Scaling Knowledge Processing from 2D Chips to 3D Brains

Kwabena Boahen

Bioengineering, Electrical Engineering, & Computer Science Departments
Bio-X, Human-Centered AI, & Wu Tsai Neurosciences Institutes
Stanford University

Cofounder of Femtosense and advisor to Radical Semiconductor

3rd July 2025

Microsoft and OpenAI Plot 100 Billion Dollar AI Supercomputer

Launching as soon as 2028 and expanding through 2030, it will use as much as 5GW

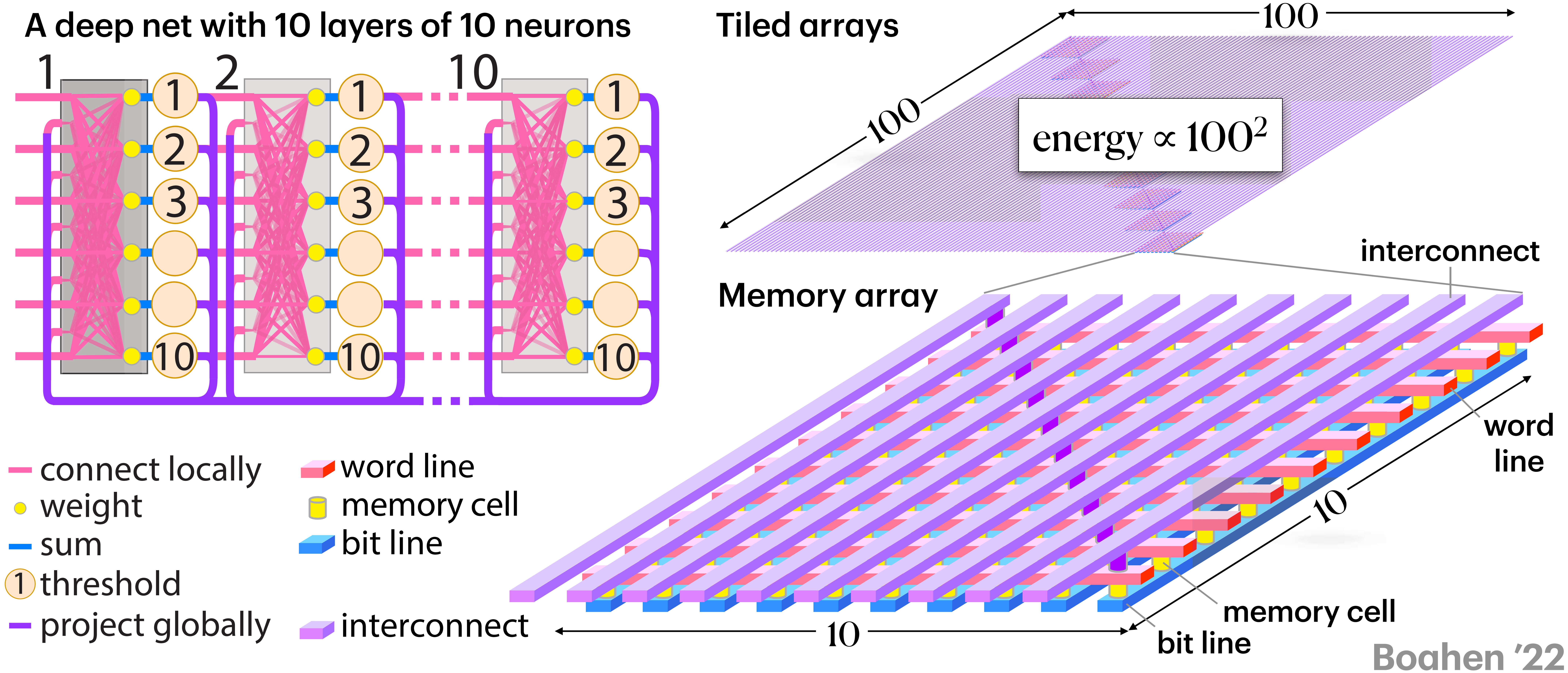
OpenAI CEO Sam Altman, left, and Microsoft CEO Satya Nadella. Photos via Getty. Art: Shane Burke

- Supplied by 5 nuclear power plants
- To run millions of GPUs

The Information '24

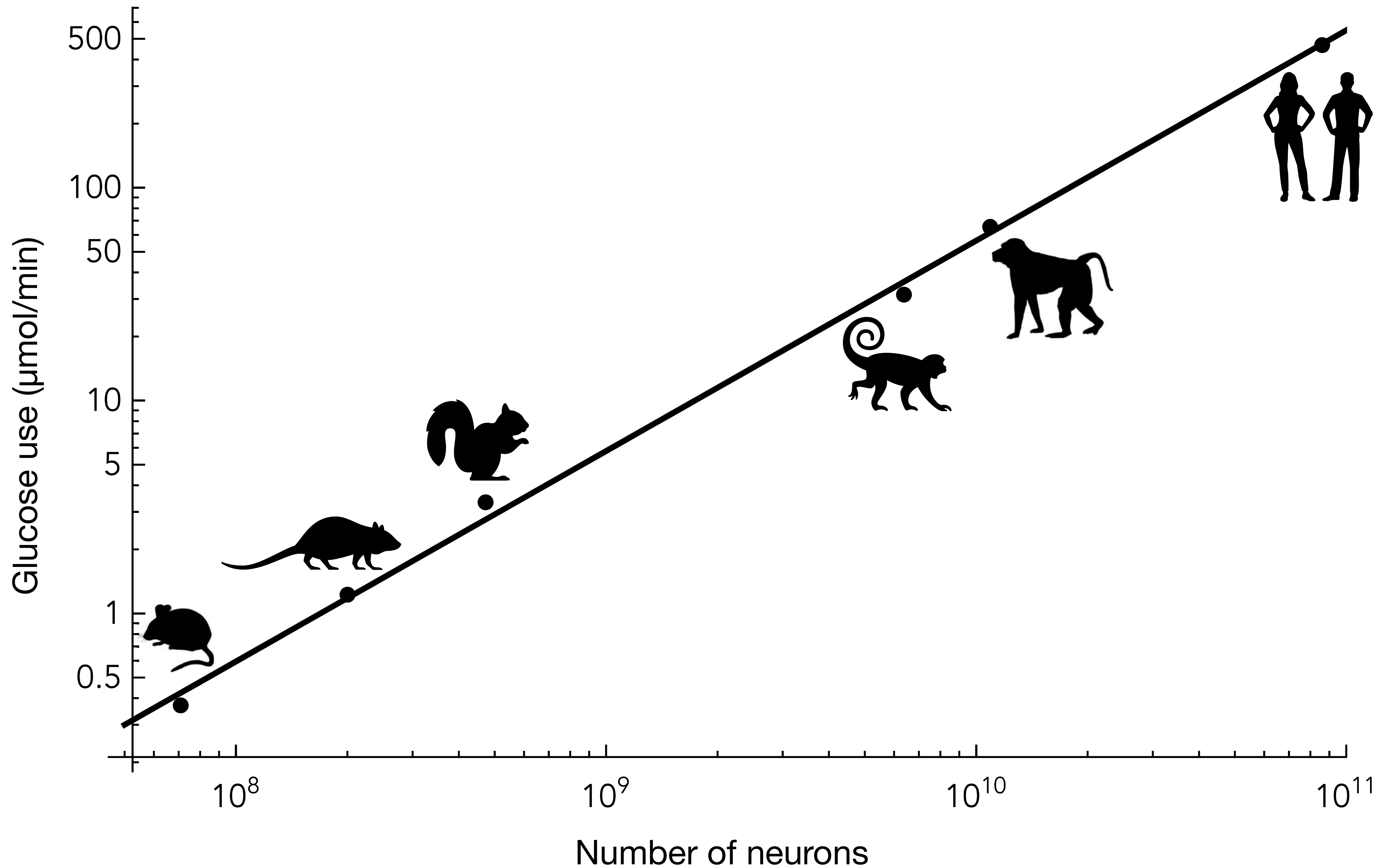
Energy-use scales quadratically with number of units

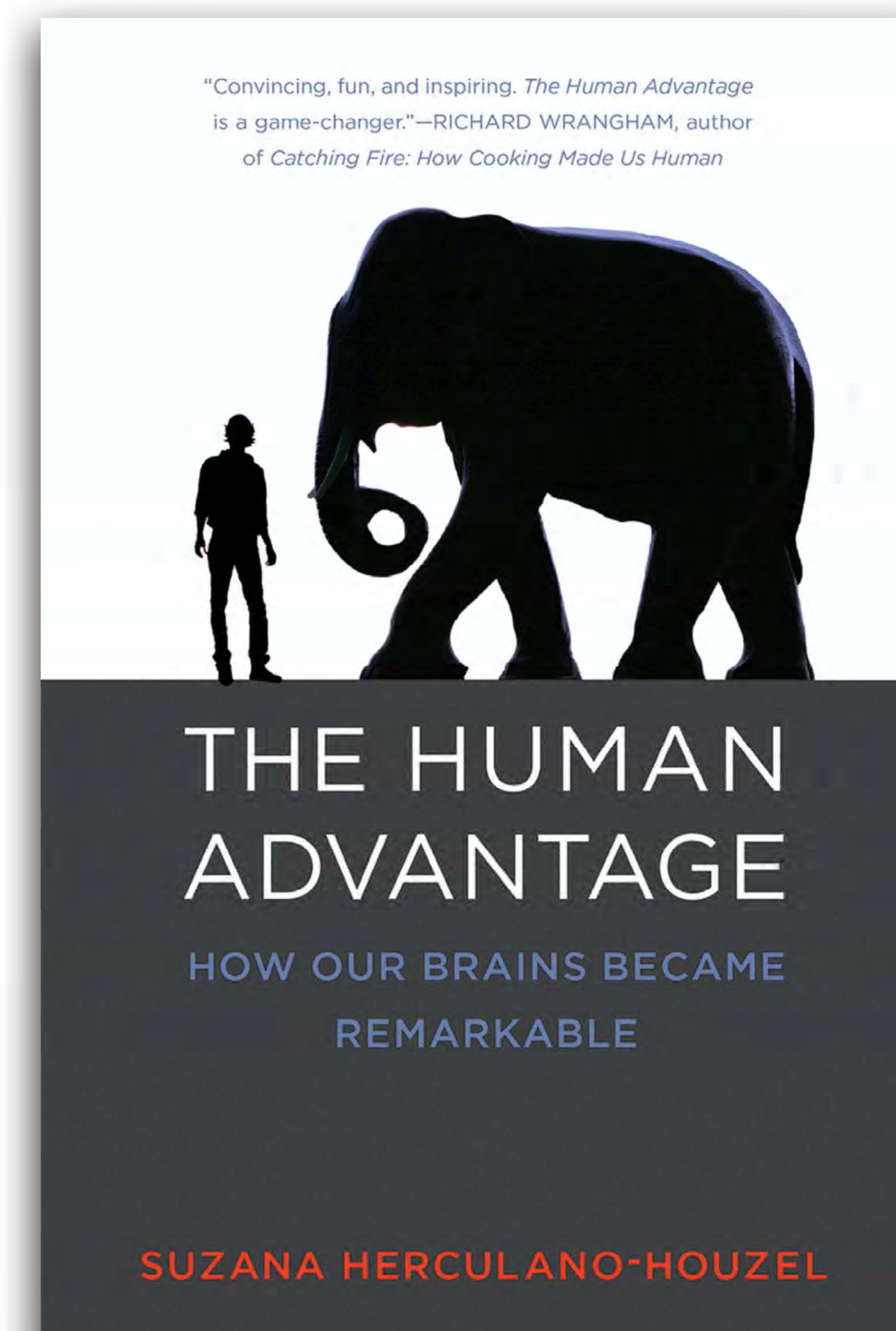
When a fixed fraction of signals travel a distance that scales linearly



A brain's energy-use scales linearly with its neurons

From a 70M-neuron mouse brain to a 86B-neuron human brain

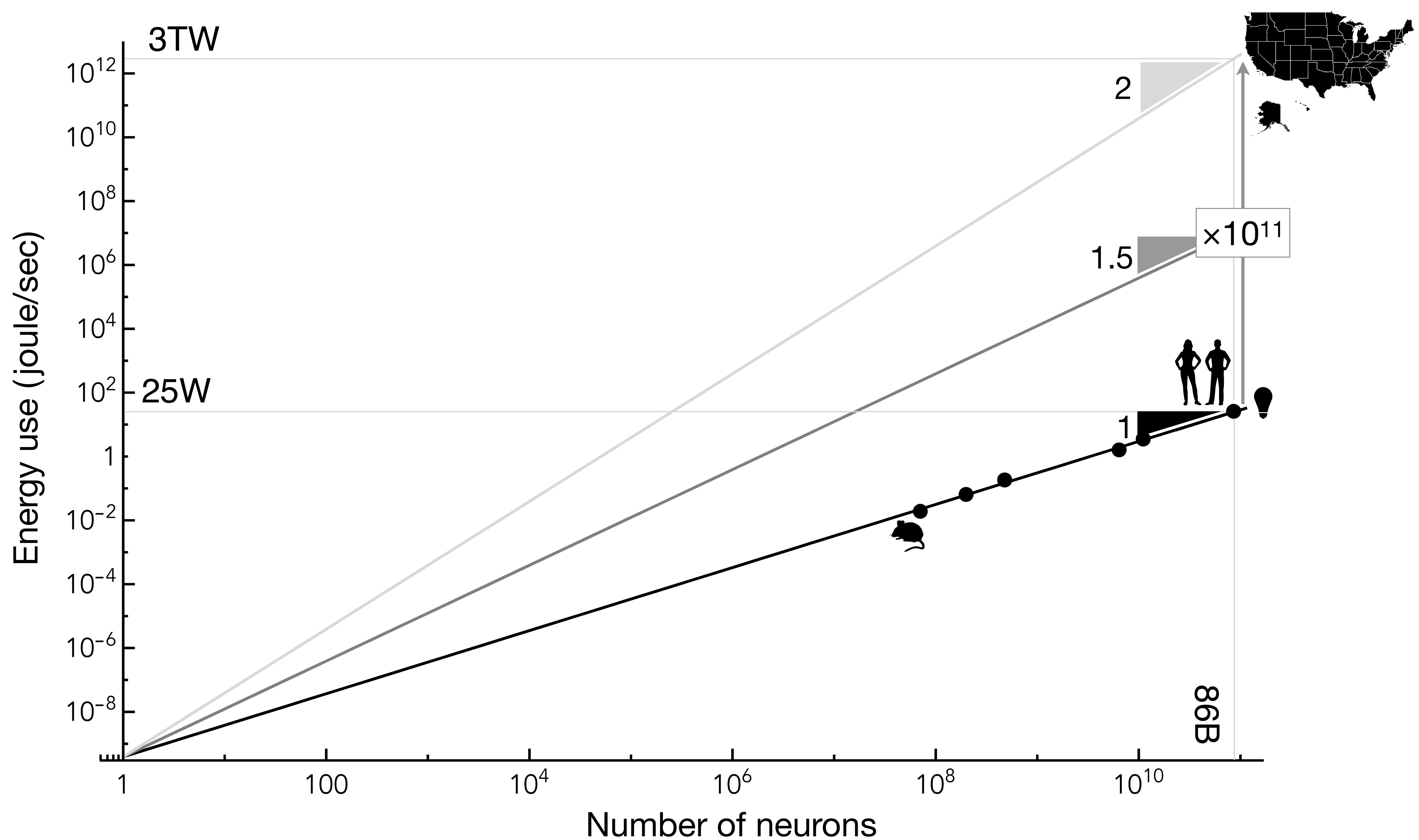




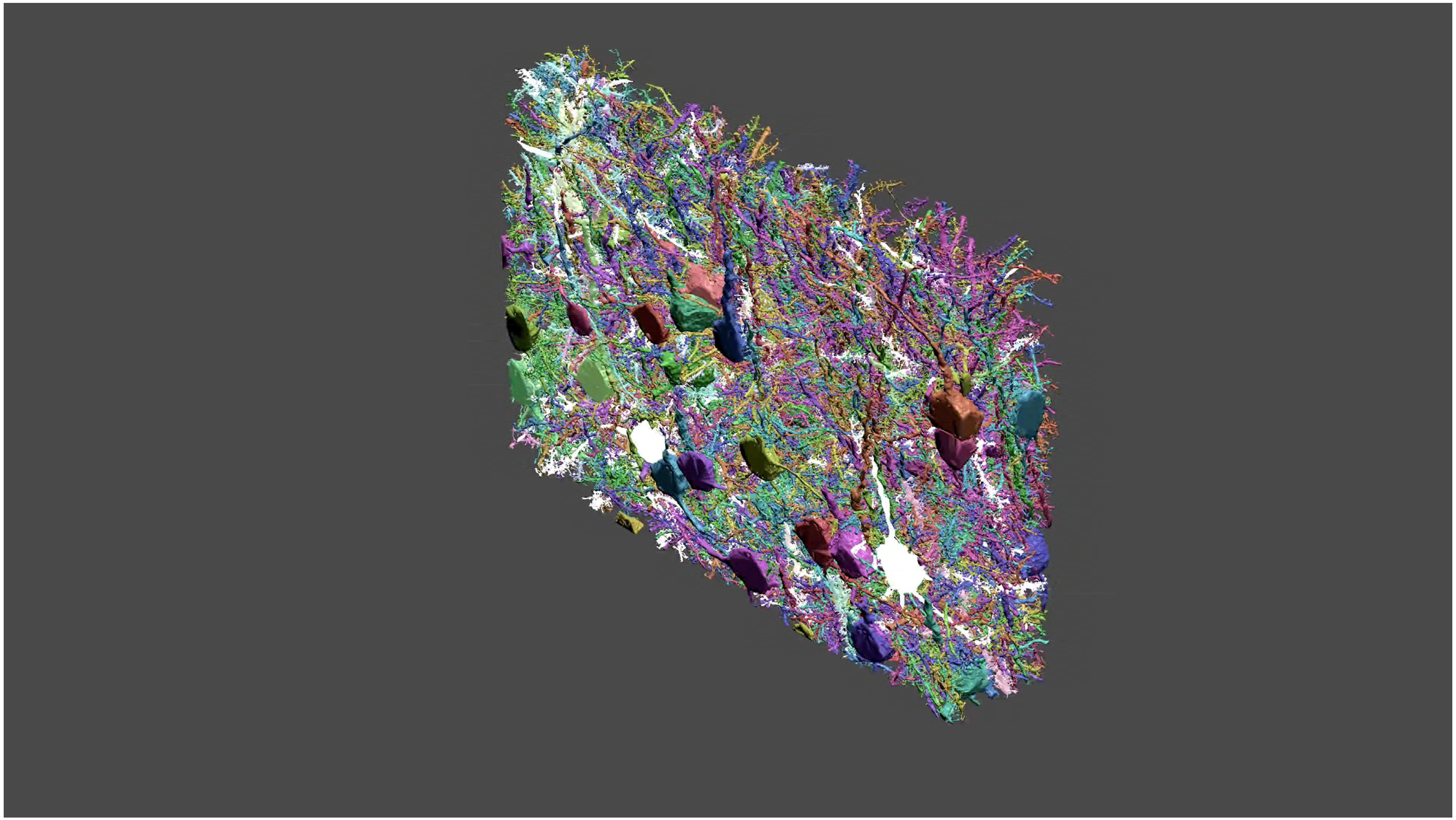
Herculano-Houzel '11,'16

A human brain would as much electricity as the US

if energy-use scaled quadratically



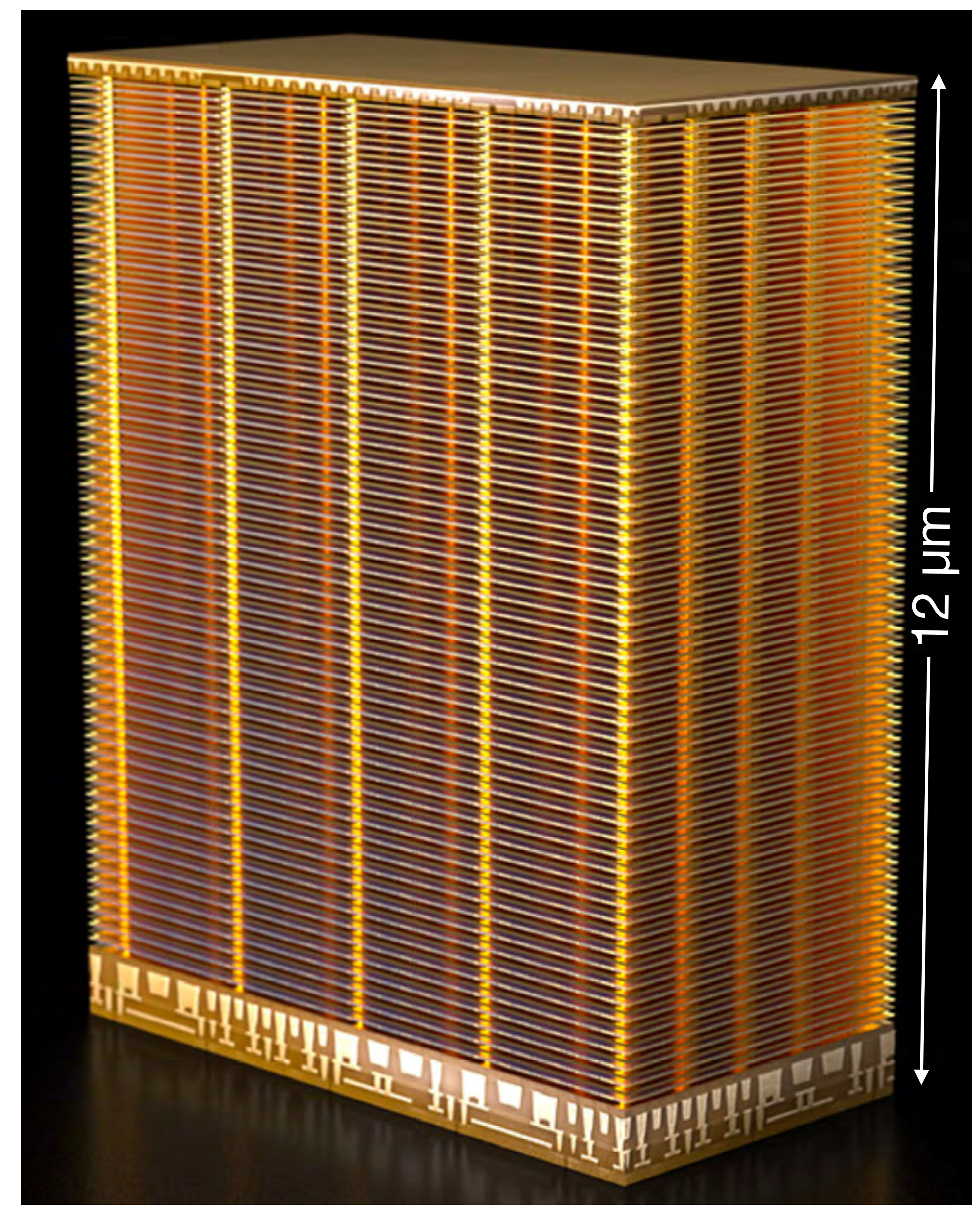
Boahen '22

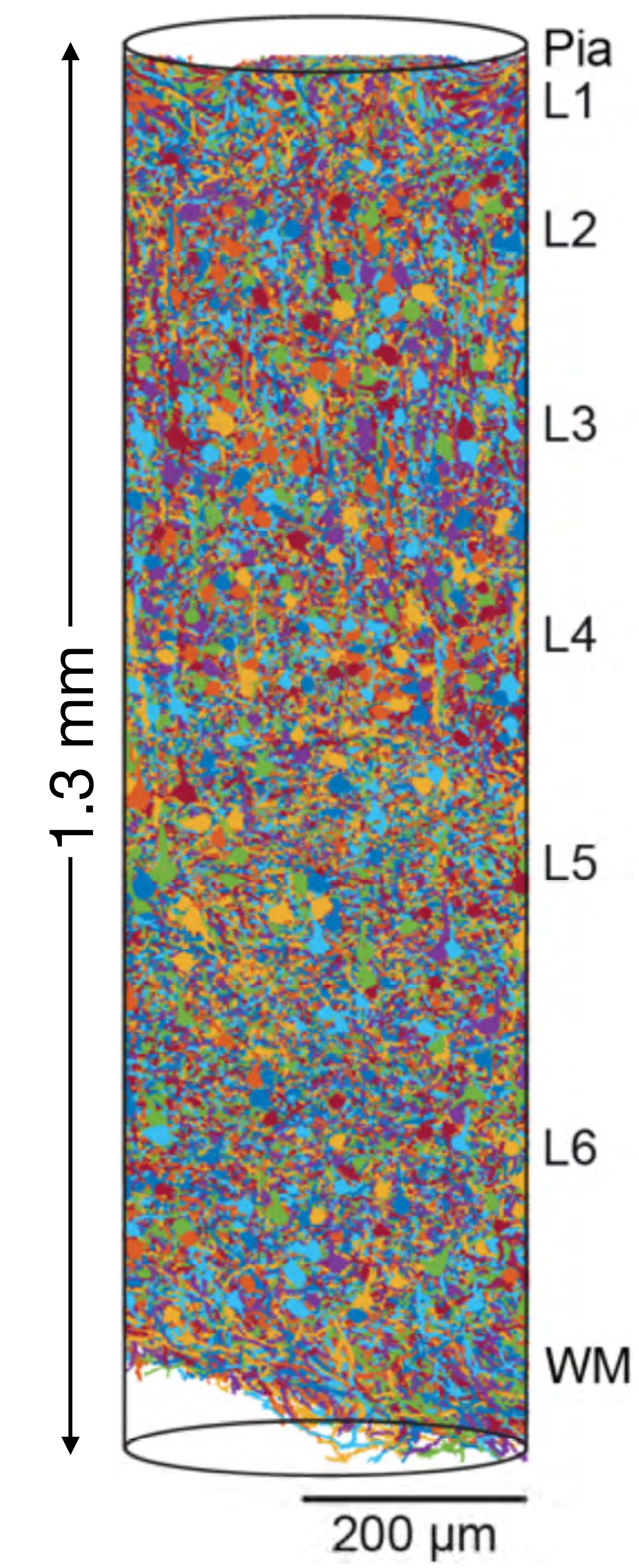


Memory cells are now 3,600 times more dense than cortical synapses

Stacked 10-fold more arrays in a decade: from 24 in 2013 to 232 in 2023

	Transistor	Synapse
Unit volume (µm^3)	1.33×10^{-3}	4.8
Number of layers	232	765
Density (G/mm^2)	4.9	0.27



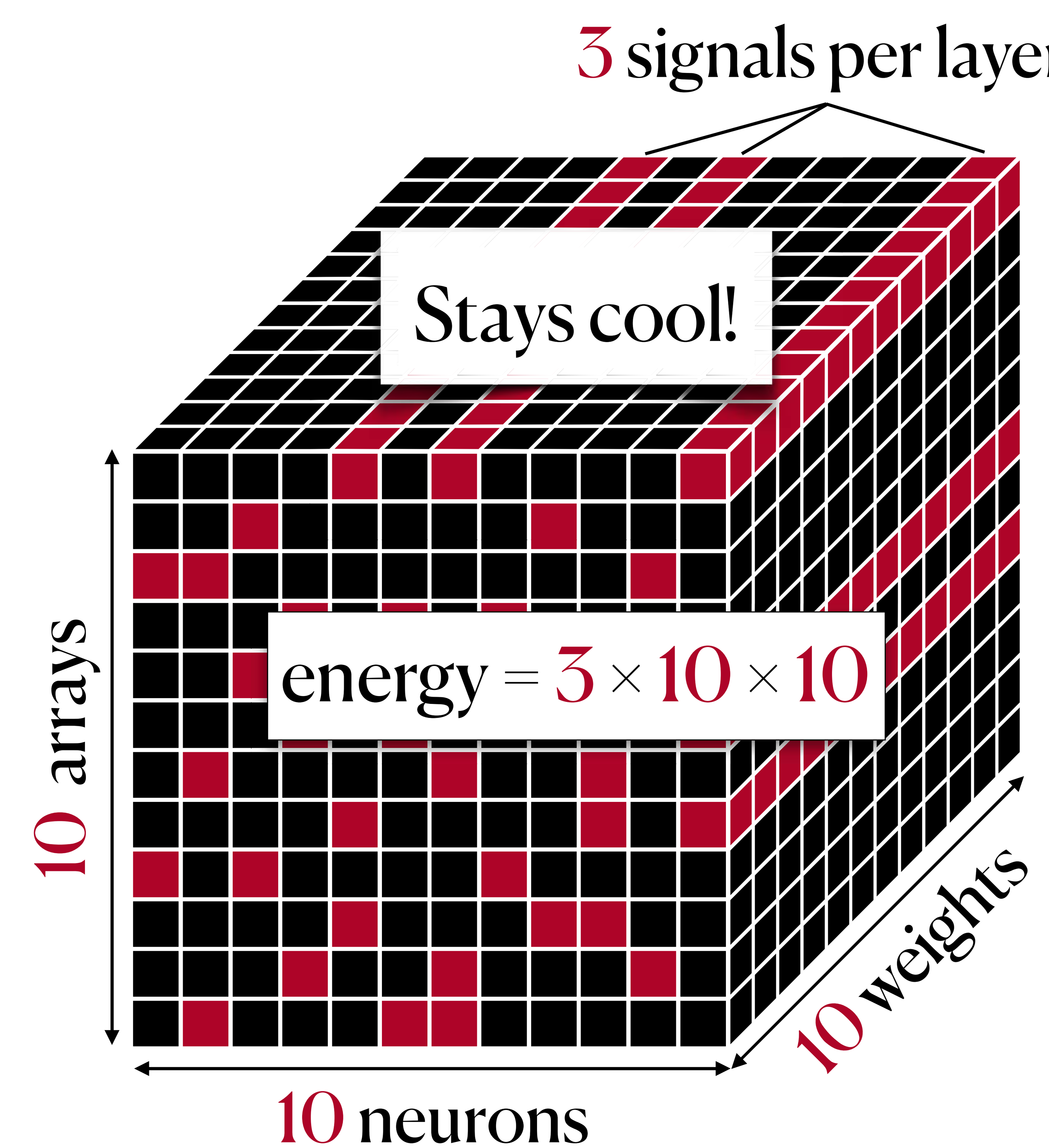


Tech Insights 2019, Meyer et al. 2022, Sievers et al. 2024

Matching brains' linear energy-scaling

With a network of $\sqrt{N}$ layers of $\sqrt{N}$ neurons that emit D signals per inference

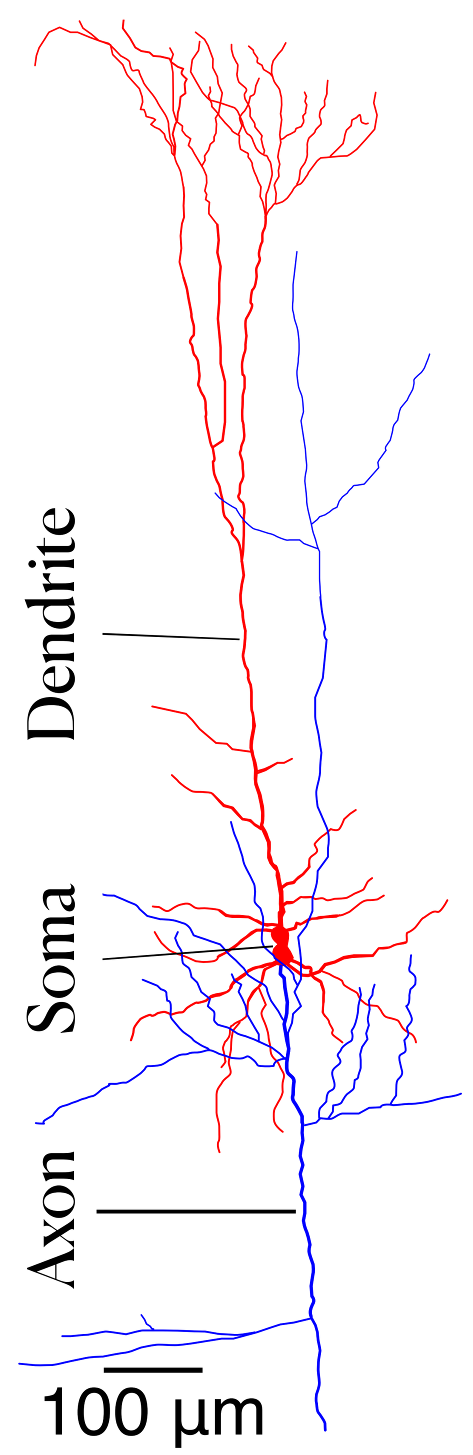
- Stack $\sqrt{N}$ memory arrays:
  - $\sqrt{N}$ arrays $\times$ $\sqrt{N}$ neurons $\times$ $\sqrt{N}$ weights
- Keep a layer's D signals constant:
  - Signal $\sqrt{N}$ times less frequently
  - Respond $\sqrt{N}$ times more selectively



Boahen '22

Conceptions of the learning brain

What coding scales communication linearly?

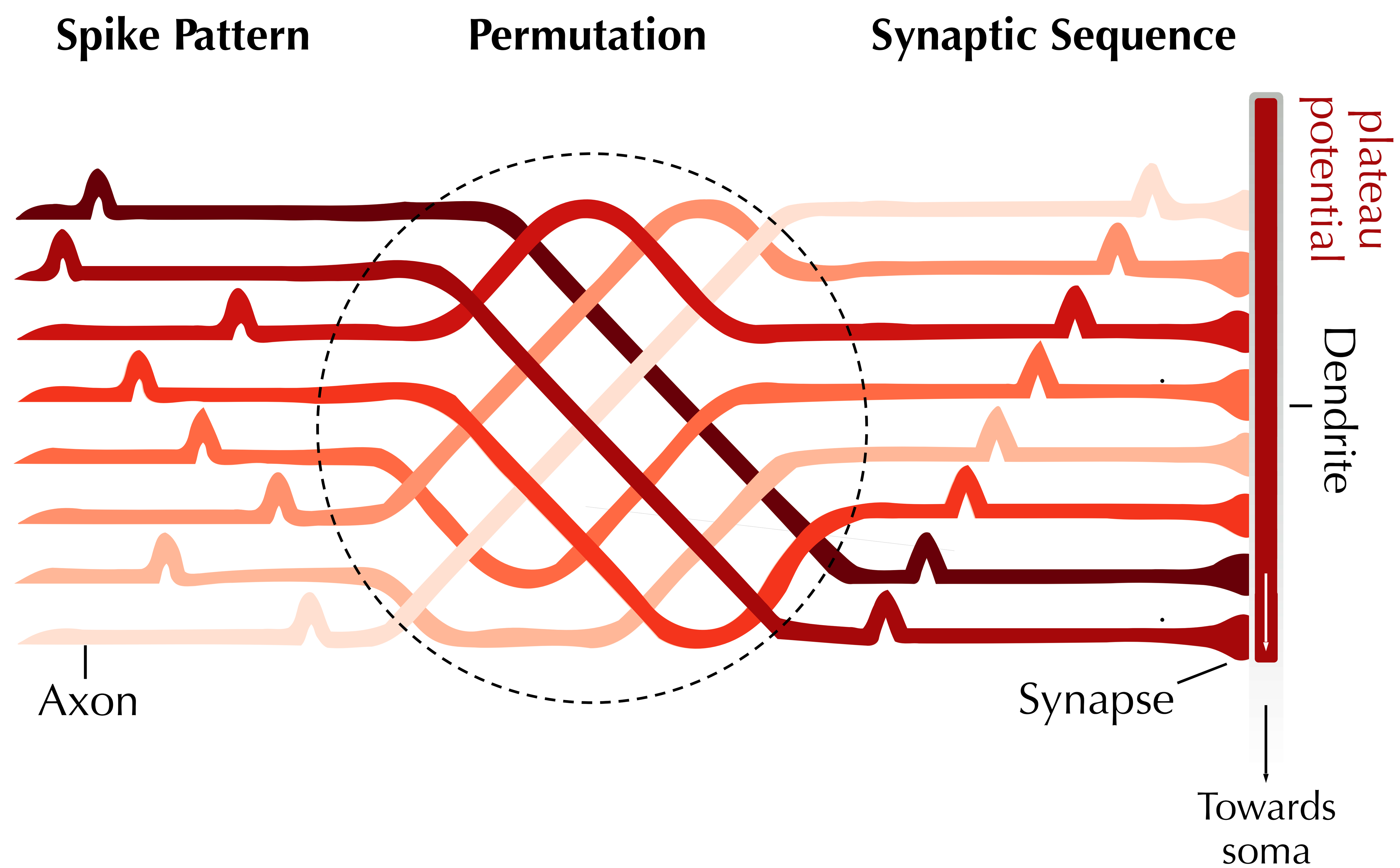


Conception	Communication	Computation	Realization
Synaptocentric	Nonnegative part	Aggregate spatially	GPU or TPU (2D)
Axocentric	Spike rate or timing	Integrate temporally	Neuromorphic chip (2D)
Dendrocentric	Spike sequence	Sort spatiotemporally	Silicon brain (3D)

Boahen ’22

Learning permutes a pattern into a sequence

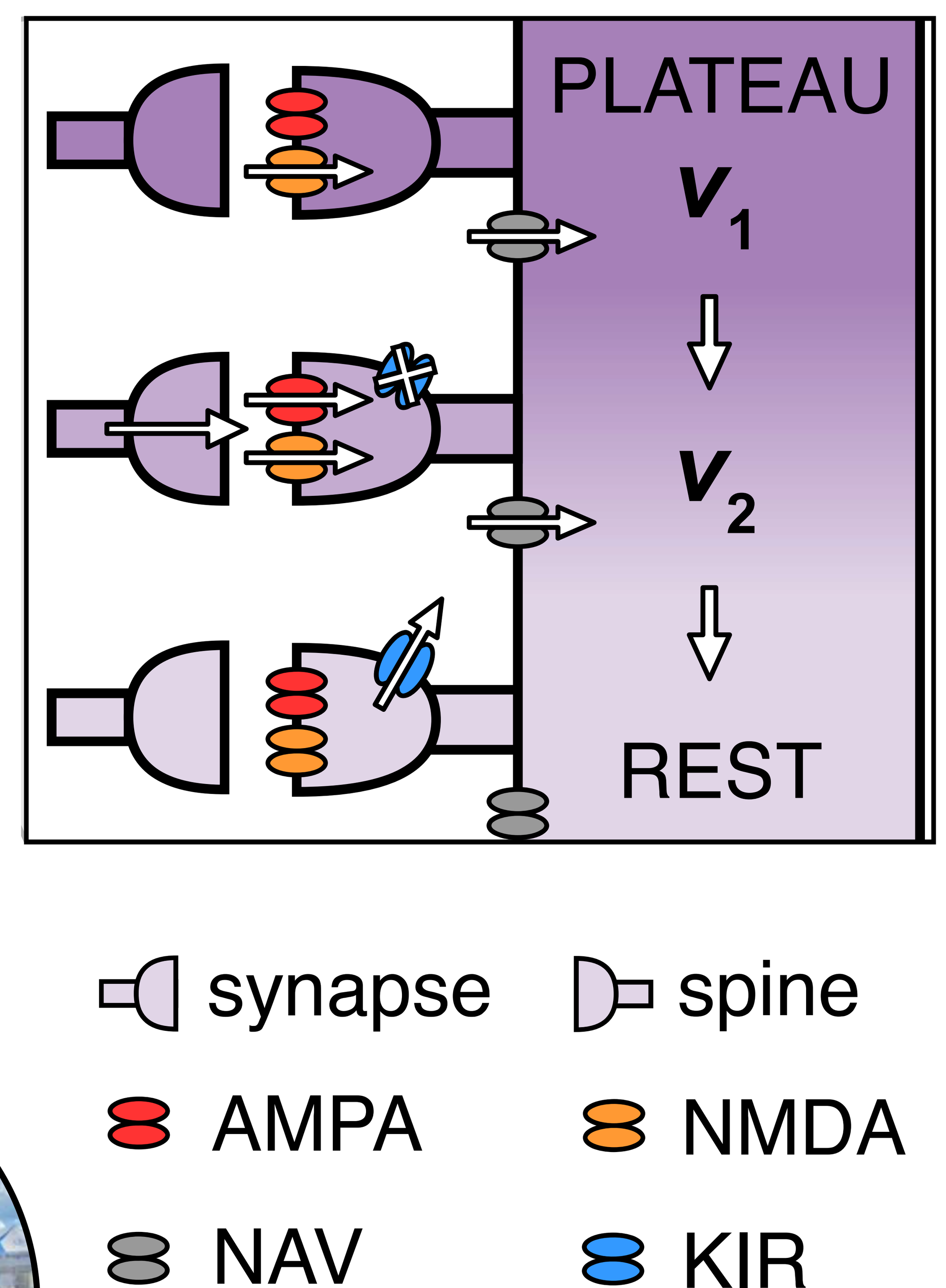
Axons arrange their synapses along a dendrite to deliver ordered input

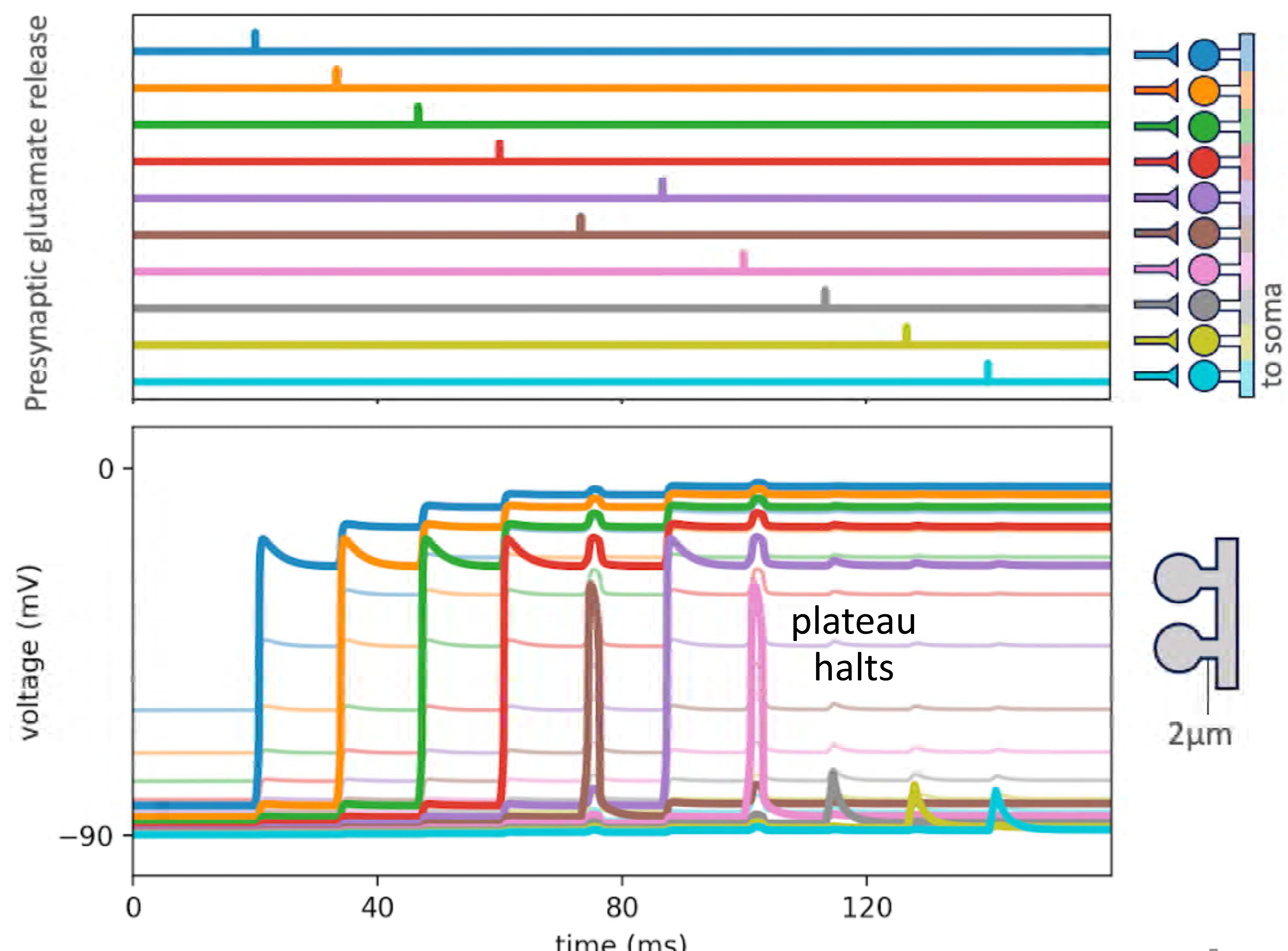


Boahen '22, Le Coeur et al. '23

Sequence selective responses in a stretch of dendrite

Ligand and voltage-gated ion-channels discriminate an out-of-sequence spike



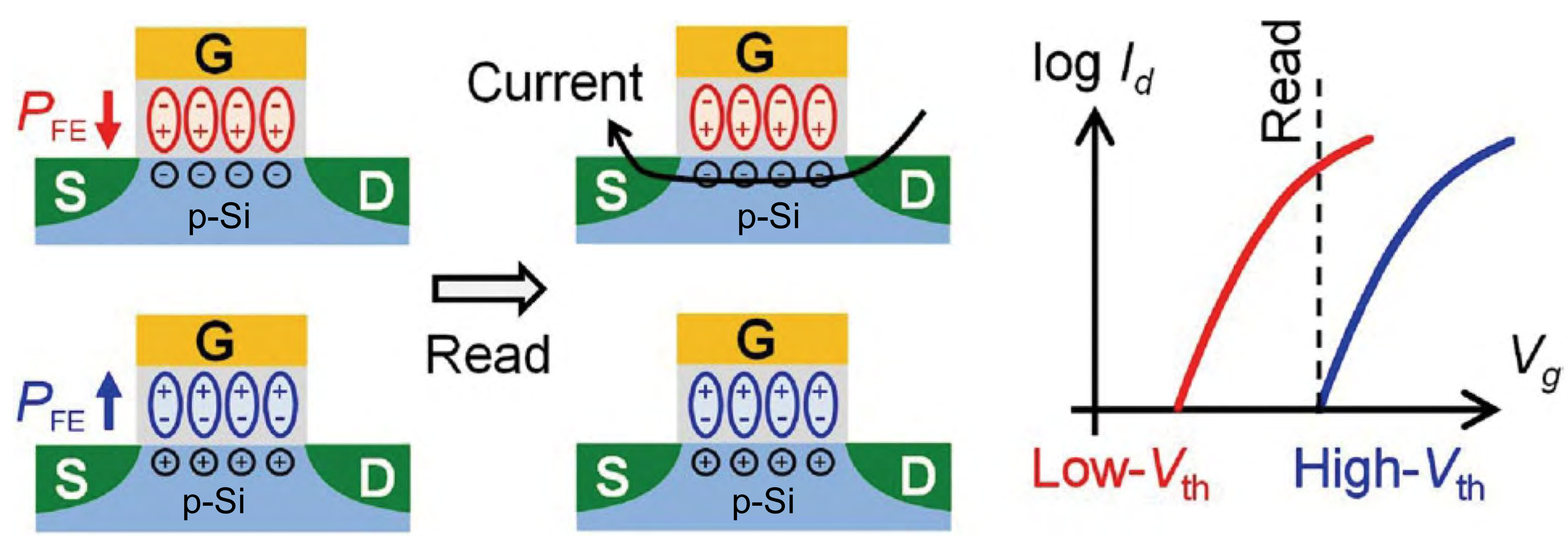


Boahen '22

Basic concepts of Si-based n-FeFET

Polarity of charge dipoles in ferroelectric insulator control conductivity of semiconductor

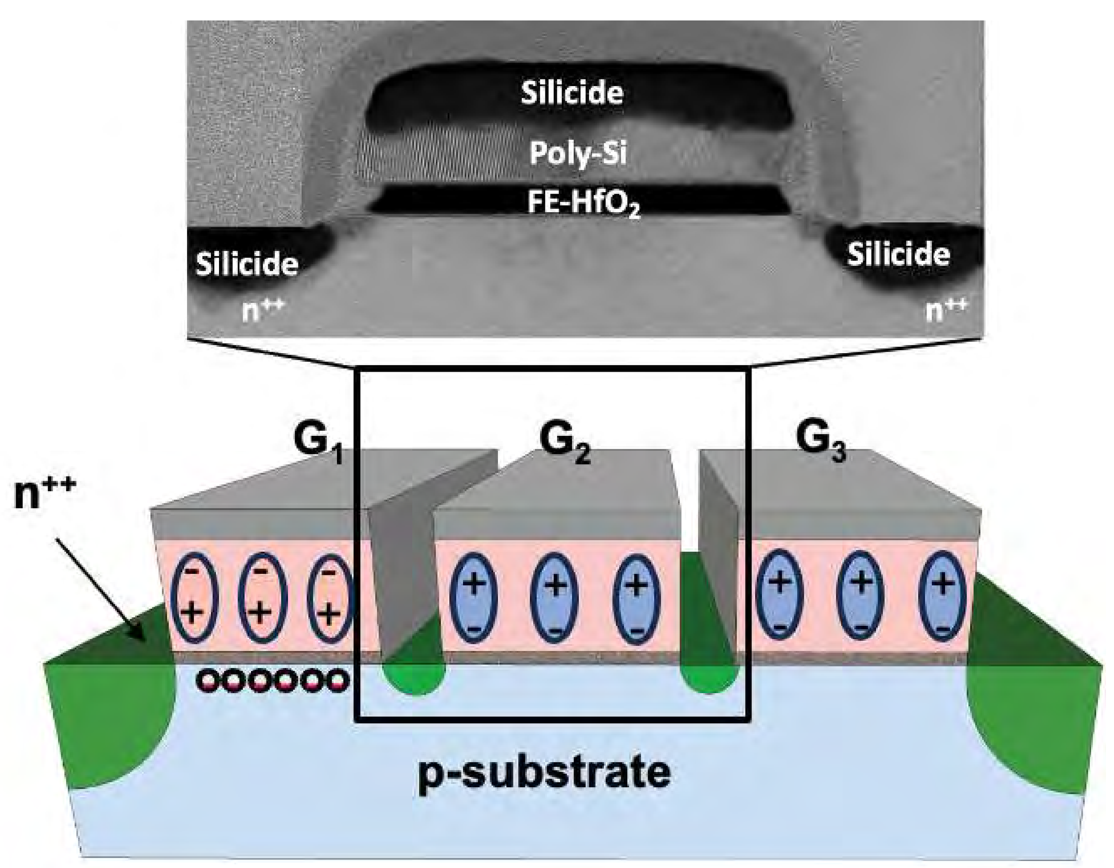
- Conductive state (low V_T): Polarization vector points toward channel
- Nonconductive state (high V_T): Polarization vector points away from channel

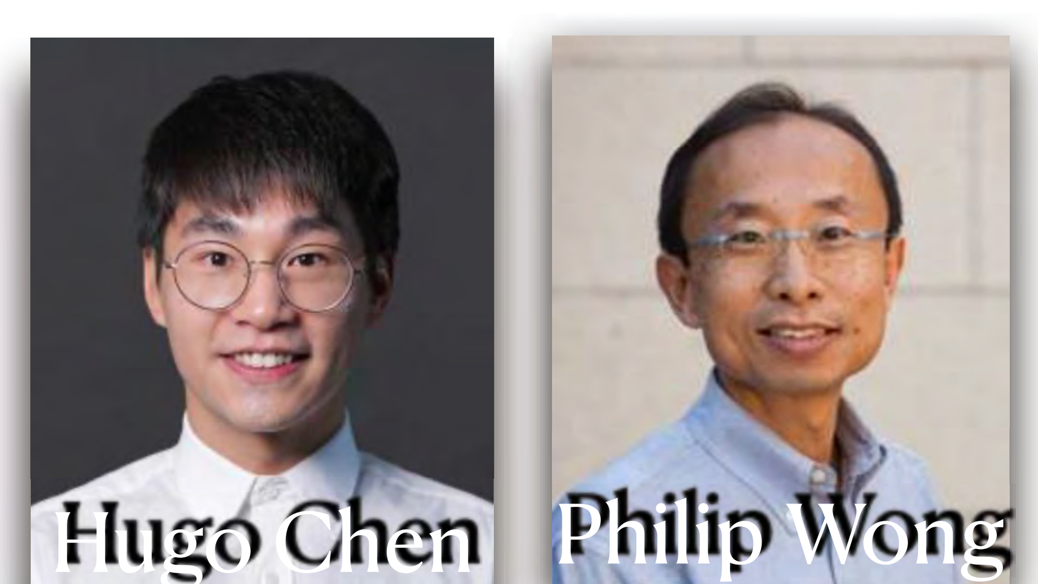


Toprasertpong, K., Takenaka, M. & Takagi, S. On the strong coupling of polarization and charge trapping in HfO2/Si-based ferroelectric field-effect transistors: overview of device operation and reliability. Appl. Phys. A 128, 1114 (2022). https://doi.org/10.1007/s00339-022-06212-6

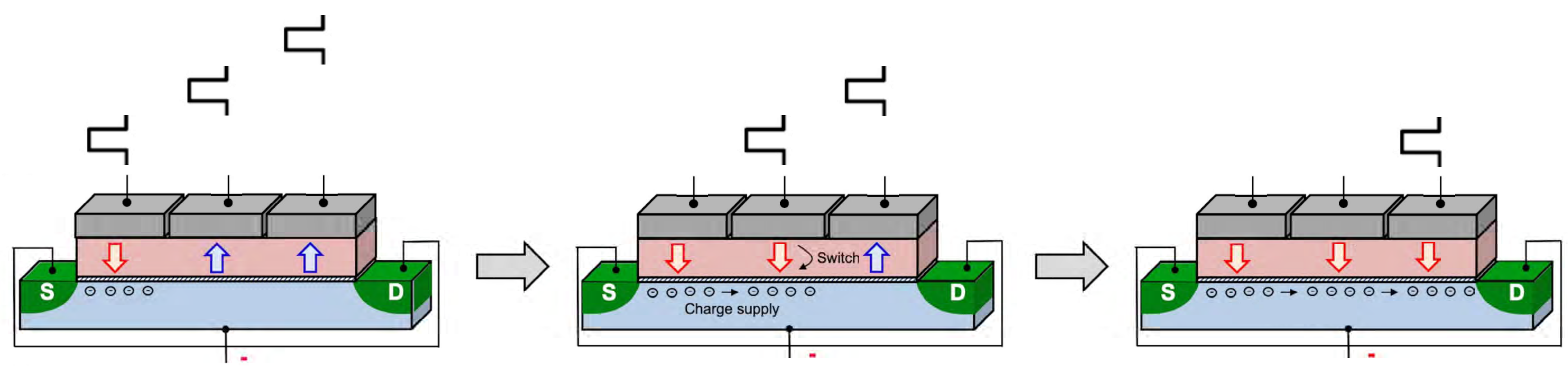
A string of FeFETs emulates a stretch of dendrite

Consecutively applied voltage pulses flip all charge dipoles in ferroelectric layer

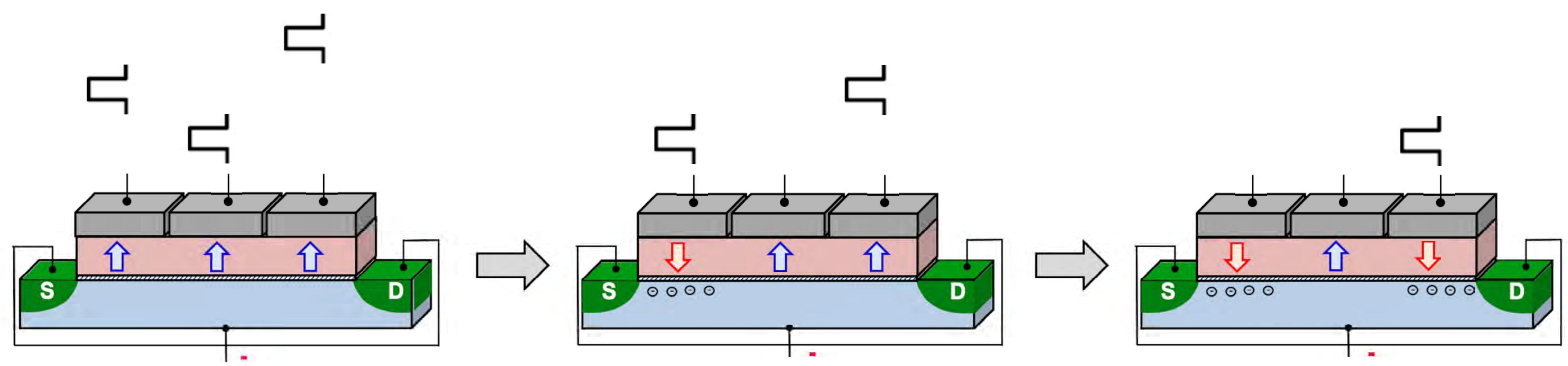




Correct sequence: Gate1 → Gate2 → Gate3



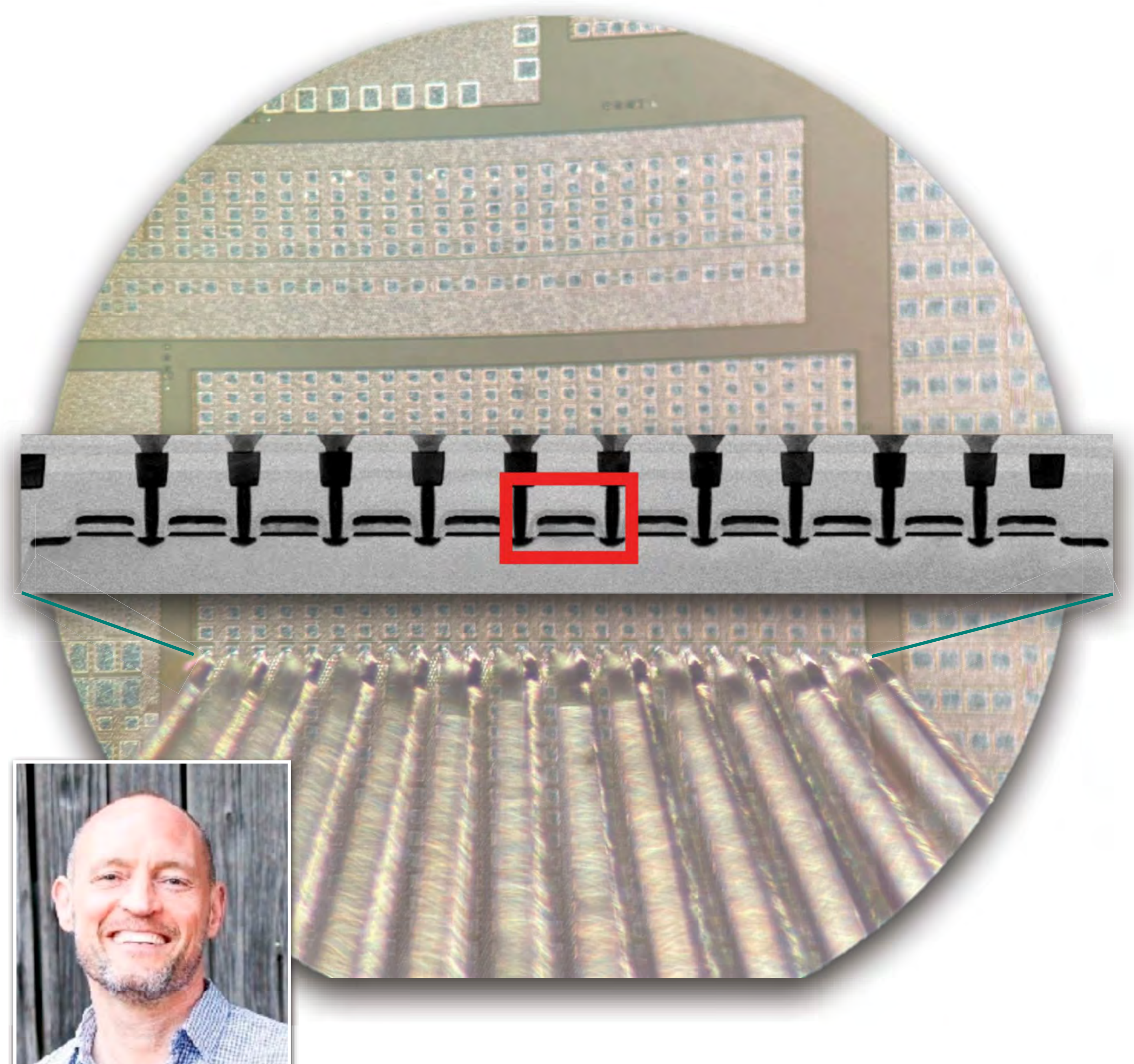
Incorrect sequence: Gate2 → Gate1→ Gate3



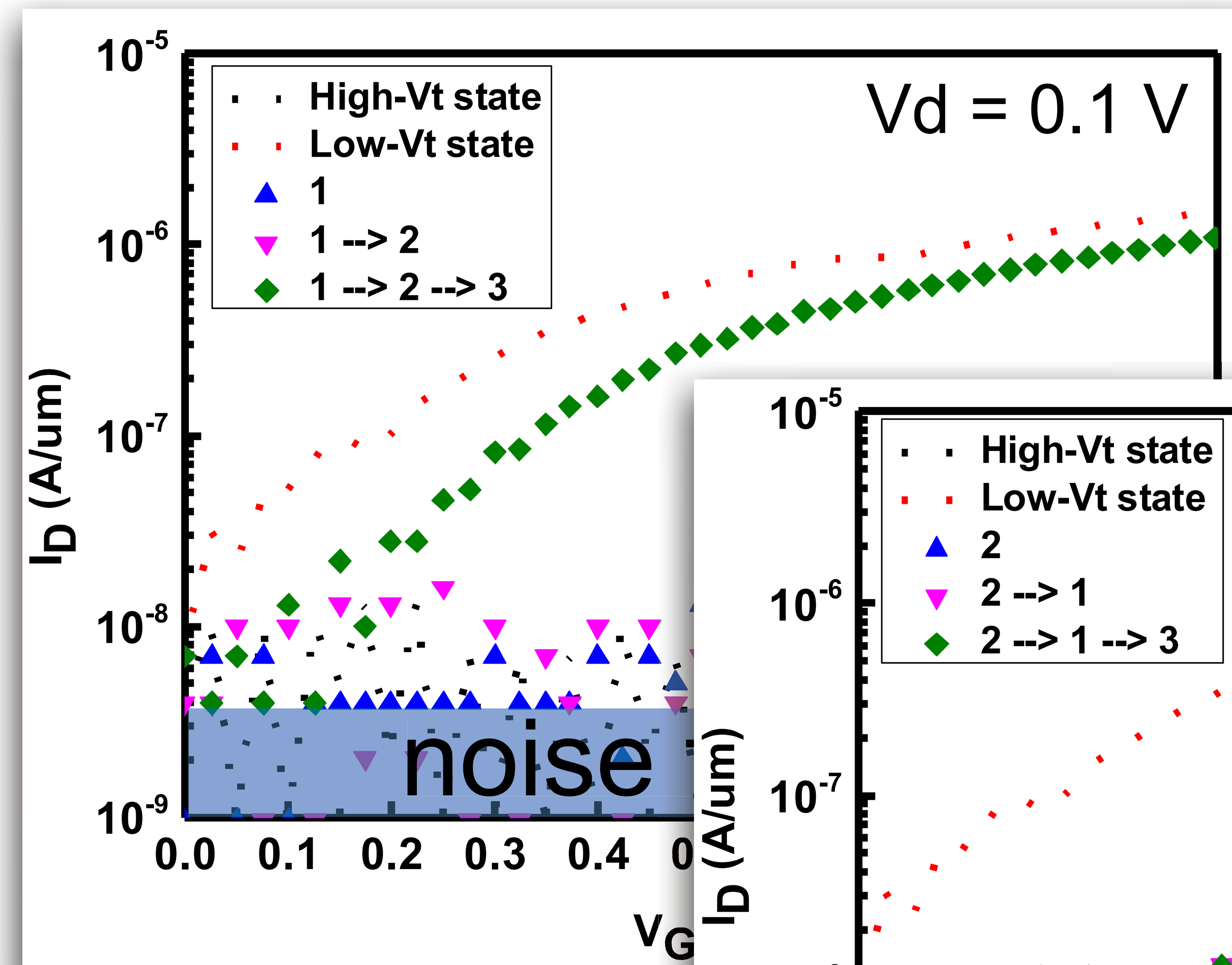
Chen, Beauchamp, et al. ‘24

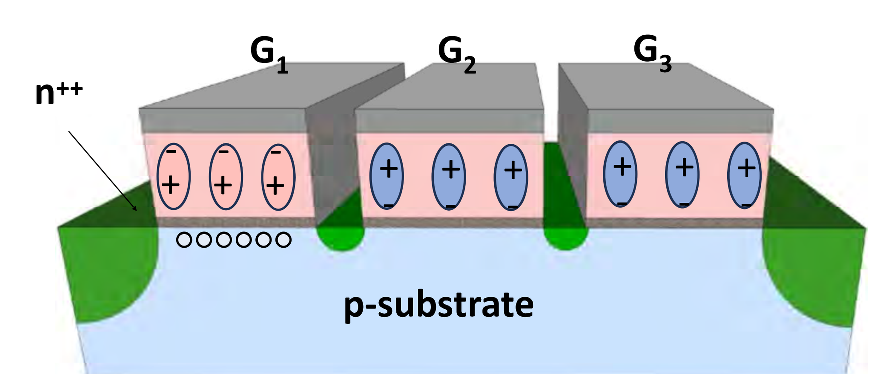
FeFET string discriminates pulse patterns

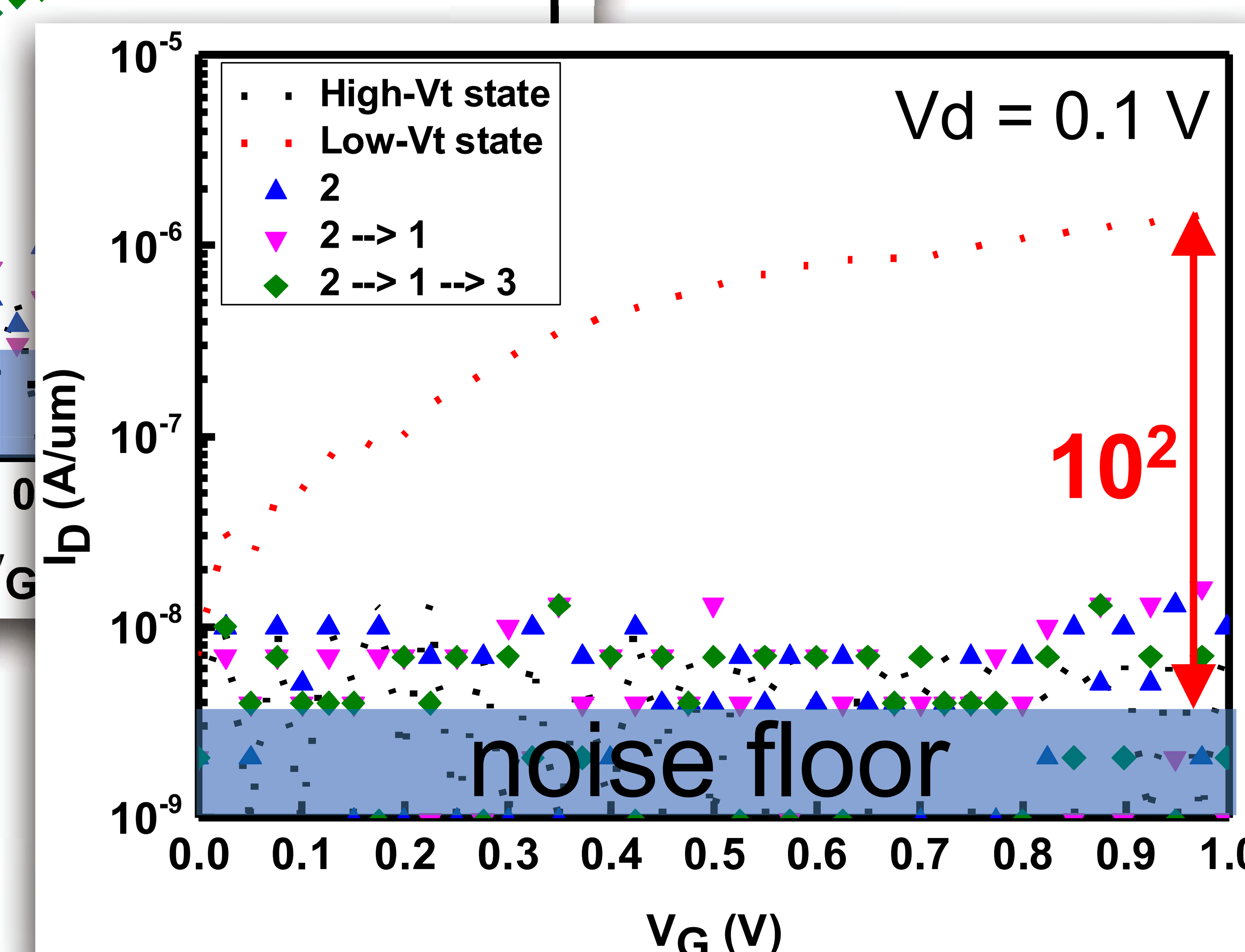
GlobalFoundries fabricated our design in a production-ready 28-nm process



Sven Beyer, DMTS



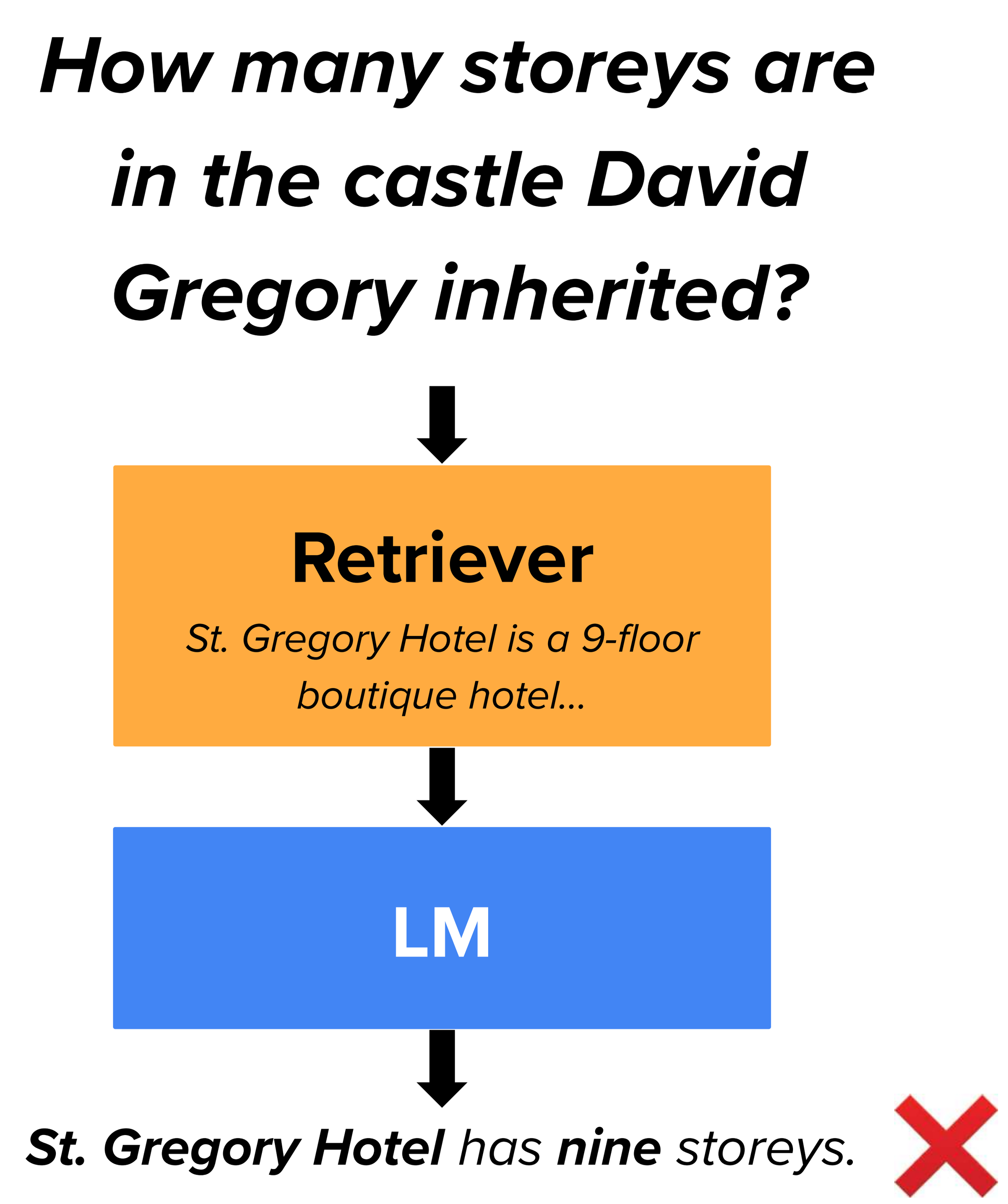


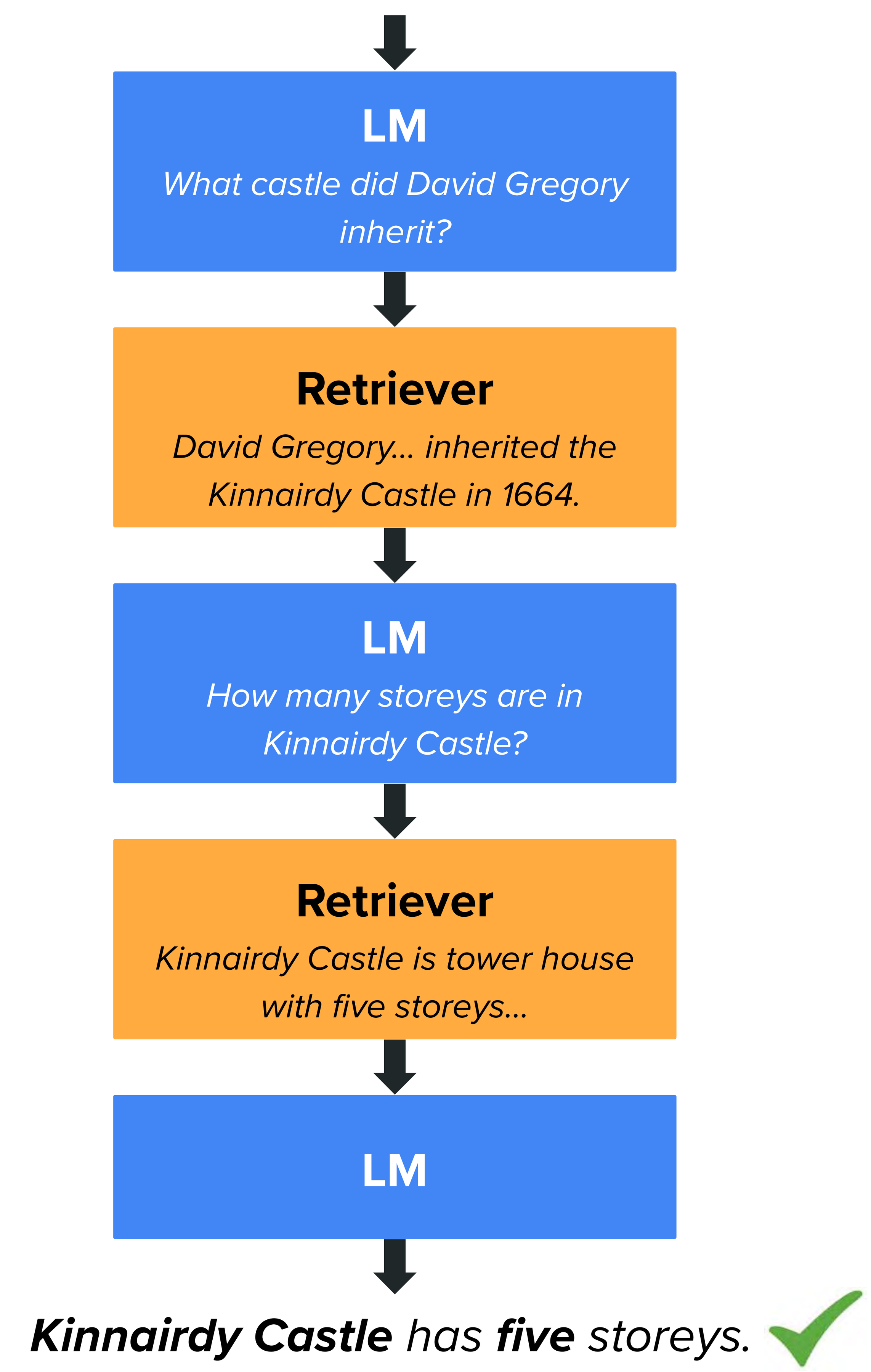


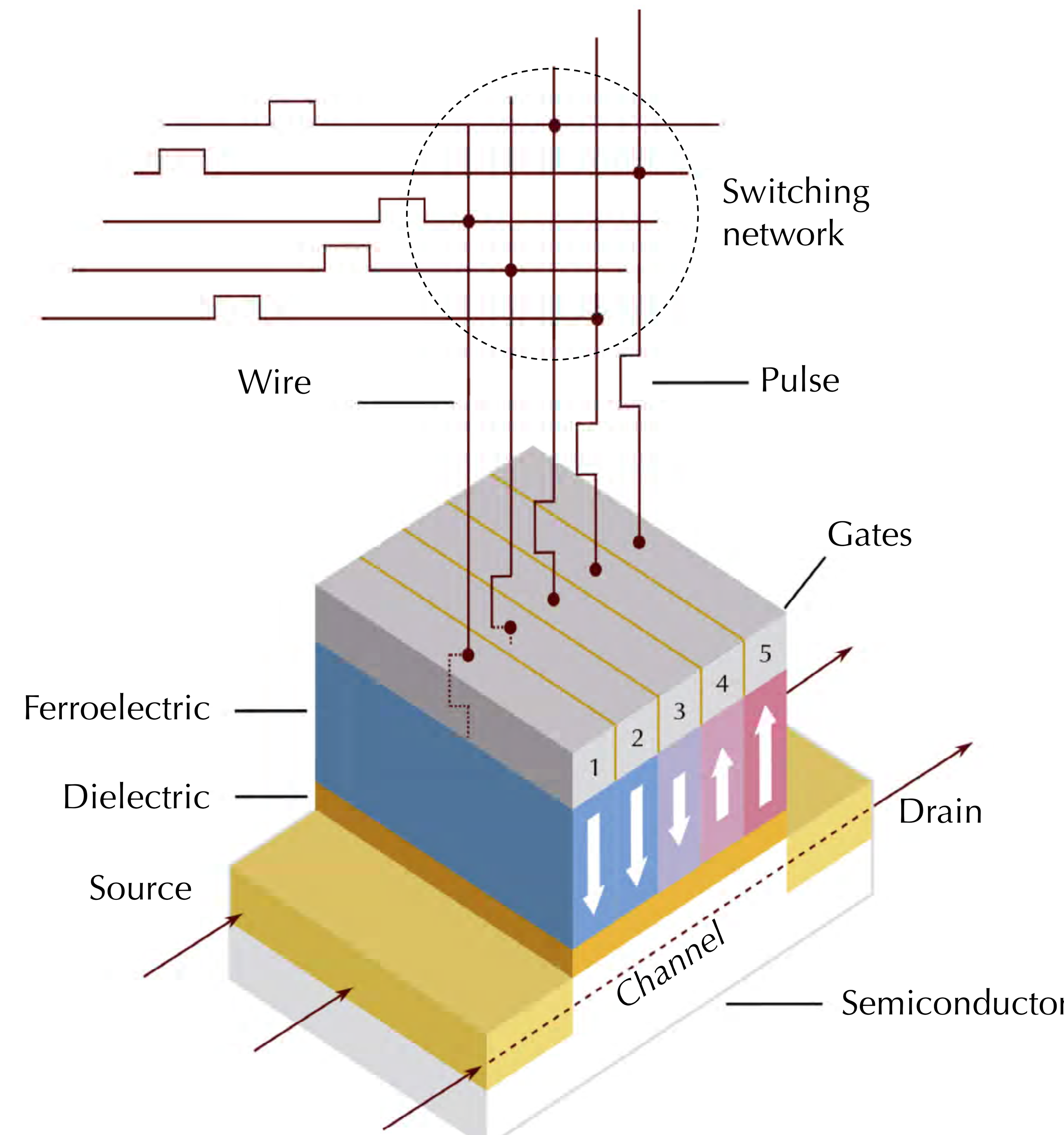
Chen, Beauchamp, et al. ‘24

Retrieval Augmented Generation (RAG)

Lightweight language models (LMs) rely heavily on retrieval of facts from memory







Khattab et al. '24

Moving away from synaptocentric to dendrocentric learning would enable AI to run not with megawatts in the cloud but rather with watts on a phone.

Acknowledgements

EFRI/BRAID

Hoffman-Yee

Wu-Tsai Neuro

Perspective

Dendrocentric learning for synthetic intelligence

https://doi.org/10.1038/s41586-022-05340-6

Received: 22 December 2020

Accepted: 12 September 2022

Published online: 30 November 2022

Check for updates

Kwabena Boahen[1,2,3,4,5,6] ✉

Artificial intelligence now advances by performing twice as many floating-point multiplications every two months, but the semiconductor industry tiles twice as many multipliers on a chip every two years. Moreover, the returns from tiling these multipliers ever more densely now diminish because signals must travel relatively farther and farther. Although travel can be shortened by stacking tiled multipliers in a three-dimensional chip, such a solution acutely reduces the available surface area for dissipating heat. Here I propose to transcend this three-dimensional thermal constraint by moving away from learning with synapses to learning with dendrites. Synaptic inputs are not weighted precisely but rather ordered meticulously along a short stretch of dendrite, termed dendrocentric learning. With the help of a computational model of a dendrite and a conceptual model of a ferroelectric device that emulates it, I illustrate how dendrocentric learning artificial intelligence—or synthetic intelligence for short—could run not with megawatts in the cloud but rather with watts on a smartphone.

Boahen '22