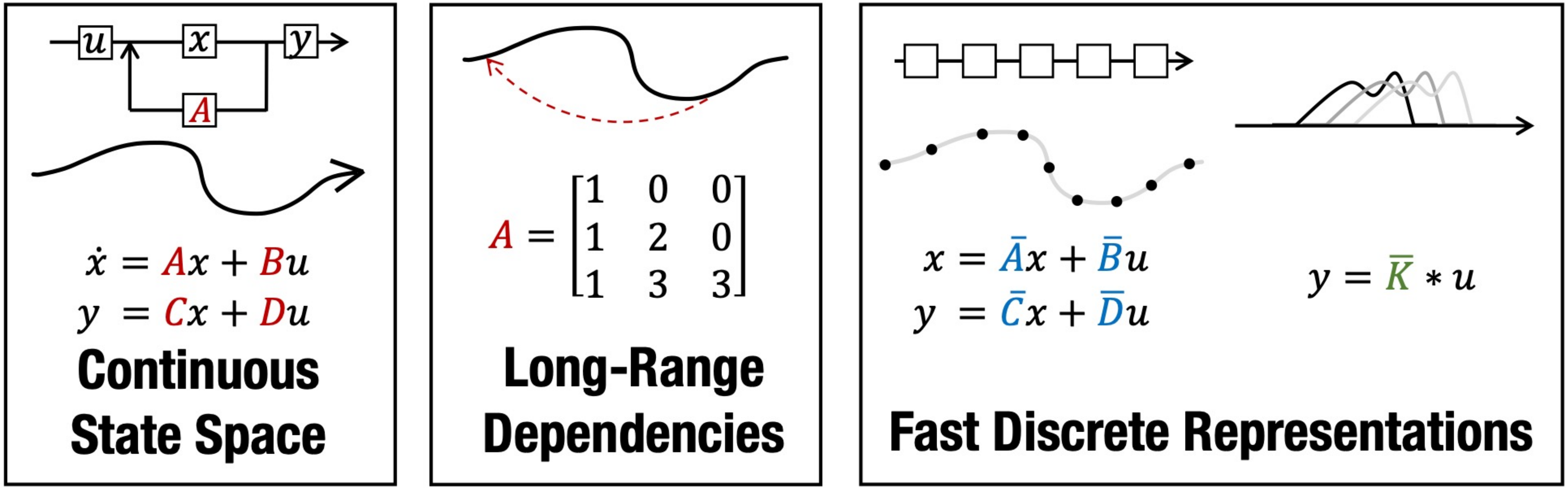
Mamba: Linear-Time Sequence Modeling with Selective State Spaces

S4 (easy mode)

Why S4? Not a basic SSM?



(1) SSM can model long sequence

(2) However, a basic SSM has prohibitive computation and memory requirements!

(3) S4 computes efficiently and require less memory (30x faster + 400x less usage than conventional SSM)

S4 introduces a novel parameterization that efficiently swaps between these representations,
allowing it to handle a wide range of tasks, be efficient at both training and inference, and excel at long sequences.

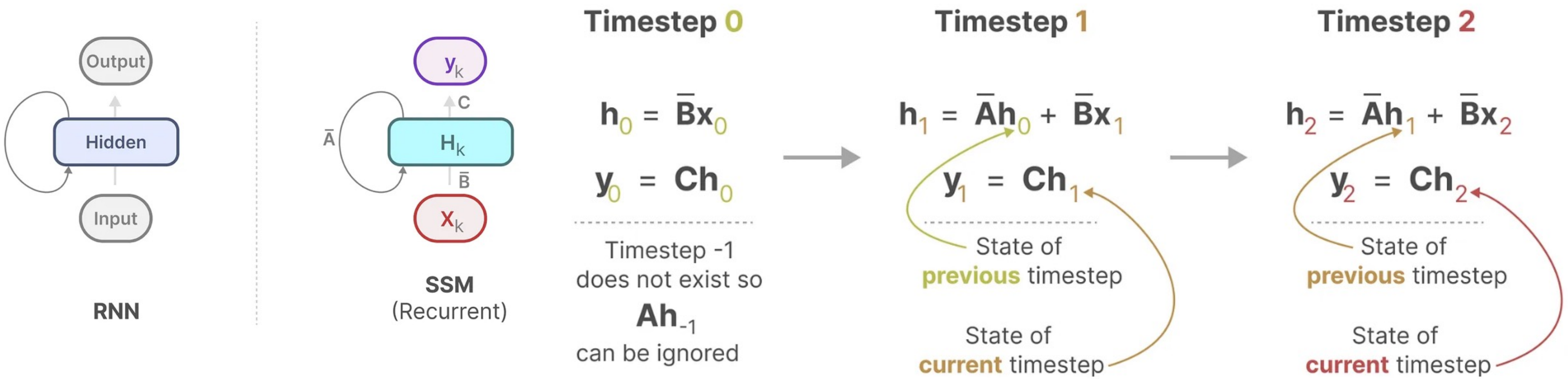
Three representations: Recurrent

$$x = \bar{A}x + \bar{B}u$$
$$y = \bar{C}x + \bar{D}u$$

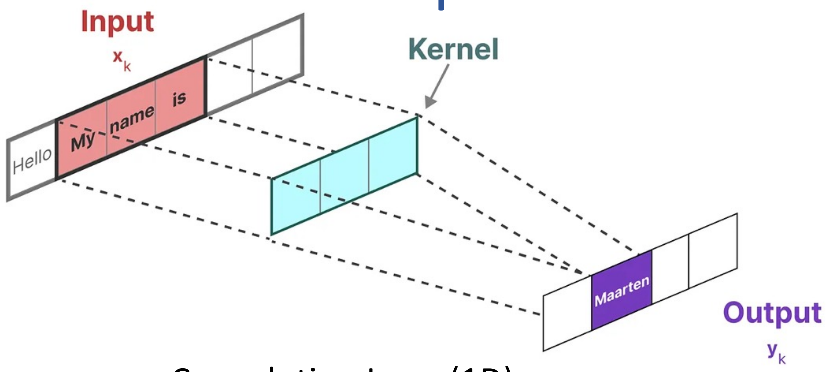
Recurrent Representation

Bilinear method:

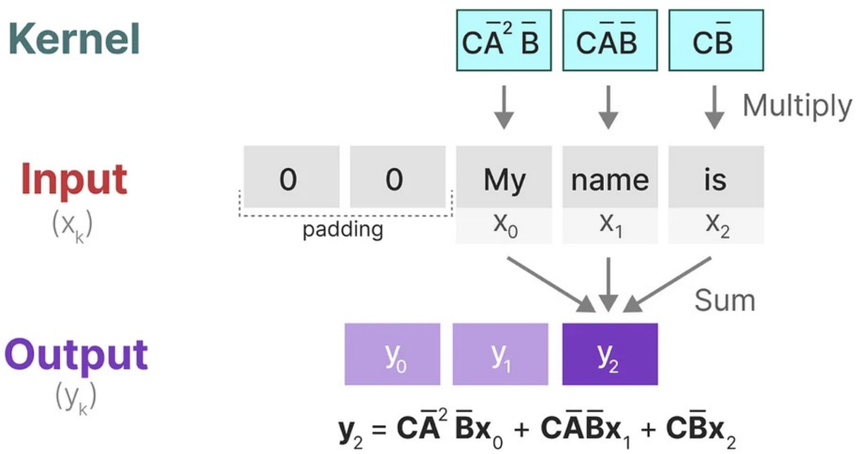
$$\bar{A} = \left(I - \frac{\Delta}{2} * A\right)^{-1}\left(I + \frac{\Delta}{2} * A\right)$$
$$\bar{B} = \left(I - \frac{\Delta}{2} * A\right)^{-1} \Delta B$$
$$\bar{C} = C$$



Three representations: Convolutional



Convolution Layer(1D)



$$y = \bar{K} * u$$

Convolutional Representation

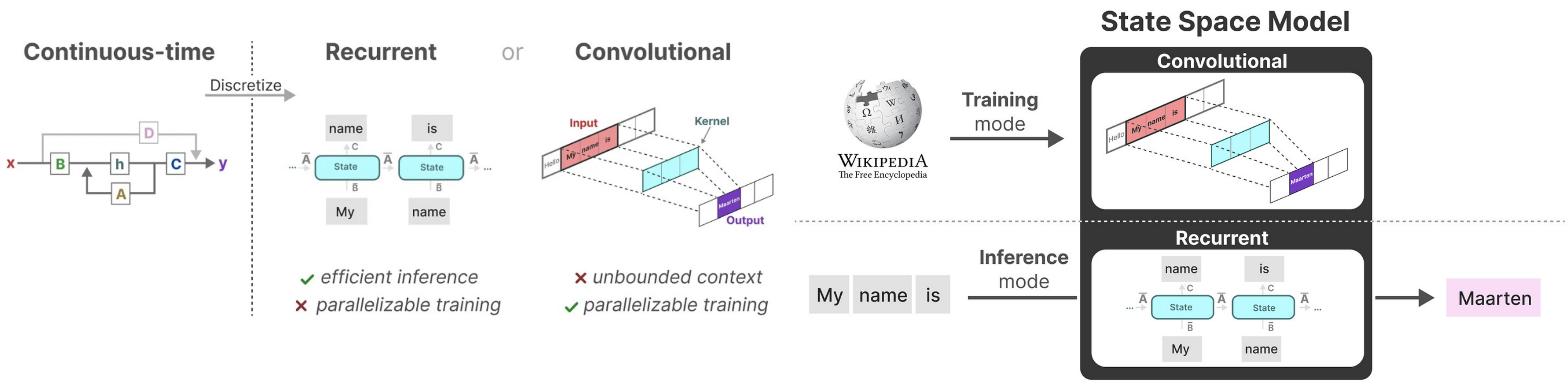
time	hidden	prediction
0	$\bar{B}x_0$	$C\bar{B}x_0$
1	$\bar{A}h_0 + \bar{B}x_1$	$C\bar{A}\bar{B}x_0 + C\bar{B}x_1$
2	$\bar{A}(\bar{A}h_0 + \bar{B}x_1) + \bar{B}x_2$	$C\bar{A}^2\bar{B}x_0 + C\bar{A}\bar{B}x_1 + C\bar{B}x_2$
…	…	…

If conv size is fixed, like conv size = 3,
Conv kernel should be: $\bar{K} = (C\bar{A}^{L-1}B, \ldots, C\bar{A}^2\bar{B}, C\bar{A}\bar{B}, C\bar{B})$

Difficulties: L times for computing $y = \bar{K} * x$
Solve: truncated generating function (convert the power of matrix to the inverse of matrix, like

$$y = fun_1(A) * fun_2(A^{-1}) * x$$

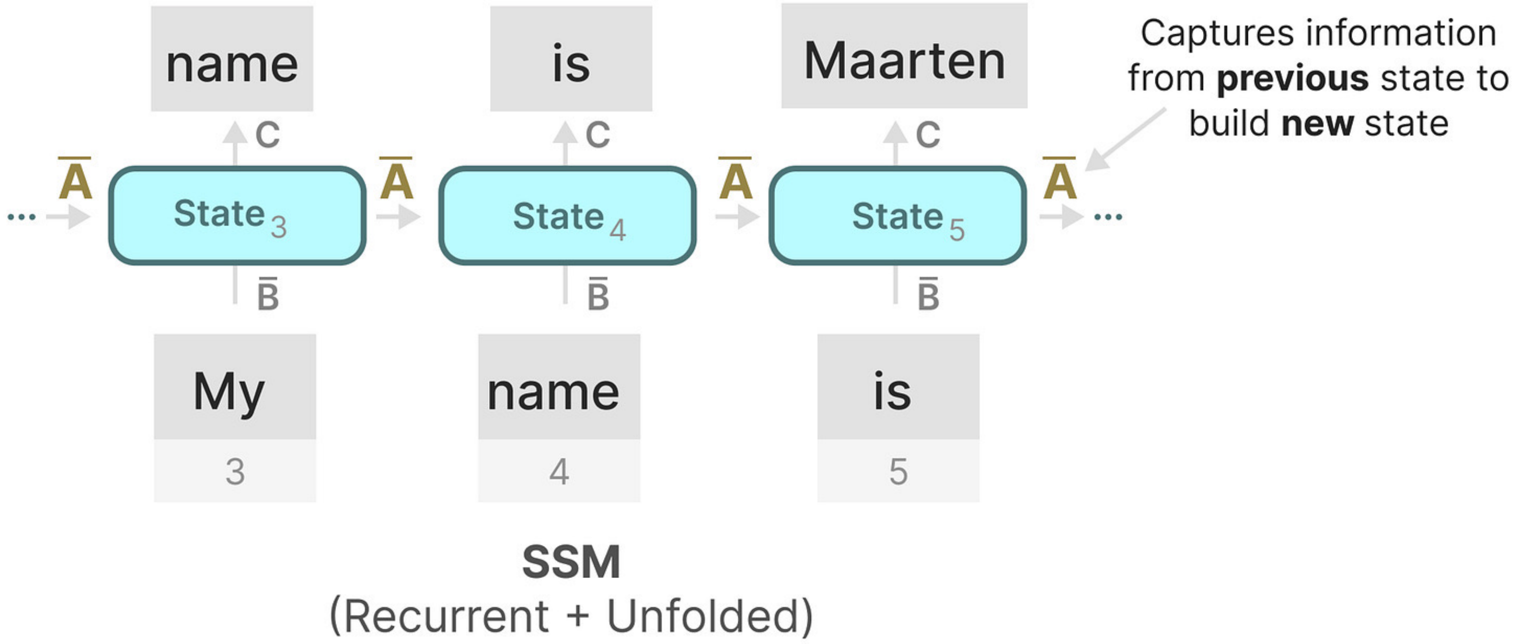
Summary of three representations



LSSL(Linear State-Space Layer)

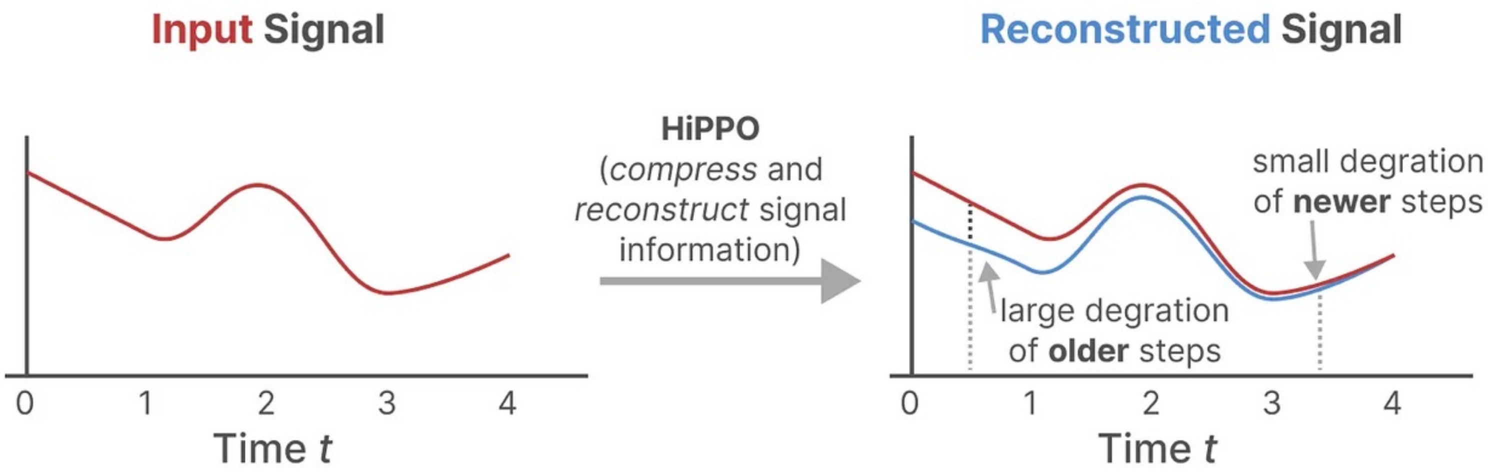
Property: Linear Time Invariance(LTI)

The importance of Matrix A



How can we get a Matrix A to retain the large memory (since it only look at previous states)

HIPPO Matrix



$$(\textbf{HiPPO Matrix}) \qquad \boldsymbol{A}_{nk} = -\begin{cases} (2n+1)^{1/2}(2k+1)^{1/2} & \text{if } n > k \\ n+1 & \text{if } n = k \\ 0 & \text{if } n < k \end{cases}.$$

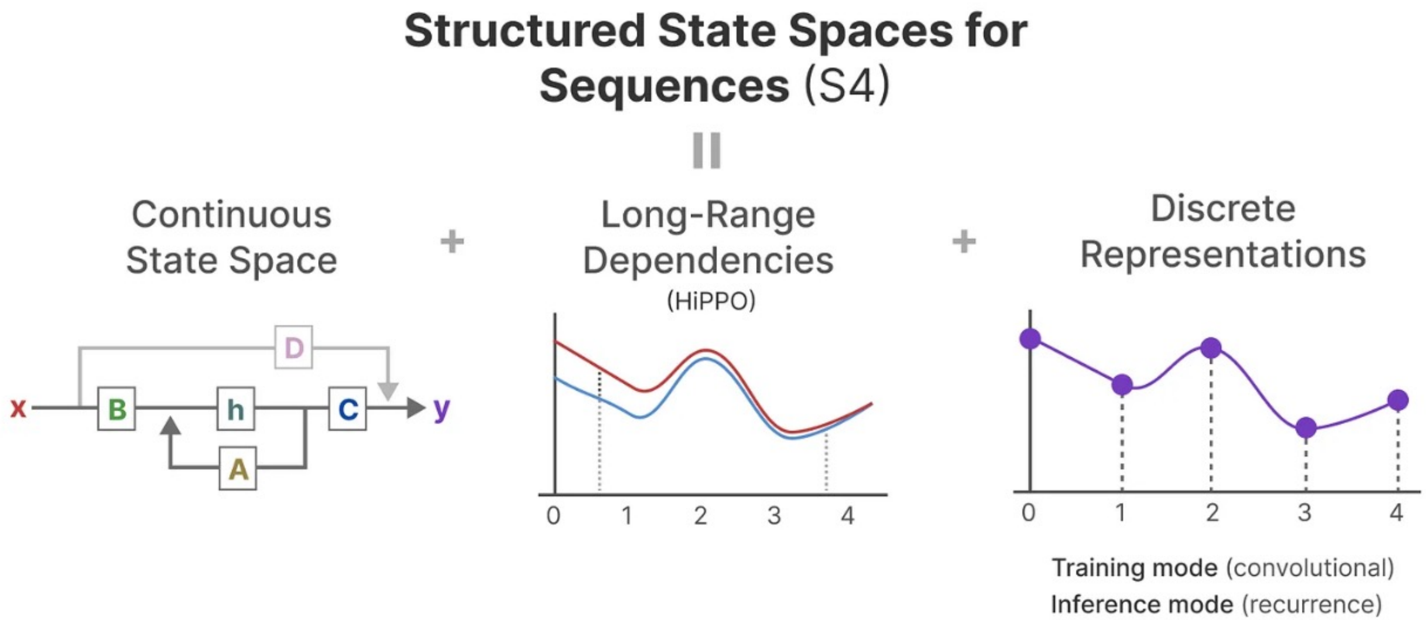
Mathematically, it does so by tracking the coefficients of a Legendre polynomial which allows it to approximate all of the previous history.

https://hazyresearch.stanford.edu/blog/2020-12-05-hippo

https://proceedings.neurips.cc/paper/2019/hash/952285b9b7e7a1be5aa7849f32ffff05-Abstract.html

Summary of S4

- State Space Model
- HiPPO for handling long-range dependencies
- Discretization for creating recurrent and convolution representations



Differ from basic SSMs:

(1) Modeling challenge for LRD: using a special formula for the A matrix (HiPPO).

(2) Computational Challenge: introducing a special representation and algorithm to be able to work with this matrix (truncated generating function)