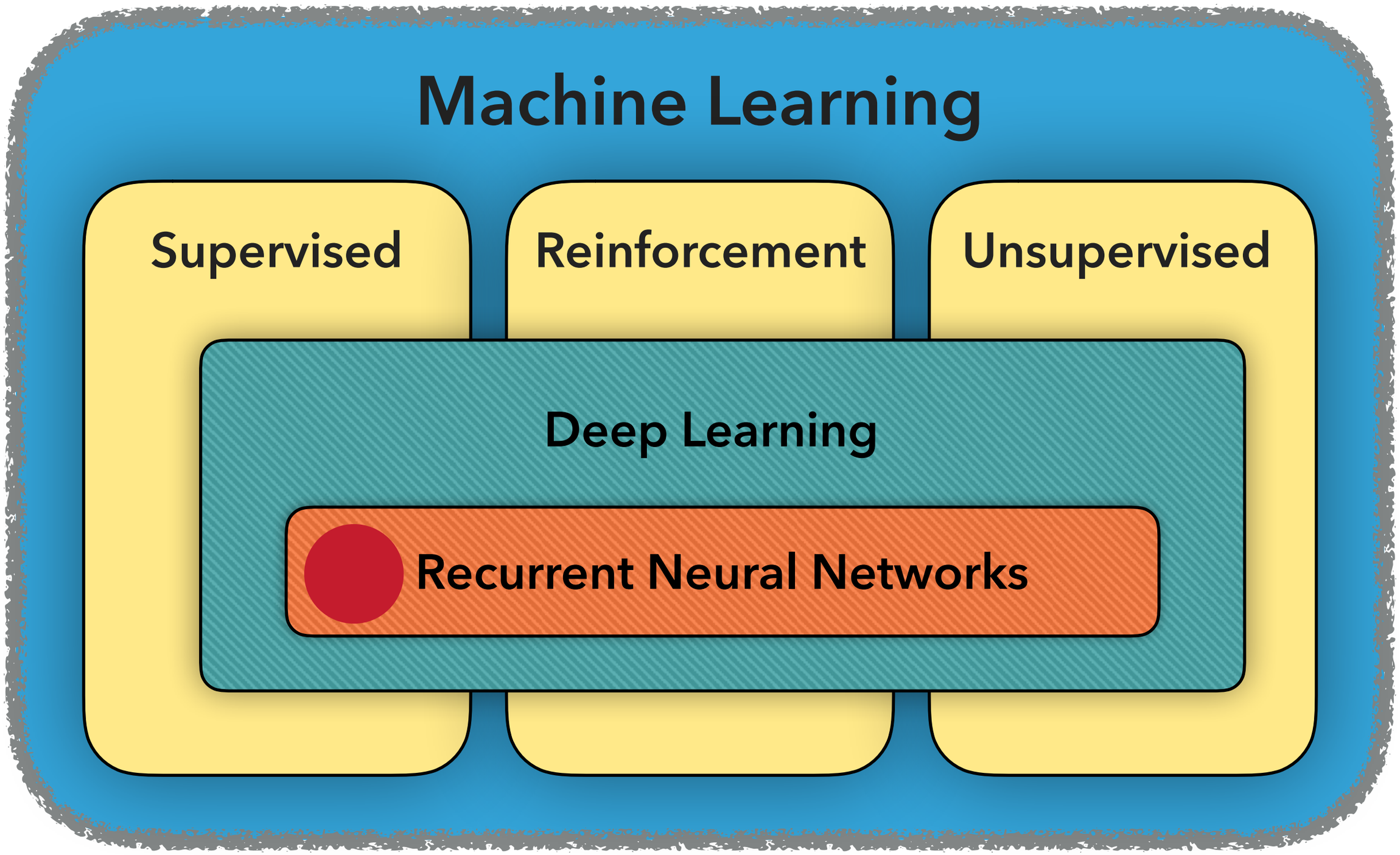
JOHANNES KILIAN LANGER

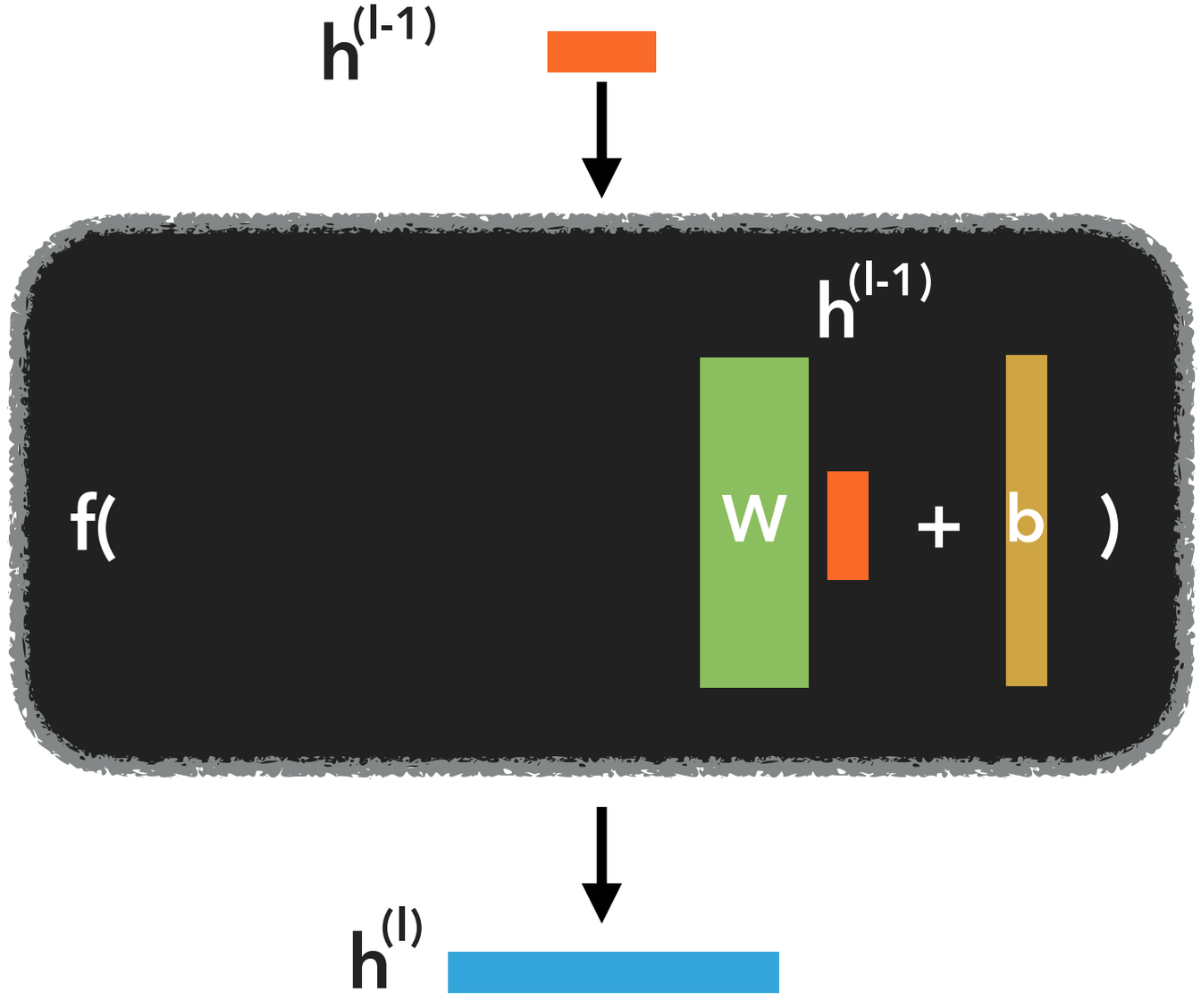
MASTER THESIS DISPUTATION

TABLE OF CONTENTS


Machine Learning
Supervised
Reinforcement
Unsupervised
Deep Learning
Recurrent Neural Networks

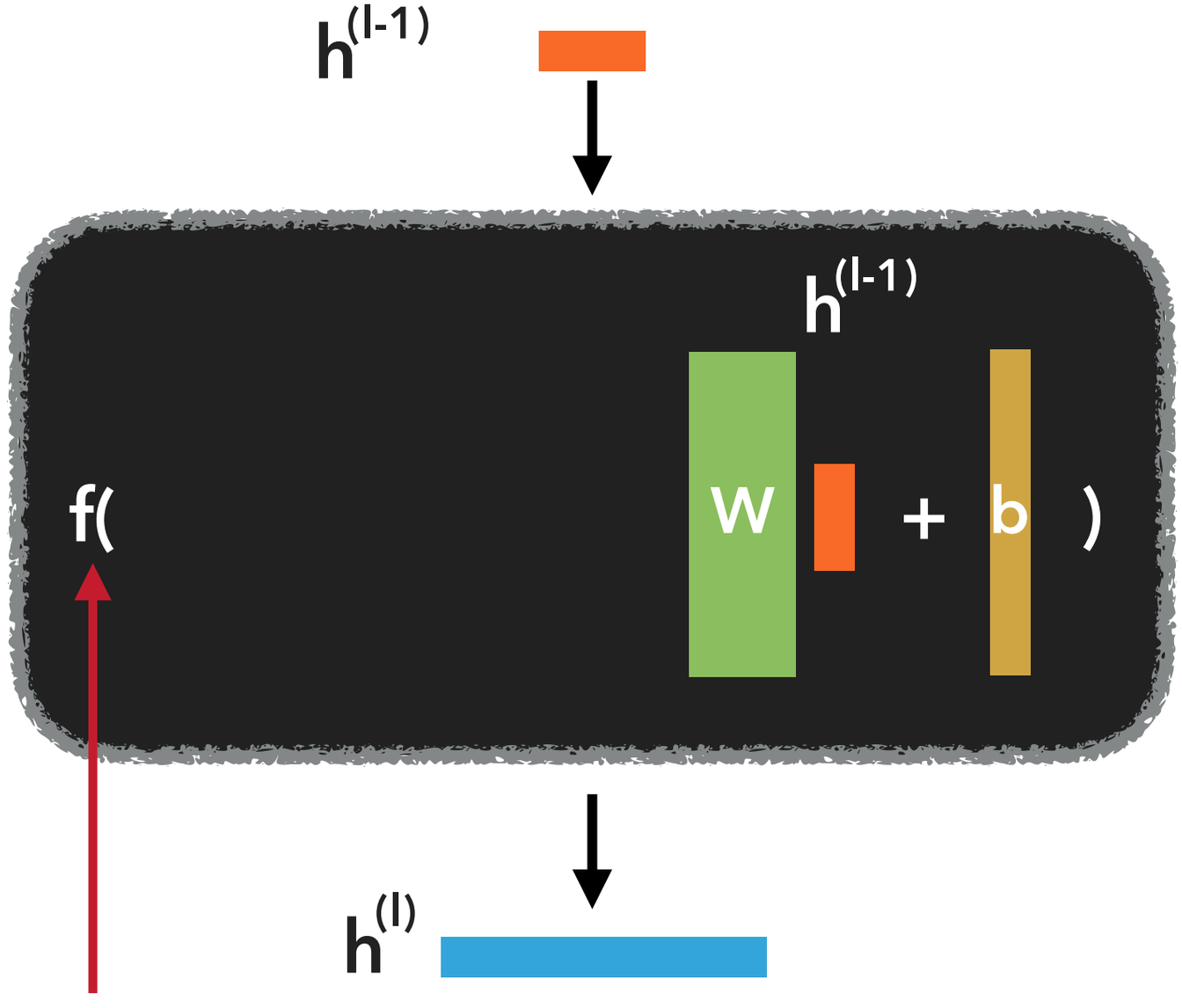
FF-LAYER

$$h^{(l)} = f(Wh^{(l-1)} + b)$$

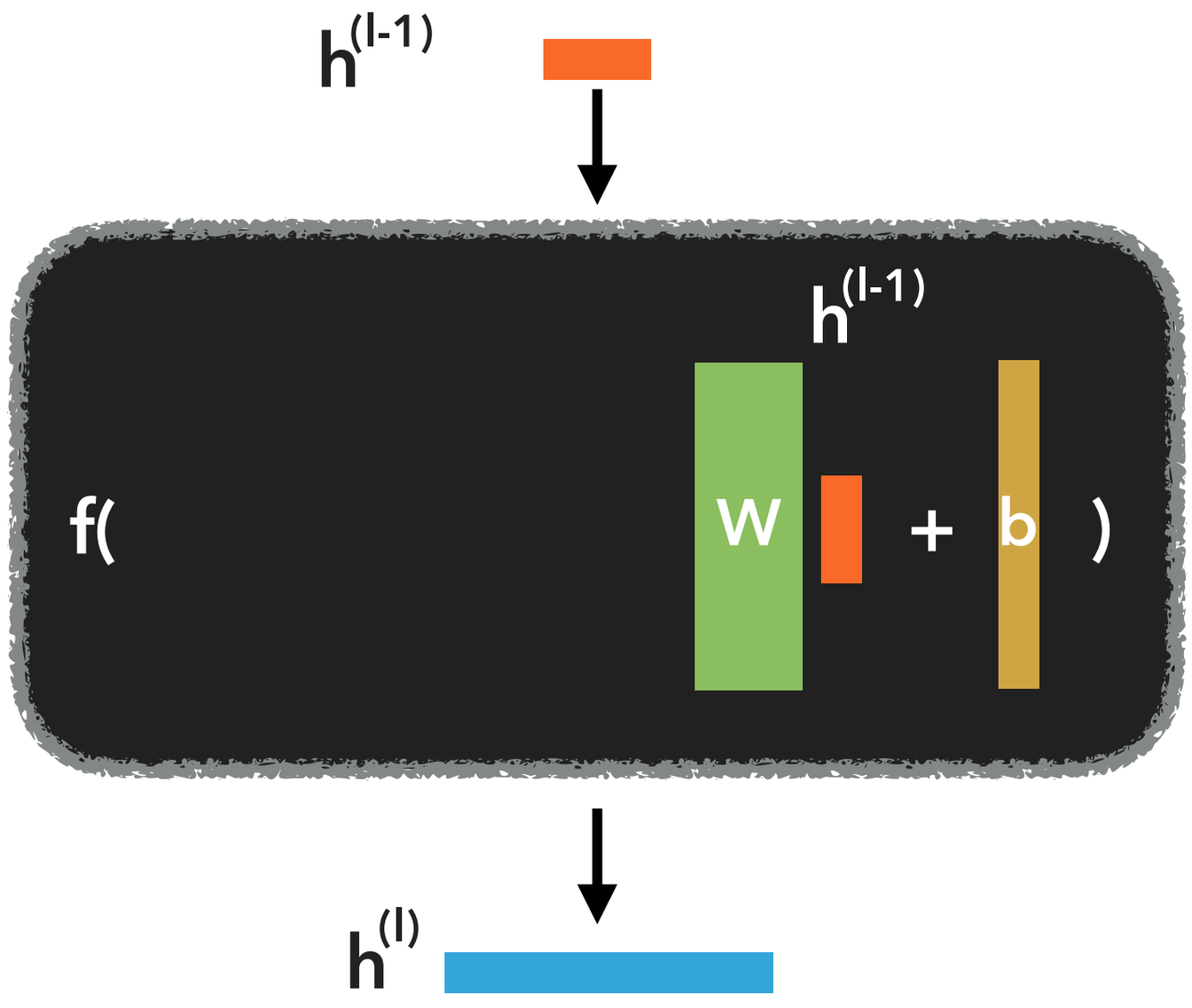


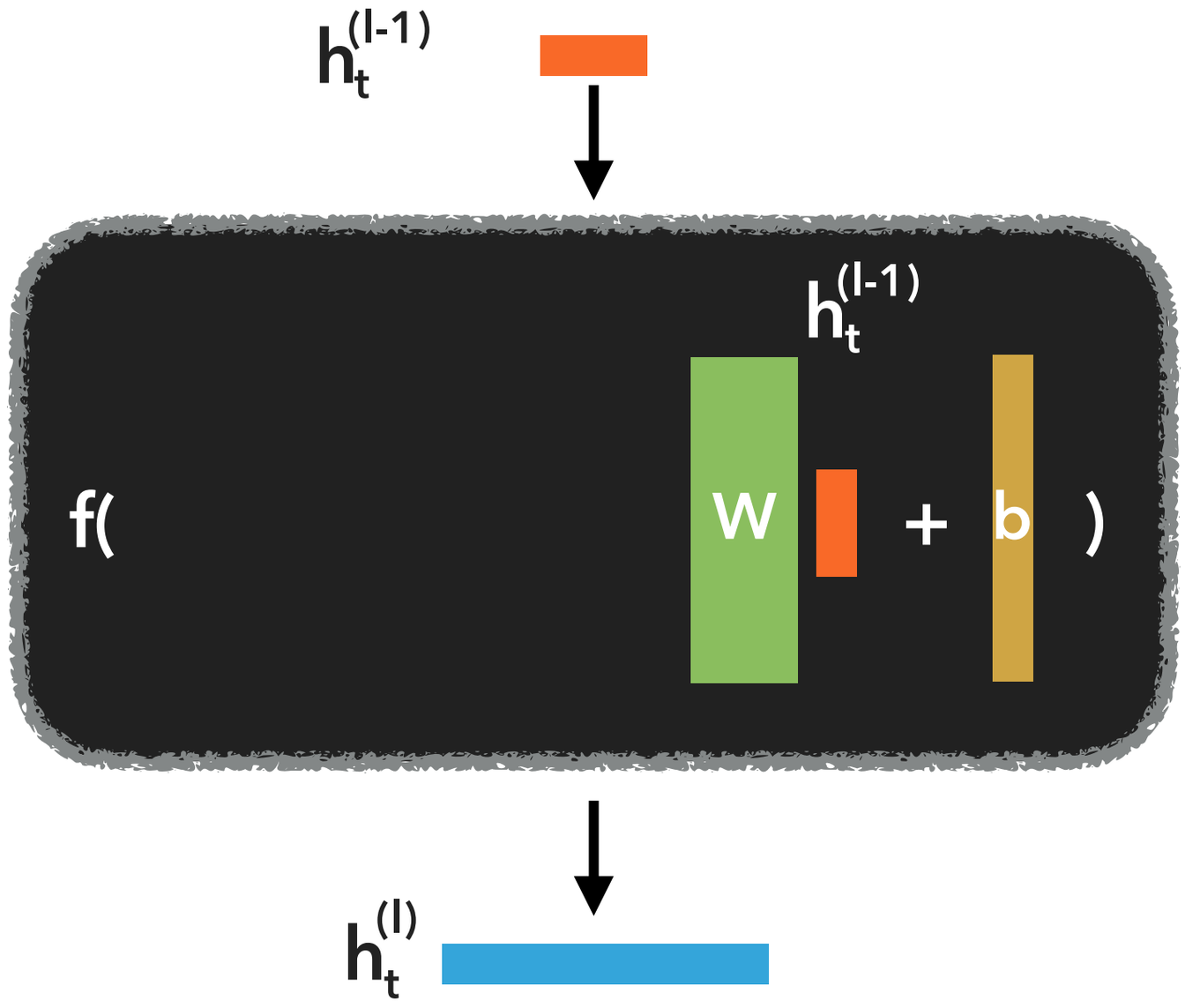
FF-LAYER

$$h^{(l)} = f(Wh^{(l-1)} + b)$$

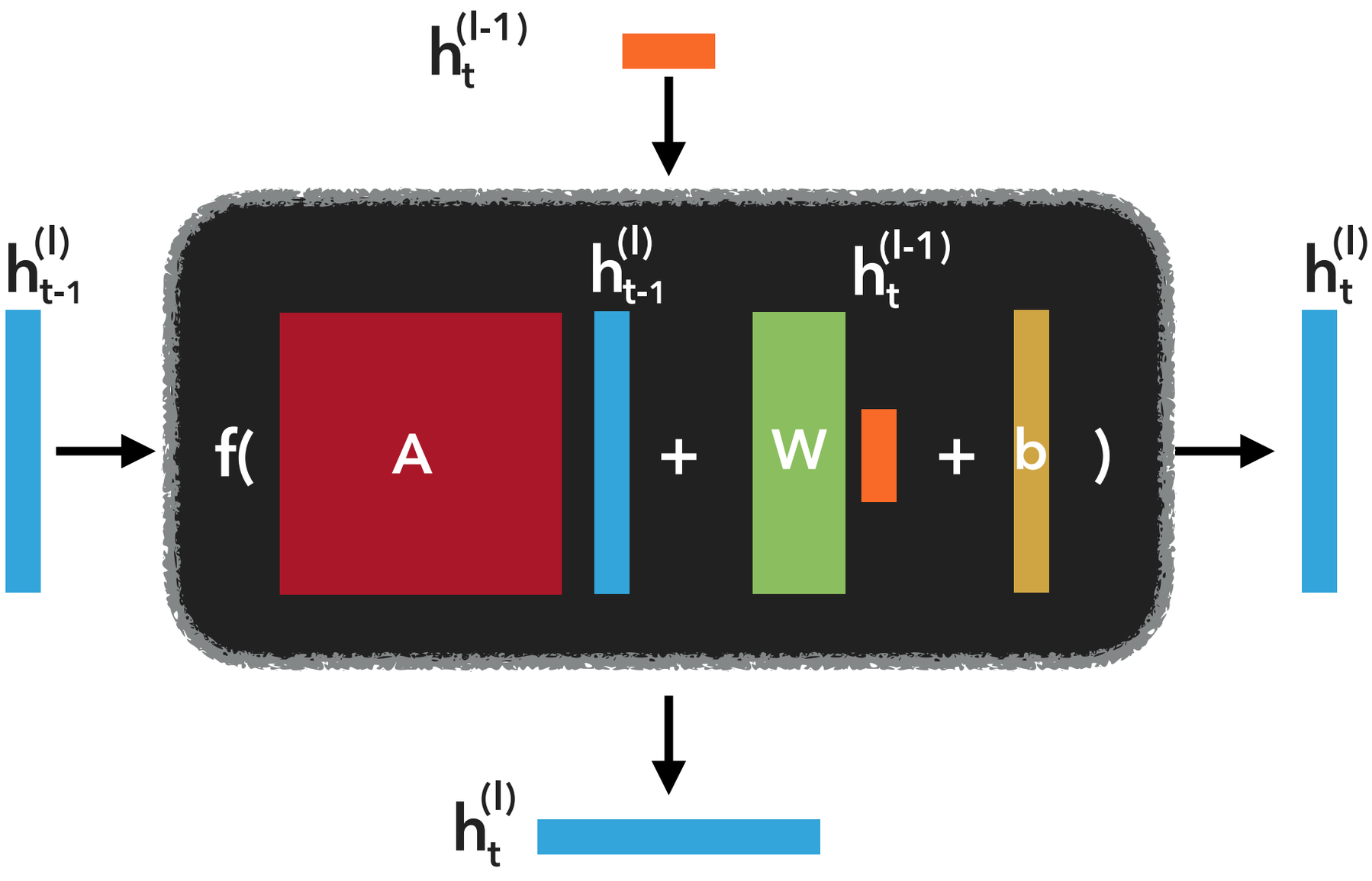


non-linear function required to achieve universal approximator property


h(l-1)
h(l-1)
f(
W
+
b
)
h(l)

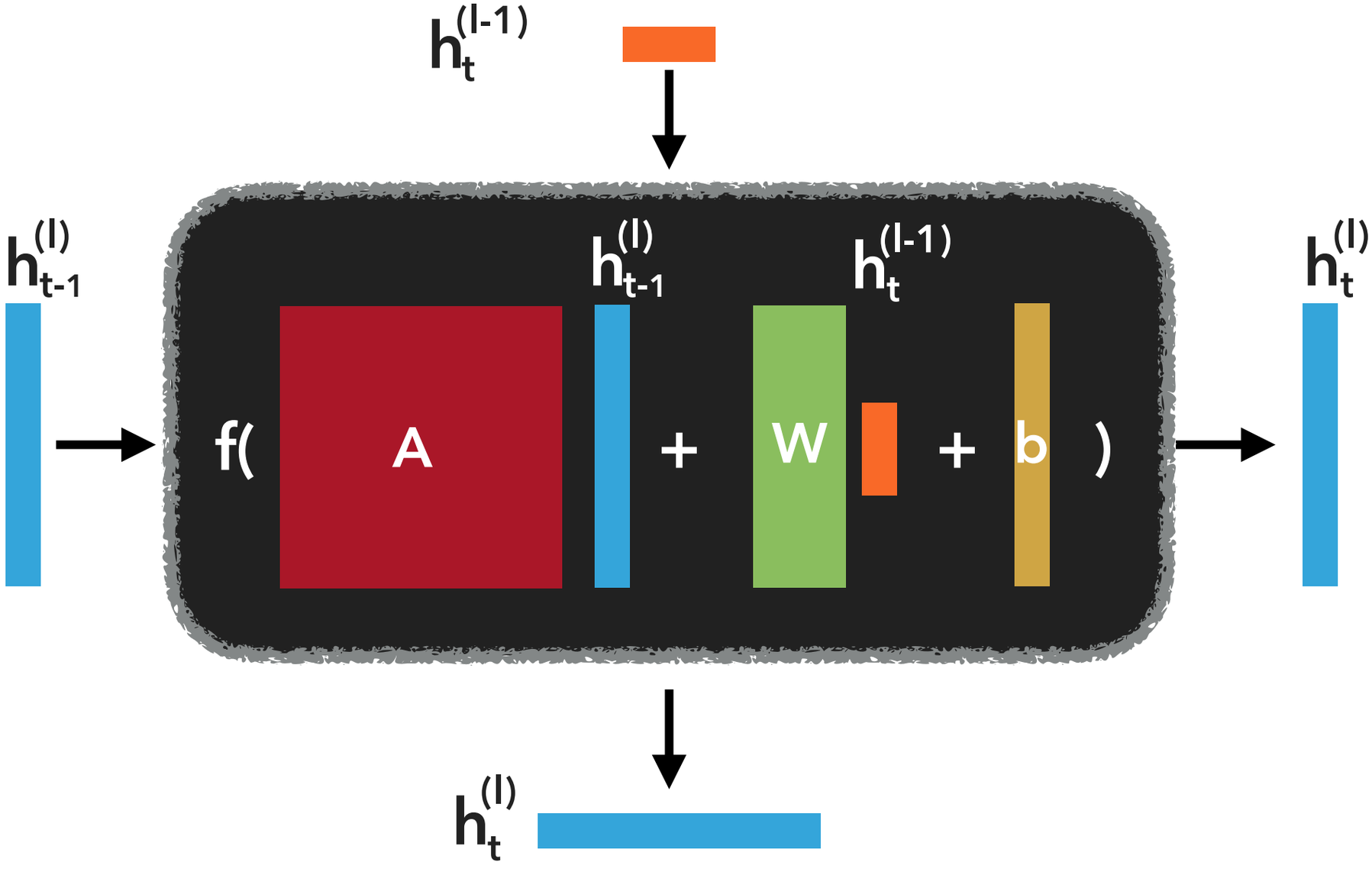

$h_t^{(l-1)}$
$h_t^{(l-1)}$
f(
W
+
b
)
$h_t^{(l)}$

RECURRENT NEURAL NETWORKS

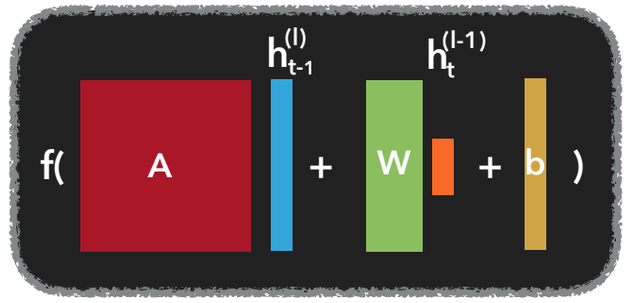


RNN-CELL

$$h_t^{(l)} = f(Ah_{t-1}^{(l)} + Wh_t^{(l-1)} + b)$$



VANISHING AND EXPLODING GRADIENT

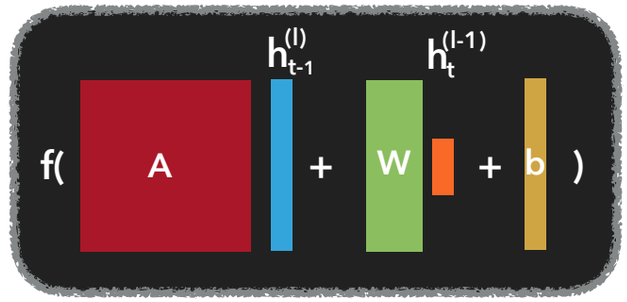


$$h_t^{(l)} = f(Ah_{t-1}^{(l)} + Wh_t^{(l-1)} + b)$$

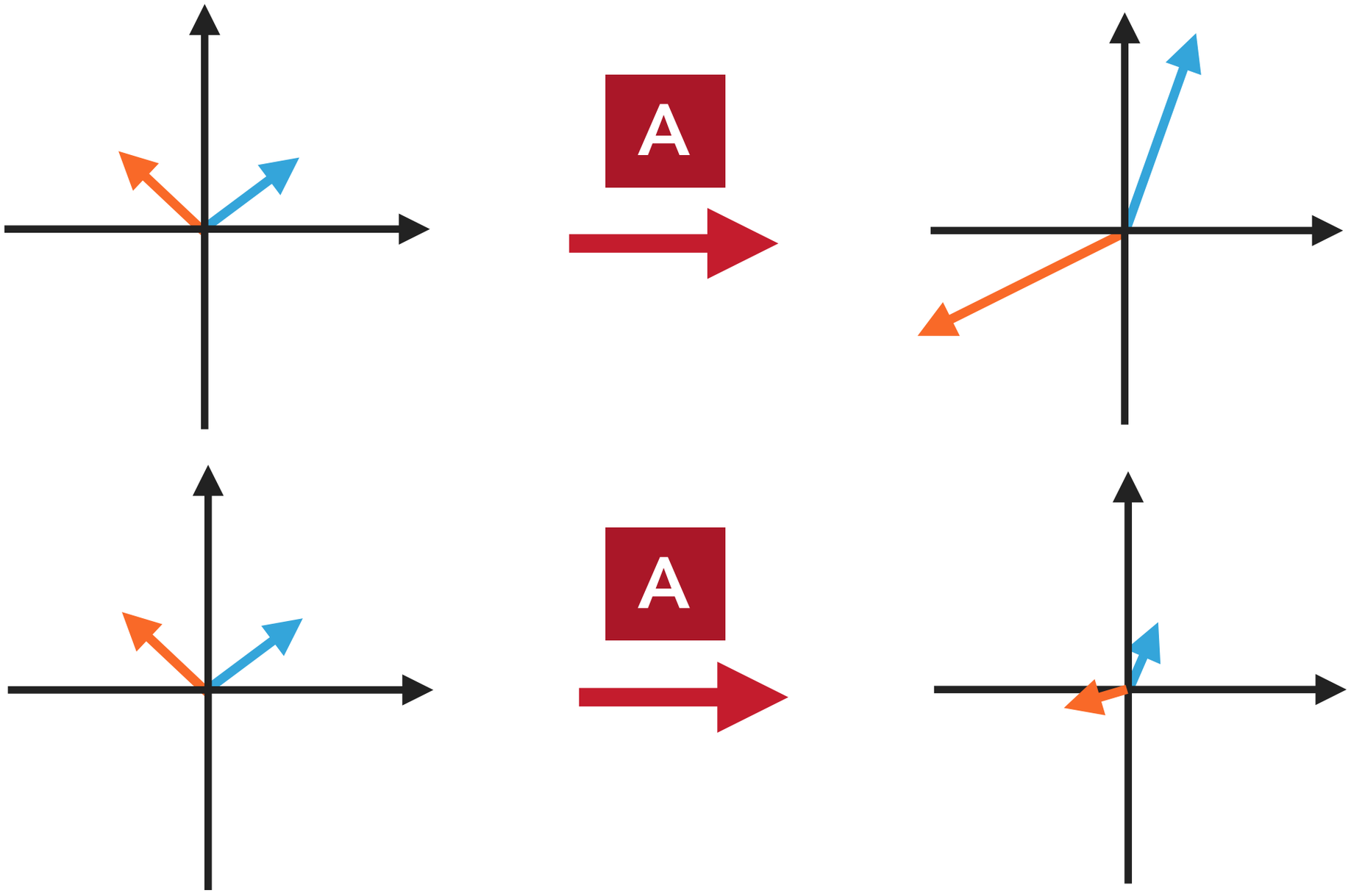
backpropagation

$$\frac{\partial h_T^{(l)}}{\partial A_{a,b}} = \ldots + \ldots A^t \frac{\partial h_{T-t}^{(l)}}{\partial A_{a,b}}$$

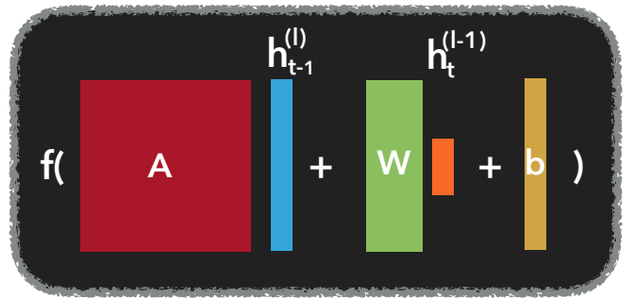
VANISHING AND EXPLODING GRADIENT



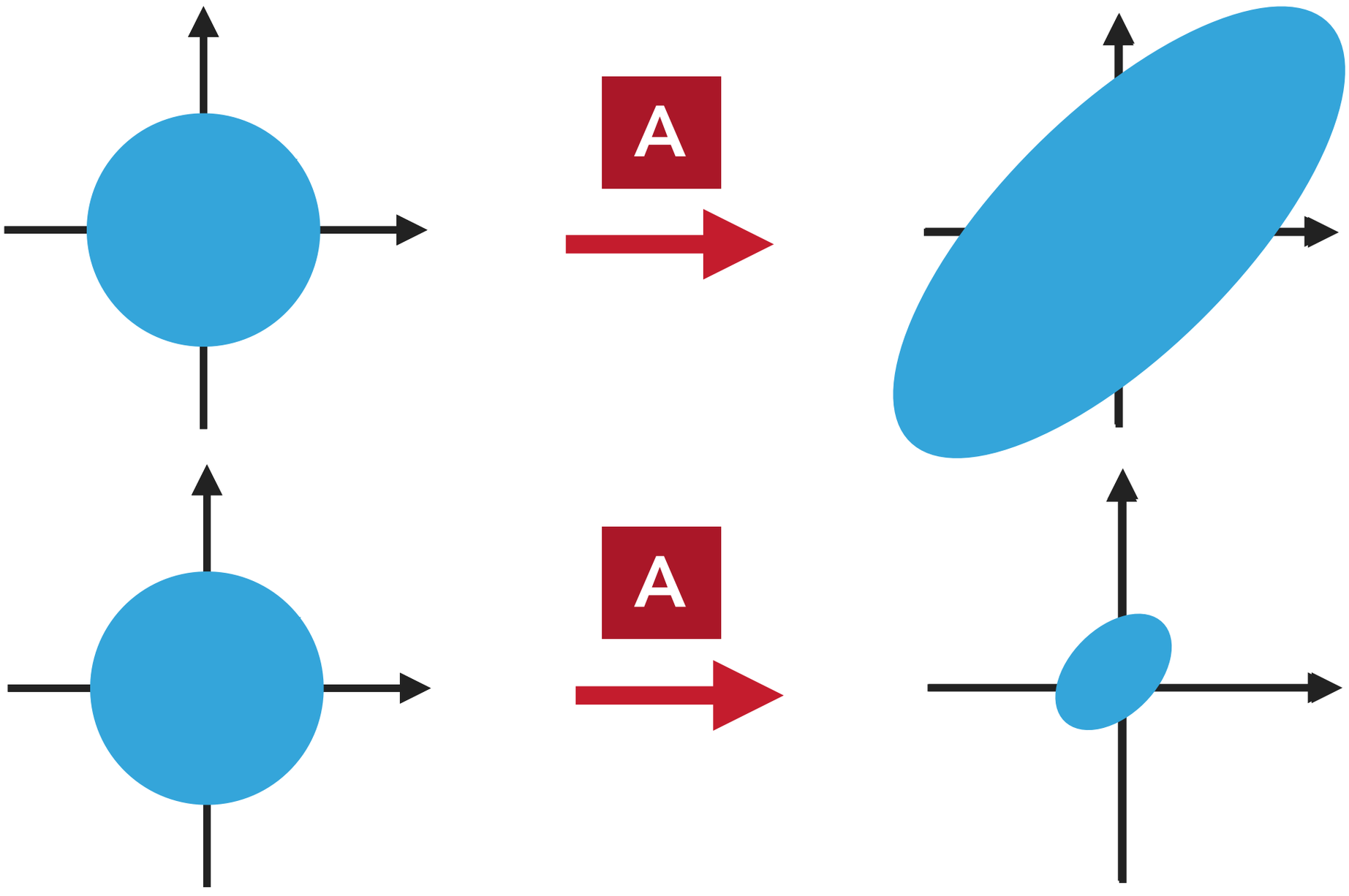
Matrix Multiplication



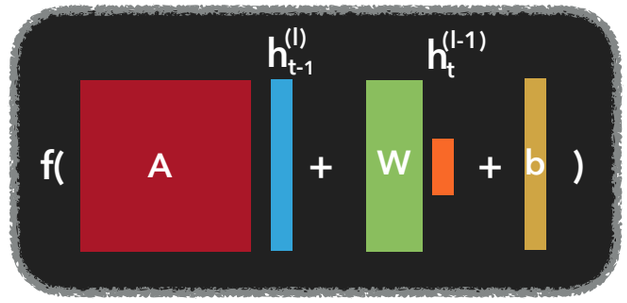
VANISHING AND EXPLODING GRADIENT



Matrix Multiplication



VANISHING AND EXPLODING GRADIENT

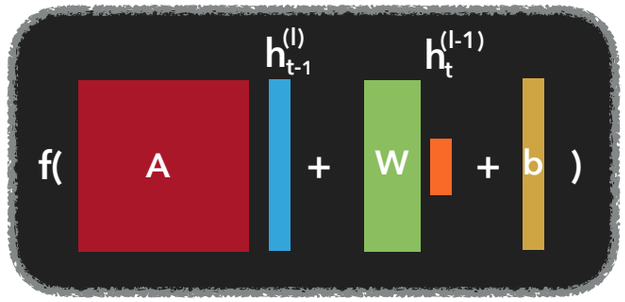


$$h_t^{(l)} = f(Ah_{t-1}^{(l)} + Wh_t^{(l-1)} + b)$$

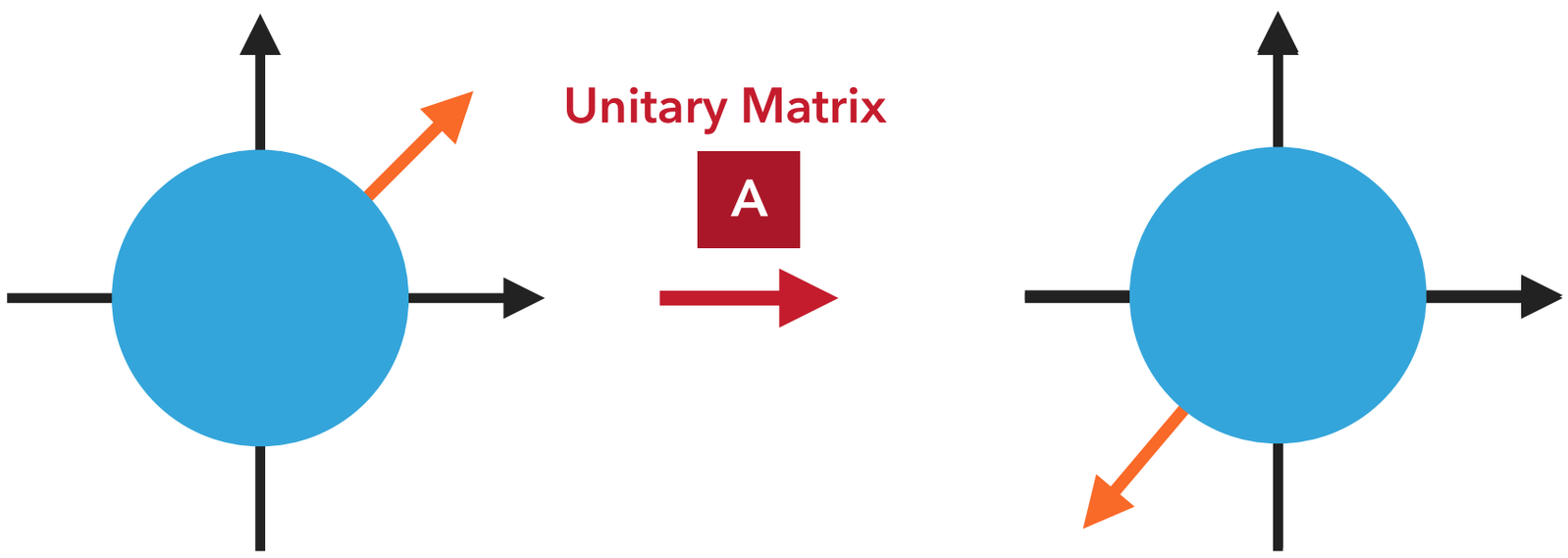
backpropagation

$$\frac{\partial h_T^{(l)}}{\partial A_{a,b}} = \ldots + \ldots A^t \frac{\partial h_{T-t}^{(l)}}{\partial A_{a,b}}$$

UNITARY RNN

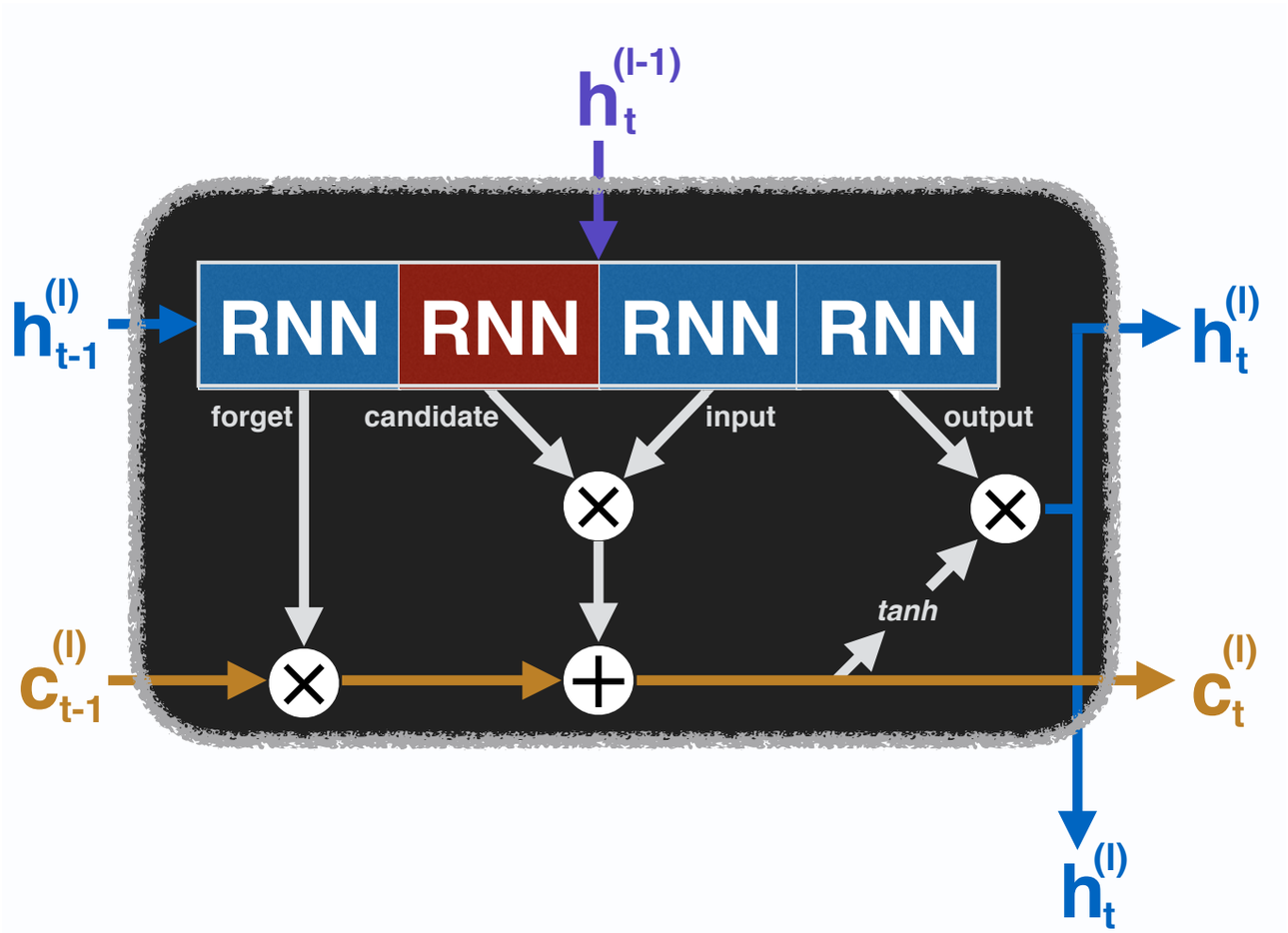


$$\frac{\partial h_T^{(l)}}{\partial A_{a,b}} = \ldots + \ldots A^t \frac{\partial h_{T-t}^{(l)}}{\partial A_{a,b}}$$

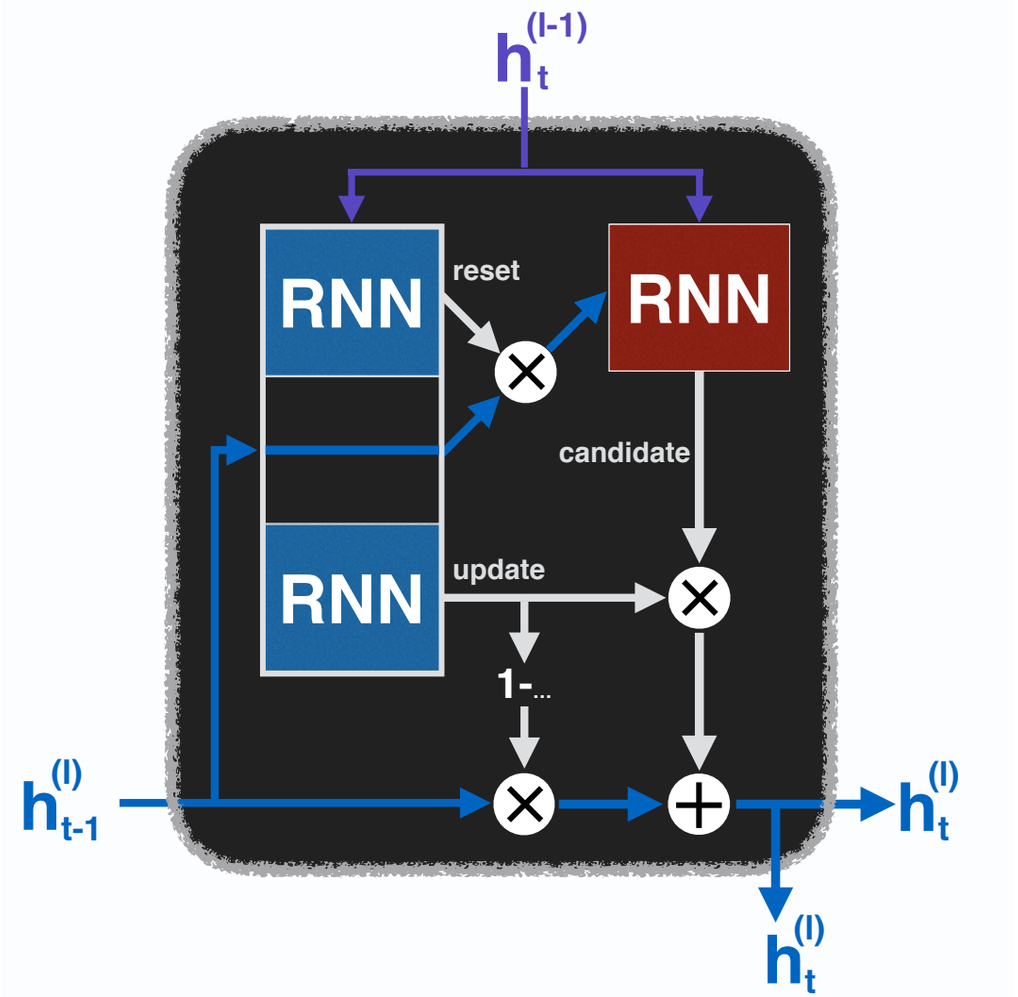


GATED SOLUTIONS

LSTM



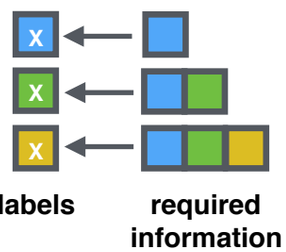
GRU



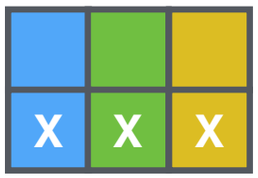
TASKS REQUIRING HIGH INPUT HISTORY STORAGE CAPACITY

non sequential

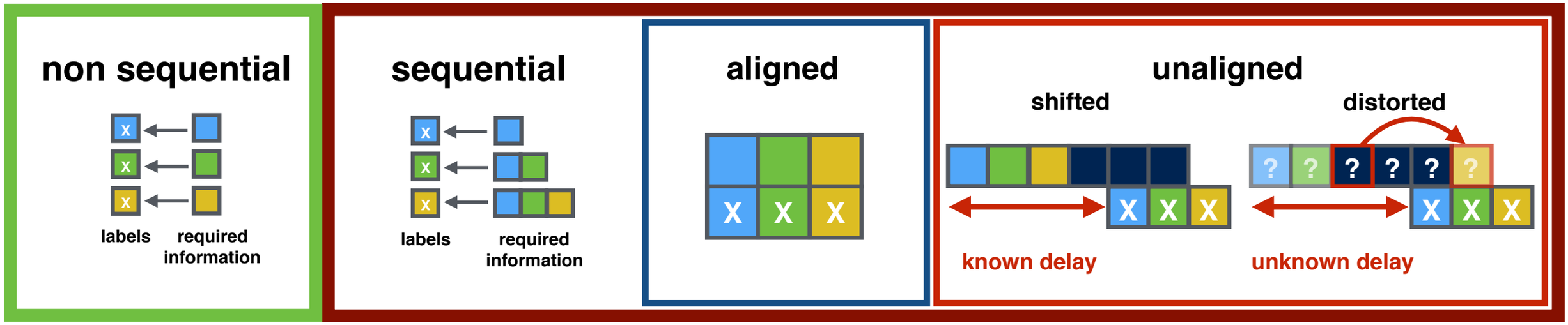
sequential



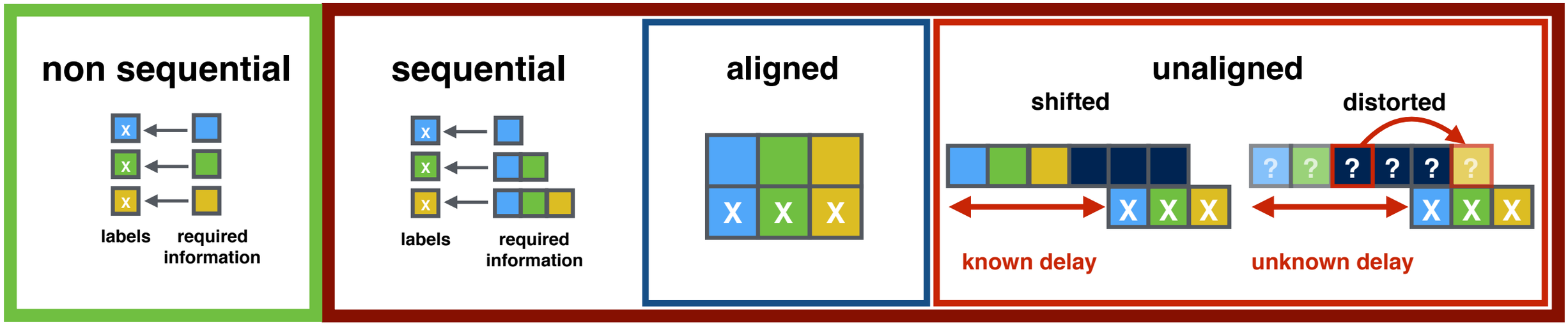
aligned



TASKS REQUIRING HIGH INPUT HISTORY STORAGE CAPACITY

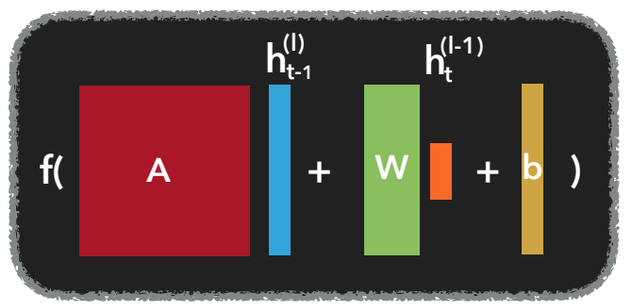


TASKS REQUIRING HIGH INPUT HISTORY STORAGE CAPACITY



high numbers of hidden units

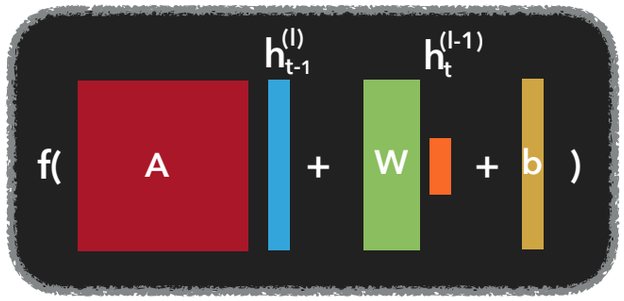
- intractability
- overcomplexity



high delays

- vanishing and exploding gradient problem

BAND RNN



Band Matrix Restriction

intractability → linear growth in necessary storage and operations: O(n)

overcomplexity → linear growth in number of parameters\task storage capacity: O(n)

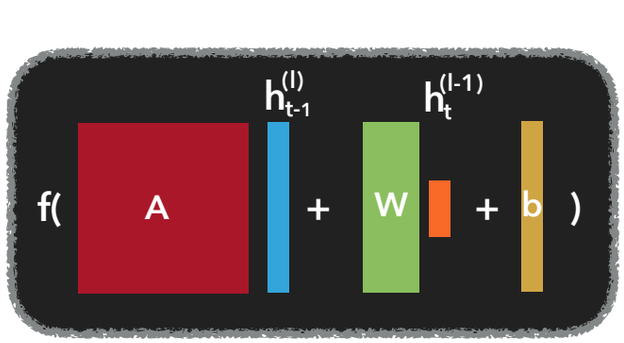
vanishing and exploding gradient problem NOT solved BUT

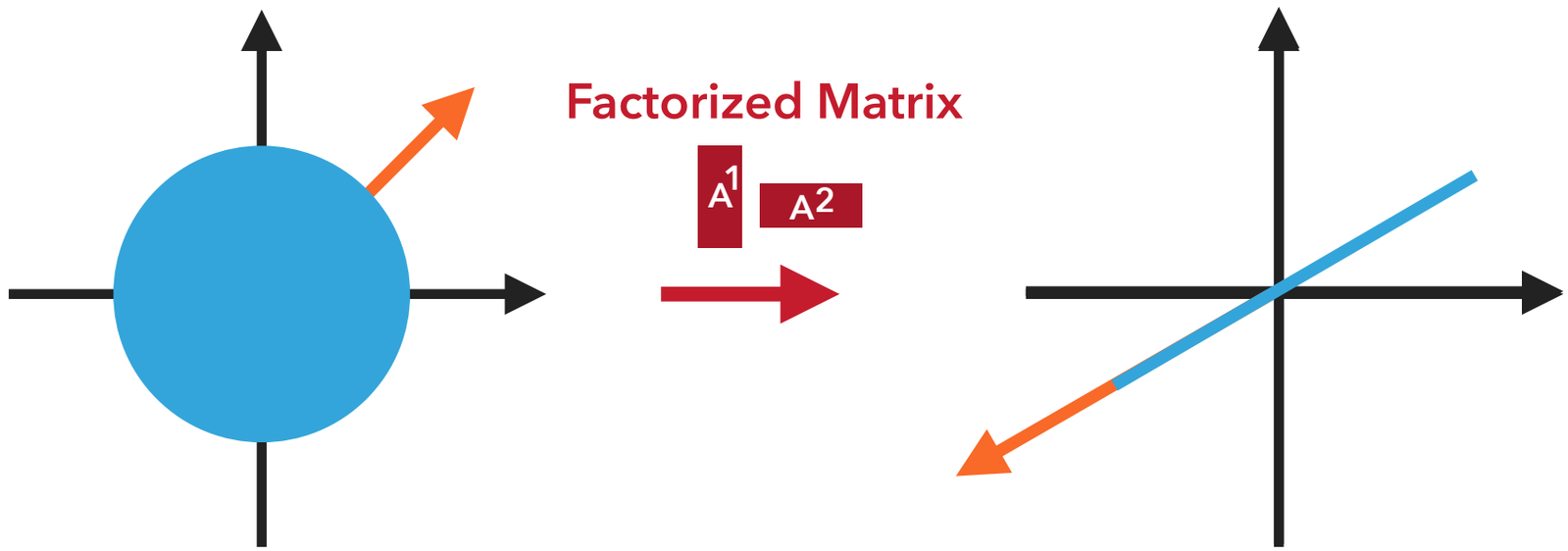
- → close to but lower than full rank
- → less chaotic
- → shift initialization

MEMORY CAPACITY OF OTHER MATRIX RESTRICTIONS

Factorized

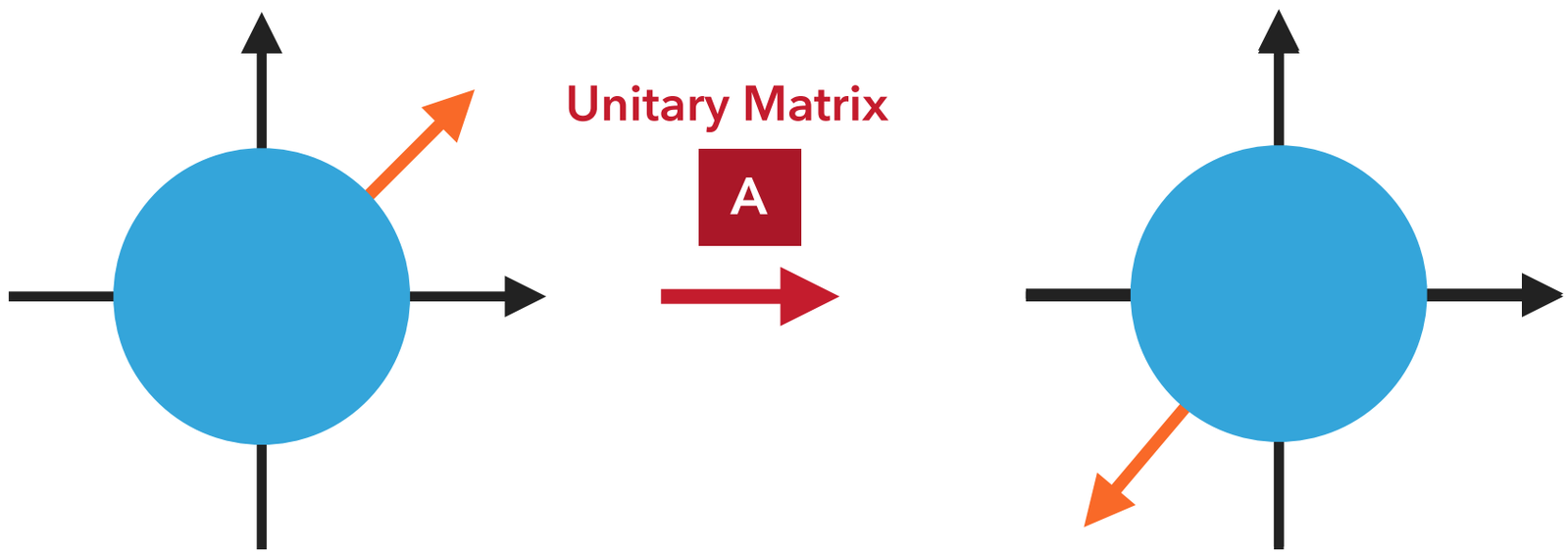
forgets much information





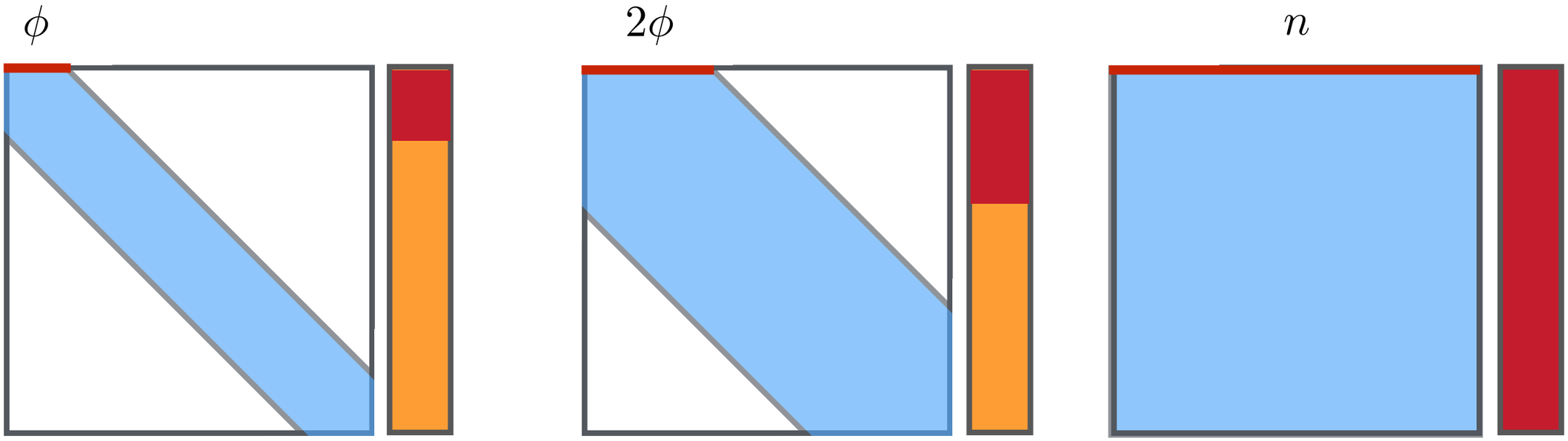
URNN

needs input to forget



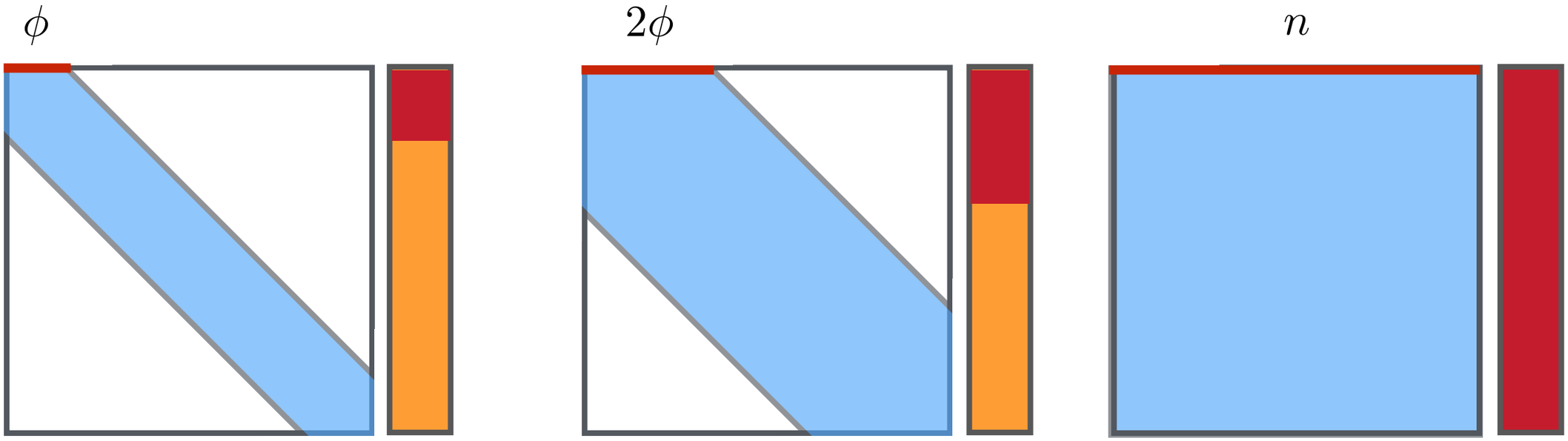
MESSAGE PASSING

A A^2 $A^{\frac{n}{\phi}}$



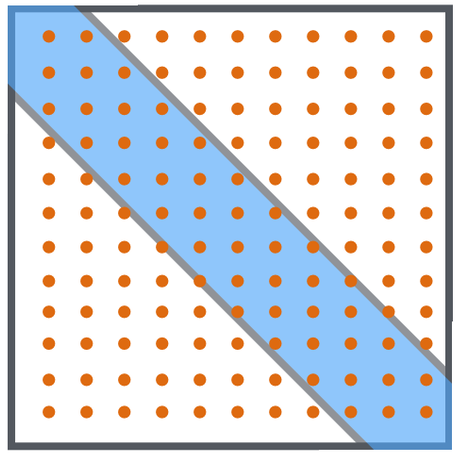
MESSAGE PASSING

A A^2 $A^{\frac{n}{\phi}}$



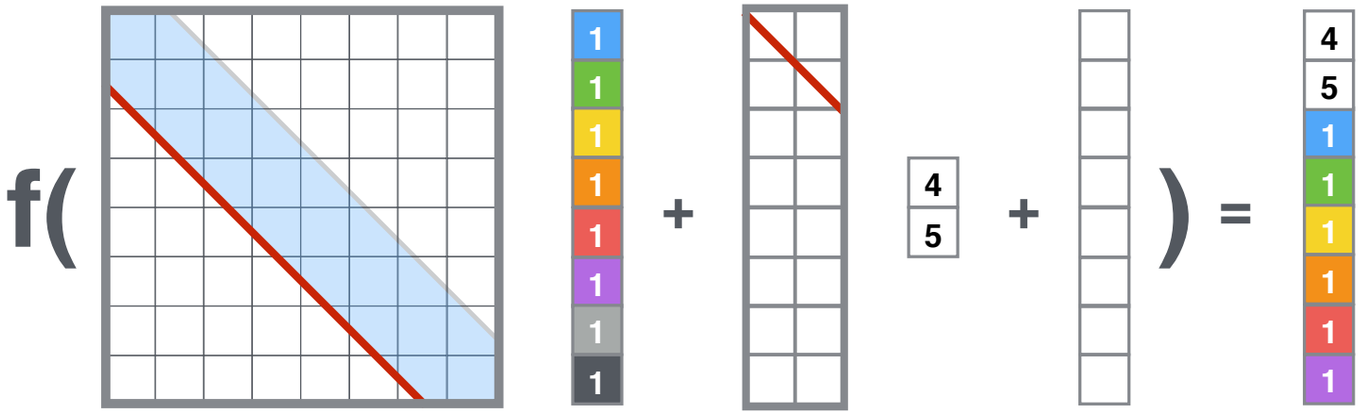
idea: adding a sparse grid on top of the band

- speeds up the message passing
- adds complexity
- complex system need long-range interactions



SHIFT INITIALIZATION

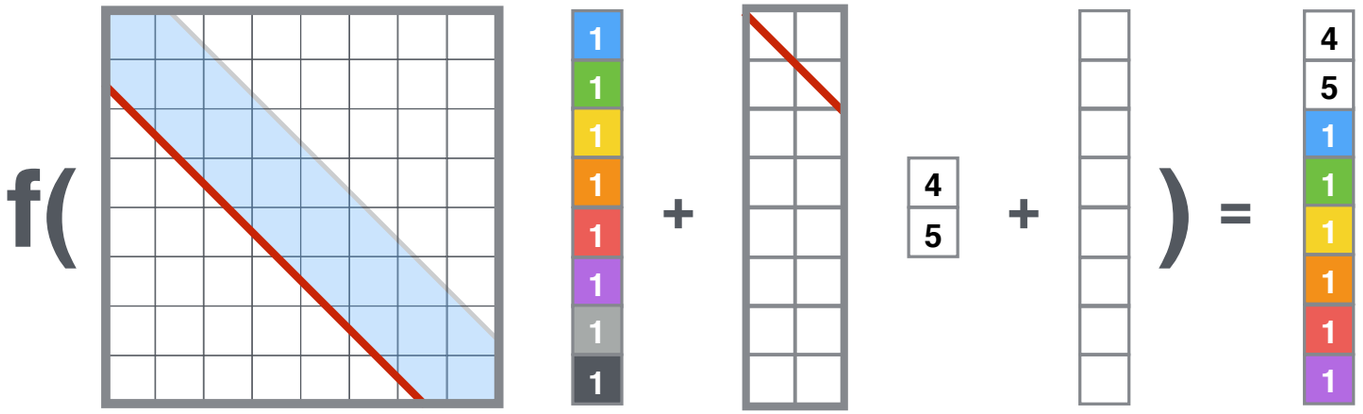
Shift initialization



W_h h_{i-1} W_i x_i b h_i

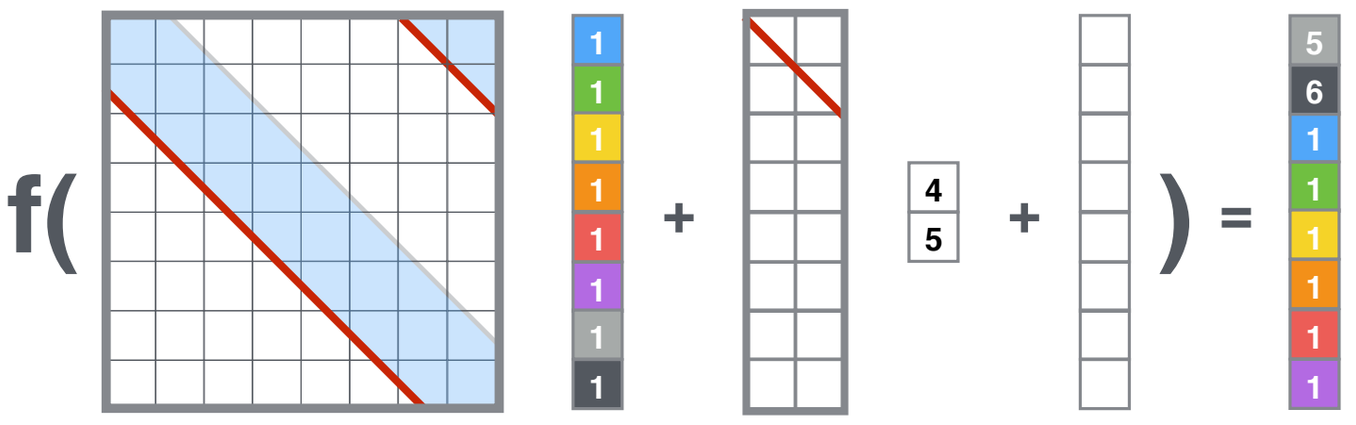
SHIFT INITIALIZATION

Shift initialization



W_h h_{i-1} W_i x_i b h_i

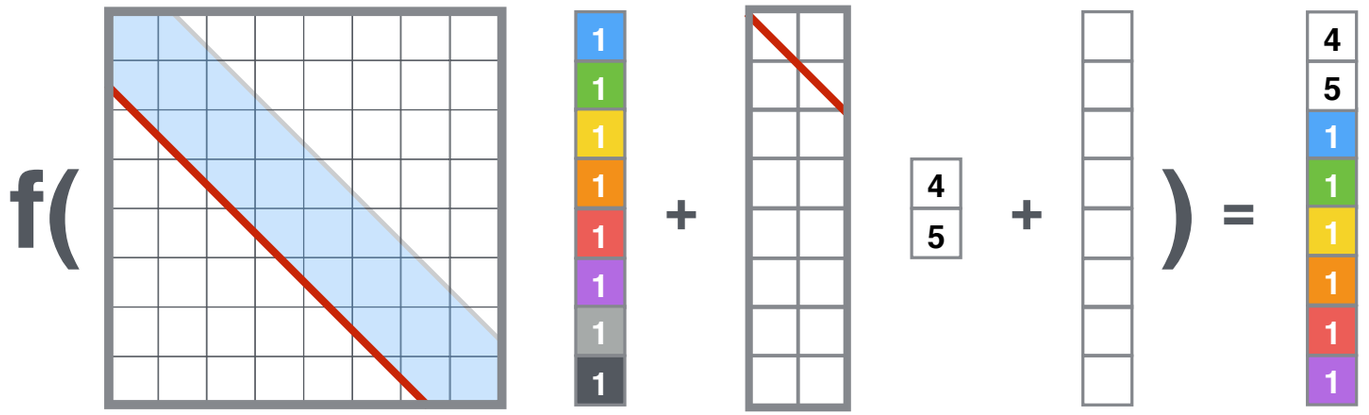
Closed Shift initialization



W_h h_{i-1} W_i x_i b h_i

SHIFT INITIALIZATION

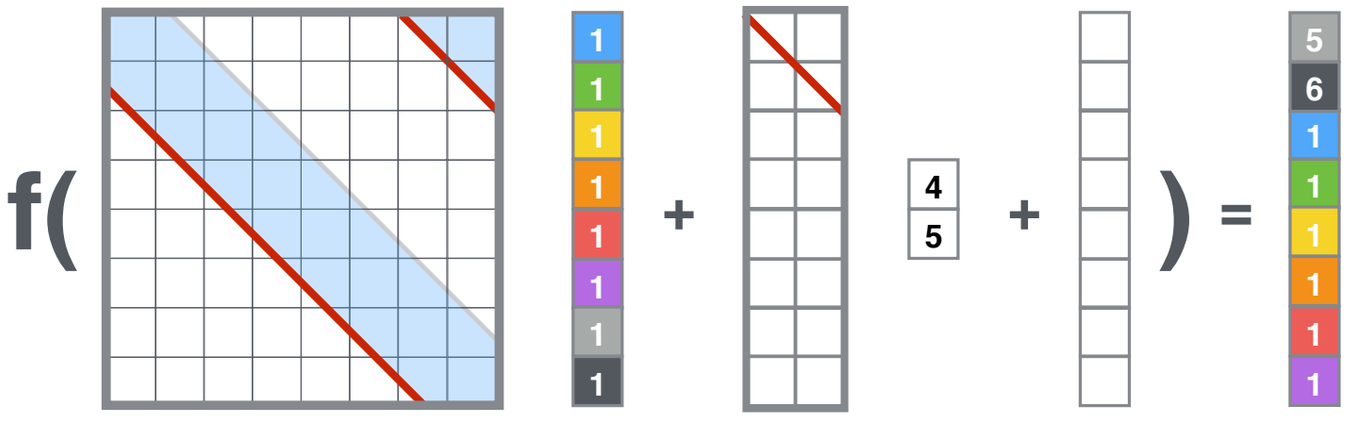
Shift initialization



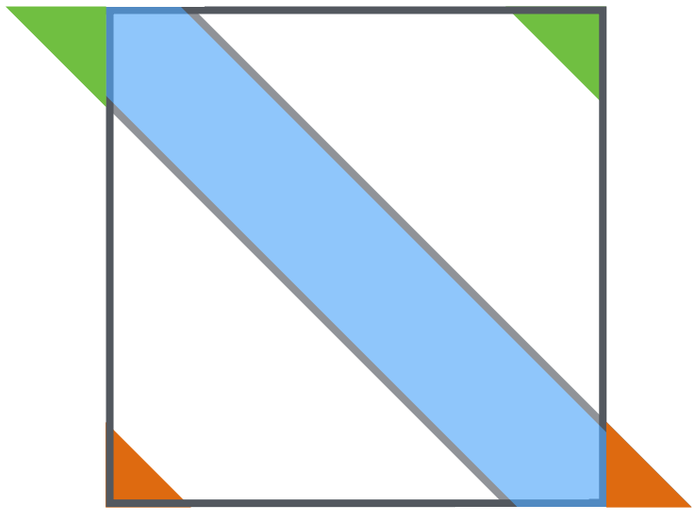
$W_h \quad h_{i-1} \quad W_i \quad x_i \quad b \quad h_i$

idea: adding values to the corners of the band matrix

Closed Shift initialization



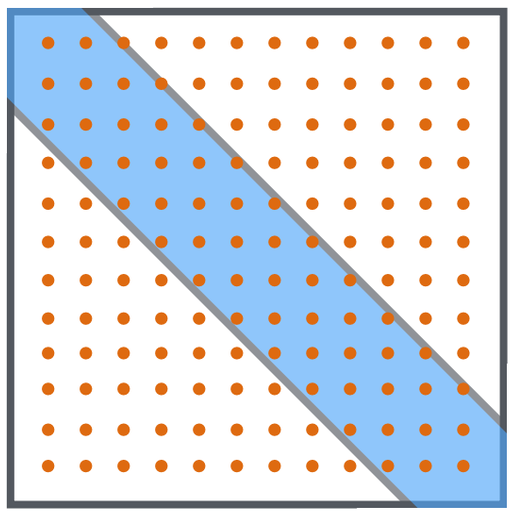
$W_h \quad h_{i-1} \quad W_i \quad x_i \quad b \quad h_i$



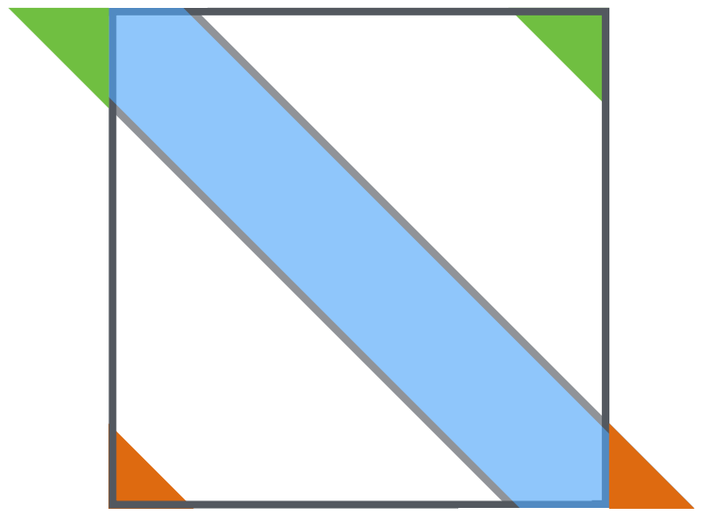
FINAL MODELS

band

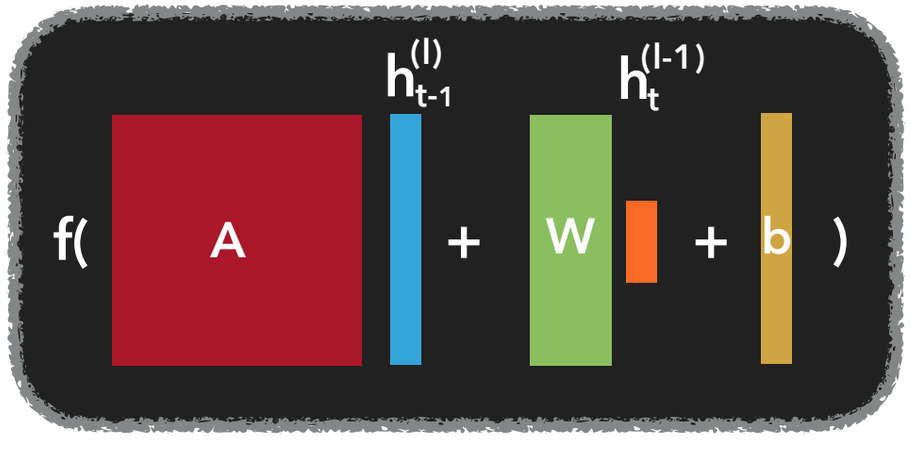
band grid combination



closed band

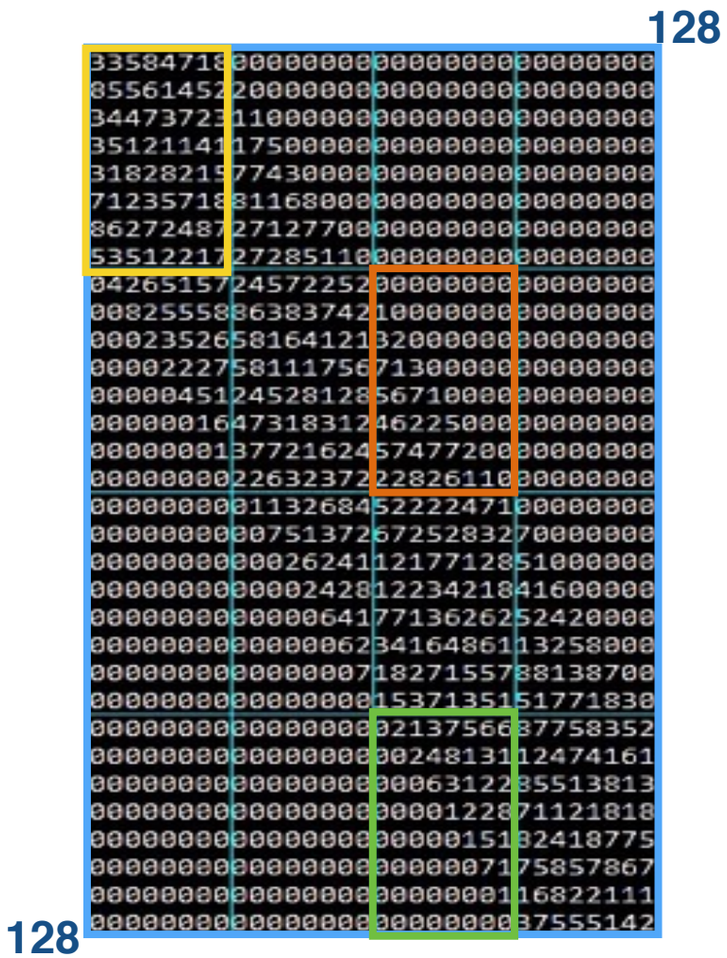


Band Matrix Restrictions



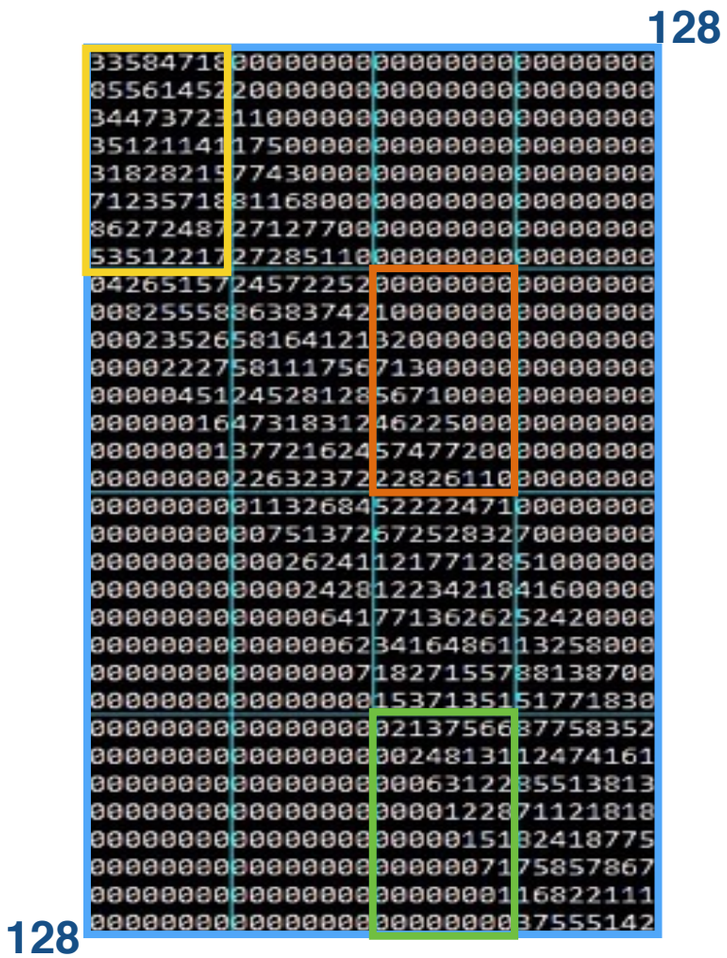
SPARSE AND BLOCK FORMAT

standard matrix



SPARSE AND BLOCK FORMAT

standard matrix

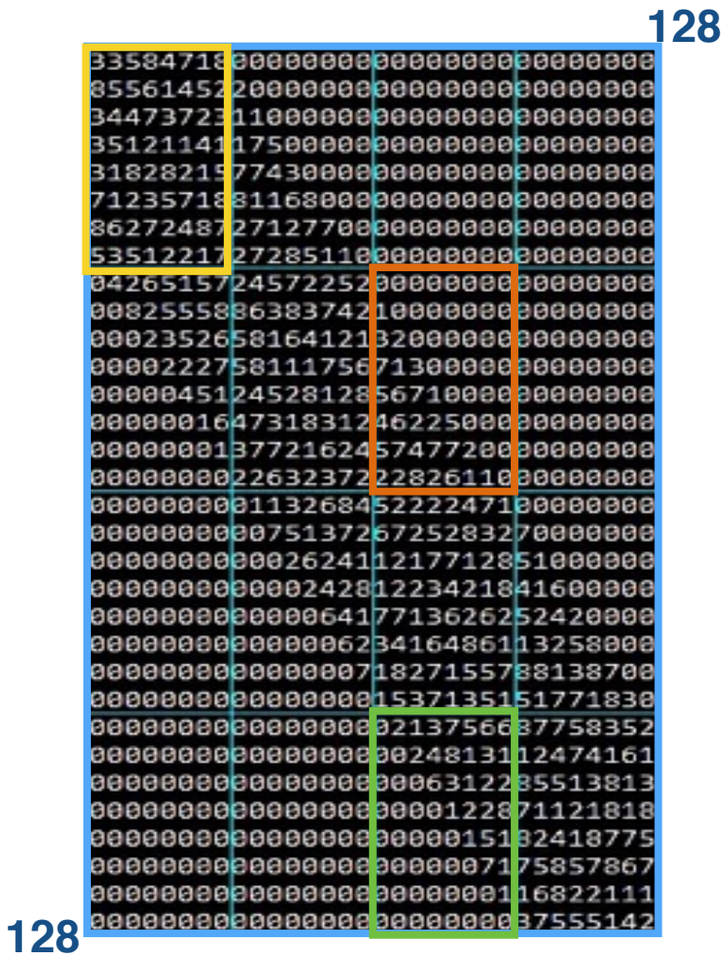


sparse format

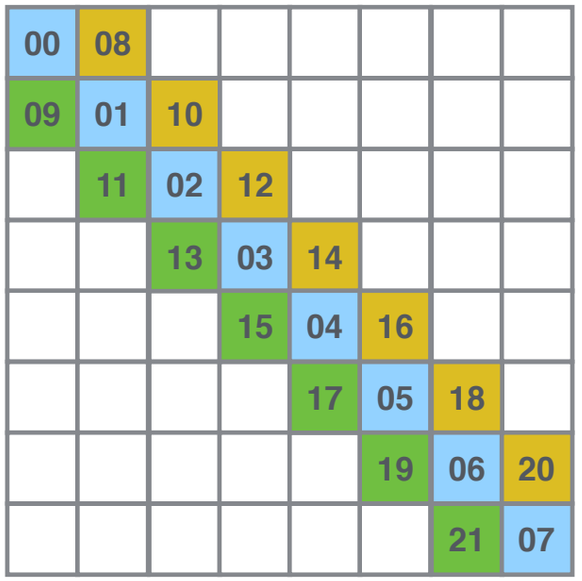
00	08						
09	01	10					
	11	02	12				
		13	03	14			
			15	04	16		
				17	05	18	
					19	06	20
						21	07

SPARSE AND BLOCK FORMAT

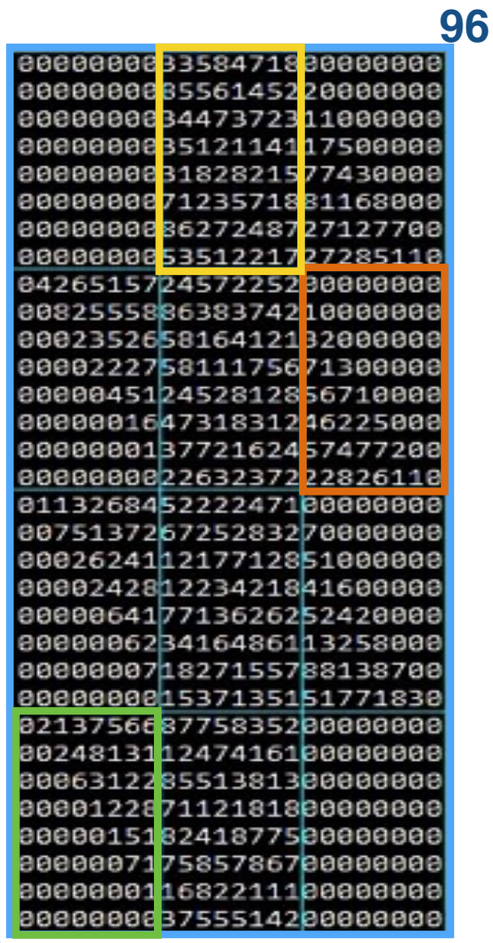
standard matrix



sparse format

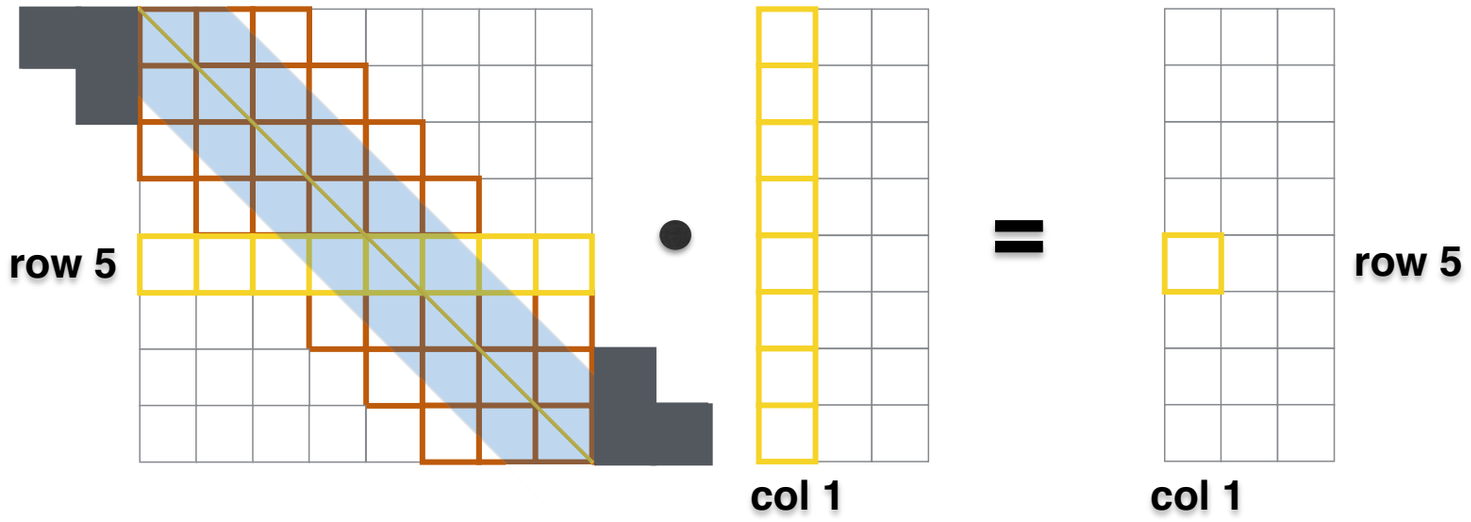


block format



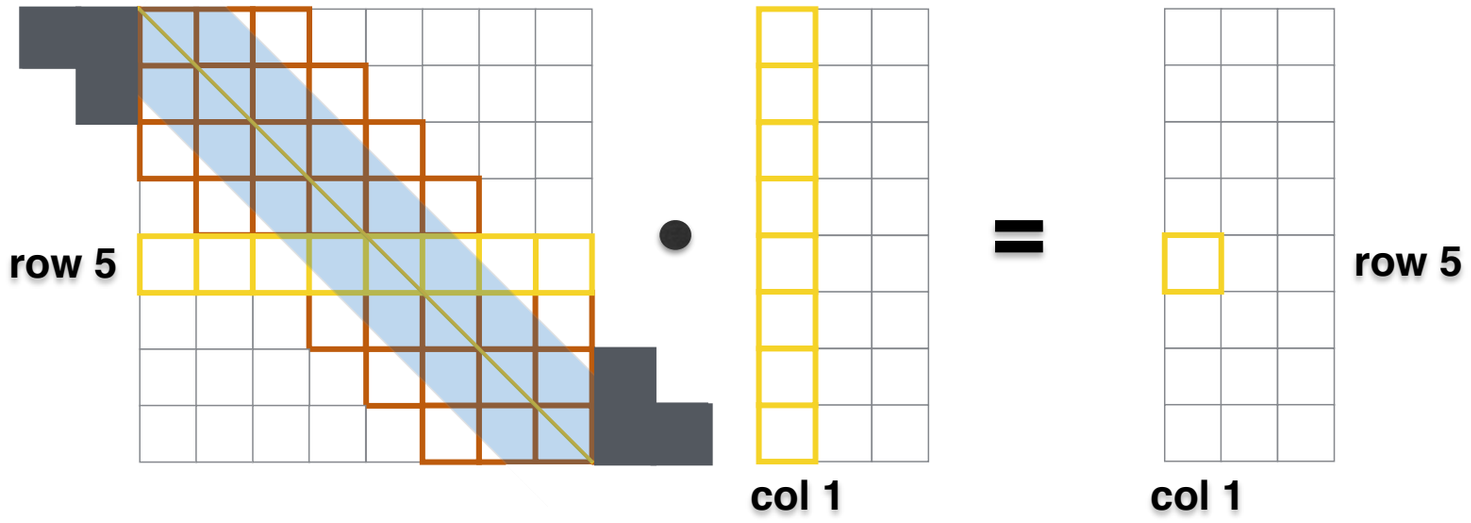
GENERAL BAND-MATRIX MATRIX MULTIPLICATION IDEA

GEMM loops through all blocks of the full matrices A and B

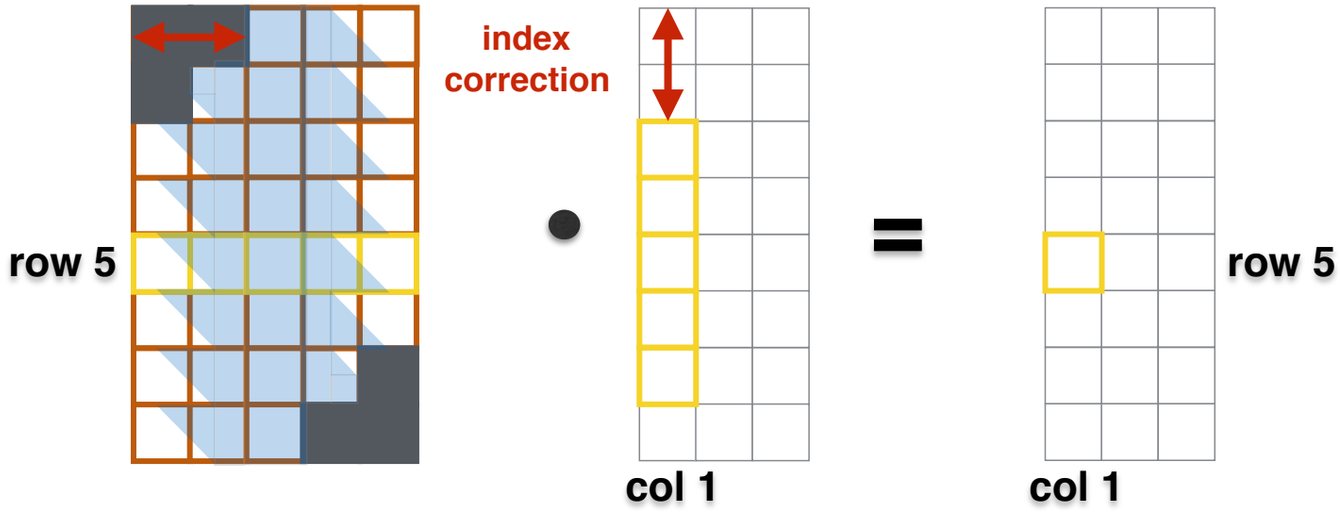


GENERAL BAND-MATRIX MATRIX MULTIPLICATION IDEA

GEMM loops through all blocks of the full matrices A and B

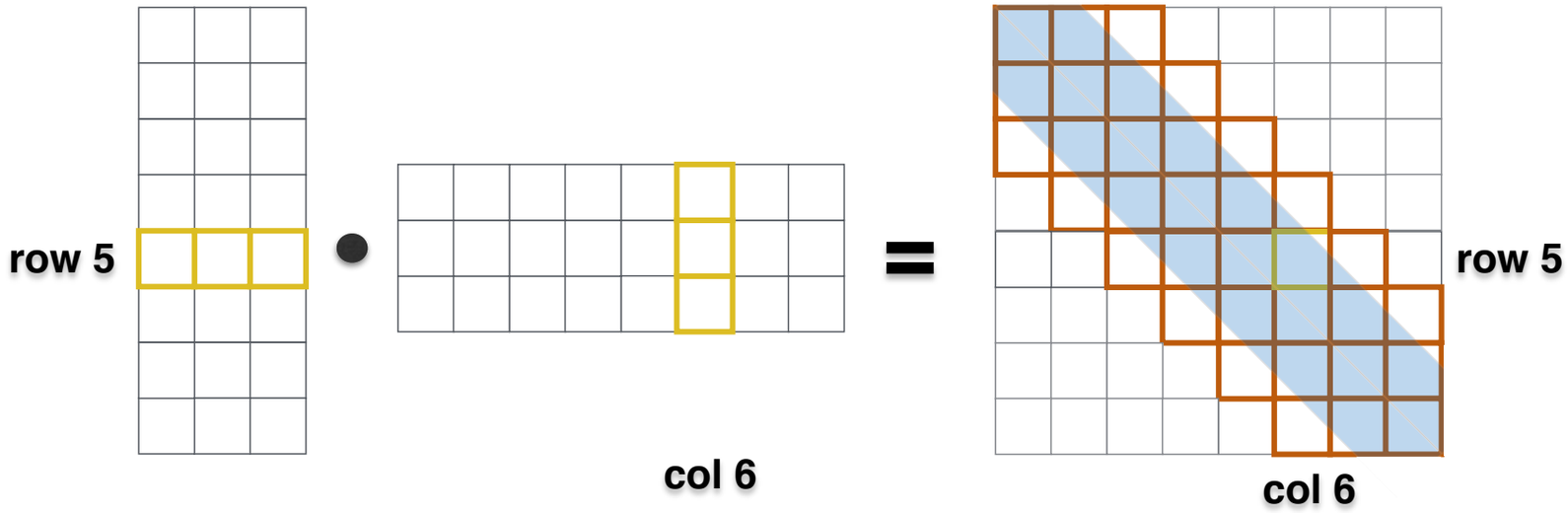


GBMM cuts the loop using a sparse blocked band matrix for A and corrects in block indices in B



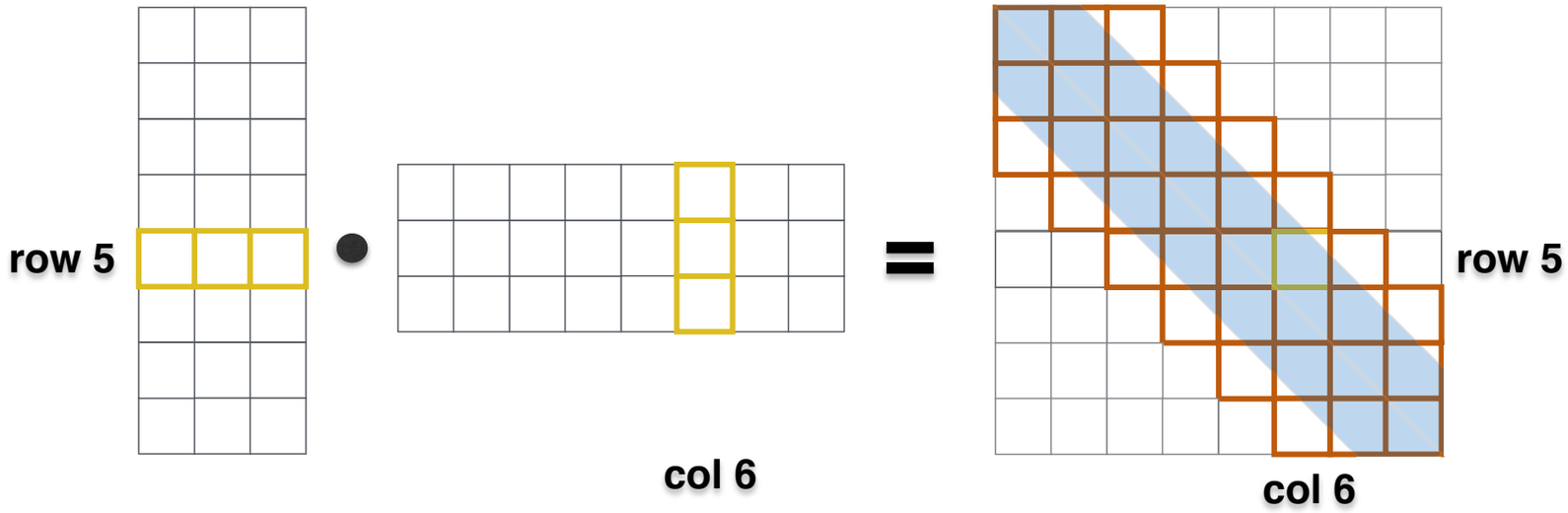
GENERAL TO BAND-MATRIX MATRIX MULTIPLICATION IDEA

GEMM calculates all blocks of the resulting matrix C

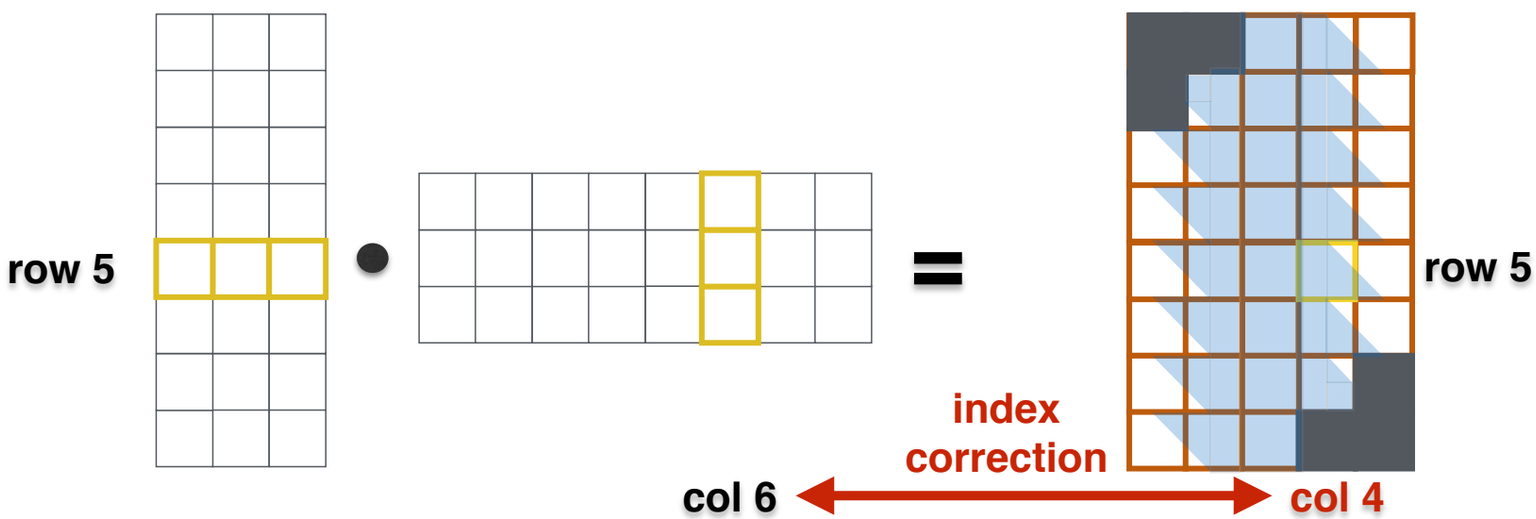


GENERAL TO BAND-MATRIX MATRIX MULTIPLICATION IDEA

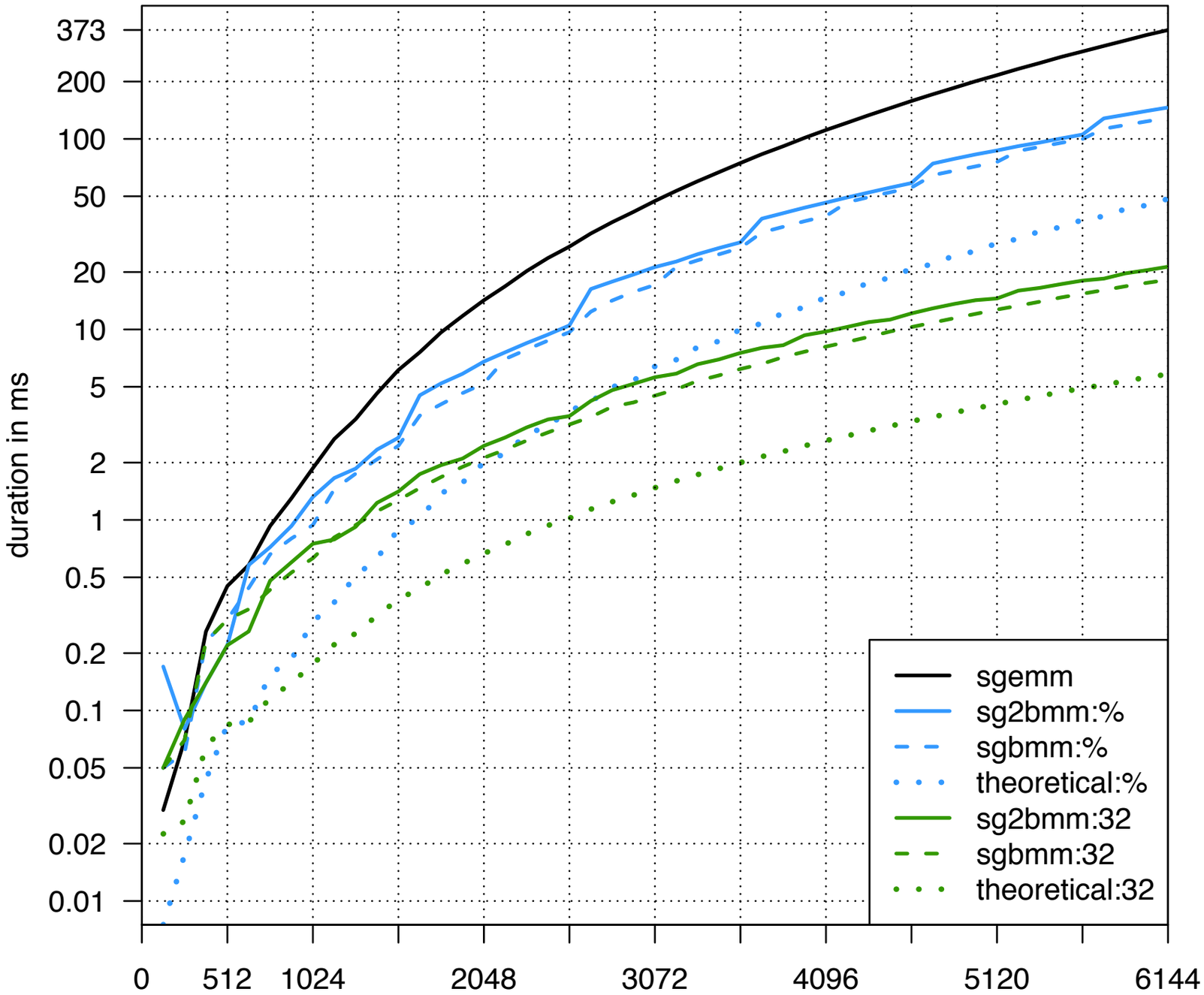
GEMM calculates all blocks of the resulting matrix C



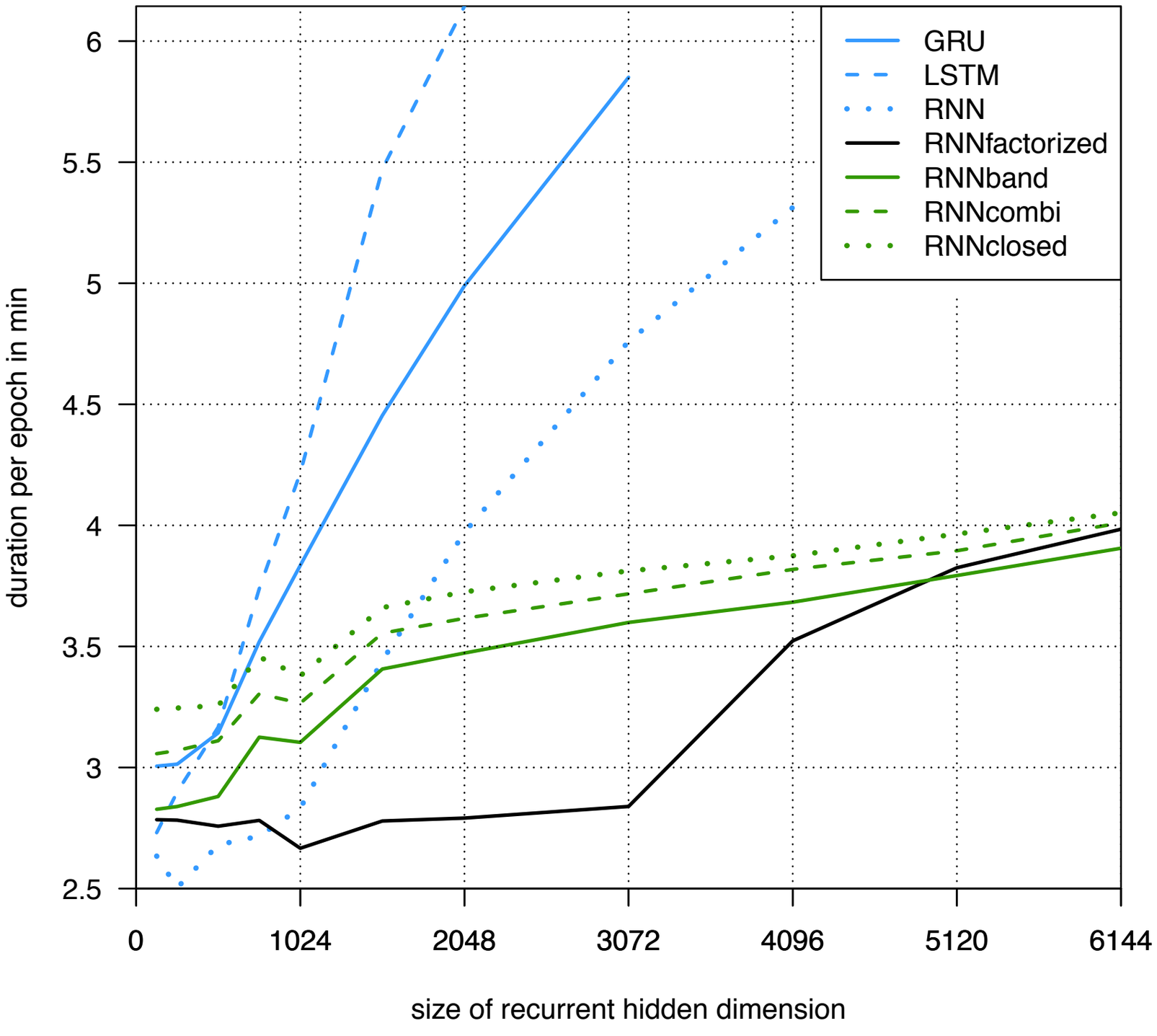
SG2BMM only calculated the blocks of C stored in blocked band format



KERNEL BENCHMARK



MODEL BENCHMARK



COMPARABILITY

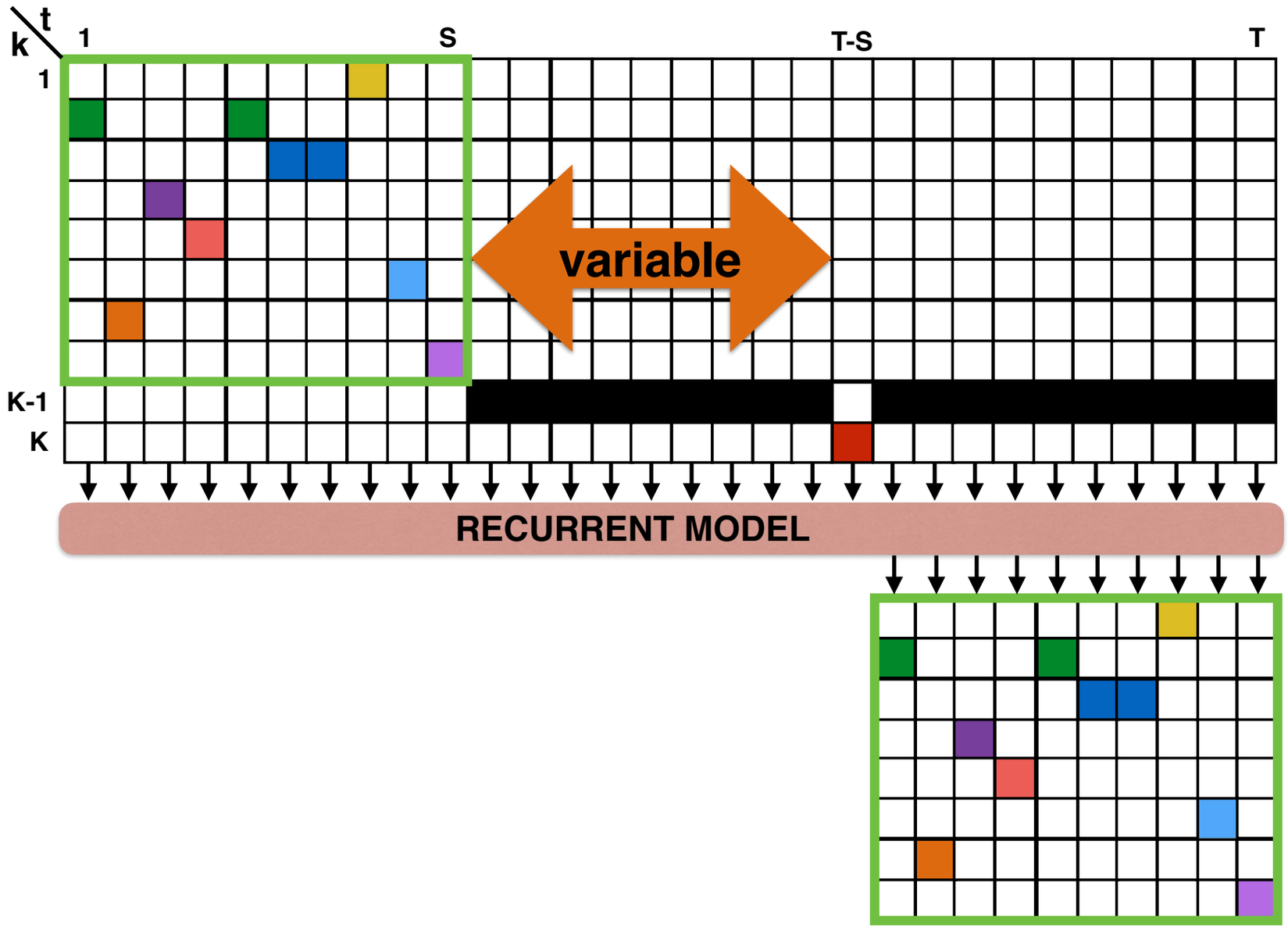
- **fixed number of trainable parameters: 80k (540k for big model)**
- **band-half-width and factorizing dimension: 32**

Example for Braille Task

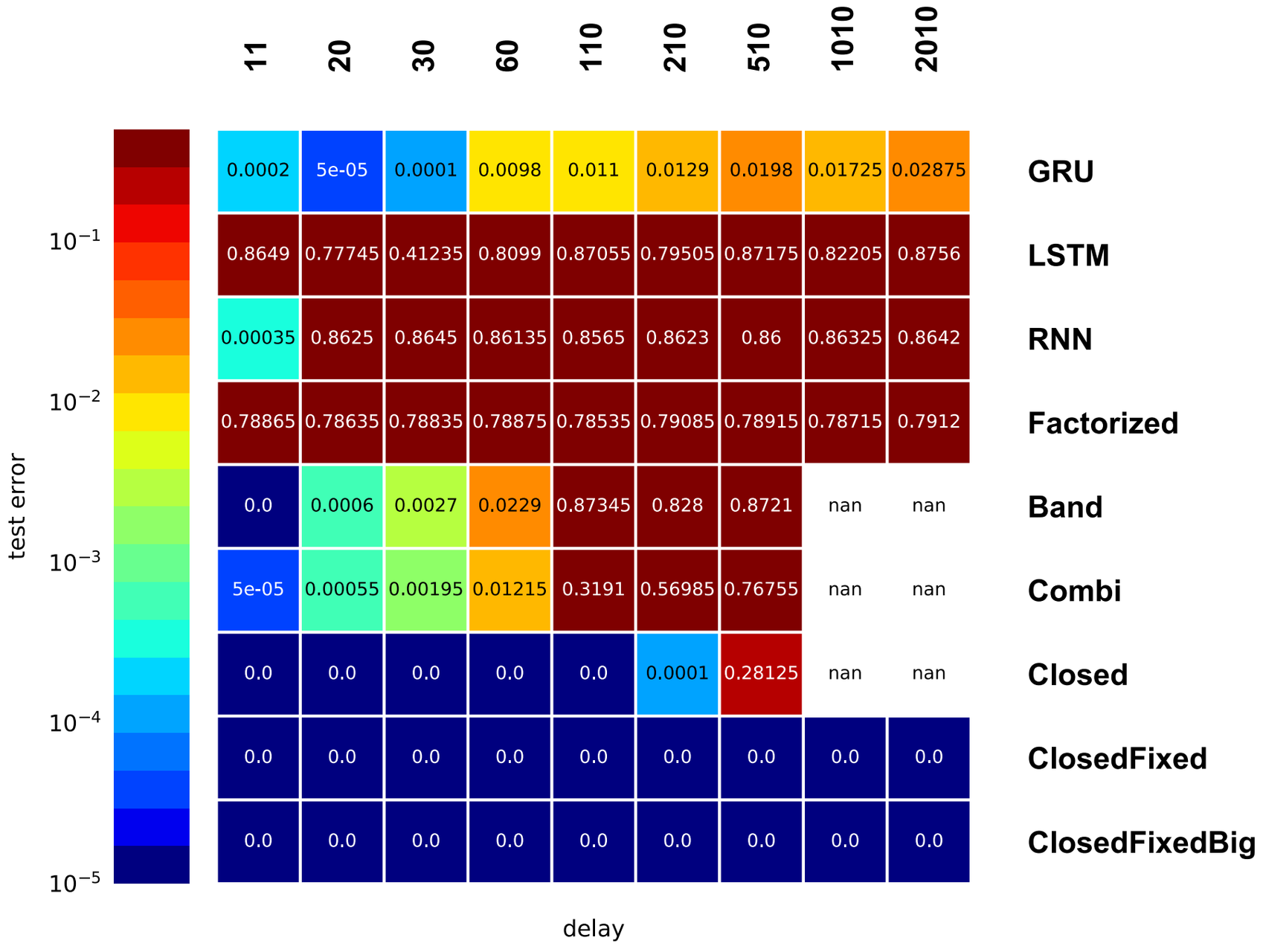
model	hidden units
GRU	160
LSTM	128
RNN	256
Factorized	864
Band	896
Combi	864
Closed	864
Closed Fixed	864
Big Closed Fixed	5888

Note: an LSTM with 1000 hidden units has as much parameters as a band model (band-half-width of 32) with 55000 hidden units

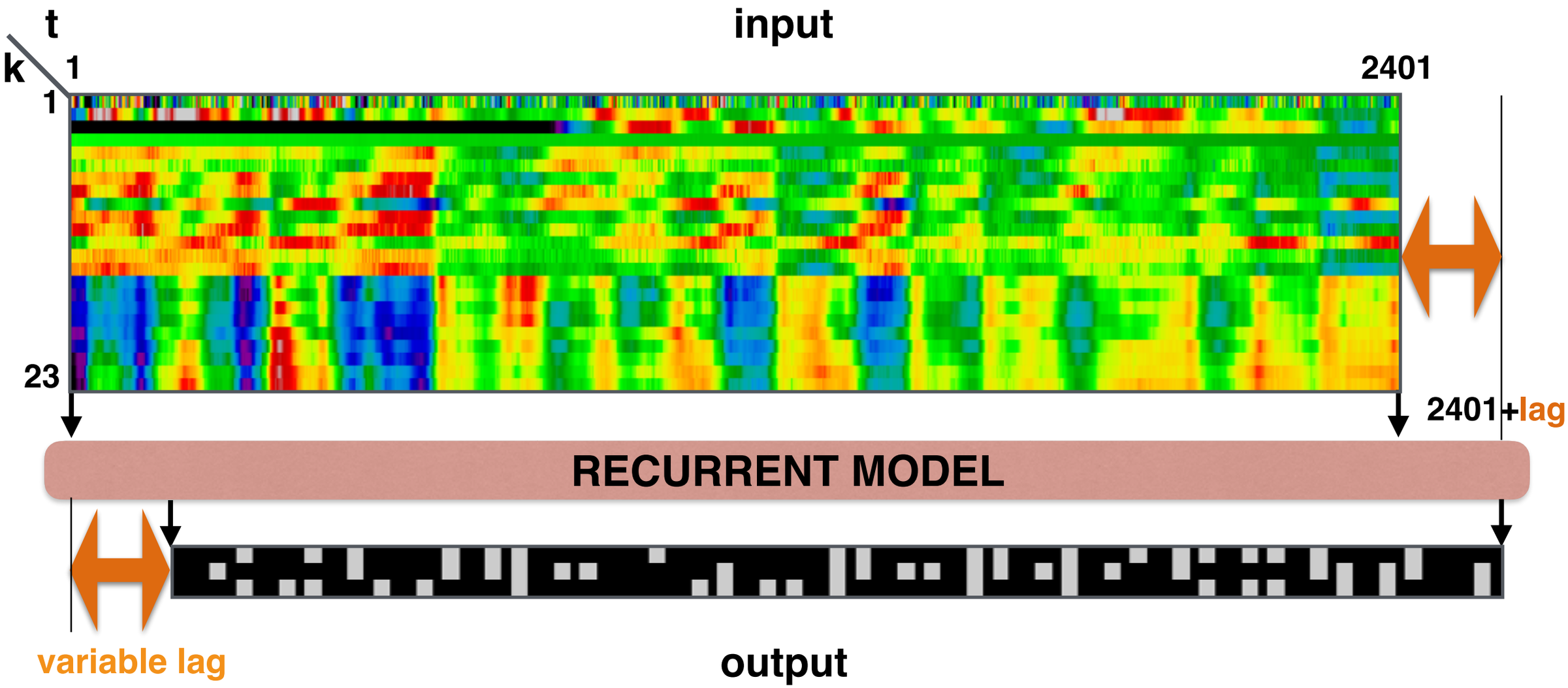
COPY TASK – INPUT



COPY TASK – RESULTS

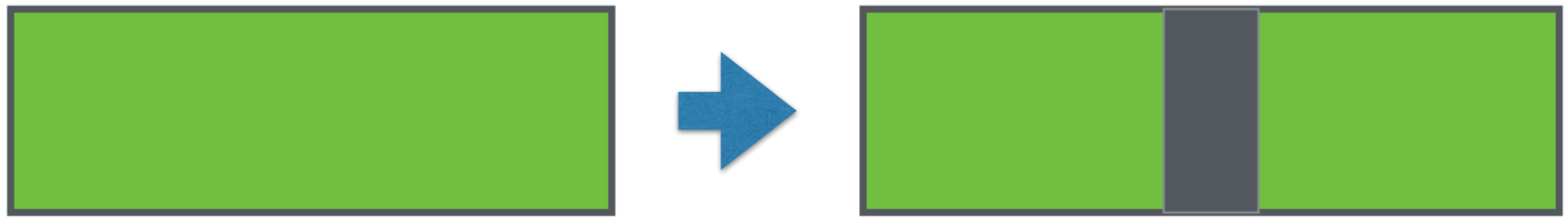


BRAILLE TASK – INPUT



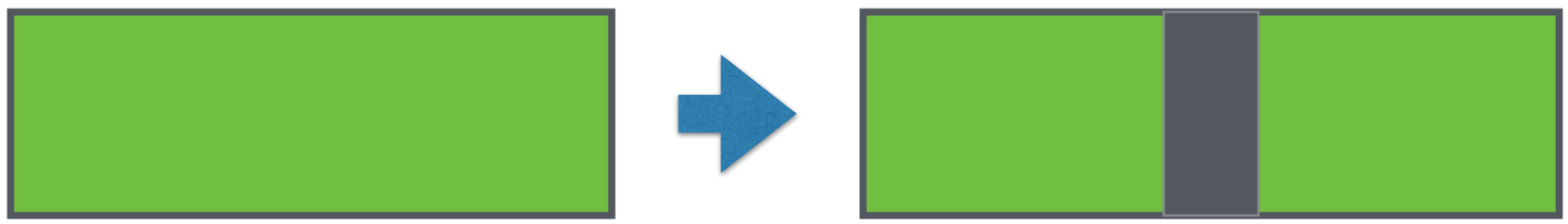
BRAILLE TASK – DISTORTIONS

paused

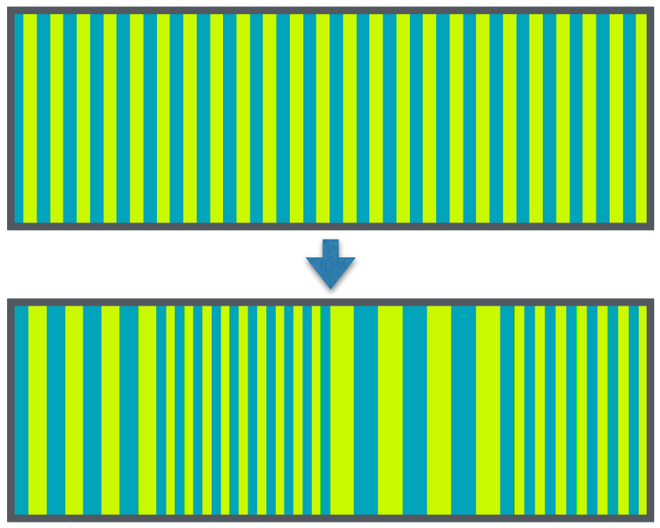


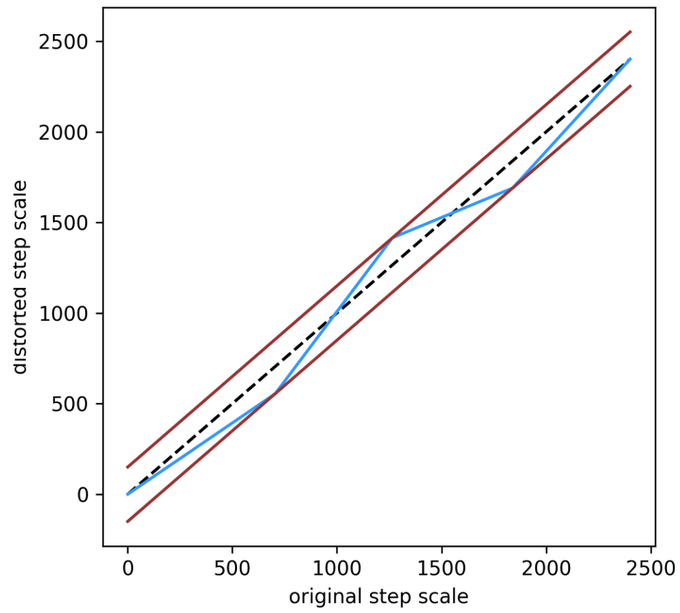
BRAILLE TASK – DISTORTIONS

paused

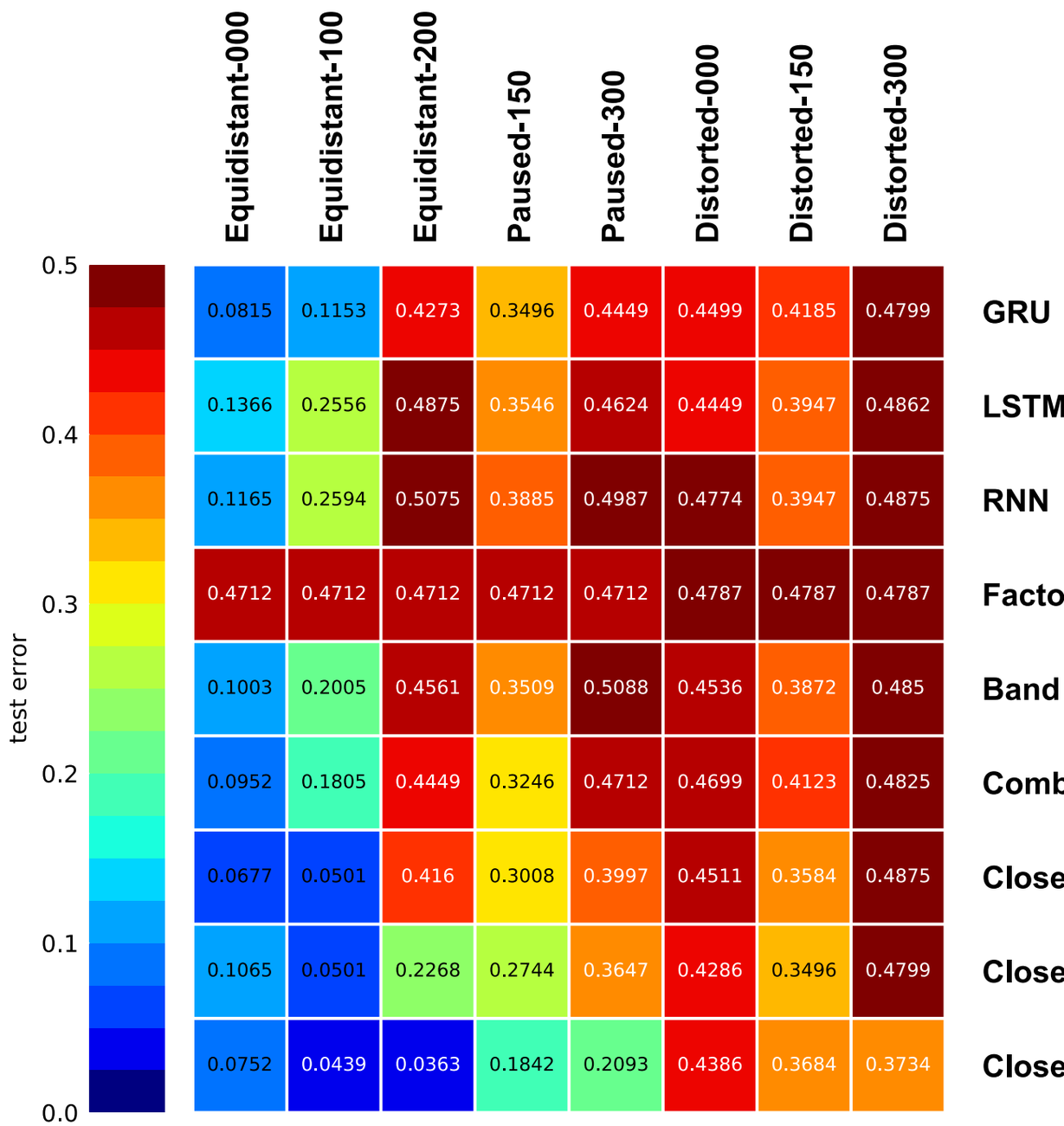


distorted

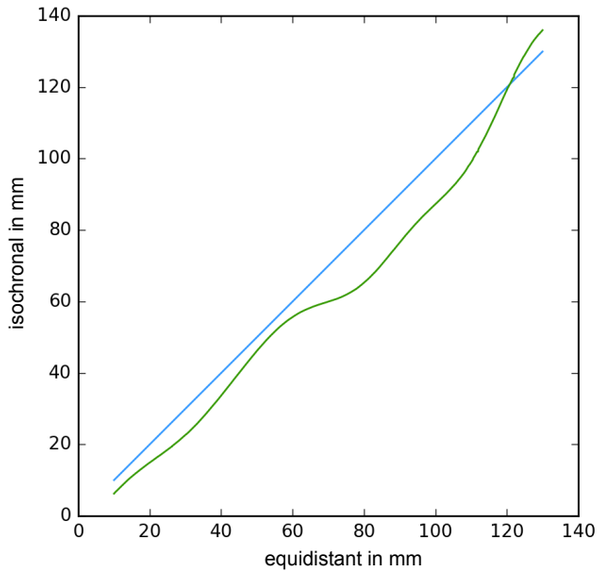


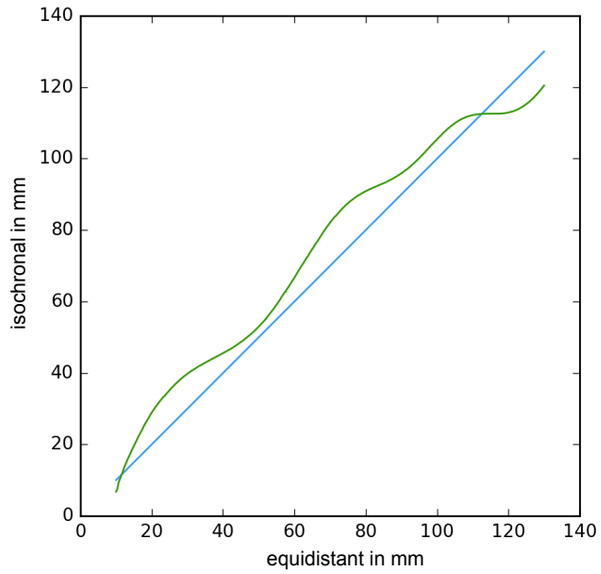


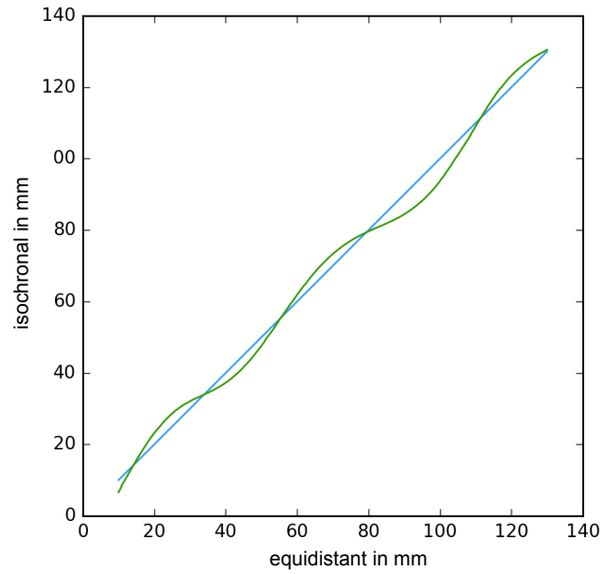
BRAILLE TASK – CONSTANT SPEED RESULTS

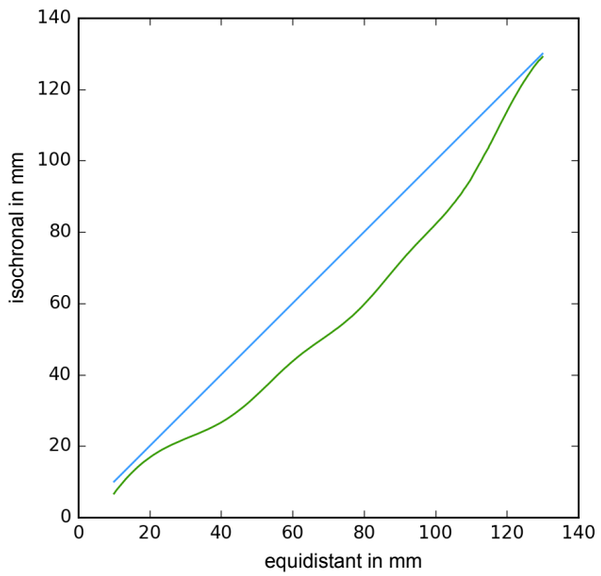


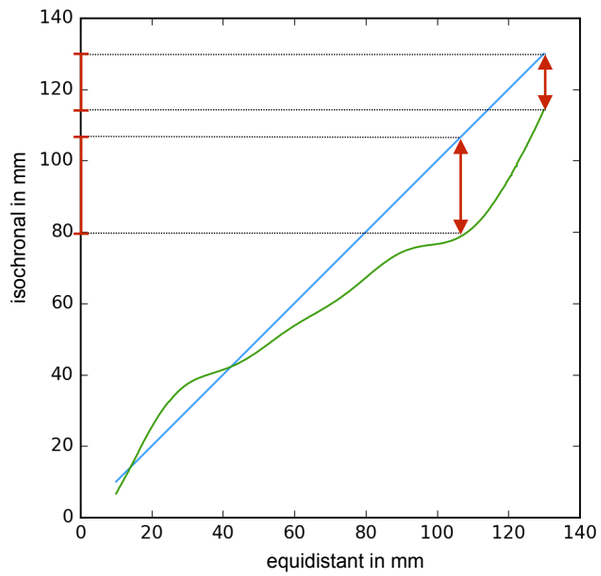
BRAILLE TASK – VARIABLE SPEED

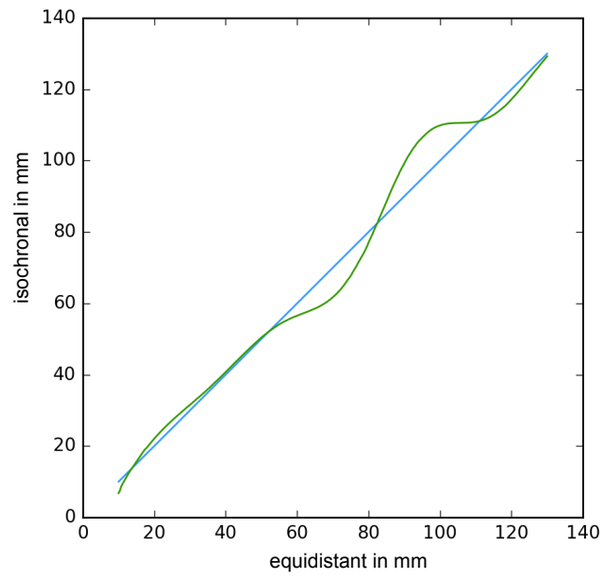




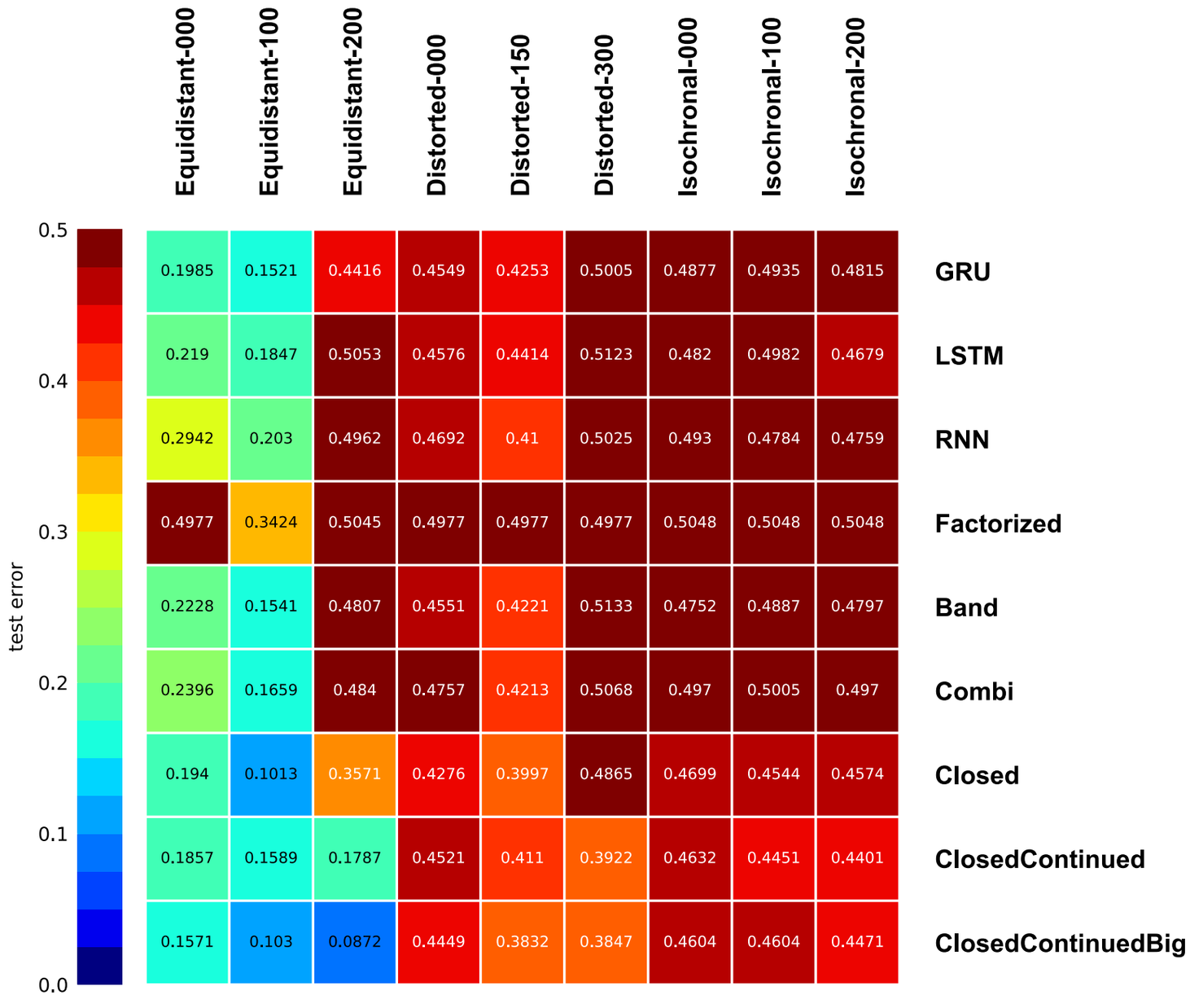








BRAILLE TASK – VARIABLE SPEED RESULTS



EVALUATION

- classical models suffered overfitting for high delays
  → regularization (dropout)

EVALUATION

- classical models suffered overfitting for high delays
  → regularization (dropout)

- closed band RNNs showed best performance (increased with hidden units and delay)
  → comparison to the unitary RNN

EVALUATION

- classical models suffered overfitting for high delays
  → regularization (dropout)

- closed band RNNs showed best performance (increased with hidden units and delay)
  → comparison to the unitary RNN

- fixed shift initialization required for high delays
  → epoch wise increasing band-half-with

EVALUATION

- classical models suffered overfitting for high delays
  → regularization (dropout)

- closed band RNNs showed best performance (increased with hidden units and delay)
  → comparison to the unitary RNN

- fixed shift initialization required for high delays
  → epoch wise increasing band-half-with

- band RNNs shift capacity from task to input history storage
  → combinations of band RNNs and classical RNNs

EVALUATION

- classical models suffered overfitting for high delays
  → regularization (dropout)

- closed band RNNs showed best performance (increased with hidden units and delay)
  → comparison to the unitary RNN

- fixed shift initialization required for high delays
  → epoch wise increasing band-half-with

- band RNNs shift capacity from task to input history storage
  → combinations of band RNNs and classical RNNs

- braille task required high delays and input history storage capacity
  → more tests on natural data sets

THANK YOU