

23 March 2016 — Berkeley CA

---

# Neuromorphic Chips: Combining Analog Computation with Digital Communication

---

Kwabena Boahen

Bioengineering and Electrical Engineering (by courtesy)  
Stanford University



*Eliasmith, 2013*



# Acknowledgments



## Students

Ben Benjamin  
Alex Neckar  
Sam Fok  
Samir Menon  
Tatiana Engel  
Nick Oza  
John Aguayo  
Eric Kauderer  
Ashok Cutkosky  
**Alumni (Recent)**  
Peiran Gao  
Nick Steinmetz  
John Arthur  
Paul Merolla  
Rodrigo Alvarez

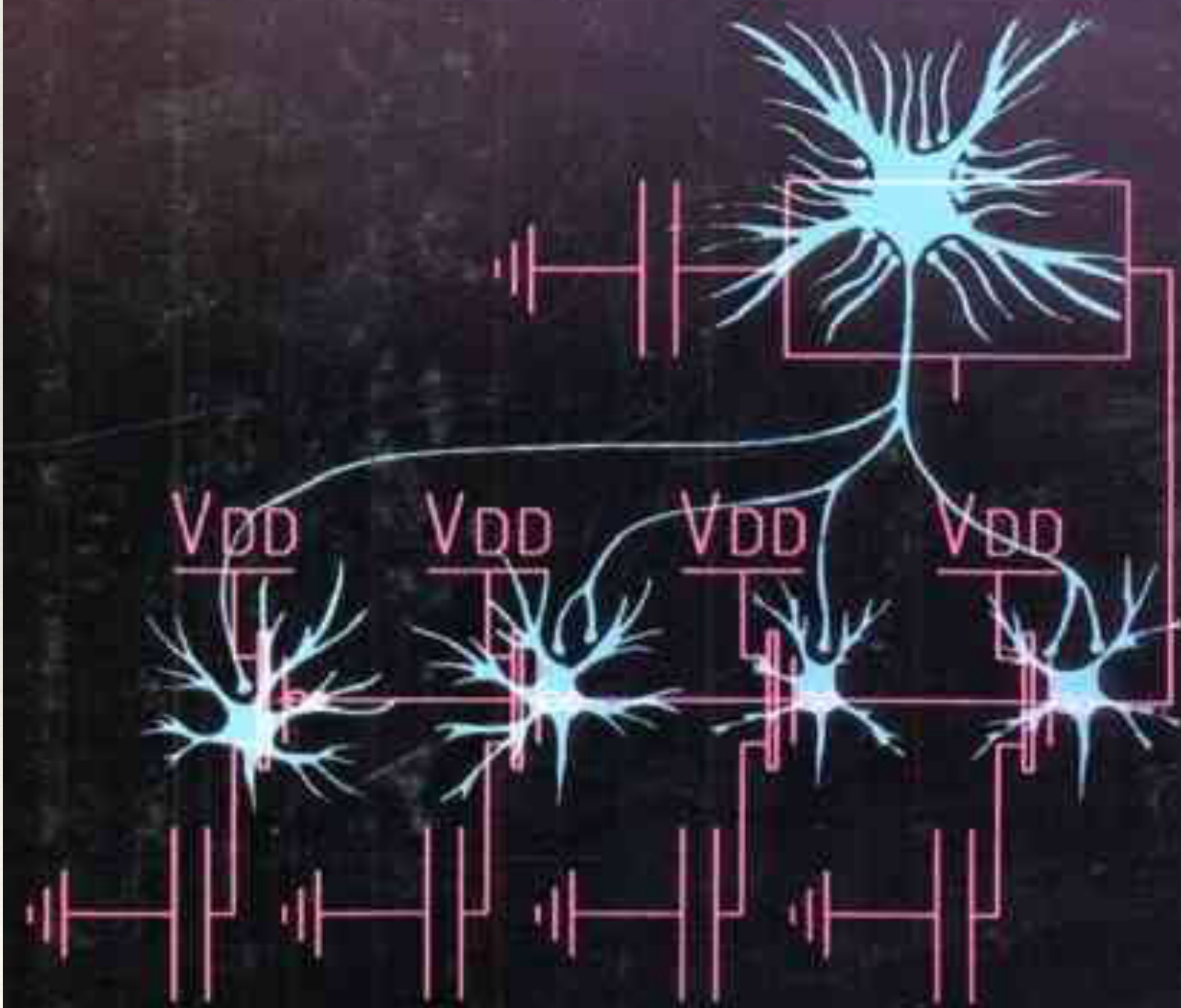
## Collaborators

**Stanford**  
Krishna Shenoy  
Tirin Moore  
Oussama Khatib  
**Waterloo**  
Chris Eliasmith  
Terry Stewart  
**Cornell**  
Rajit Manohar  
Ned Bingham  
**Funding**  
NIH Pioneer, TR01  
ONR: C. Baatar



# Analog VLSI and Neural Systems

Carver Mead



“Listen to the technology and find out what it is telling you.”

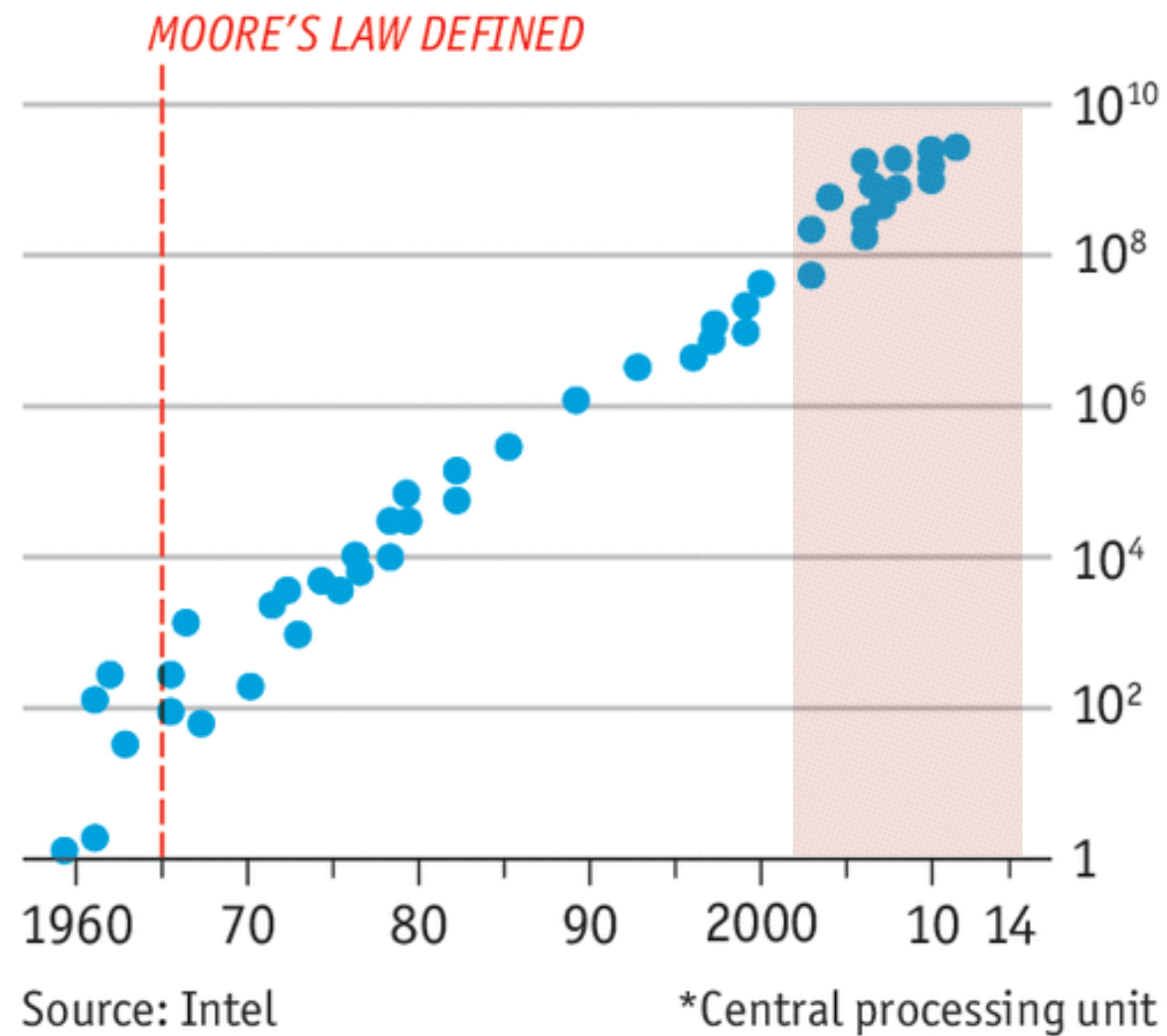
– Carver Mead, 1999



# It's getting really hard to shrink transistors

## A persevering prediction

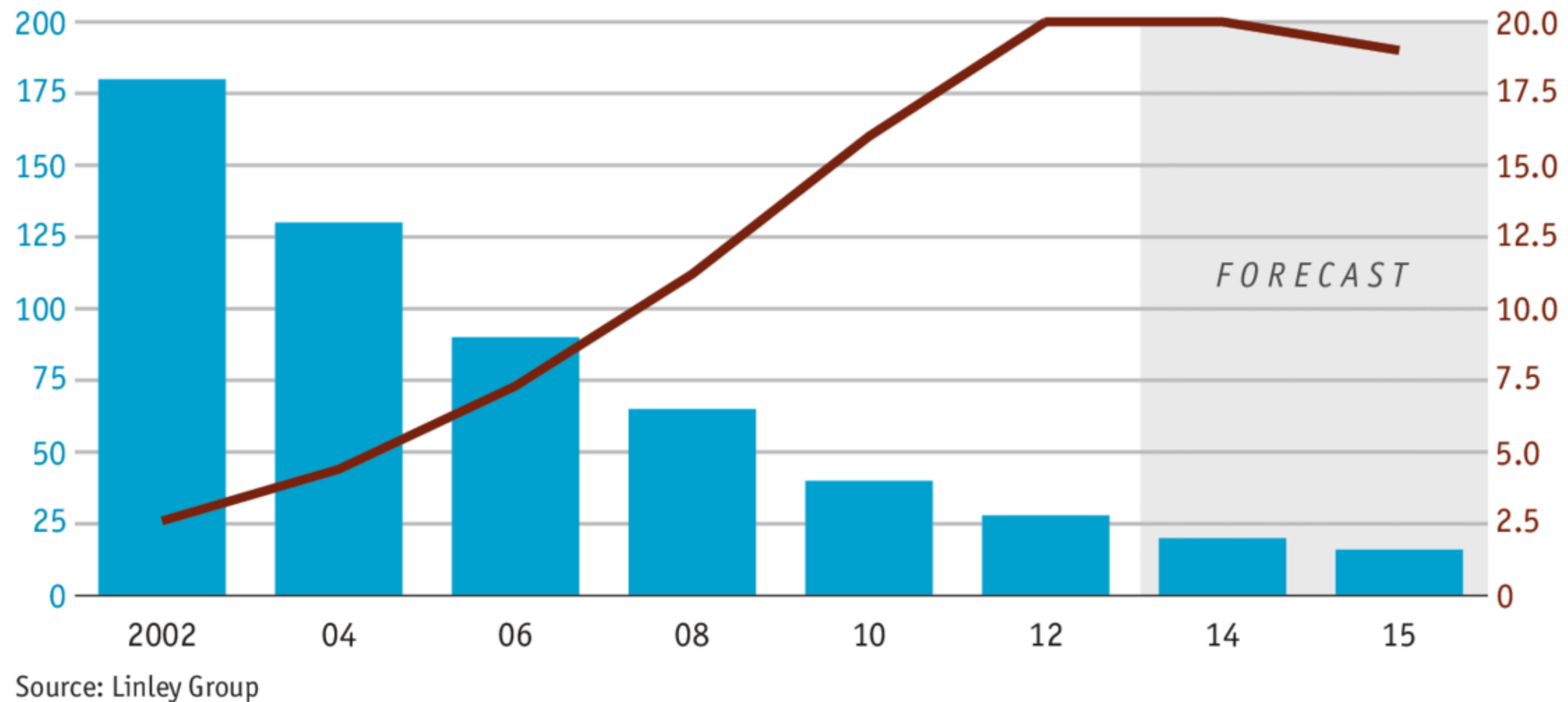
Number of transistors in CPU\*  
Log scale



## Shrinking chips

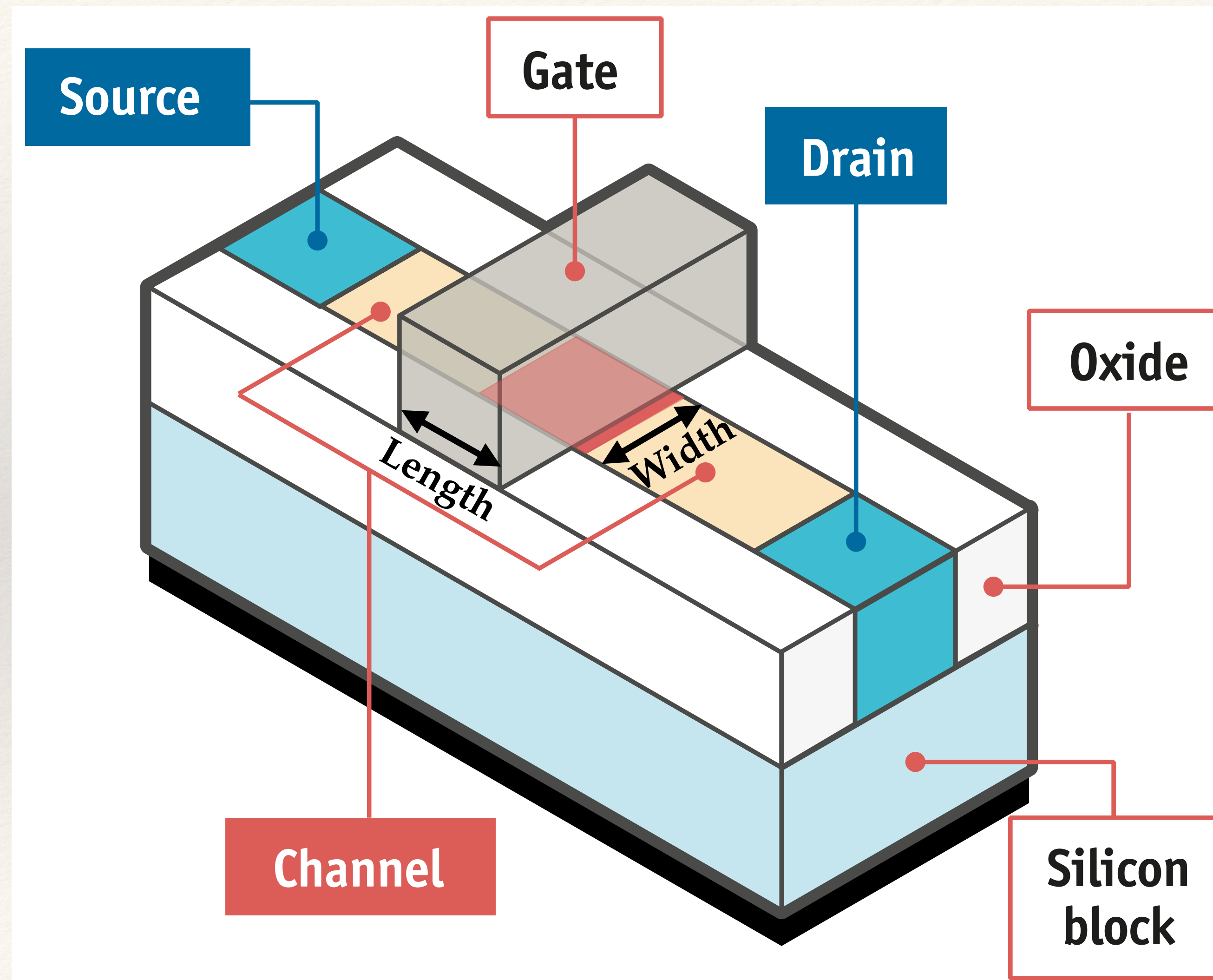
Number and length of transistors bought per \$

*Transistor size, nanometres (nm)*



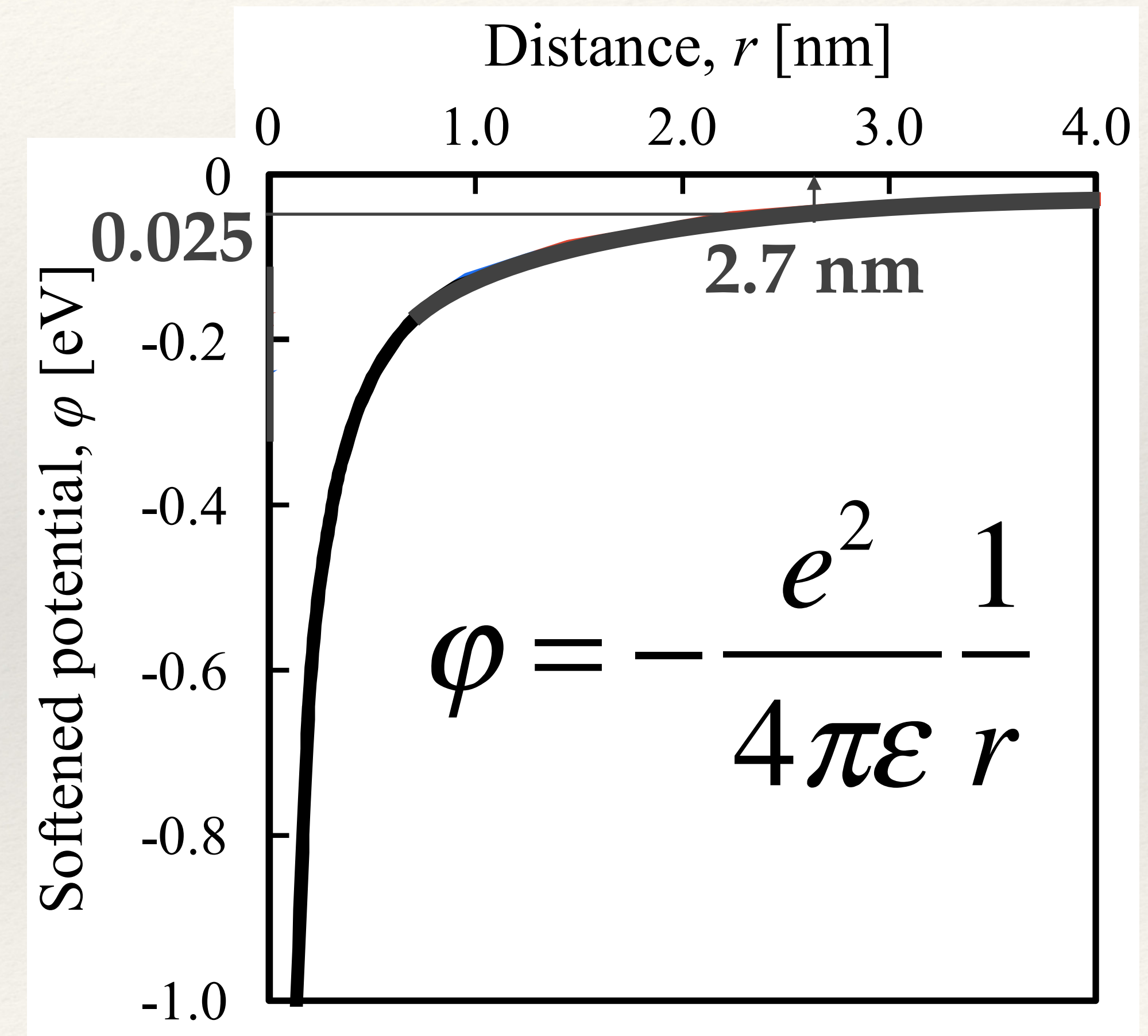


# The MOS Transistor



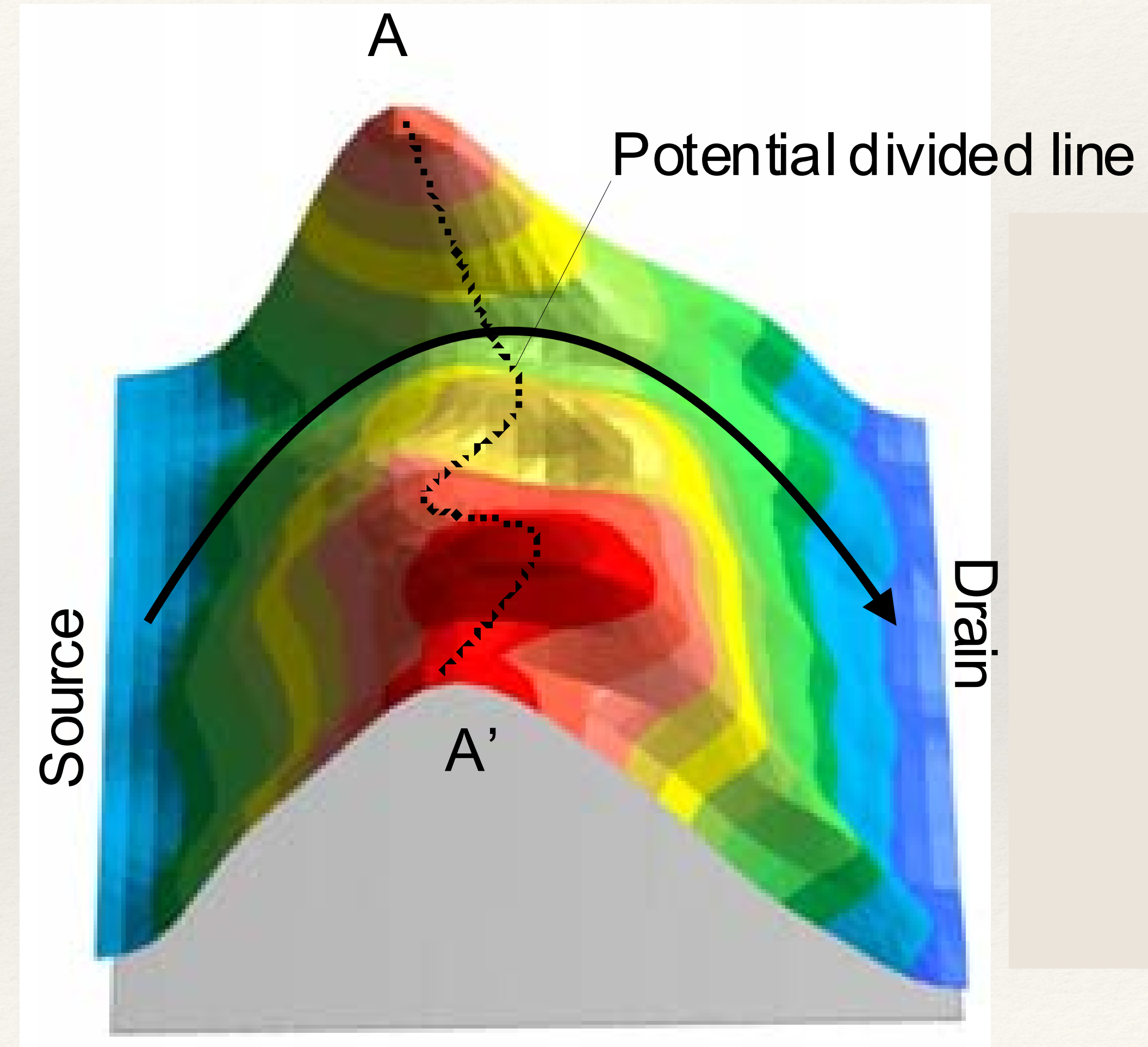


# How many lanes on the transistor's freeway?





# Beware of potholes!



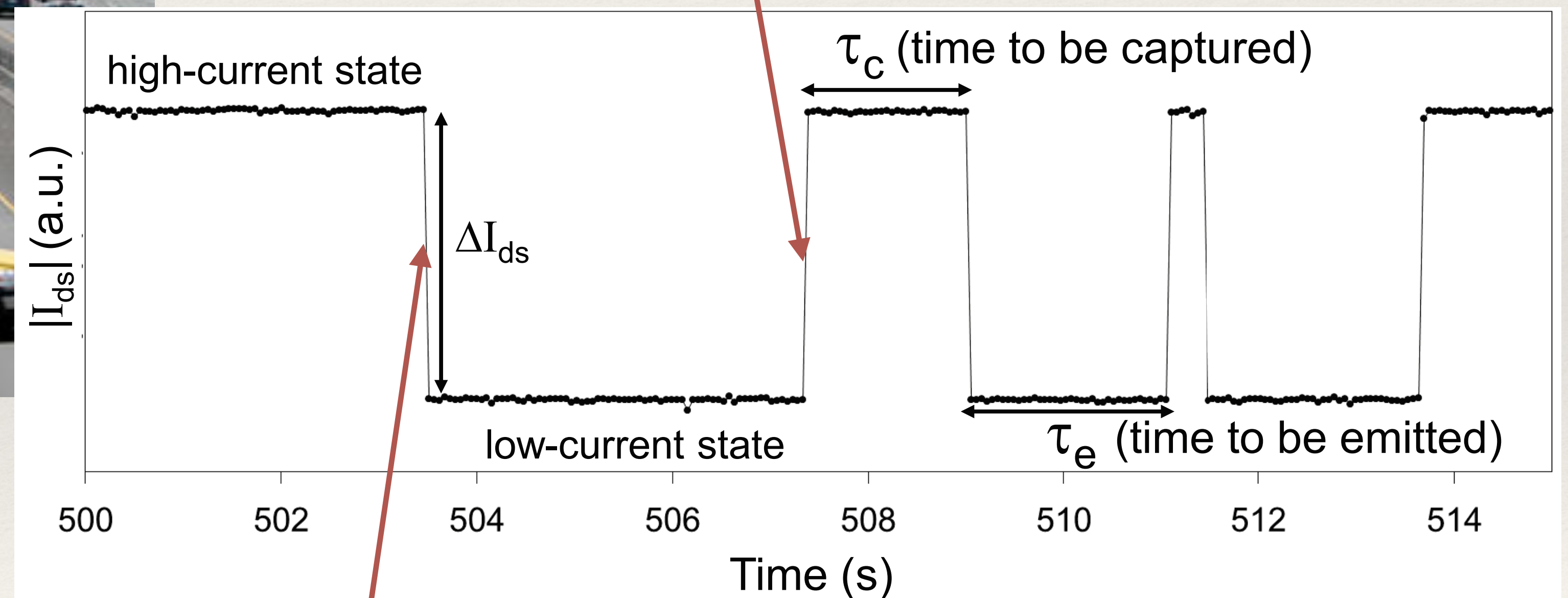
*Saraya et al 2011*



# Accidents occur!



Electron escapes



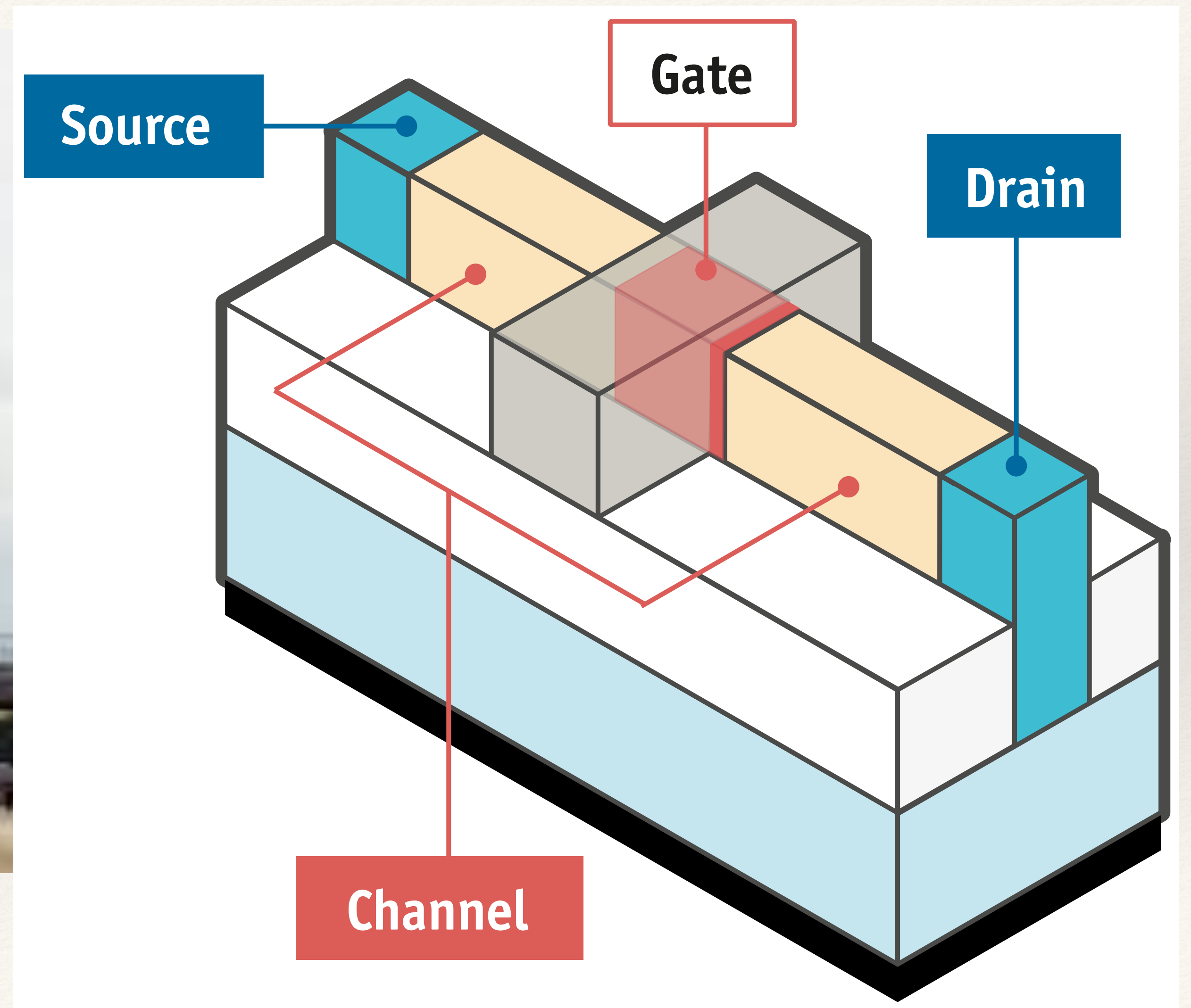
Electron trapped



# 2014: Industry went 3D to get more lanes

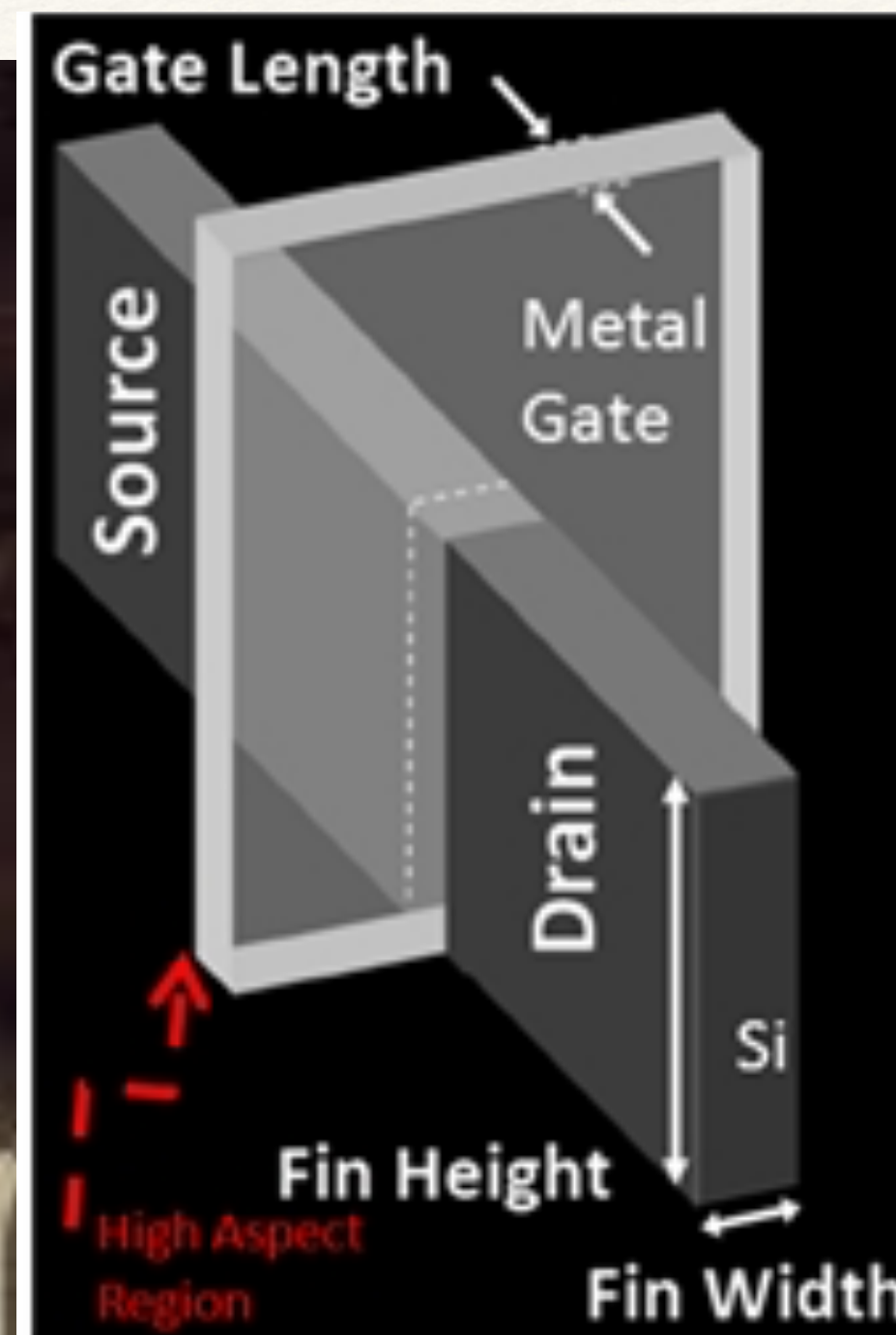


*The Economist* 2016

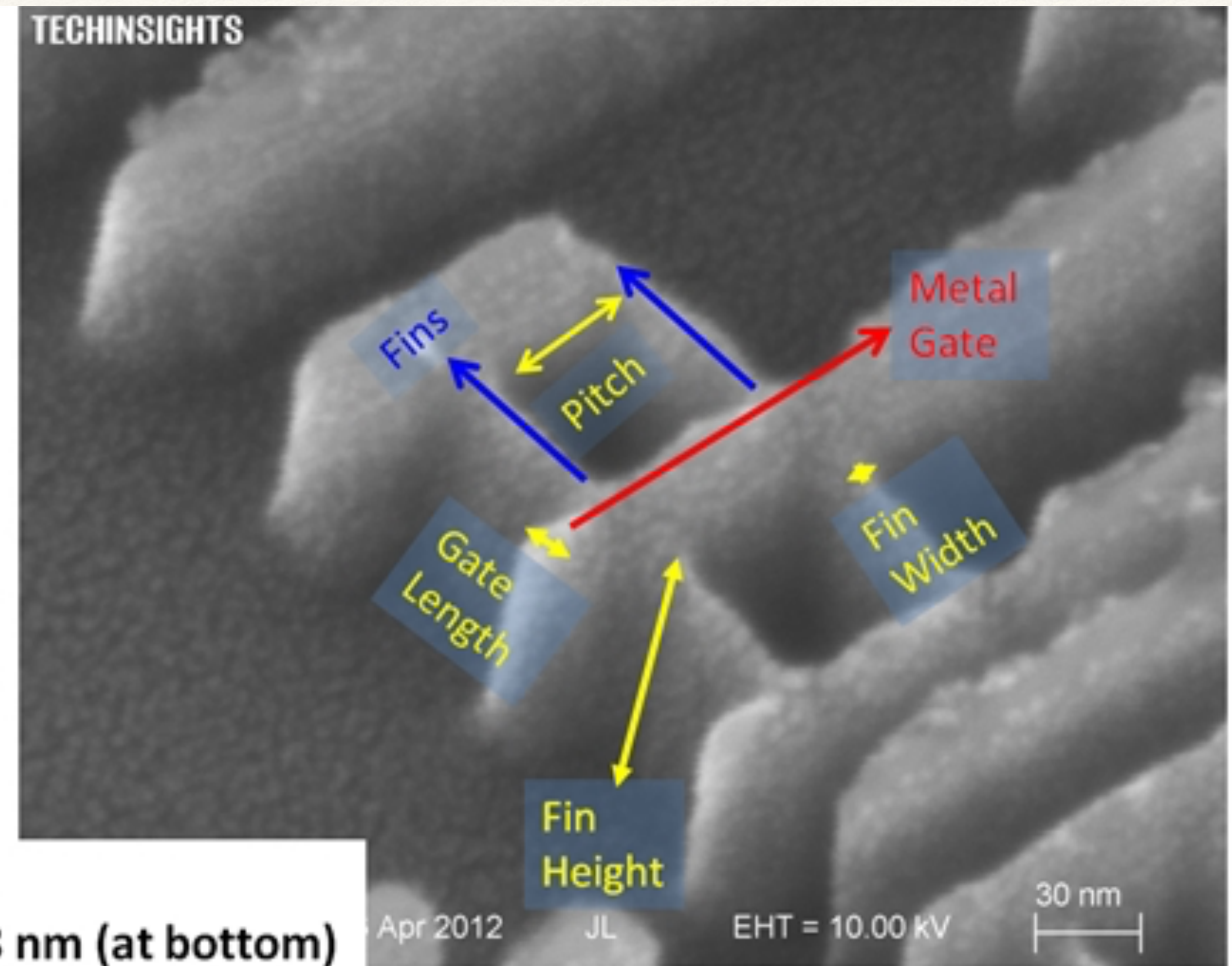




# Intel's 28nm FinFET: 30 lanes (13+4+13)



Average Gate height:  
Average Fin width: 18 nm (at bottom)  
Average Fin height: 35 nm  
Average Fin pitch: 60 nm

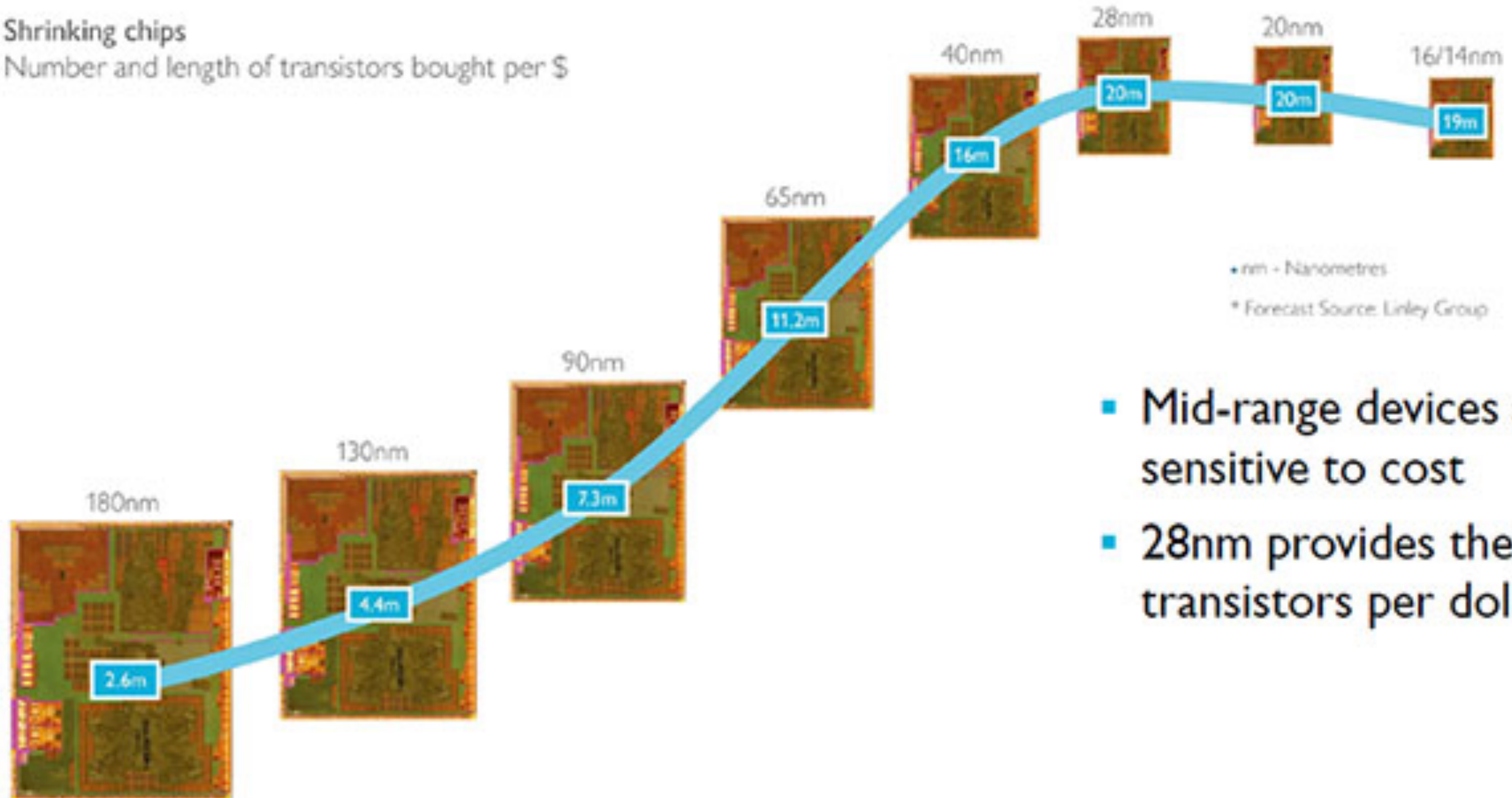




# 28nm: Optimal Balance of Cost and Power for 2015 Devices

Shrinking chips

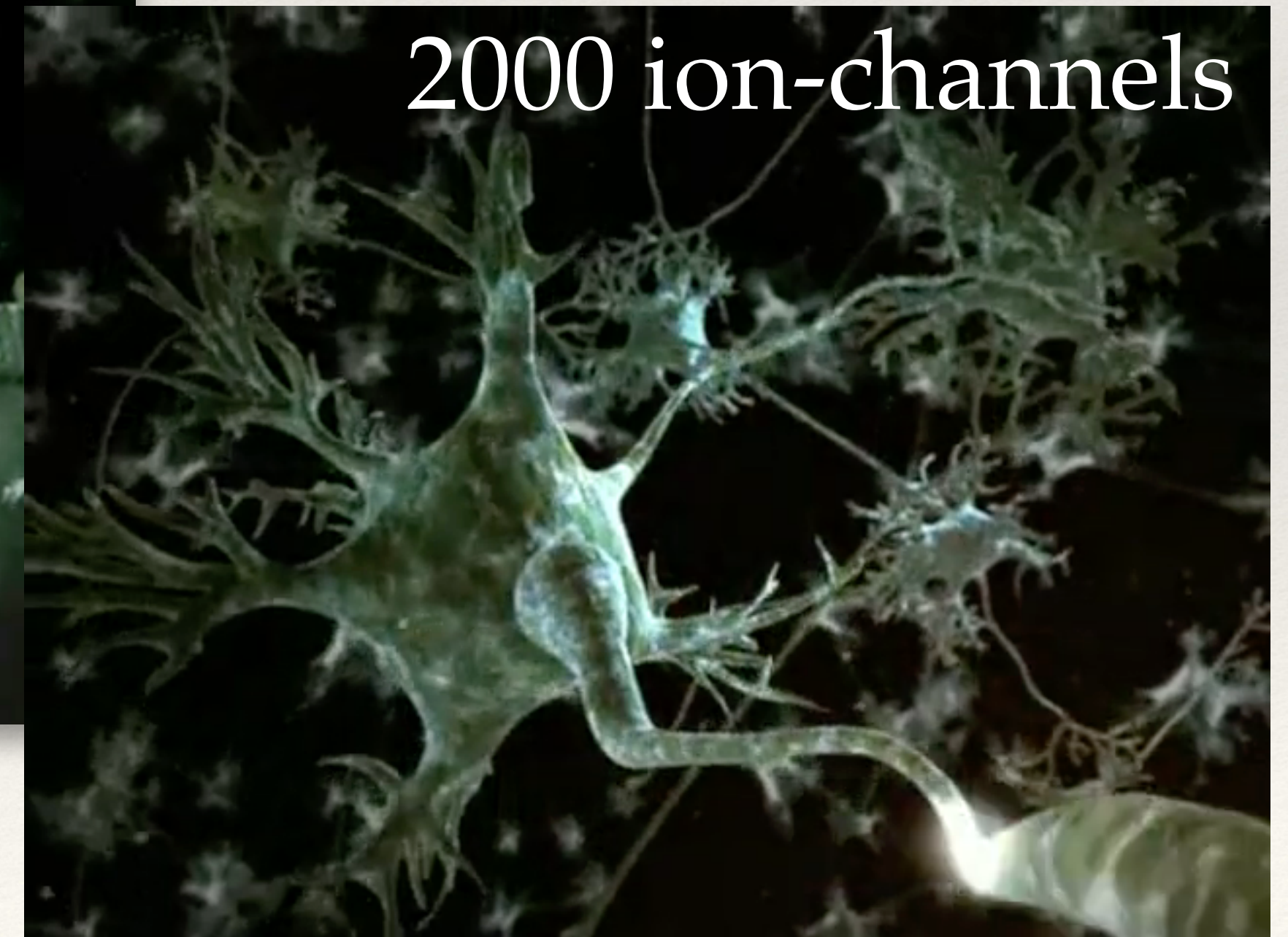
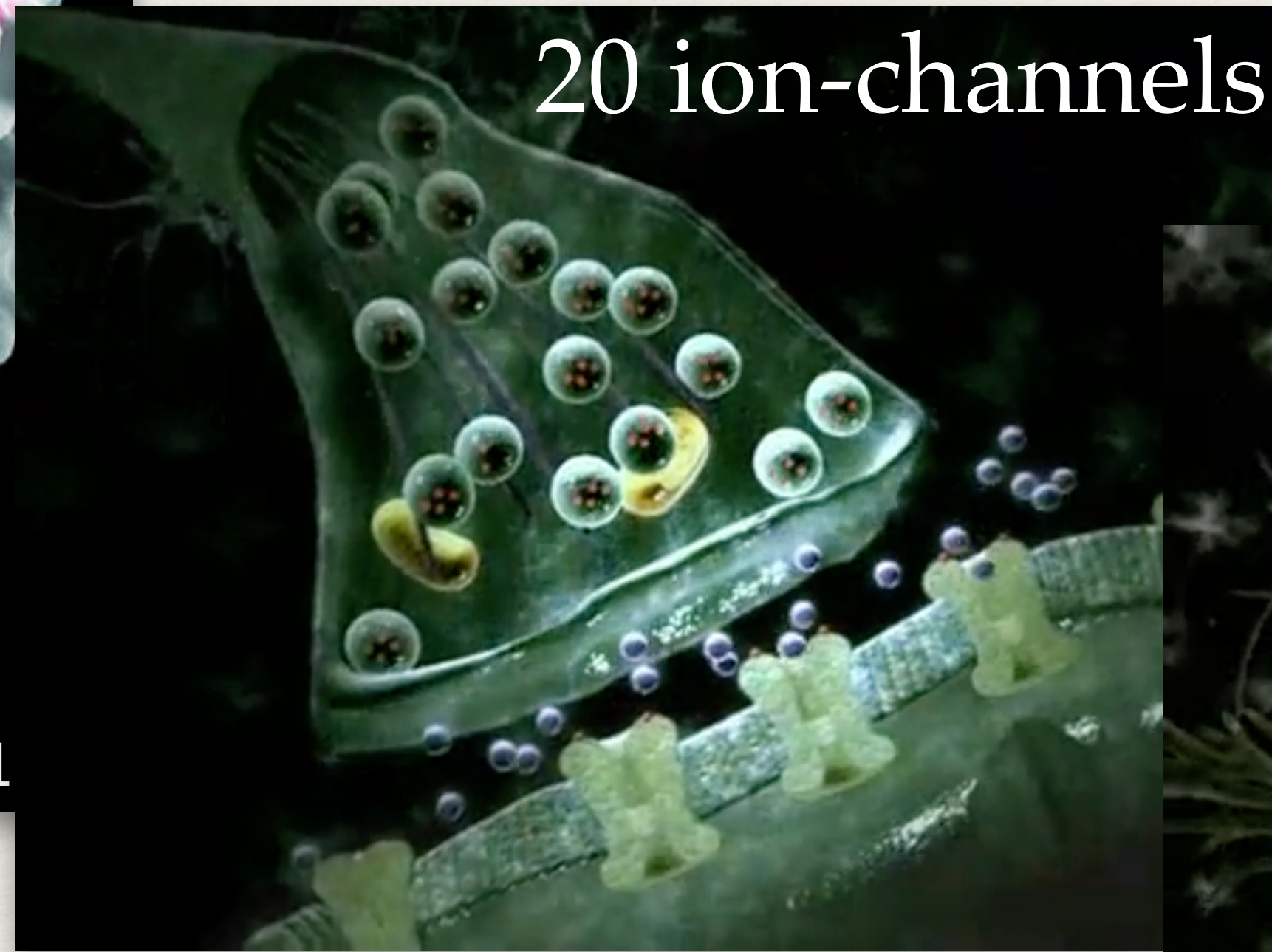
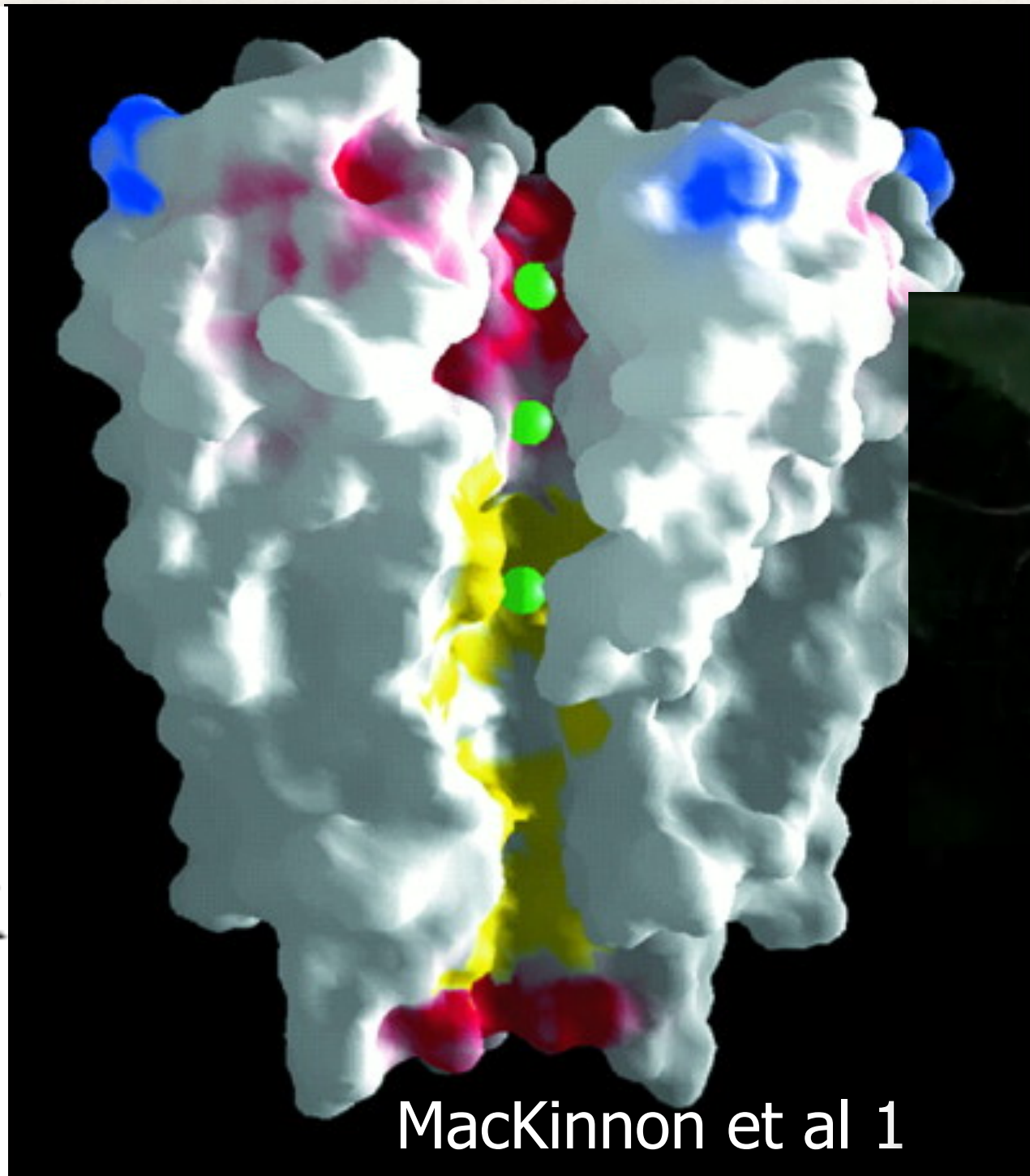
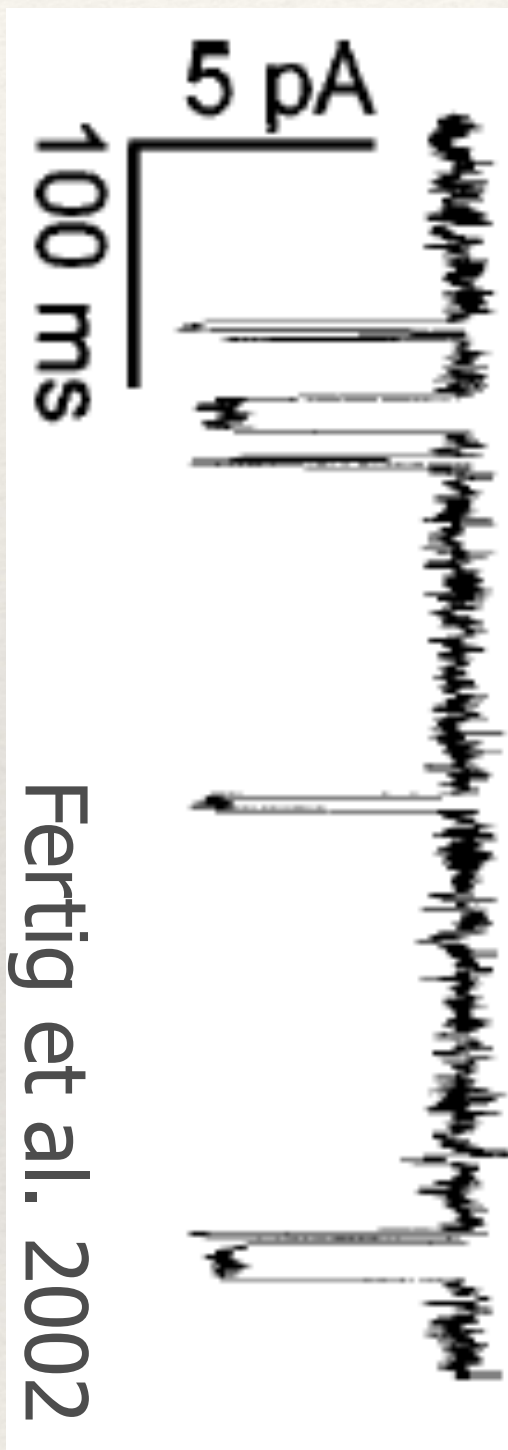
Number and length of transistors bought per \$



- Mid-range devices are highly sensitive to cost
- 28nm provides the most transistors per dollar

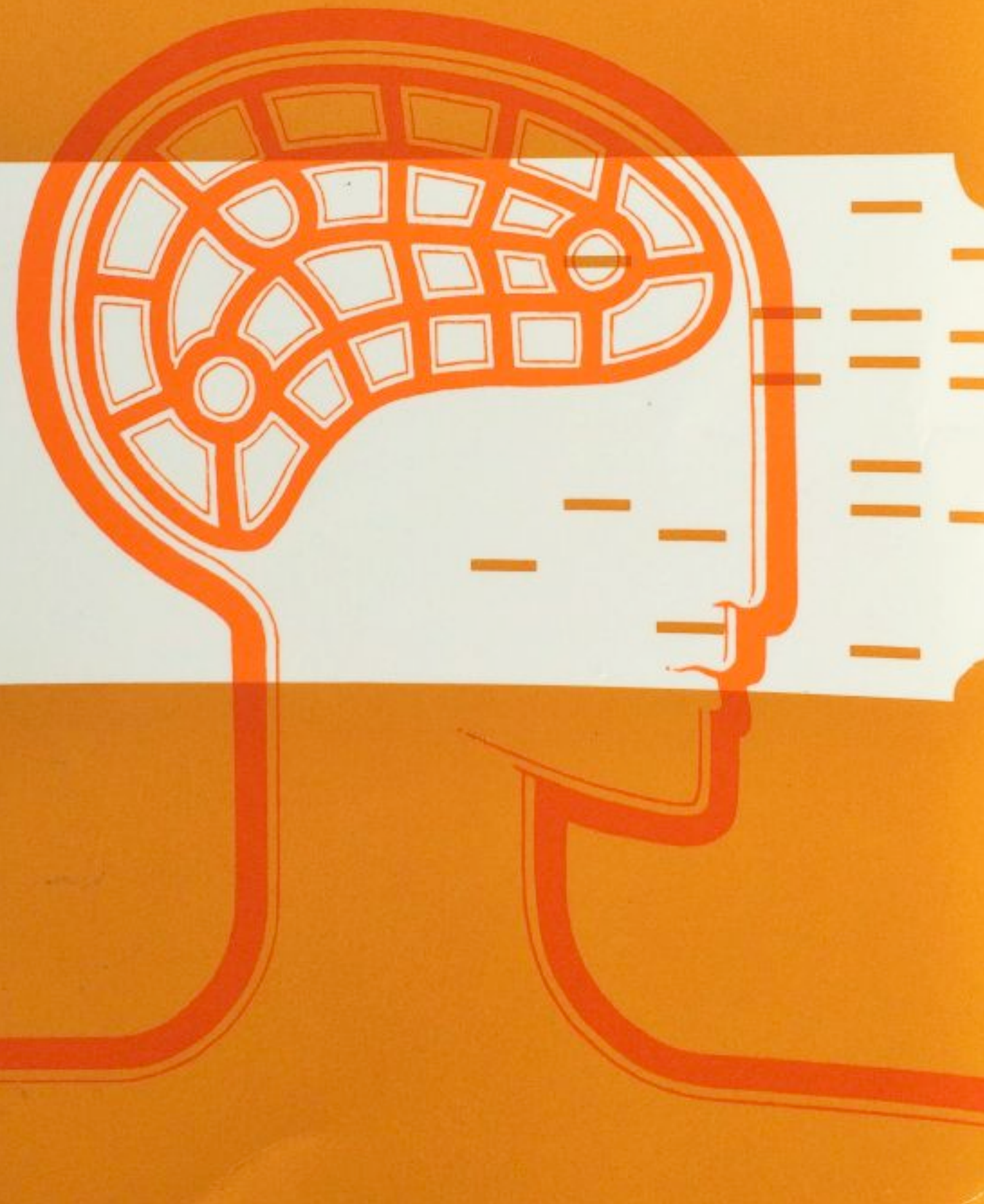


# Brains operate fine with single-lane ion-channels





# JOHN VON NEUMANN THE COMPUTER AND THE BRAIN

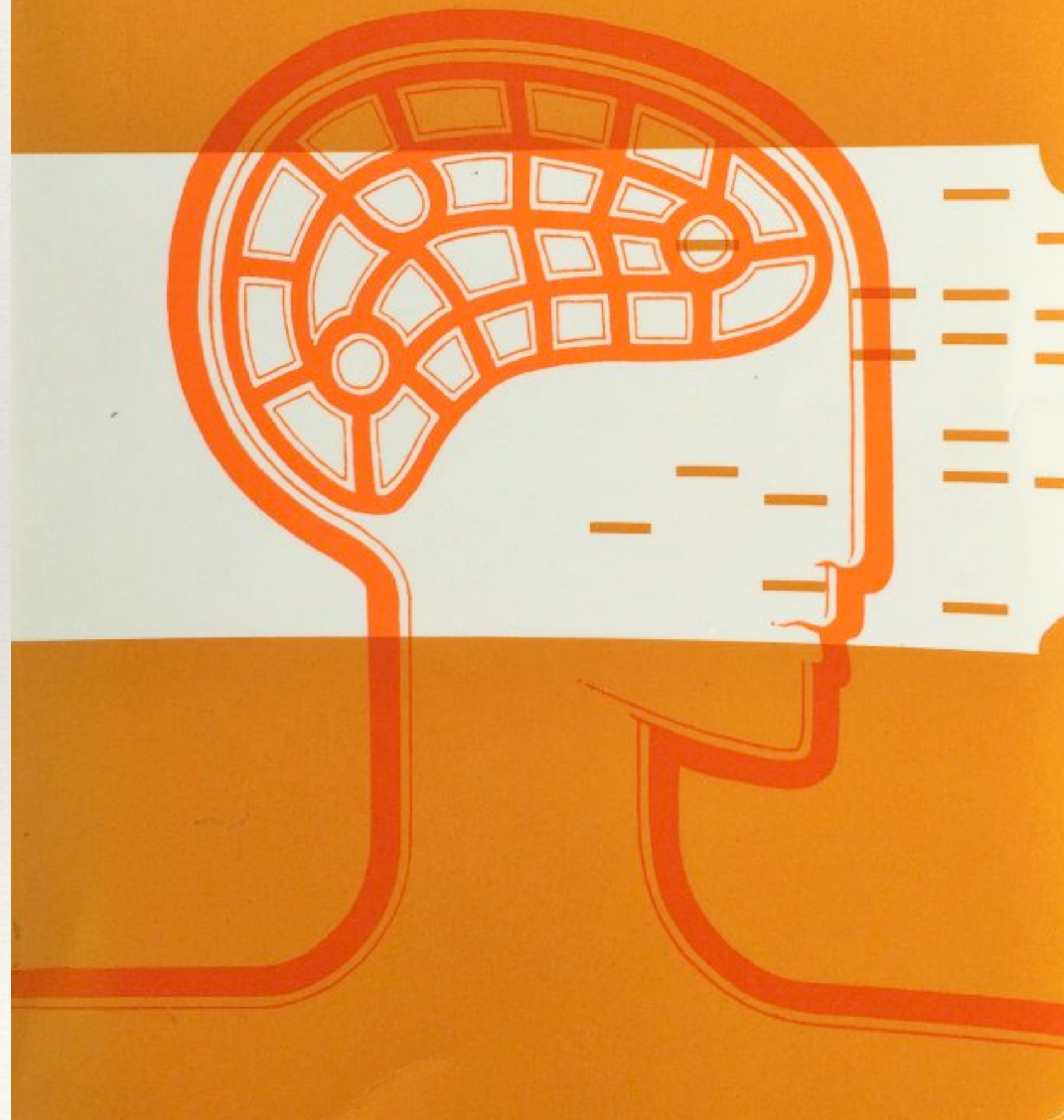


“The most immediate observation regarding the nervous system is that its functioning is *prima facie* digital.”

– John von Neumann 1957



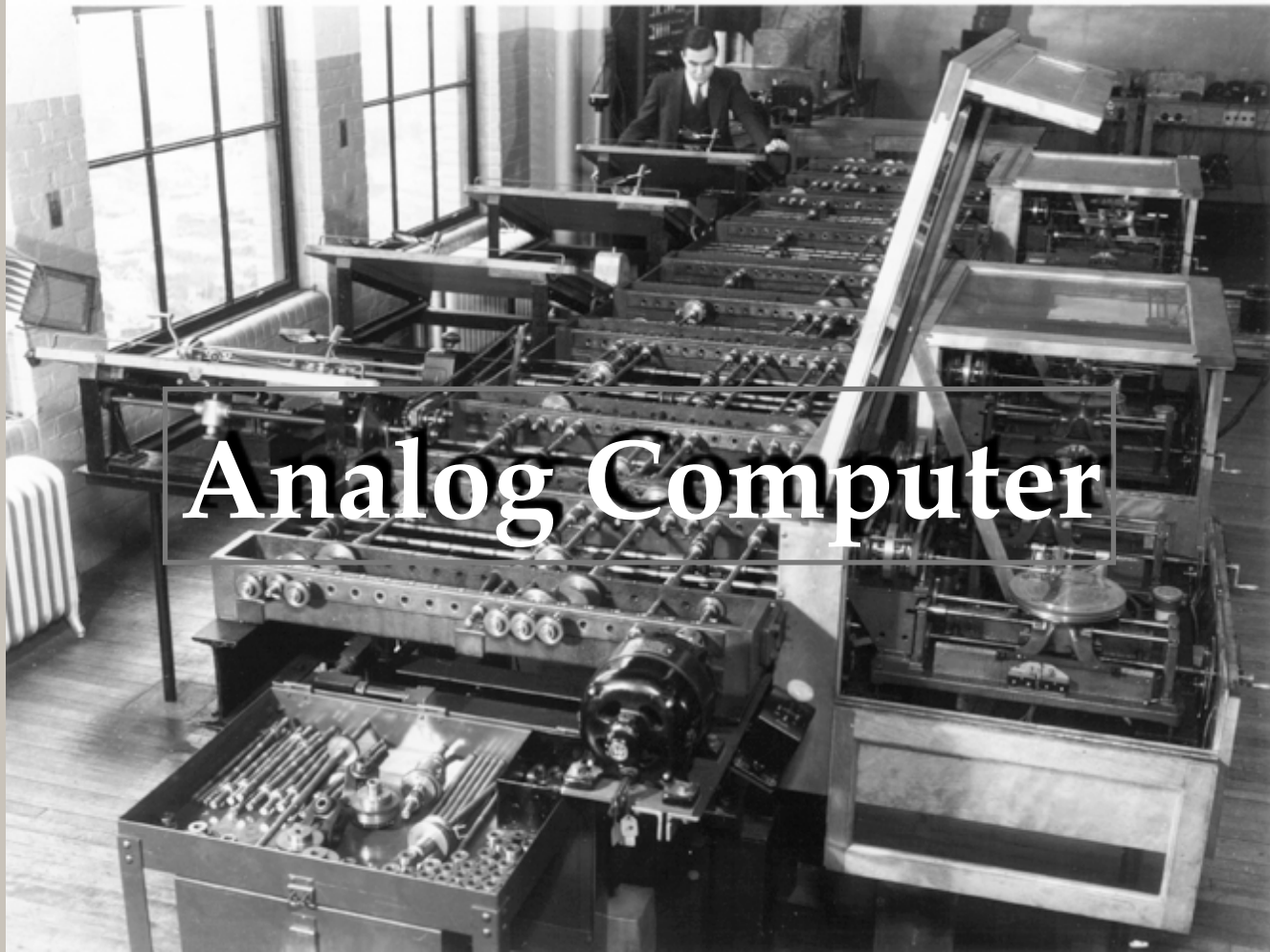
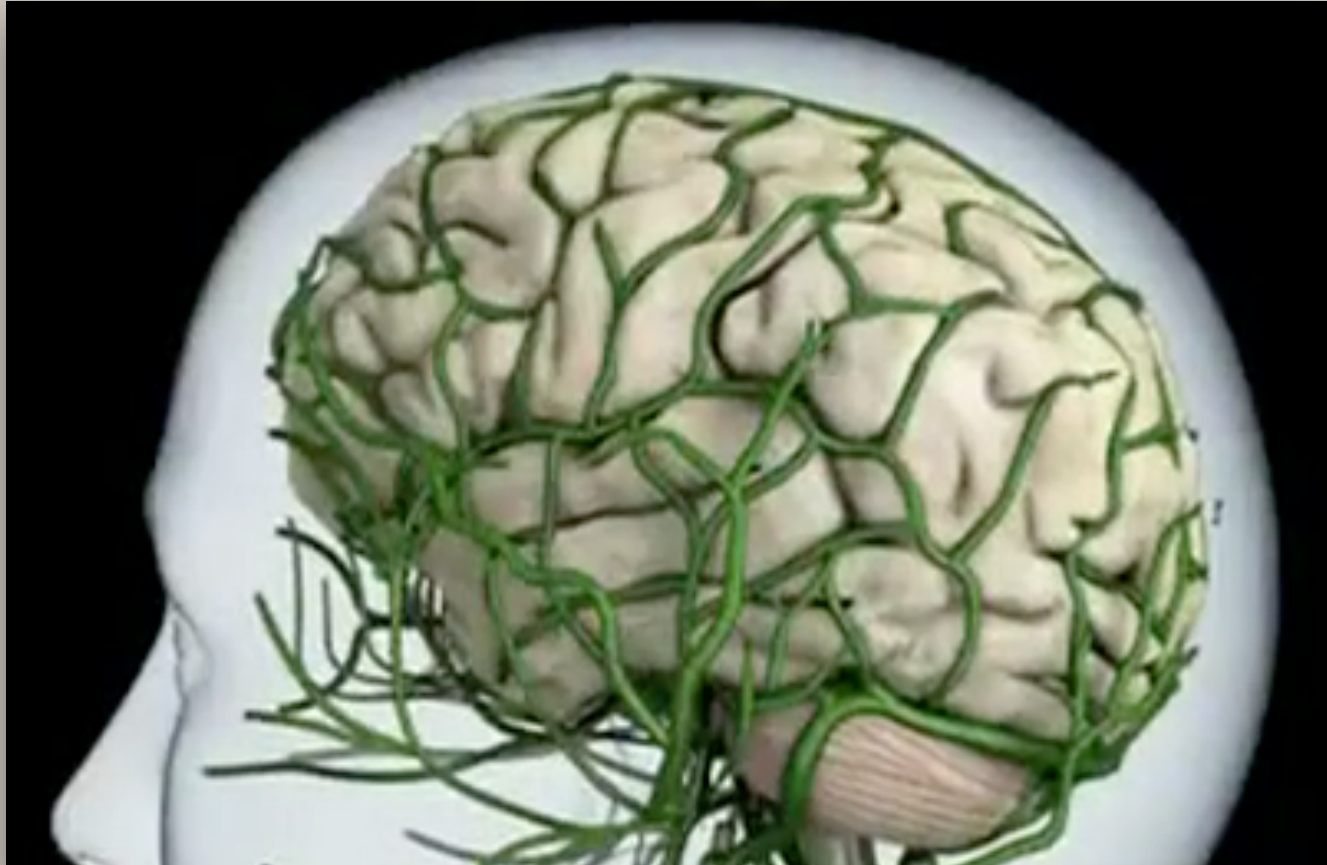

# JOHN VON NEUMANN THE COMPUTER AND THE BRAIN



“Thus all the complexities referred to here may be irrelevant, but they may also endow the system with a (partial) analog character, or a ‘mixed’ character.”

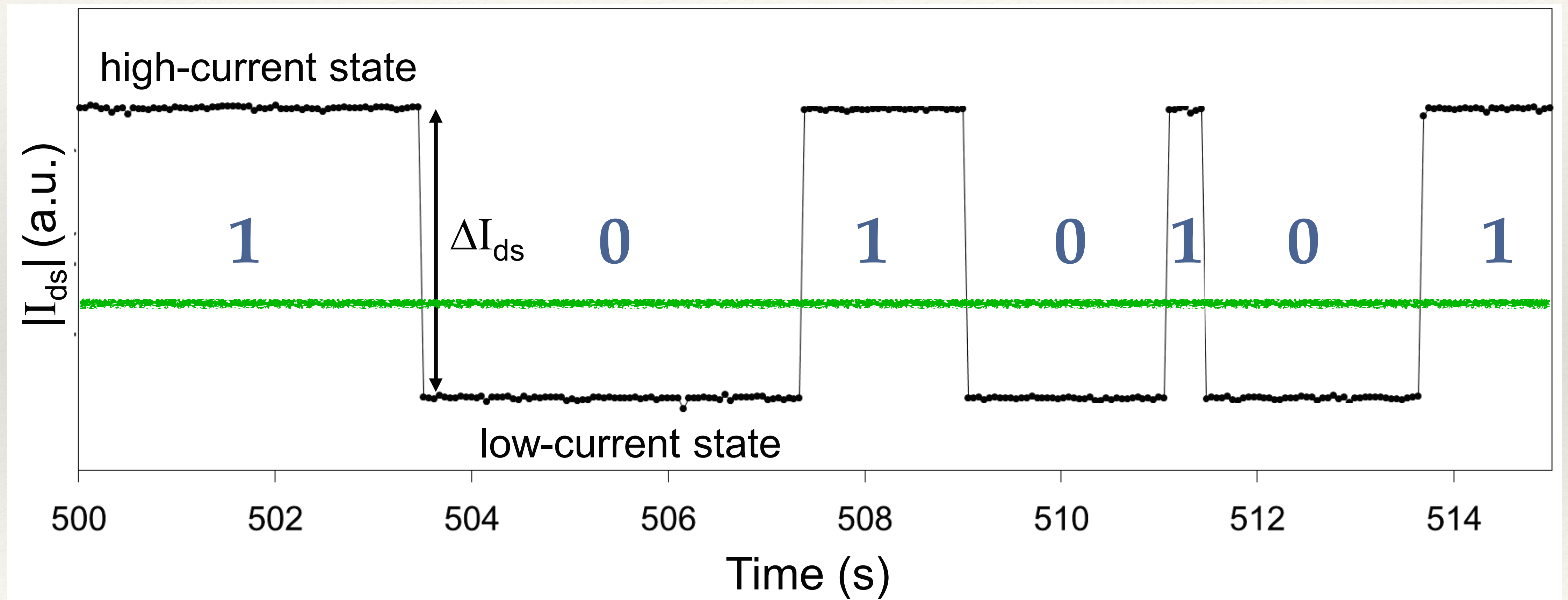
– *John von Neumann 1957*



Communication		Computation	
		Analog	Digital
Communication	Analog	 <p>Analog Computer</p>	
	Digital		 <p>Digital Computer</p>

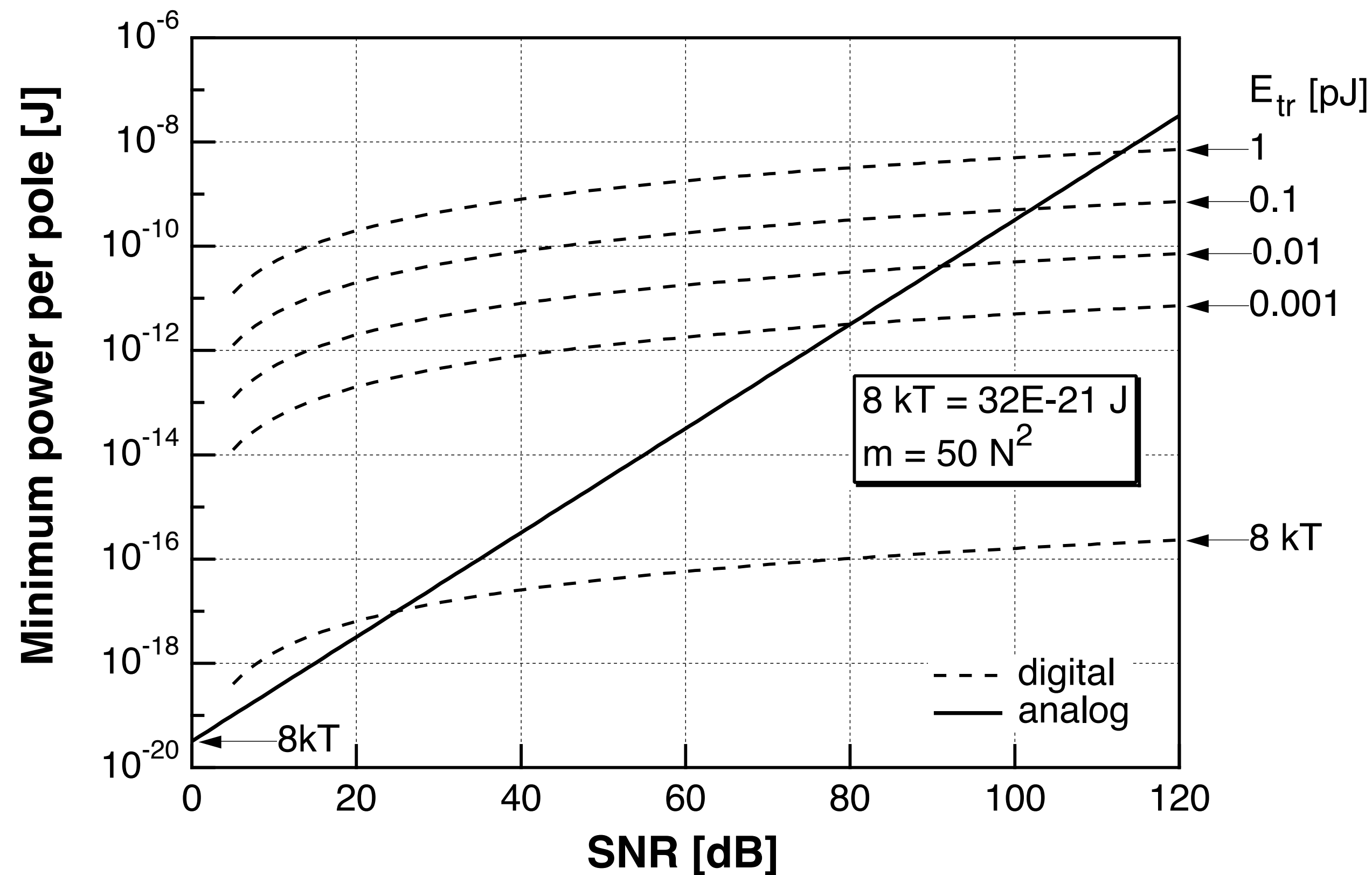


# Difference between Digital and Analog

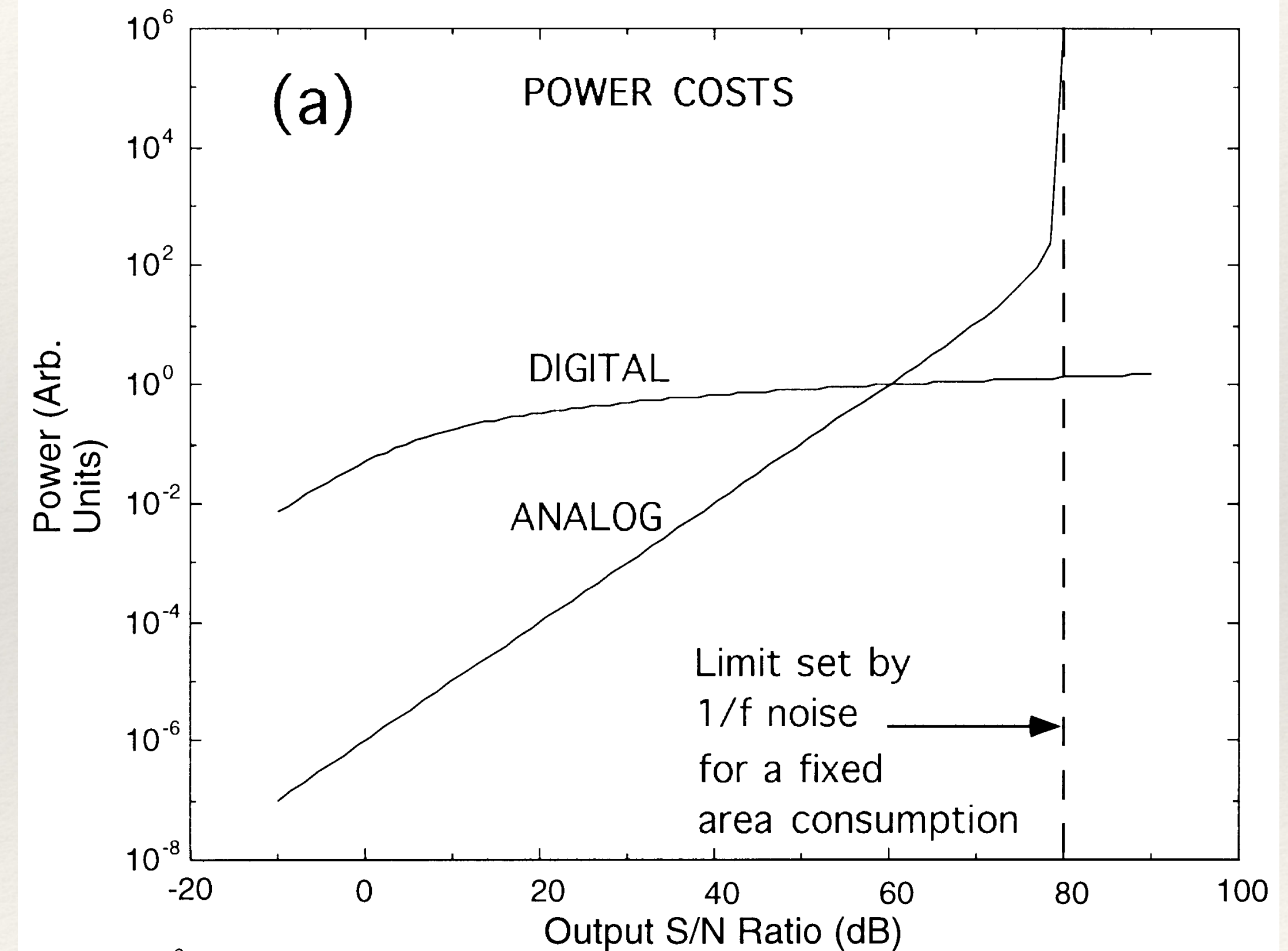




# Analog versus Digital Computation: Energy Cost



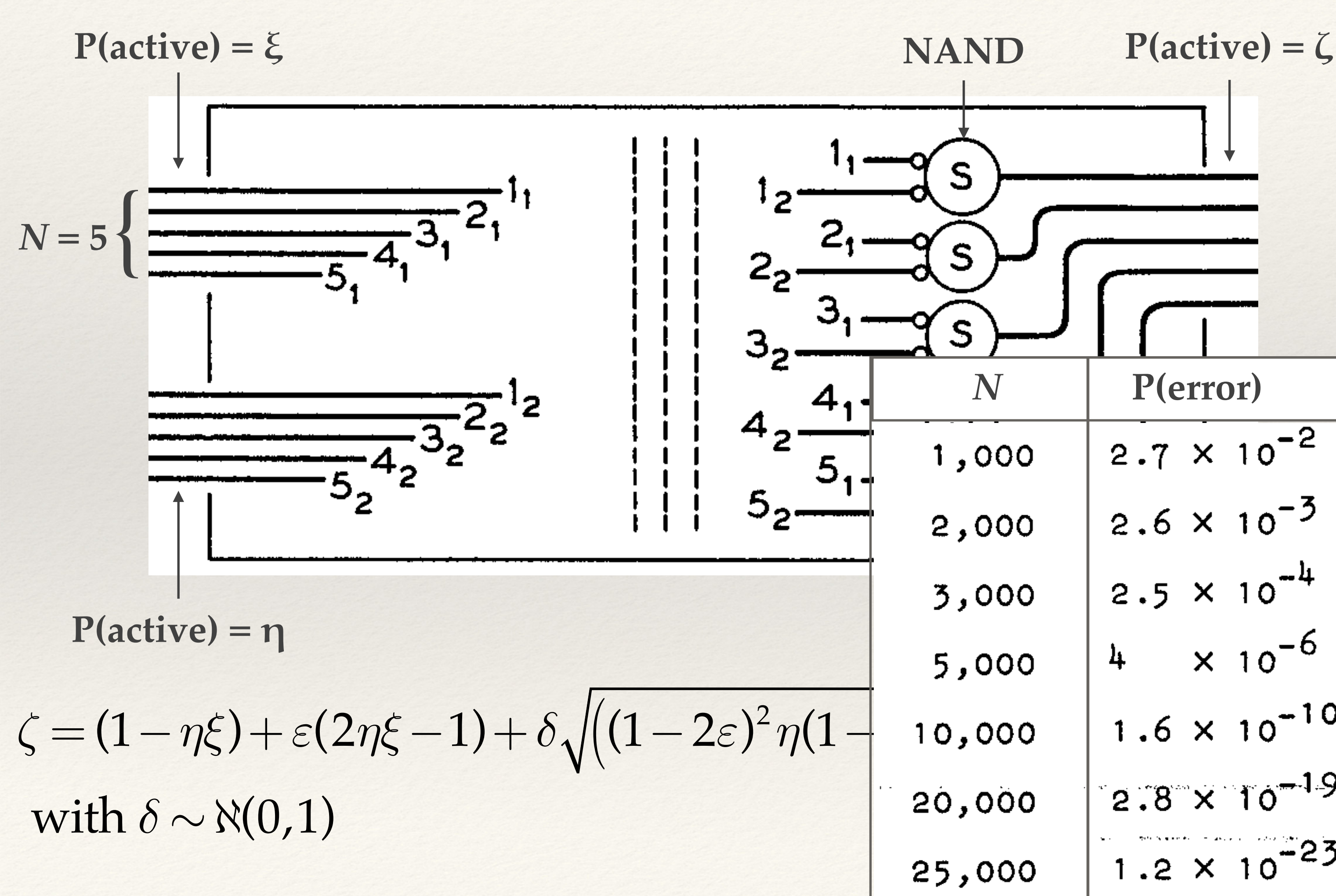
Enz & Vittoz 1996



Sarpeshkar 1998



# Robust Digital Computation: N-way Redundancy



❖ Binary code:

$\zeta > 0.97 \Rightarrow \text{bit} = 1$

$\zeta < 0.03 \Rightarrow \text{bit} = 0$

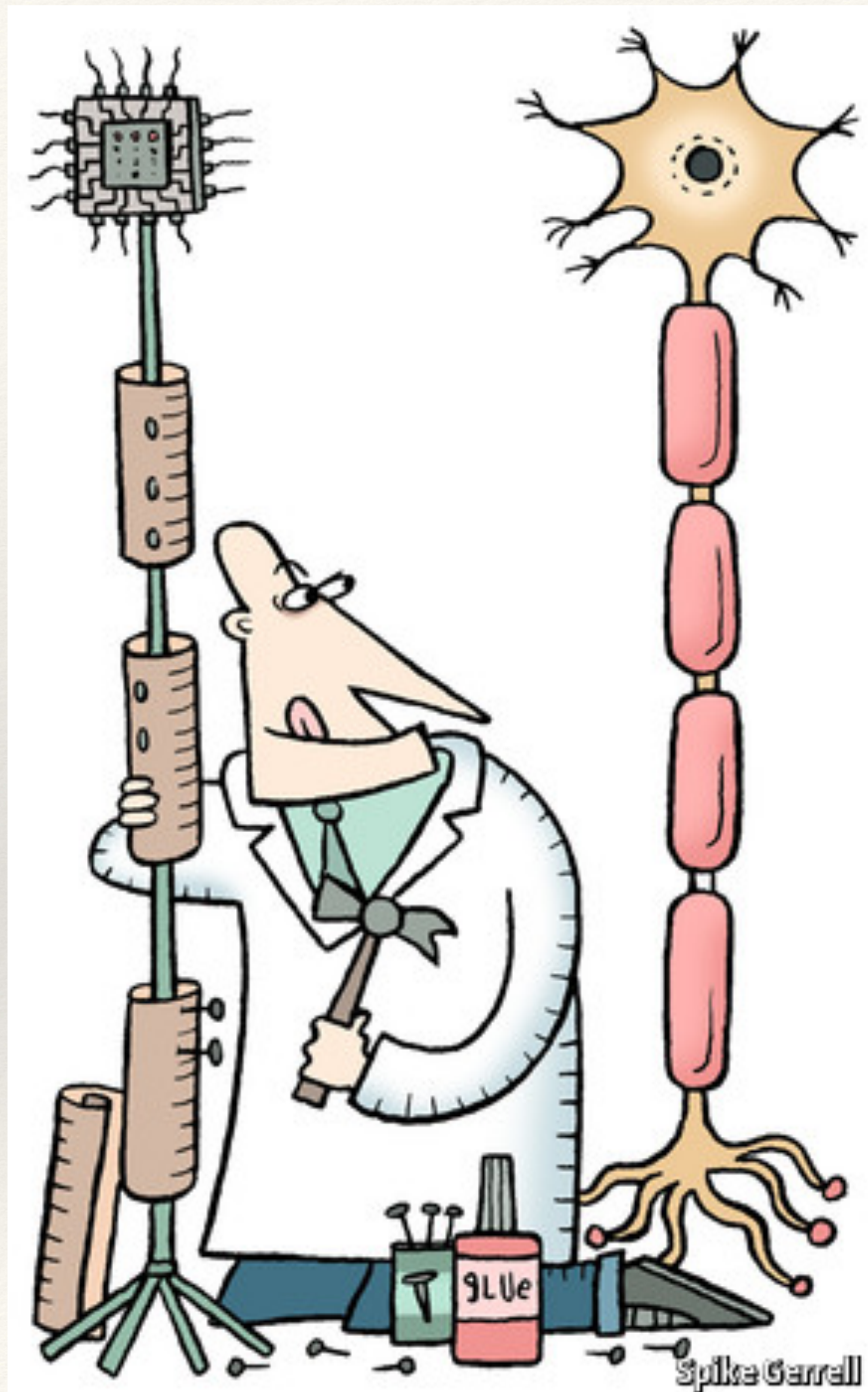
othw  $\Rightarrow$  error

❖ For  $\varepsilon > 1.07 \times 10^{-2}$ , increasing  $N$  doesn't help

❖ For  $\varepsilon = 5 \times 10^{-3}$ :

$$P(\text{error}) = \frac{6.4}{\sqrt{N}} 10^{-\frac{8.6N}{10,000}}$$





How can we unleash the computational power and energy efficiency of nanoscale transistors using analog computation and digital communication?

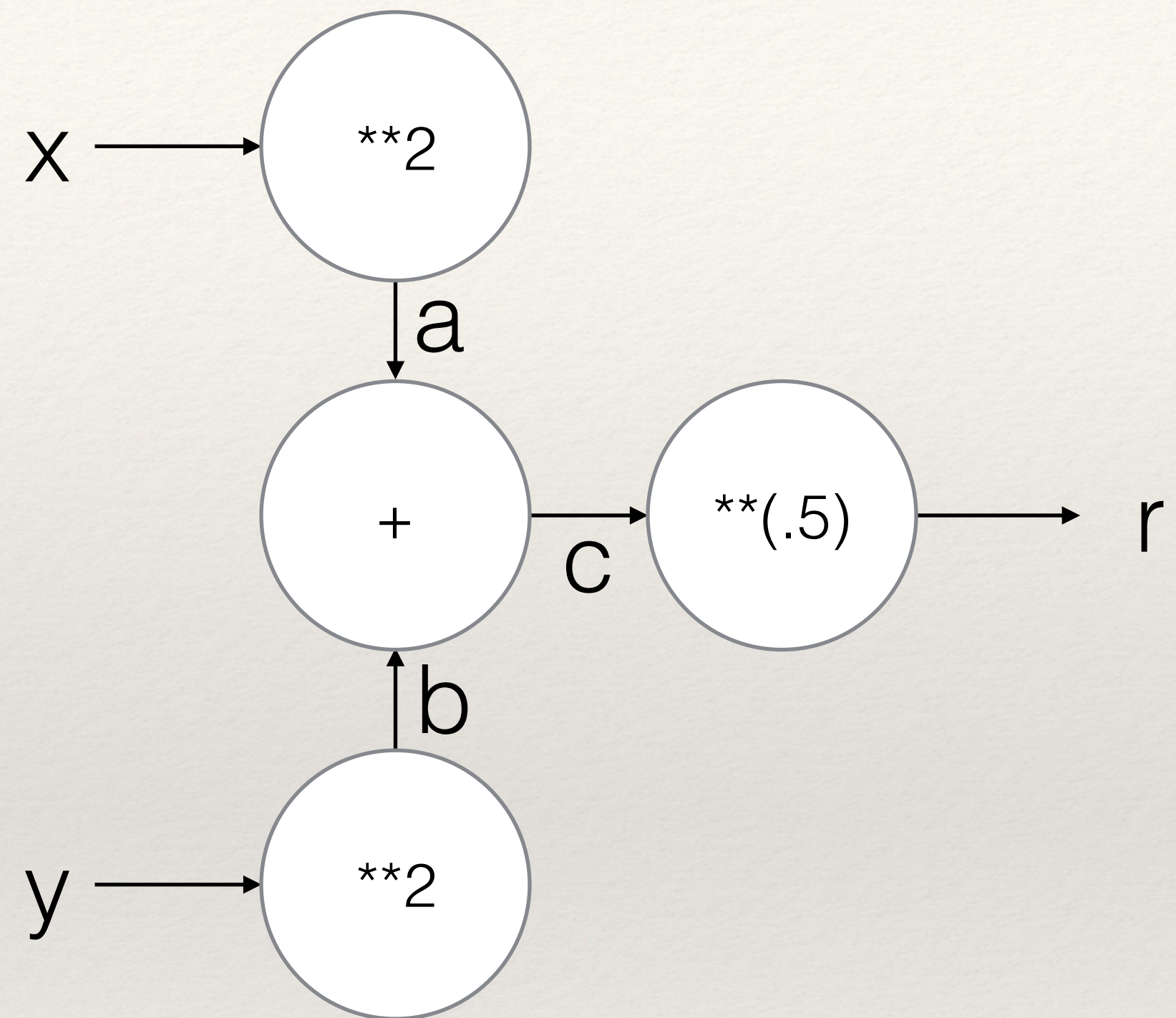


```
while (True):  
    r = (x**2 + y**2)**(.5)
```

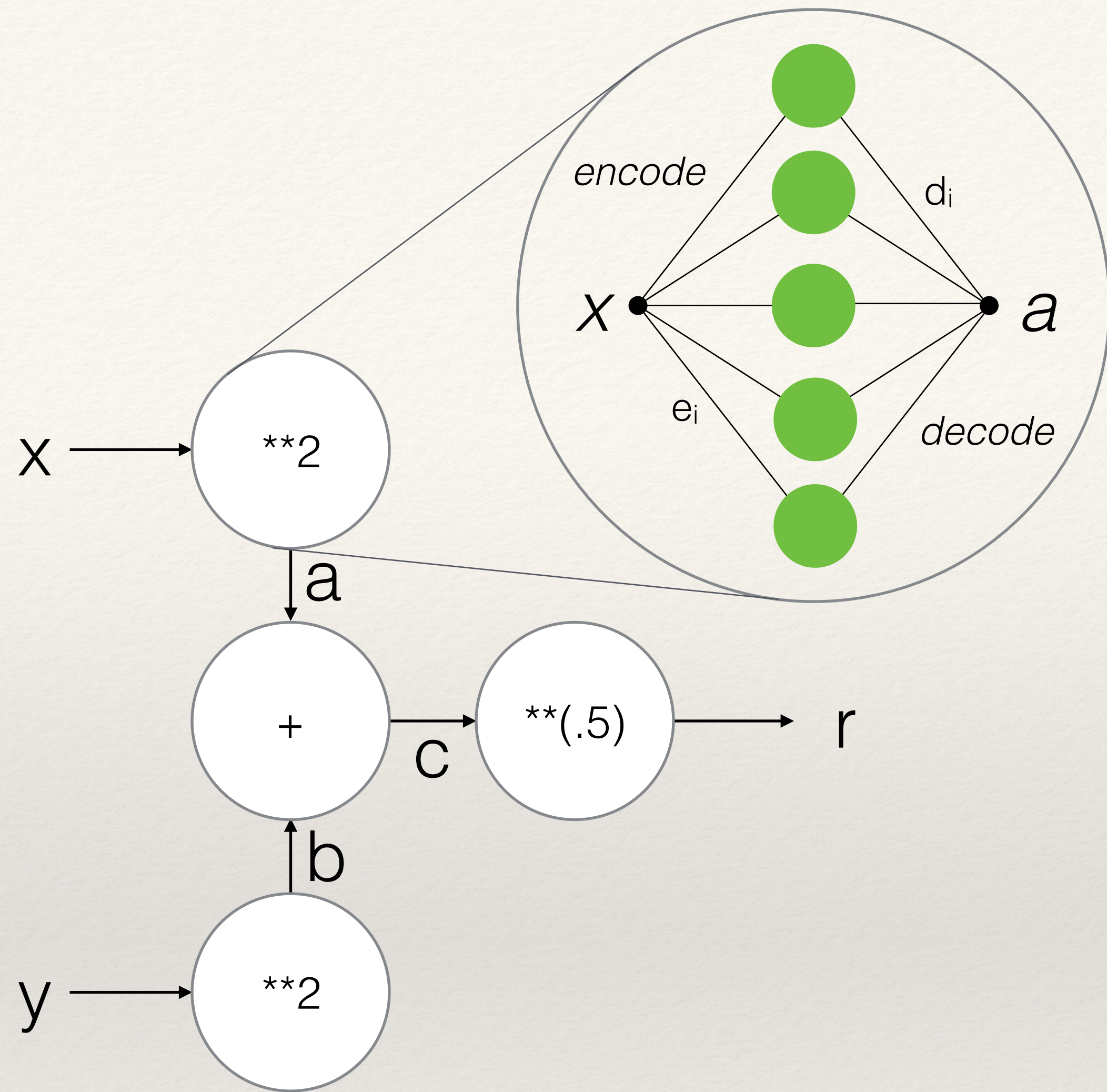


```
while (True):  
    a = x**2  
    b = y**2  
    c = a + b  
    r = c**(.5)
```

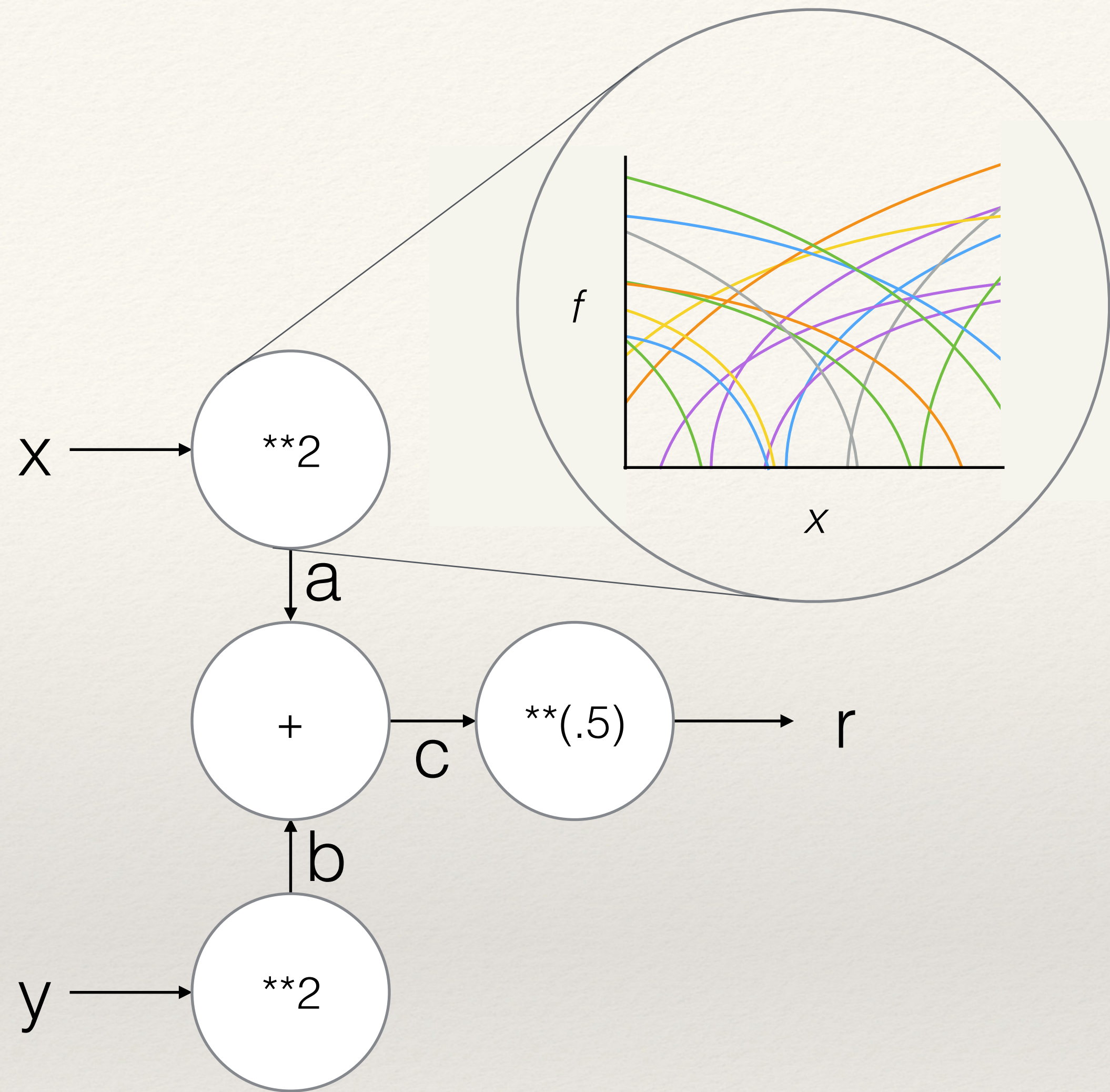




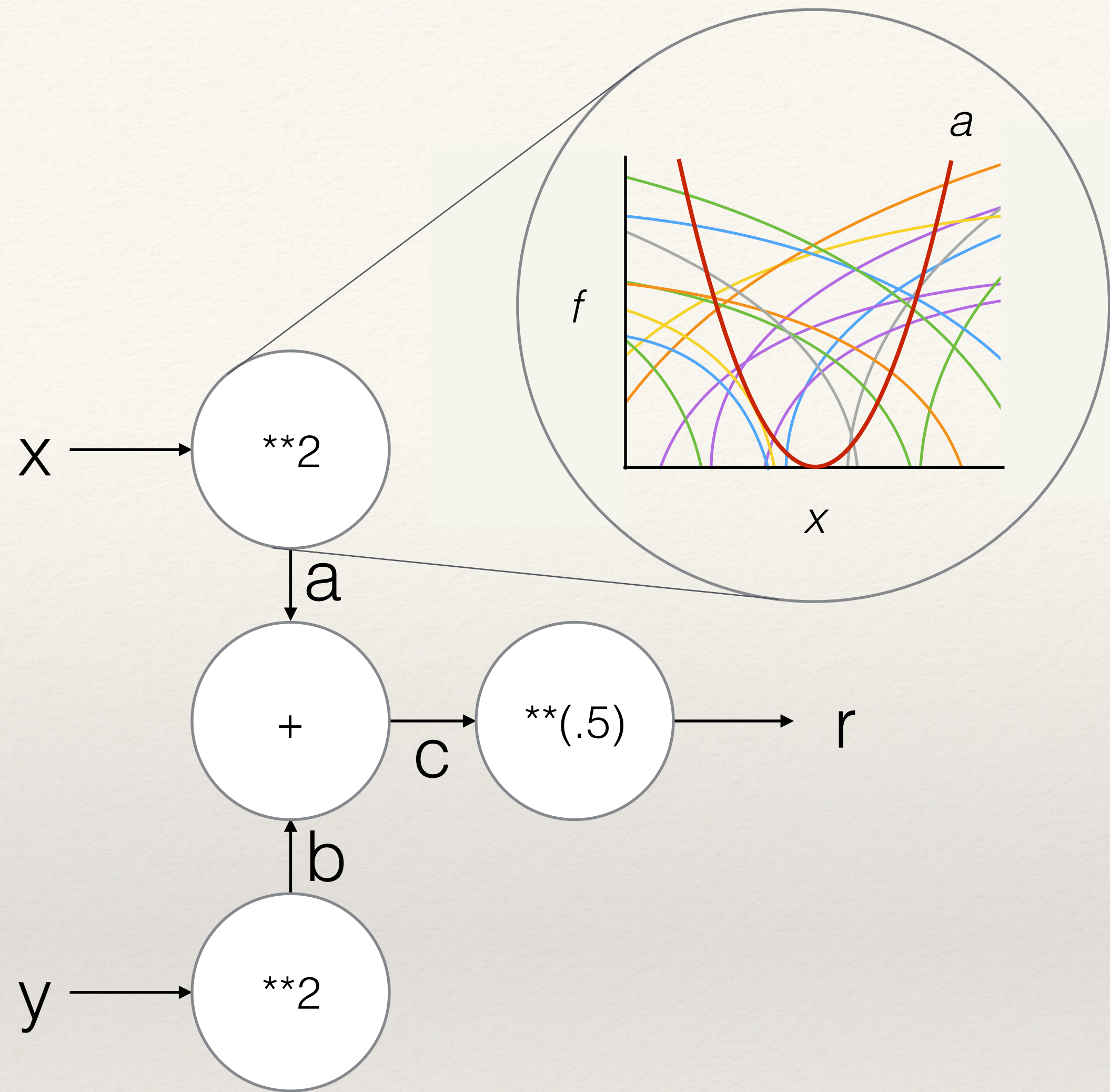




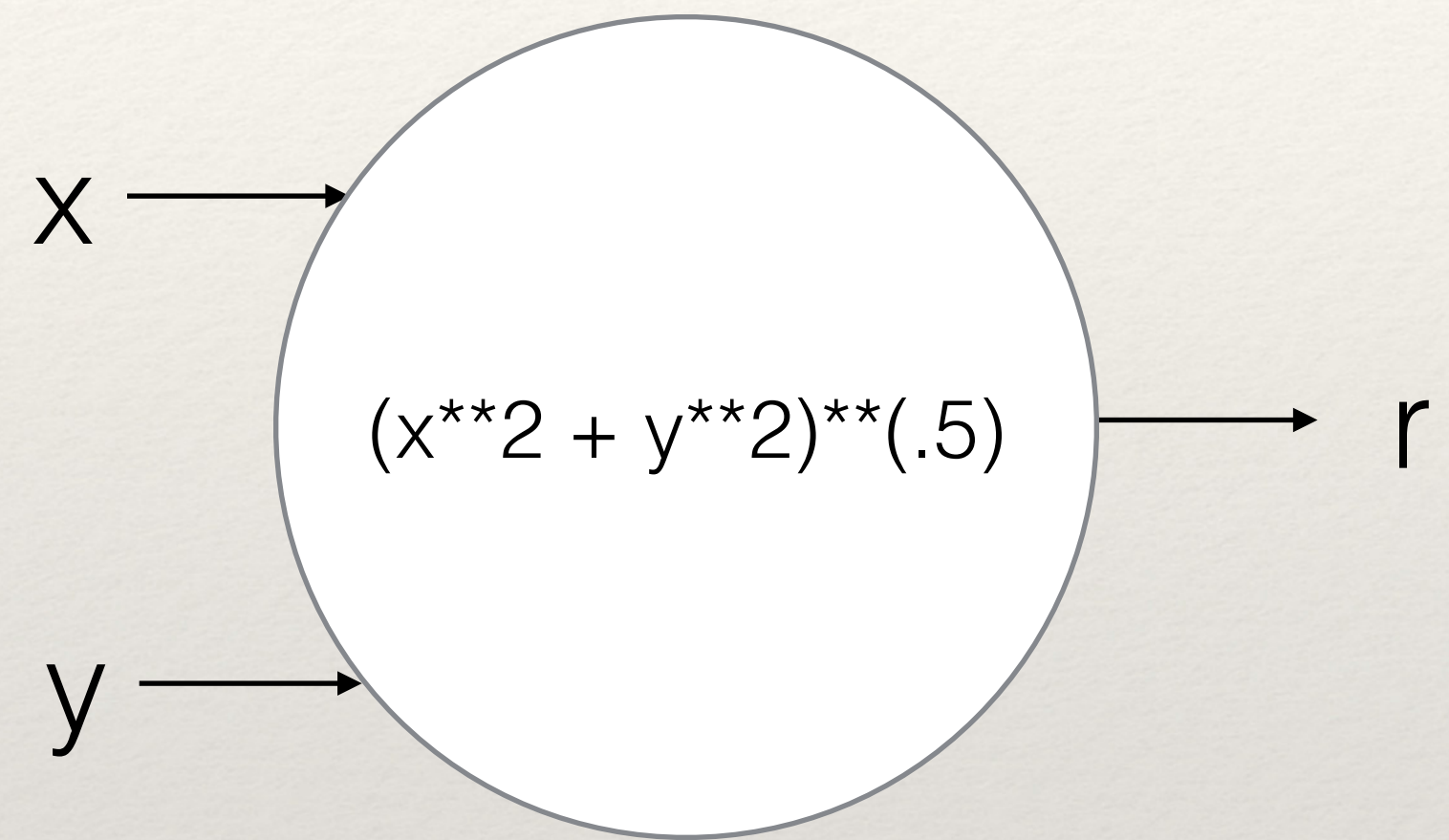










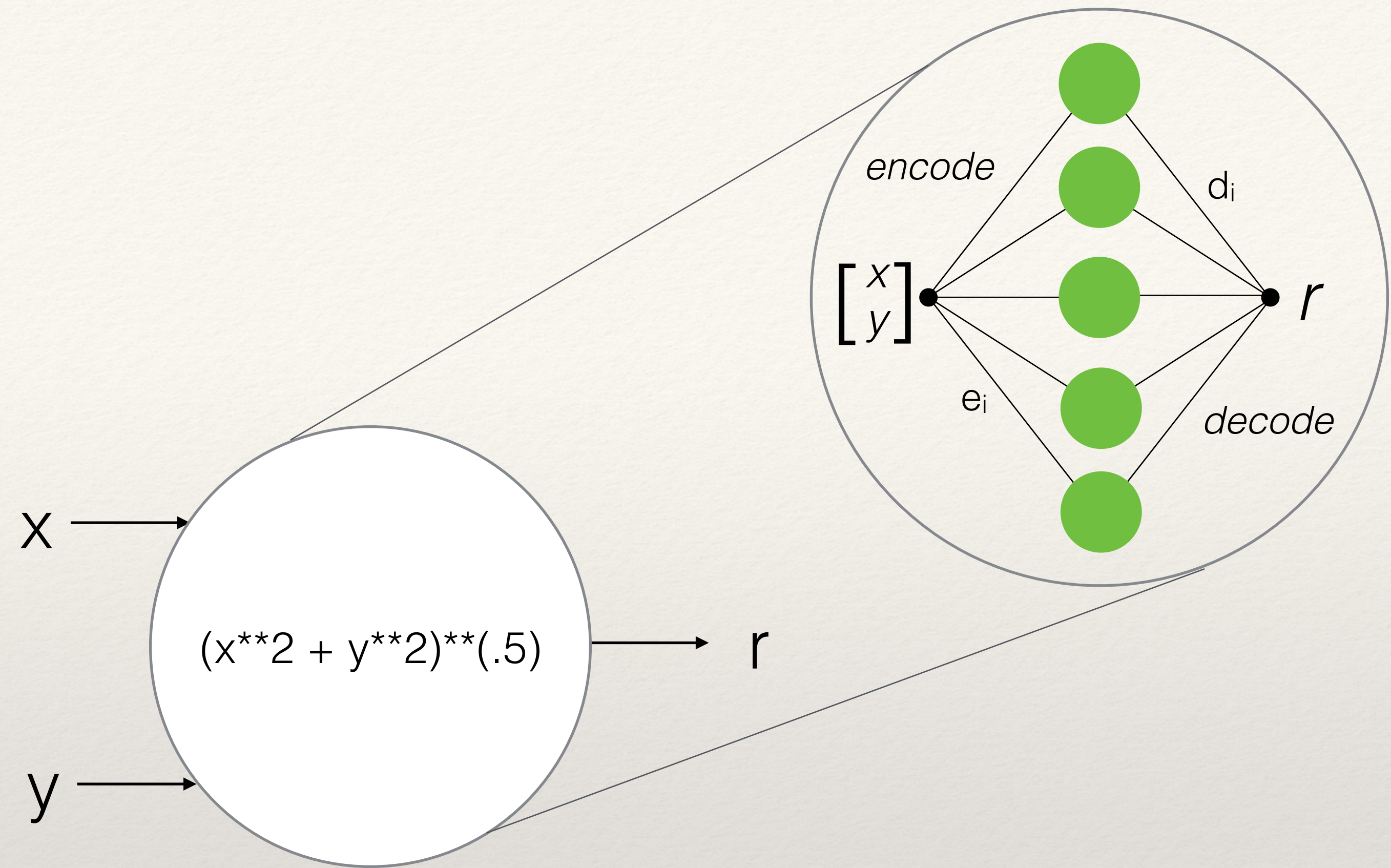
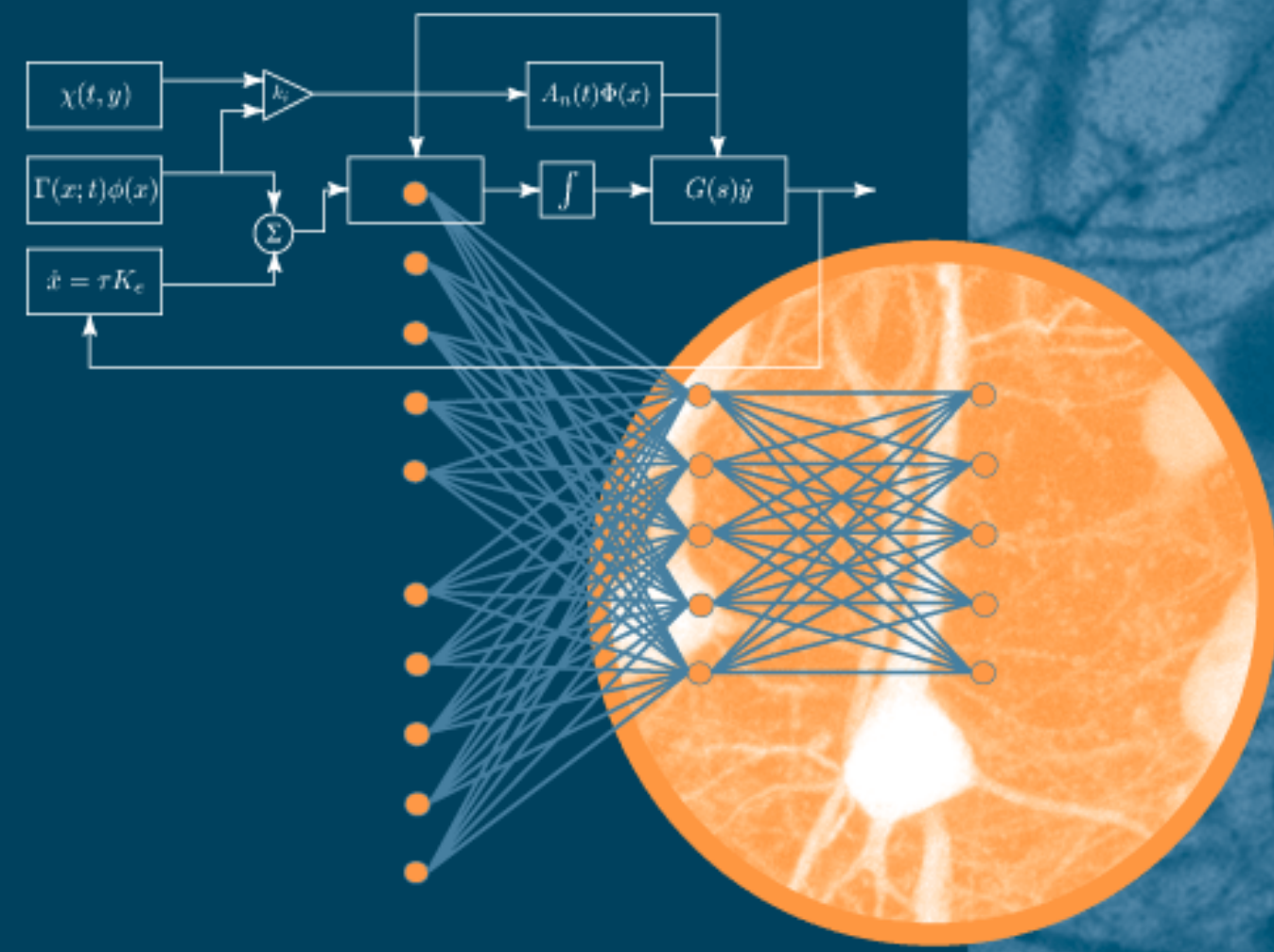




# Neural Engineering

COMPUTATION, REPRESENTATION, AND DYNAMICS  
IN NEUROBIOLOGICAL SYSTEMS

Chris Eliasmith and Charles H. Anderson

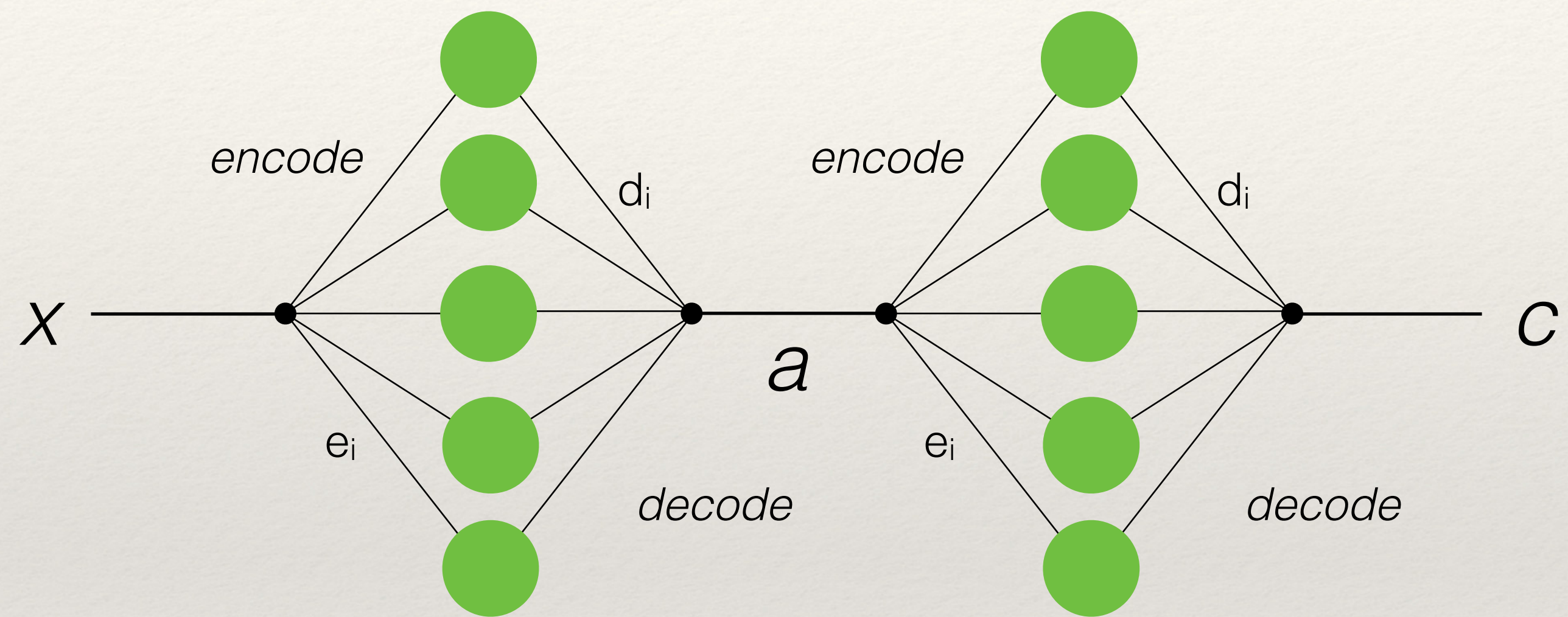




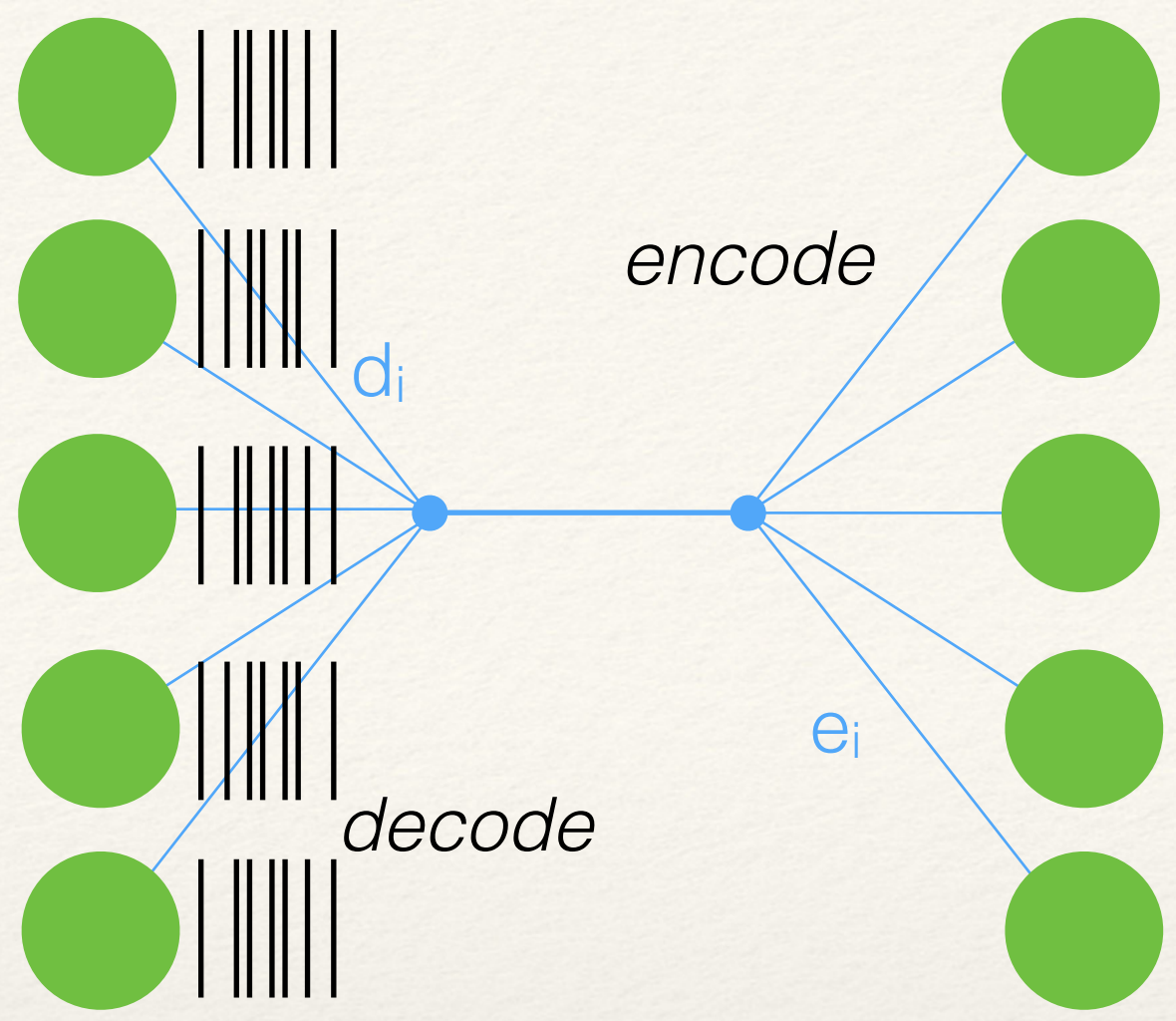
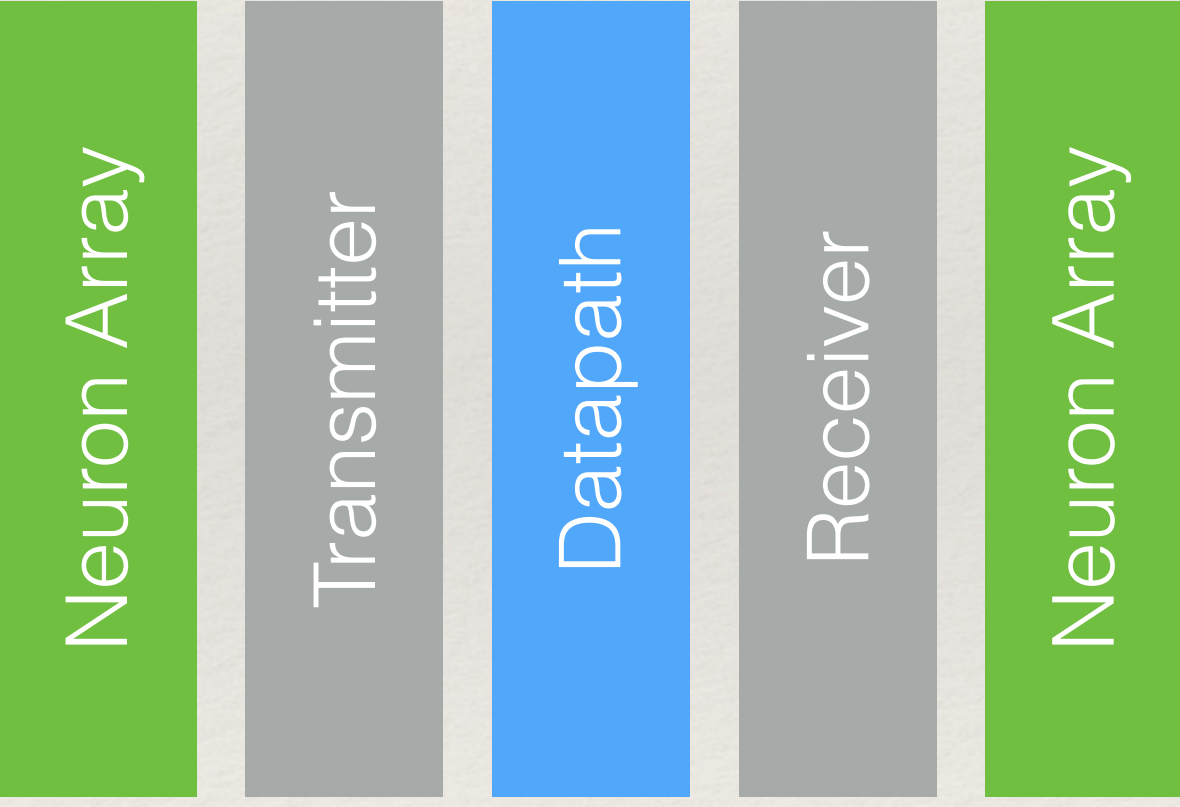
$$\dot{\mathbf{x}} = \alpha \mathbf{I} \mathbf{x}(t) + \mathbf{B} \mathbf{u}(t)$$

*Courtesy of  
Chris Eliasmith  
2013*



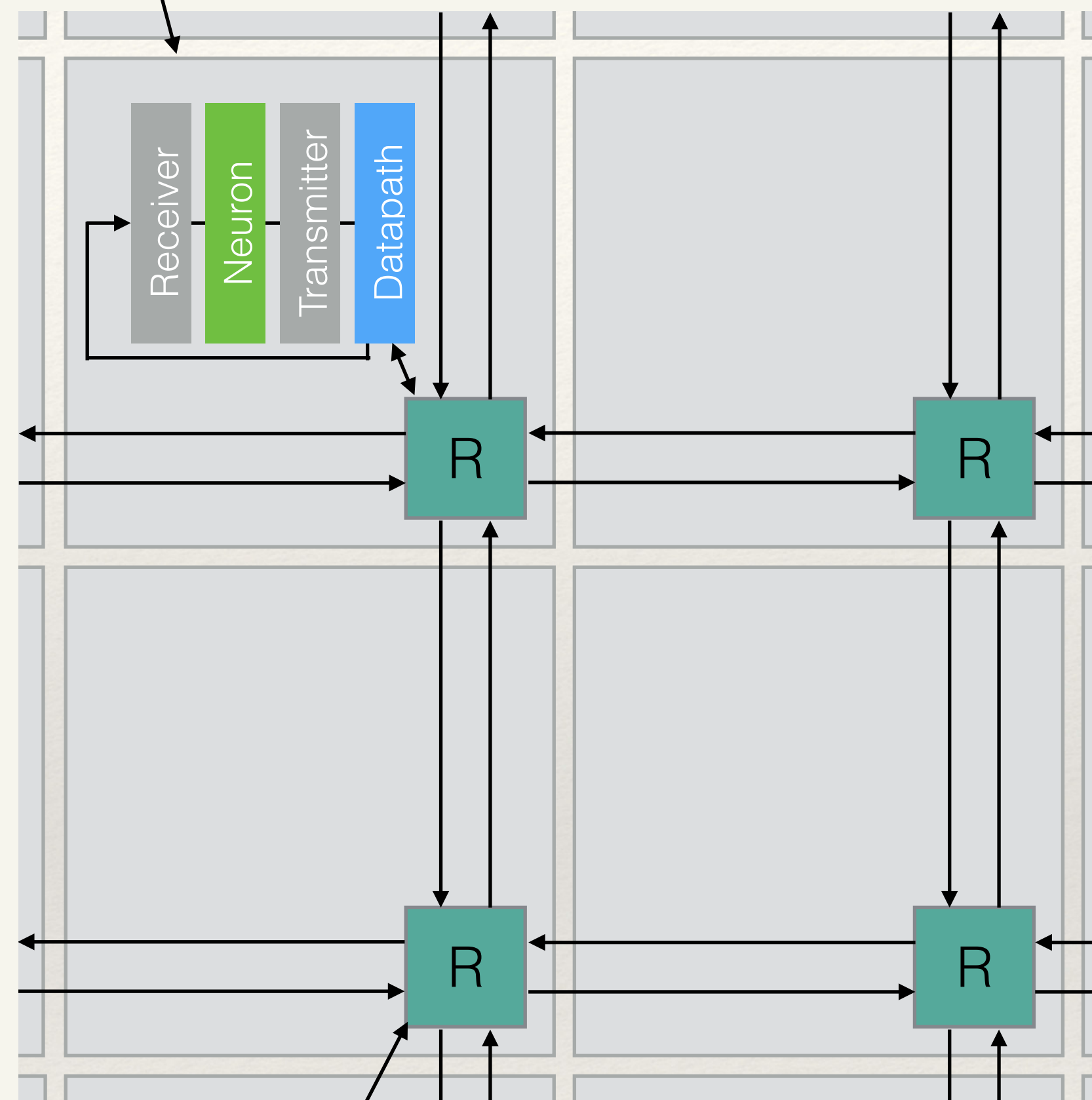








Core

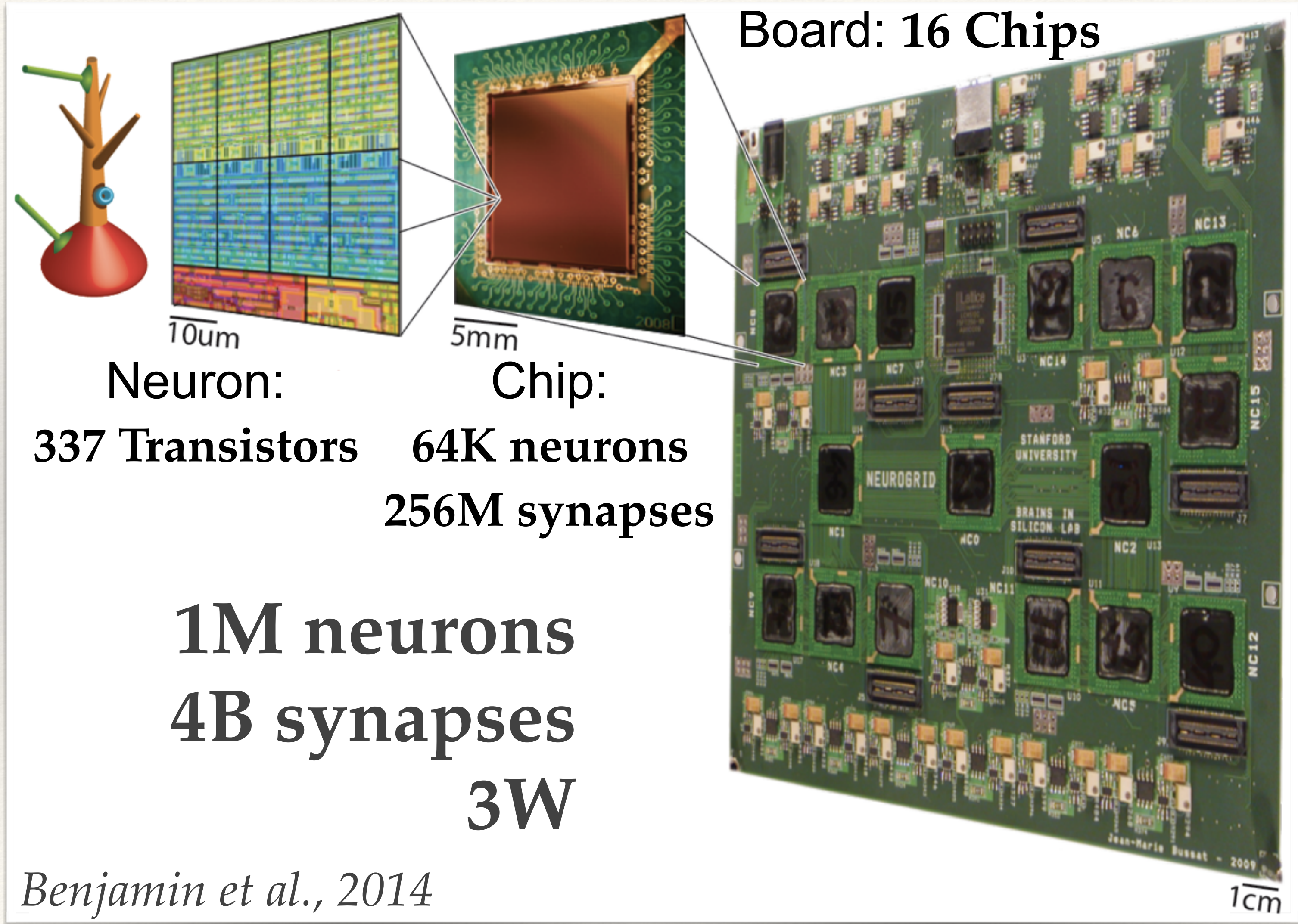


Router



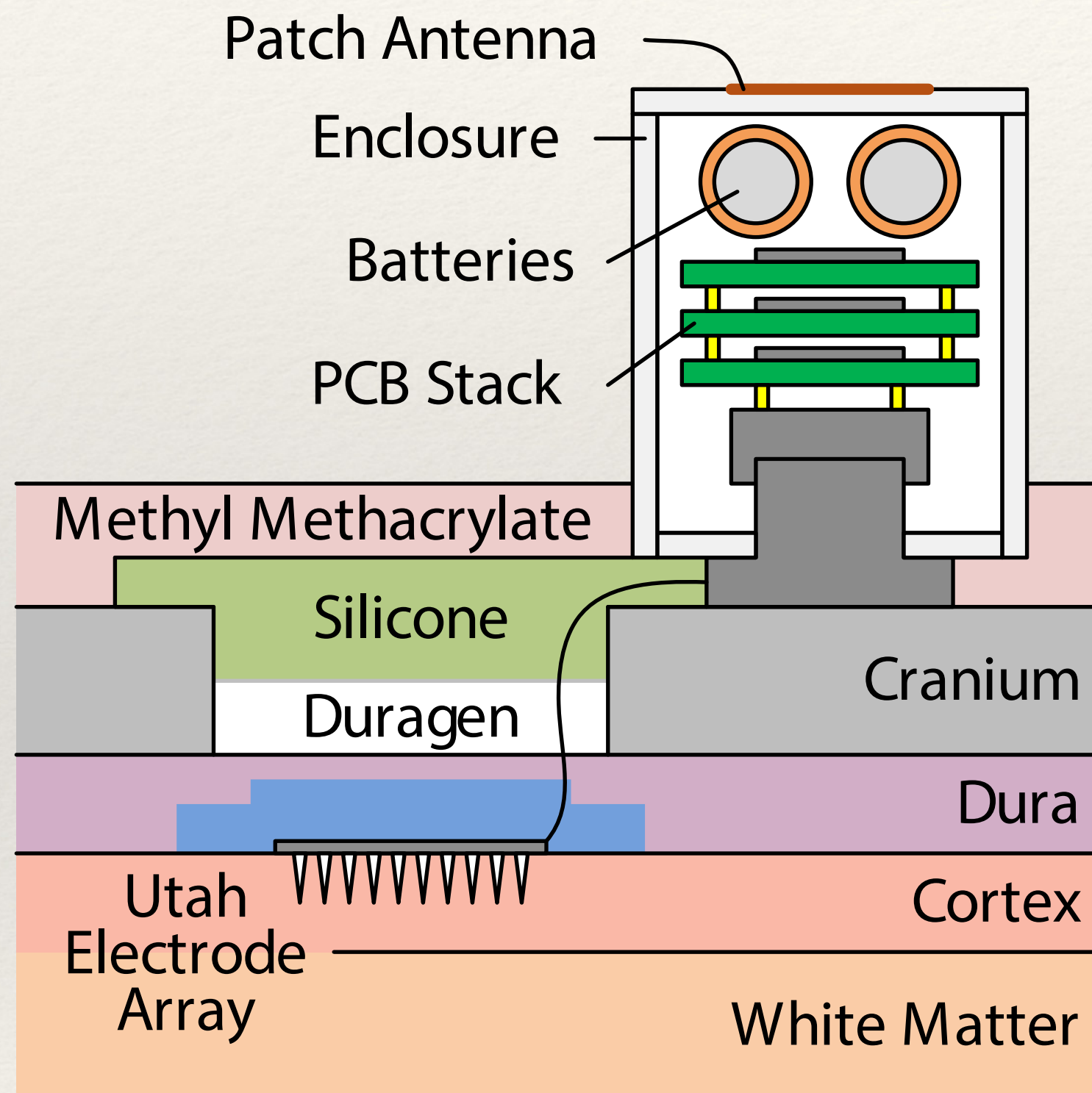
# Neurogrid

- ❖ Dendrites modeled with **subthreshold** analog circuits
- ❖ Axons modeled with **asynchronous** digital logic
- ❖ Connects each neuron to thousands of others with **clustered** synaptic connections
- ❖ Real-time operation
- ❖ 180 nm



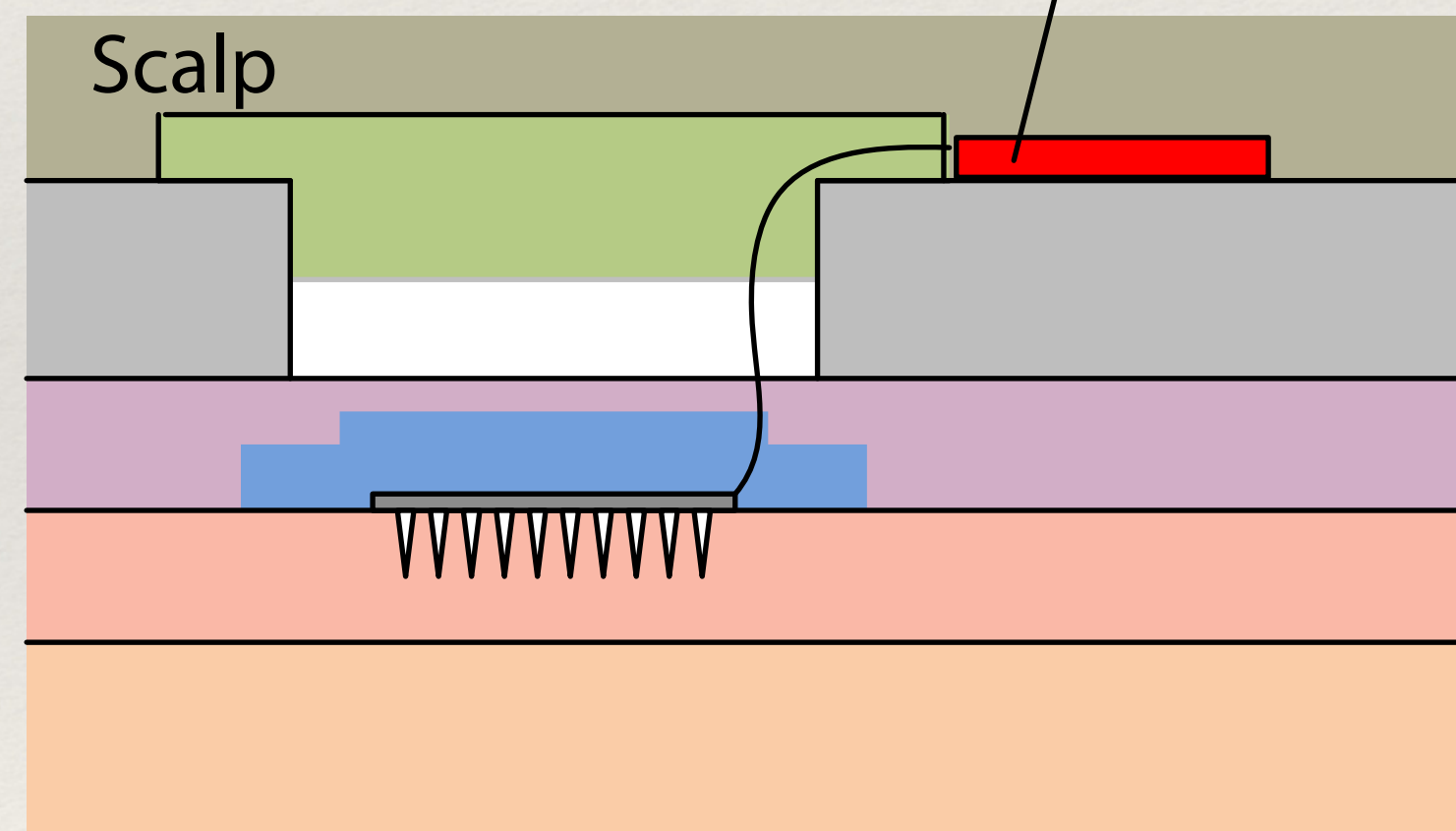


# Prosthesis power challenge



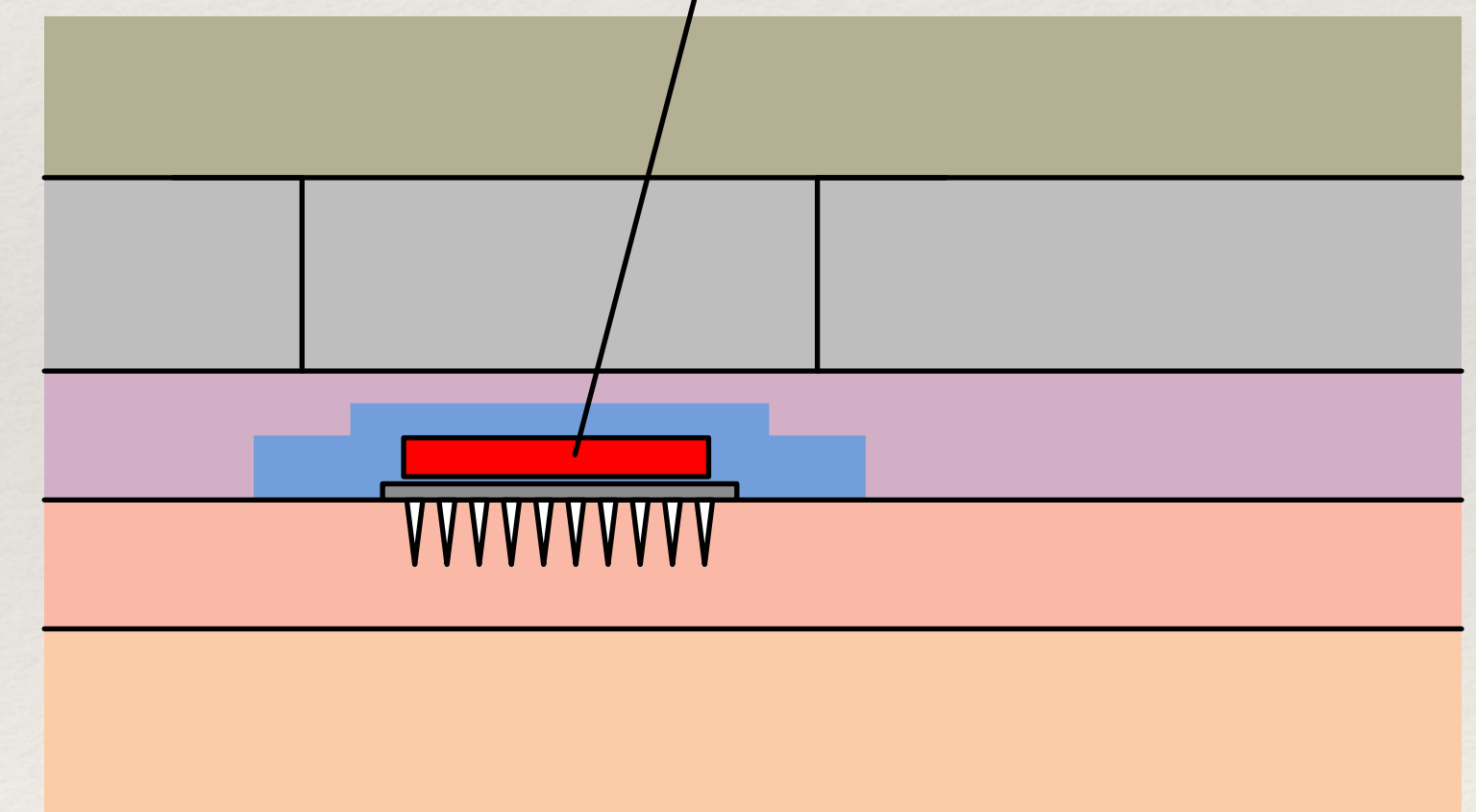
Head Mounted Electronics

IC with LNAs, ADCs, neuromorphic decoder circuit, telemetry, and inductive powering



Subcutaneous Mounted Electronics

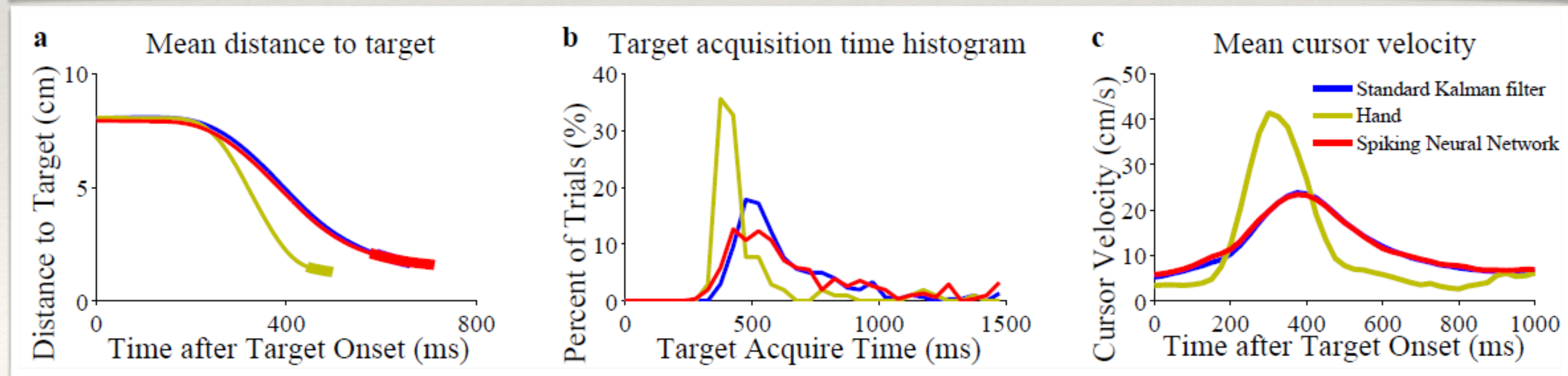
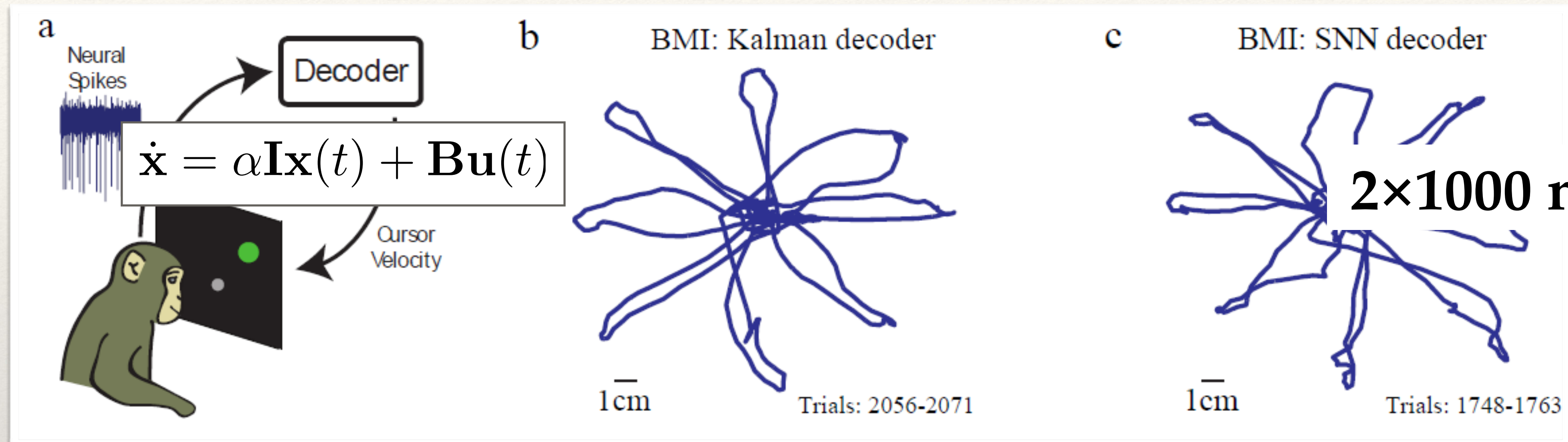
As shown to the left, but with IC mounted to backside of electrode array and fully-implanted beneath cranium



Fully-implanted Electronics (with array)



# Spiking neural network decoder



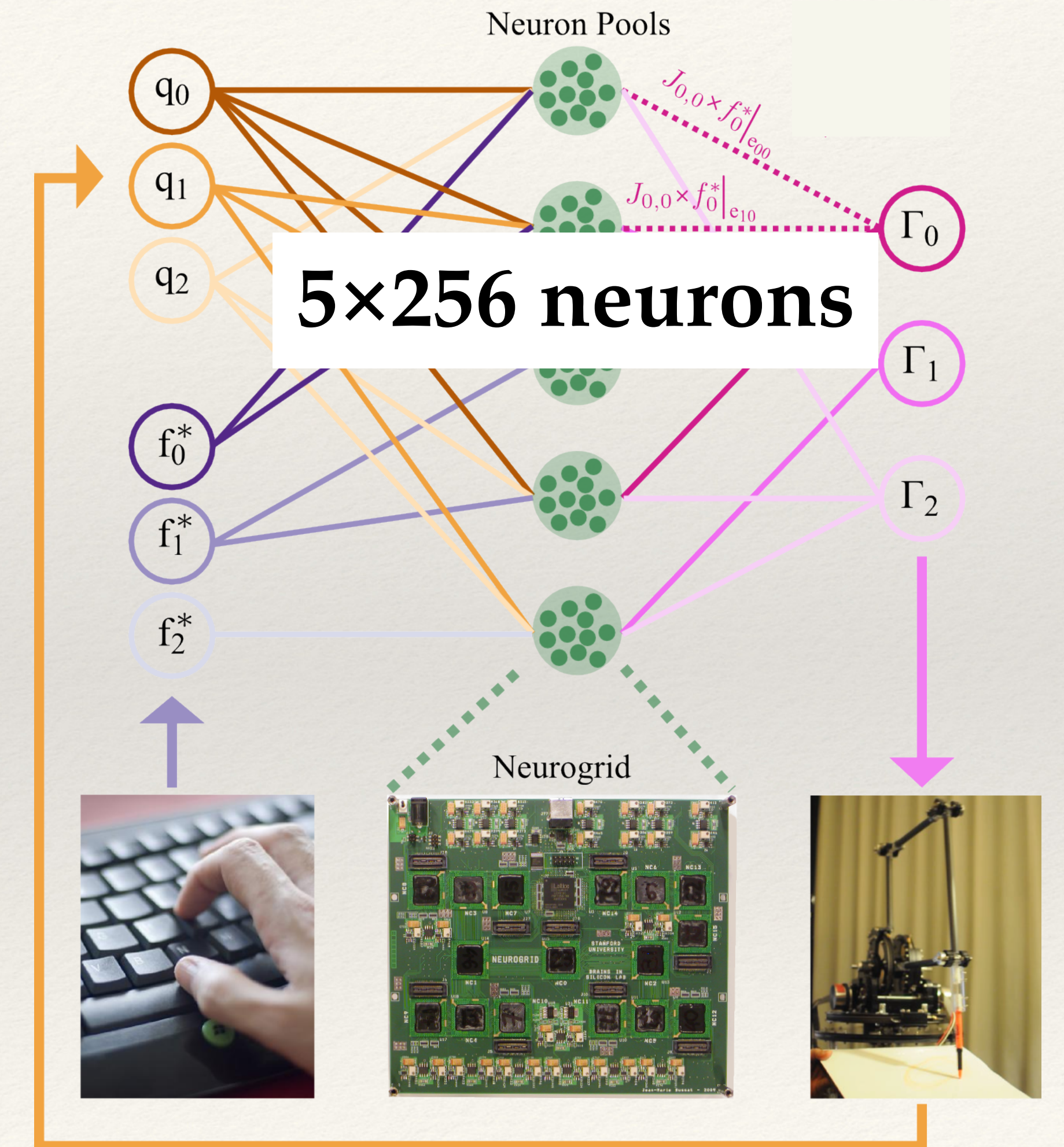


# Robot-arm controller

$$\Gamma = J_x^T f_x^* + \sum_{i=0}^4 m_i J_{com_i}^T g$$

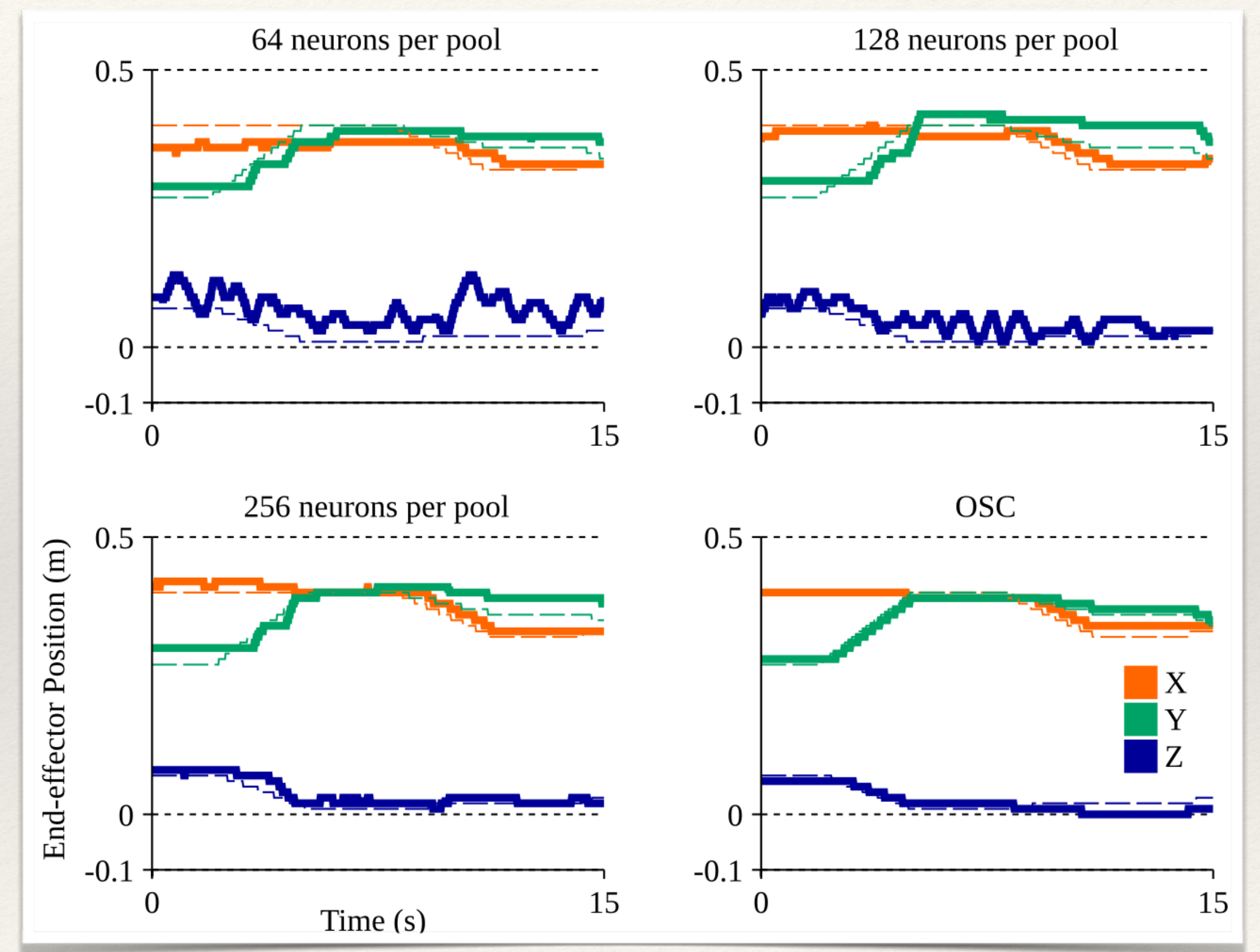
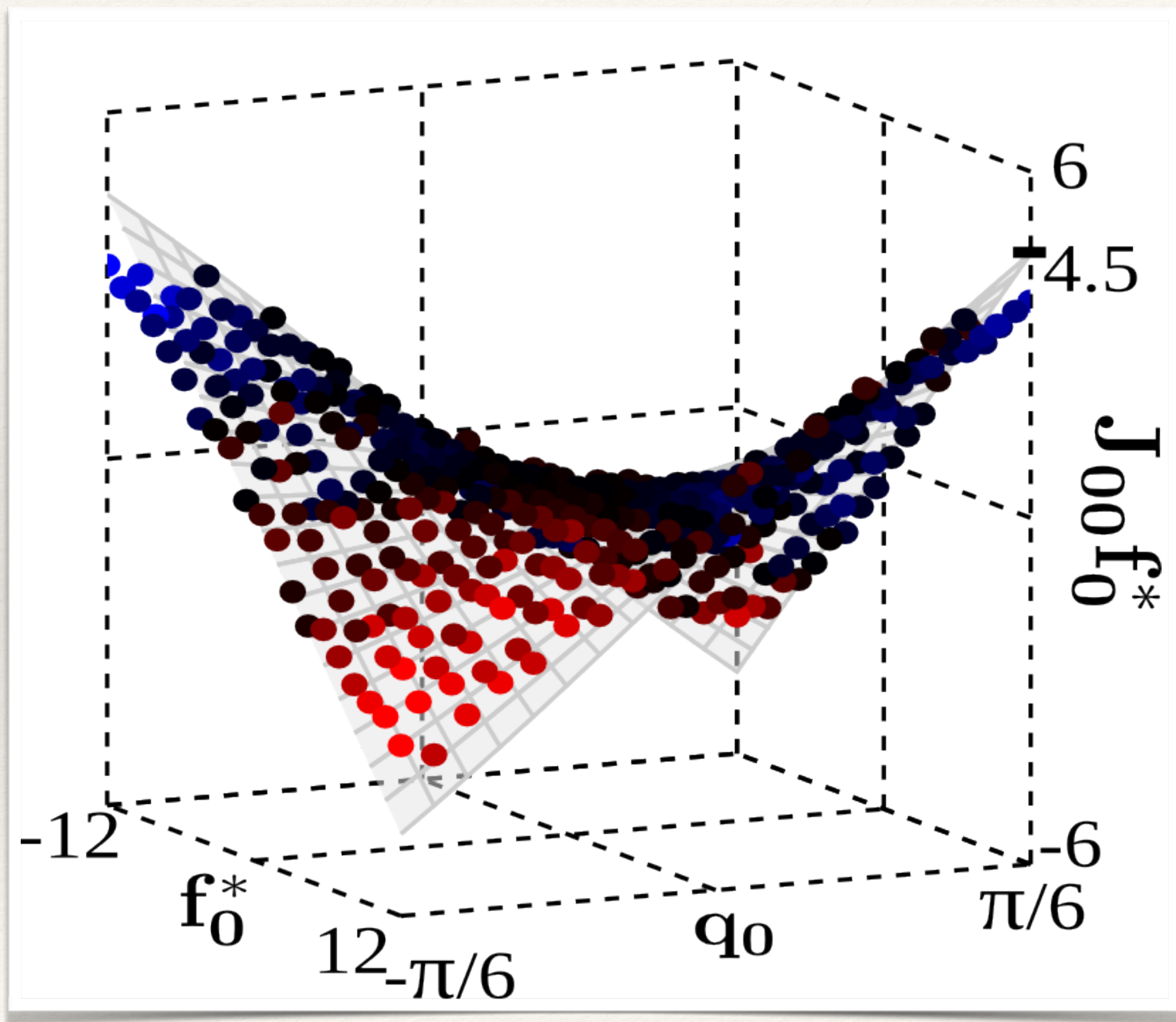
$$\begin{bmatrix} J_{0,0} & J_{0,1} & J_{0,2} \\ J_{1,0} & J_{1,1} & J_{1,2} \\ J_{2,0} & J_{2,1} & J_{2,2} \end{bmatrix}^T \begin{bmatrix} f_0^* \\ f_1^* \\ f_2^* \end{bmatrix} \rightarrow J_{0,0} \times f_0^*$$

$$J_{0,0} \times f_0^* \Big|_{e_{00}} = -0.35 \sin(q_0) \cos(q_2) f_0^*$$





# Robot-Arm controller performance



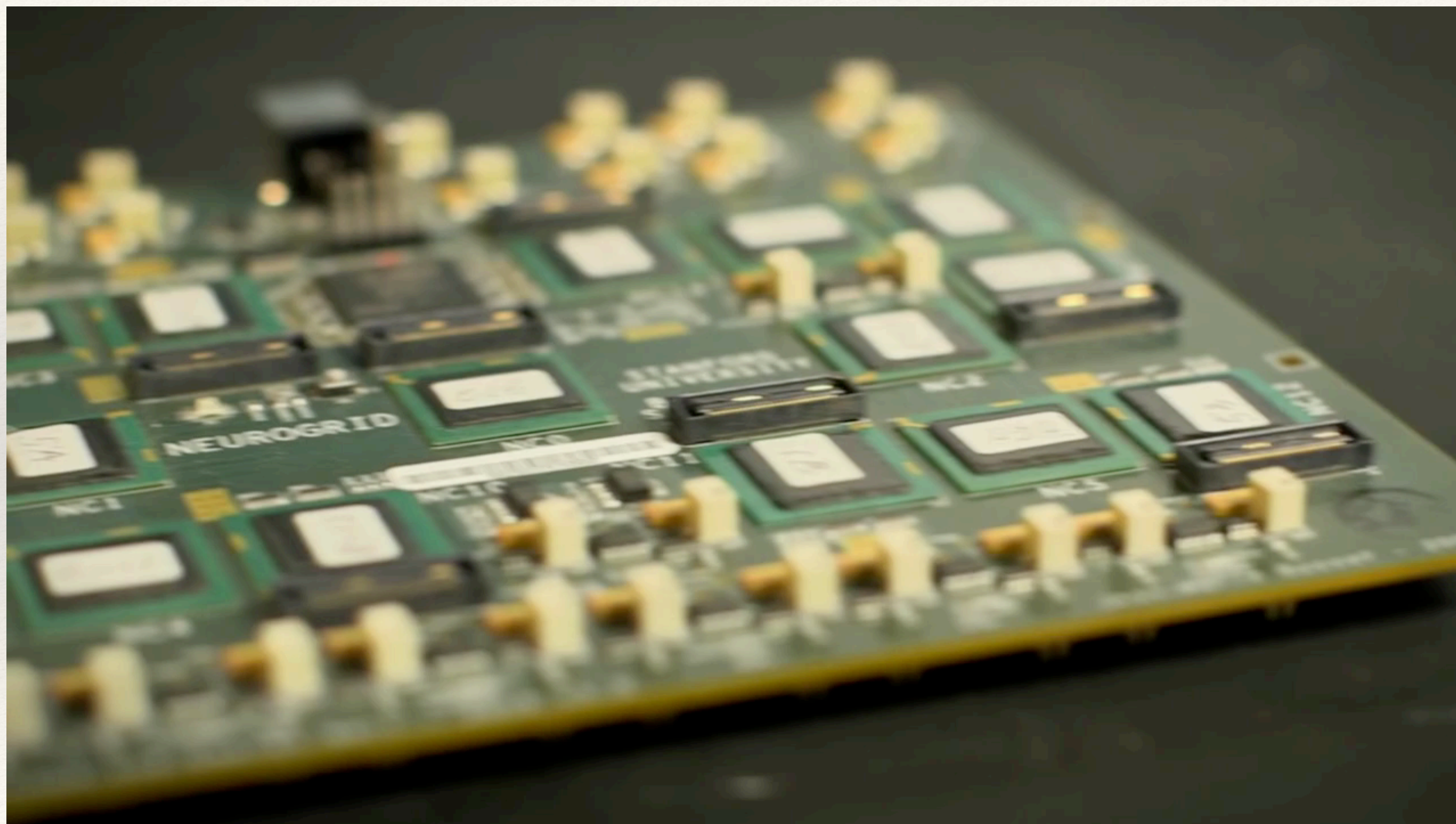
$$J_{0,0} \times f_0^* \Big|_{e_{00}} = -0.35 \sin(q_0) \cos(q_2) f_0^*$$



---

# Robot-arm controller video

---



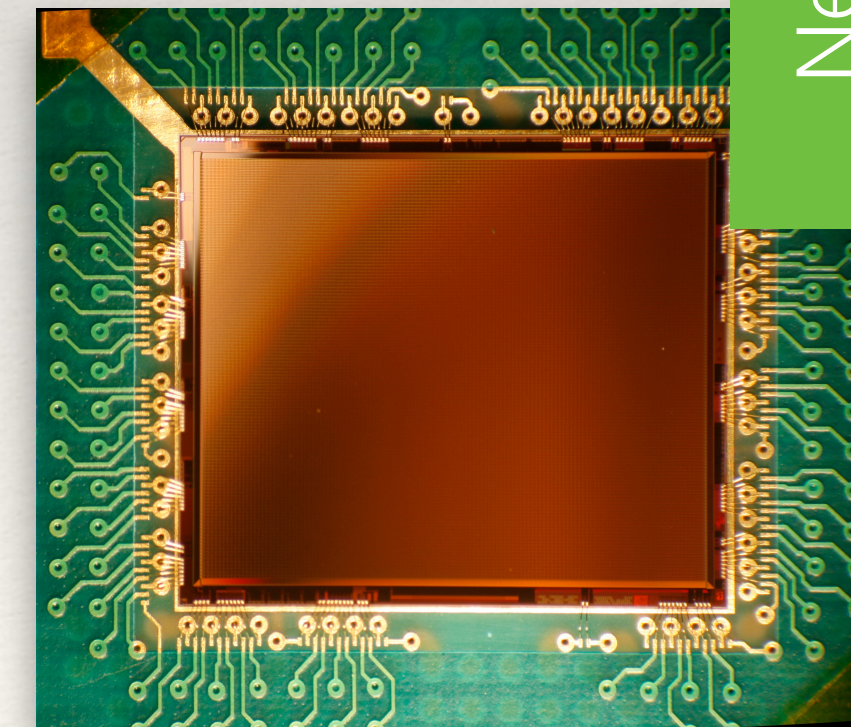


# Summary

- ❖ Combining **analog computation** with **digital communication** proved energy-efficient and noise-robust
- ❖ Building the first neuromorphic chip (*Brainstorm*) that **implements spiking neural networks** specified at a **higher level of abstraction**
- ❖ Writing software tool (*Neuromorph*) that **automatically synthesizes network** from high-level specification
- ❖ **ONR-funded** collaboration with colleagues at **Cornell** and **Waterloo**



Neuron Array



Transmitter

Datapath

Receiver

Neuron Array