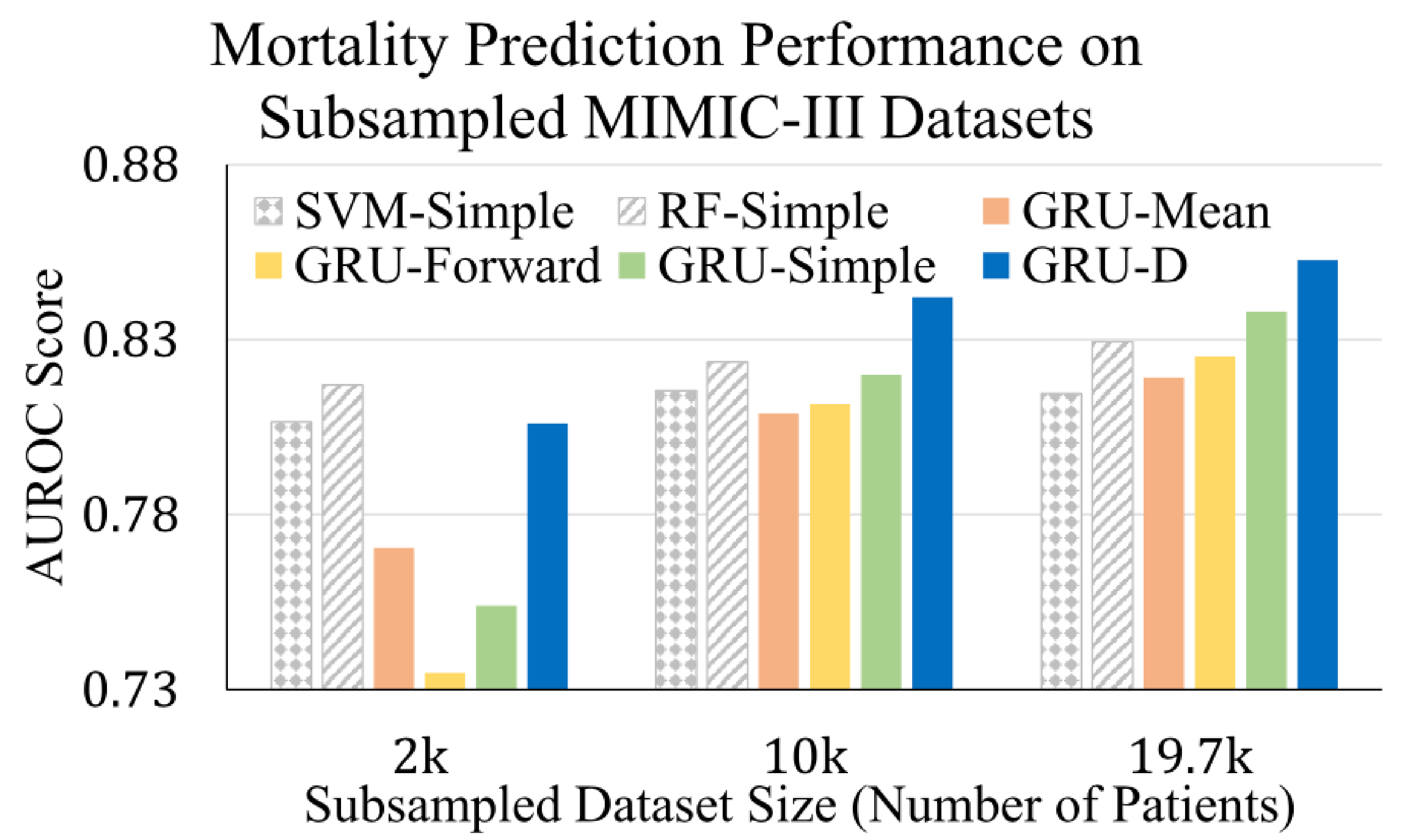
MODEL	AUROC
GRU - Mean	0.8162 ± 0.014
GRU - Forward	0.8195 ± 0.004
GRU - Simple	0.8226 ± 0.010
GRU - D	0.8424 ± 0.012

MIMIC (ICD 9 20 tasks)

MODEL	AUROC
GRU - Mean	0.7070 ± 0.001
GRU - Forward	0.7077 ± 0.001
GRU - Simple	0.7105 ± 0.001
GRU - D	0.7123 ± 0.003

Strengths

- Scalable
- Extensive evaluation
- Solution to data not missing-completely-at-random



(b) Performance for predicting mortality on subsampled datasets.

Limitations

- **Uninformative missingness**
  - No clear inherent correlation between the missingness pattern and prediction task.
- **Decay mechanism**
  - Needs to be explicitly designed for each domain.
  - Should it always decay?
- **Unsupervised settings**
  - Labels are required.

Questions?