Foundation Models

Instructor - Simon Lucey

RVSS - 2026

AUSTRALIAN
INSTITUTE FOR
MACHINE LEARNING

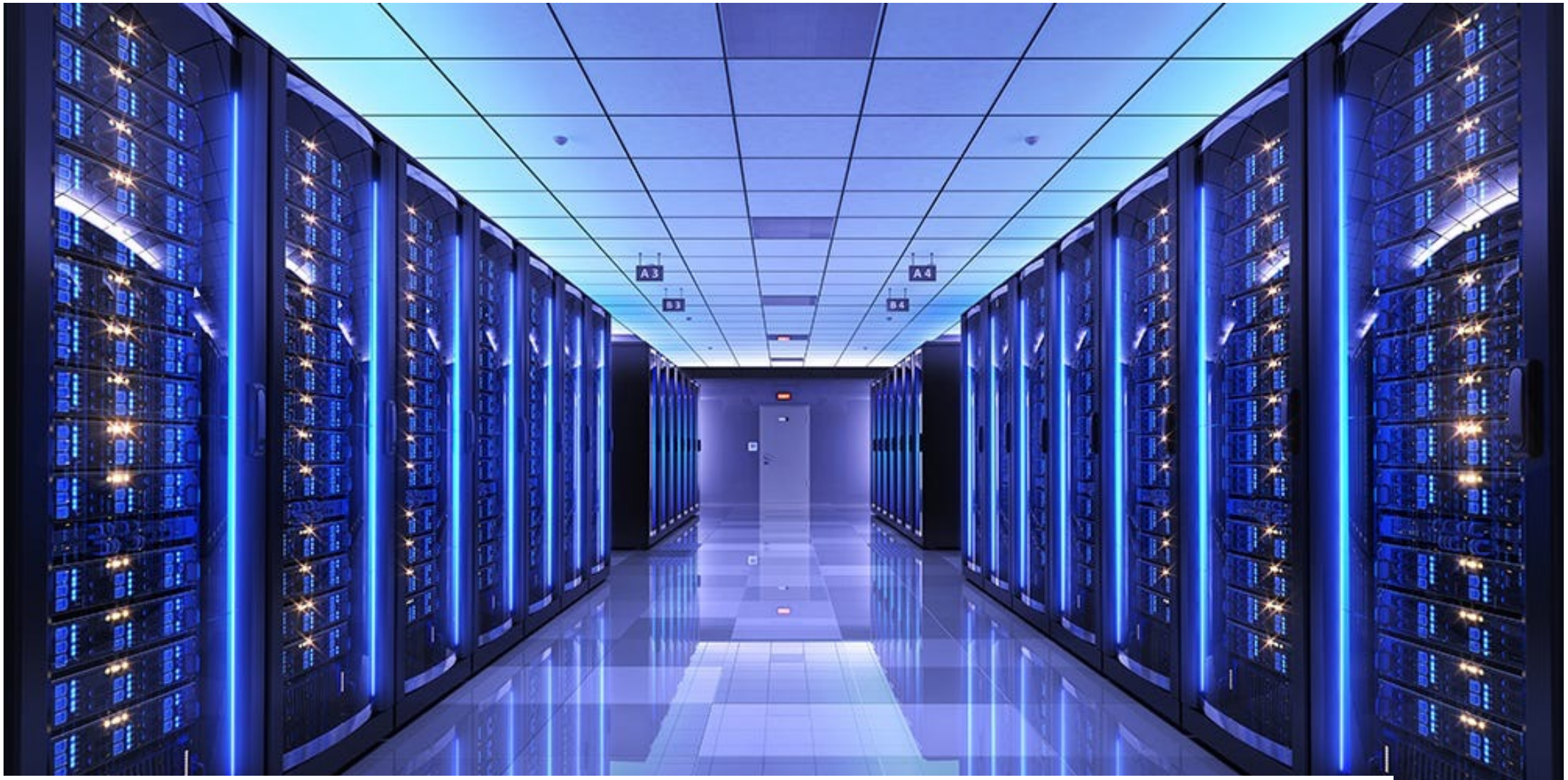
TECH· ELON MUSK

Elon Musk's just fired up Colossus—the world's largest Nvidia GPU supercomputer built in just three months from start to finish

BY CHRISTIAAN HETZNER
September 3, 2024 at 11:48 PM GMT+9:30

xAI founder Elon Musk aims to double the capacity of his Memphis investors could end up benefiting as well thanks to Optimus.
RICHARD BORD—WIREIMAGE/GETTY IMAGES

https://fortune.com/2024/09/03/elon-musk-xai-nvidia-colossus/



Contains 180,000 NVIDIA GPU (H100,H200,GB200) processors, estimated cost $U4.5 billion.

Will use 1.3 million gallons of water per day to cool servers, and consume 280 megawatts of power.

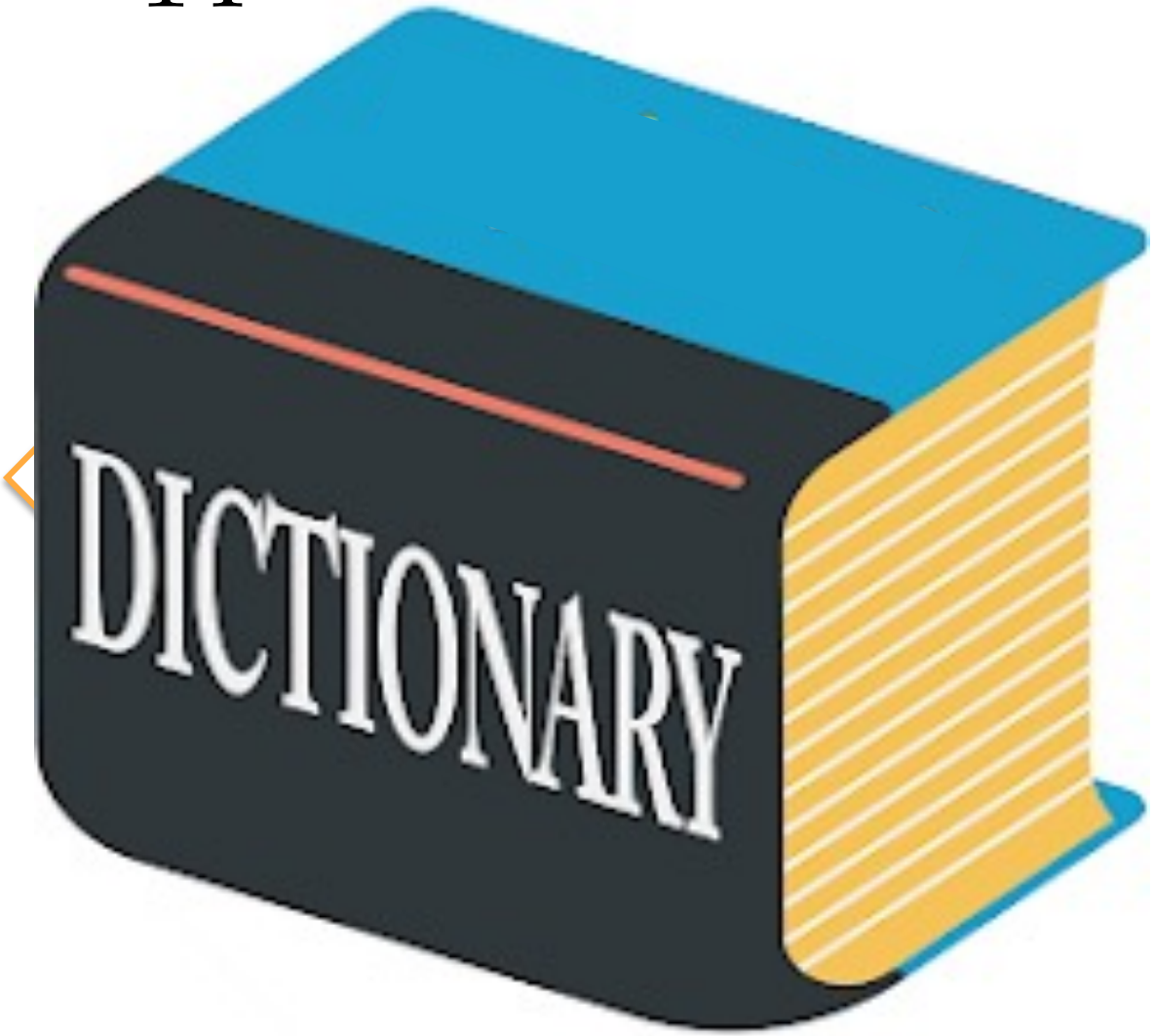
https://www.reuters.com/markets/deals/constellation-inks-power-supply-deal-with-microsoft-2024-09-20/

What is Large Language Model?

$D = 14$ "dictionary size"

"tokens"

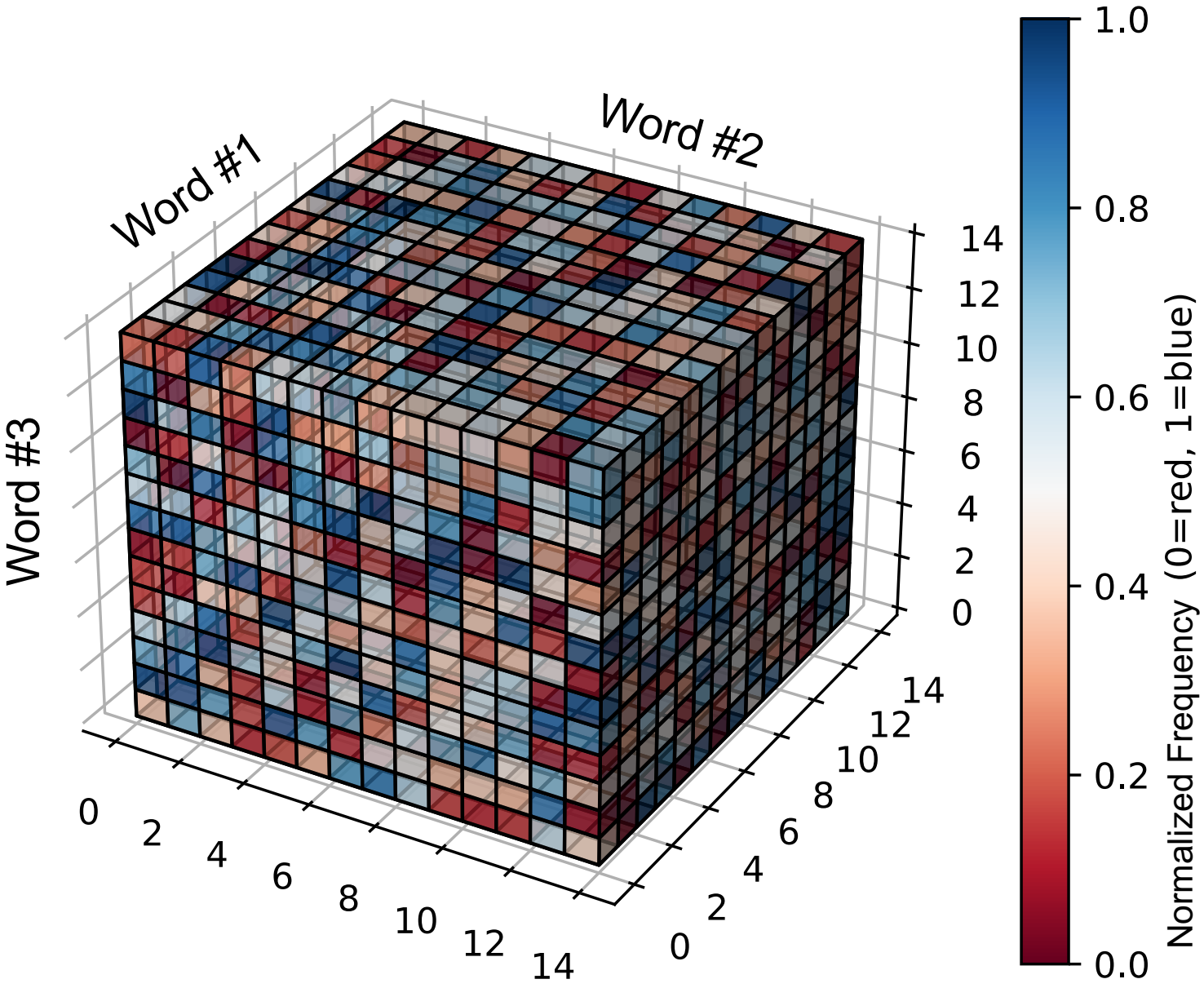
axiom
ate
aspic
are
anvil
antelope
ant
an
am
ally
abbatoir
abate
aardvark
a



Aardvark ate ant aspic abattoir. $W = 3$ "context window" rdvark am ant ally. Antelope are anvil axiom. Ant aspic ate, aardvark am. Antelope ally abate abattoir. Ant anvil axiom, aardvark ate. Antelope are aspic ally. Abattoir ant aardvark am. Antelope axiom ally abate. Aardvark ant aspic a telope ate a attoir a ally ant a Antelope ally abate anvil. Aardvark ant axiom are. Antelope aspic ate abatt am anvil ally. Aardvark ate antelope axiom. Antelope are abate as Aardvark am ant abattoir. Antelope ally anvil axiom. Ant ate aspic, are. Antelope abate ally am. Aardvark ant anvil ate. Antelope axio abattoir. Ant aspic ally aardvark. Antelope am abate anvil. Aardva axiom. Antelope are ally abattoir. Ant aspic am, aardvark ate. Ante anvil axiom ally.

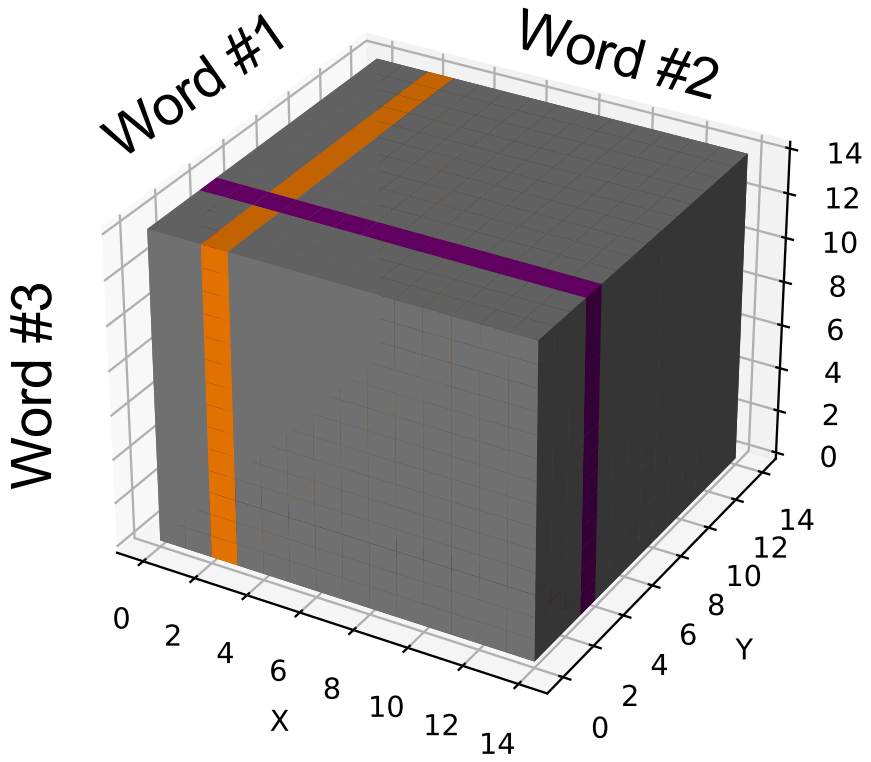
"The future state depends only on the past W-1 states."

Andrey Markov

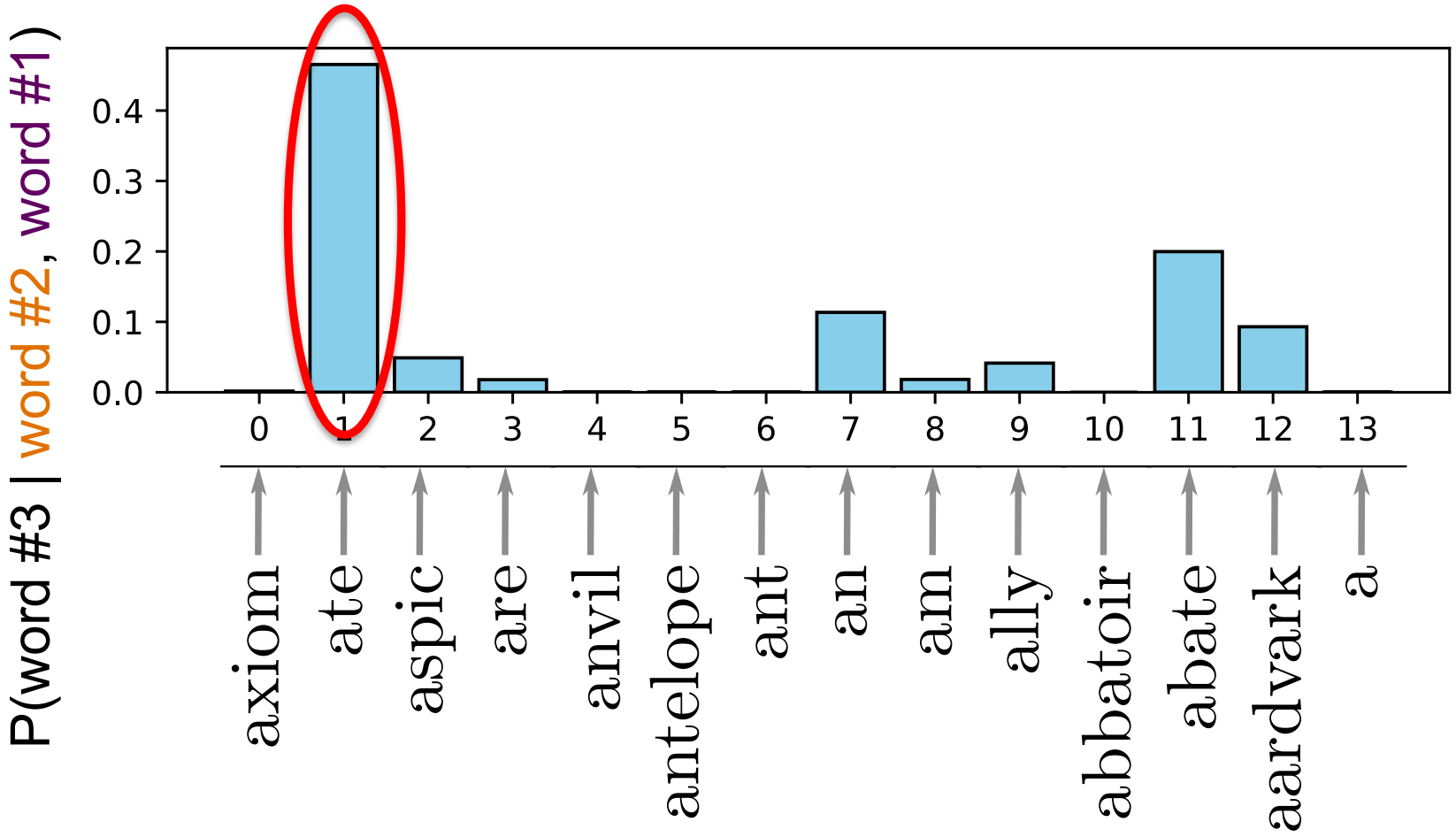

Word #1
Word #2
Word #3
0
2
4
6
8
10
12
14
Normalized Frequency (0=red, 1=blue)
0.0
0.2
0.4
0.6
0.8
1.0

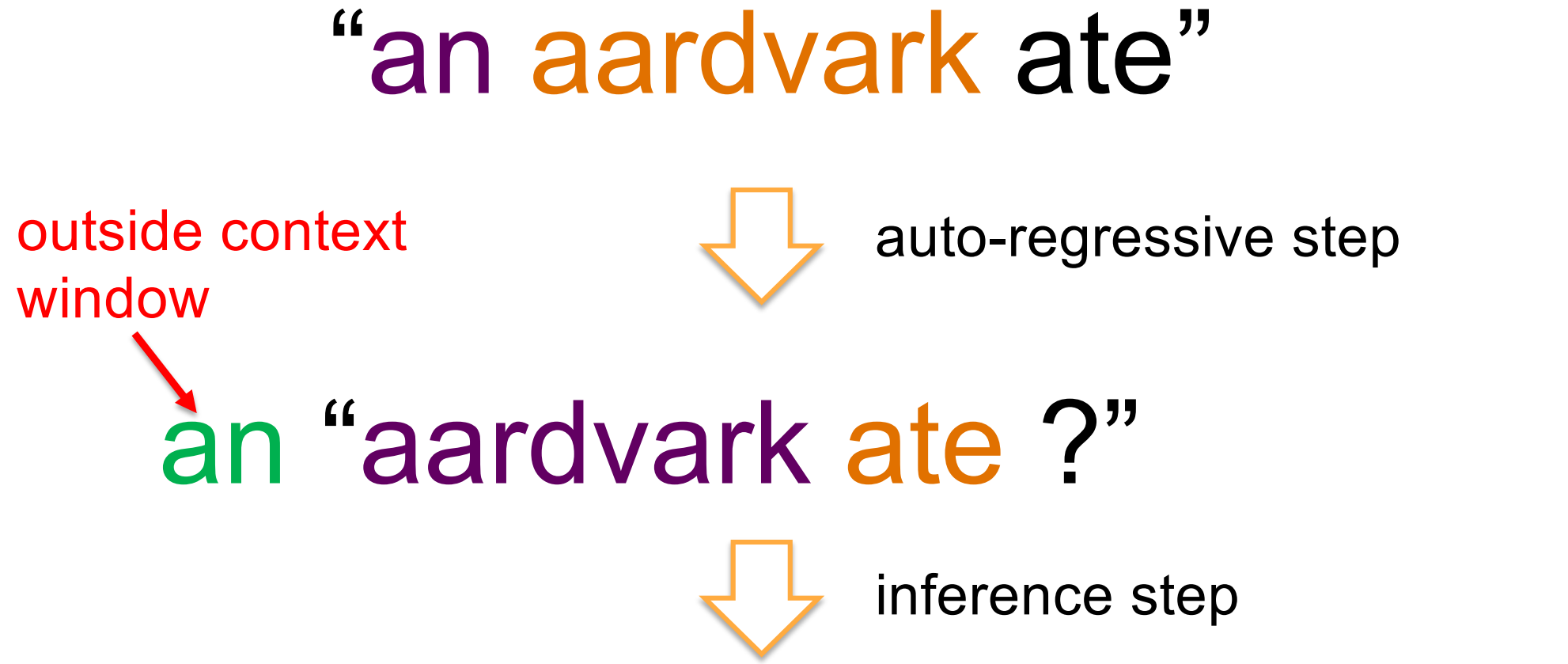
“an aardvark ?”

P(word #3 | word #2, word #1)



“an aardvark ate”




“an aardvark ate”
outside context window
auto-regressive step
an “aardvark ate ?”
inference step
word #3 ← max P(word #3 | word #2, word #1)
word #3


“an aardvark ate”
outside context window
auto-regressive step
an “aardvark ate ?”
inference step
an “aardvark ate an”

auto-regressive step

inference step

If this works why do we need AI?

“Dictionary Size”

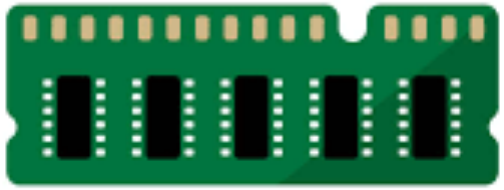
$D = 14$

“Context window”

Aardvark ate ant

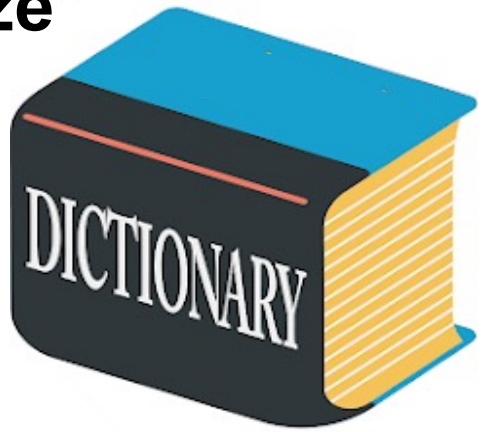
$W = 3$

“Memory”



$EBWKb$

“GPT-3 Dictionary Size”



$$D = 50,000$$

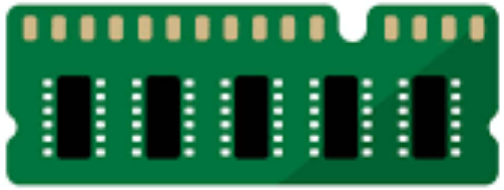
“GPT-3 Context window”

Mind! I don't mean to say that I know, of my own knowledge, what there is particularly dead about a door-nail. I might have been inclined, myself, to regard a coffin-nail as the deadest piece of ironmongery in the trade. But the wisdom of our ancestors is in the simile; and my unhallowed hands shall not disturb it, or the Country's done for. You will therefore permit me to repeat, emphatically, that Marley was as dead as a door-nail.

$$W = 2048$$

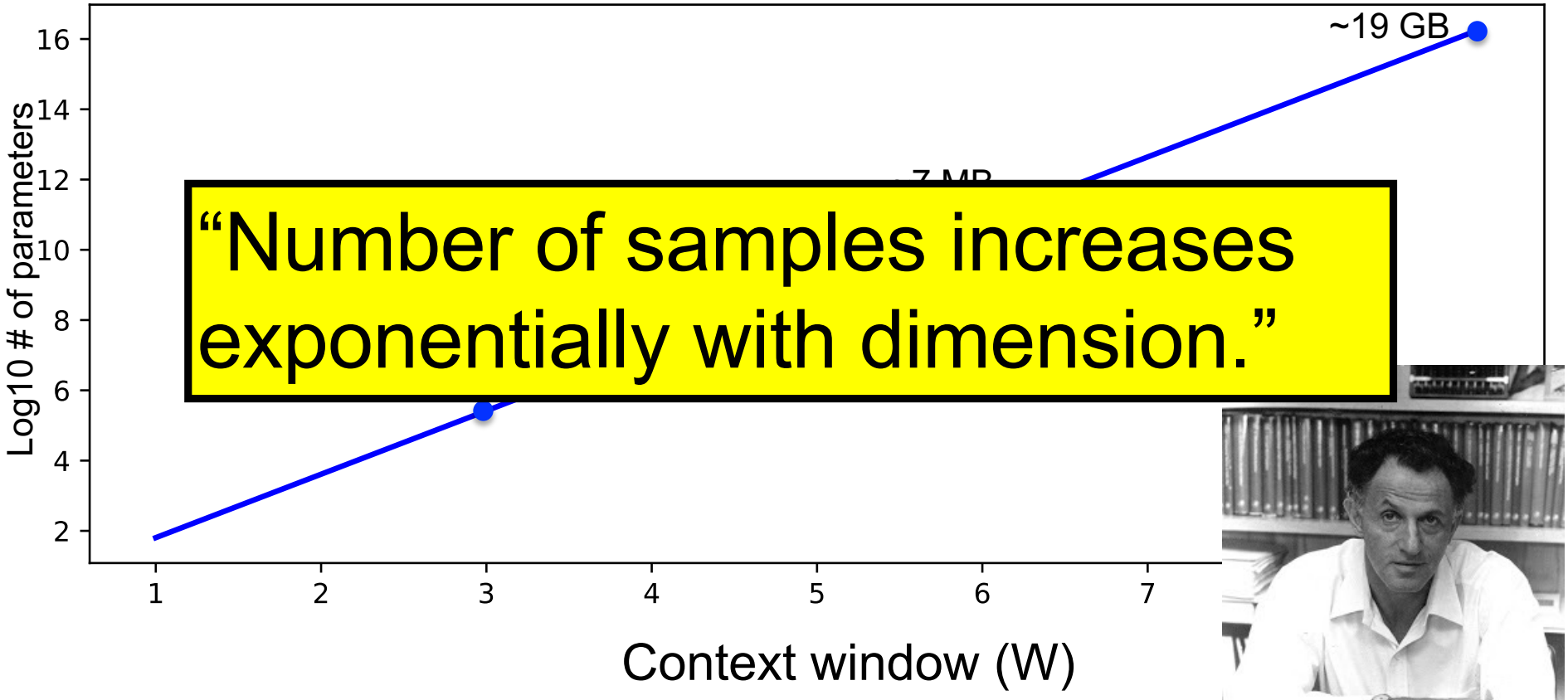
Only 10^{80} atoms in the universe!!

“Memory”



$$10^{9600}\text{bytes}$$

Curse of Dimensionality (D = 14)

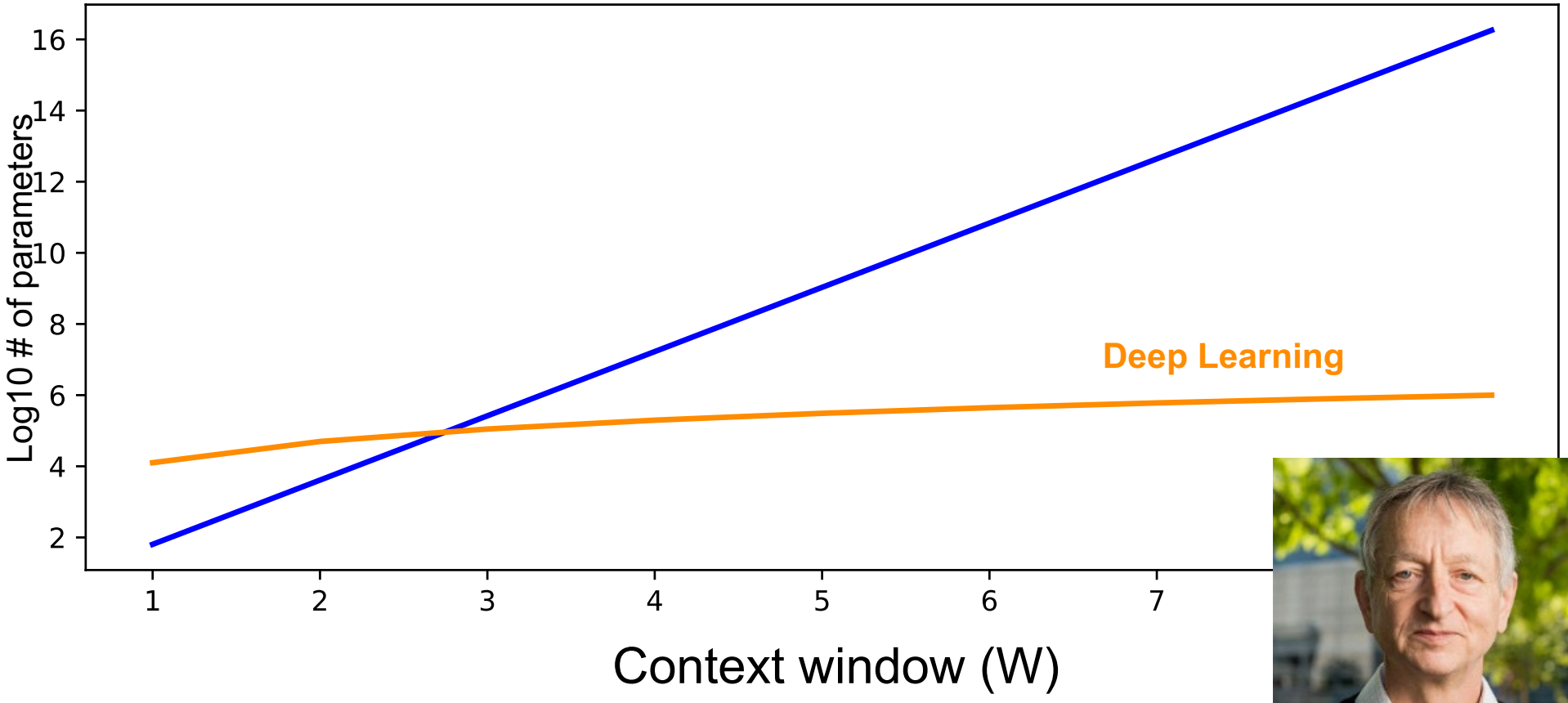


"Number of samples increases exponentially with dimension."

# of parameters = D^W

Richard E. Bellman

Deep Learning – A breakthrough in sampling?

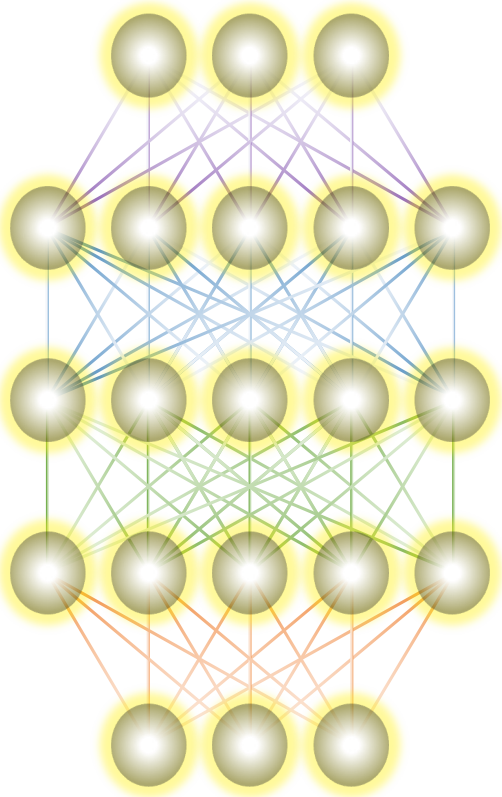


$$\text{\# of parameters} = D^W$$

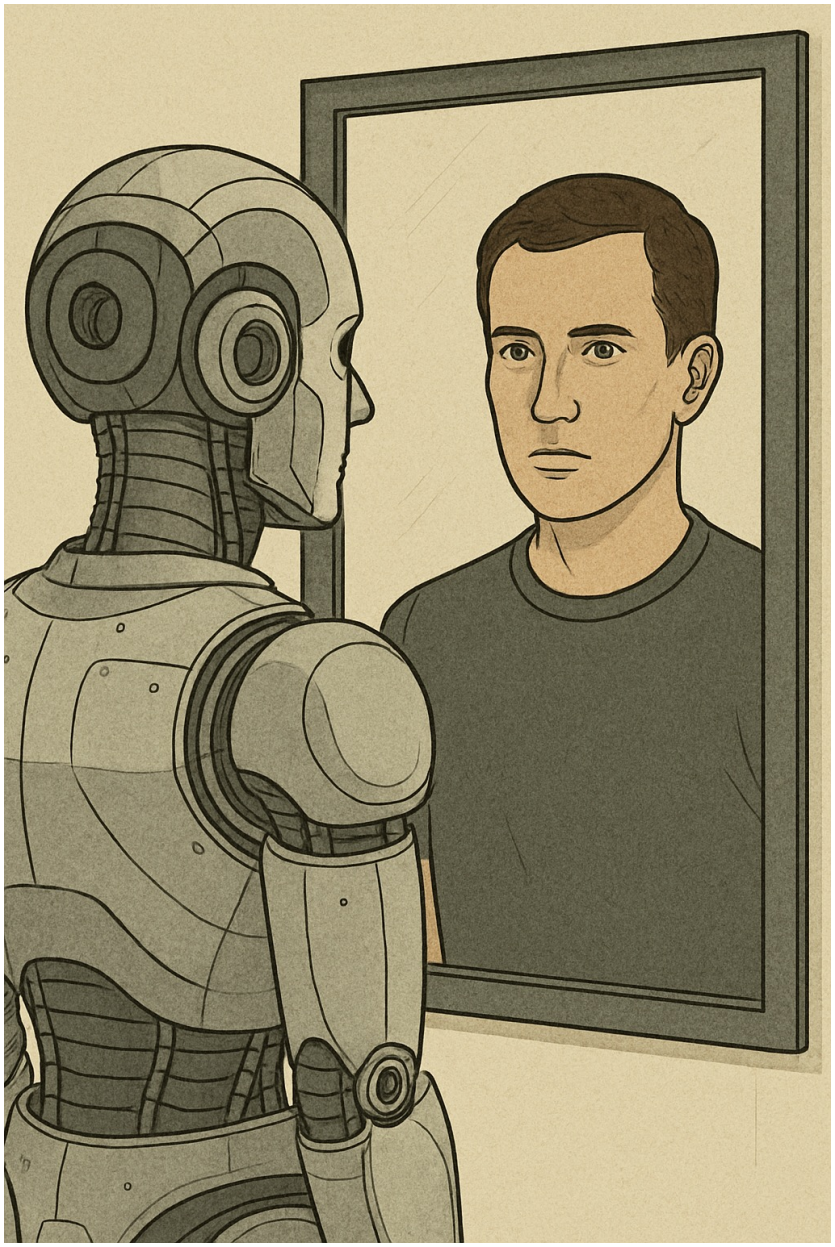
"Geoffrey Hinton"

Deep Learning – Current Open Questions?

- **Why does it generalize so well?**
- **Why does it not get caught in local minima?**
- **What are the limits of scaling laws?**
- **Why does it hallucinate?**
- **How can we train more efficiently?**

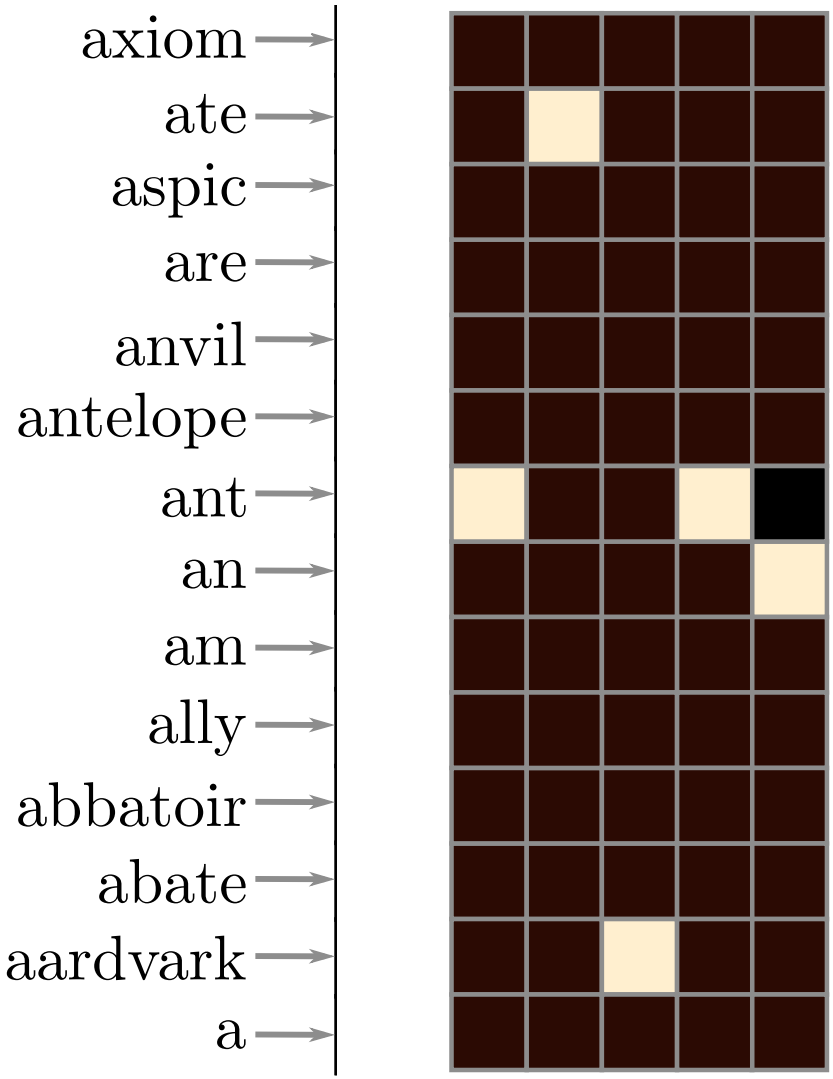


- **Is AI just a reflection of our own intelligence?**



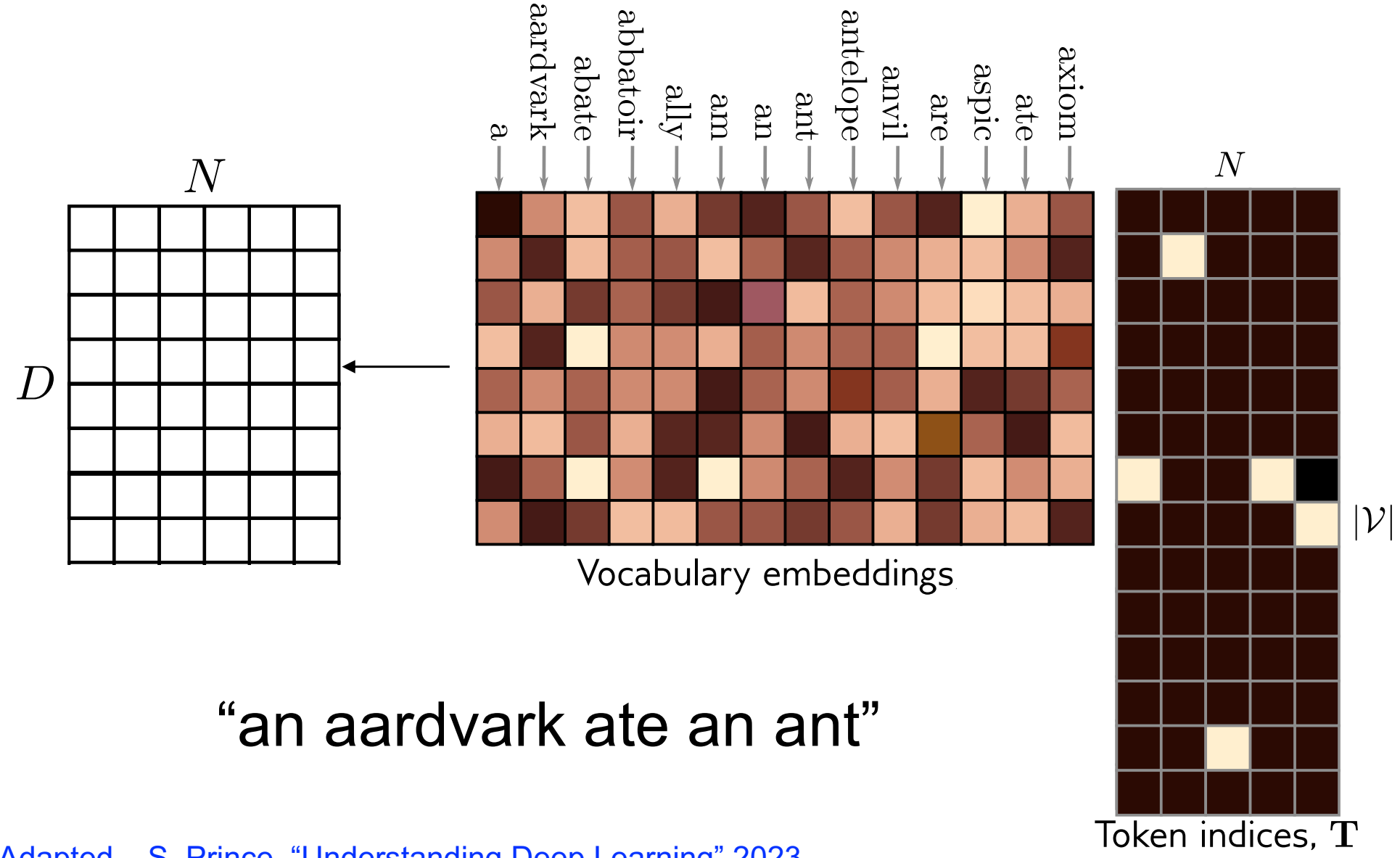
Efficiency through Understanding

“an aardvark ate an ant”



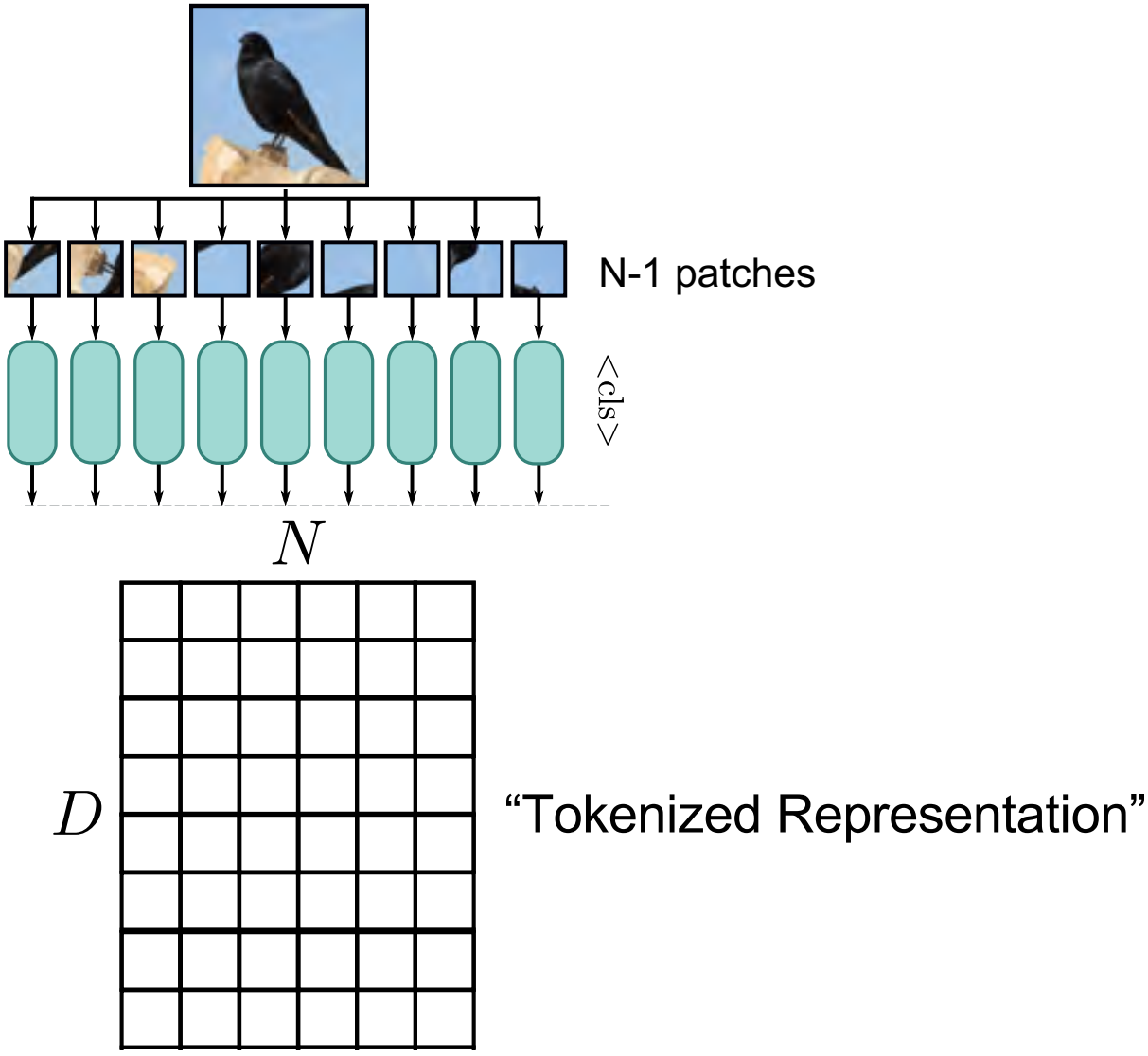
“an aardvark ate an ant”

Adapted – S. Prince. “Understanding Deep Learning” 2023.

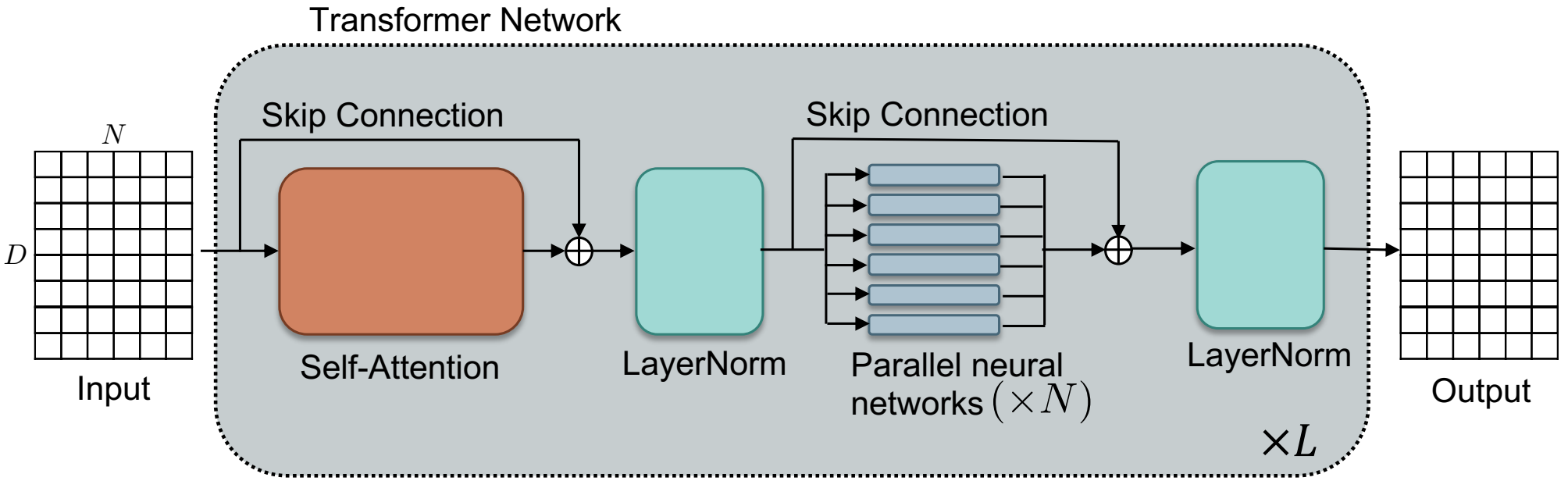


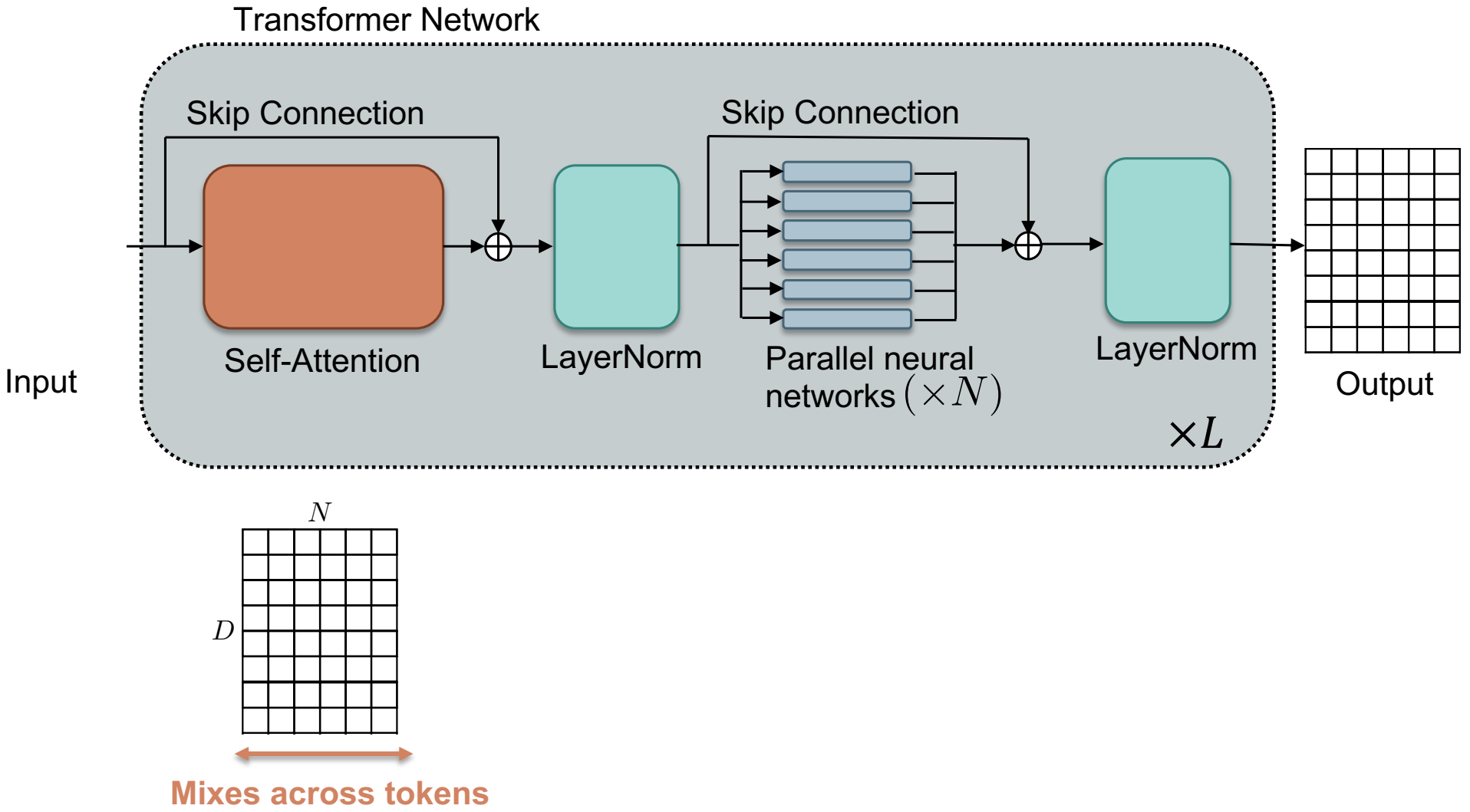
"an aardvark ate an ant"

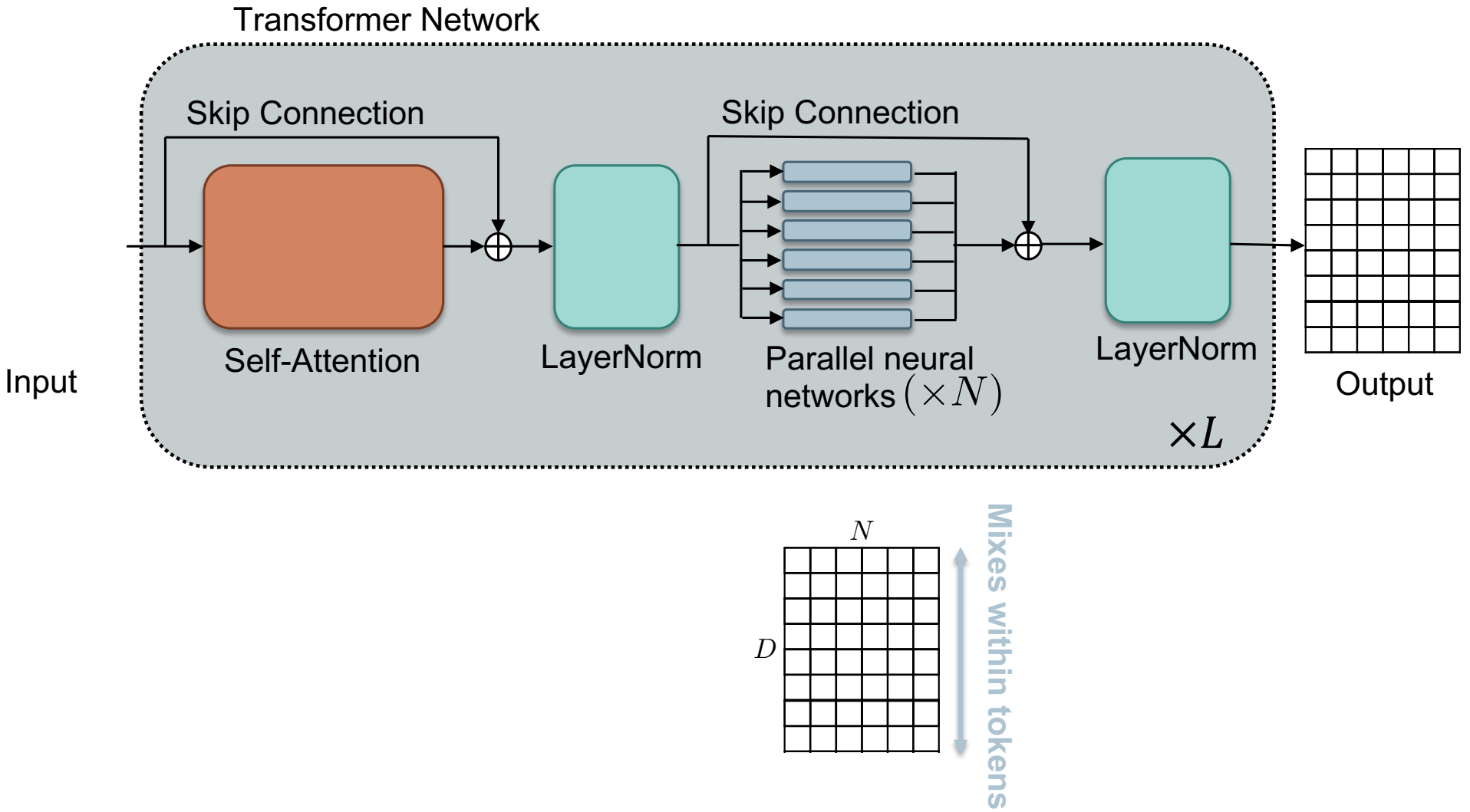
Adapted – S. Prince. "Understanding Deep Learning" 2023.

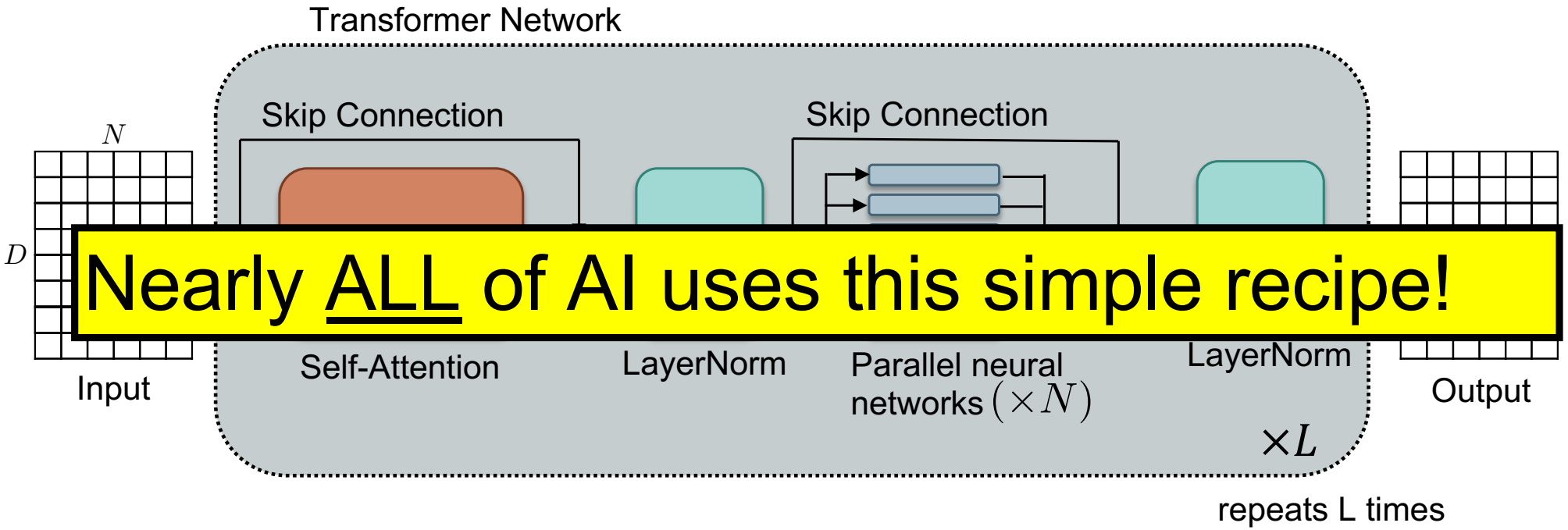


Adapted – S. Prince. “Understanding Deep Learning” 2023.

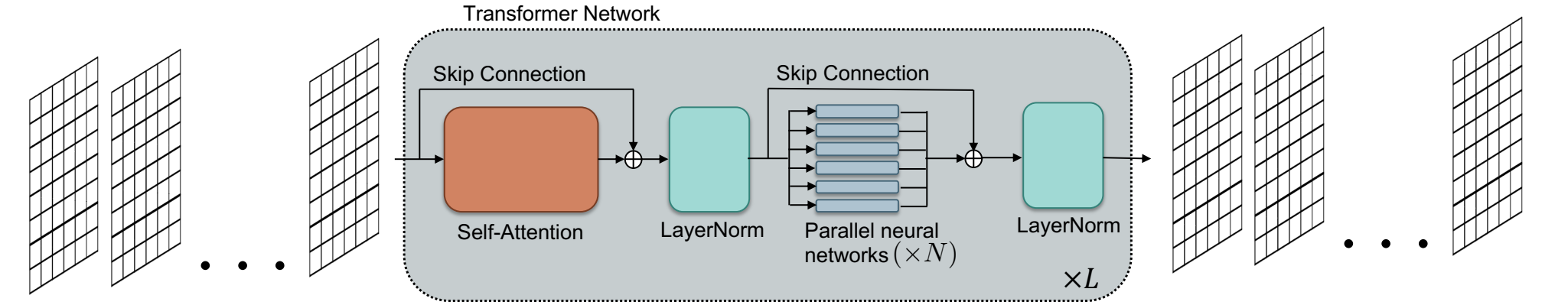

Transformer Network
N
D
Input
Skip Connection
Self-Attention
LayerNorm
Skip Connection
Parallel neural networks $(\times N)$
LayerNorm
$\times L$
Output


Transformer Network
Skip Connection
Skip Connection
Input
Self-Attention
LayerNorm
Parallel neural networks $(\times N)$
LayerNorm
Output
$\times L$
N
D
Mixes across tokens


Transformer Network
Skip Connection
Skip Connection
Input
Self-Attention
LayerNorm
Parallel neural networks $(\times N)$
LayerNorm
Output
$\times L$
N
D
Mixes within tokens



Can I know in advance how much it costs to train?



Minimize loss

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \mathcal{L}\{\mathbf{y}_m; f(\mathbf{x}_m; \boldsymbol{\theta})\}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \mathcal{L}\{\mathbf{y}_m; f(\mathbf{x}_m; \boldsymbol{\theta})\}$$

Gradient Descent

$$\boldsymbol{\theta}_{k+1} \rightarrow \boldsymbol{\theta}_k - \alpha \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_k)$$

"Learning Rate"

$$\nabla \mathcal{L}(\boldsymbol{\theta}_k) = [\nabla f(\mathbf{x}_1; \boldsymbol{\theta}_k), \ldots, \nabla f(\mathbf{x}_M; \boldsymbol{\theta}_k)] \begin{bmatrix} f(\mathbf{x}_1; \boldsymbol{\theta}_k) - \mathbf{y}_1 \\ \vdots \\ f(\mathbf{x}_M; \boldsymbol{\theta}_k) - \mathbf{y}_M \end{bmatrix}$$

Jacobian matrix

error vector

$$\nabla \mathcal{L}(\boldsymbol{\theta}_k) = \mathbf{J}_k \boldsymbol{\epsilon}_k$$

Jacobian matrix error vector

“Least-Squares Objective”

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \frac{1}{2} ||\mathbf{y}_m - f(\mathbf{x}_m; \boldsymbol{\theta})||_2^2$$

Gradient Descent

$$\boldsymbol{\theta}_{k+1} \rightarrow \boldsymbol{\theta}_k - \alpha \cdot \mathbf{J}_k \boldsymbol{\epsilon}_k$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{1}{2} ||\mathbf{y}_n - f(\mathbf{x}_n; \boldsymbol{\theta})||_2^2$$

Gradient Descent

$$\boldsymbol{\theta}_{k+1} \rightarrow \boldsymbol{\theta}_k - \alpha \cdot \mathbf{J} \boldsymbol{\epsilon}_k$$

"Static Jacobian"

If we start with $\boldsymbol{\theta}_0 = \mathbf{0}$

$$\boldsymbol{\theta}_1 \rightarrow -\alpha \cdot \mathbf{J}\boldsymbol{\epsilon}_0$$

$$\boldsymbol{\theta}_2 \rightarrow -(\mathbf{I} - \alpha\mathbf{J}\mathbf{J}^T)\alpha\mathbf{J}\boldsymbol{\epsilon}_0 - \alpha\mathbf{J}\boldsymbol{\epsilon}_0$$

$$\boldsymbol{\theta}_3 \rightarrow -(\mathbf{I} - \alpha\mathbf{J}\mathbf{J}^T)^2\alpha\mathbf{J}\boldsymbol{\epsilon}_0 - (\mathbf{I} - \alpha\mathbf{J}\mathbf{J}^T)\alpha\mathbf{J}\boldsymbol{\epsilon}_0 - \alpha\mathbf{J}\boldsymbol{\epsilon}_0$$

at the q-th iteration generalizes to,

$$\boldsymbol{\theta}_q = -\sum_{k=0}^{q}(\mathbf{I} - \alpha\mathbf{J}\mathbf{J}^T)^{k-1}\alpha\mathbf{J}\boldsymbol{\epsilon}_0$$

$$\boldsymbol{\theta}_q = -\sum_{k=0}^{q}(\mathbf{I} - \alpha\mathbf{J}\mathbf{J}^T)^{k-1}\alpha\mathbf{J}\boldsymbol{\epsilon}_0$$

$$\mathbf{J} = \mathbf{U}\,\mathrm{diag}(\mathbf{s})\,\mathbf{V}^T$$ "SVD"

$$\boldsymbol{\theta}_q = -\,\mathbf{U}\,\mathrm{diag}(\hat{\mathbf{s}}_q)\,\mathbf{V}^T\,\boldsymbol{\epsilon}_0$$ "closed form solution"

$$\boldsymbol{\theta}_q = -\,\mathbf{U}\,\mathrm{diag}(\hat{\mathbf{s}}_q)\,\mathbf{V}^T\,\boldsymbol{\epsilon}_0$$

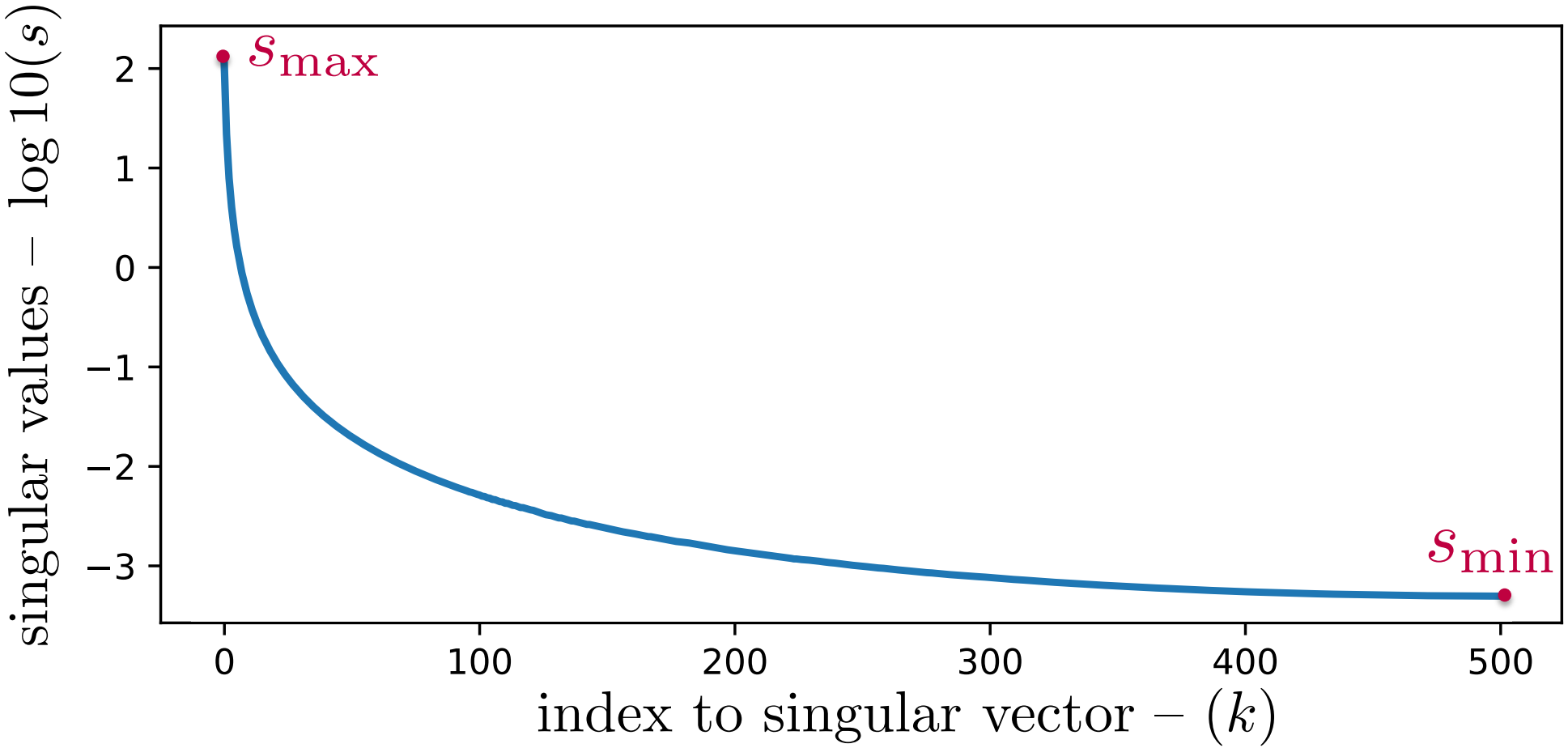
$$\hat{\mathbf{s}}_q[k] \equiv \alpha\,\mathbf{s}[k]\,\frac{1-\left(1-\alpha\,\mathbf{s}[k]^2\right)^q}{\sum_{j=0}^{q-1}\left(1-\alpha\,\mathbf{s}[k]^2\right)^j}$$

"GD shrinkage"

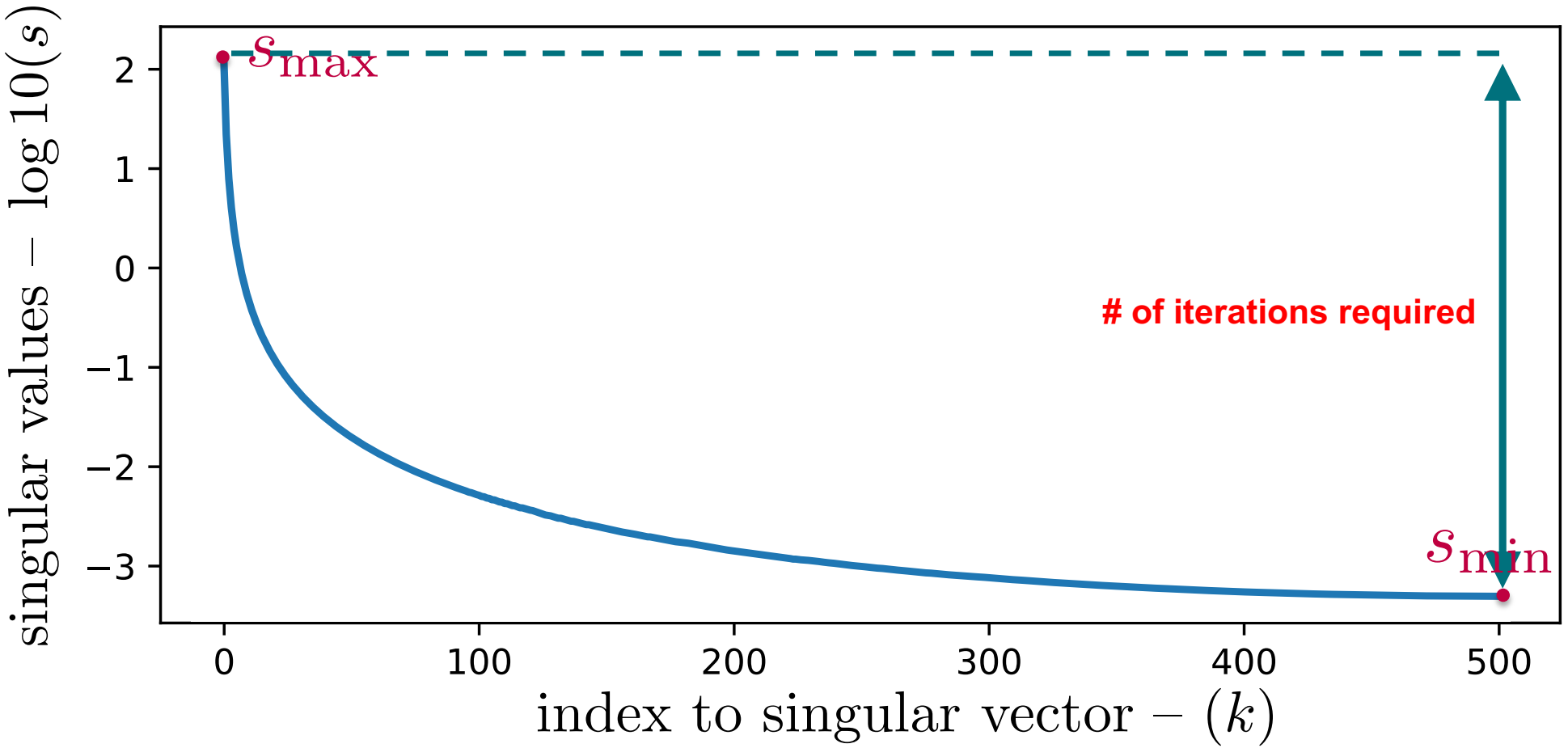
$$\left(1-\alpha\,\mathbf{s}[k]^2\right)^q \;\approx\; \exp\left(-\,q\,\alpha\,\mathbf{s}[k]^2\right)$$

A. Jacot et al. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks" in NeurIPS 2018.
S. Lucey. "Gradient Descent as a Shrinkage Operator for Spectral Bias" in arXiv 2025.

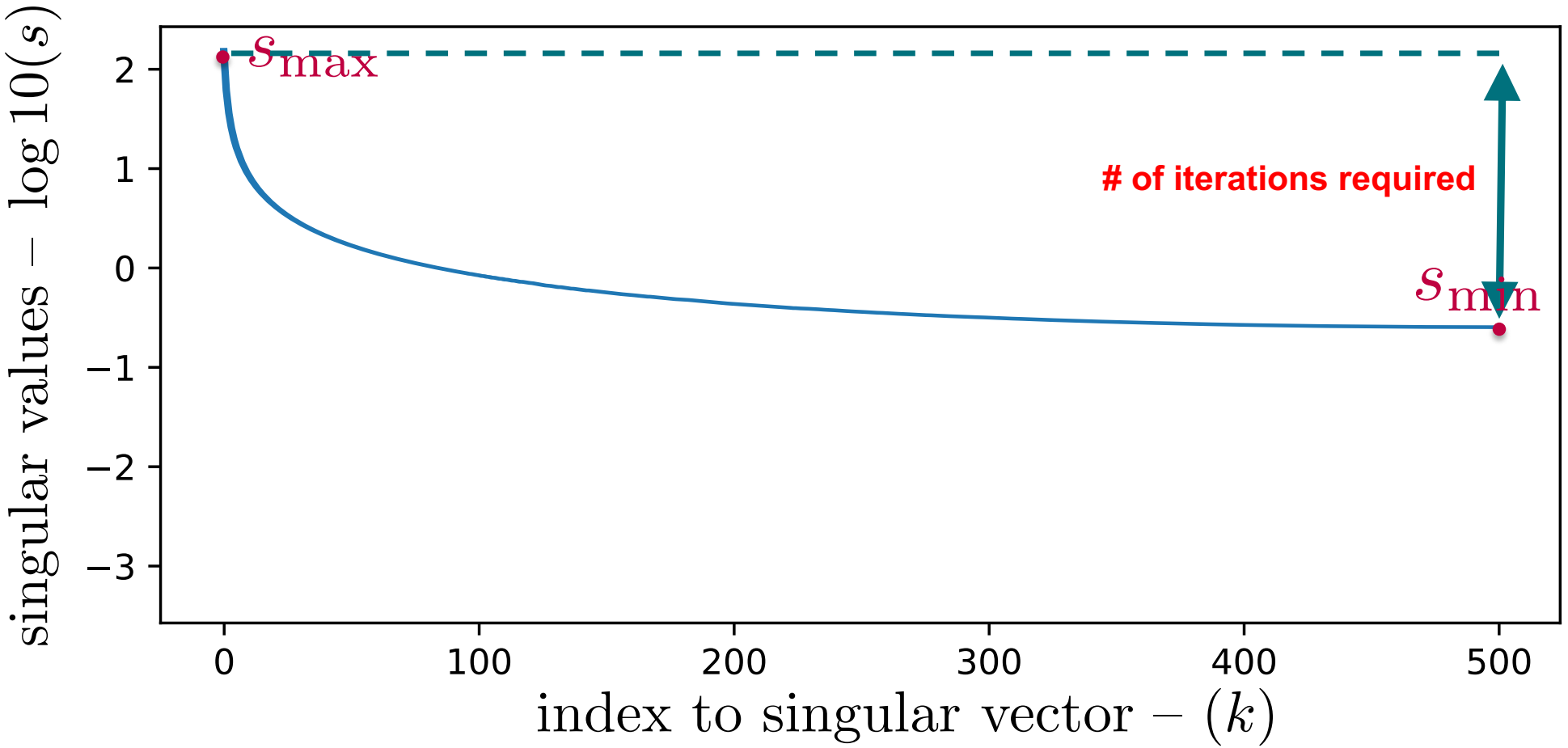
$$\mathrm{cond}(\mathbf{J}) = \frac{s_{\max}^2}{s_{\min}^2}$$



$\text{cond}(\mathbf{J}) \propto \#$ of iterations required in GD

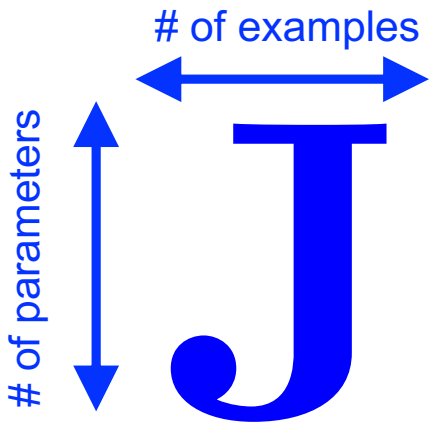


$\text{cond}(\mathbf{J}) \propto \#$ of iterations required in GD

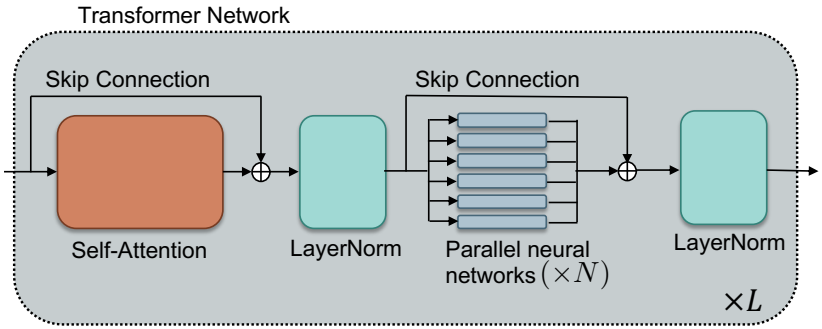


Insights….

- Condition of the Jacobian is a useful heuristic even for SGD and Adam.
- However, problematic forming the Jacobian for modern large-scale transformers.
- Could we decompose the Jacobian?



$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_0 \\ \vdots \\ \boxed{\mathbf{J}_\ell} \\ \mathbf{J}_{\ell+1} \\ \vdots \\ \mathbf{J}_{2L-1} \end{bmatrix}$$



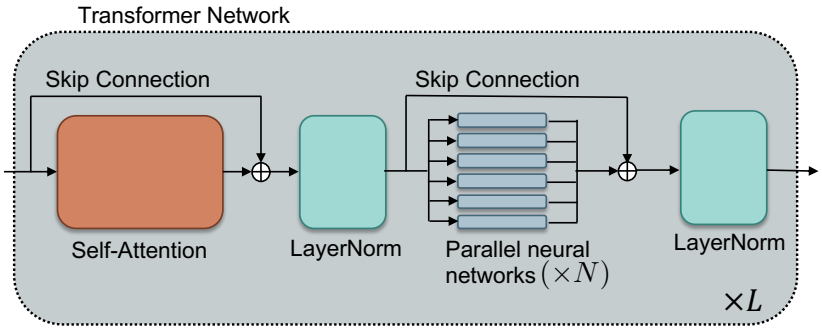
$$\mathbf{X} \rightarrow \mathrm{SA}(\mathbf{X};\boldsymbol{\theta}_\ell) \rightarrow \mathbf{X}_{\mathrm{out}}$$

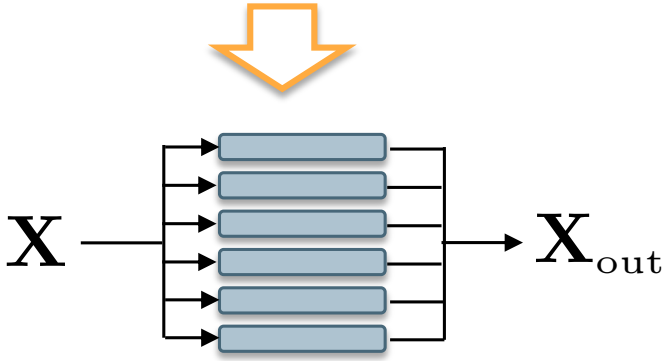
Self-Attention

$$\boxed{\mathbf{J}_\ell} = \nabla \mathrm{SA}(\mathbf{X};\boldsymbol{\theta}_\ell)$$

Jacobian of Self-Attention Block (SAB)

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_0 \\ \vdots \\ \mathbf{J}_\ell \\ \boxed{\mathbf{J}_{\ell+1}} \\ \vdots \\ \mathbf{J}_{2L-1} \end{bmatrix}$$





$$\boxed{\mathbf{J}_{\ell+1}} = \nabla \text{FFN}(\mathbf{X}; \boldsymbol{\theta}_{\ell+1})$$

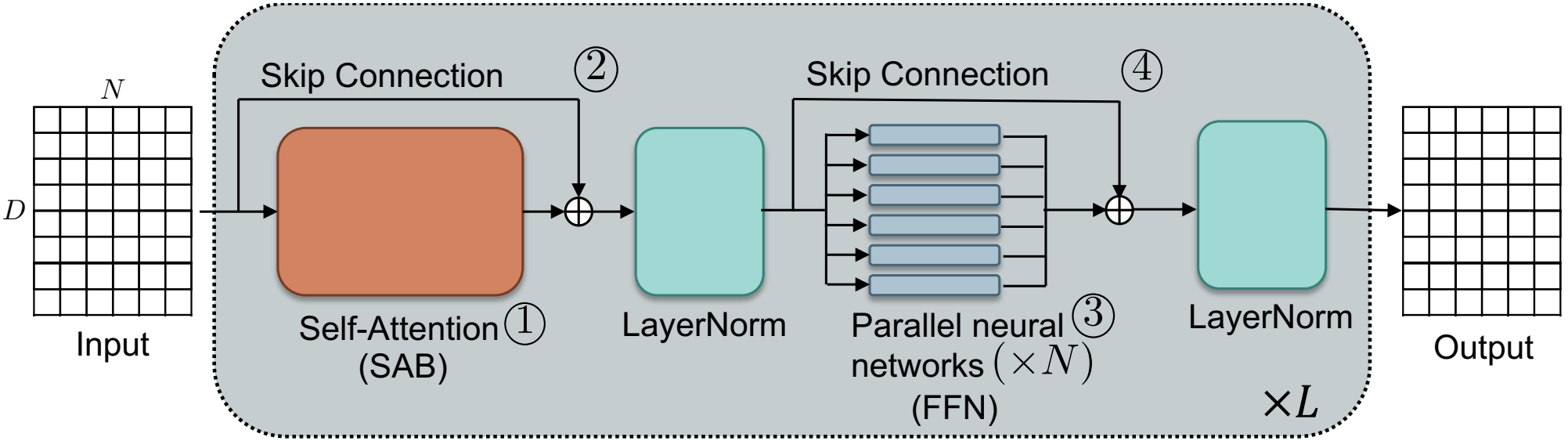
Jacobian of Feed Forward Network (FFN)

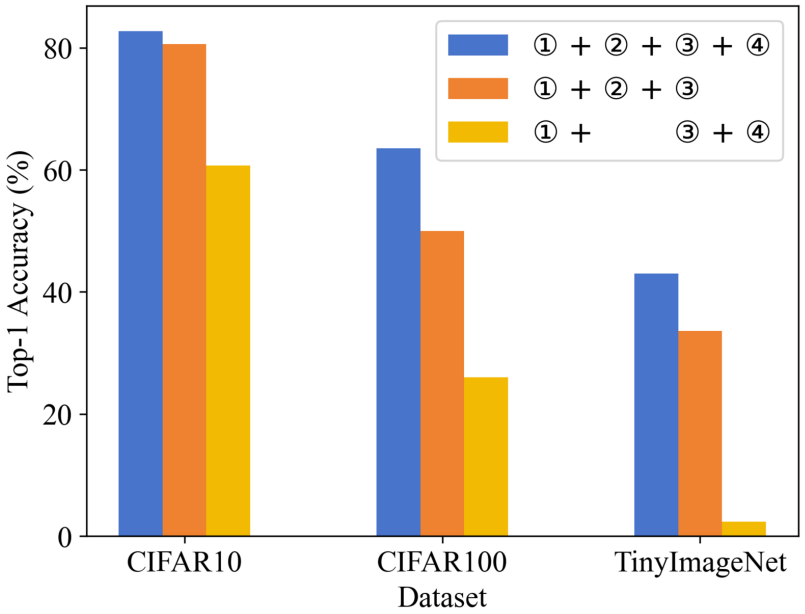
$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_0 \\ \vdots \\ \mathbf{J}_\ell \\ \mathbf{J}_{\ell+1} \\ \vdots \\ \mathbf{J}_{2L-1} \end{bmatrix}$$

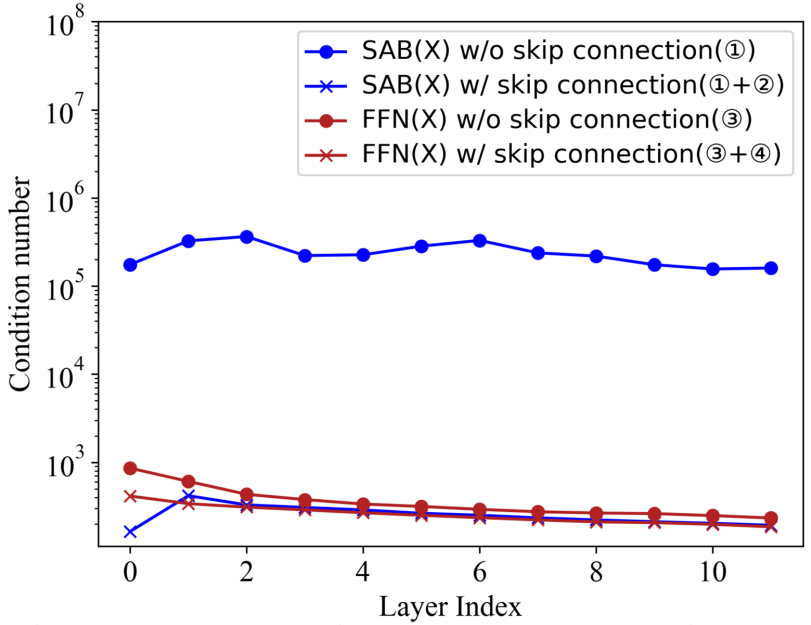
Simplifying assumption:

$$\Rightarrow \quad \mathrm{cond}(\mathbf{J}) \leq \max_{\ell} \mathrm{cond}(\mathbf{J}_\ell)$$

Y. Ji, H. Saratchandran, S. Lucey, “Always Skip Attention.” In ICCV 2025.



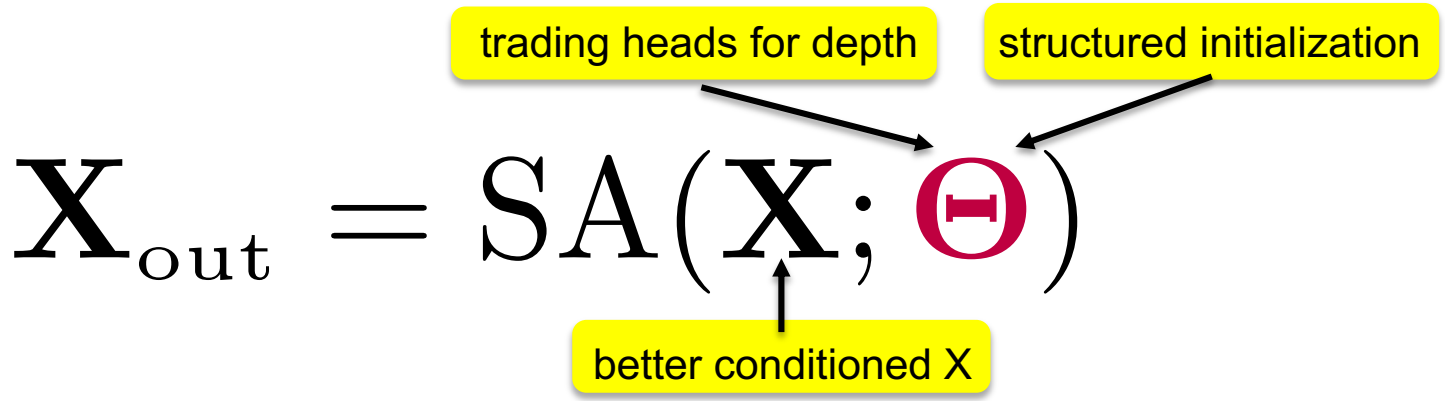




Y. Ji, H. Saratchandran, S. Lucey, “Always Skip Attention.” In ICCV 2025.

New Directions….

- Designing transformers to maximize sub-block condition in self-attention has opened up brand new-directions.

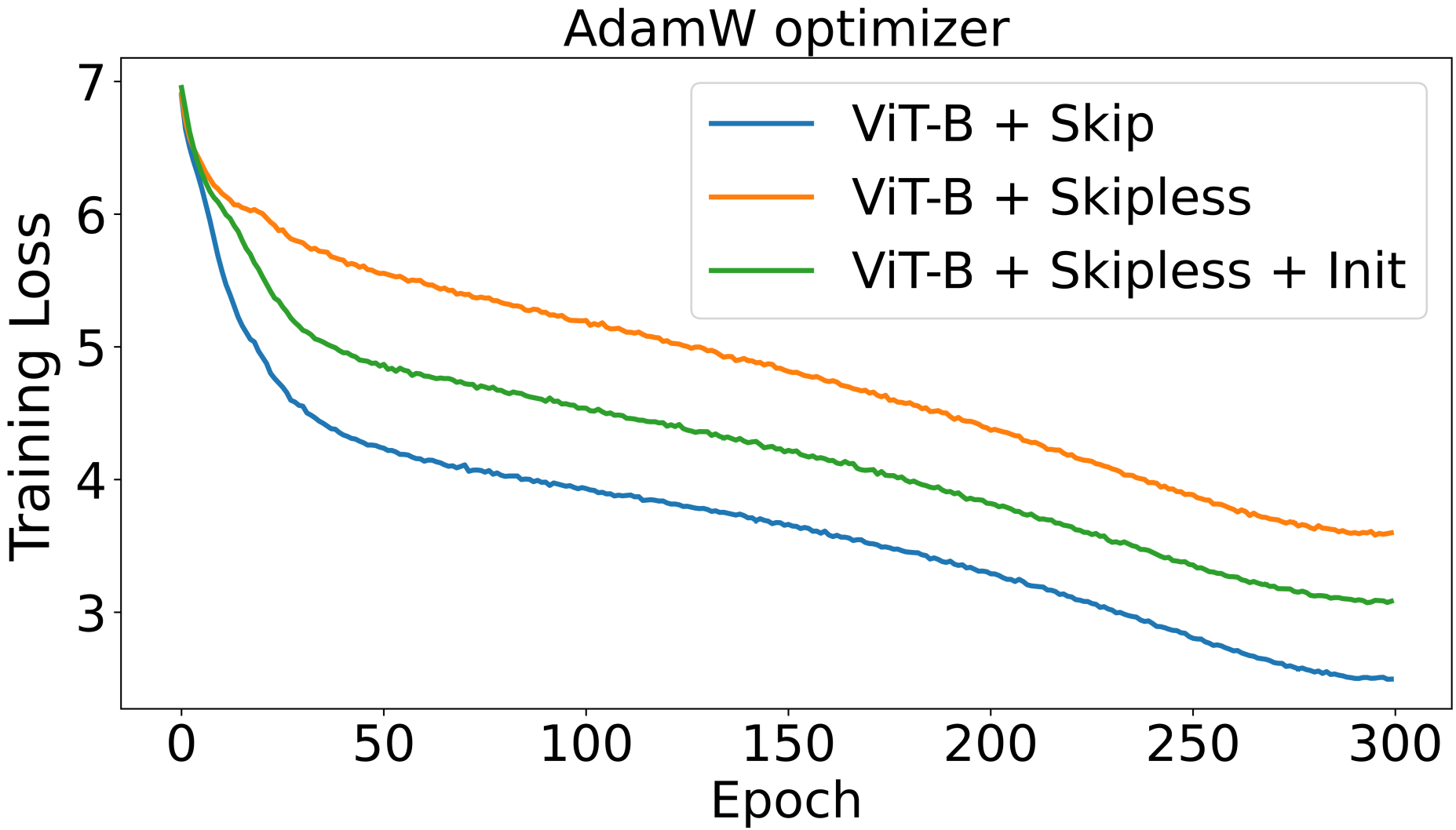


H. Saratchandran, D. Teney, S. Lucey. “Leaner Transformers: More Heads Less. Depth.” In arXiv 2025.

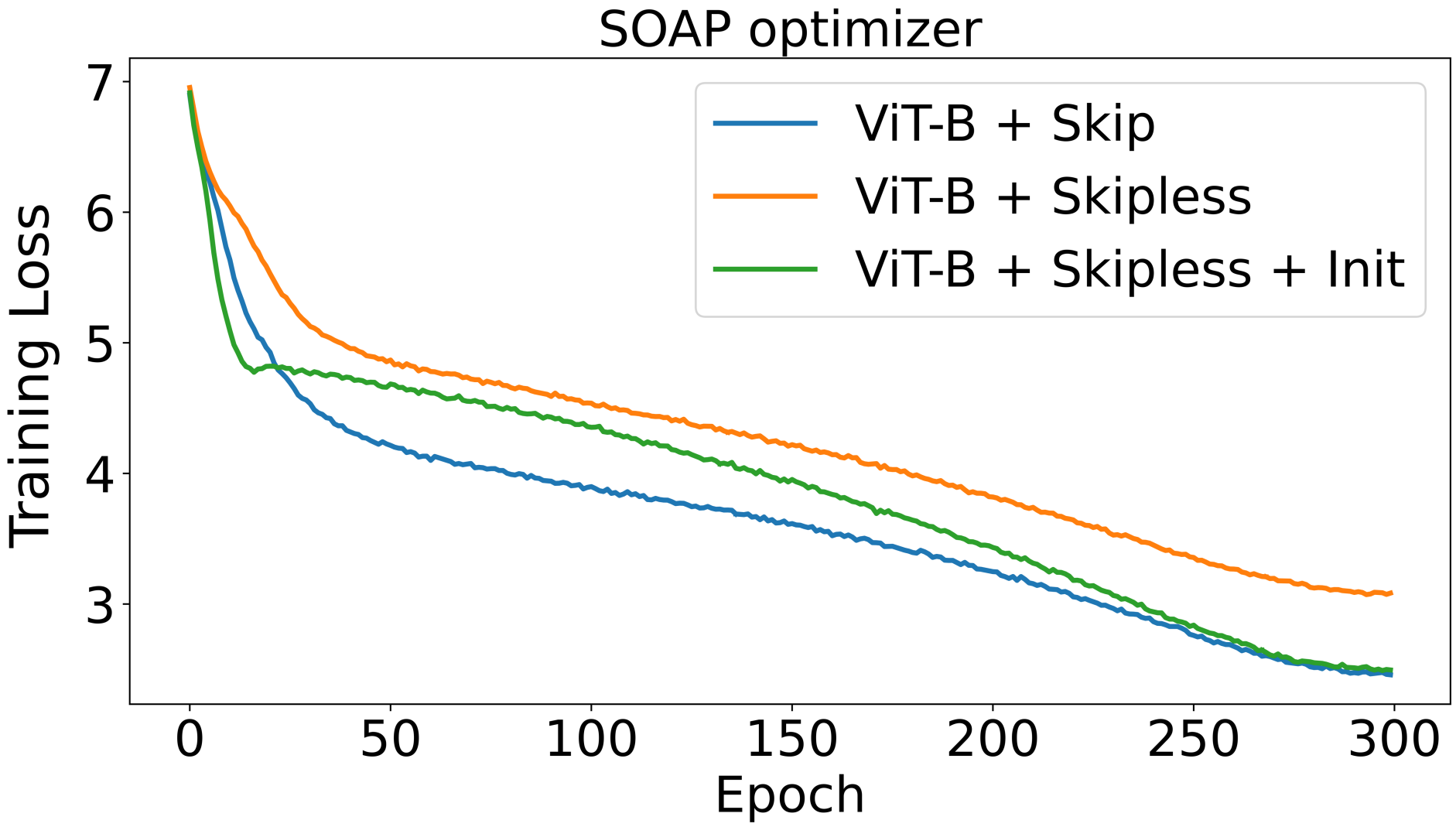
J. Zheng, X. Li, H. Saratchandran, S. Lucey. “Structured Initialization for Vision Transformers.” In NeurIPS 2025.

Y. Ji, S. Lucey, et al. “Cutting the skip: Training residual-free transformers.” In ICLR 2026.

Y. Ji, H. Saratchandran, P. Moghadam, and S. Lucey. “Always Skip Attention.” In ICCV 2025.



Y. Ji, S. Lucey, et al. “Cutting the skip: Training residual-free transformers.” In arXiv 2025.



Y. Ji, S. Lucey, et al. “Cutting the skip: Training residual-free transformers.” In arXiv 2025.

	VOC2012			COCO20k		
Epoch →	300	300	200	300	300	200
Optimizer ↓	skip	skipless	skipless	skip	skipless	skipless
AdamW	32.3	53.5	**54.0**	21.2	36.5	**38.5**
SOAP	49.4	63.2	**68.1**	27.5	46.7	**54.1**

Object Discovery

Y. Ji, S. Lucey, et al. “Cutting the skip: Training residual-free transformers.” In arXiv 2025.

Object Discovery

Y. Ji, S. Lucey, et al. “Cutting the skip: Training residual-free transformers.” In arXiv 2025.

Conclusions….

- Self-attention is fundamentally poorly conditioned, increasing the cost and complexity of training modern transformers.
- Numerous interventions are possible, but in some circumstances it is best to “skip” attention all together.
- Acts as a reminder that the field of AI could be up-ended by better understanding the foundational principles of modern deep learning.