

THE PSYCHOLOGY OF ARTIFICIAL INTELLIGENCE

TONY PRESCOTT

COMPUTER SCIENCE
UNIVERSITY OF SHEFFIELD

@tonyjprescott.bsky.social



iCub by IIT

British Computer Society, Edinburgh, 2nd September 2026.



The
University
Of
Sheffield.



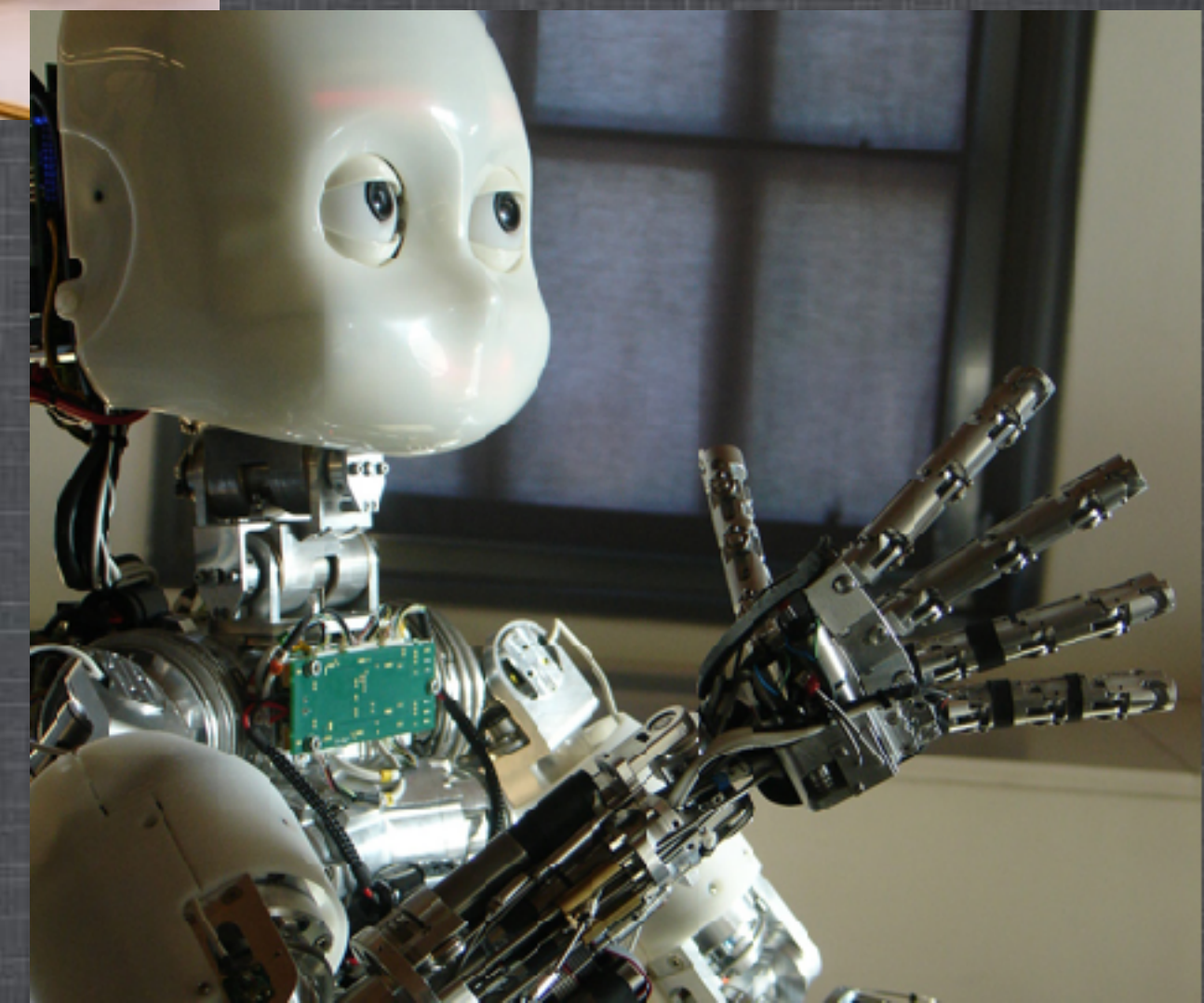
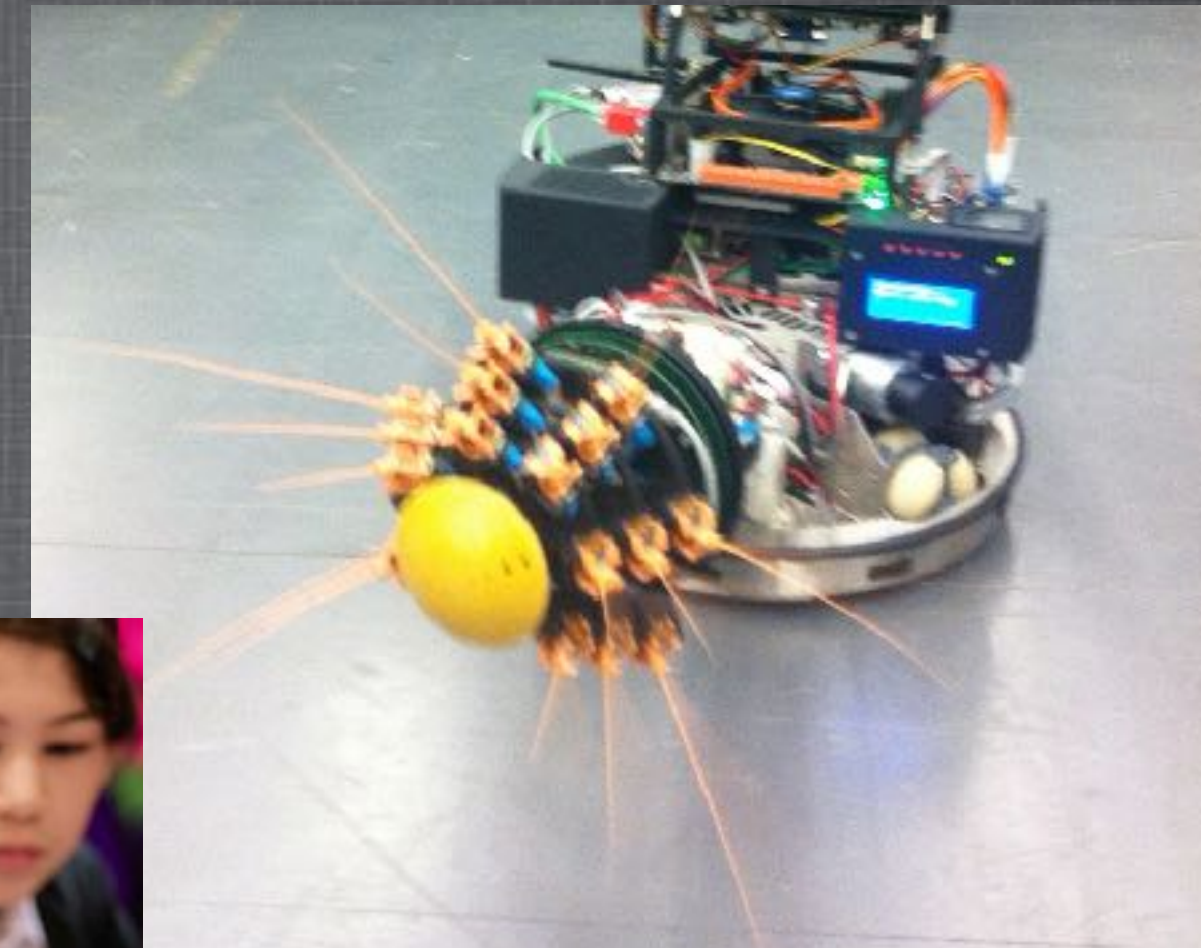
Centre for Machine Intelligence

THE
ROYAL
SOCIETY



MY RESEARCH

- Biomimetic brain-inspired robotics
- Cognitive architecture
- Active touch sensing
- Memory and sense of self
- Human-robot interaction
including how people see robots
- Robot applications
in education and therapy
- Ethics and responsible innovation

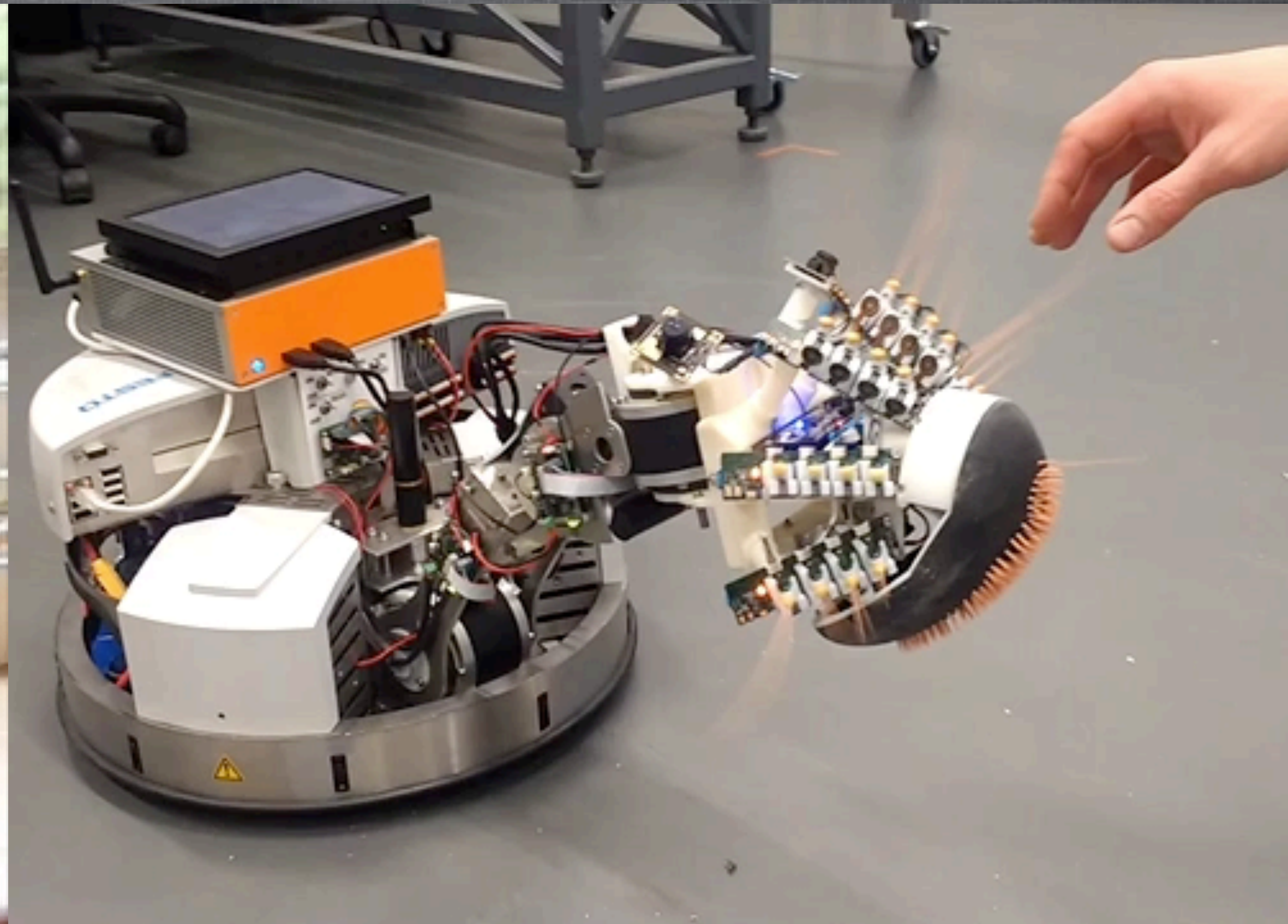


USING BRAIN-INSPIRED BIOMIMETICS TO CREATE SYNTHETIC ACTIVE TOUCH FOR ROBOTS



1/7th real
time

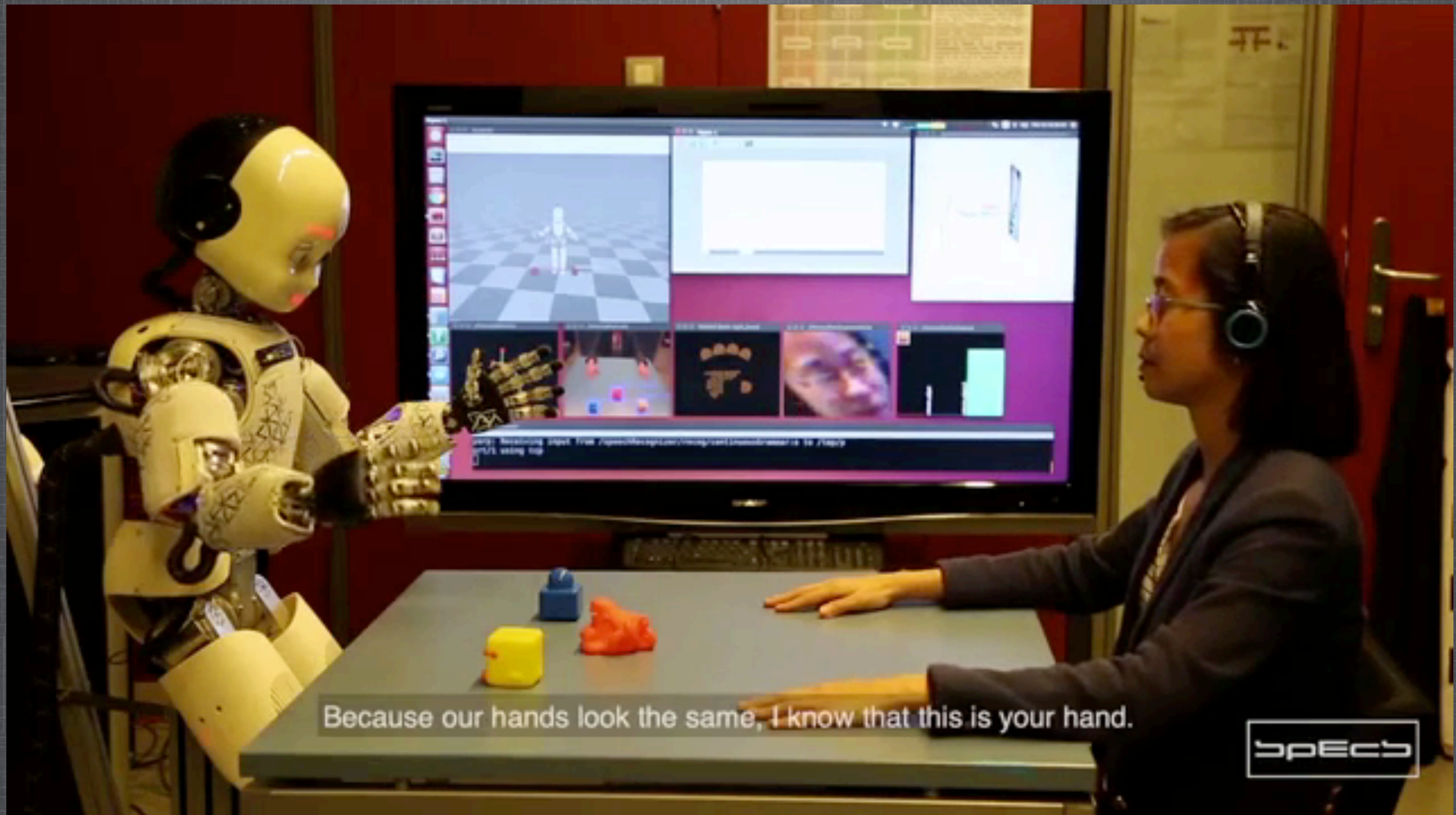
Video by Tony Prescott



Video by Martin Pearson

Developed with Bristol Robotics Lab

AN EMBODIED MODEL OF THE SELF

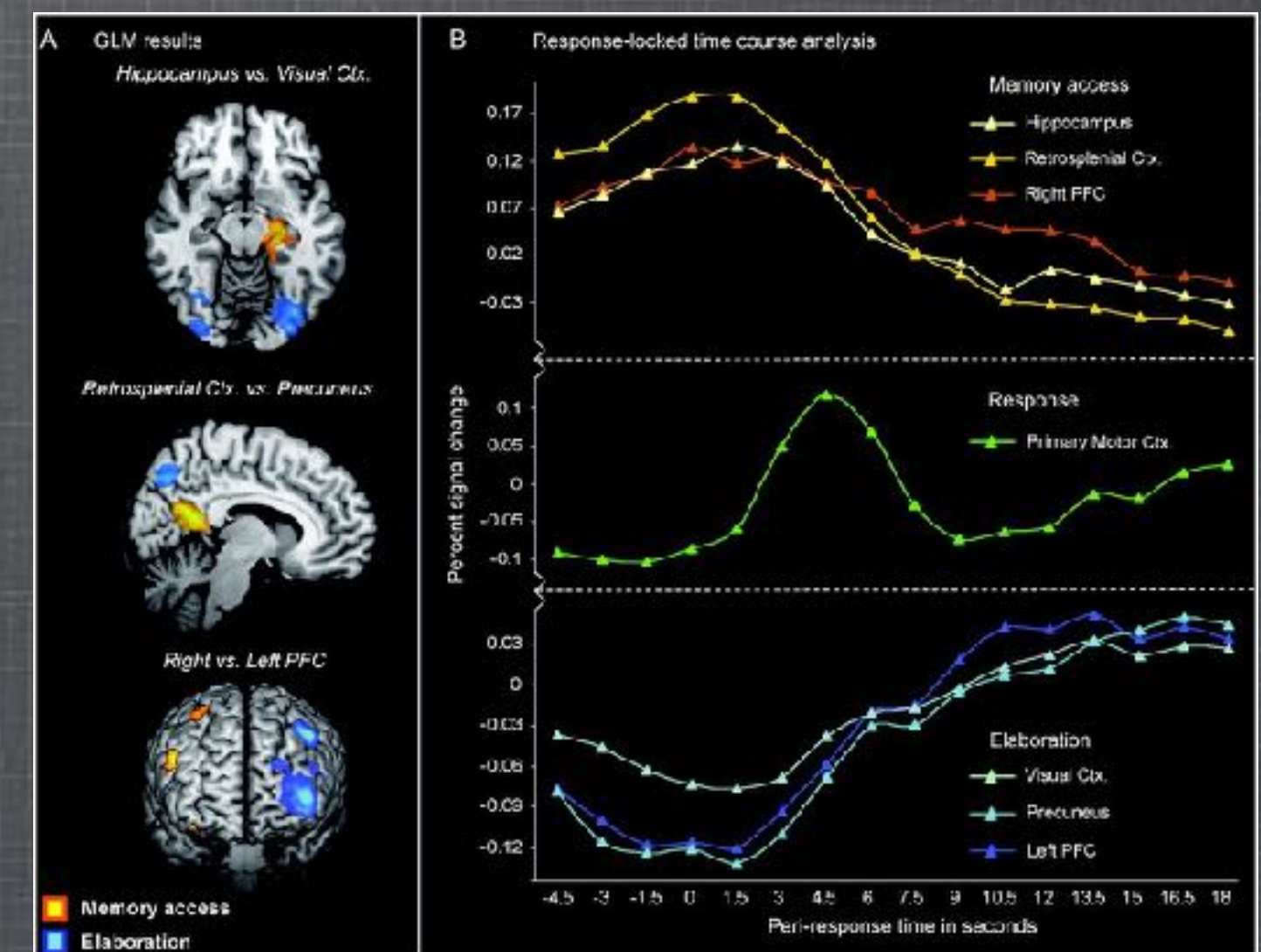


The robot has a model of itself and is able to learn about the relationship between its body parts and sensations it receives, and how its own body relates to another. These are some of the building blocks for a sense of self.

Moulin-Frier et al. (2017) *IEEE Transactions in Cognitive and Developmental Systems*.

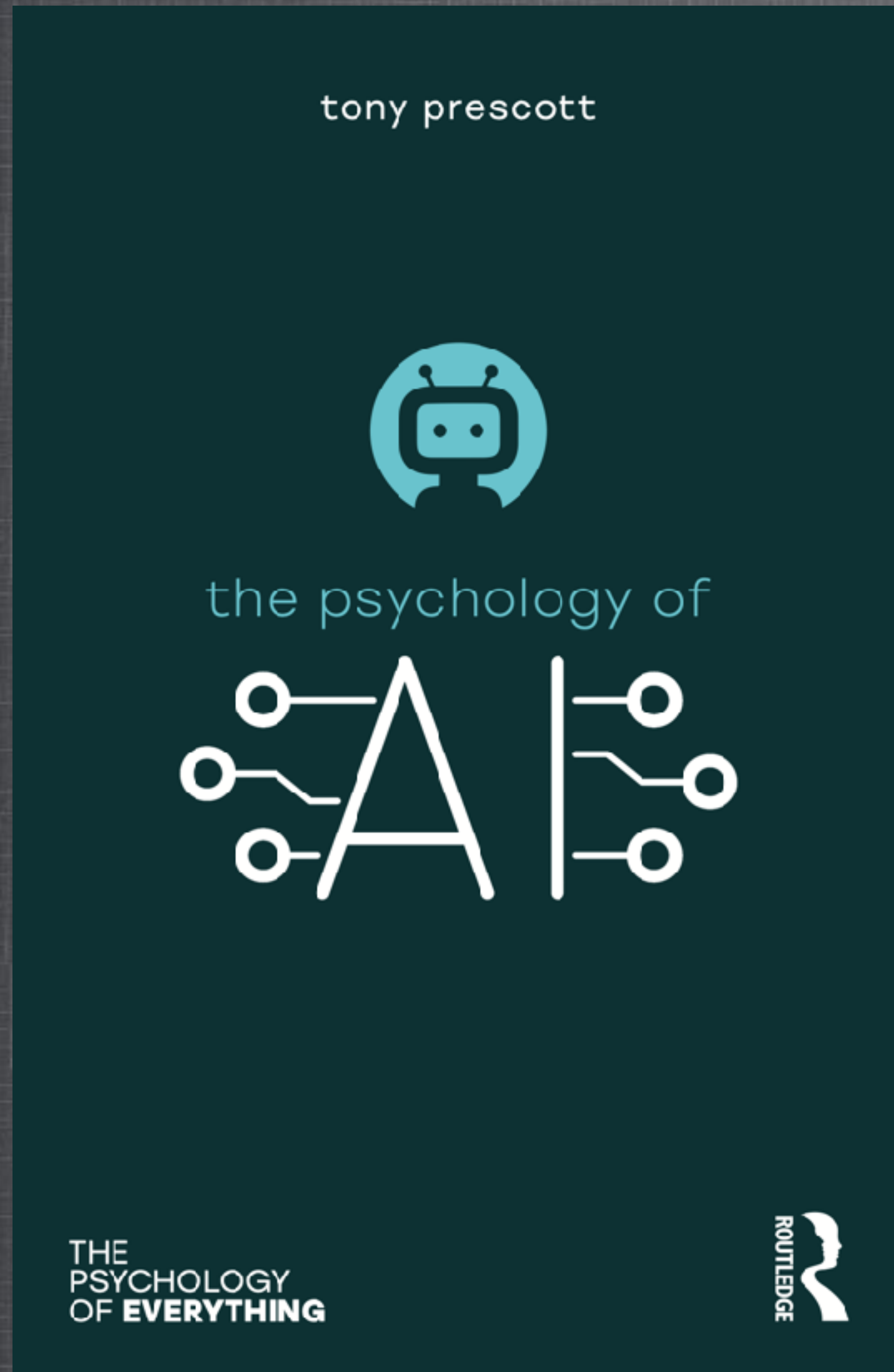
GROUNDING MEMORIES IN SENSORIMOTOR EXPERIENCE

Use generative capabilities of GP models to “recreate” the experience



D. Camilleri, A. Damianou, H. Jackson, N. Lawrence, T. Prescott.
“iCub Visual Memory Inspector: Visualising the iCub's Thoughts.”
5th International Conference on Biomimetic and Biohybrid Systems (Living Machines).

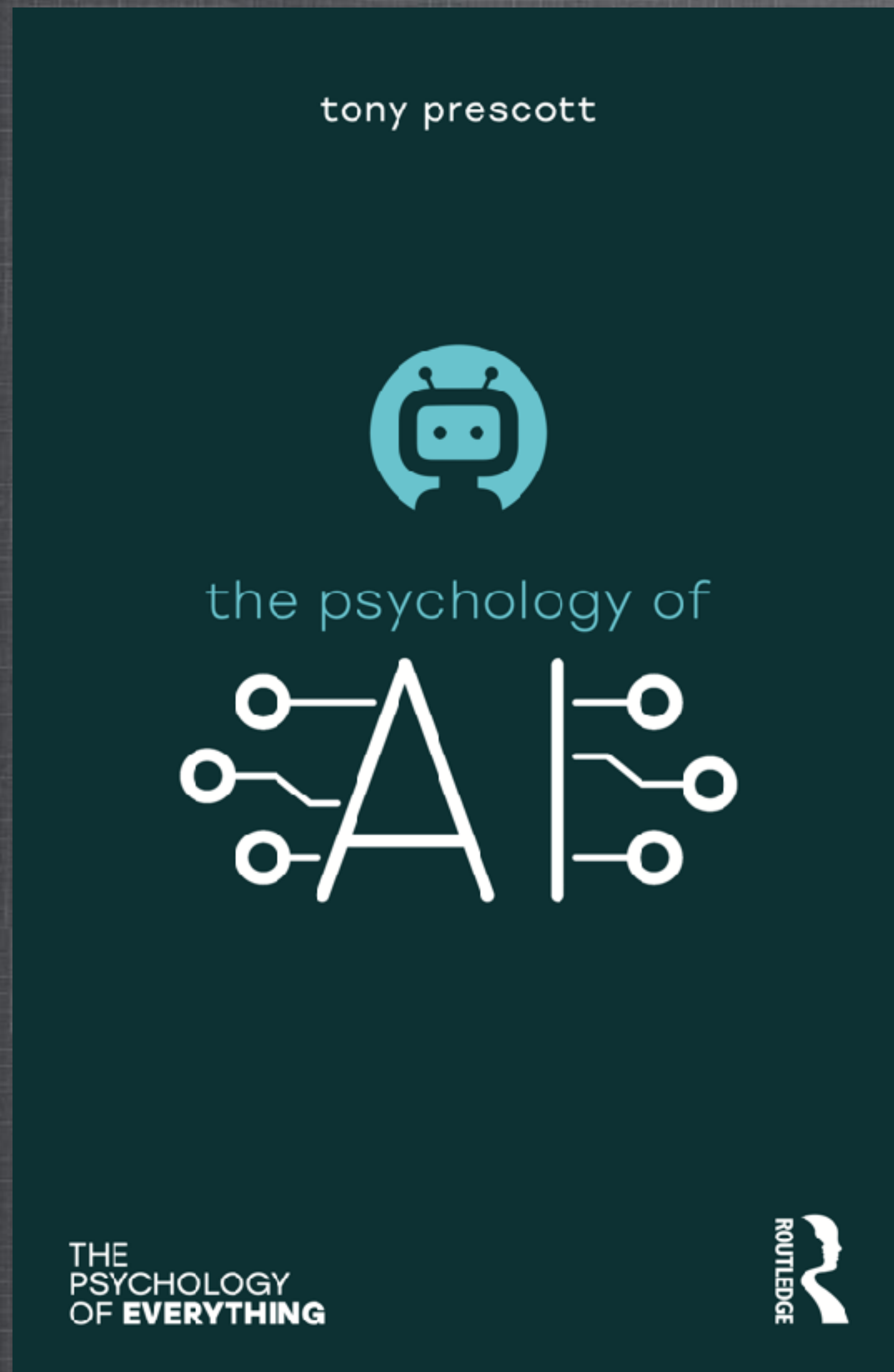
THE PSYCHOLOGY OF AI



- What is intelligence?
- Brains and computers
- The building blocks of intelligence
- Learning in neural networks
- Towards Artificial General Intelligence (AI)
- Living with Artificial Intelligence

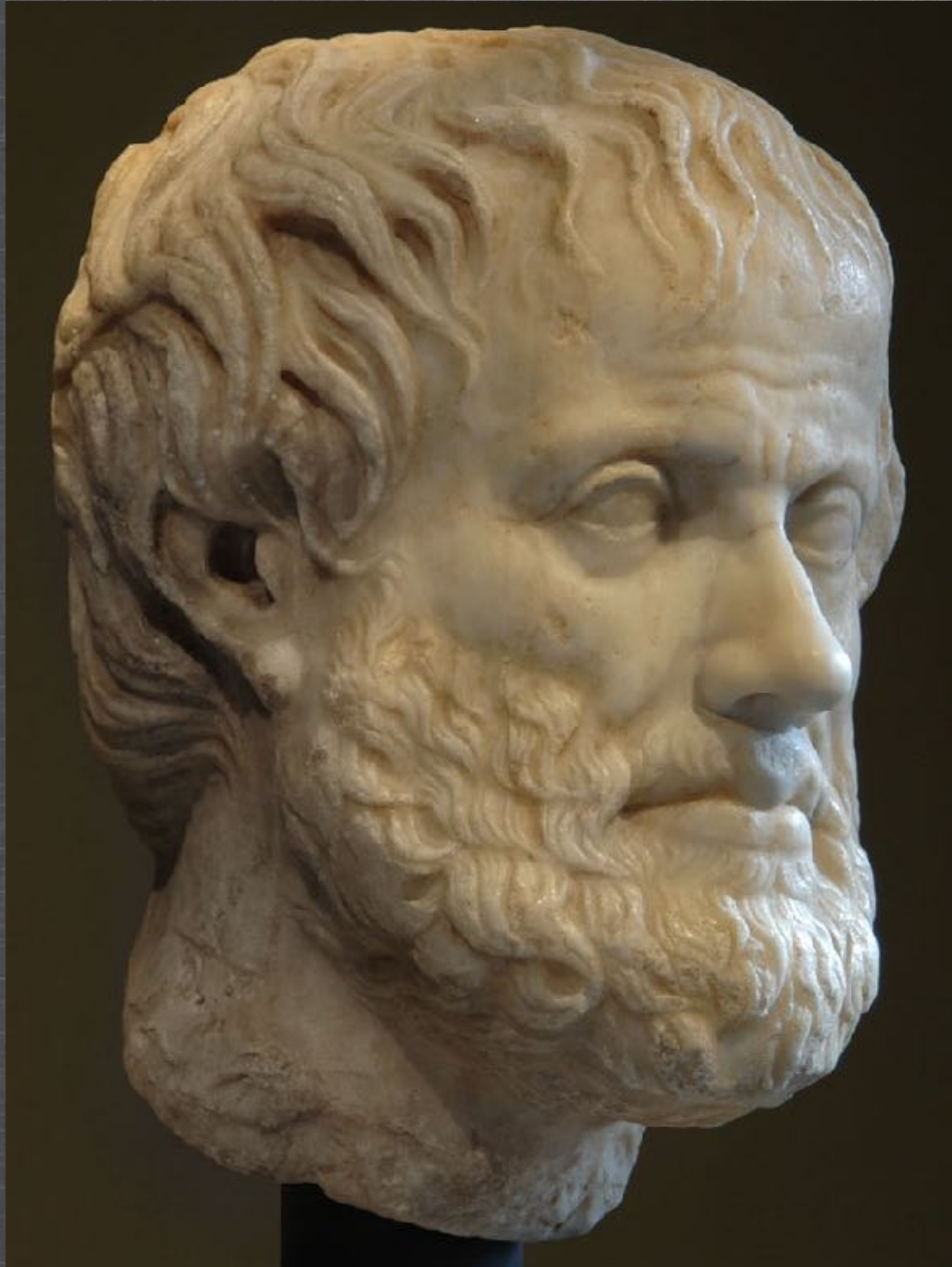
AVAILABLE NOW 11.39 Euros

THE PSYCHOLOGY OF ARTIFICIAL INTELLIGENCE



- *Overview*
- What is intelligence?
- Are brains computers?
- Do AIs understand anything?

ARISTOTLE, 4TH C. BCE



***Aesthesis* = perception** (seeing, hearing, touching, tasting and so on)

is different from

***Nuos* = intelligence/intellect/mind**

All animals having *aesthesis* (at least a sense of touch). In contrast, only humans have *nuos*

Aristotle thought that the **heart was the seat of intelligence**, the role of the brain was to cool the blood

INTELLIGENCE: ONE THING OR MANY?

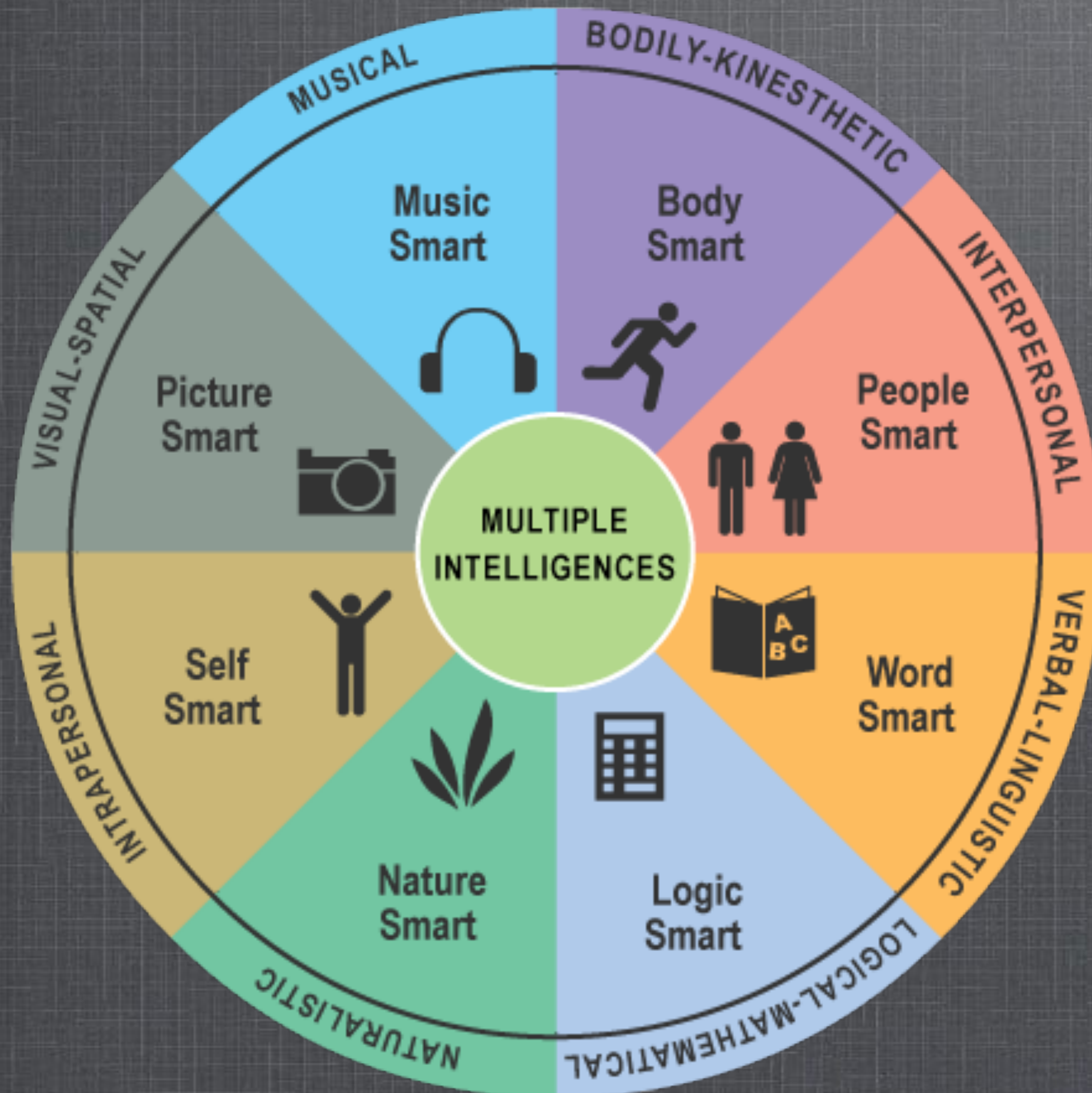
From the Latin *intelligentia* which has two roots, *inter* meaning between, and *legere*—to pick out or read.

Some psychologists think intelligence is a single underlying capacity, *g*



INTELLIGENCE: ONE THING OR MANY?

others think that there are multiple intelligences



- Analytical, creative, practical (Sternberg)
- Fast and slow thinking (Kahneman)
- Emotional intelligence (e.g. Damasio)
- Social intelligence (e.g. Thorndike)

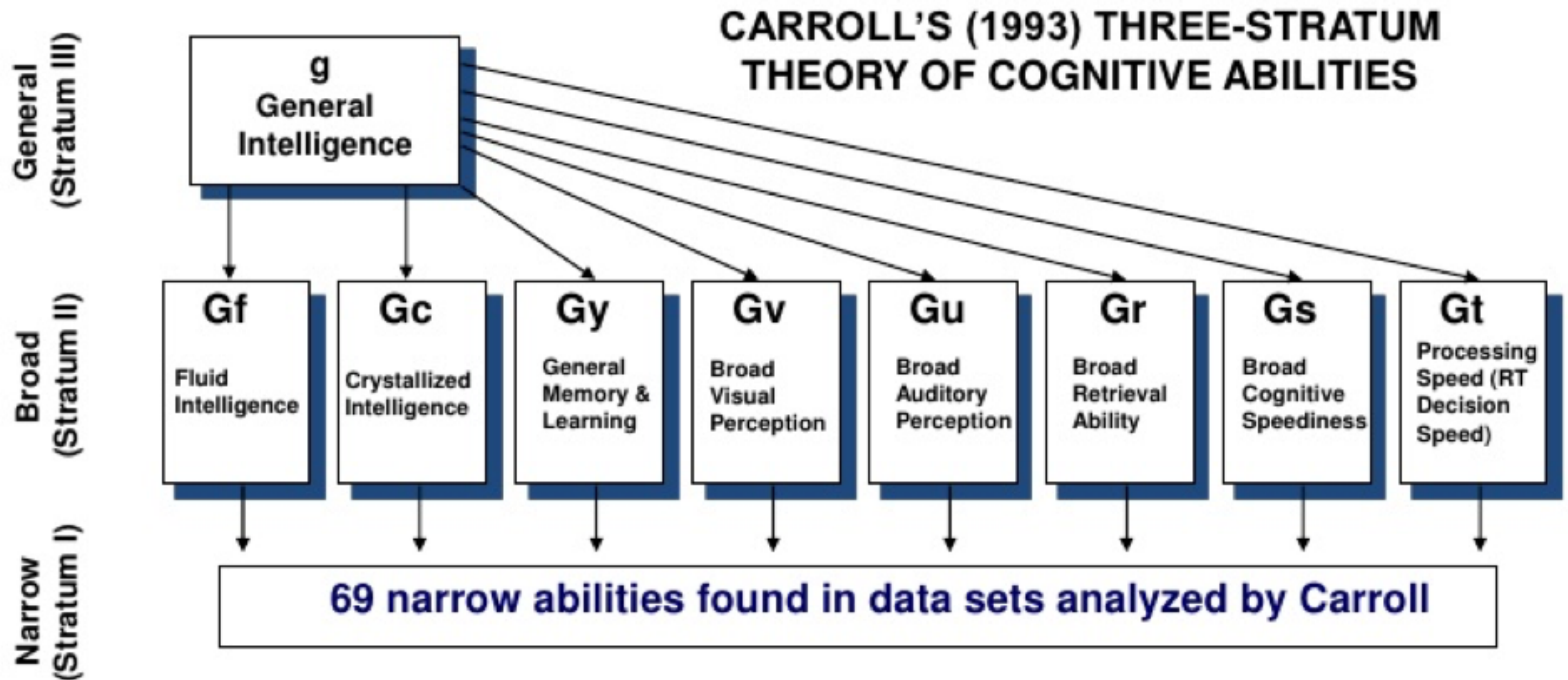
Gardner, H. (2006). Multiple Intelligences: New Horizons. New York: Basic Books.

EXAMPLE: BODILY-KINESTHETIC INTELLIGENCE



Christiano Ronaldo's brain can compute the flight of a ball so accurately he can score from a pass in complete darkness

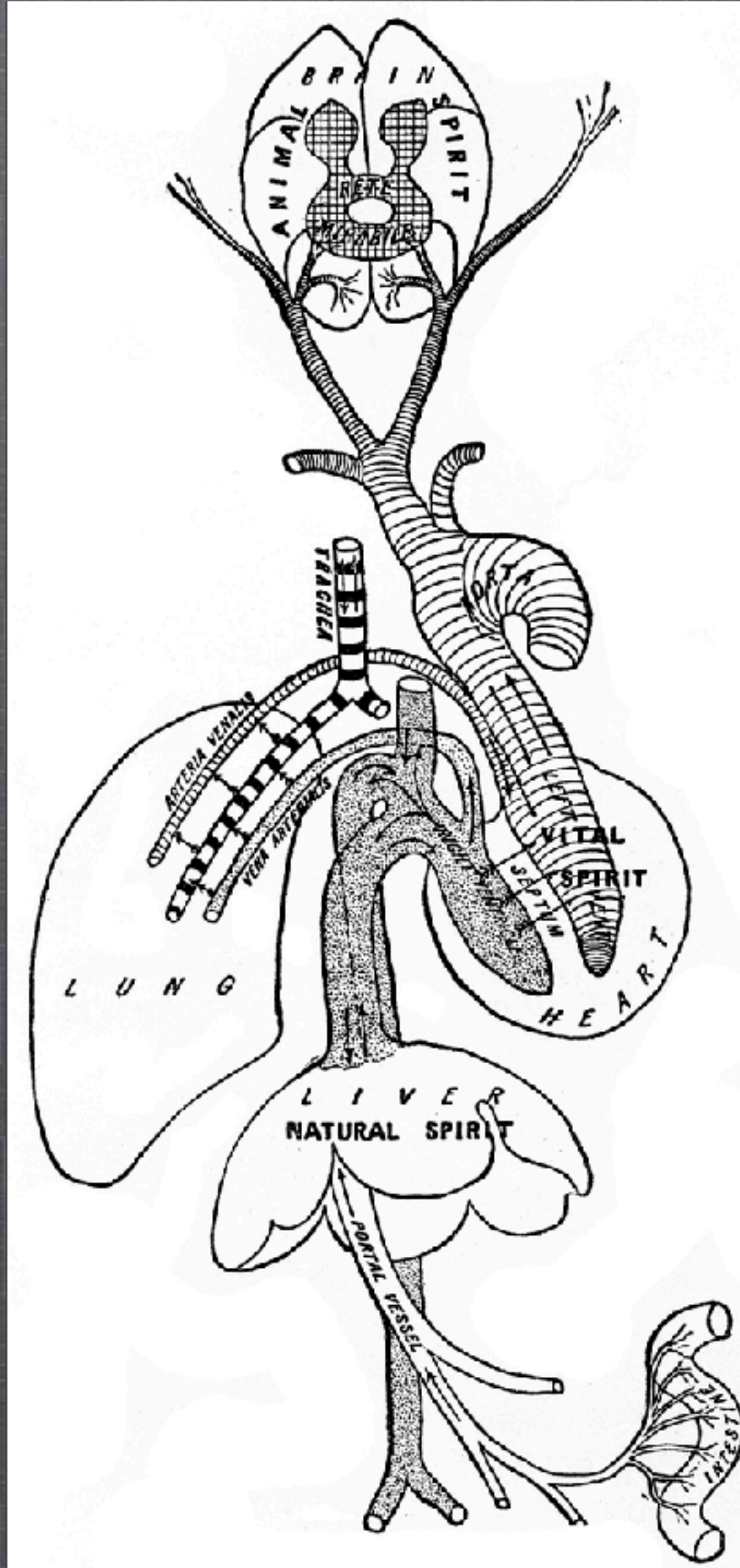
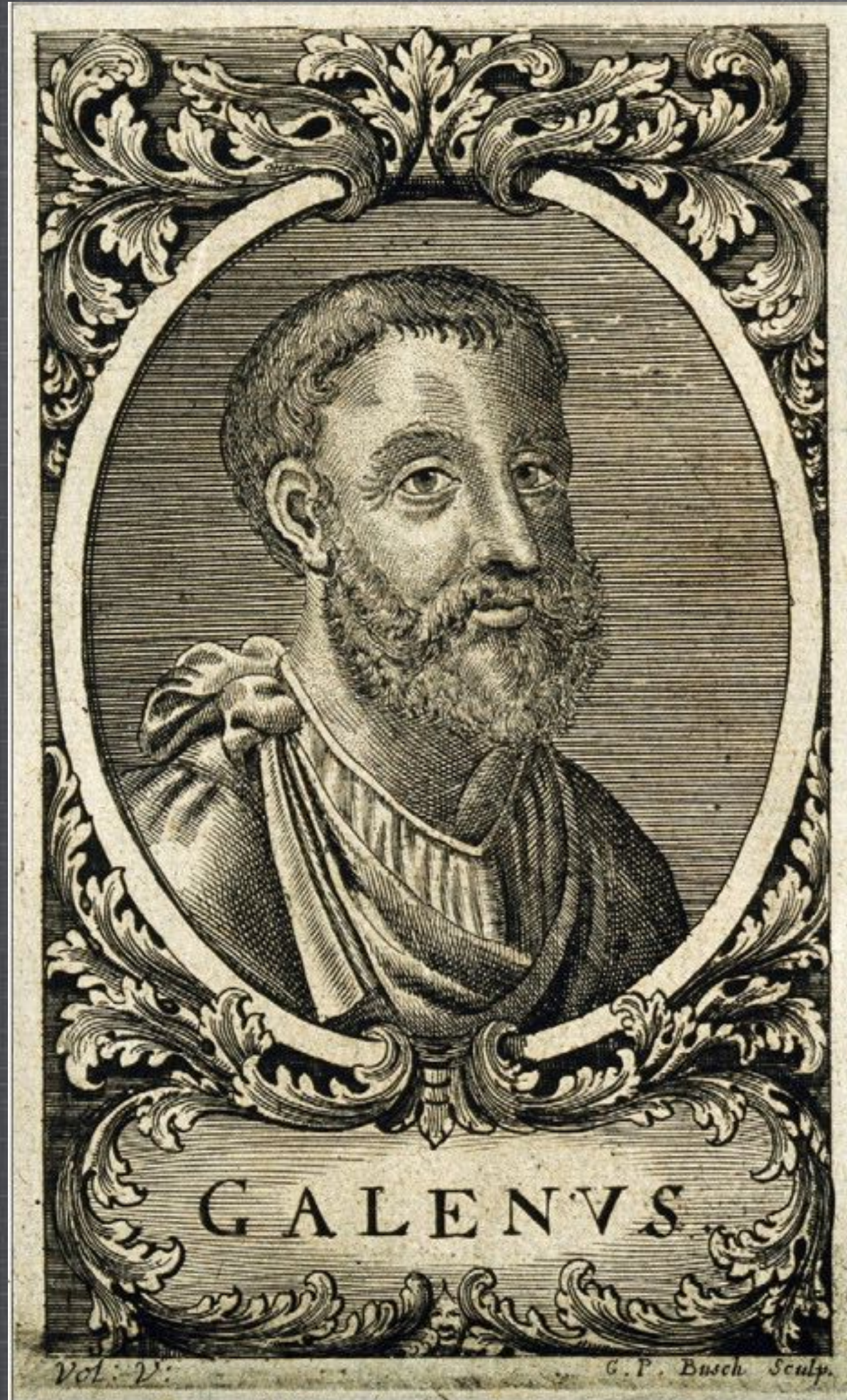
A CONSENSUS VIEW?



Intelligence is “an aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment”

David Weschler, developer of the WAIS IQ test.

GALEN. 2ND C. BCE



Established that the intelligence resided in the **brain not the heart**

Mental functions depend on the **structure and condition of the brain**
Damage to the brain impairs cognition

Proposed a **pneumatic theory** (animal spirits) of how the brain controlled the body

DESCARTES MIND-BODY DUALISM

“it is certain that I, that is, my **mind**, by which I am what I am, is entirely and truly distinct from my **body**, and may exist without it.”

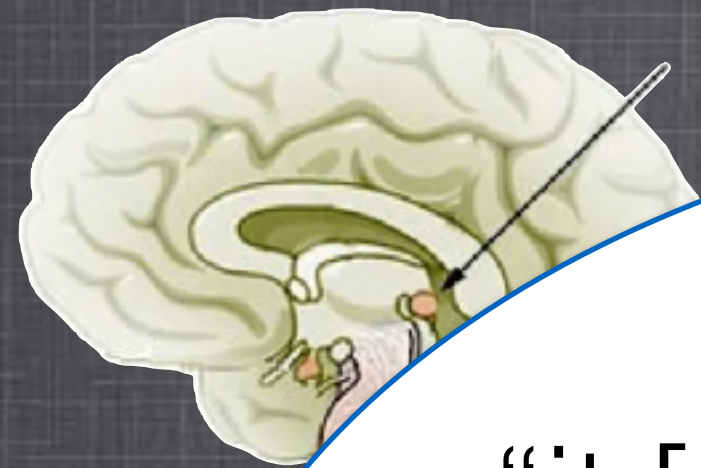
Meditations 6



Renee Descartes, 17th c.

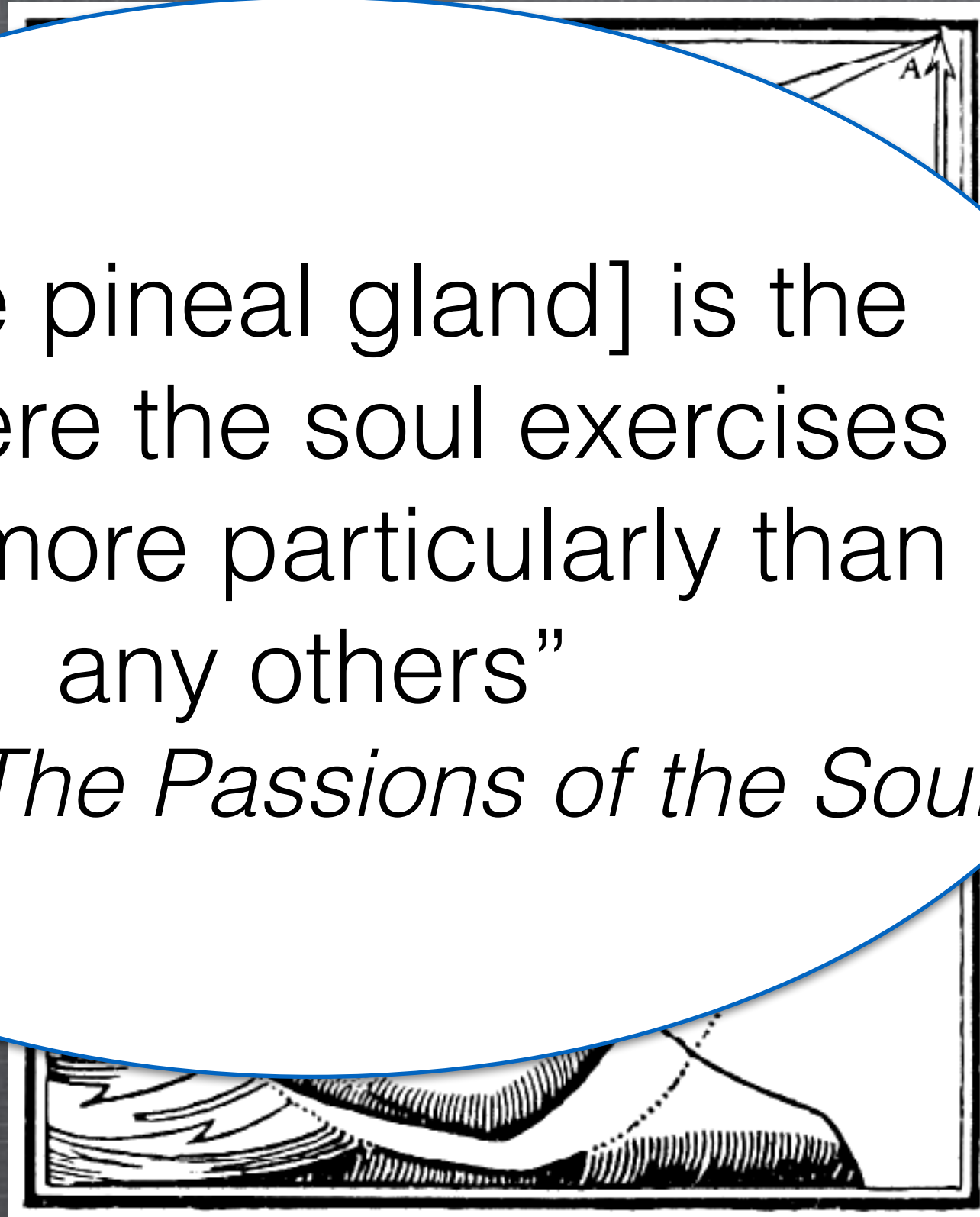
DESCARTES MIND-BODY DUALISM

pineal gland



“it [the pineal gland] is the place where the soul exercises its function more particularly than in any others”

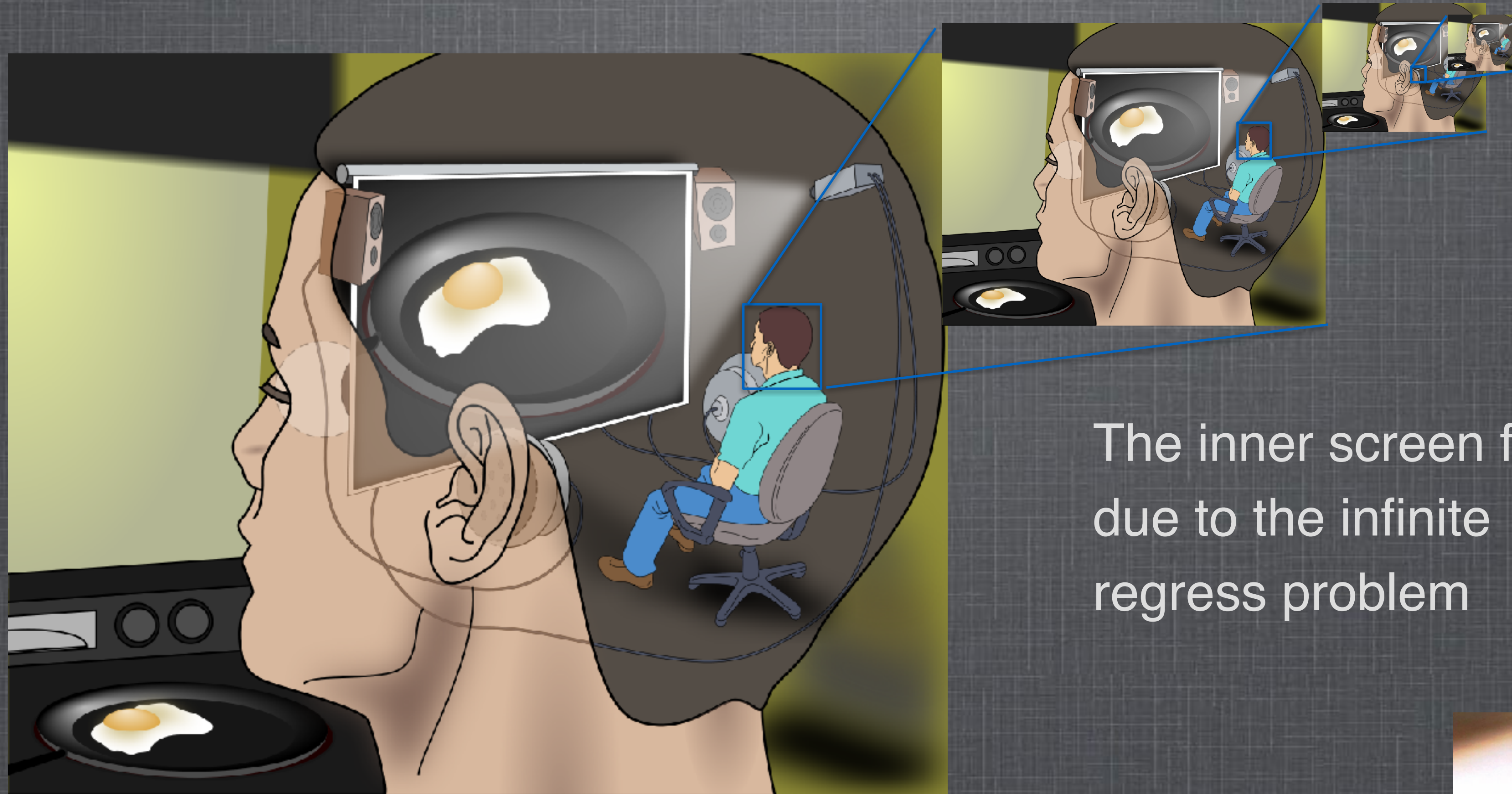
The Passions of the Soul



Renee Descartes, 17th c.

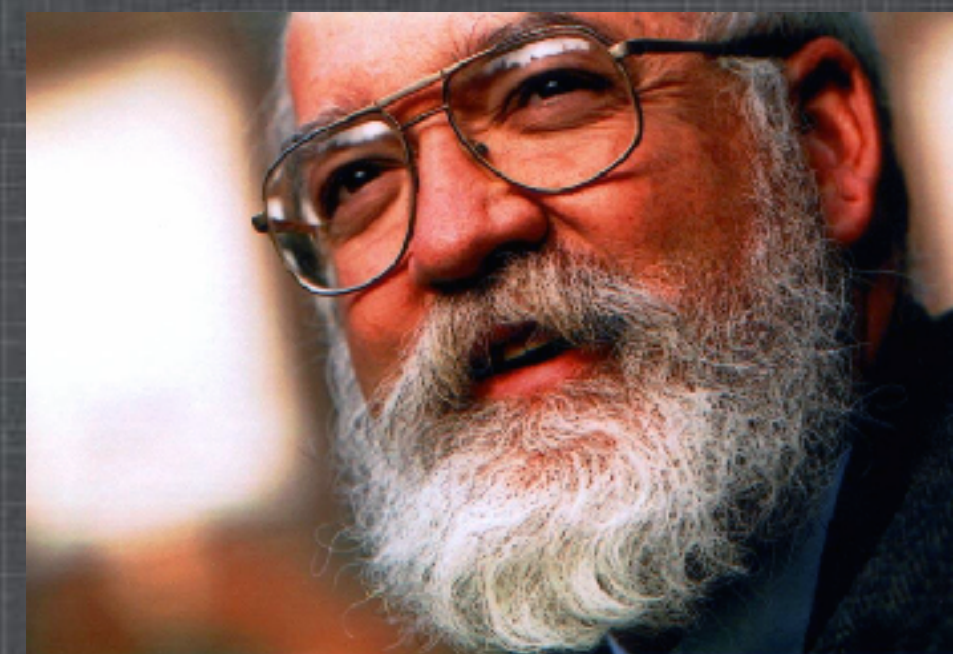
Descartes famously described the pineal gland [a small gland in the middle of the brain] as the "principal seat of the soul" and the site of mind-body interaction

THE “INNER SCREEN” THEORY OF SEEING



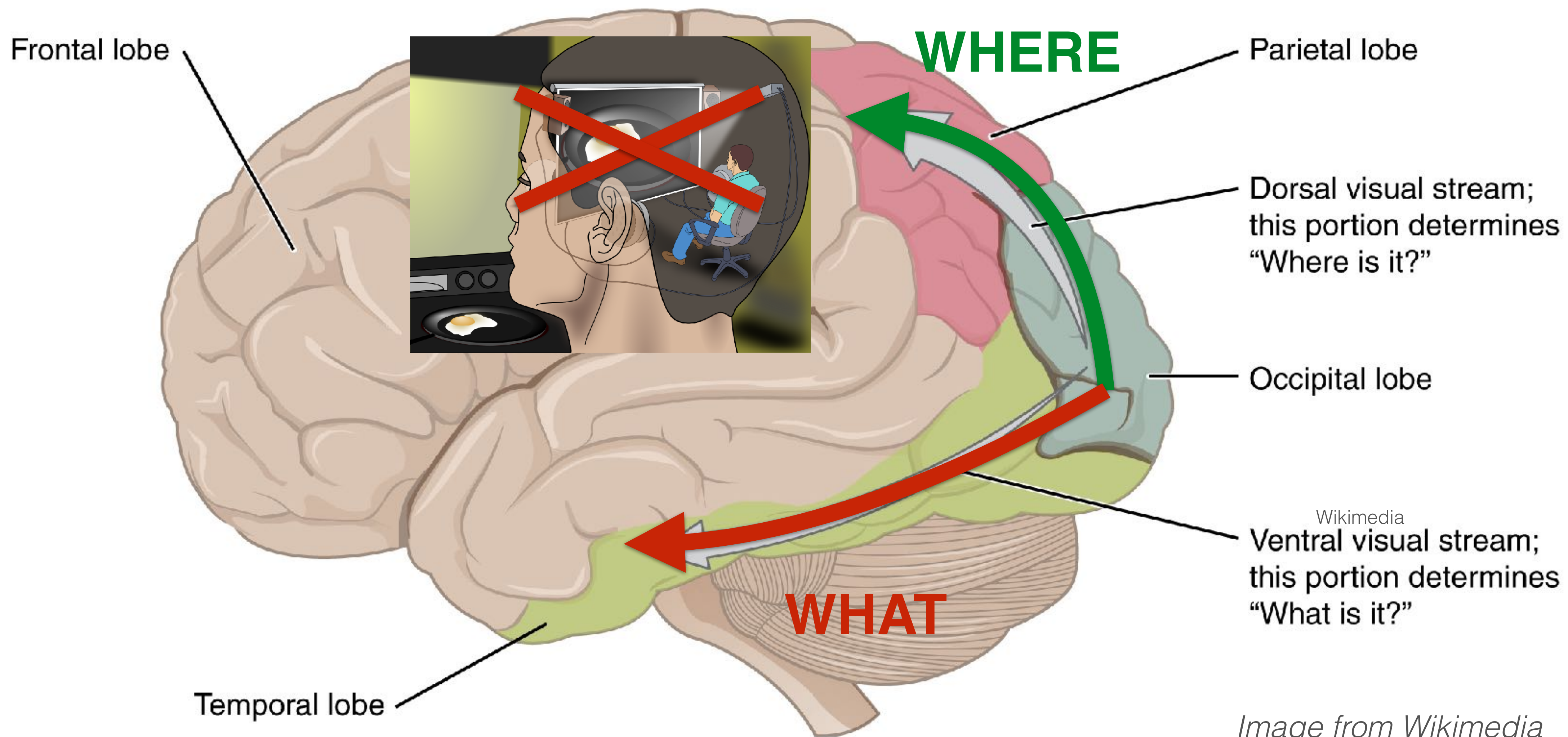
Jennifer Garcia, Wikimedia, CC-BY

The inner screen ‘theory’ proposes that the brain represents the brightness of points in the scene like a being in a movie theatre (Daniel Dennett called this the “Cartesian Theatre”). Here it is imagined upright to match our perceptions of the world.



Daniel Dennett, *Consciousness Explained*, 1991.

VISION PROCESSING IN THE BRAIN FOLLOWS MULTIPLE PATHWAYS



There is no single location where everything comes together...

WHAT IS INTELLIGENCE? THE VIEW FROM AI

Intelligence is Computation

“The computational part of the ability to achieve goals in the world”. (John McCarthy)

“I have always been convinced that the only way to get artificial intelligence to work is to do the computation in a way similar to the human brain” (Geoffrey Hinton)



By OLCF at ORNL, CC BY 2.0

In 2023, the HPE Frontier supercomputer can do one million trillion floating point operations per second. More than enough to simulate the 100 trillion connections in the human brain if we knew how to write that program.

THE ROOTS OF COMPUTATION ARE MULTICULTURAL

Many civilizations have developed:

- Systematic numerical representation
- Rule-based transformation procedures
- Large-scale numerical prediction
- Mechanical or physical aids for calculation



Babylonian clay tablets, c. 1800 BCE



Chinese
number
system c.
2500 BCE

—	==	≡	≡	≡
1	2	3	4	5
↗↗	†	) (	ㄥ	
6	7	8	9	10
∪	∩	∩	⋈	↗
20	30	40	50	60
⊖	⊖	⊖	⊖	⊖
100	200	300	400	500
ㄣ	ㄣ	ㄣ	ㄣ	ㄣ
1000	2000	3000	4000	5000



Astrolabe.

A tool to aid navigation by the
sun and stars, 13th. C.

Quipu, Incan counting tool,
c. 2500 BCE

BABYLONIAN ALGORITHMS, 1,800 BCE

Clay tablets
show evidence
of algorithmic
thinking ~3,800
years ago!



BABYLONIAN ITERATIVE METHOD FOR $\sqrt{N}$

Algorithm:

1. Choose an initial guess x_0
2. Compute: $x_{i+1} = \frac{1}{2} (x_i + N / x_i)$
3. Repeat until stable

- Example for $N=2$:

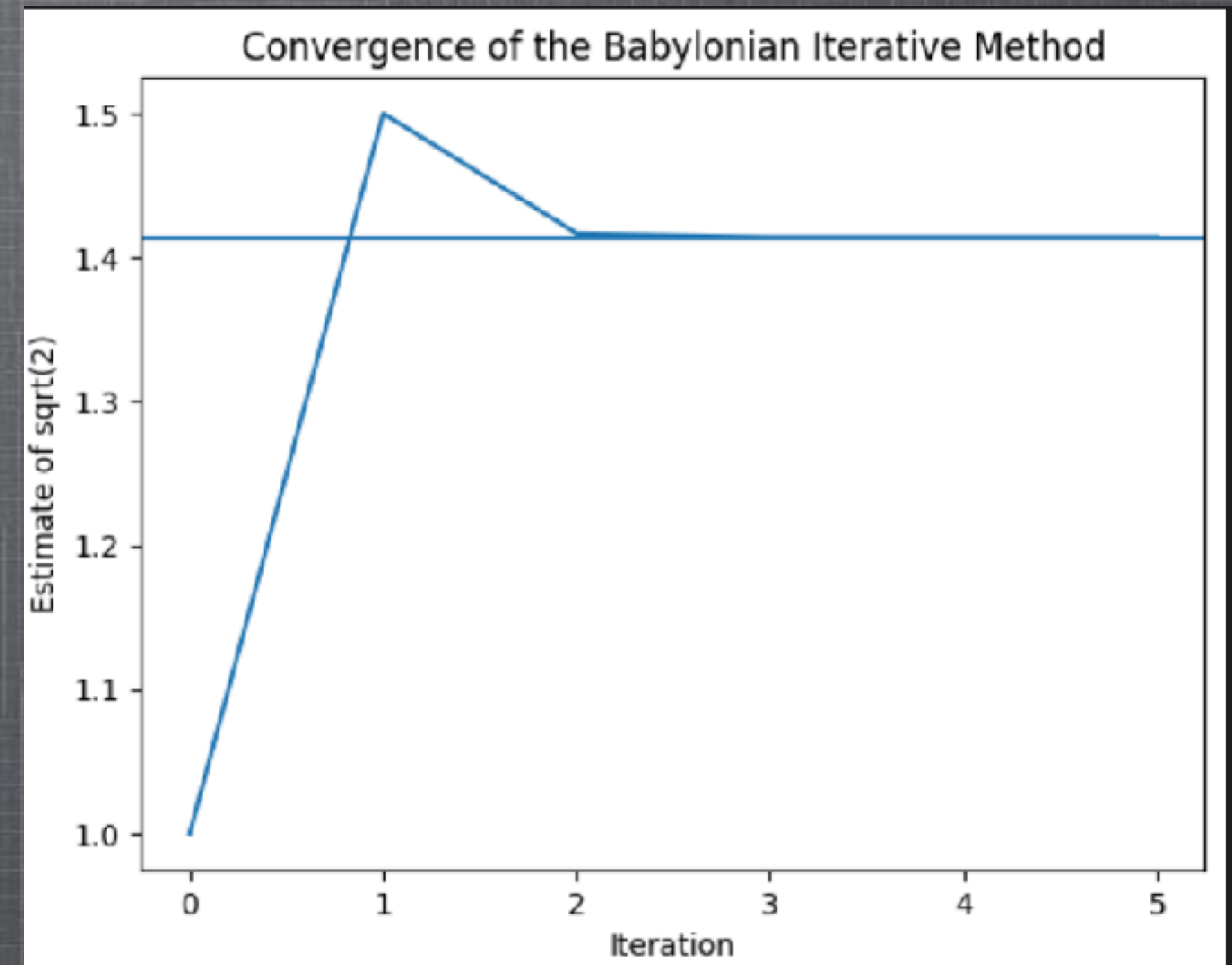
- $x_0 = 1$

- $x_1 = \frac{1}{2}(1 + 2/1) = 1.5$

- $x_2 = \frac{1}{2}(1.5 + 2/1.5) \approx \frac{1}{2}(1.5 + 1.333) \approx 2.833/2 \approx 1.4167$

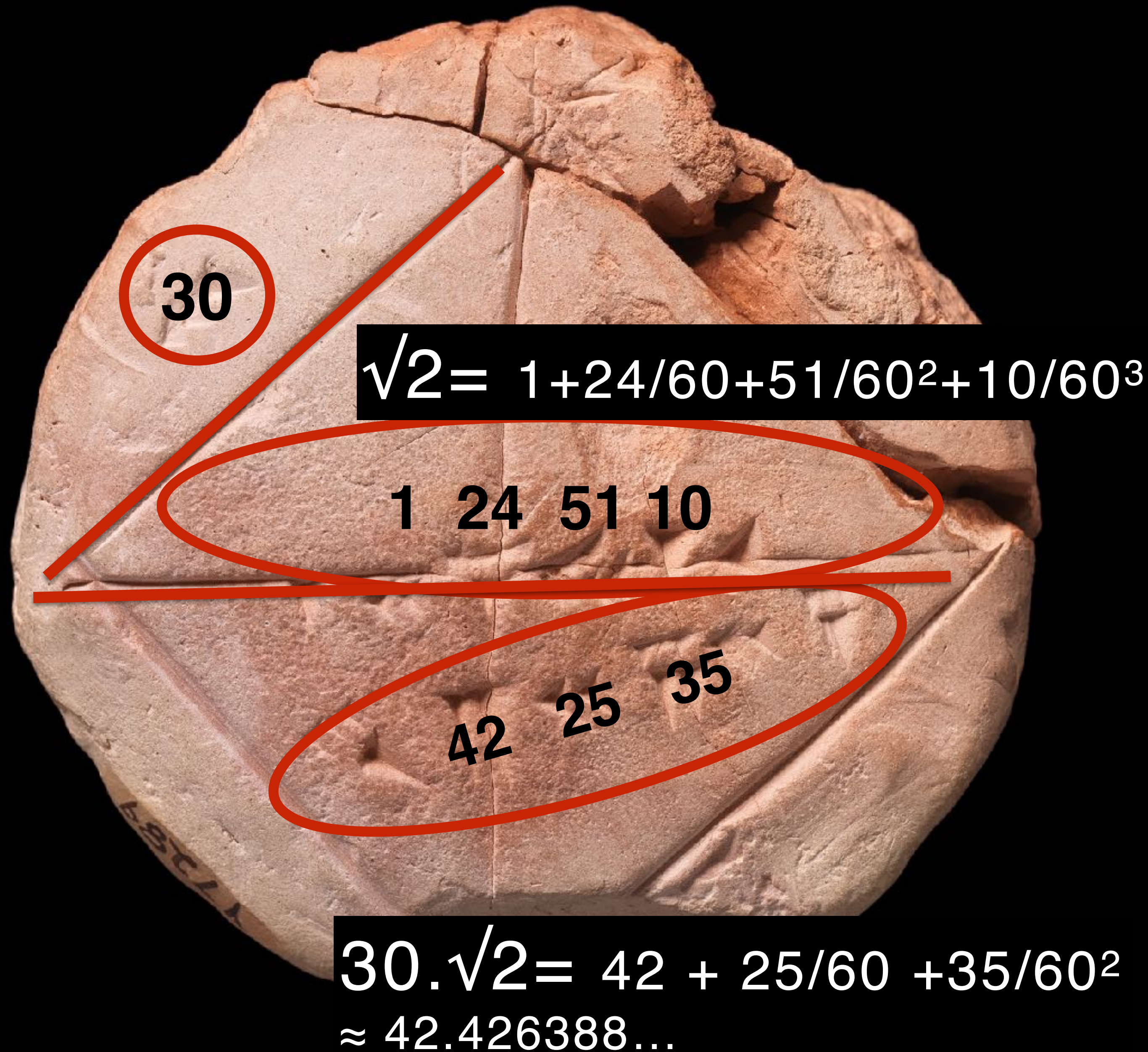
- $x_3 \approx \frac{1}{2}(1.4167 + 2/1.4167) \approx \frac{1}{2}(1.4167 + 1.4118) \approx 2.8285/2$

- ≈ 1.4142

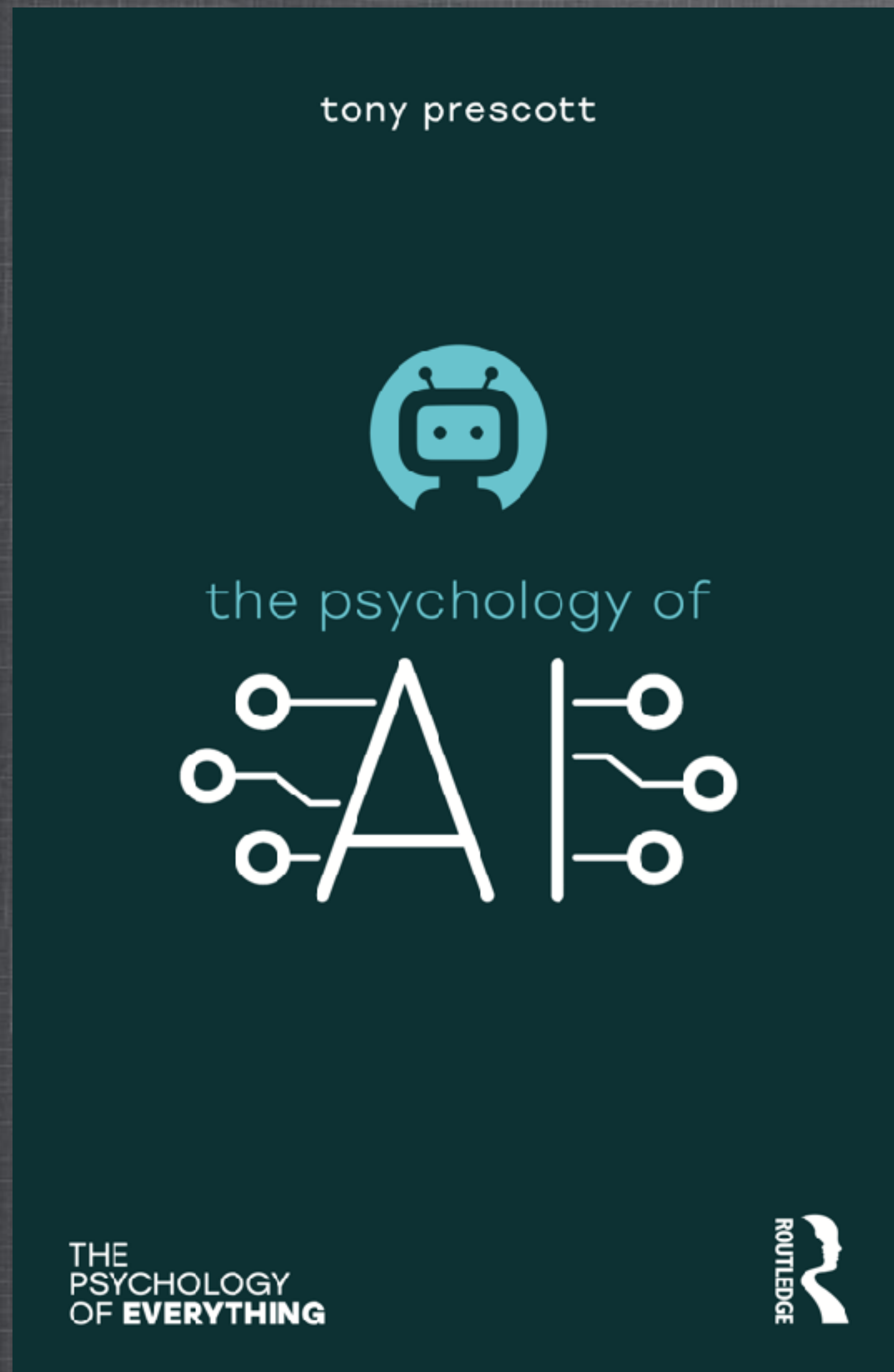


BABYLONIANS WERE ACCURATE TO SIX DECIMAL PLACES!

- Tablet YBC 7289 records:
- 1; 24, 51, 10 (base-60)
- ≈ 1.414213
- (accurate to ~6 decimal places)
- Because base-60 makes it easy to represent fractions, e.g. $2/1.5 = 1.333... = 1; 20$



THE PSYCHOLOGY OF ARTIFICIAL INTELLIGENCE



- *Overview*
- What is intelligence?
- **Are brains computers?**
- Do AIs understand anything?

A DEEP DIVE INTO A STATE-OF-THE-ART COMPUTER

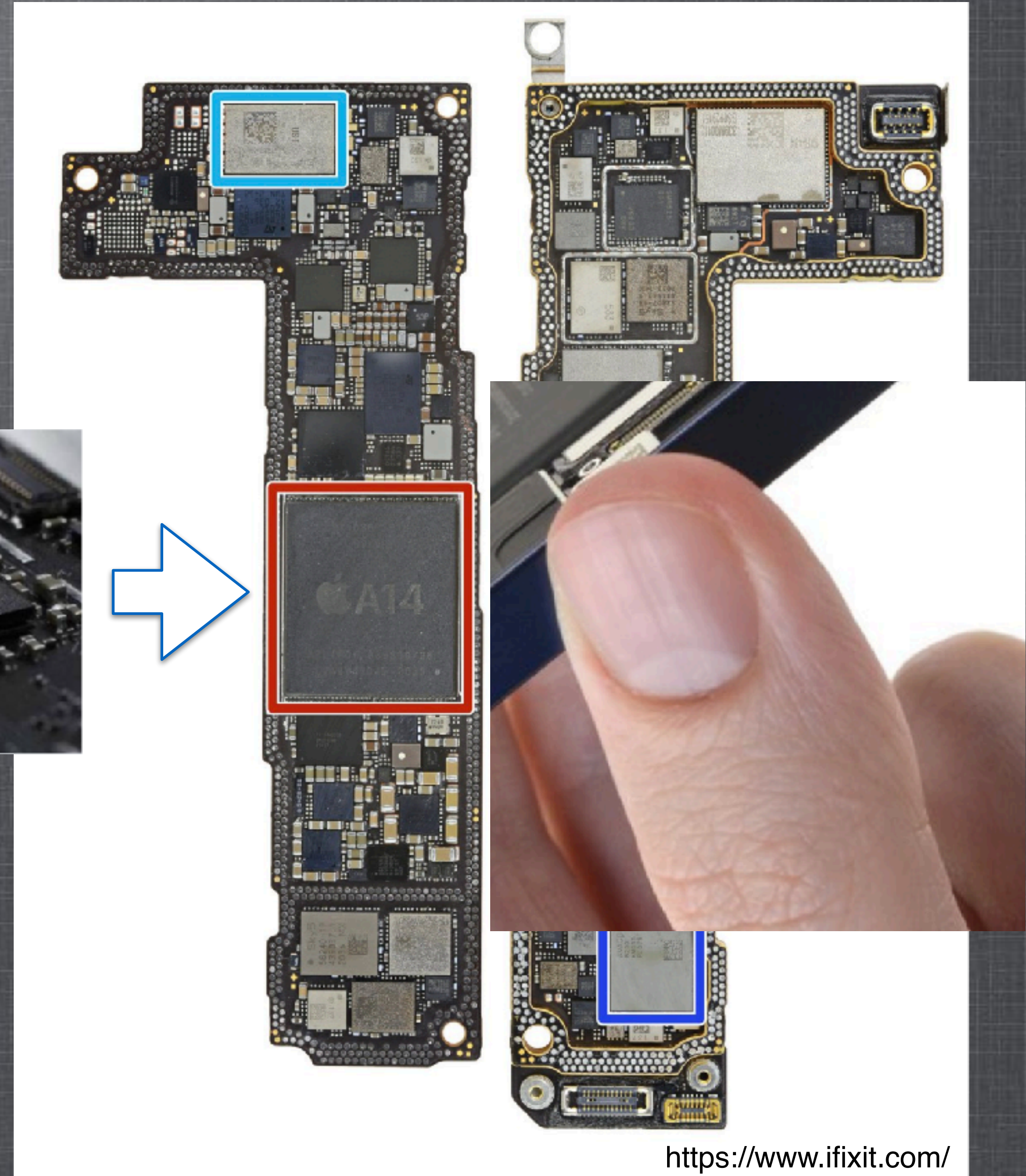


<https://www.ifixit.com/>

12: H: 146.7 mm (5.78 in) W: 71.5 mm (2.81 in) D: 7.4 mm (0.29 in)

The iPhone 12

10cm



<https://www.ifixit.com/>

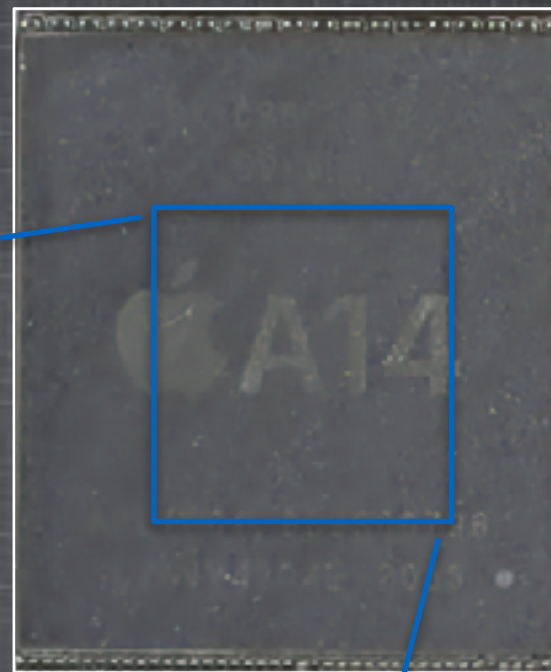
[from www.ifixit.com/](https://www.ifixit.com/)

A14 BIONIC SYSTEM ON A CHIP

- **4-core** graphics processing unit (GPU)
- **6-core** 64-bit CPU implementing ARMv8 ISA,
 - 4x 1.8 GHz energy-efficient *Icestorm* cores
 - 2x 3.1 GHz high-performance *Firestorm* cores
- **16-core** *Neural Engine* 11 trillion operations/s
- **11.8B** transistors at a density of 134M/mm²



Die Photo



88mm²

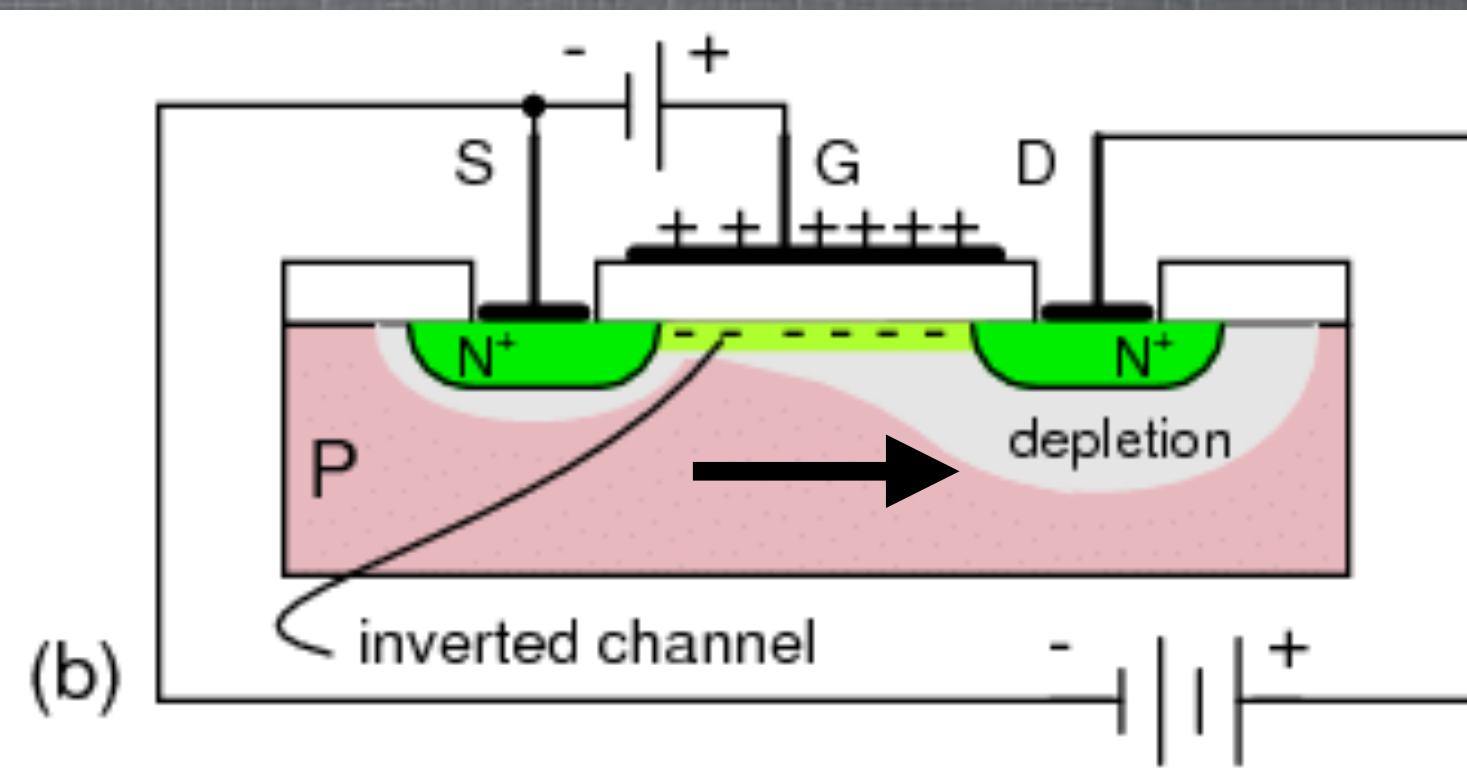
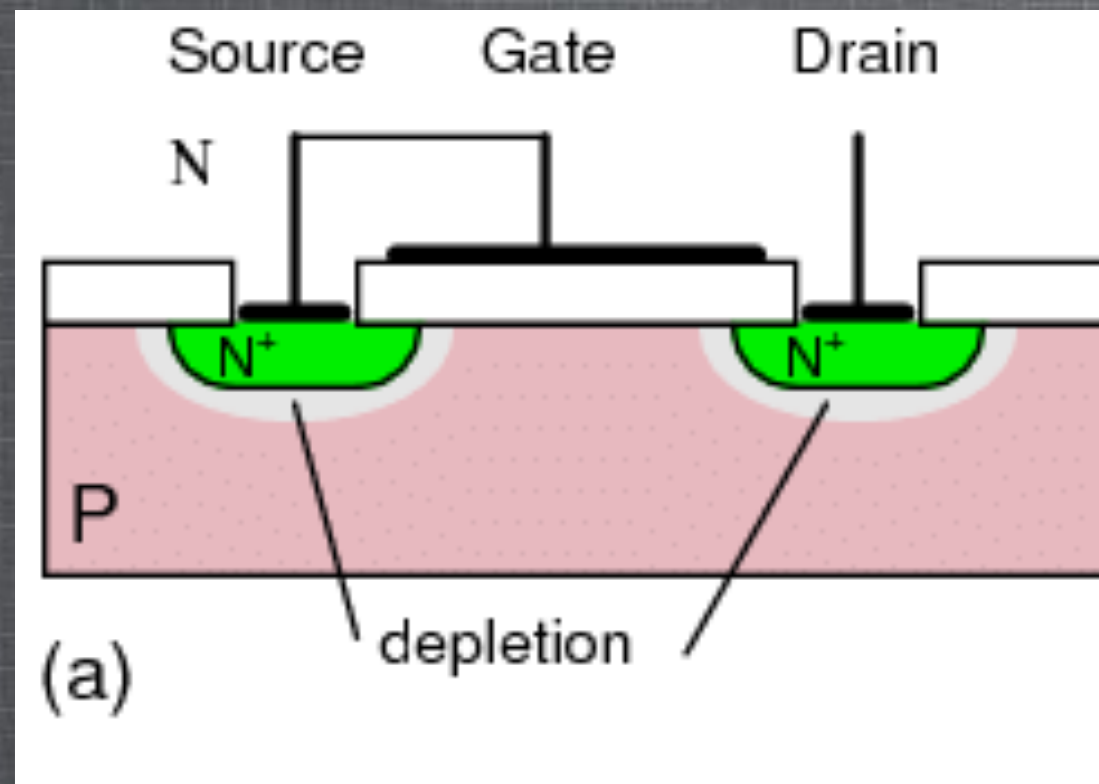
0.5cm

TRANSISTORS

Has three terminals: Gate, Source, Drain. Voltage at the Gate controls current flow between Source and Drain

Off State (0)

- Gate voltage is low
- No current flows from Source → Drain
- Acts like an open switch



On State (1)

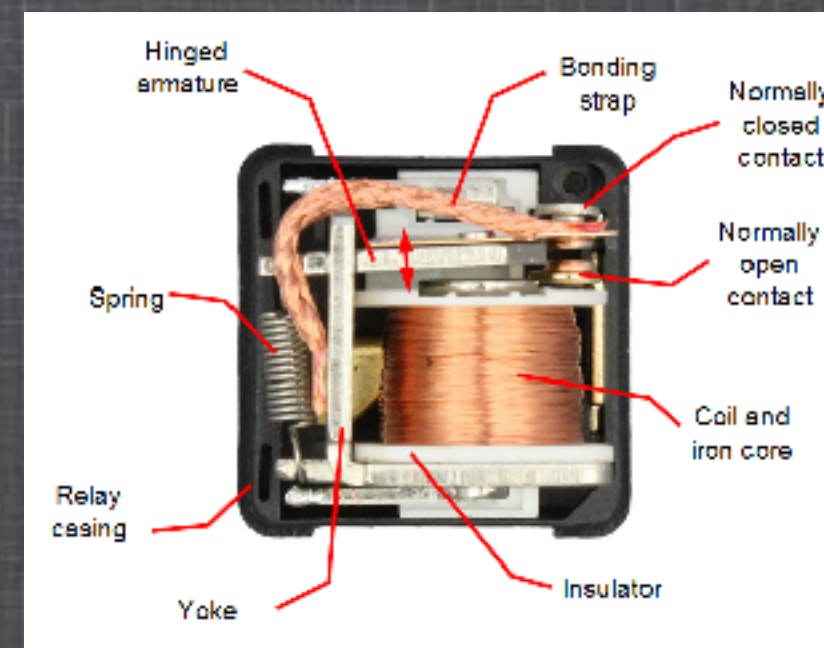
- Gate voltage is high
- Charge is pushed away from the drain (depletion)
- This creates an (inverted) conductive channel in the semiconductor
- Current flows from Source → Drain
- Acts like a closed switch

Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) — most common in CPUs

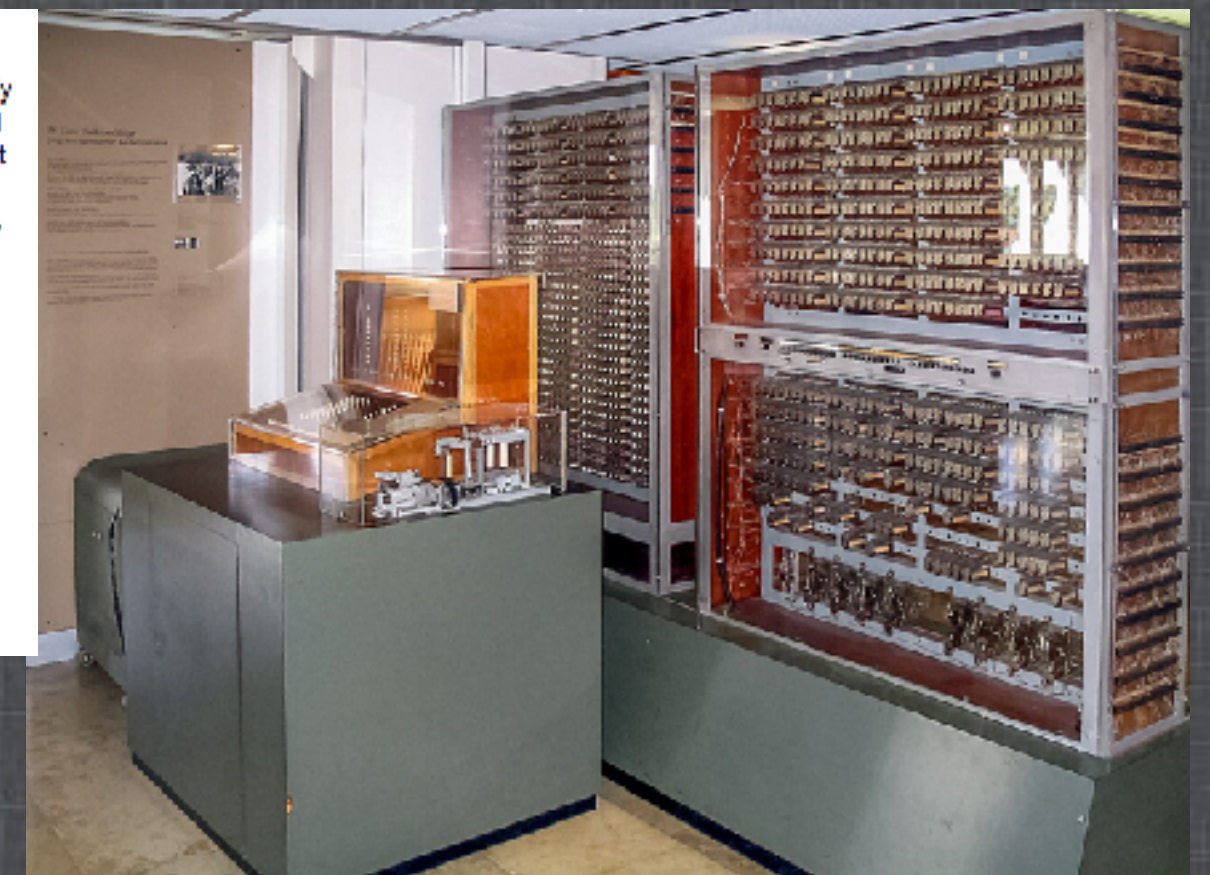
By controlling many transistors in circuits, we can perform logic operations (AND, OR, NOT) and build registers, ALUs, caches, and eventually an entire CPU core.

Functionally a transistor is just like a relay, but nanoscale, with no moving parts and 1 billion times faster

Exactly as Claude Shannon showed in 1937!



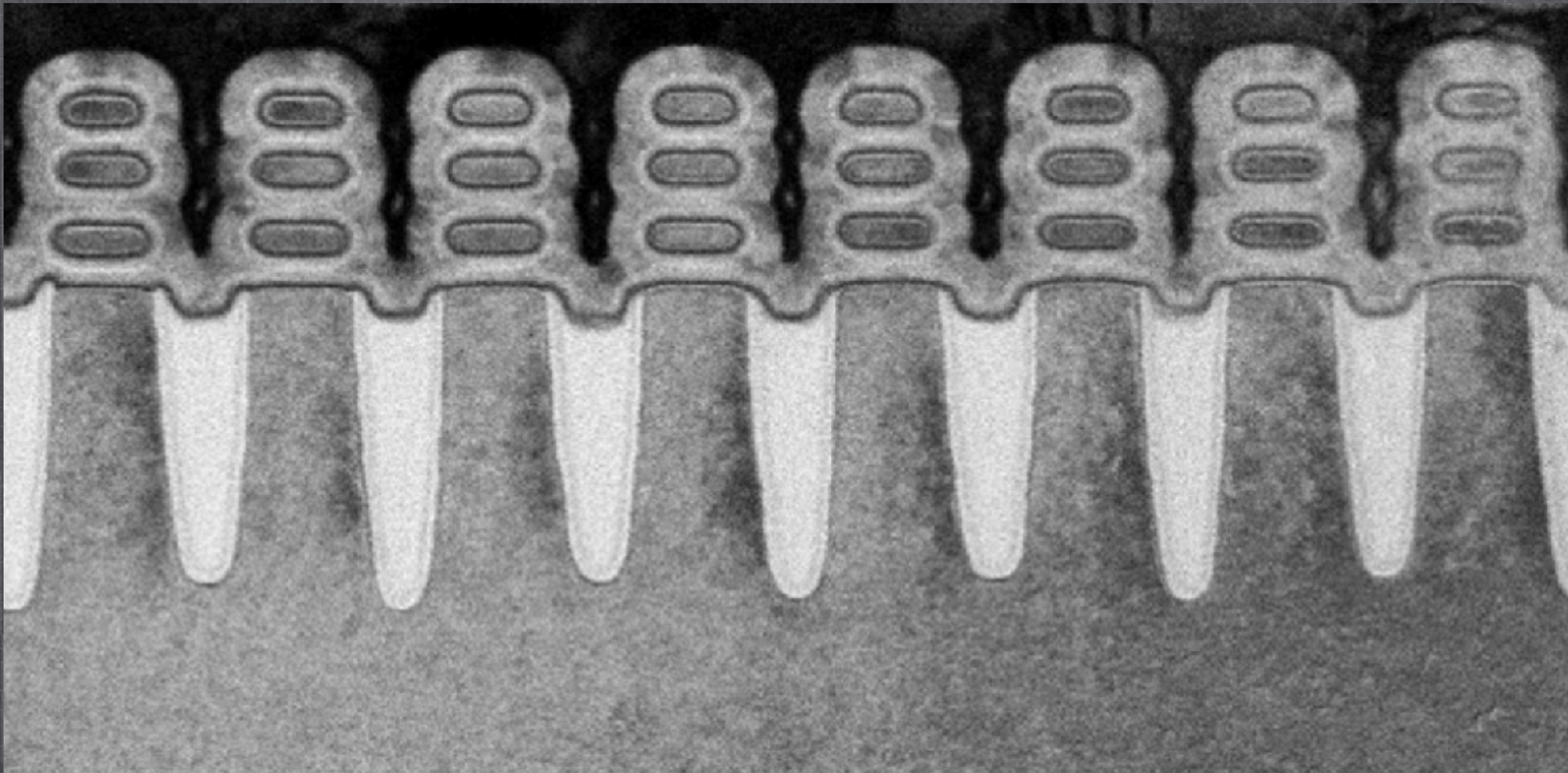
≈ 100 ops/s



Z3 computer, 1941, 2,000 relays

TRANSISTORS AT NANOSCALE

TEM image of FinFET transistors similar to those found in the A14.



Transmission Electron Micrograph (TEM)

~30,000 of these extreme close-up transistor slices would fit linearly across the width of the A14 chip

<https://www.top500.org/news/new-five-nanometer-transistor-unveiled-by-ibm-and-cohorts/>

IF THE A14 USED RELAYS....



Visualised by ChatGPT

A14 CPU CORES

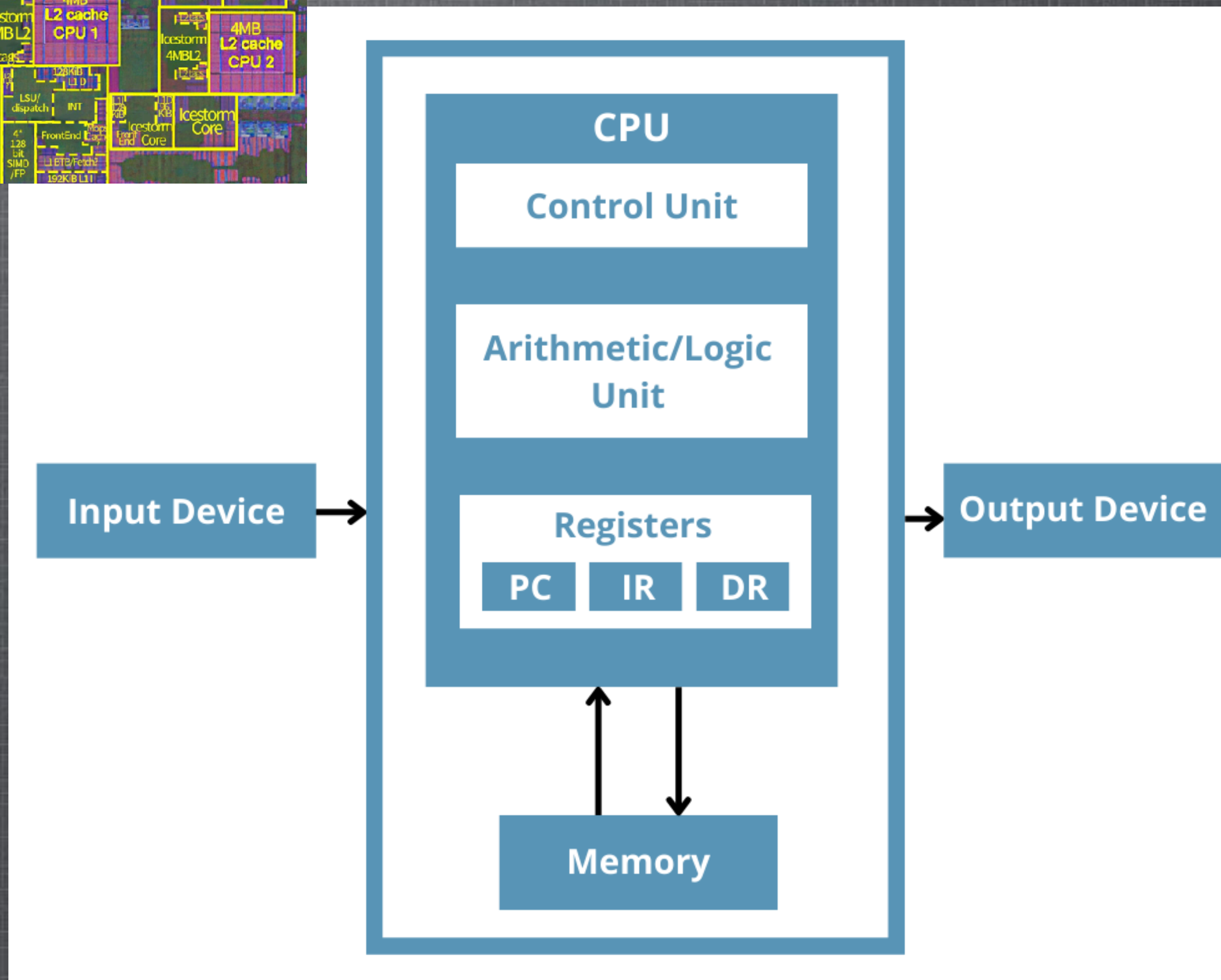
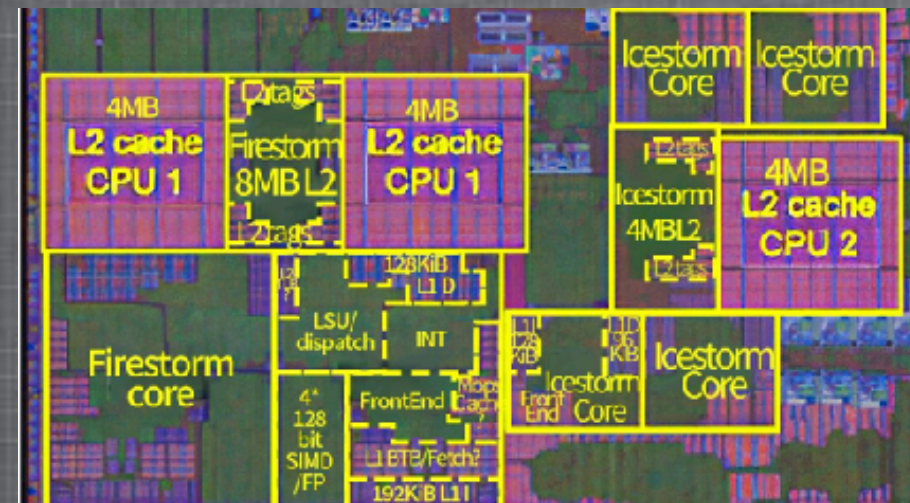
Follows the classic von Neumann design

Each CPU core:

- Fetches instructions (like “open app” or “calculate this”)
- Decodes them
- Executes math or logic operations
- Writes results back to memory

It contains:

- ALUs (Arithmetic Logic Units)
- Control units
- Registers (fast memory)
- Cache memory



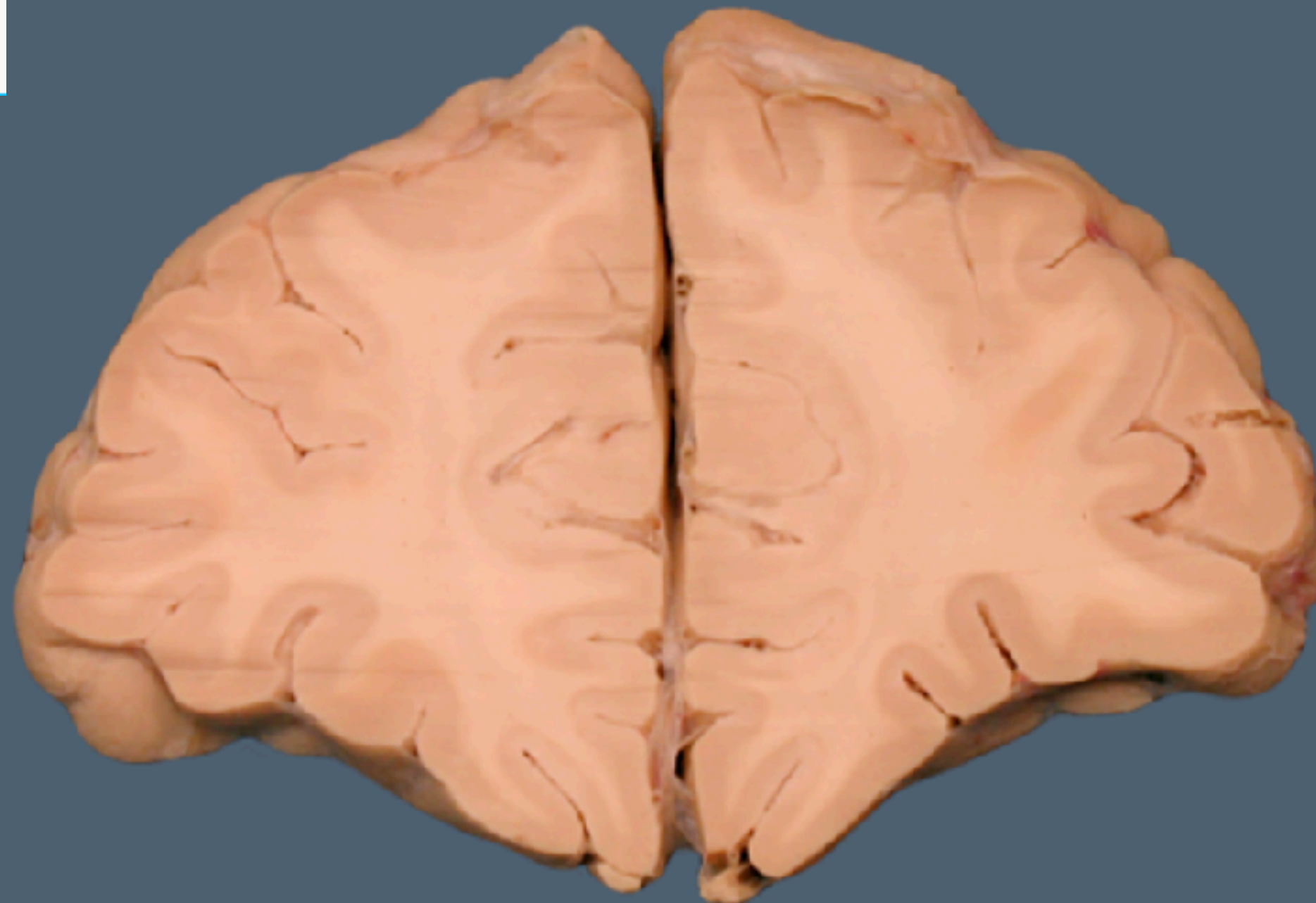
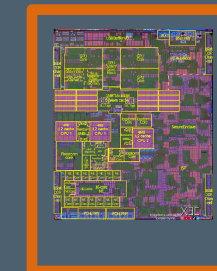
Fetch-execute-write loop

A SLICE OF BRAIN

The adult brain is approximately 16x wider than the A14 chip



A14



10cm



<https://www.neuroanatomy.ca/coronals.html>

DESCARTES' MECHANICAL PHILOSOPHY, 17TH C.

Like Galen, and inspired by some of the mechanical machines of his era. Descartes, thought the the brain function as some kind of hydraulic pump...

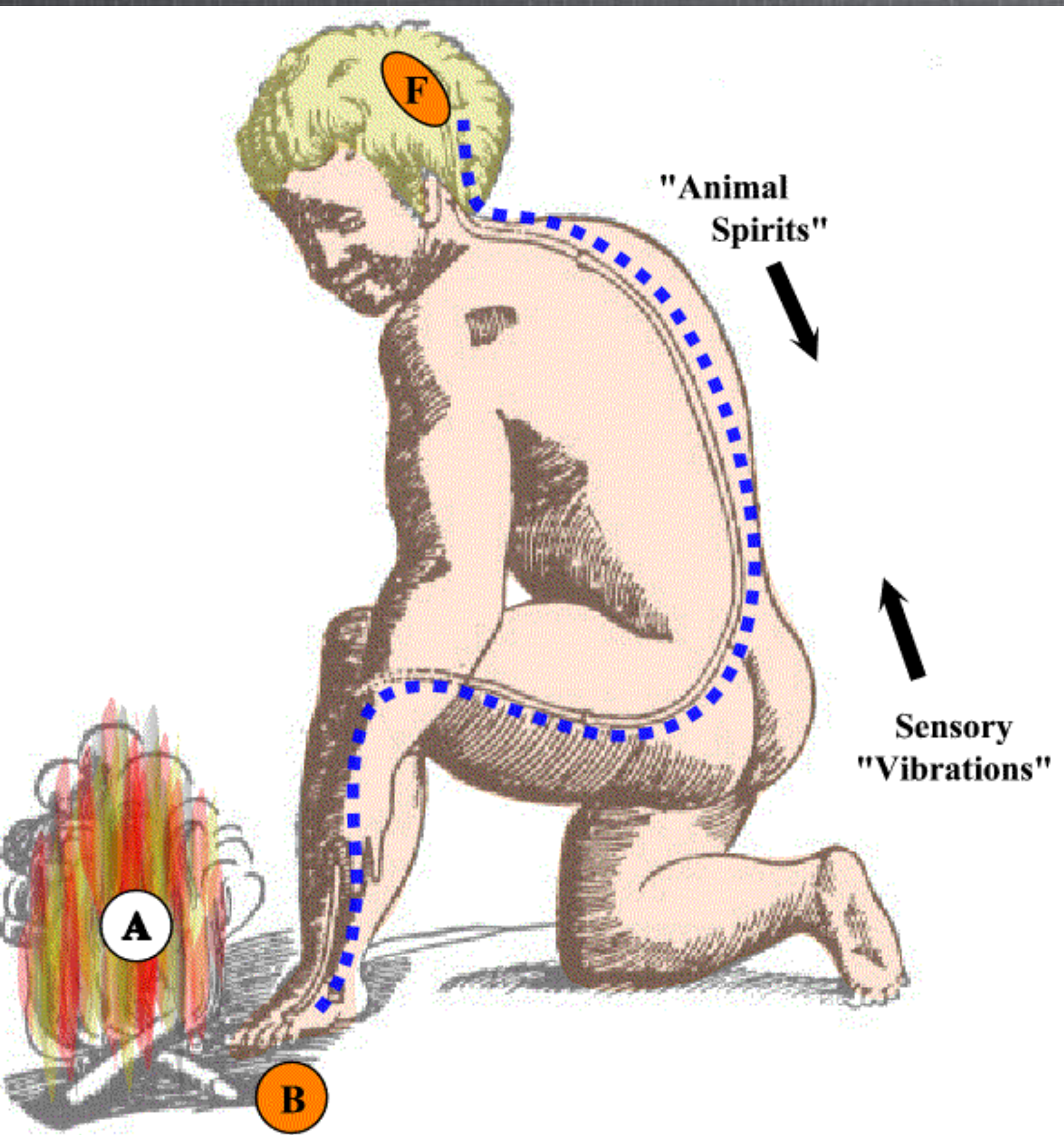
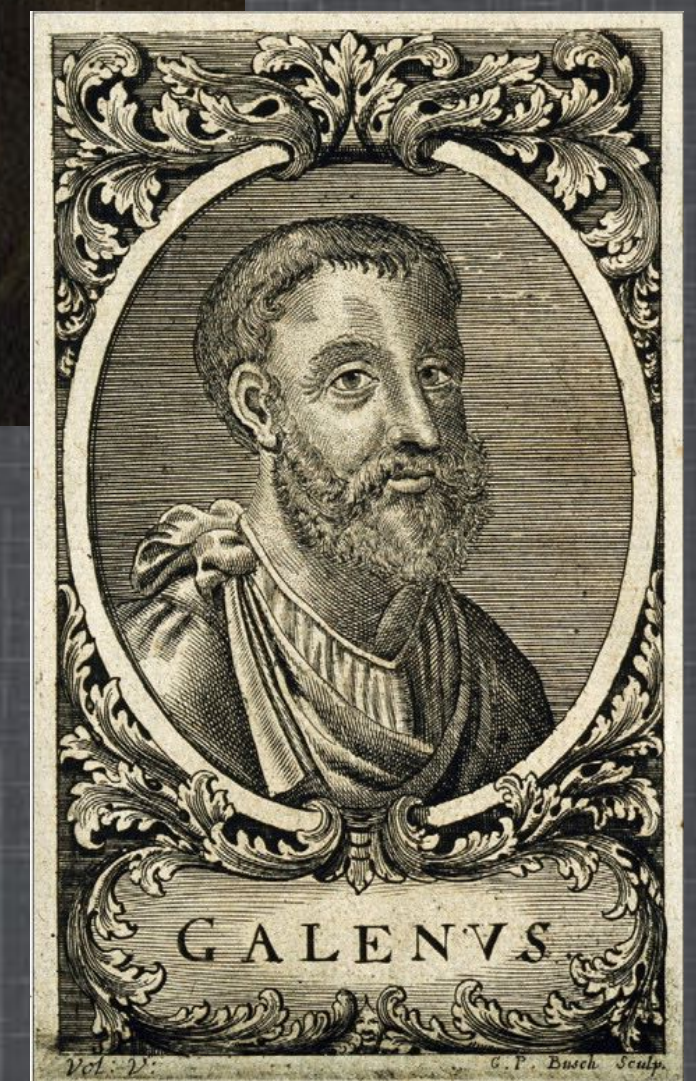
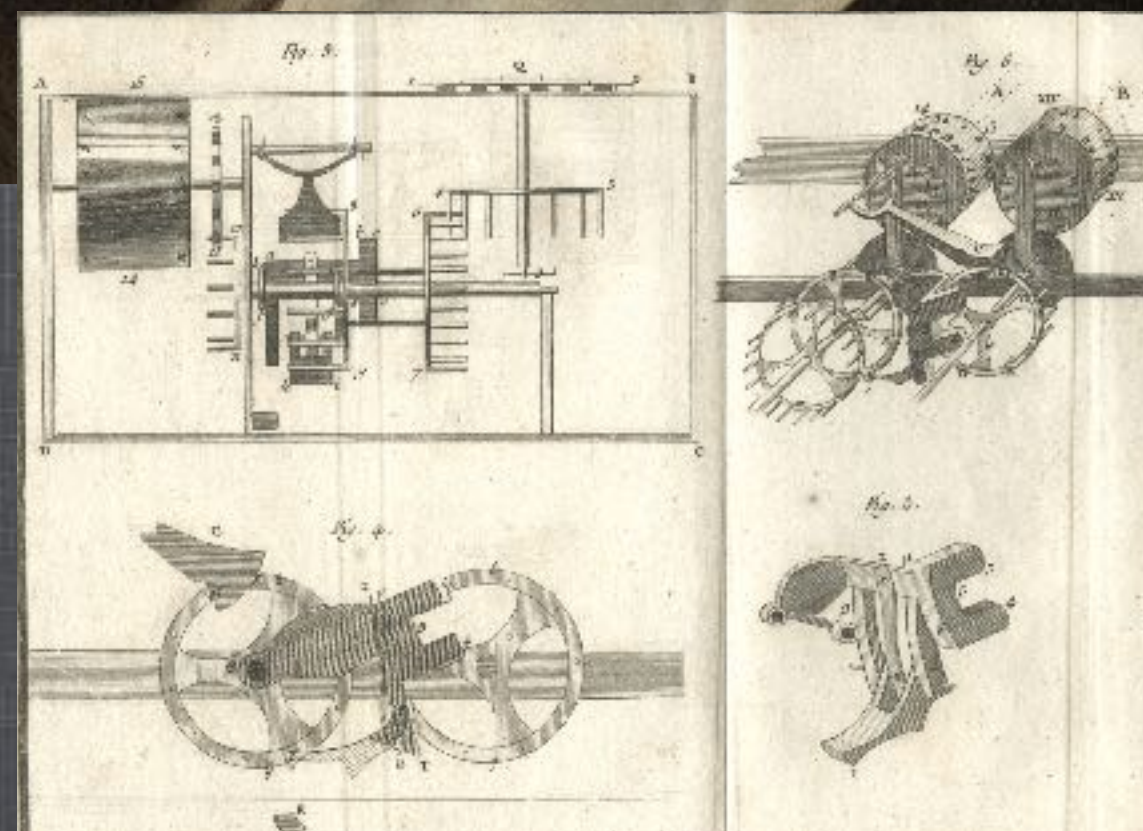
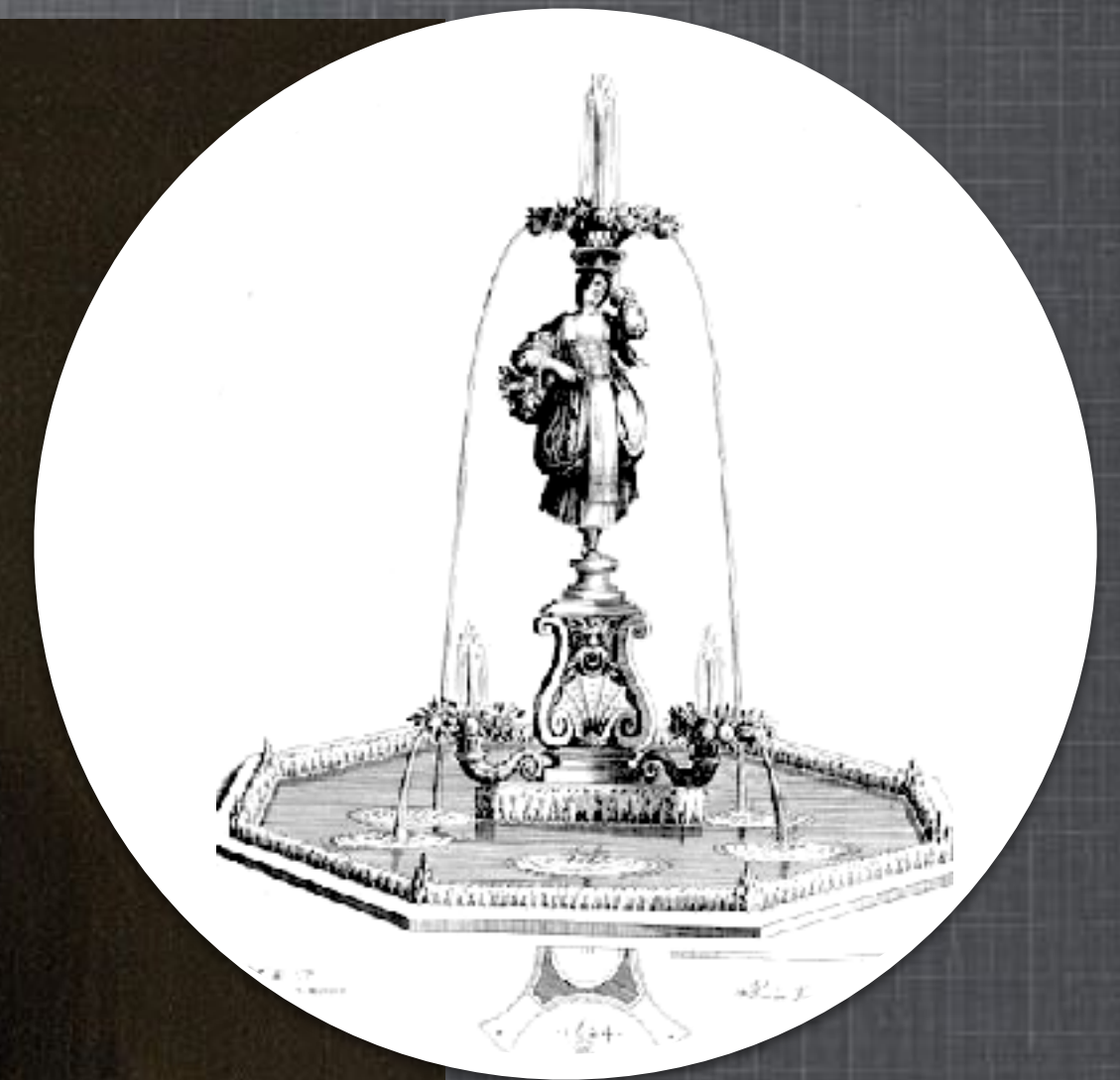
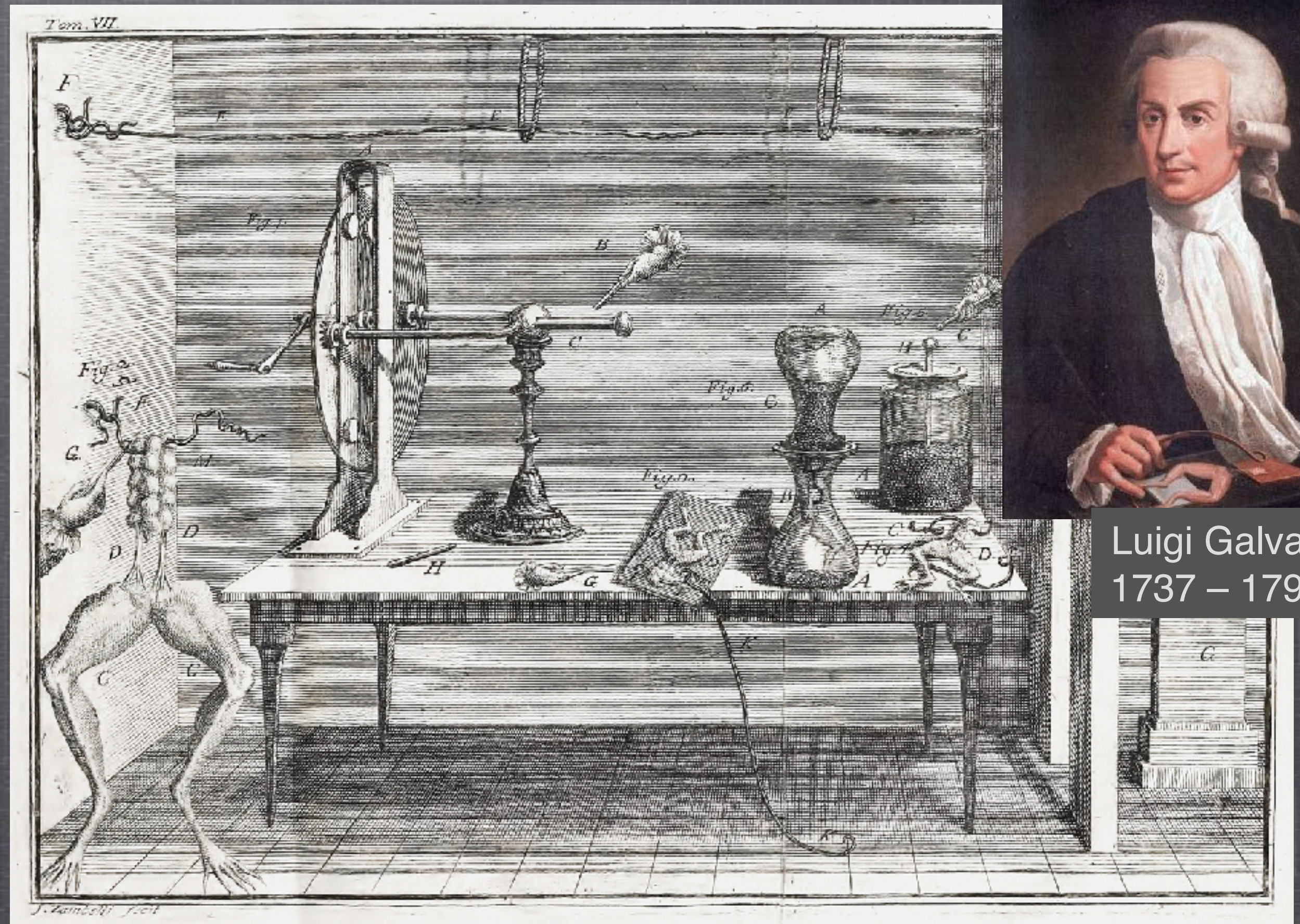
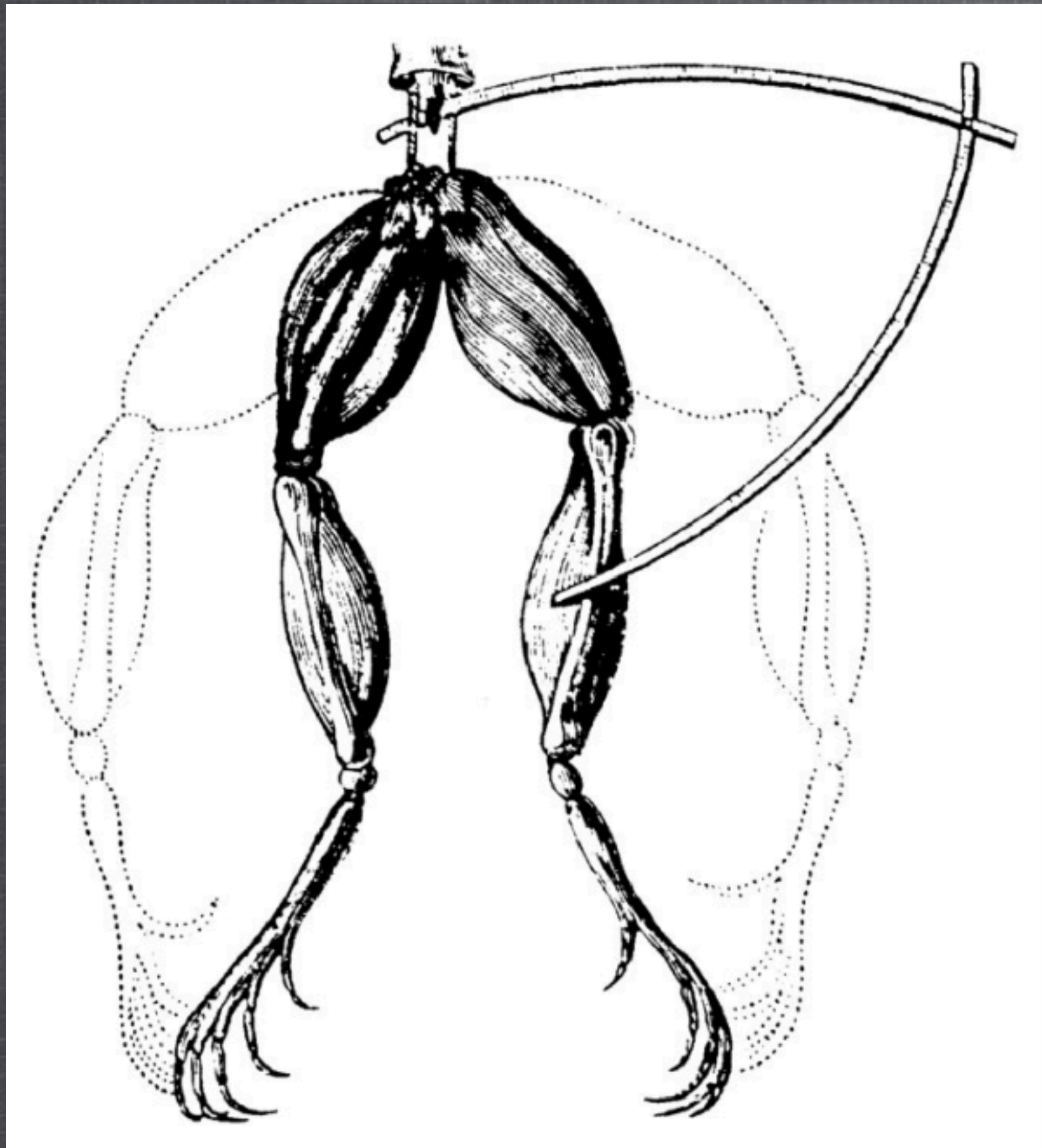


Diagram of the reflex arc,
from *Treatise on Man*, 1662



THIS CHANGED WITH THE DISCOVERY AND HARNESSING OF ELECTRICITY

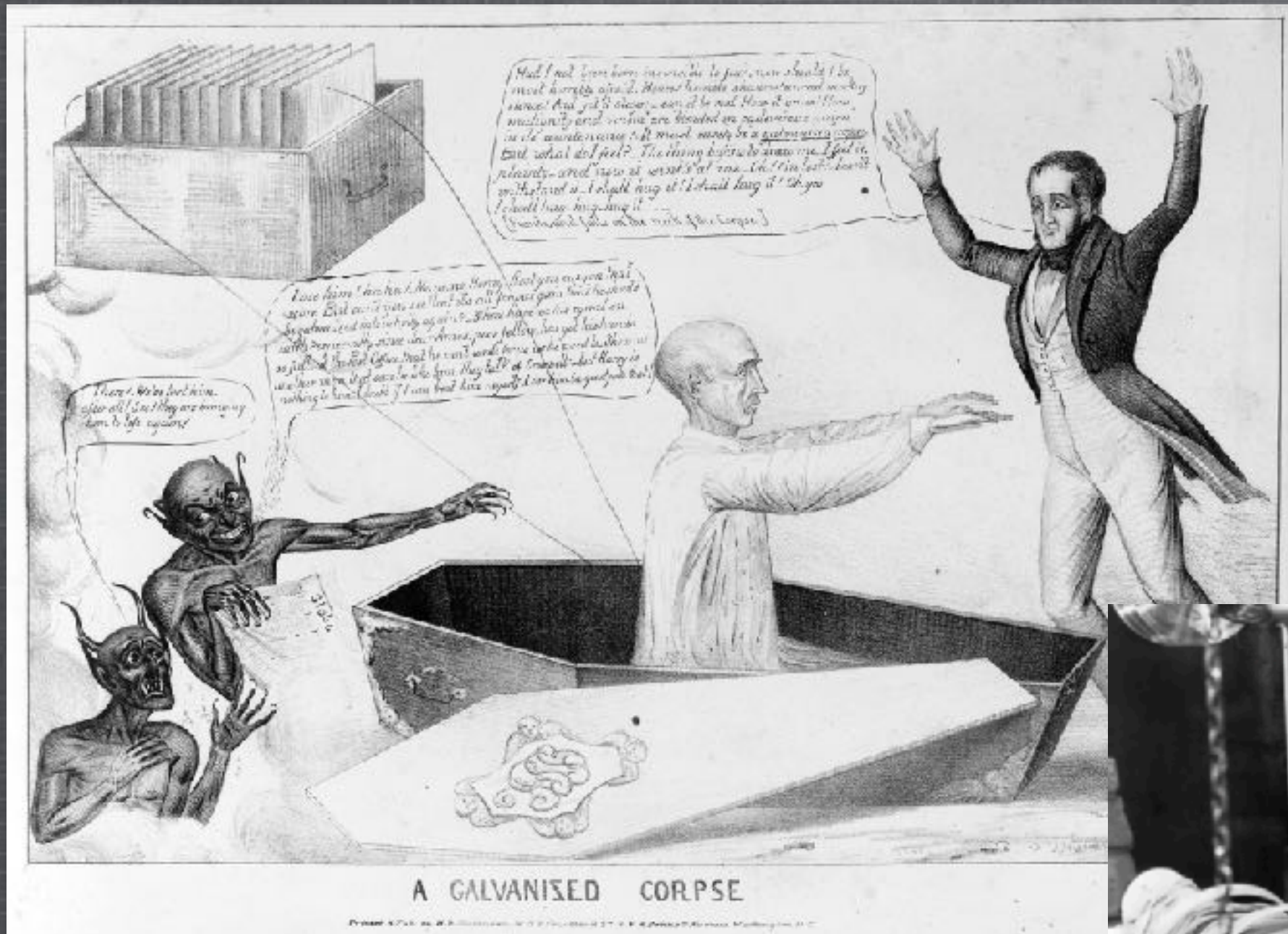


Luigi Galvani
1737 – 1798

In 1780, Luigi Galvani accidentally discovered that a frog's leg will contract if the circuit between the leg's muscle and the frog's spinal cord is closed by means of a connection involving different metals.

Galvani thought a vital life force—an electric fluid—was stored in the tissues of living creatures and discharged through the nerves, he called this “animal electricity”.

AND LED TO SOME MORE WEIRD SCIENCE



A contemporary illustration



Giovanni Aldini, 1762–1834



Still from the 1937 Frankenstein movie

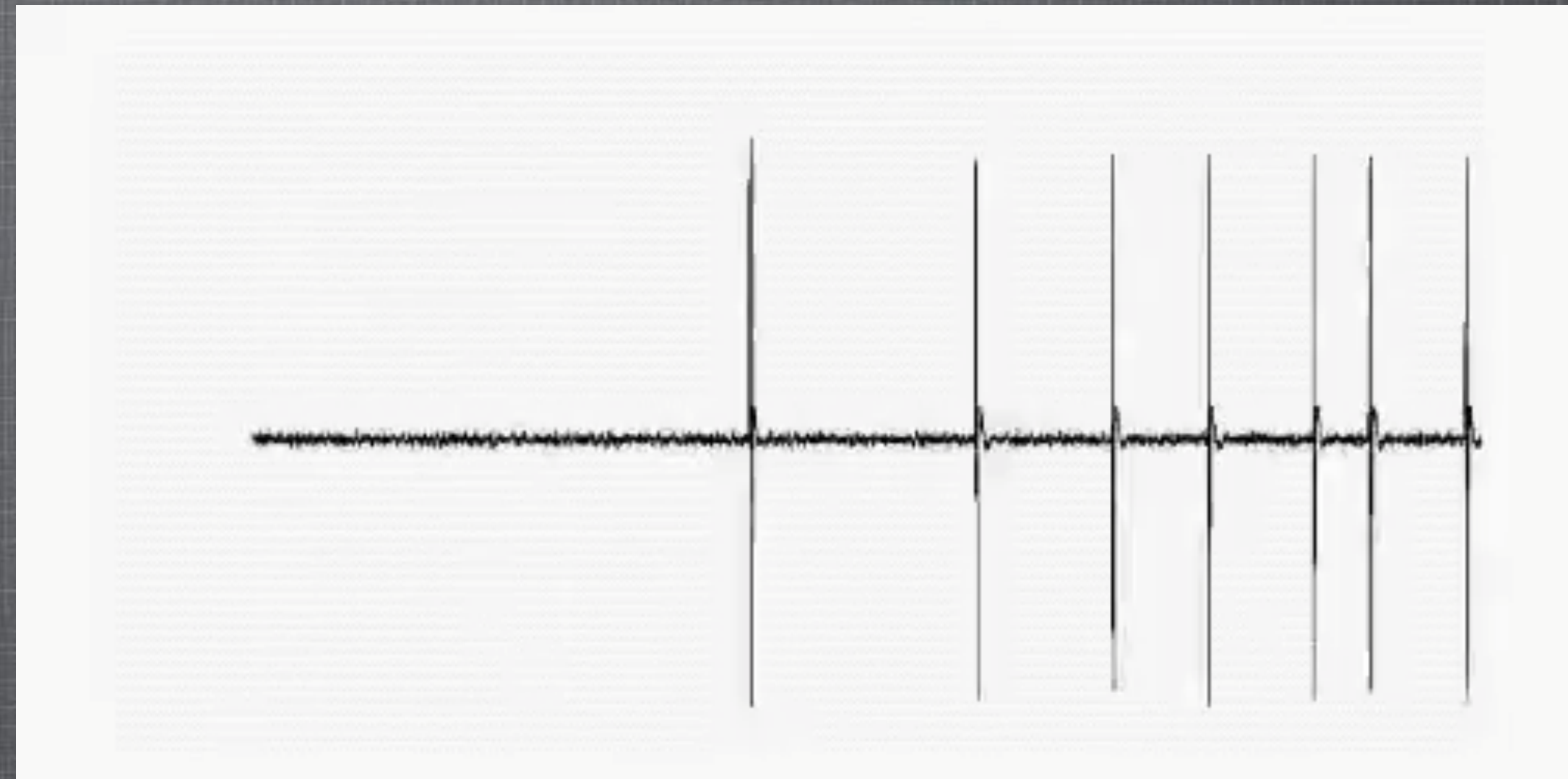
Giovanni Aldini was Galvani's nephew. In a famous experiment at Newgate Prison, London, in 1803, Aldini applied a powerful battery to the corpse of executed murderer George Forster. The jaw quivered, the eyes opened, the limbs moved...

The experiment was said to have inspired Mary Shelley's *Frankenstein*.

1926, RECORDING A NEURONAL SPIKE: EDGAR ADRIAN



Edgar Adrian,
1889-1977

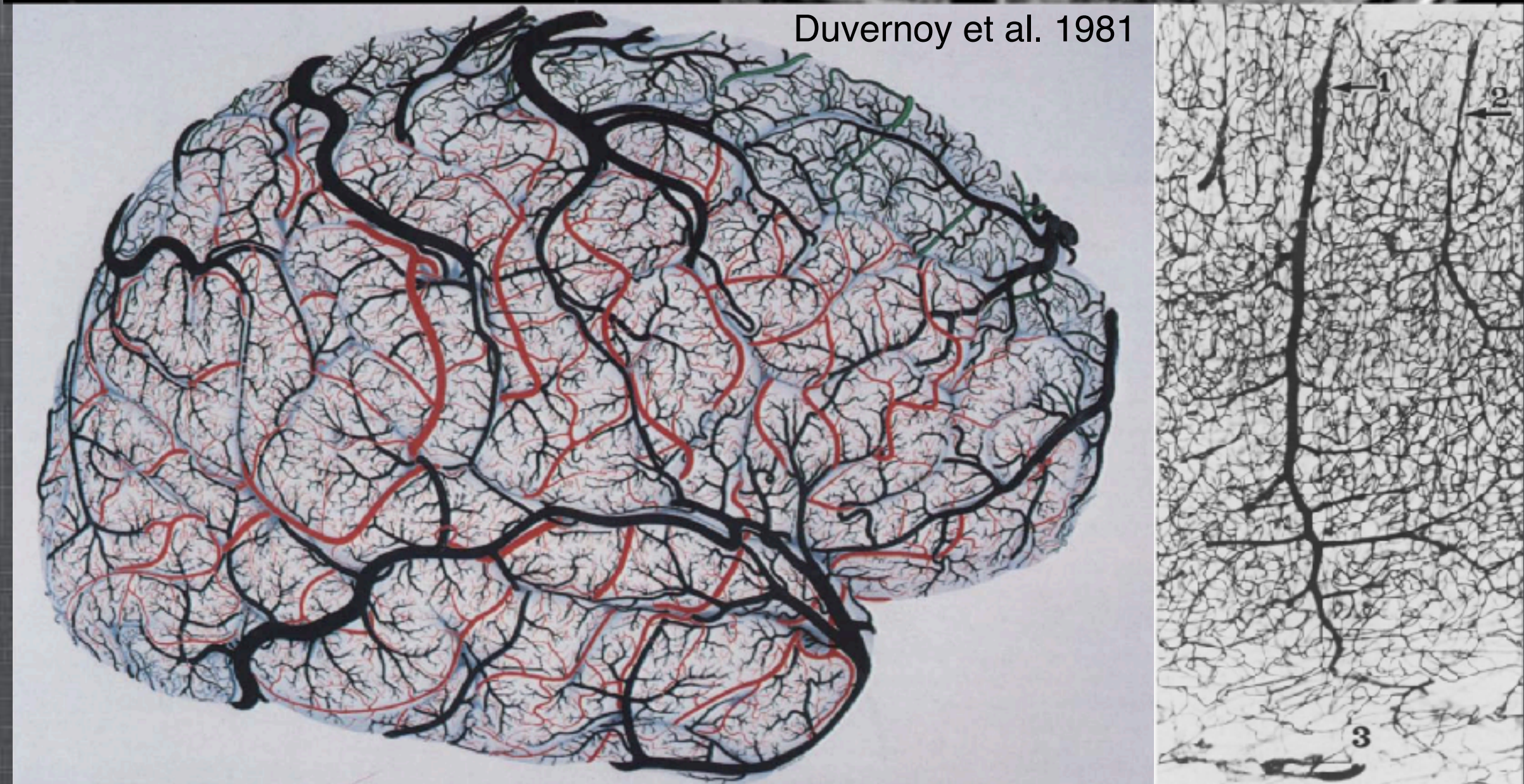
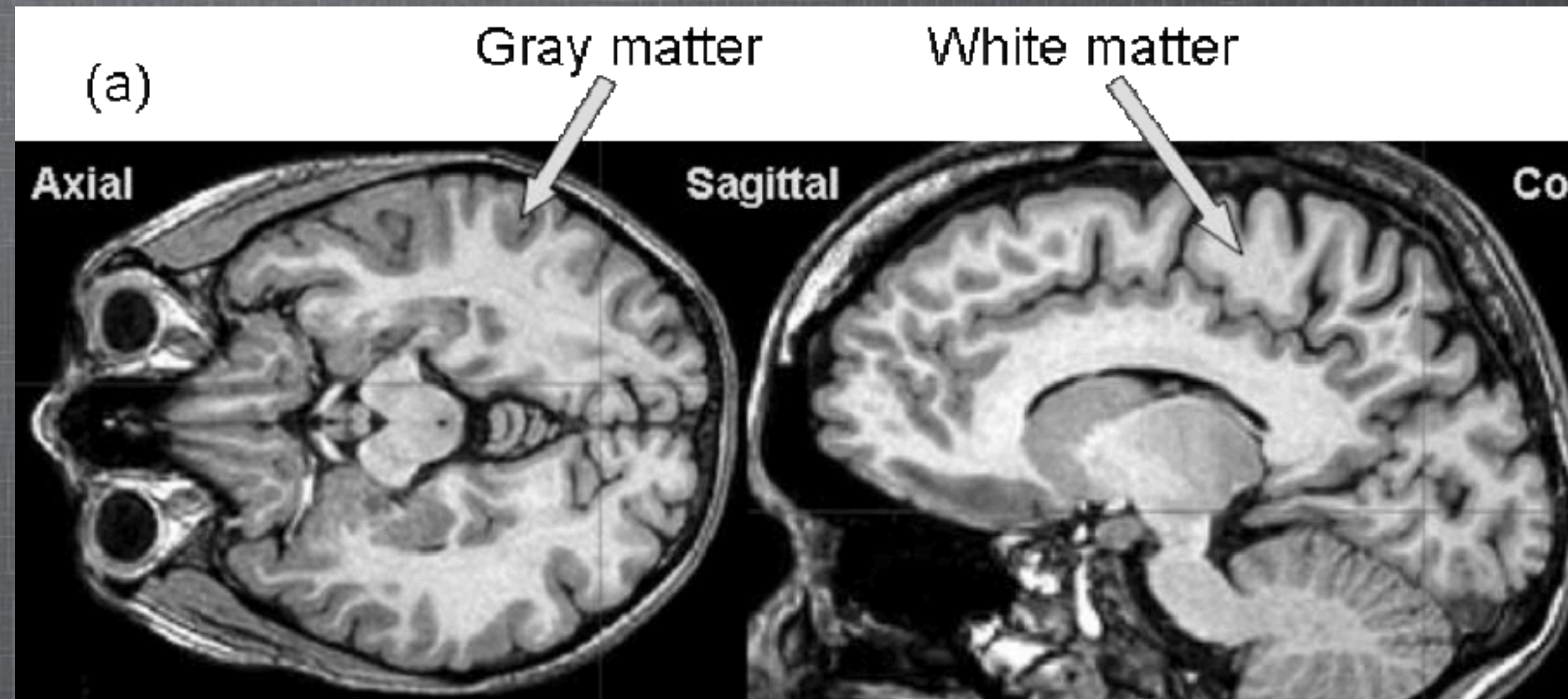


Sound recording of a single neuron
(from Faculty of Life Sciences, University of Bristol)

Edgar Adrian used the newly invented vacuum tube (valve) amplifier, which had been developed for radio technology. He placed an electrode in a sensory nerve fiber connected to a muscle in a frog's leg. When he stretched the muscle, he recorded the electrical “pops” that became known as “spikes”

THE BIOLOGICAL BRAIN

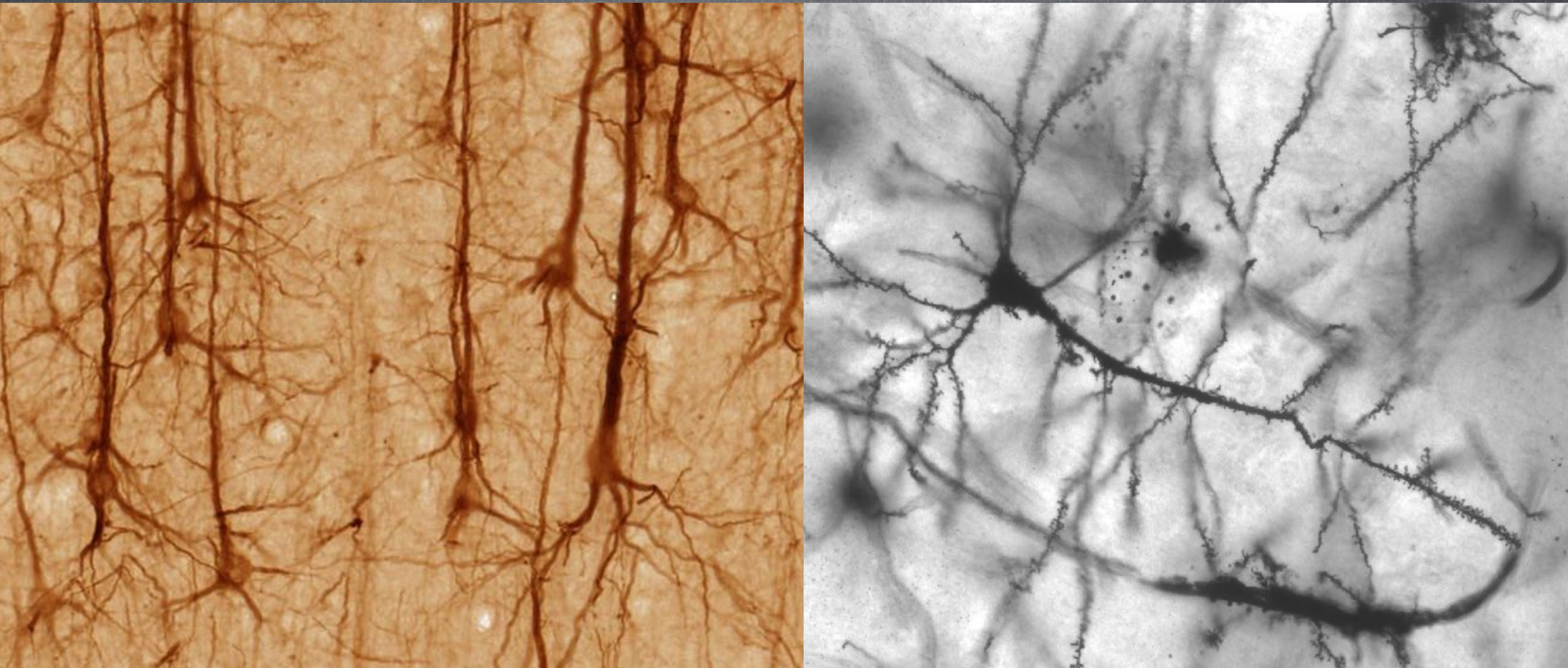
- Your brain weighs around 1.3kg and is roughly 60% fat, with the remaining 40% composed of water, protein, carbohydrates, and salts
- The brain contains ~86M **neurons** a similar number of non-neural cells, blood vessels etc.
- Cortical grey matter is just 2-4mm thick
- Like all other organs the brain is primarily driven by metabolising glucose with oxygen
- Whilst the brain accounts for only 2% of total body weight, it consumes roughly 20% of the daily energy budget

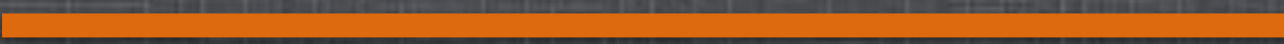


Blood vessels in the brain: arteries (red), veins (black)

LIGHT MICROSCOPE IMAGES OF NEURONS

Modern photographs use similar staining techniques to Golgi and Cajal

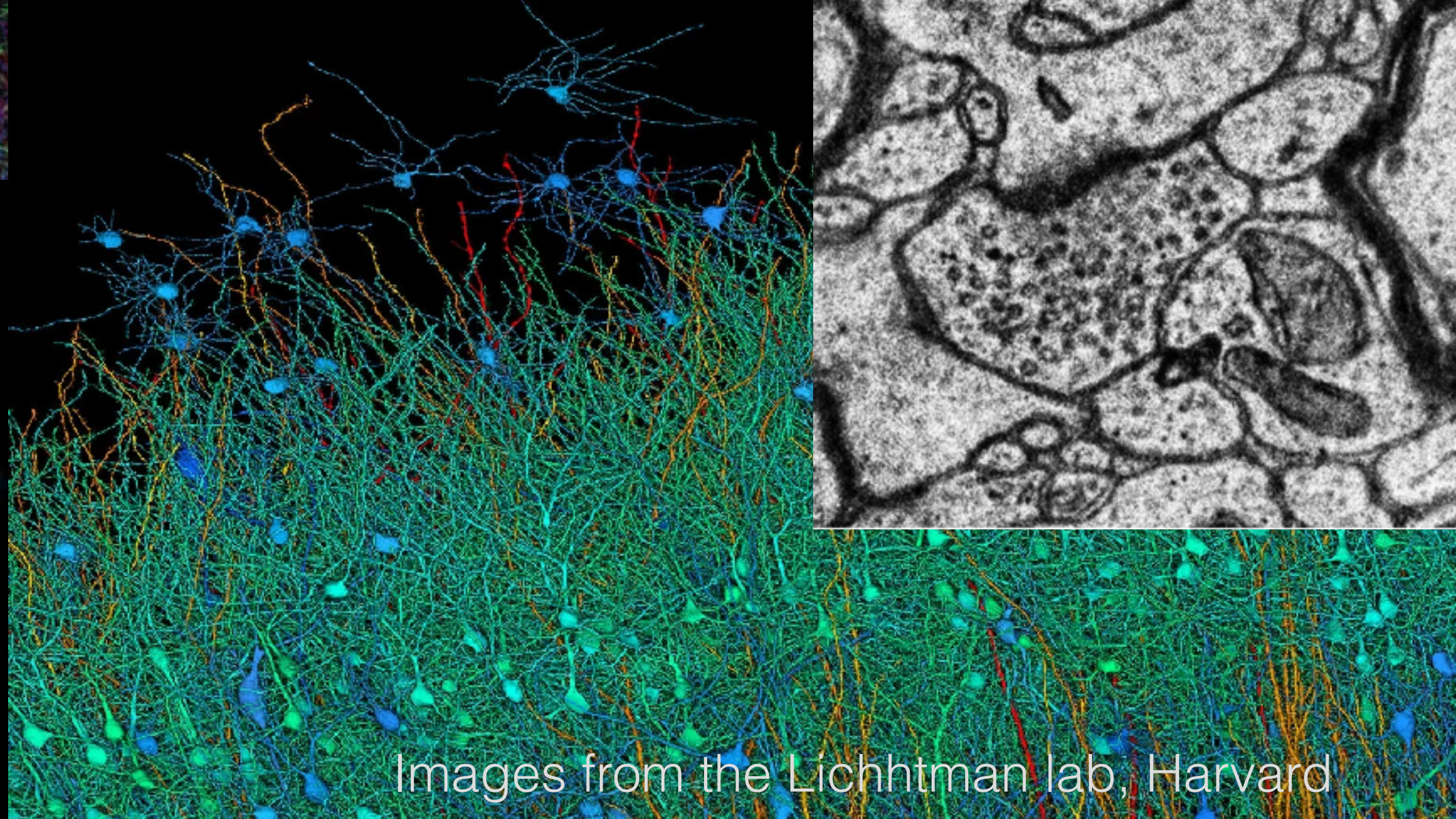
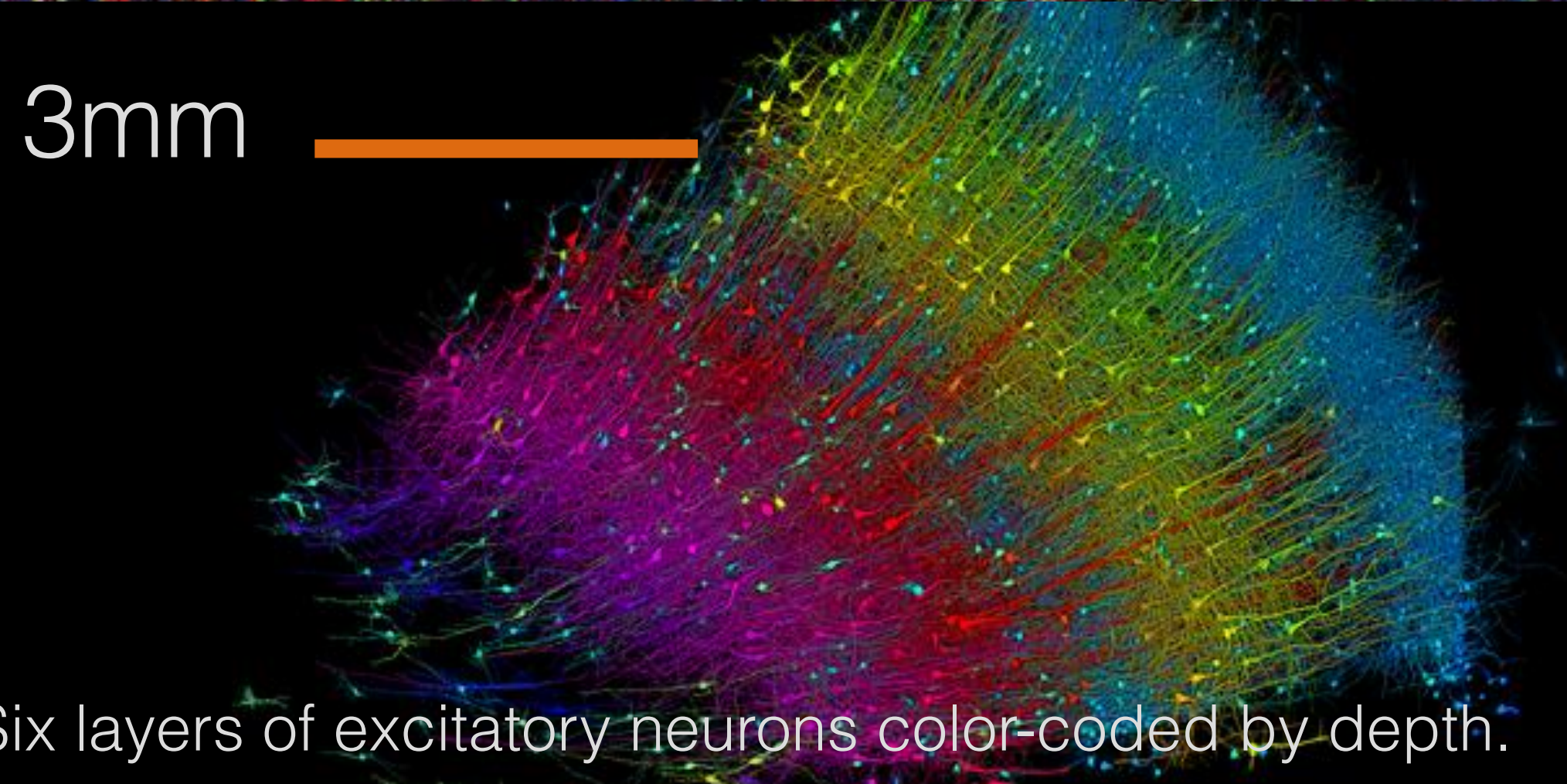
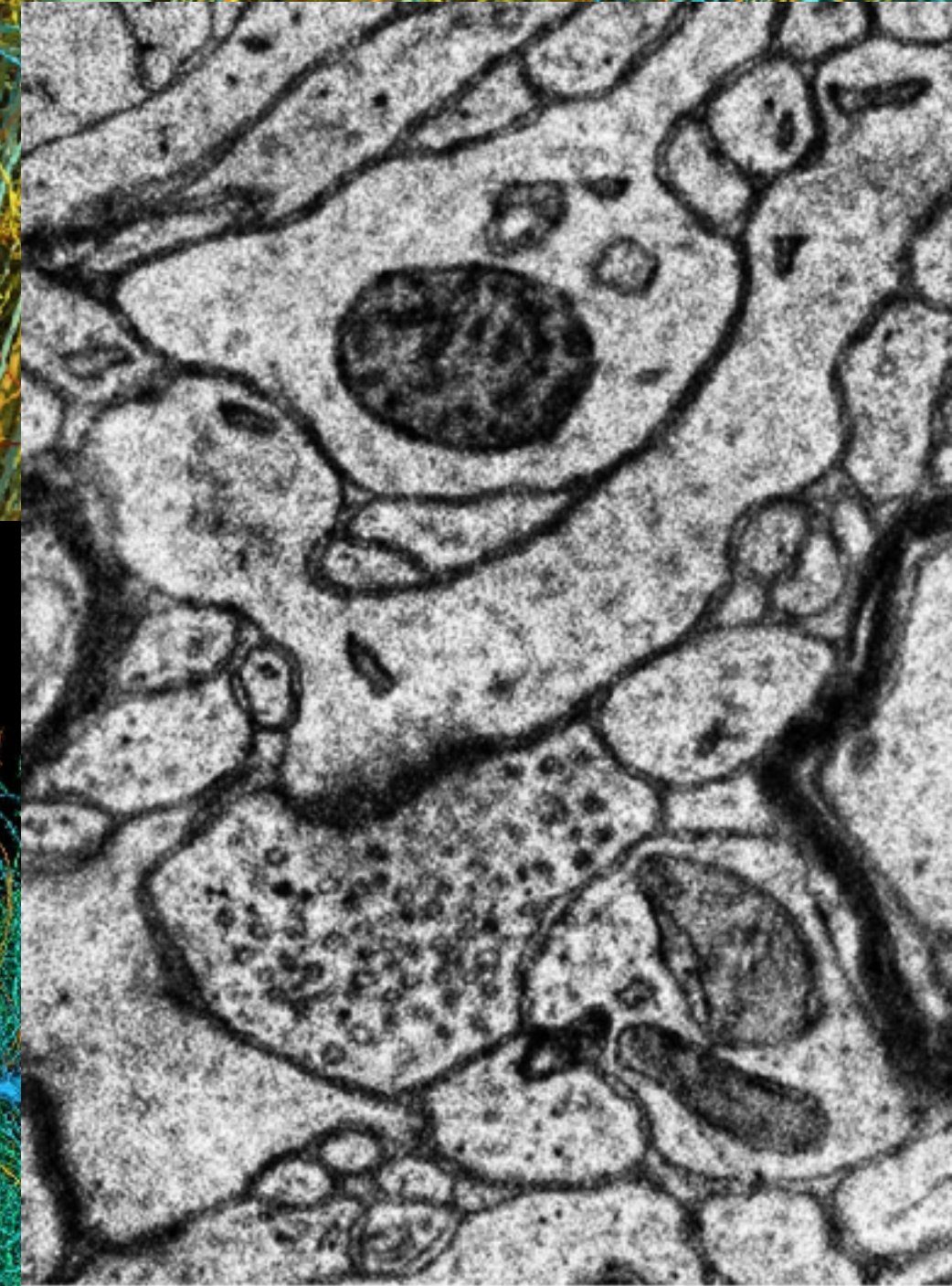
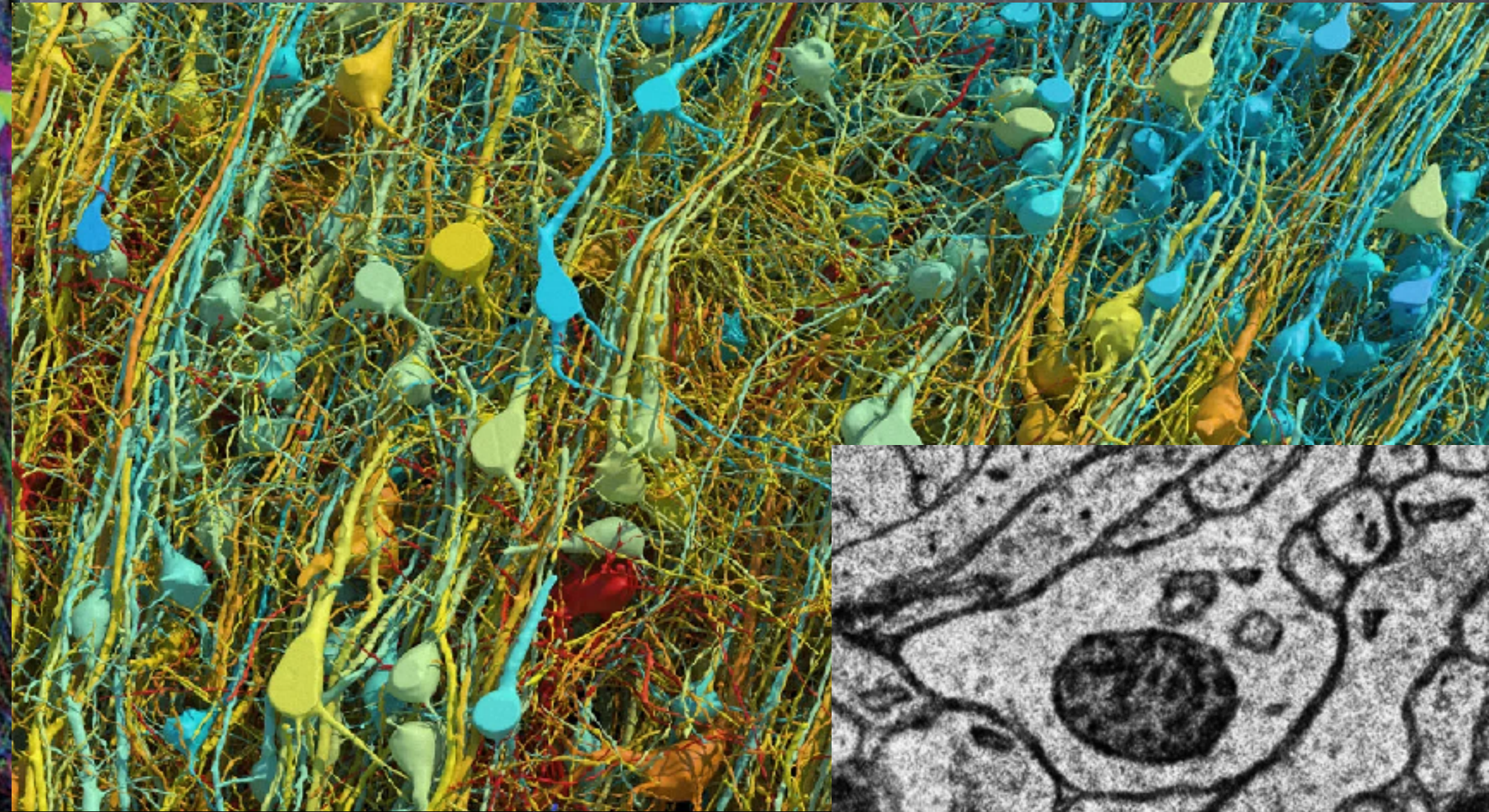
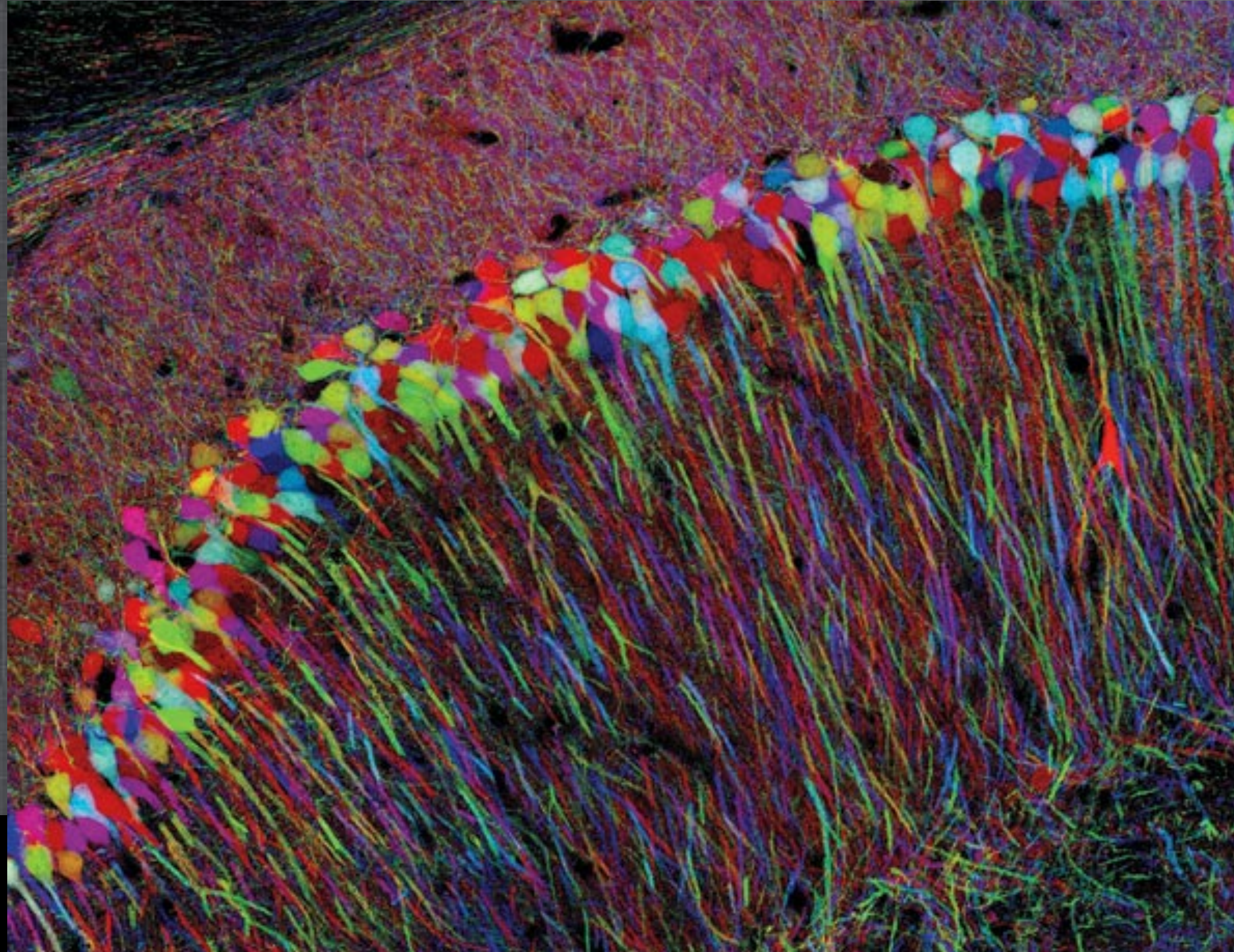


0.5mm 

Neurons are surrounded by a supportive fluid filled extracellular space
The Golgi stain used for these photographs picks out only 1-3% of neurons

ELECTRON MICROSCOPE RECONSTRUCTIONS OF NEURONS

Electron microscopy shows how densely-packed the brain actually is

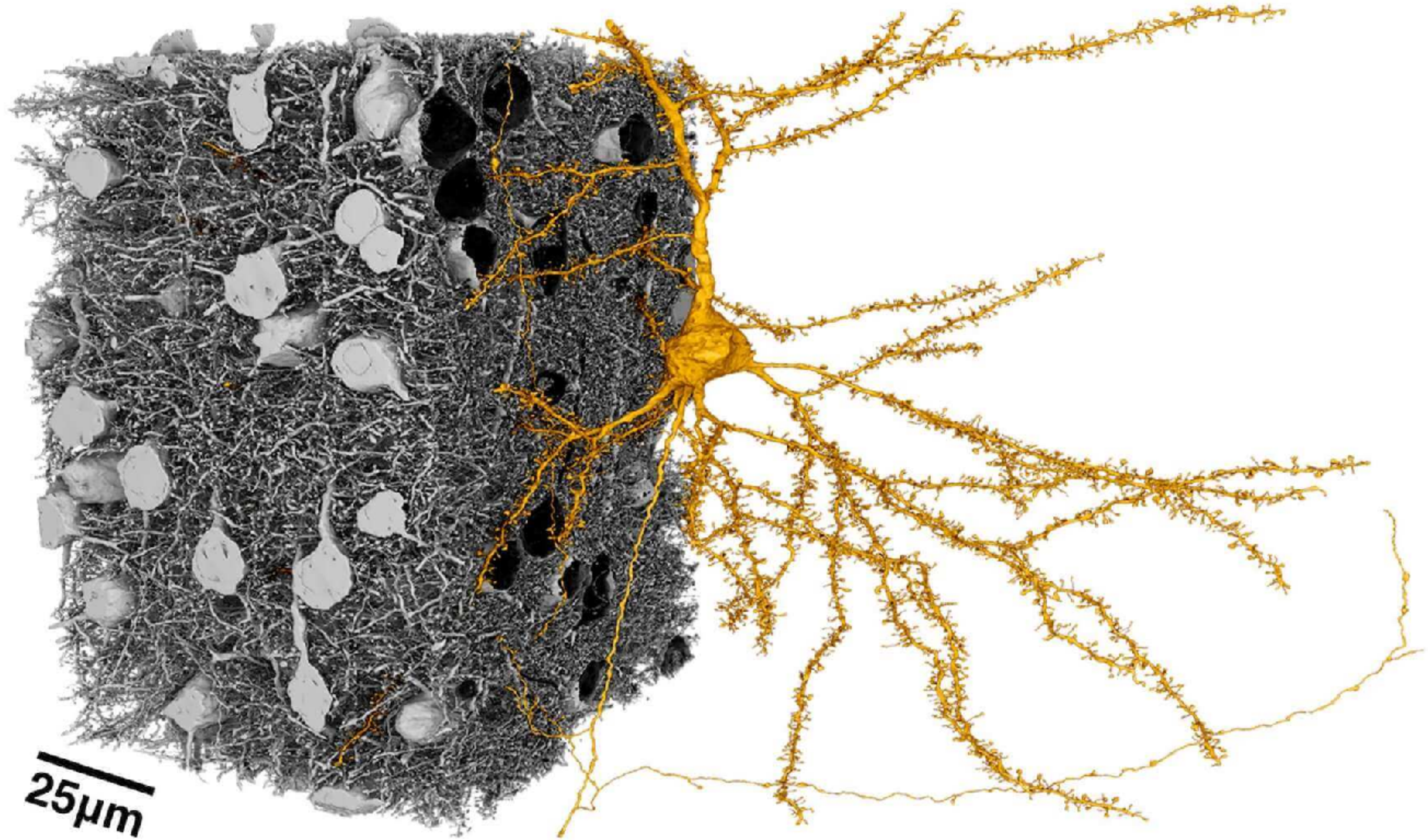


3mm

Six layers of excitatory neurons color-coded by depth.

Images from the Lichtman lab, Harvard

ELECTRON MICROSCOPE 3D RECONSTRUCTION OF MOUSE CORTEx

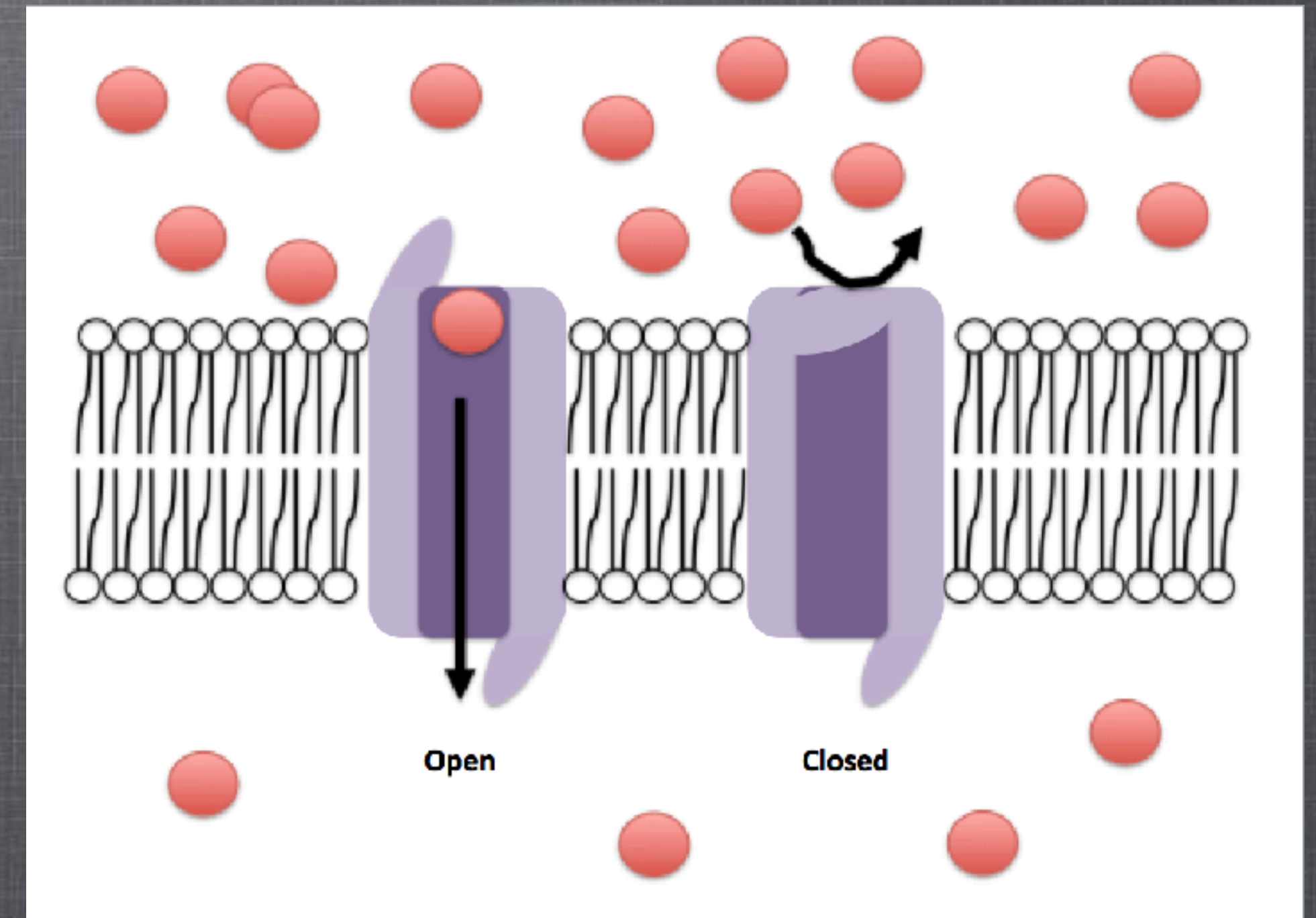


Neurons in mouse cortex

Image from the Allen Institute

ELECTRIC CURRENT IN A NEURON

Each neuron has a “resting” membrane potential of -0.07 V (70 mV), sodium, potassium and chlorine ions move through the membrane via ion channels



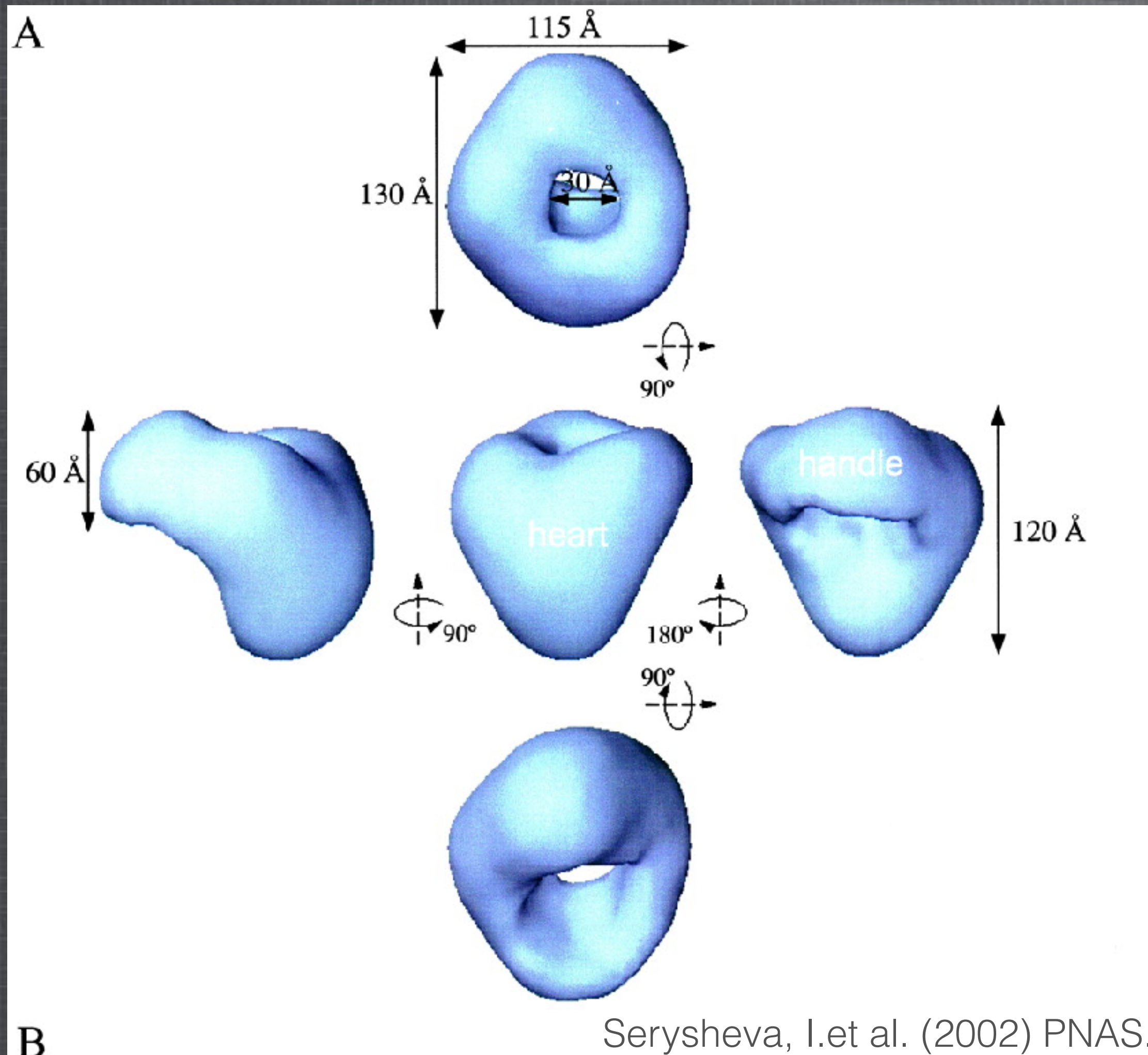
The neuron is full of large **negatively-charged** molecules that can't get out, but it lives in a soup containing **positively-charged sodium ions**, this creates the -70 mV electrical potential

Neurons also have a **chemical gradient** across the cell membrane related to concentrations of ions. This is maintained by molecular ion “pumps”

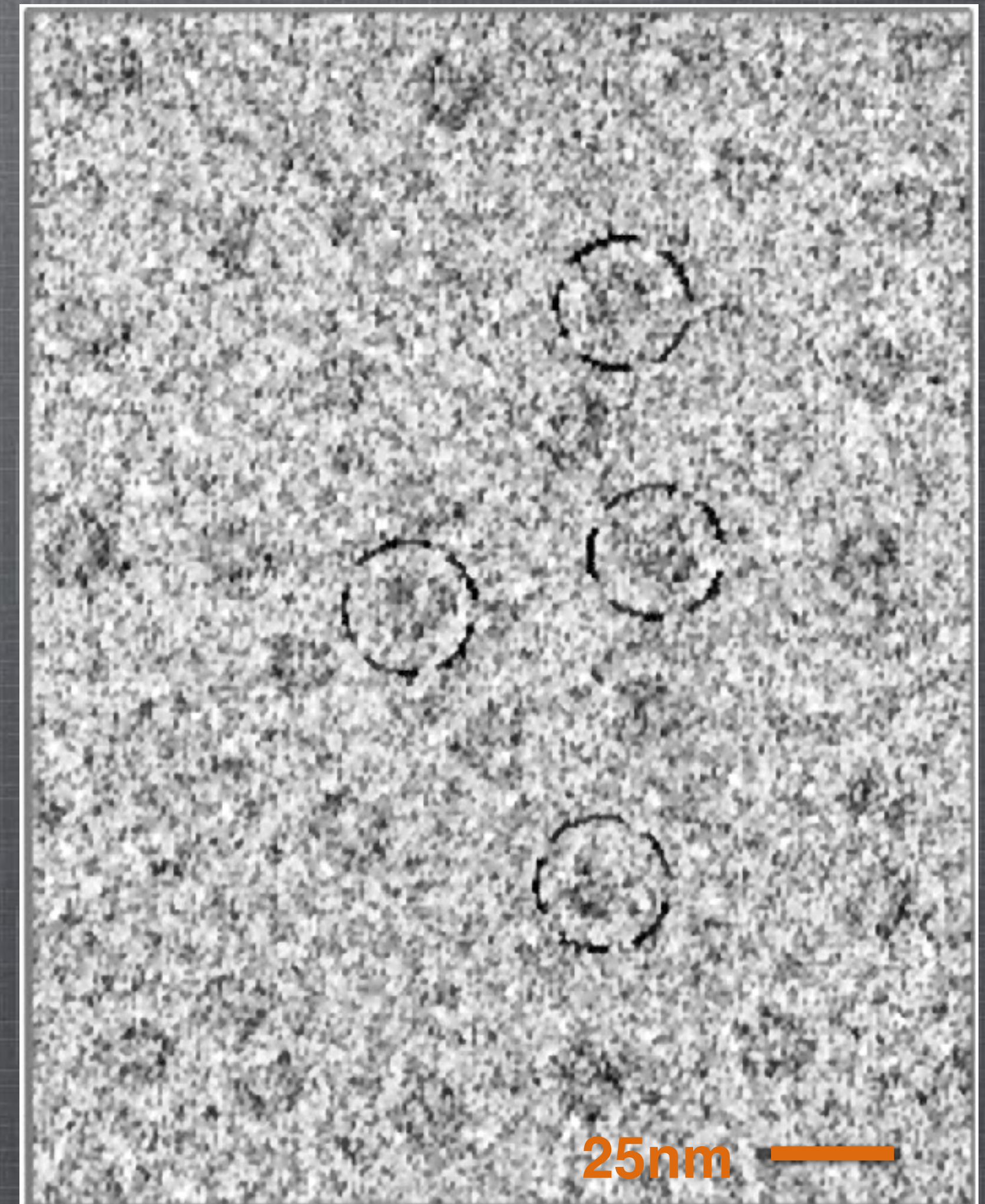
Movements of ions across the membrane depend on both the electrical and chemical gradients and is regulated by the opening and closing of **voltage-gated** ion channels. A typical neuron might have around 10 million ion channels

ION CHANNELS IMAGED BY CRYO-ELECTRON MICROSCOPY

Electrical activity in ion channels was recorded in the '70s, and they were imaged from the late '90s.



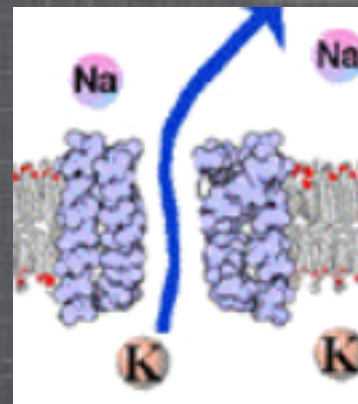
Reconstructed shape of a Calcium channel, the channel is composed of protein



Electron micrograph of ice-embedded voltage-gated L-type Ca^{2+} channels. Some individual channel particles are circled. (Bar = 250 Å = 25nm)

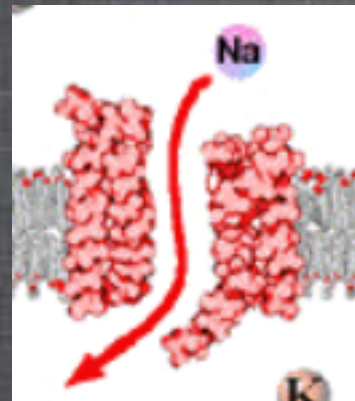
THE ACTION POTENTIAL IS GENERATED BY MOVEMENT OF POTASSIUM (K^+) AND SODIUM (Na^+) IONS ACROSS THE MEMBRANE

"leak"
potassium
channel



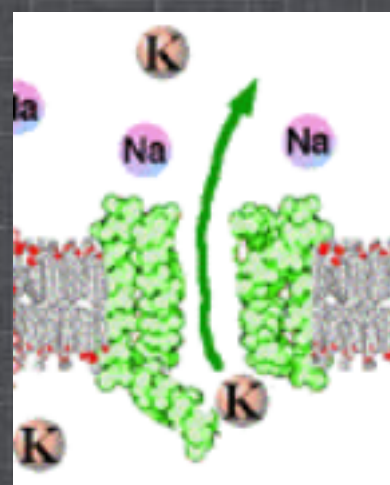
K

voltage-
gated
sodium
channel

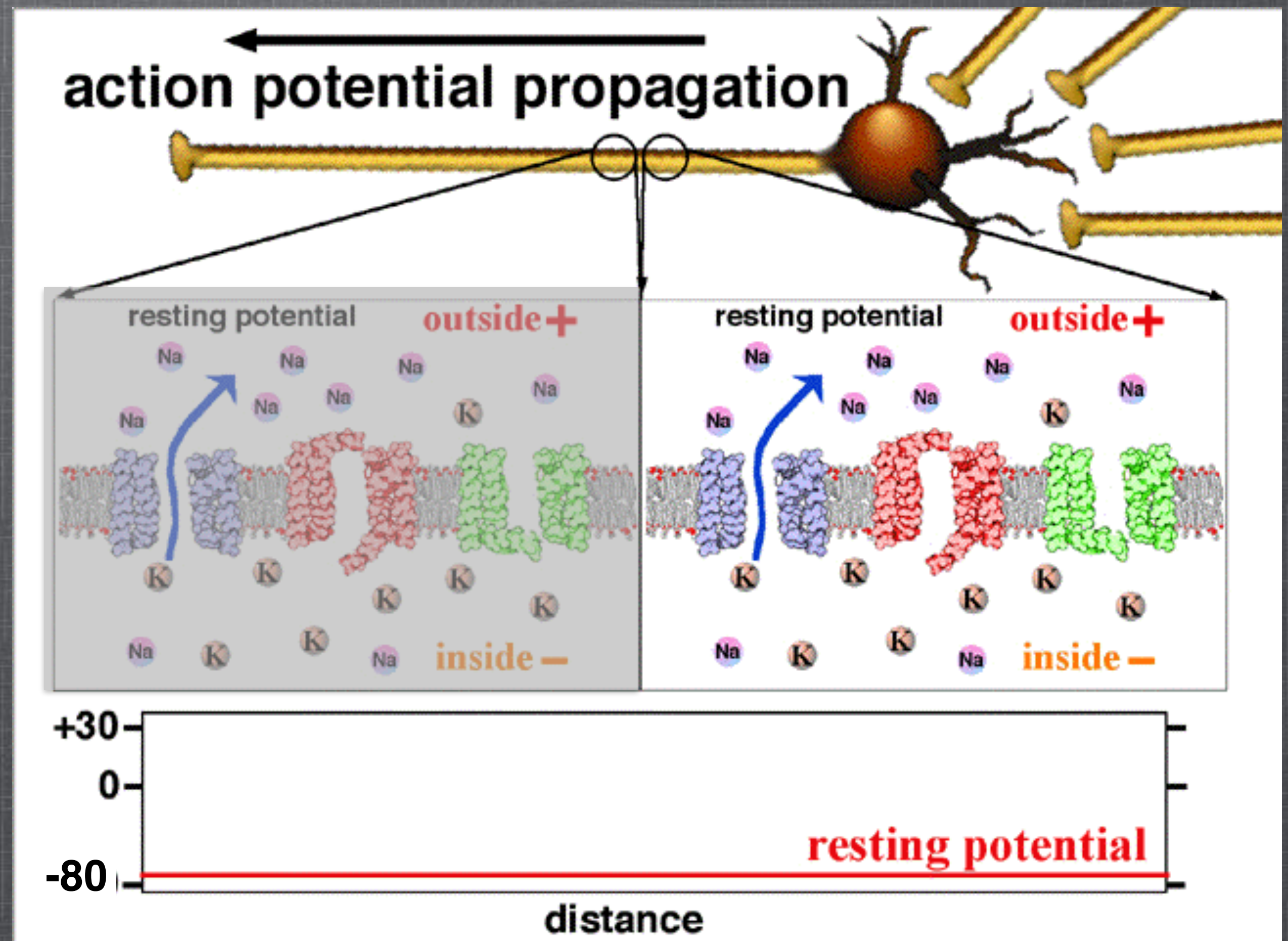


Na

voltage-
gated
potassium
channel



K

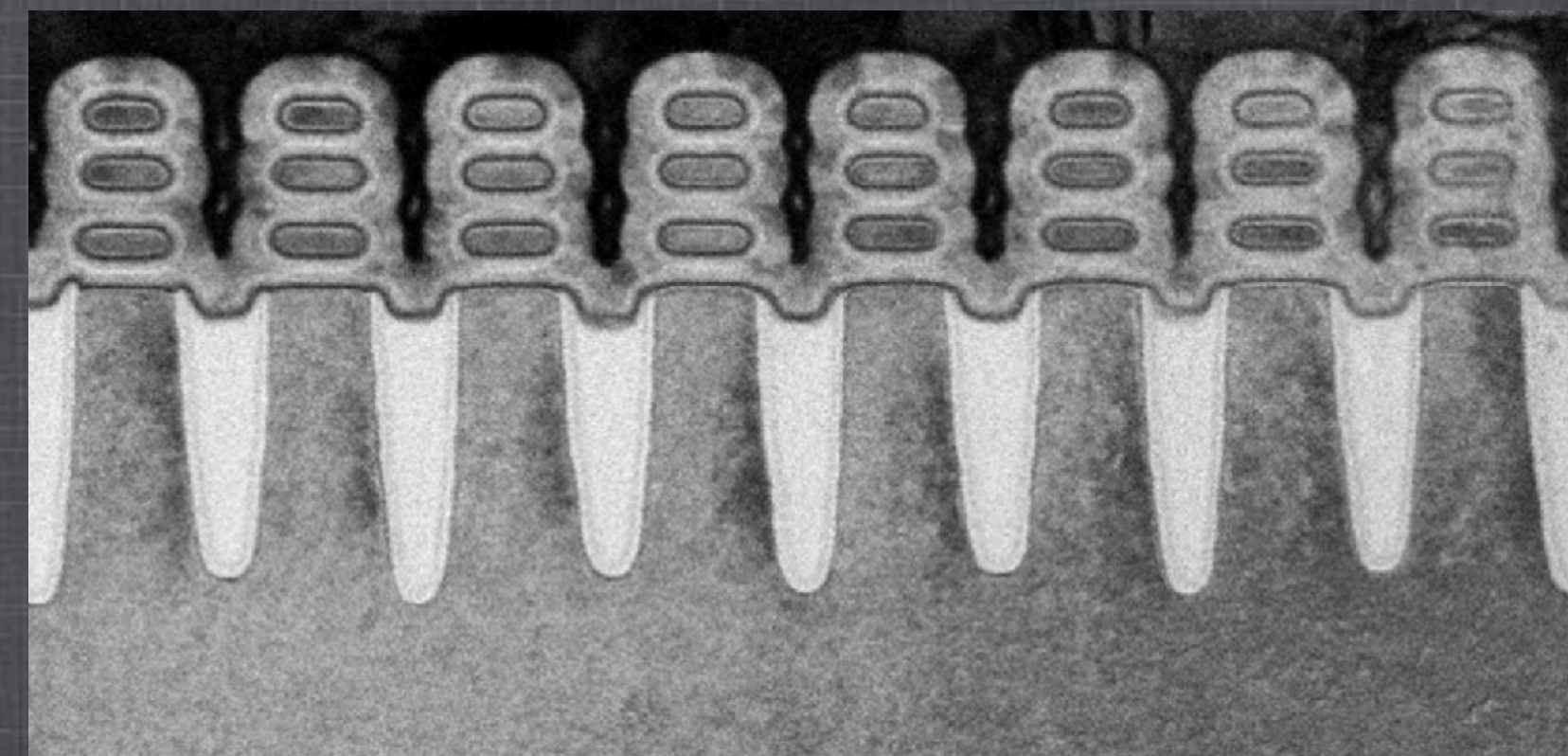
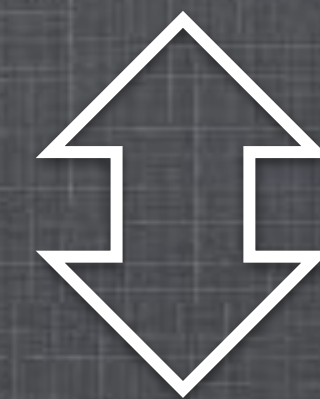
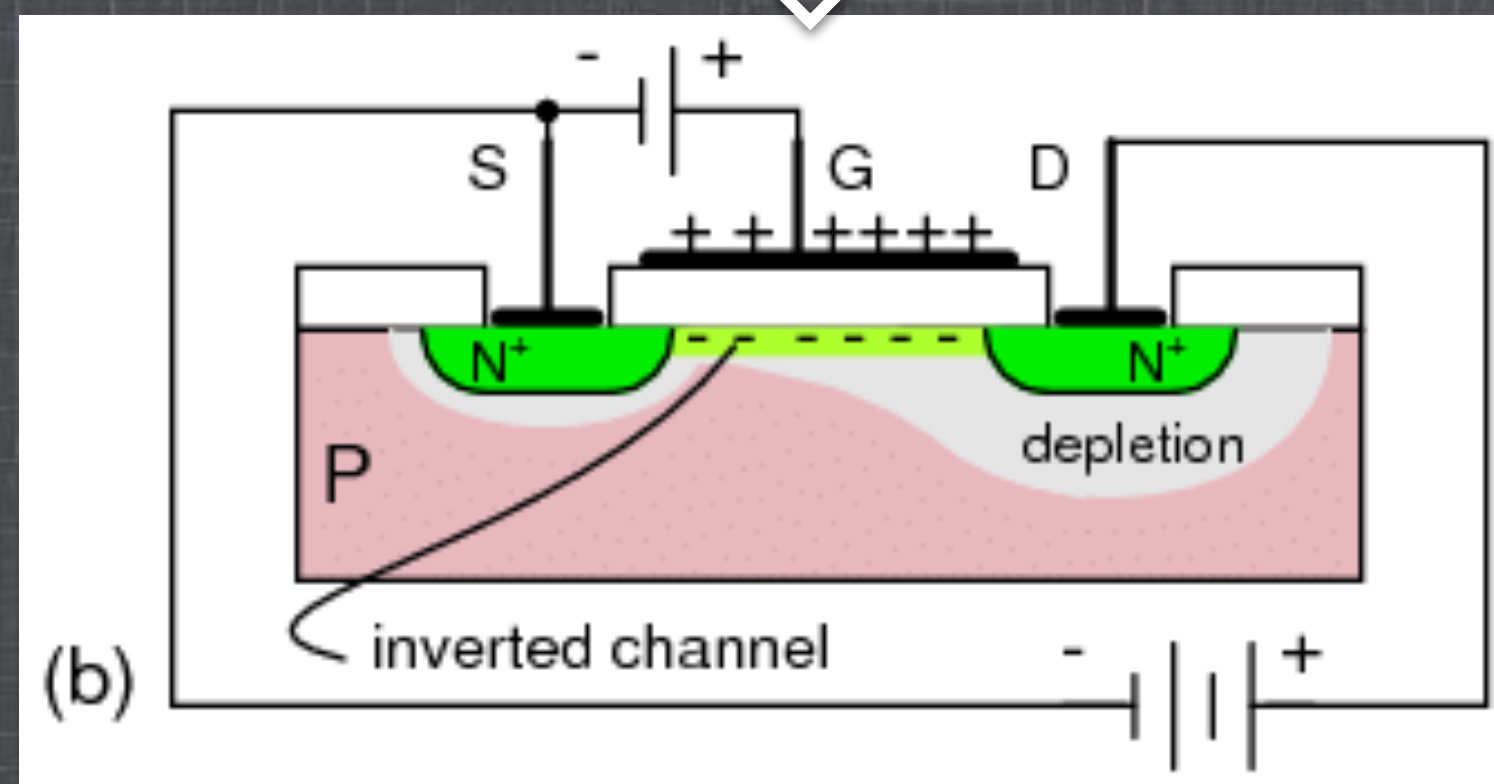
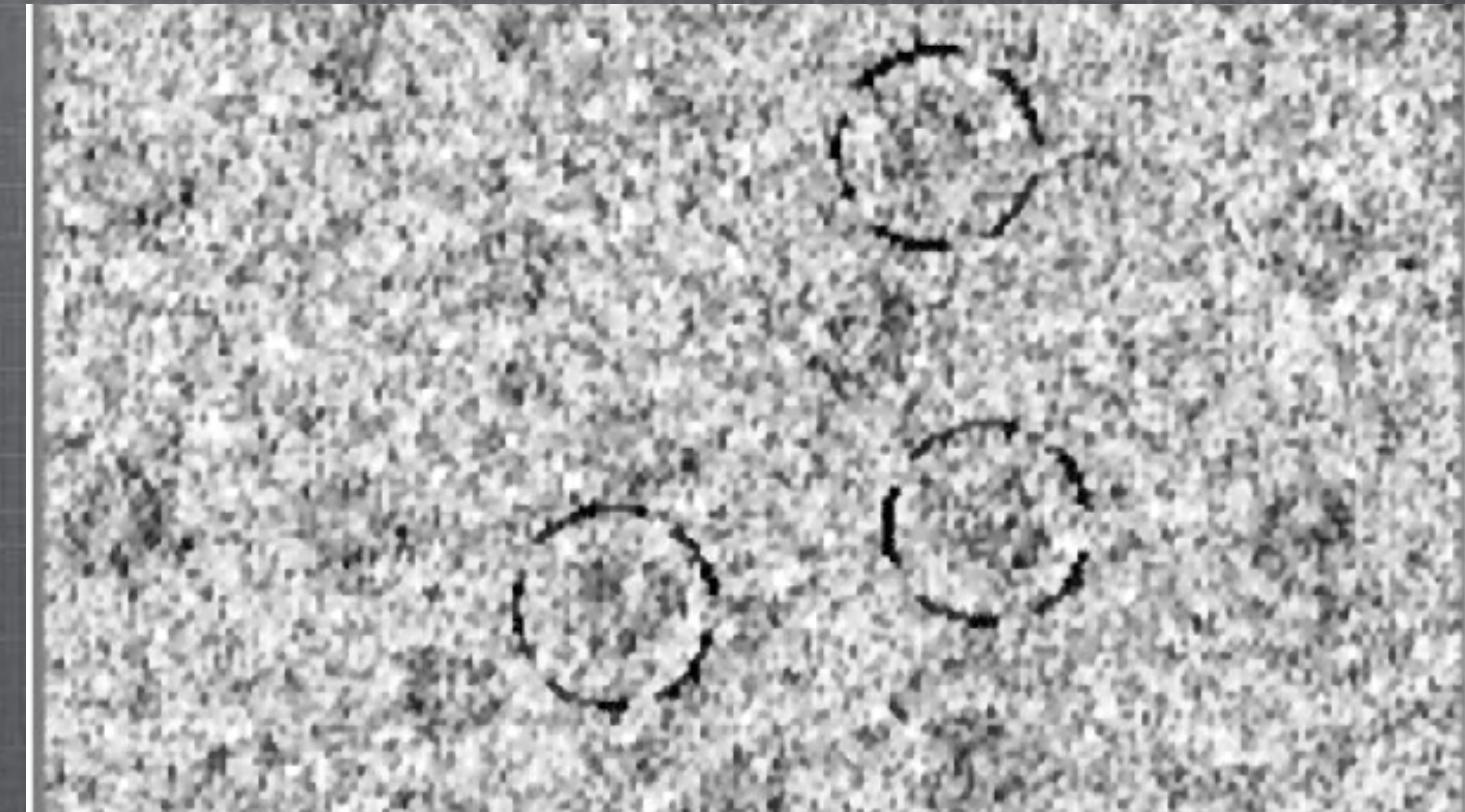
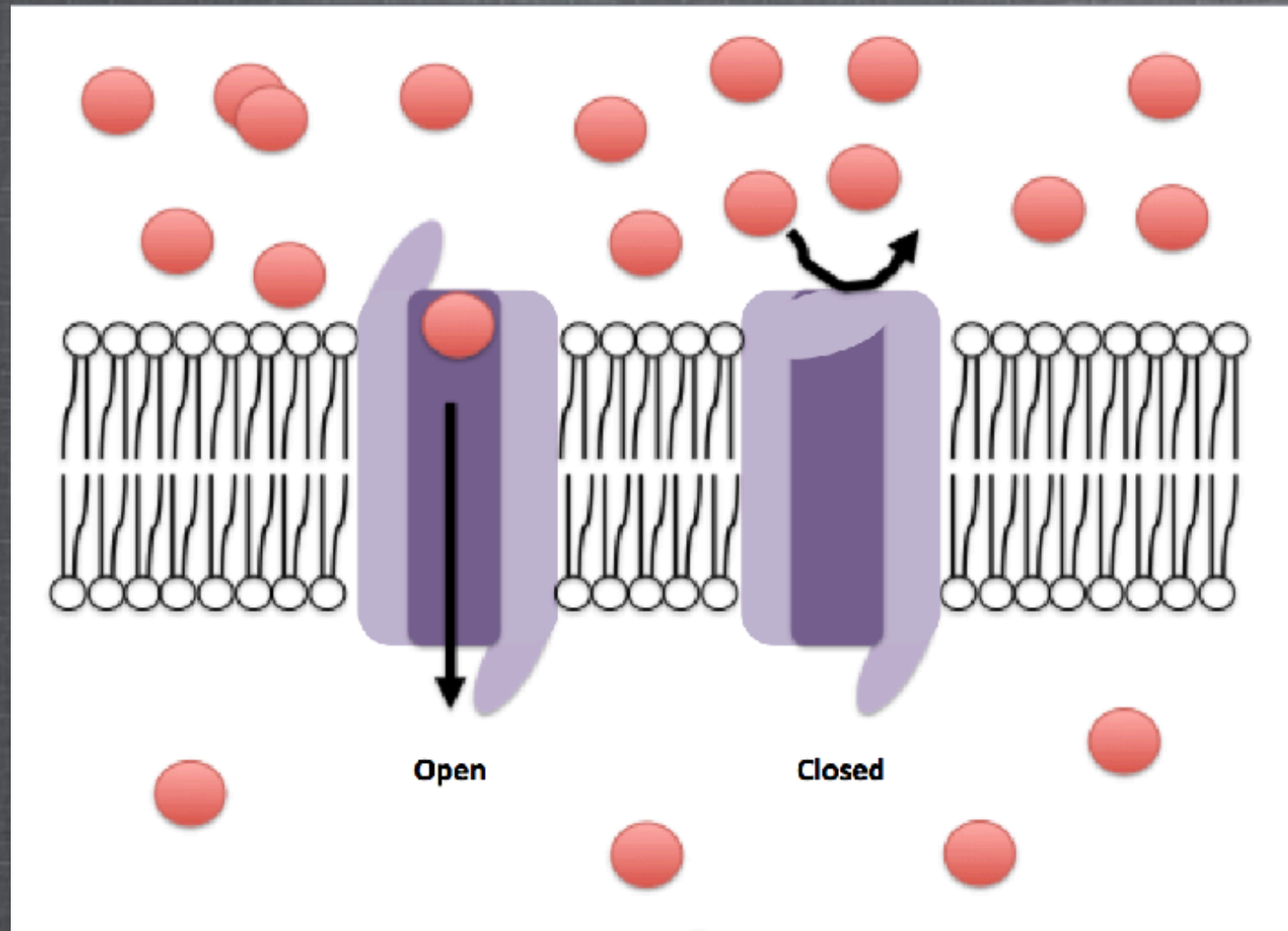


John Schmidt, CC BY-SA 4.0

When the local area of the membrane reaches threshold (-55 mV) the Na^+ gates open and Na^+ moves into the axon. This has a snowball effect and the potential shoots up to +30 mV (depolarisation). 1ms later the K^+ gates open and K^+ moves outside the axon, and the potential drops. The membrane potential overshoots (hyperpolarization) before returning to rest.

ION CHANNELS ARE NATURE'S TRANSISTORS

Both ion channels and transistors operate through the flow of electric charge and can be controlled to be in on and off states



ION CHANNELS AND TRANSISTORS COMPARED

Ion channels are at a similar scale to state-of-the-art transistors

Feature	Ion Channel	State-of-the-Art Transistor
Diameter	~5-10 nm (pore 0.3-0.7 nm)	~3–5 nm
Length	5 nm	~3–5 nm (vertical scaling may vary)
Material	Protein in lipid bilayer	Silicon + metal contacts
Function	Pass individual ions	Switch electrical current in circuits
Speed	Opens/closes in μs –ms	Switches in picoseconds (10^{-12} s)
Voltage	70 mV (0.07 V)	0.7-0.8 V

Cell metabolism powers ion channels, external power supplies are needed for transistors

HOW MANY IPHONES TO MATCH A BRAIN?

Apple's latest chip the A19 pro has 28 processors, 25 billion transistors (the A14 has 11.8 billion) with clock speeds of 1.8-4.25GHz



HOW MANY IPHONES TO MATCH A BRAIN?

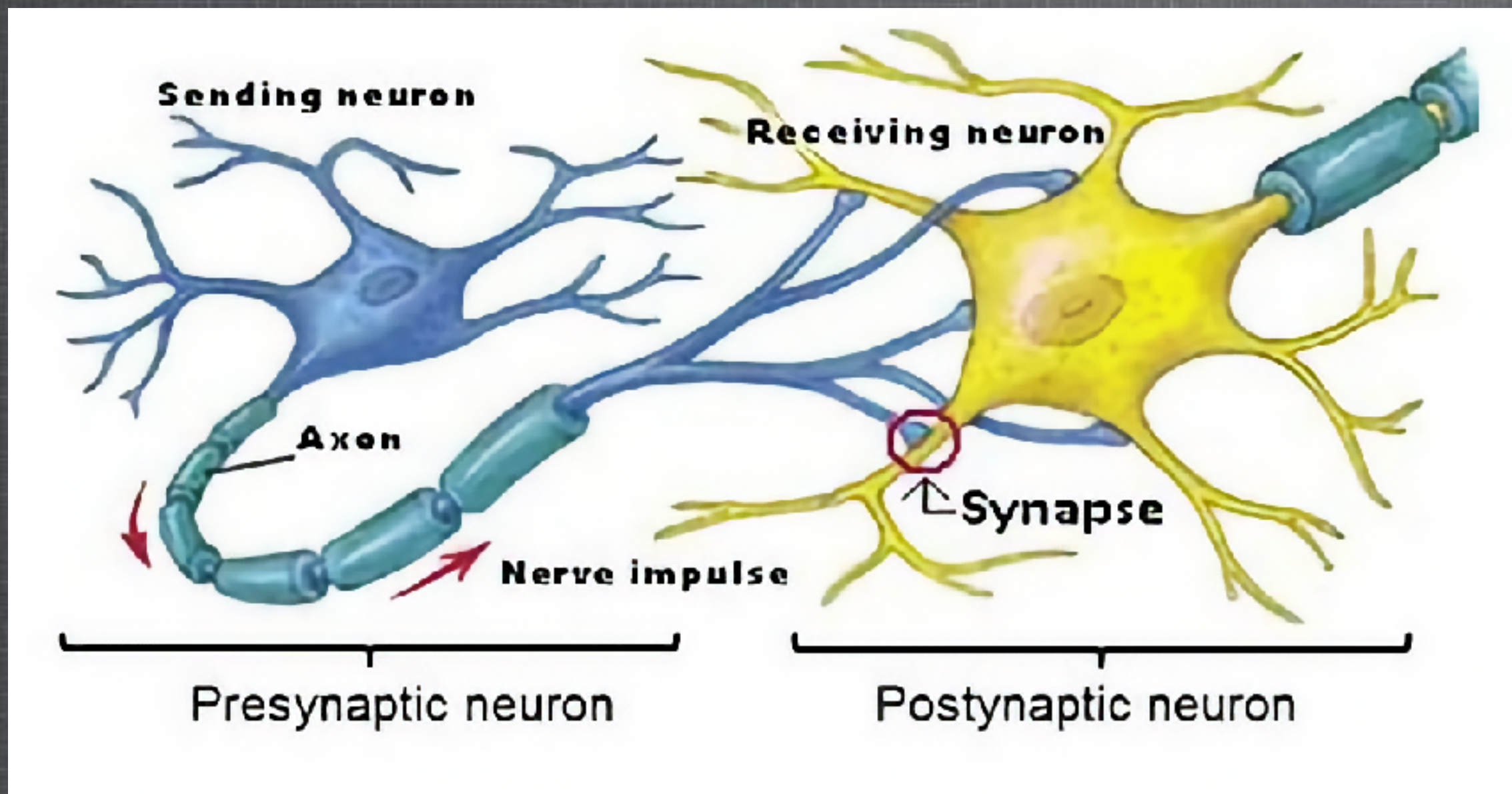
	<i>Neurons</i>	<i>Switches/neuron</i>	<i>Total switches</i>	<i>Switches/s</i>	<i>Switch capacity</i>
Human brain	8.60E+10	1.00E+07	8.60E+17	1.00E+03	8.60E+20
A19 chip			2.50E+10	3.00E+09	7.50E+19
			A19s needed to match the brain		11.47

- Apple's latest chip the A19 pro has 28 processors, 25 billion transistors (the A14 has 11.8 billion) with clock speeds of 1.8-4.25GHz
- The brain has 86 billion neurons if each has 10 million channels, then it has roughly 860 quadrillion (a quadrillion is 1,000,000 billion) molecular switches, but switch speed is less than 1,000Hz
- Allowing for the much faster switch frequency of transistors you could emulate the switch capacity of the brain with just 12 A19 chips!
- BUT THIS DOES NOT MATCH THE BRAIN FOR PROCESSING POWER because all iPhones still have the **von Neumann bottleneck**
 - Most transistors are unused (idling) most of the time ("dark silicon")
 - More than 99% of energy use is in moving data to and from the processor rather than on computation itself, and moving data is slower than processing



SYNAPSES

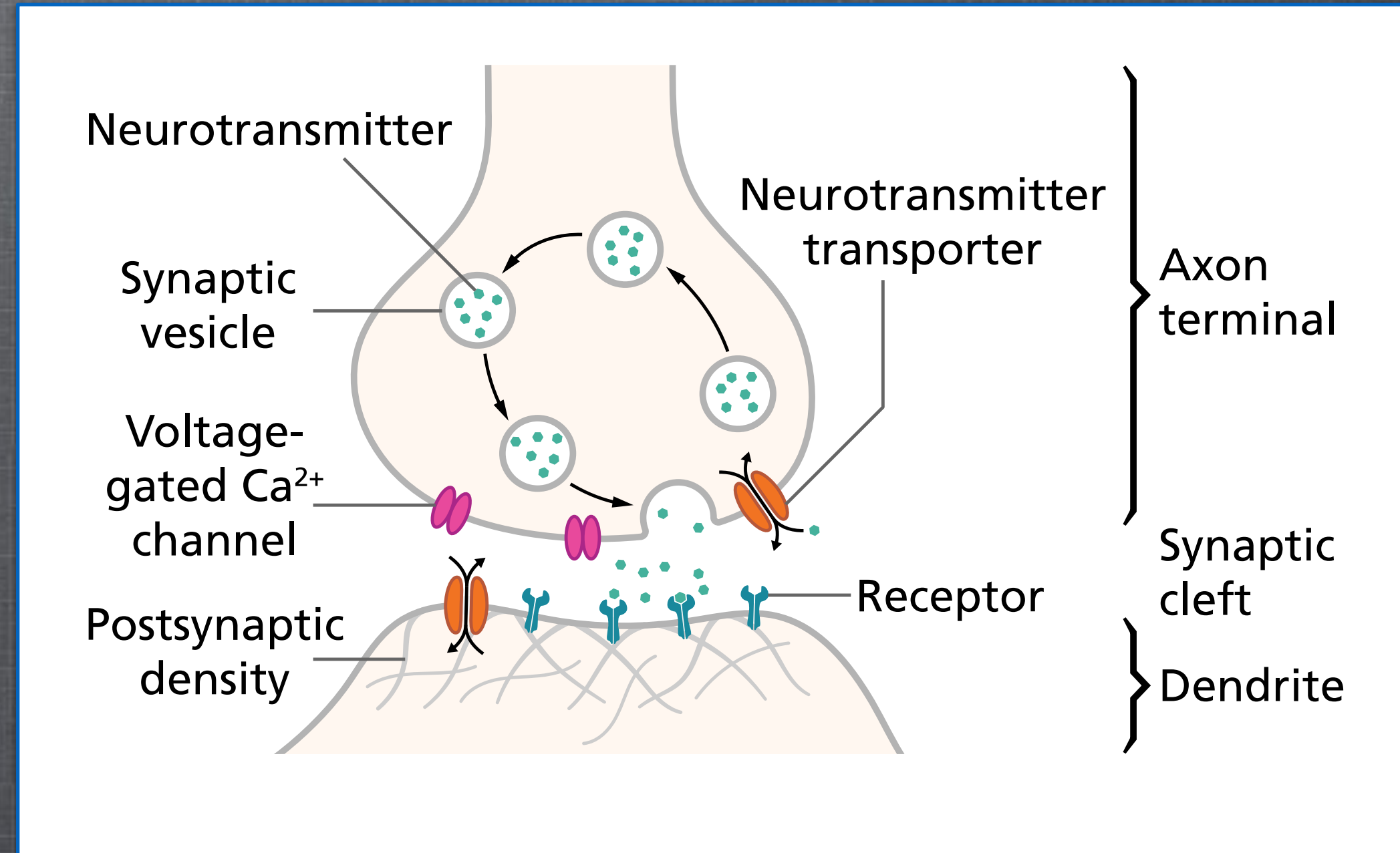
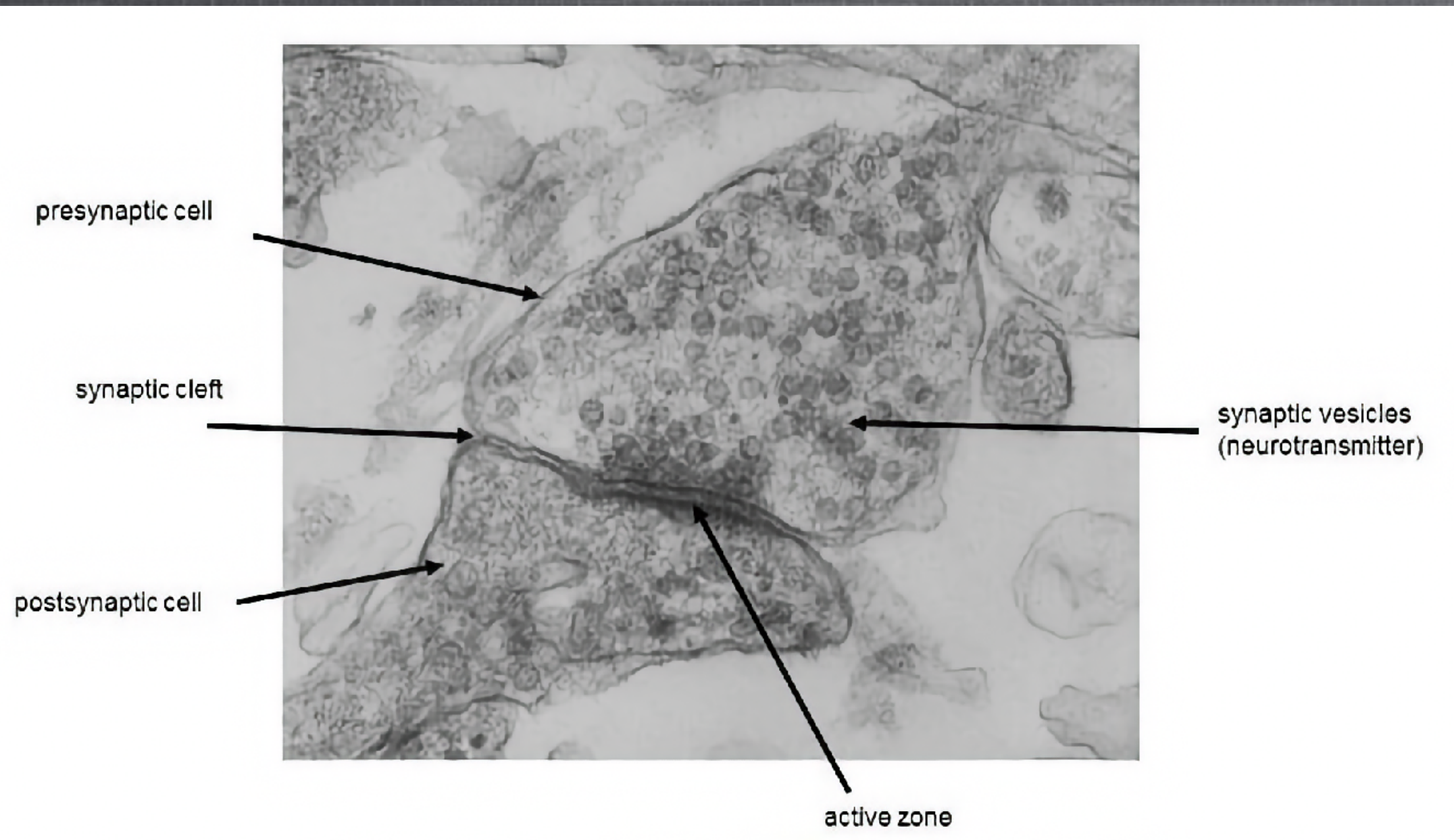
Synapses are the principal site of learning and memory in the brain



Signals are passed between neurons at **synapses** by chemical messengers called **neurotransmitters**. These can be excitatory (e.g. glutamate) or inhibitory (e.g. GABA). Synapses are also a critical site for **learning**

1954, VISUALISING THE SYNAPSE

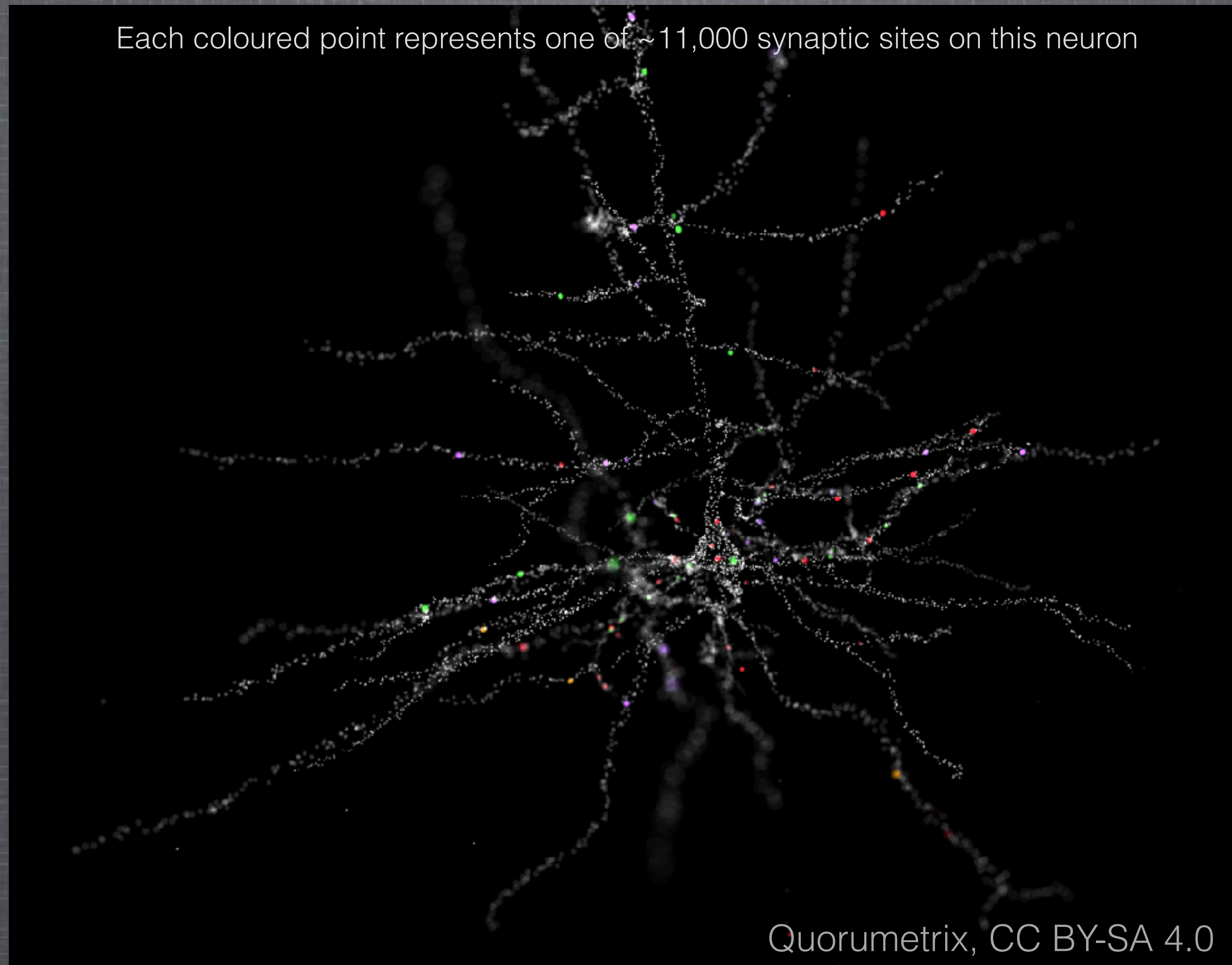
Because the synaptic gap is only about 20 to 40 nanometers it was physically impossible to see it using light. Although predicted and named in 1897 (by Charles Sherrington), the synapse remained a "theoretical ghost" until it was observed using an electron microscope in the 1950s



When the first high-resolution images of the synapse were published, they didn't just show a gap; they saw the synaptic vesicles—tiny bubbles filled with chemicals (neurotransmitters) waiting to be released. This proved that the signal wasn't just electrical; it was electrochemical. The action potential triggers a mechanical process that pushes neurotransmitter into the gap where it binds with receptors on the other side.

SYNAPSES ON A CORTICAL NEURON

Each coloured point represents one of ~11,000 synaptic sites on this neuron

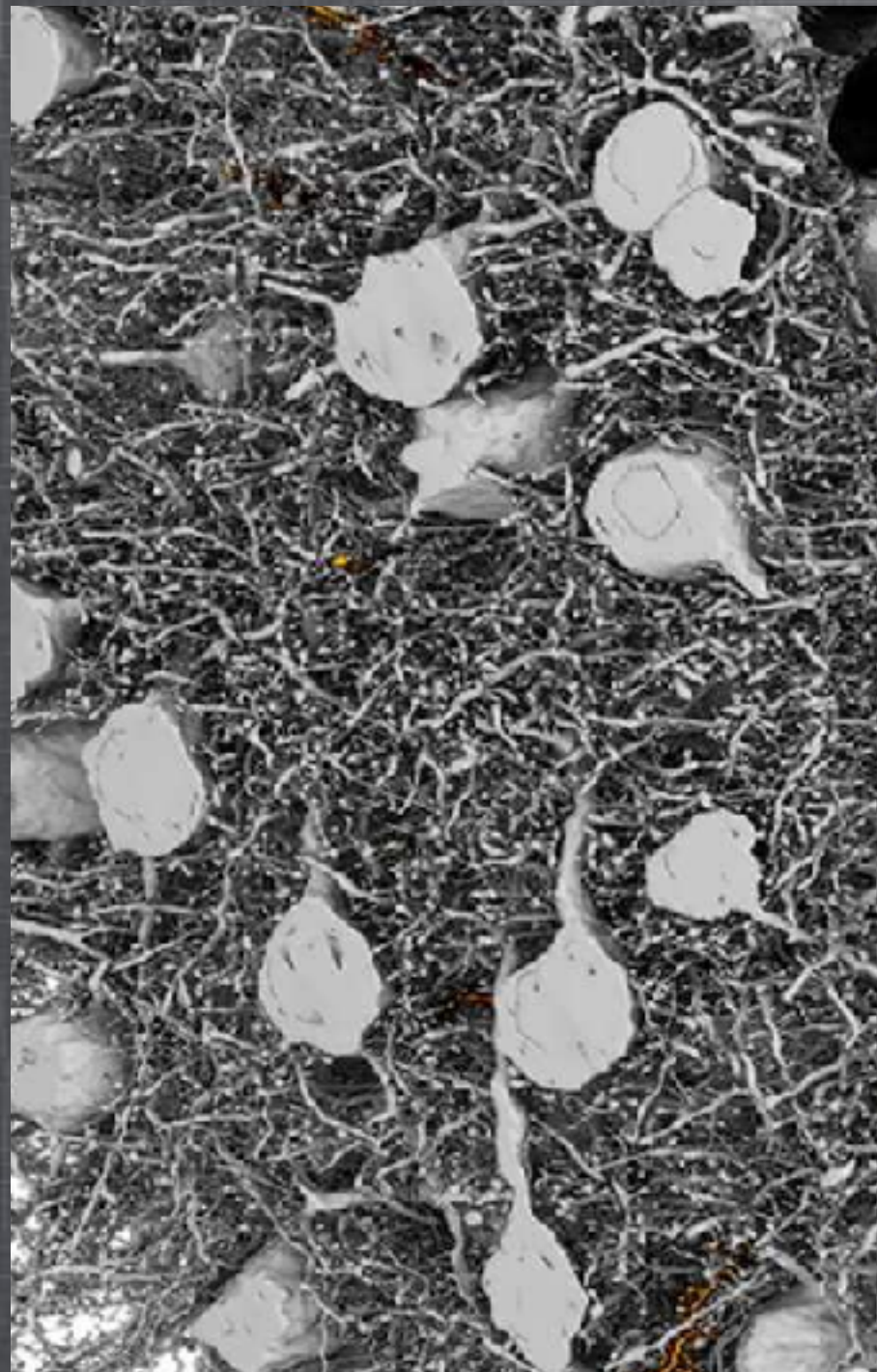


Quorumetrix, CC BY-SA 4.0

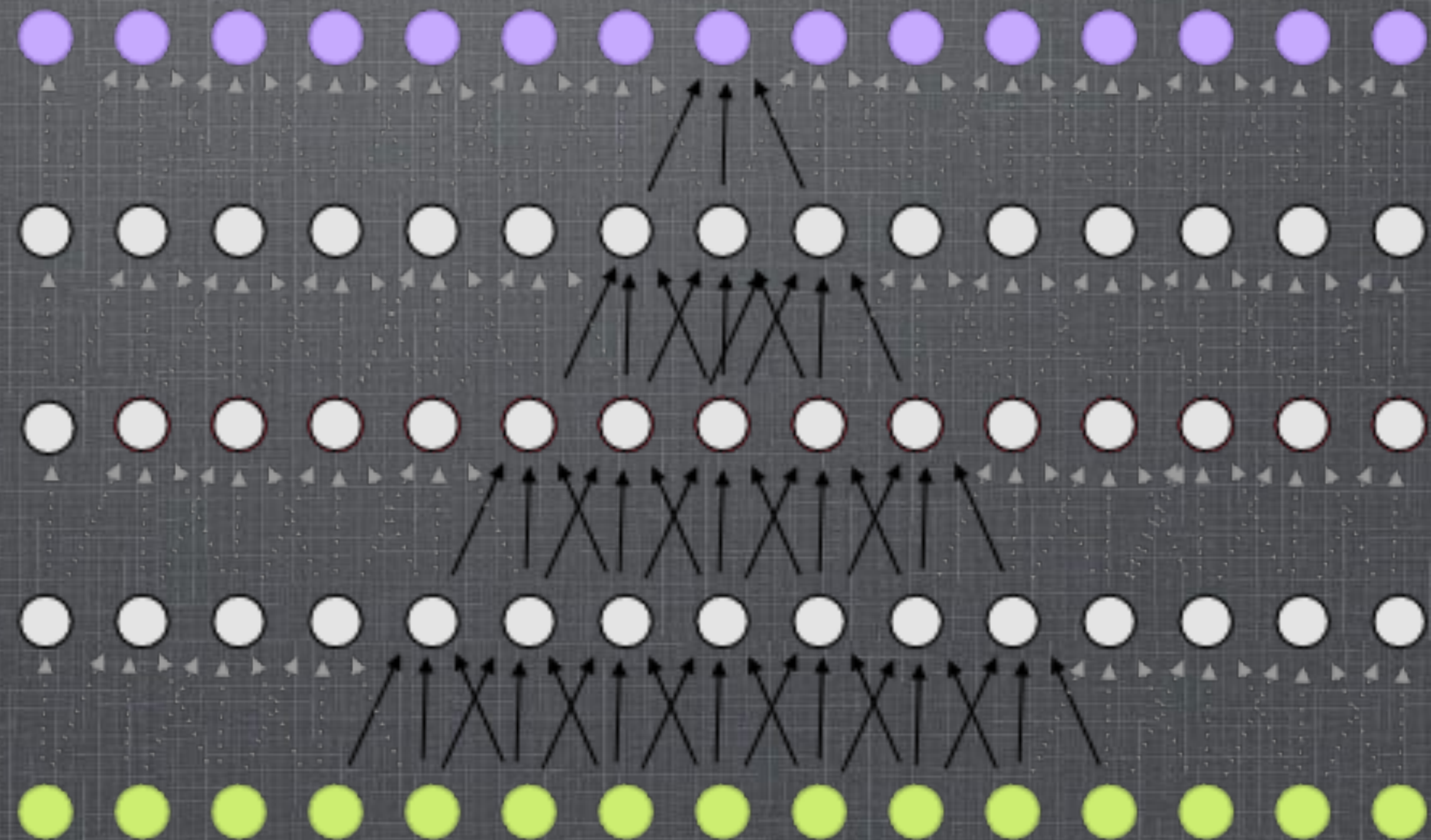
Transmission at the synapse turns the sharp digital spike of the action potential (1ms) into a graded **postsynaptic potential** that can last 10-50ms. This acts to integrate the incoming spike pattern. The signal in the dendrite travels passively and slowly ($\sim 0.5\text{m/s}$) to the cell body and can lose strength as it travels

BIOLOGICAL AND ARTIFICIAL NEURAL NETWORKS COMPARED

Modern AIs approximate biological neural networks at the level of the neuron and synapse
(not the ion channel)



Mouse cortex

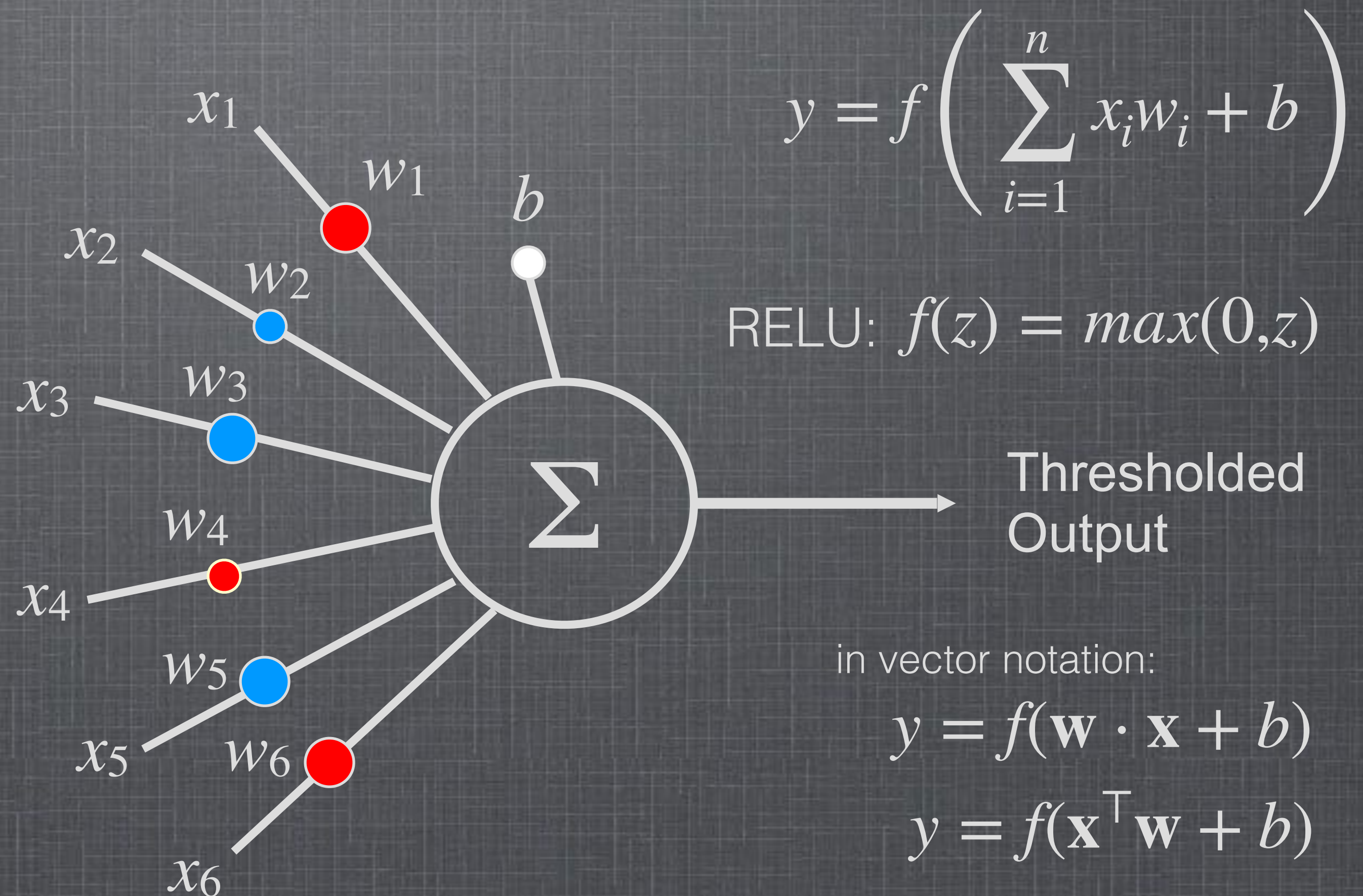


Feed-forward convolutional neural network

AI TYPICALLY USES A (VERY) SIMPLIFIED MODEL OF THE NEURON

In artificial neural networks we typically use an approximation of neural computation treated as an arithmetic-like operation:

1. Signals arriving at the synapses are modelled as continuous values (floating point numbers) that are weighted according to the positive (excitatory) or negative (inhibitory) synaptic strength (connection weight)
2. These modulated signals are summed approximating the role of the cell body
3. The model cell outputs a non-zero (typically floating point) value if this weighted sum is above some variable threshold (bias)

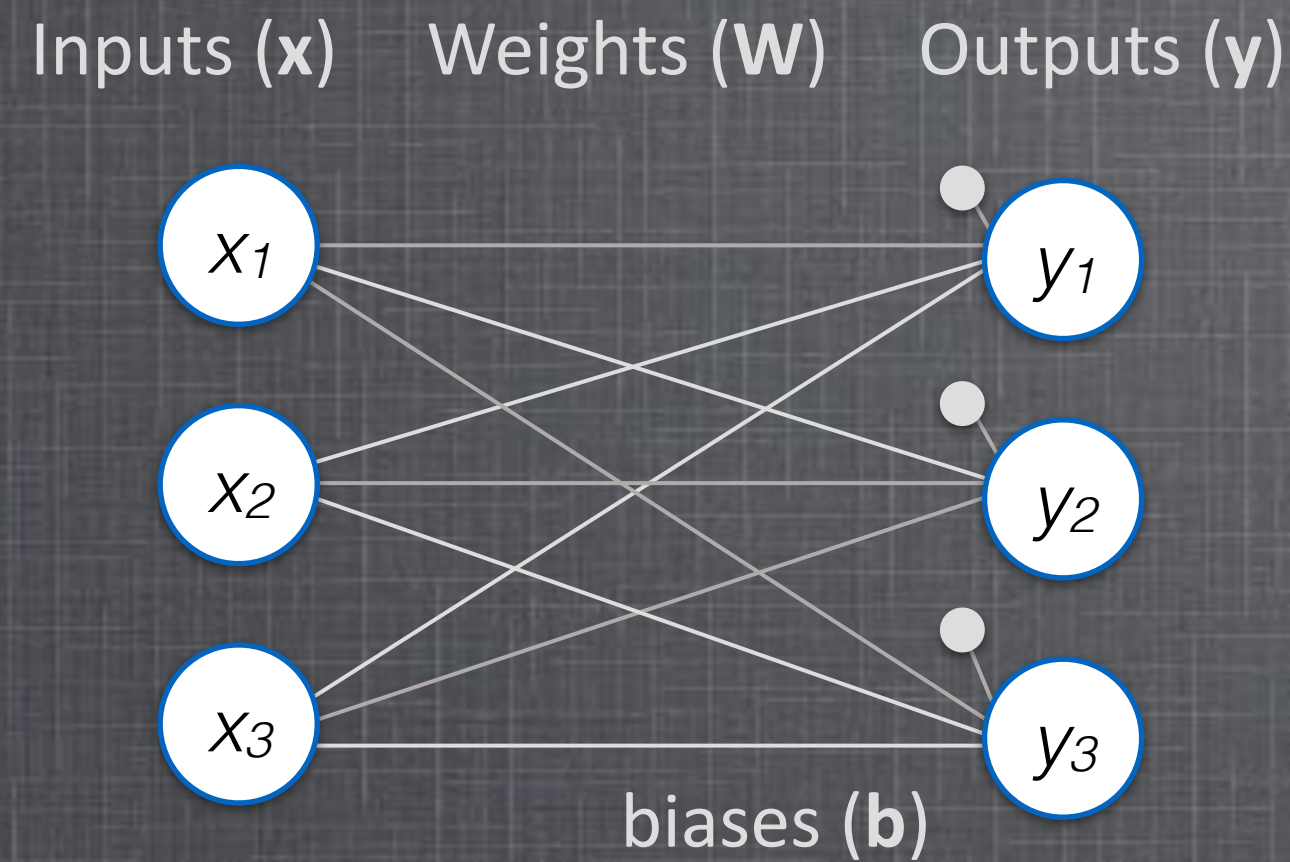


Action potential bursts are treated as continuously-valued “firing rates”

This is a loose approximation because it treats the neuron like a digital calculator, whereas a real neuron is an analog computer in which physical distance, spike timing, dendritic integration, signal decay, and chemical “exhaustion” all contribute to a complex calculation that a simple addition equation cannot replicate

ARTIFICIAL NEURAL NETWORKS BASED ON THE SIMPLIFIED MODEL

Single layer network with inputs $\mathbf{x}$, output $\mathbf{y}$, weights $\mathbf{W}$, and bias $\mathbf{b}$



$$\mathbf{y} = f(\mathbf{W}\mathbf{x} + \mathbf{b})$$

here is this equation in **matrix** form for a 3x3 network

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = f \left(\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ w_{2,1} & w_{2,2} & w_{2,3} \\ w_{3,1} & w_{3,2} & w_{3,3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \right)$$

This is an *batch* version with the inputs $\mathbf{X}$ and outputs $\mathbf{Y}$ and n input patterns

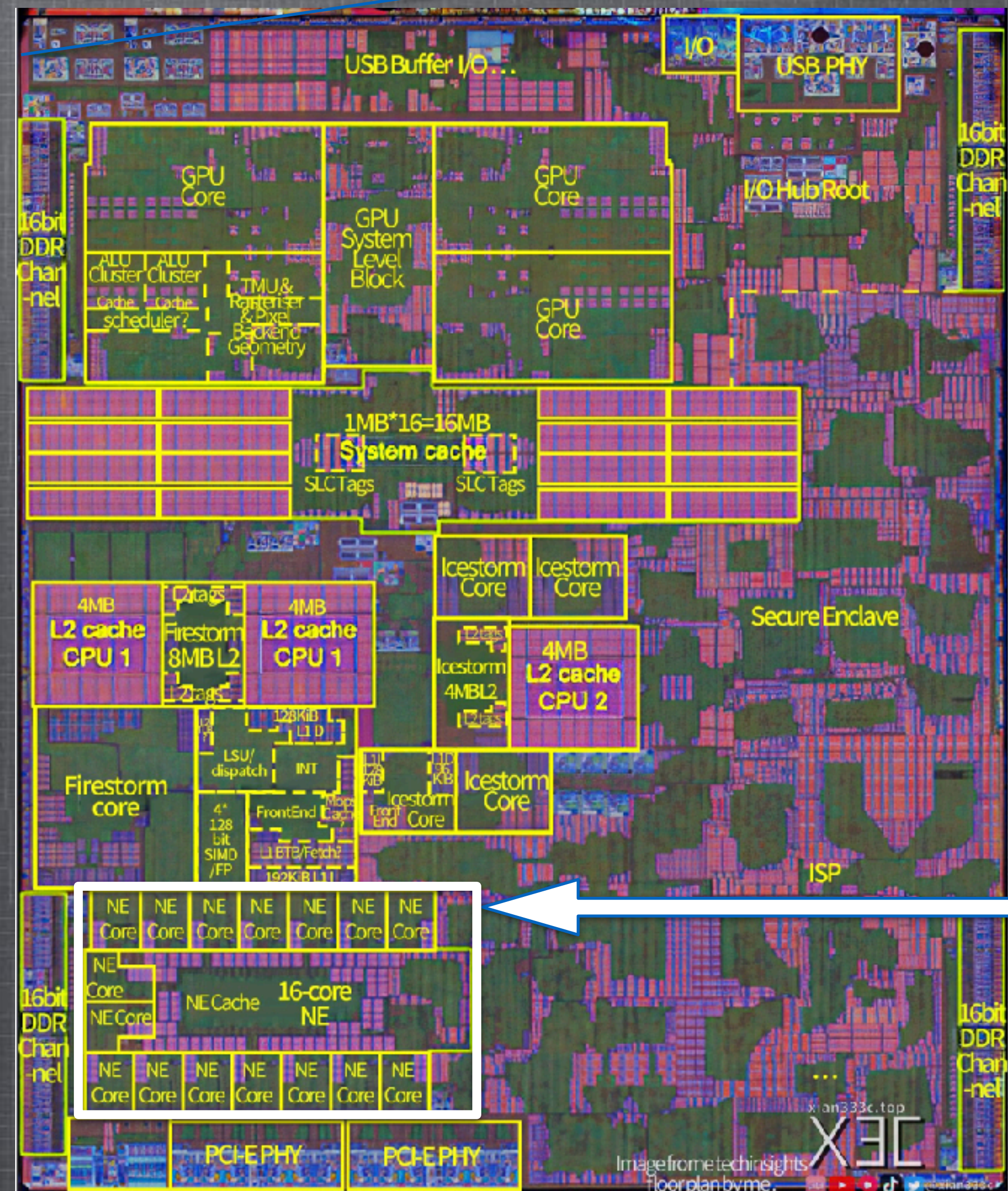
$$\mathbf{Y} = f(\mathbf{X}\mathbf{W}^T + \mathbf{b})$$

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & y_{n,3} \end{bmatrix} = f \left(\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} \end{bmatrix} \begin{bmatrix} w_{1,1} & w_{2,1} & w_{3,1} \\ w_{1,2} & w_{2,2} & w_{3,2} \\ w_{1,3} & w_{2,3} & w_{3,3} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \right)$$

For more complex networks we can use tensors (multidimensional arrays)

THE A14 BIONIC SYSTEM ON A CHIP

- The **16-core** *Apple Neural Engine* is designed for fast, energy-efficient calculation of these kinds of matrix multiplication operations
- The 16 cores are identical. For large matrices each core is assigned part of the matrix to perform calculations over
- Non-linear output functions like RELU are hard-coded
- This system can do 11 trillion operations/s (TOPS)

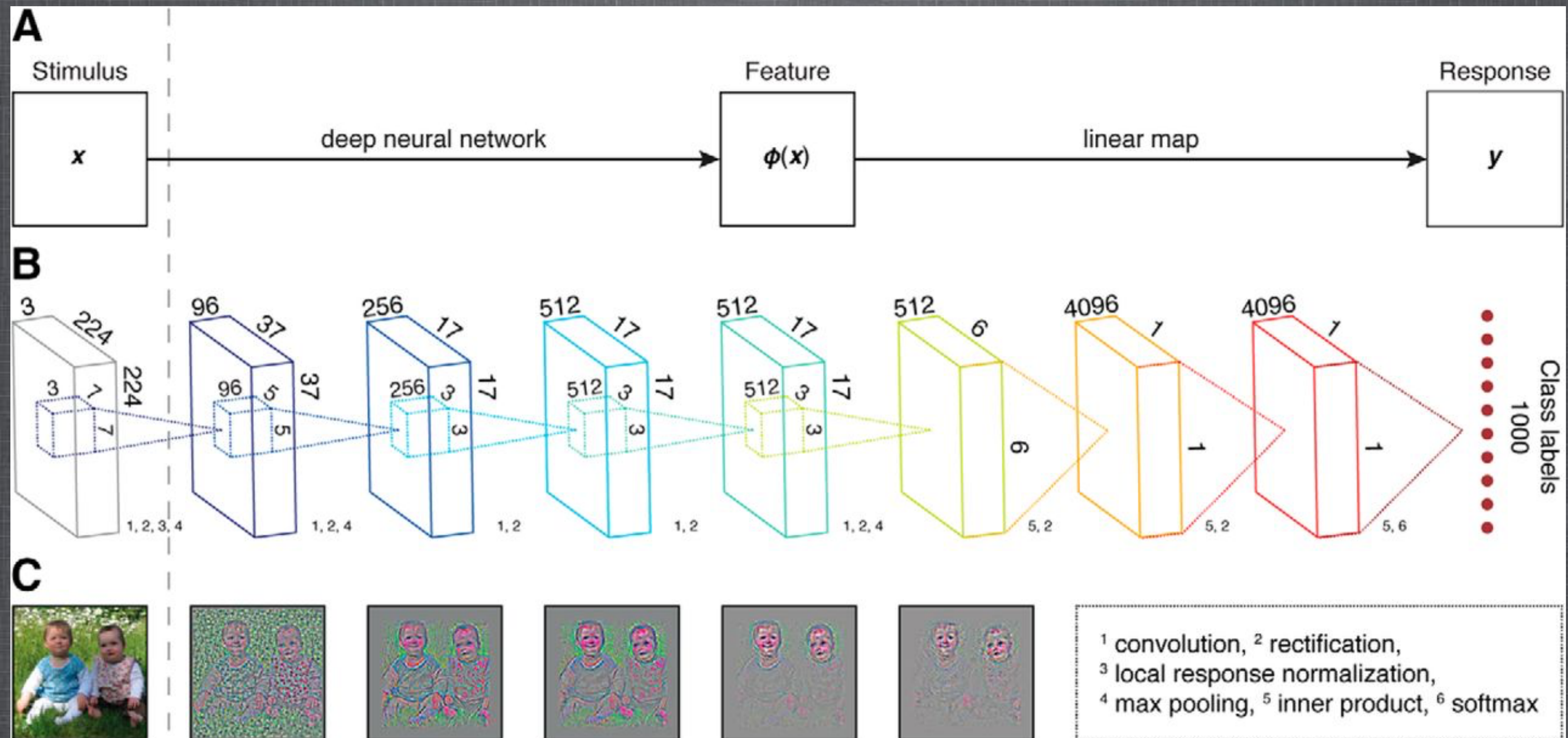


“neural engine”

88m²

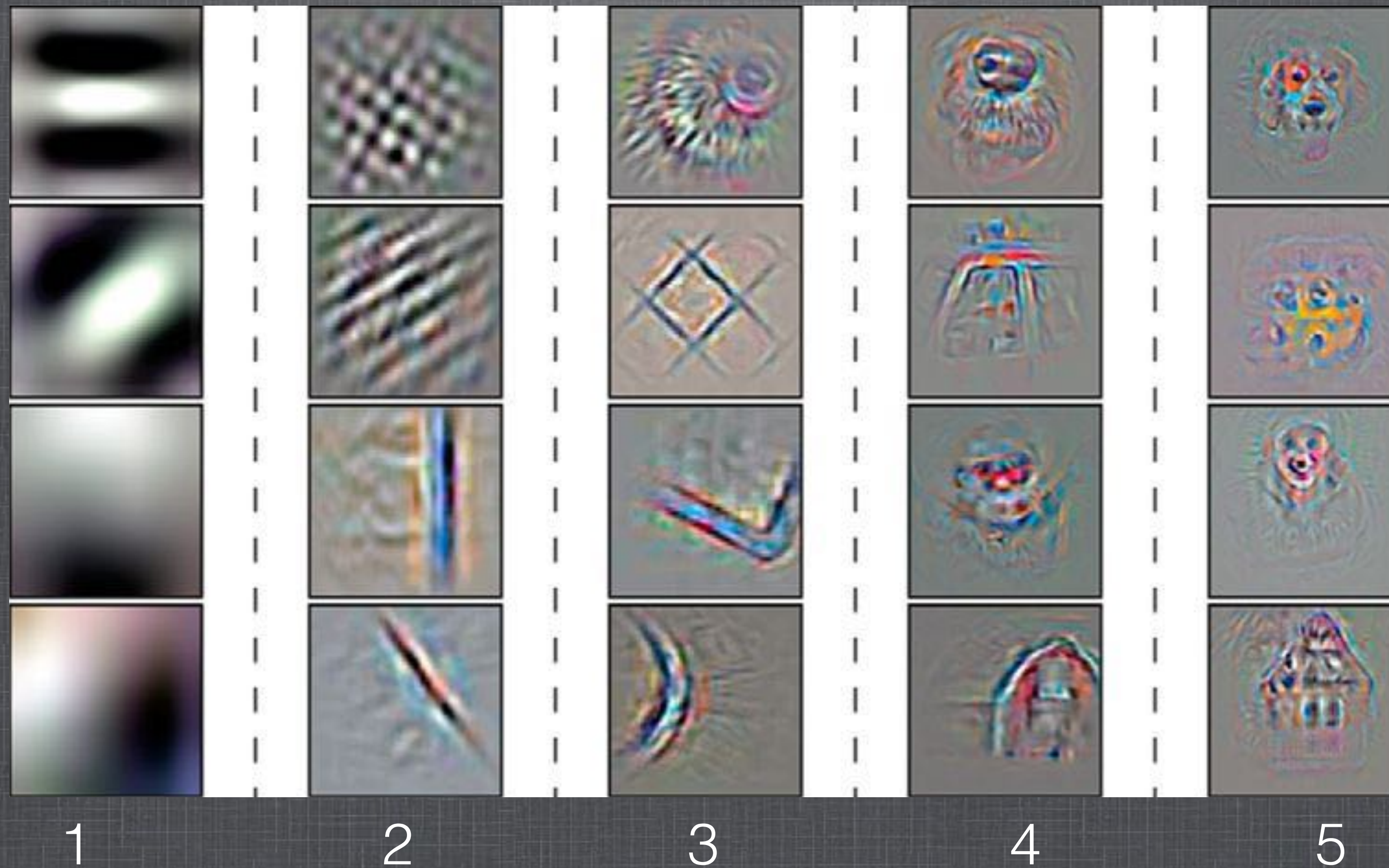
CONVOLUTIONAL NEURAL NETWORK

A convolutional neural network can be thought of as successive banks of feature detectors that where the neurons in layer 2 sample a field of neurons in layer 1 etc.

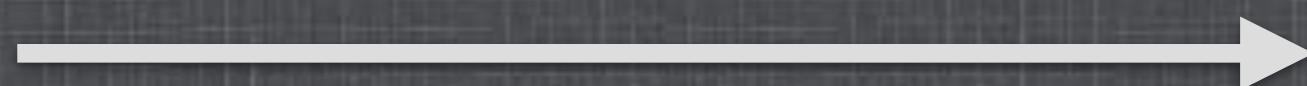


Güçlü, U, van Gerven, MAJ (2015). [The Journal of Neuroscience](#), 35(27):10005-10014.

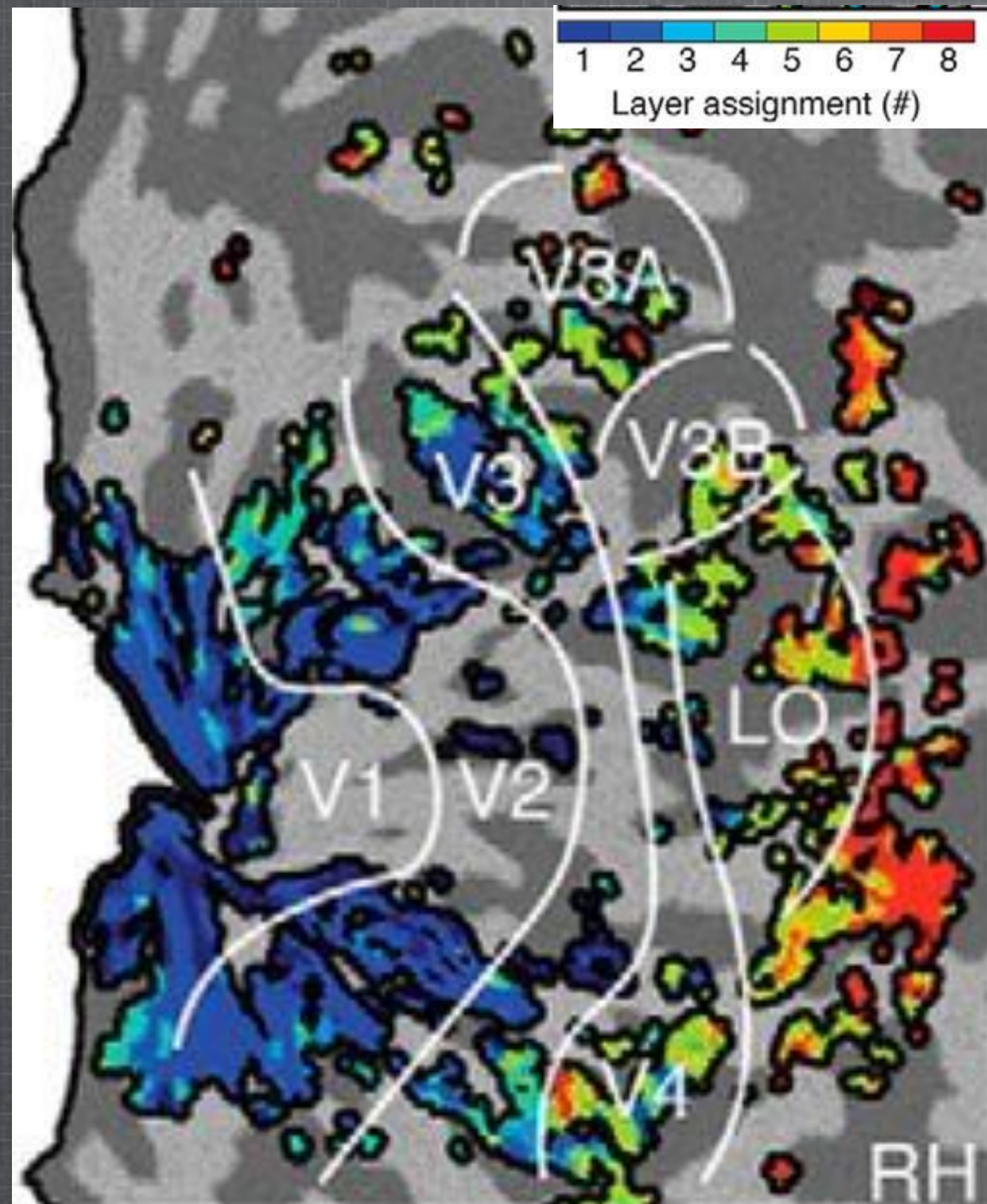
FEATURE DETECTORS BECOME MORE COMPLEX AS YOU PROGRESS THROUGH THE LAYERS OF ARTIFICIAL “DEEP” NEURAL NETWORKS



Feature complexity



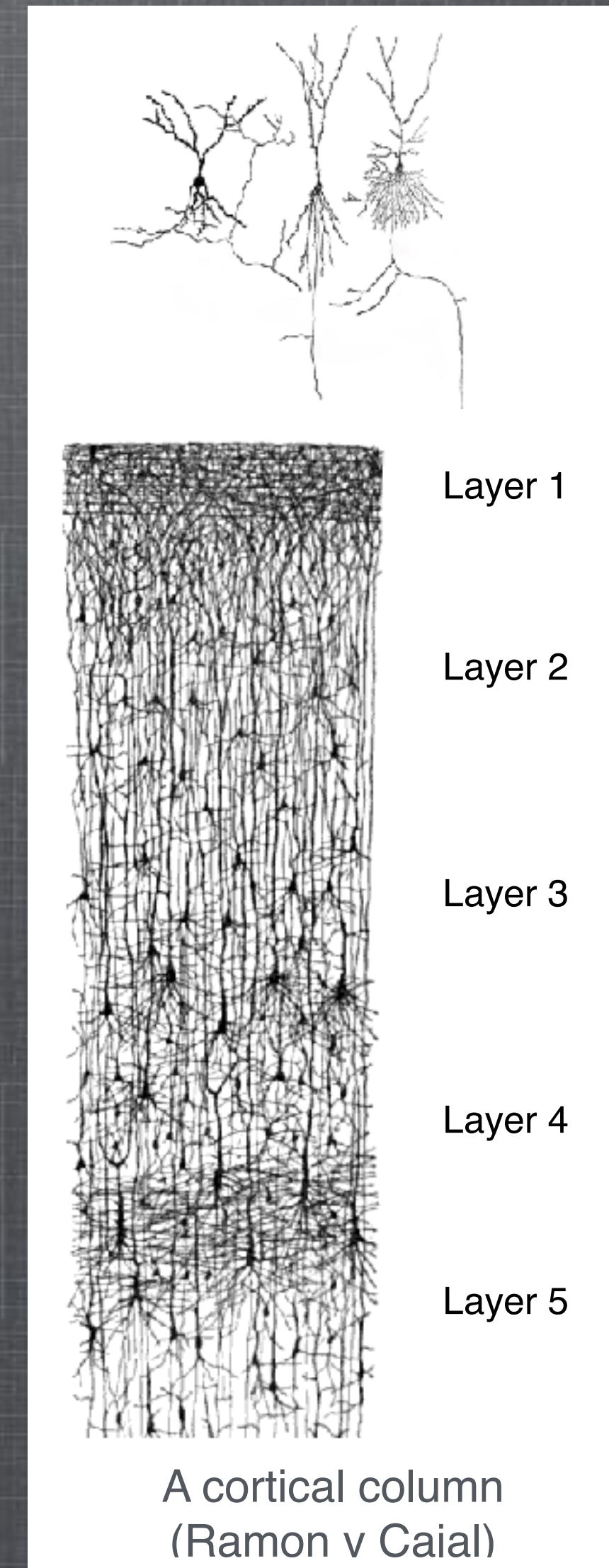
VISUAL CORTICAL AREAS SHOW SIMILAR RESPONSE PATTERNS TO DEEP NETWORKS



Activity in human visual cortex, recorded using functional brain imaging, matches that in the different layers of an artificial deep network.

Areas earlier in the visual stream (V1, V2) match lower layers in the network that have simple features; later layers (V3, V4, LO) match higher layers with more complex features.

THE CORTEX IS A NETWORK OF DEEP NEURAL NETWORKS

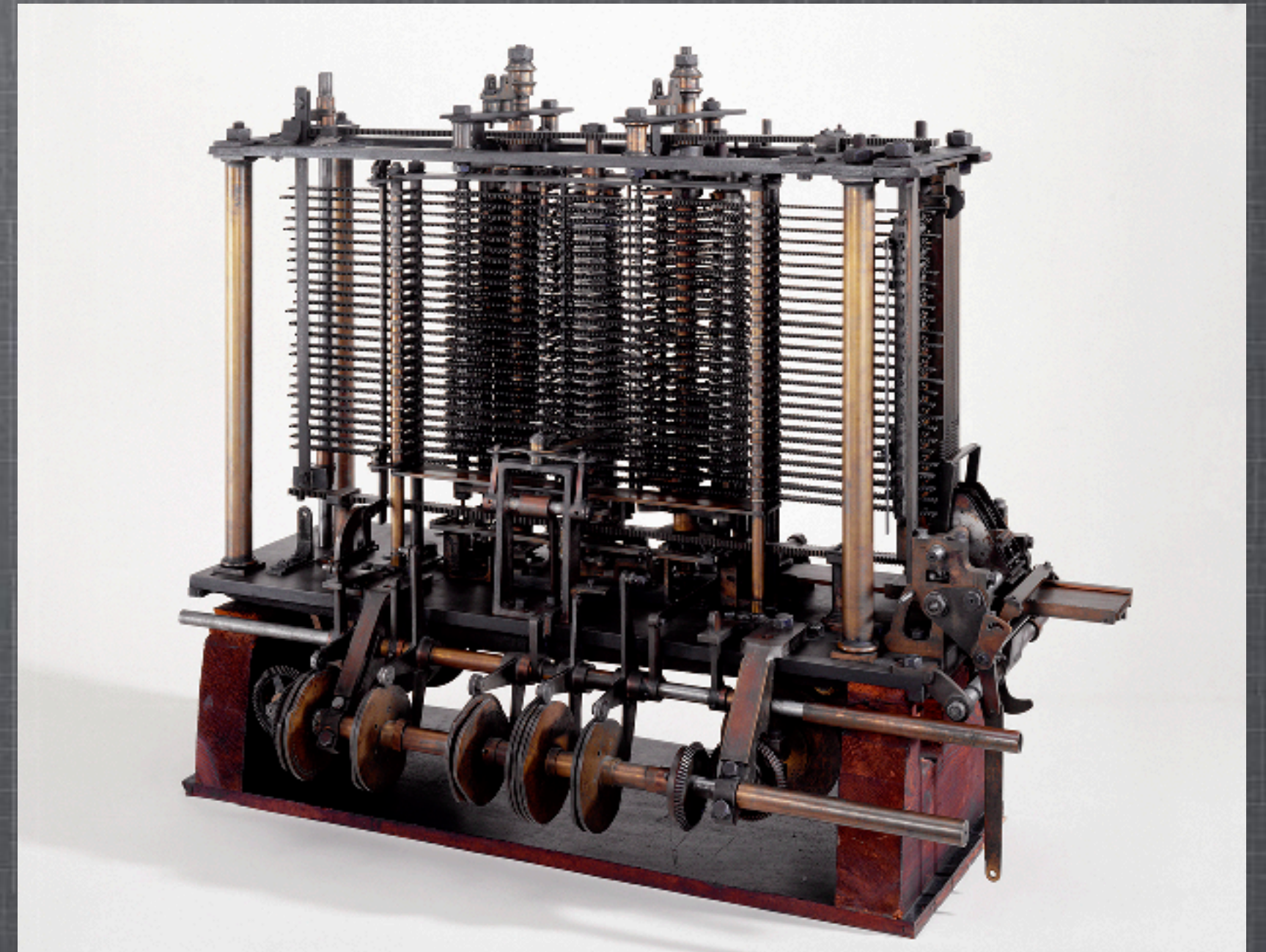


Estimates suggest 2-4 million columns in human cortex

DIGITAL COMPUTERS

Digital computation is the manipulation of strings of discrete elements, that is, strings of letters from a finite alphabet

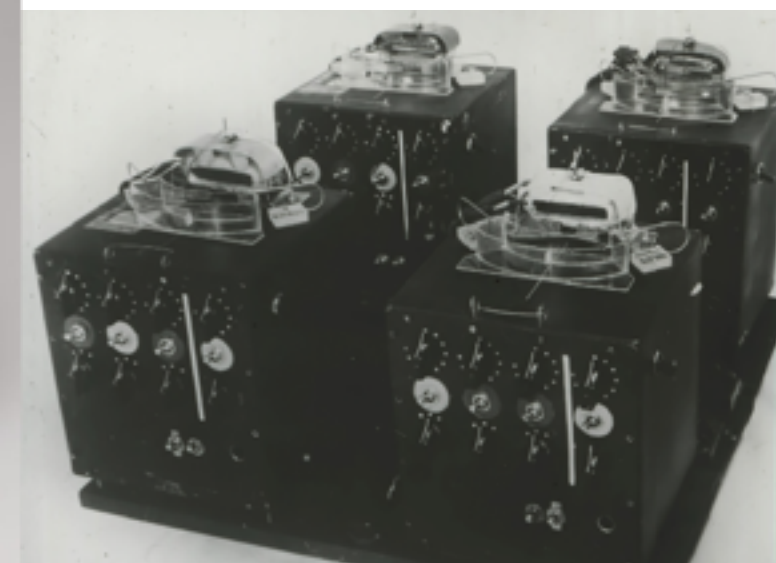
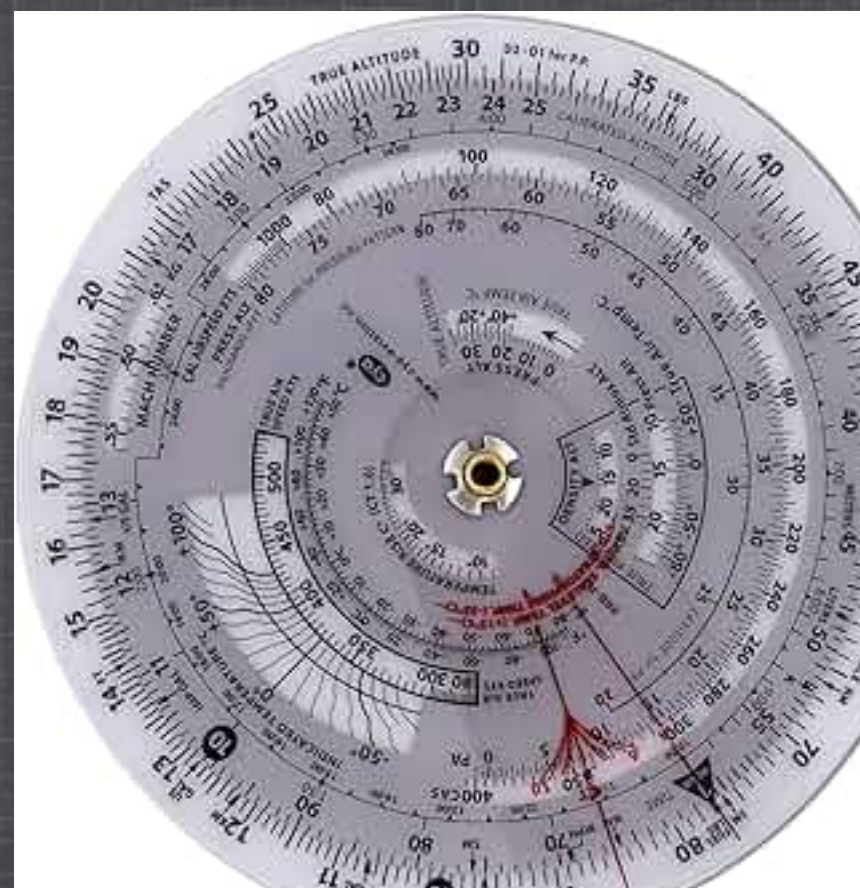
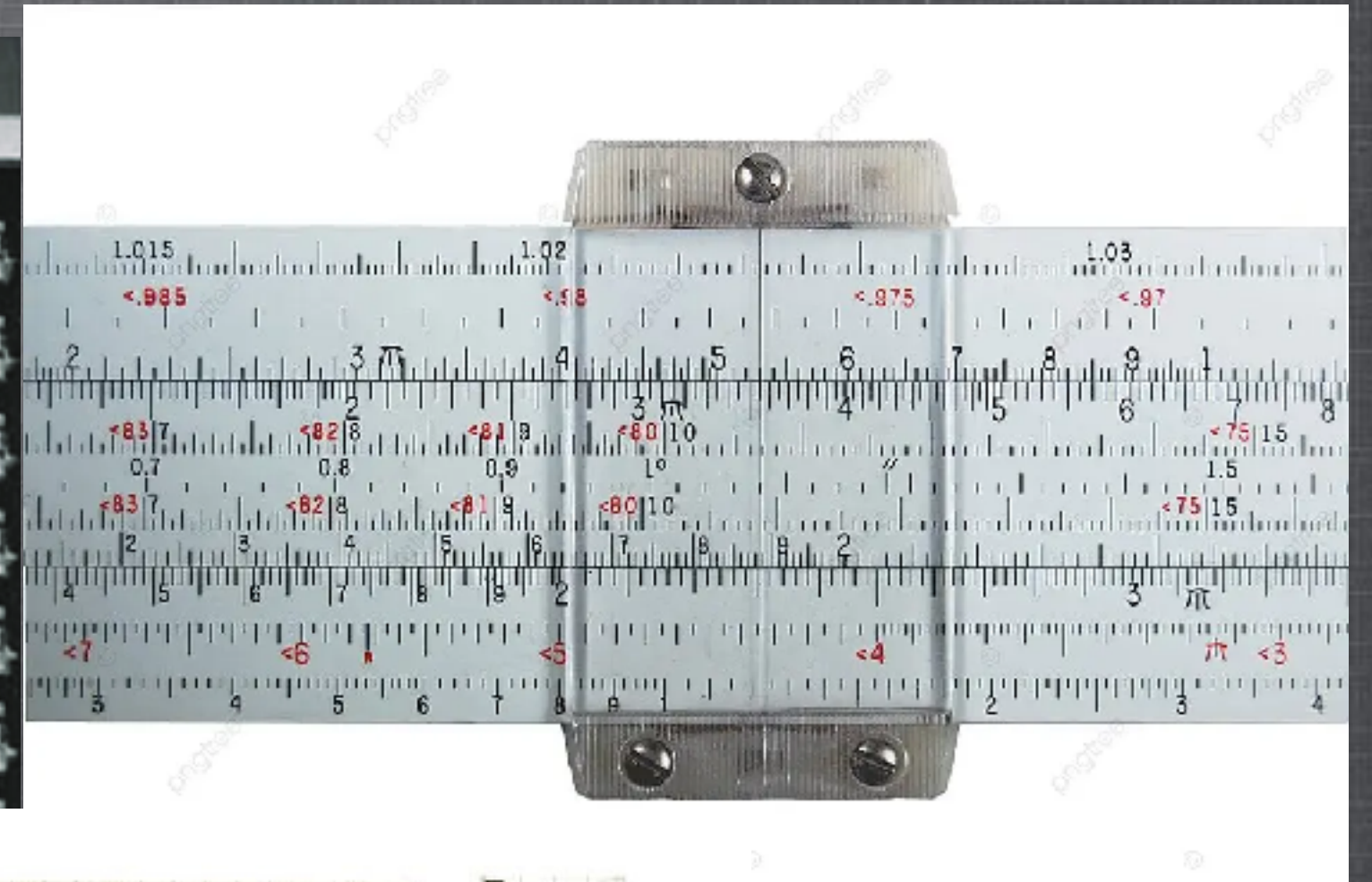
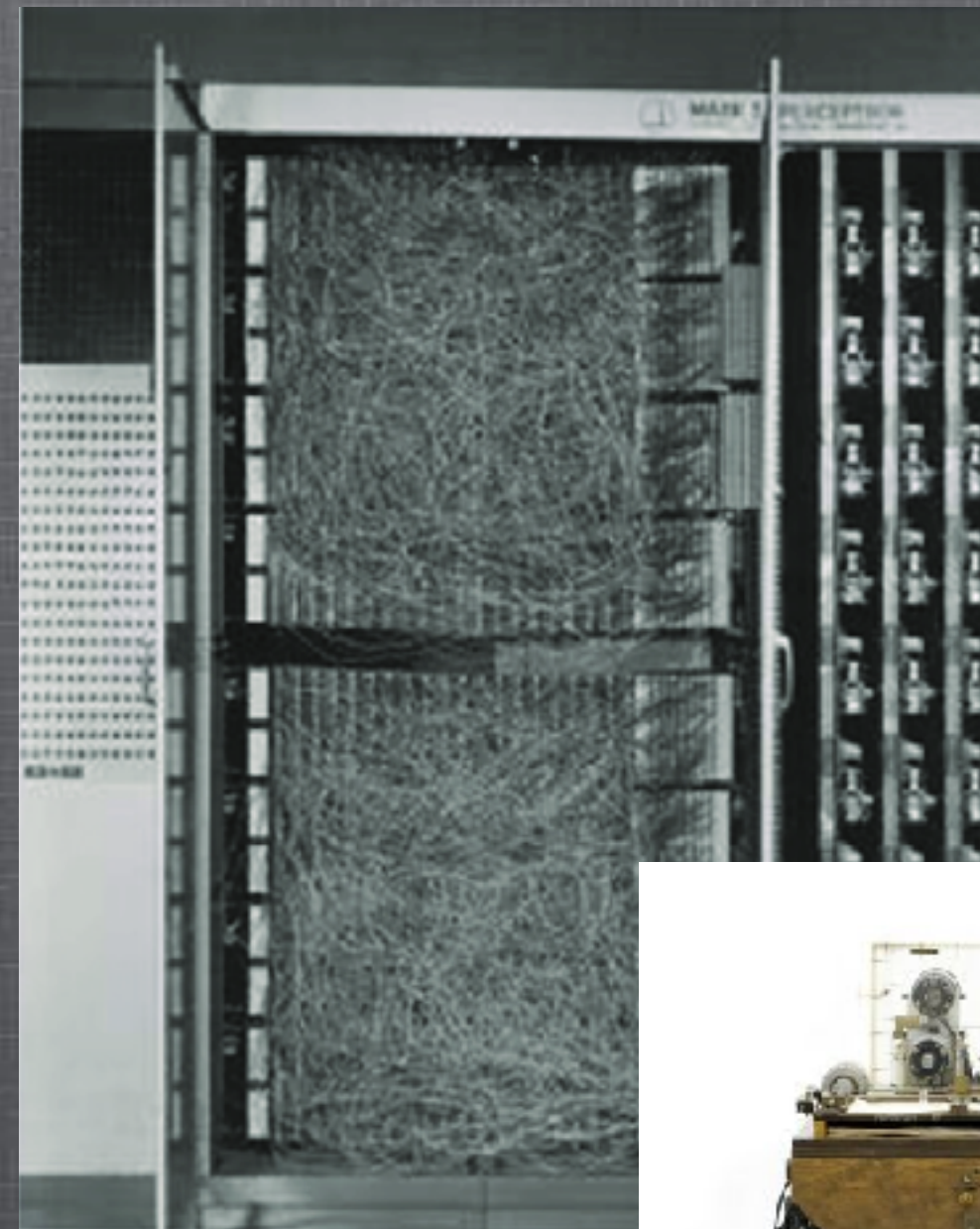
- Founded on principles of Boolean logic and Shannon's demonstration of equivalence between logic and networks of switching circuits
- This is the dominant paradigm for computation include communication technologies
- Exemplified by Babbage's analytical engine, and all modern conventional computers
- Digital computations can be implemented by any physical medium with the right degrees of freedom



ANALOG COMPUTERS

Analog computers are systems whose function is to manipulate continuous variables—variables that can vary continuously over time and take any real values within certain intervals

- Analog computers represent numbers with physical quantities
- Are also medium-independent
- Often used to solve problems involving differential equations in real time
- Widely used before digital computers up to 1970s



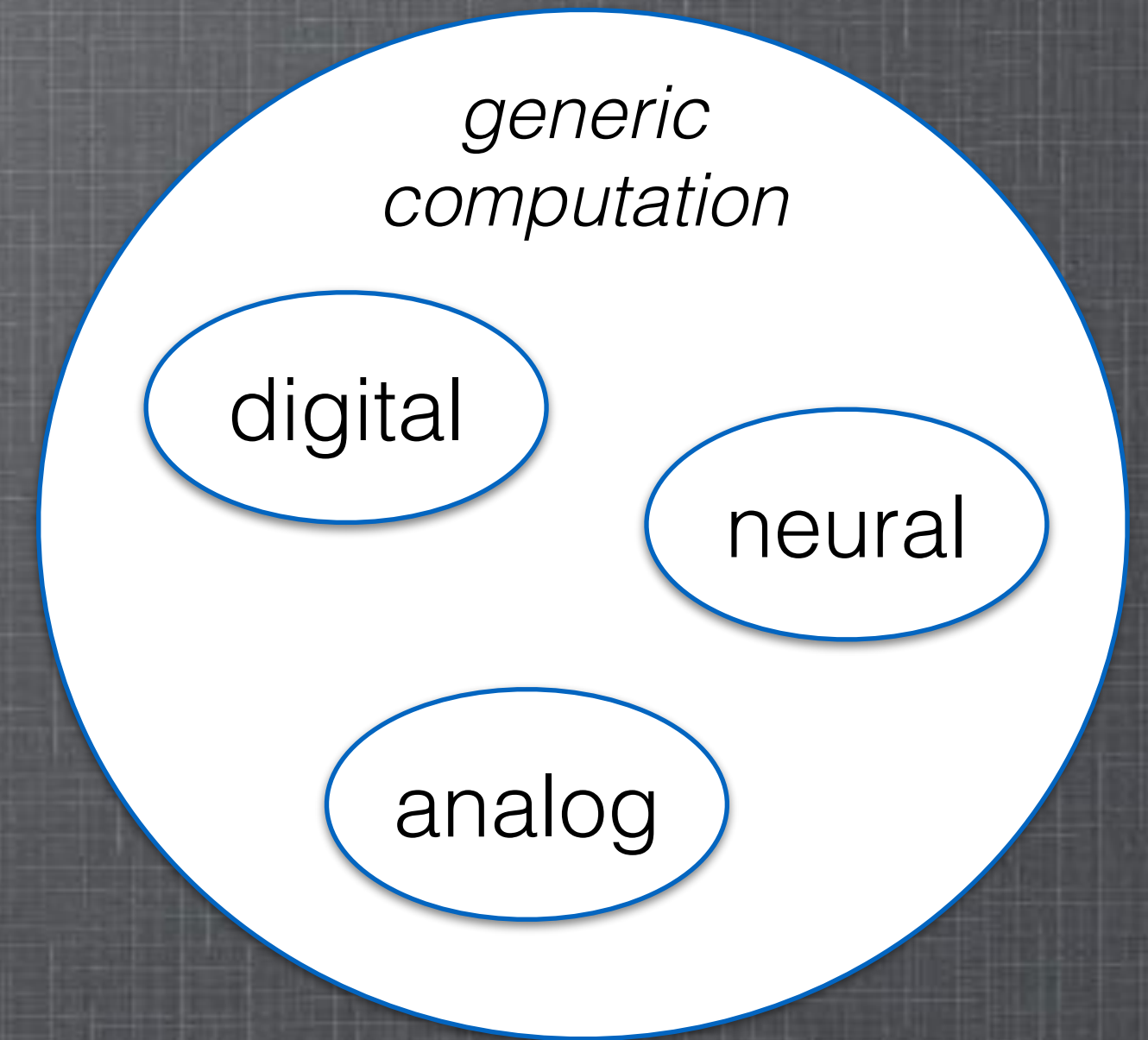
THE BRAIN IS HYBRID

Digital

- Spikes are all-or-nothing
- Ion channels are on-off

Analog

- Membrane potential varies continuously
- Synaptic neurotransmission
- Synaptic integration and dendritic transmission
- Neuromodulation



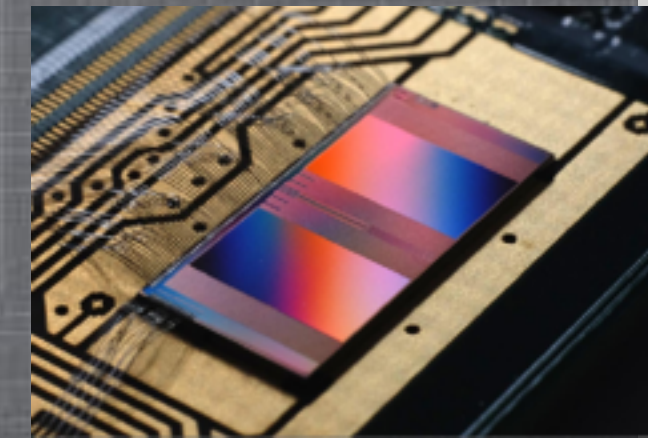
According to Piccinini and Bahar (2013):

1. The neural processes that explain cognition are manipulations of **medium-independent** vehicles
2. Manipulations of medium-independent vehicles are computations in the generic sense
3. *Therefore*, the neural processes that explain cognition **are computations in the generic sense**

BRAINSCALE—A CONTEMPORARY ANALOG COMPUTER THAT SIMULATES NEURAL CIRCUITS

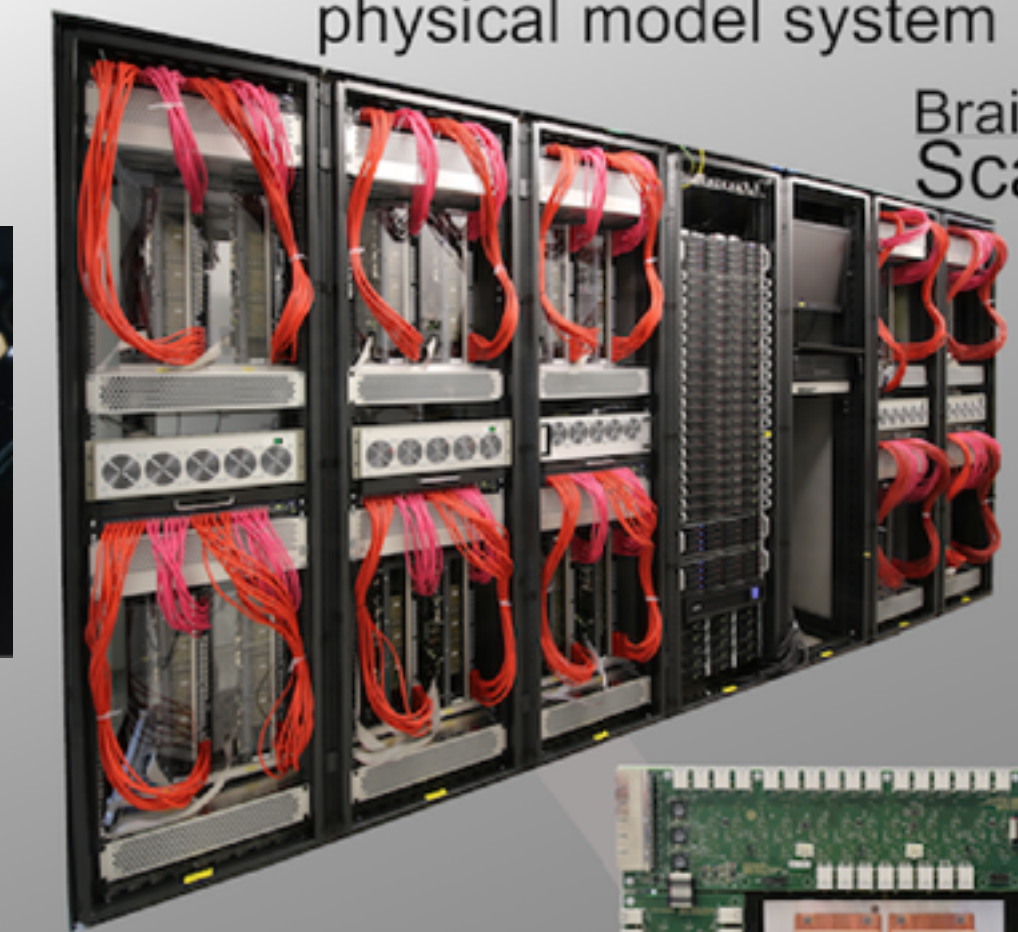
- Spiking neural networks
- Continuous-time dynamical systems
- Brain-inspired neuron models
- Primarily used as a research platform
- Each neuron is implemented as an analog circuit that physically behaves like a leaky integrate-and-fire neuron:
 - Capacitors represent membrane potential
 - Currents represent synaptic input
 - Threshold crossings trigger spikes
 - Time constants emerge from transistor physics
 - Simulates at 1000x biological speed

But digital simulation of neurons is more efficient

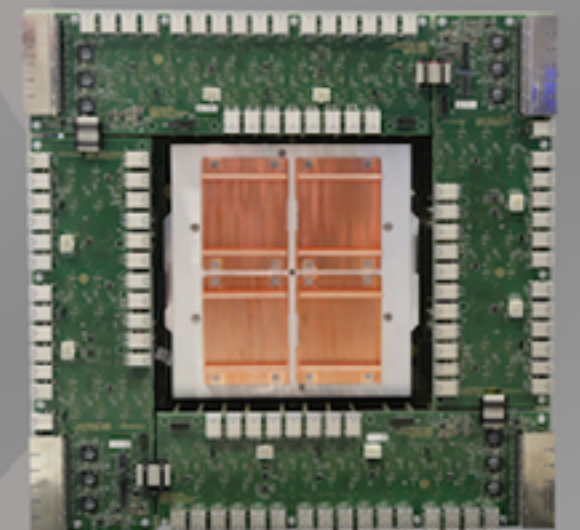


The BrainScaleS neuromorphic physical model system

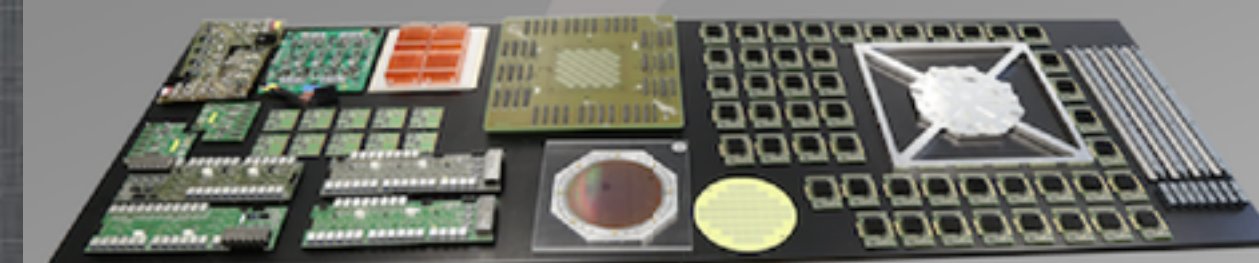
Human Brain Project
BrainScaleS



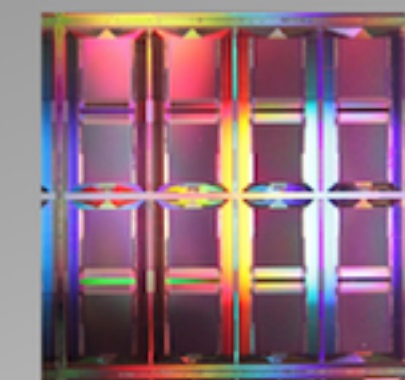
20 wafer modules
3.932.160 neurons
880.803.840 synapses



wafer module (50 cm x 50 cm)

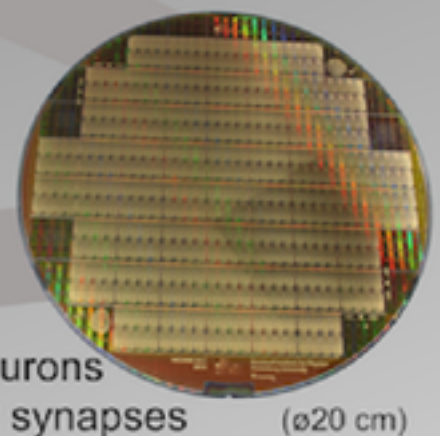


components of a wafer module



8 HICANN chips per reticle
(2 cm x 2 cm)

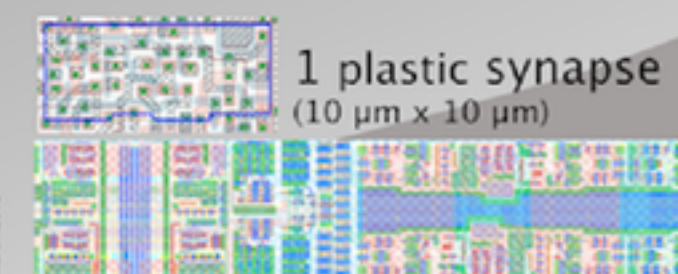
48 reticles
per wafer
196.608 neurons
44.040.192 synapses



(ø20 cm)



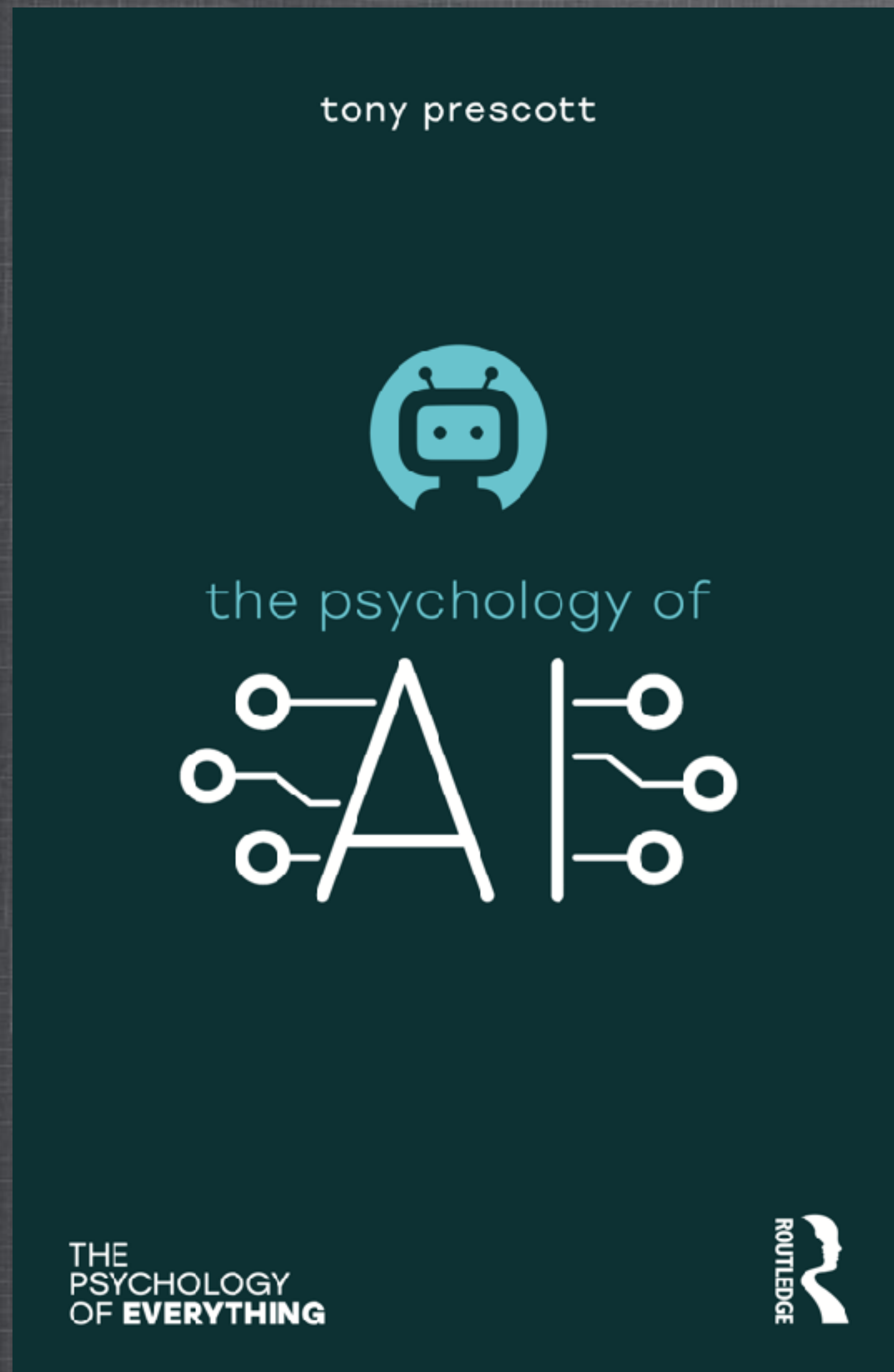
1 plastic synapse
(10 μ m x 10 μ m)



2 neurons
(150 μ m x 20 μ m)

512 neurons
114.688 synapses
per HICANN chip
(0.5 cm x 1 cm)

THE PSYCHOLOGY OF ARTIFICIAL INTELLIGENCE



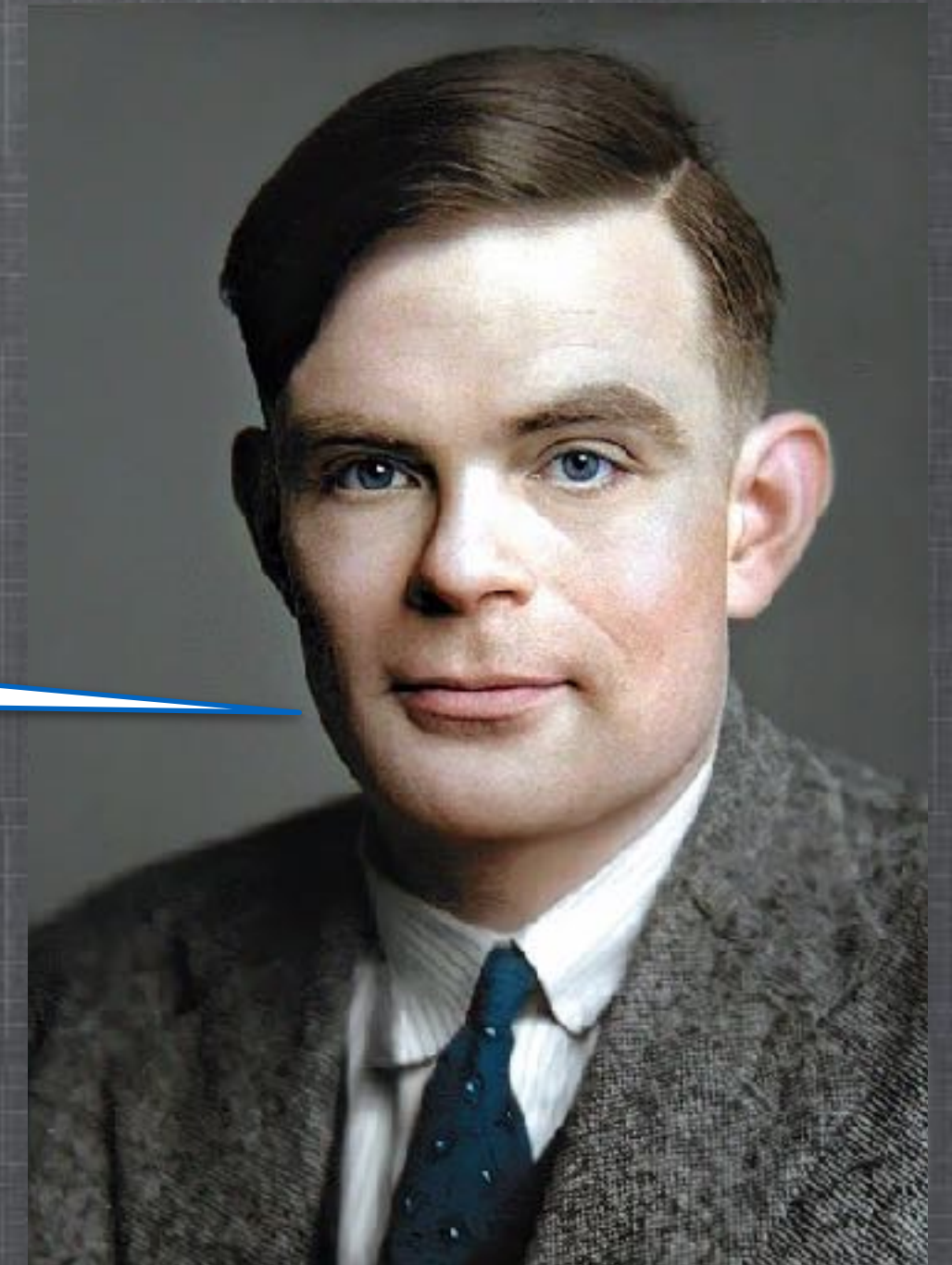
- *Overview*
- What is intelligence?
- Are brains computers?
- **Do AIs understand anything?**

CAN A MACHINE THINK?—ITS AN EMPIRICAL ISSUE

The “Turing Test” (Turing, 1950)

I believe that at the end of the century the use of words and general educated opinion will have altered so much that **one will be able to speak of machines thinking without expecting to be contradicted**

- Any subject you like, but, linguistic communication only
- The competitive element makes the test more difficult
- Not a ‘litmus’ test—a negative result proves nothing



Alan Turing, 1912-1954

DESCARTES HAD A SIMILAR THOUGHT...

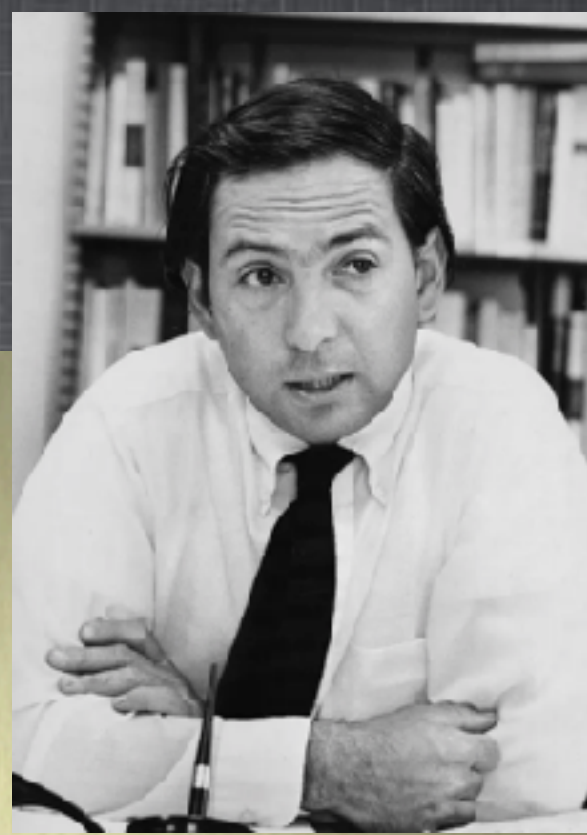
We can easily conceive of a machine so constructed that it utters words, and even utters words that correspond to bodily actions [...] **but not that such a machine should produce different sequences of words so as to give an appropriately meaningful answer to whatever is said in its presence**”



Renee Descartes, 17th c.

Descartes, 1637, Discourse on the Method, part 5.

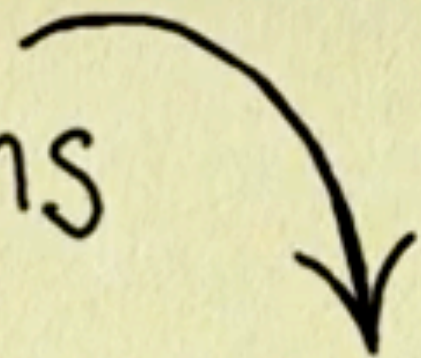
THE CHINESE ROOM



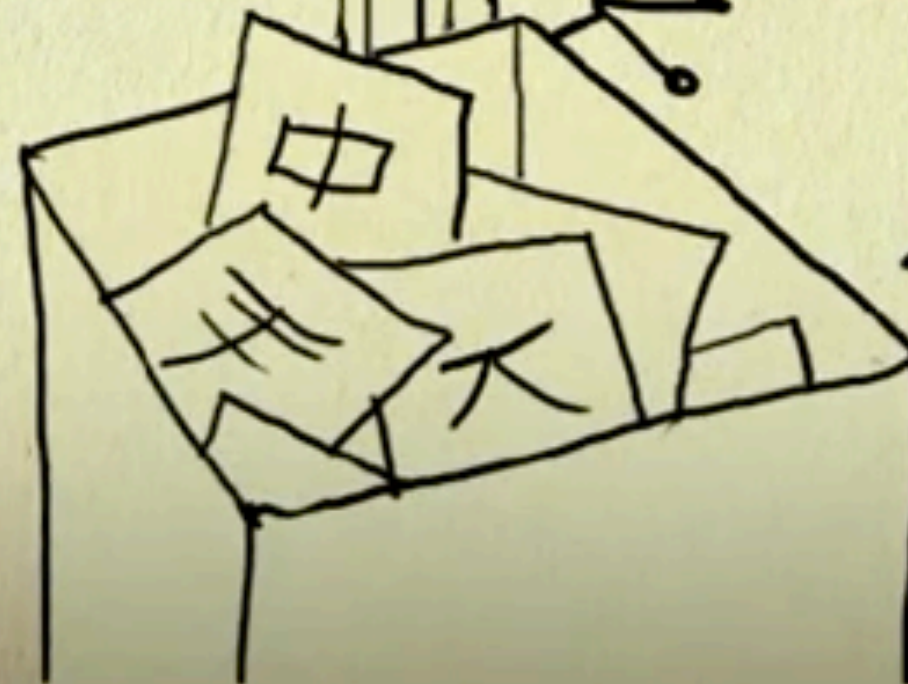
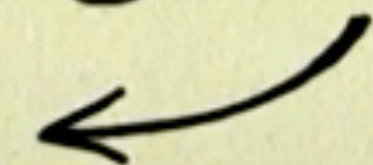
John Searle
1932-2025



book of
instructions



boxes of
Chinese characters

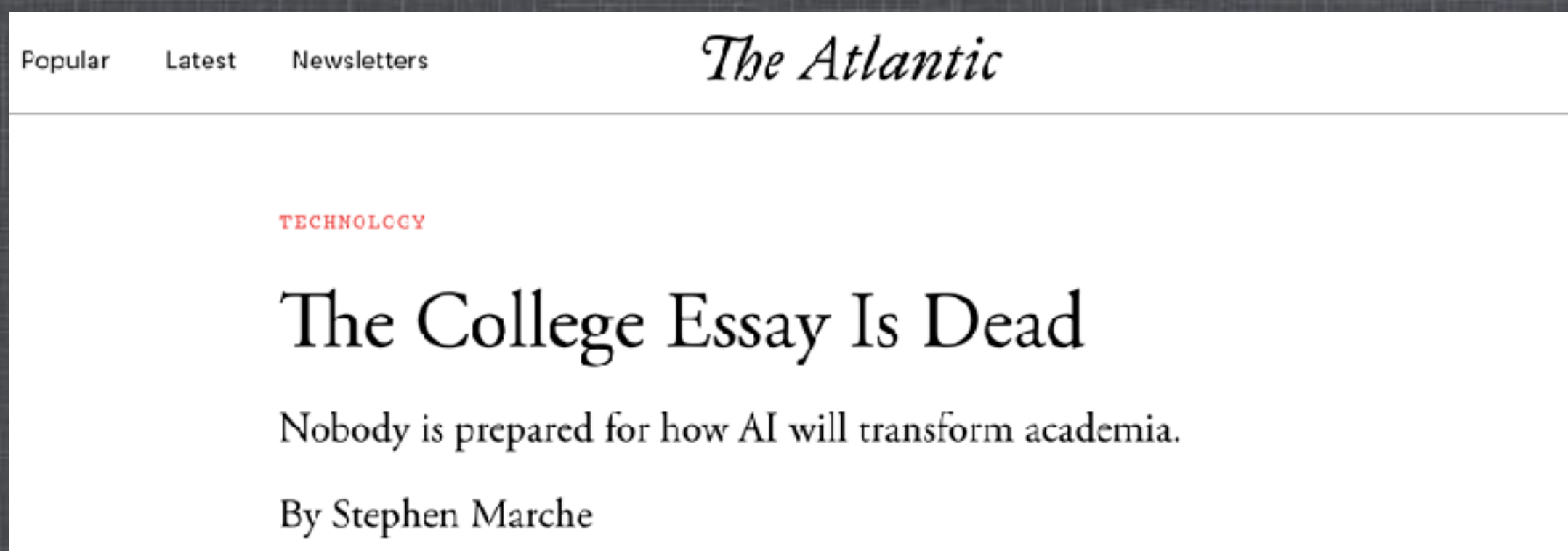
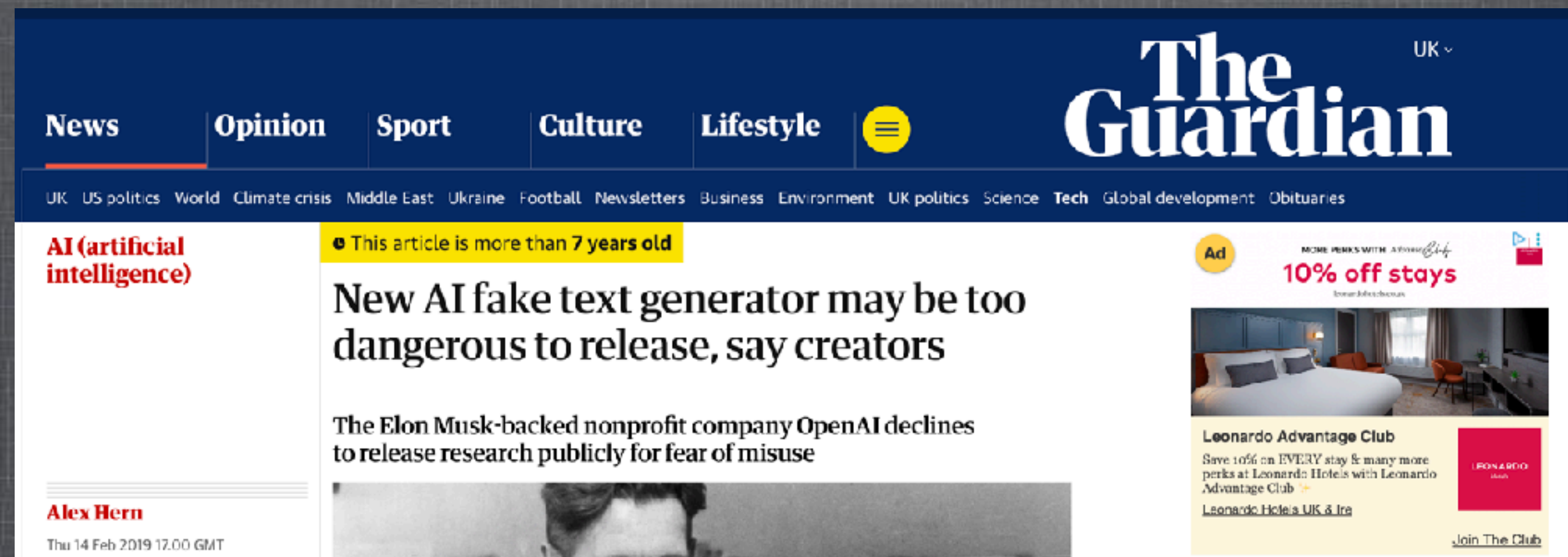


LARGE-LANGUAGE MODELS (2019–)

OpenAI built a text generator so good,
it's considered too dangerous to release

Zack Whittaker

9:17 AM PST · February 17, 2019



THE TRANSFORMER MODEL

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

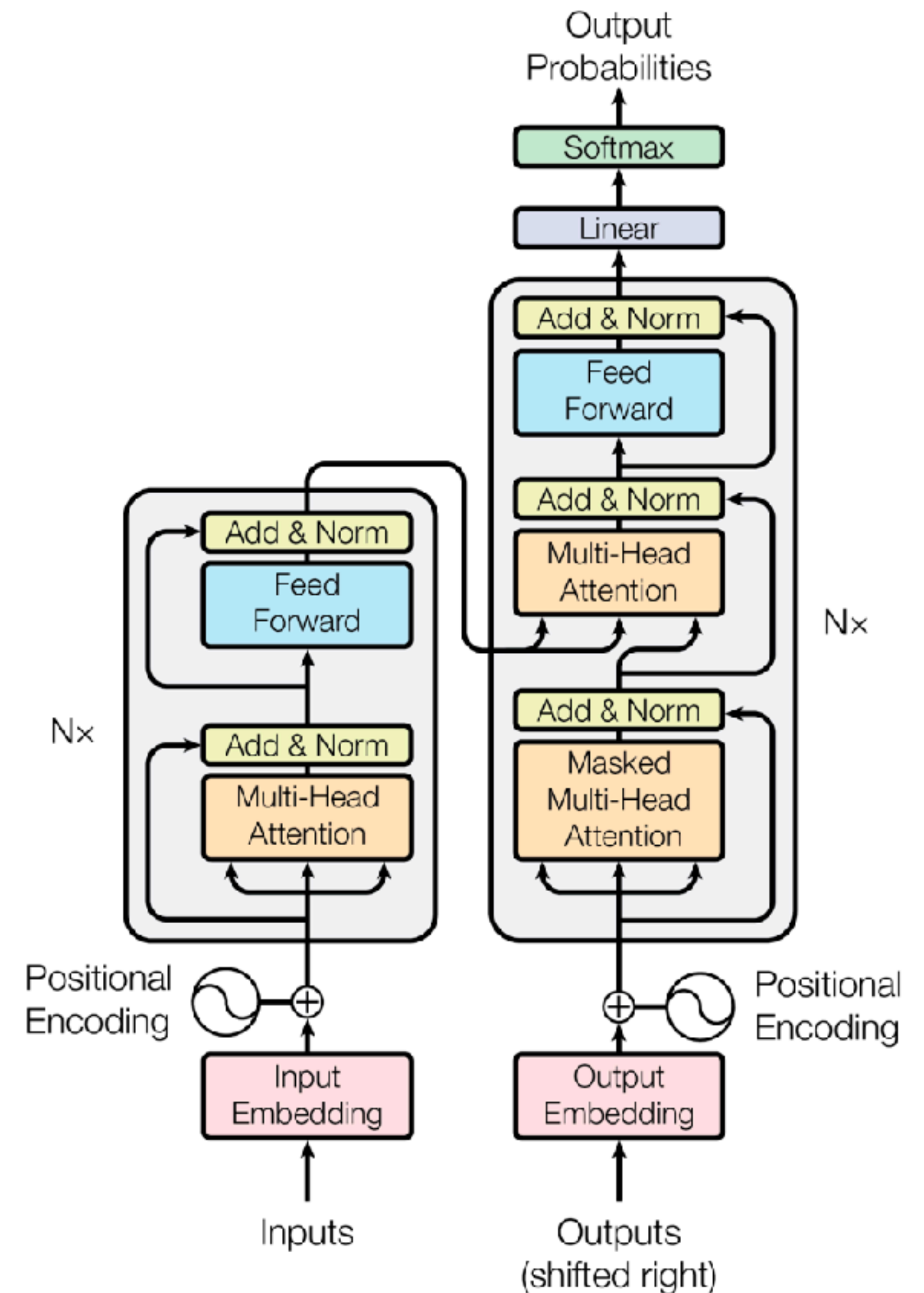
Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

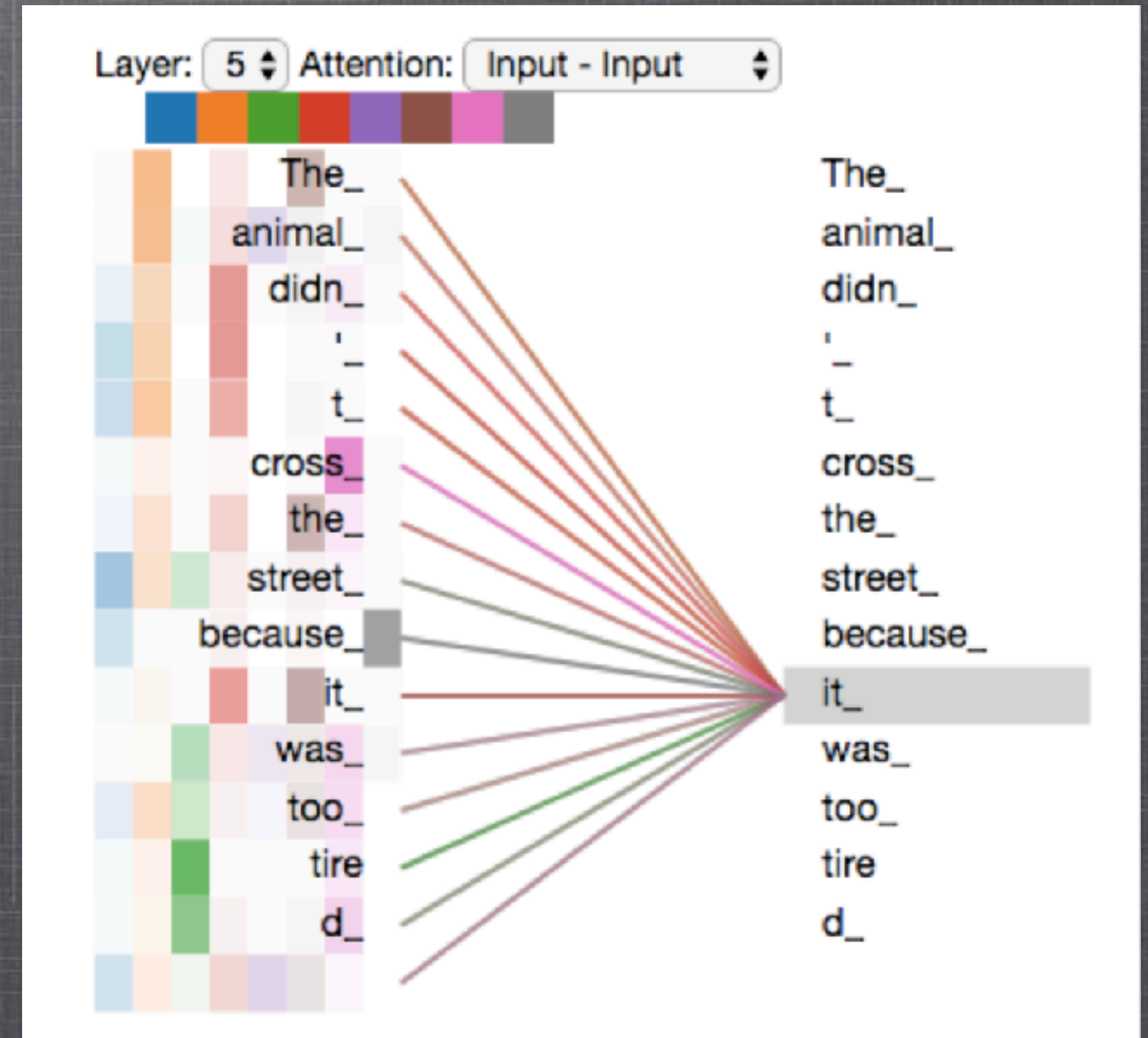
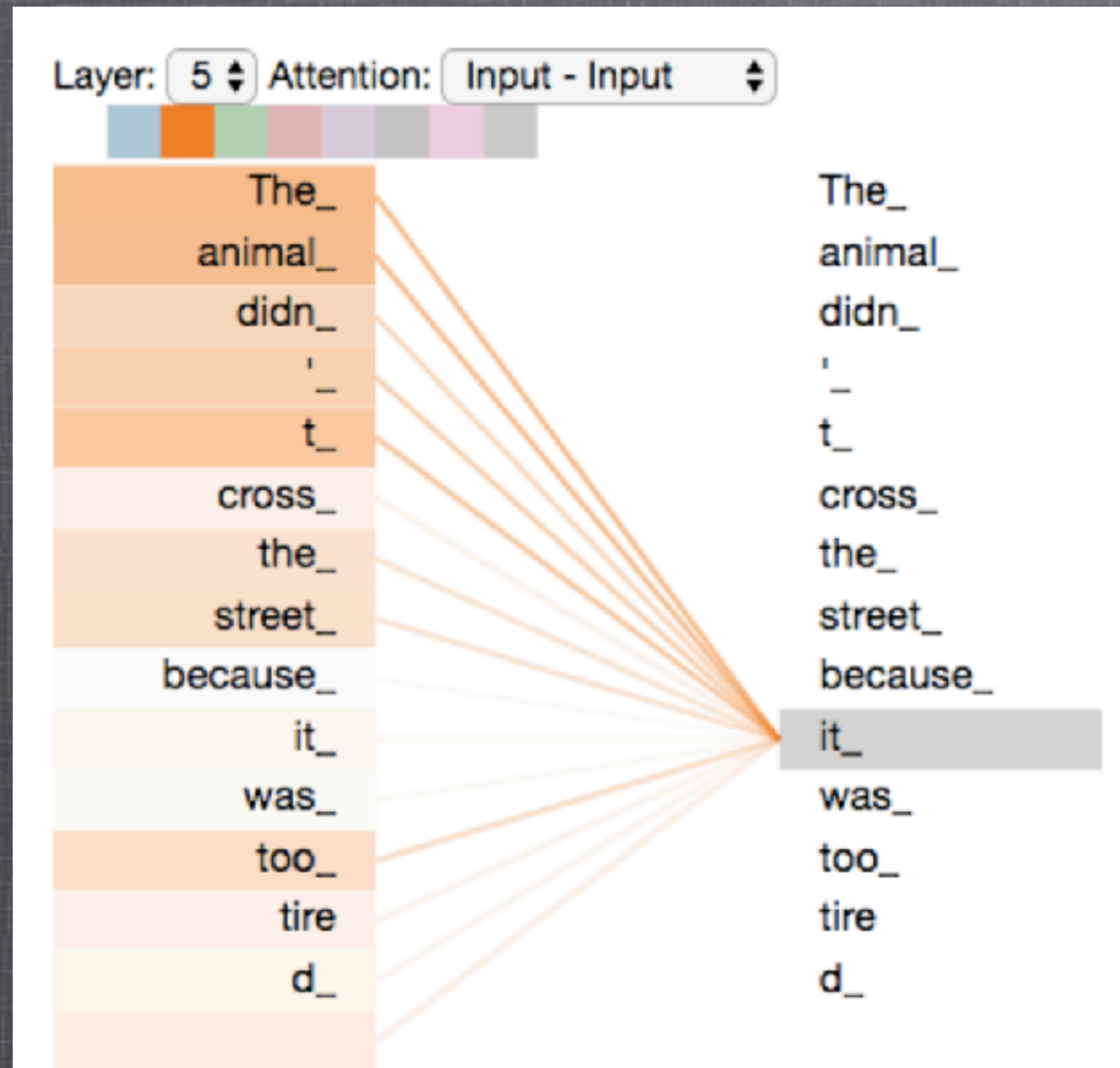
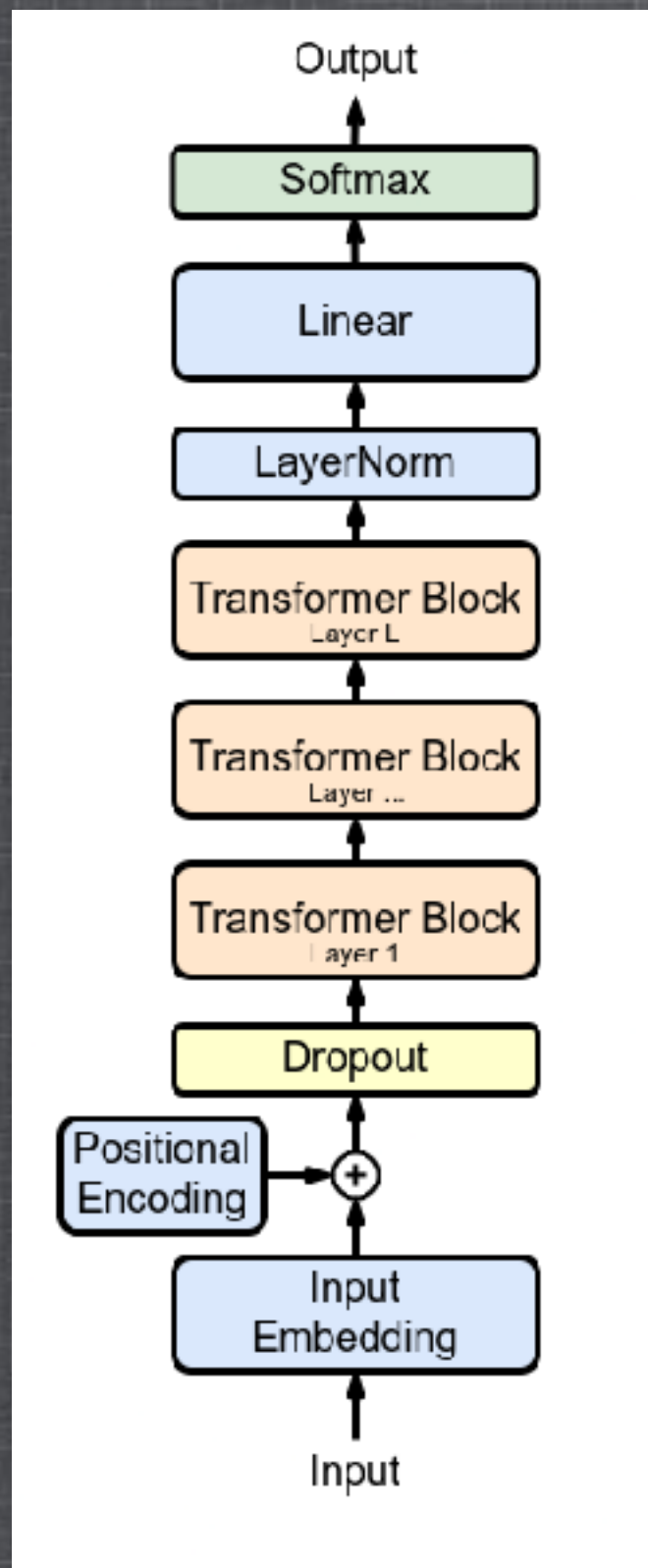
Illia Polosukhin* †
illia.polosukhin@gmail.com



MULTI-HEADED ATTENTION

”The animal didn't cross the street because it was too tired”

from Alamar, 2018



Example attention scores for the word “it”

The self-attention mechanism calculates an attention score for each pair of tokens, based on three vectors (queries, keys, and values) calculated for each token using learned weights. The attention score is calculated as the dot-product of the relevant vectors for each token in the pair. This happens for multiple attention heads and in each layer.

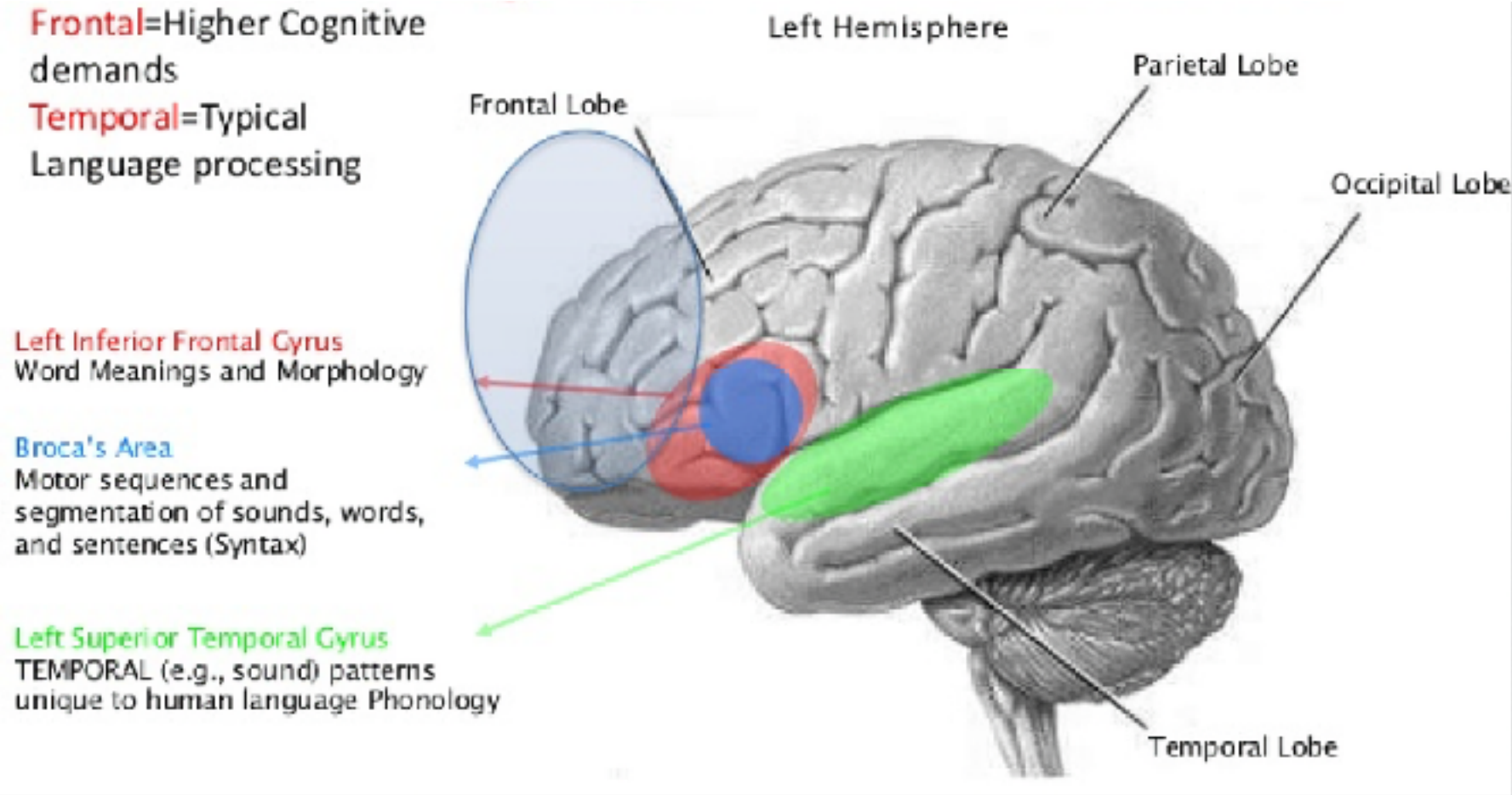
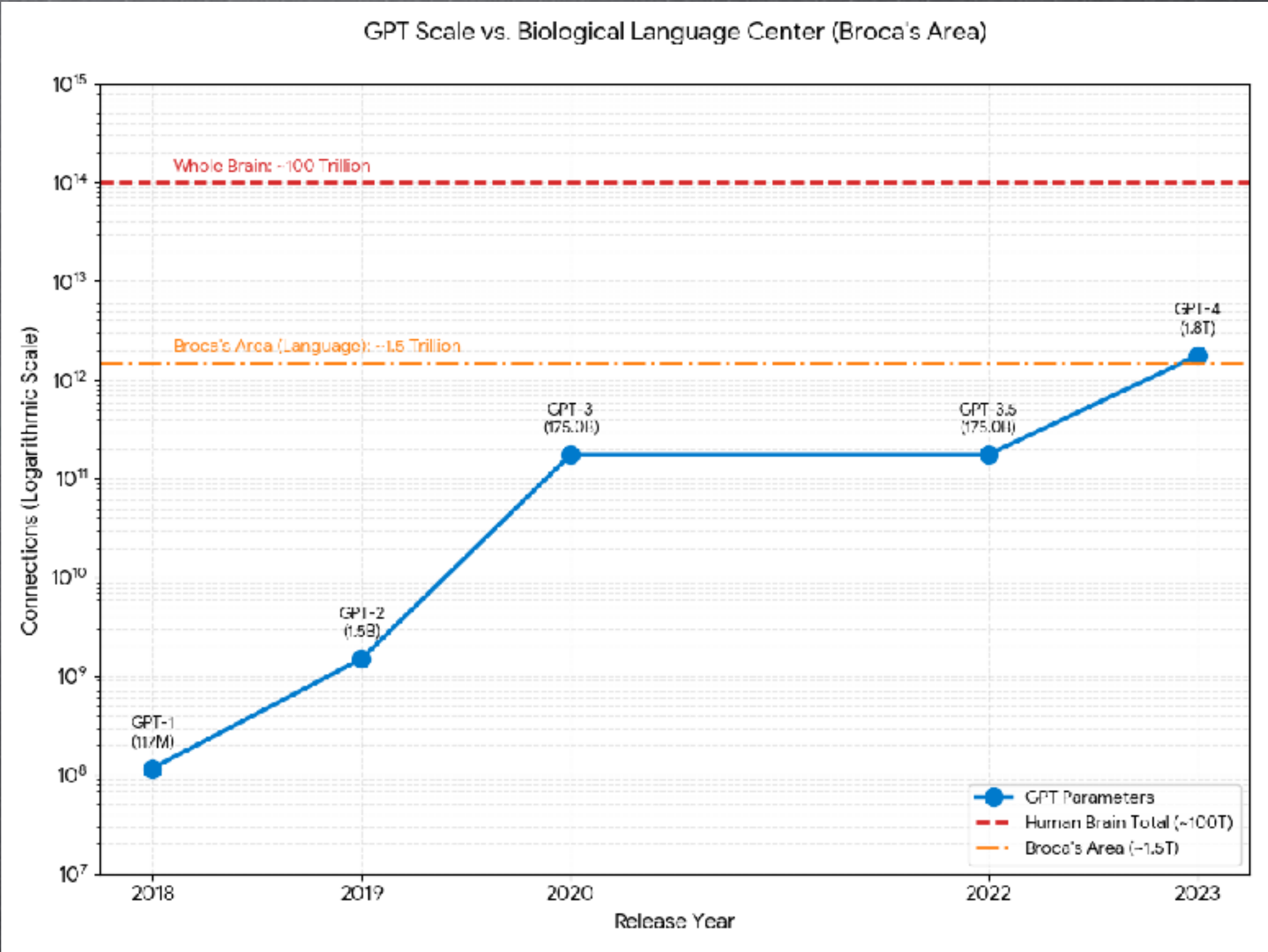
GPT MODELS COMPARED TO THE BRAIN

The latest GPT model has a similar number of parameters as there are synapses in Broca's area, one of the key language areas in the human brain.

Model	Released	Parameters	Layers	Context Window	Training Data
GPT-1	2018	117 Million	12	512 tokens	~5 GB (BookCorpus)
GPT-2	2019	1.5 Billion	48	1,024 tokens	40 GB (WebText)
GPT-3	2020	175 Billion	96	2,048 tokens	570 GB (Common Crawl+)
GPT-3.5	2022	175 Billion	96	4,096 to 16,384 tokens	Undisclosed
GPT-4	2023	~1.76 Trillion*	~120*	8,192 to 128,000 tokens	Undisclosed

*GPT-4's parameter count is widely reported by industry analysts as totalling roughly 1.76 trillion, though OpenAI has not officially confirmed this.

From GPT-3.5 on, models are improved using human feedback and reinforcement learning.
GPT-4 models are multi-modal.



LLMS PERFORM WELL ON IQ TESTS

Similar to, or above, average human performance

Table 1												
Accuracy on each question type along with the overall IQ range for each LLM, sorted by the IQ range. (Note: The highest accuracy in each column is highlighted in boldface)												
Model	Verbal questions								Numerical questions			IQ Range
	Analogies	Compreh.	Similarities	Opposites	Similar or Opposite	Odd one out	Links	Missing word	Equations	Series	Missing number	
Llama 2 (7b)	20%	25%	40%	10%	10%	10%	70%	20%	0%	15%	0%	89–110
Mistral 7B	30%	25%	50%	40%	30%	0%	20%	0%	0%	20%	0%	90–110
Mistral Small	40%	45%	20%	40%	20%	10%	30%	20%	10%	45%	0%	93–113
Mistral 8 × 7B	15%	80%	30%	20%	40%	10%	40%	0%	10%	50%	10%	94–114
Llama 2 (13b)	25%	65%	50%	20%	10%	30%	90%	0%	0%	20%	0%	95–115
Llama 2 (70b)	25%	70%	40%	30%	10%	20%	80%	40%	0%	35%	20%	96–117
Claude Haiku	45%	30%	60%	50%	30%	10%	80%	60%	20%	95%	0%	100–120
Gemini	40%	75%	80%	60%	10%	10%	40%	30%	50%	50%	0%	99–119
Mistral Medium	35%	50%	50%	20%	30%	20%	100%	70%	10%	55%	20%	99–119
GPT3.5	45%	35%	50%	40%	50%	30%	50%	20%	20%	85%	0%	99–119
Bing (Balanced)	60%	65%	50%	50%	20%	30%	100%	10%	40%	60%	0%	94–122
Mistral Large	30%	60%	60%	70%	20%	40%	100%	50%	10%	65%	10%	102–122
Bard	55%	70%	70%	40%	10%	0%	100%	70%	0%	90%	20%	103–123
Claude Sonnet	60%	40%	80%	30%	50%	30%	100%	80%	20%	80%	10%	104–124
Gemini Pro	55%	70%	70%	40%	80%	40%	80%	30%	40%	70%	0%	104–125
Bing (Precise)	85%	75%	80%	60%	40%	50%	100%	60%	50%	100%	10%	110–130
GPT4	70%	75%	80%	60%	30%	60%	100%	70%	60%	100%	0%	111–131
Claude Opus	65%	75%	90%	60%	60%	40%	100%	90%	70%	95%	20%	111–131
Bing (Creative)	90%	100%	70%	60%	60%	30%	90%	70%	60%	95%	20%	112–132

Better on verbal tests

Worse on numerical tests

Human average IQ is 100

CHATGPT IS A CHINESE ROOM

ChatGPT is a Chinese Room!



Simon Thompson

Follow

8 min read • Dec 18, 2022



That's a problem and you should know why.



What Stable Diffusion produces when asked about Chinese Rooms

Background Reading. See link on Blackboard.

THE TWO SIDES OF MEANING

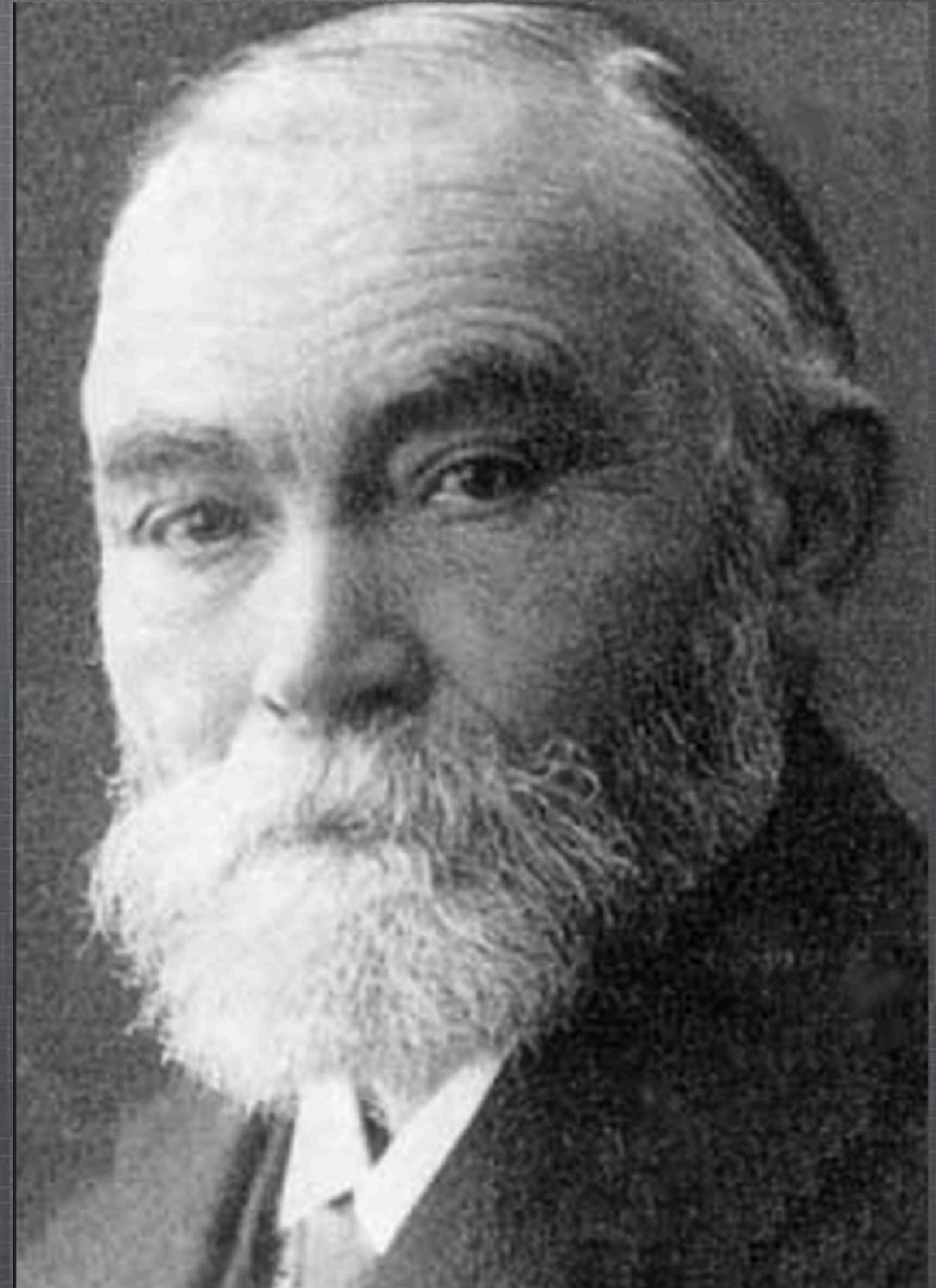
According to Frege

- There are two sides to meaning—**sense** and **reference**
- **Sense** is about the relationship between words
- **Reference** is the relationship of words to the world

ChatGPT 3.5 was trained on over 570GB of text data including more than 150,000 books

- LLMs know a lot about **sense** not much about **reference**
- Its surprising how much you can learn about the world through sense alone

The problem of reference is also known as the **symbol grounding problem**



Gottlob Frege (1848–1925)
Logician and philosopher of language

MARY AND THE BLACK AND WHITE ROOM THOUGHT EXPERIMENT

Imagine Mary, a brilliant scientist who has lived her entire life in a black-and-white room. She learns everything there is to know about the physical nature of the world through black-and-white monitors, becoming the world's leading expert on the neurophysiology of vision. She knows exactly what happens in the brain and eye when someone sees “red.”

One day, Mary leaves the room and sees a red object for the first time. The big question: Does Mary learn something new when she experiences red?

If yes, then it seems like there’s knowledge you can only get through experience (what philosophers call qualia—the subjective “what it’s like”).

If no, then all knowledge—including experience—can in principle be captured by physical facts alone.



"Mary's Room", originally described by Frank Jackson, is one of the most famous thought experiments in the philosophy of mind.

MARY AND THE BLACK AND WHITE ROOM THOUGHT EXPERIMENT

An LLM is a bit like Mary in her black-and-white room:

It has access to vast amounts of information about the world (including descriptions of experiences)

But it doesn't have first-person, lived experience—no seeing red, no feeling pain, no being situated in a body

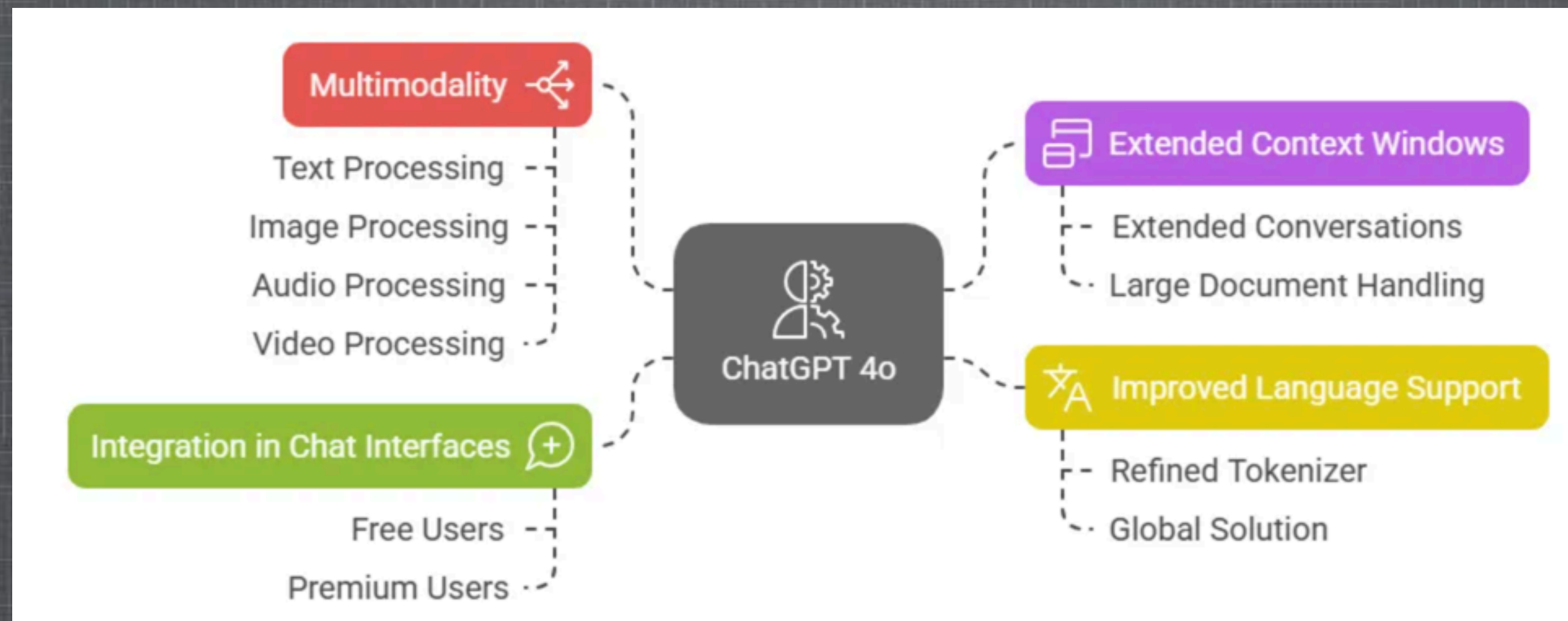
So the question becomes:

Is having all the descriptive knowledge enough for understanding, or is something like experience required?



GROUNDING THROUGH MULTIMODAL AI

Foundation models are increasingly multimodal. For instance. Since ChatGPT4o (omni) OpenAI's transformer-based models have been natively multimodal for text, speech, and images:



- **Next-Token Prediction:** Just as the model predicts the next word in a sentence, it predicts the "next patch" or pixel-set of an image based on the previous ones.
- **Hierarchical Generation:** It typically starts with a low-resolution "sketch" (coarse tokens) and progressively adds layers of detail (fine tokens). This is often called Visual Autoregressive (VAR) Modeling.

COULD MULTIMODALITY BE A ROUTE TO MEANING?

- An image-generating AI, knows the word “chair” but it can also generate images of what chairs look like
- This may not be reference in the full sense—an image of a chair is still not a chair—but there are more elements of chair-ness than before
- Is this meaning though? The philosopher Stevan Harnad would argue that the representations in all these generative AIs are still **not grounded**. Interpretation of their output is ultimately in the mind of the observer (the user of the AI) and not in the machine
- These AIs are just manipulating huge arrays of numbers, according to very complex rules, like Searle in the Chinese room manipulating meaningless squiggles...



“Chairs in space” visualised by
Google Gemini

ROBOTS AND THE CHINESE ROOM

“Searle's argument fails logically for the robot version of the Turing Test, for in simulating it [the robot] he [Searle] would either have to use its transducers and effectors (in which case he would not be simulating all of its functions) or he would have to be its transducers and effectors, in which case he would indeed be duplicating their causal powers (of seeing and doing).” (Harnad)



Stevan Harnad, 1945-

Video from 2018

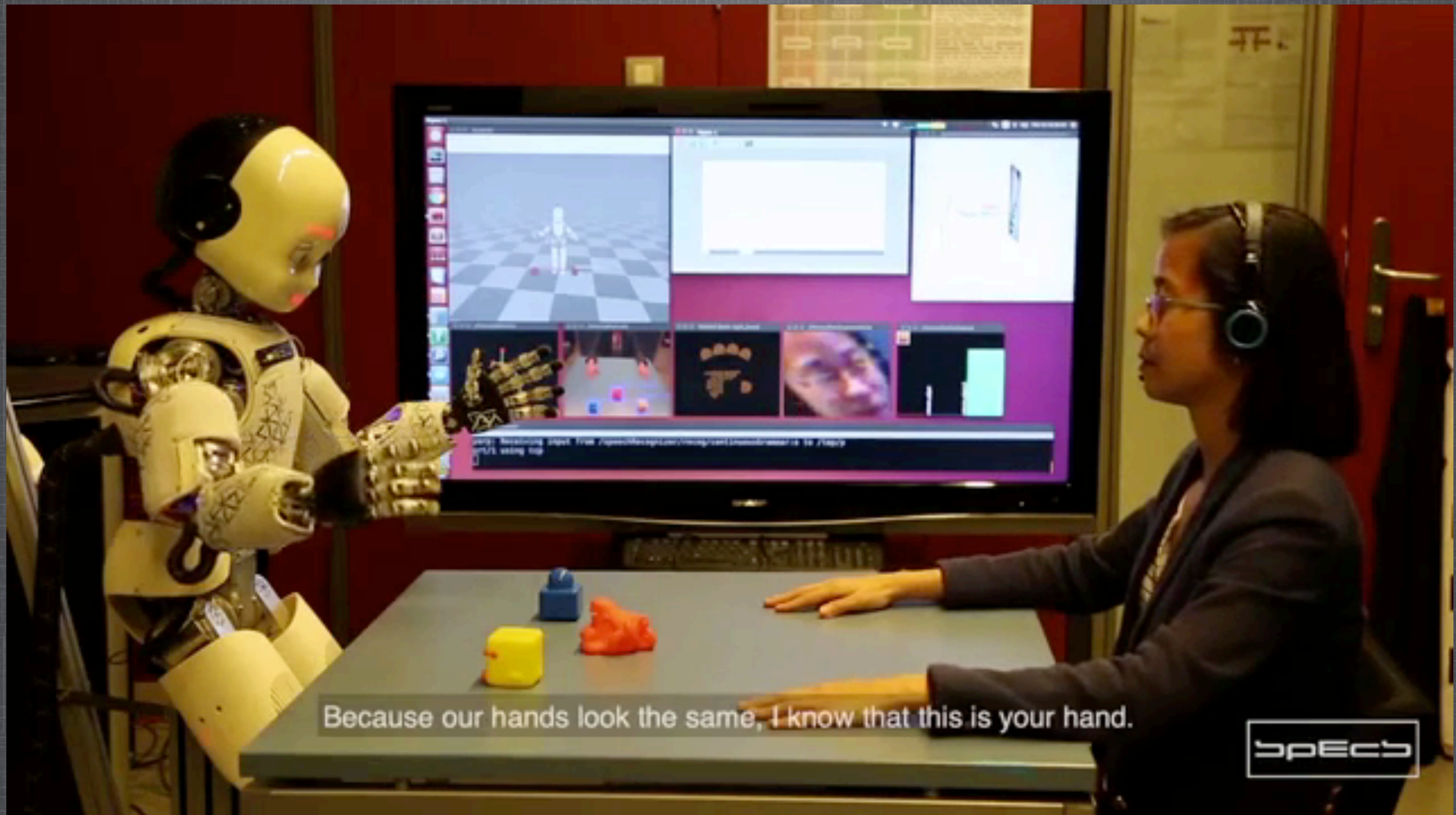
[Click to view on Youtube](#)

As AI becomes increasingly capable of acting in the real-world without human mediation, it may become more difficult to sustain any claim that AI is just **simulating** intelligence and is not a form of **actual** intelligence.



Using robotic transducers and effectors via telepresence

REFERENCE THROUGH ROBOTICS

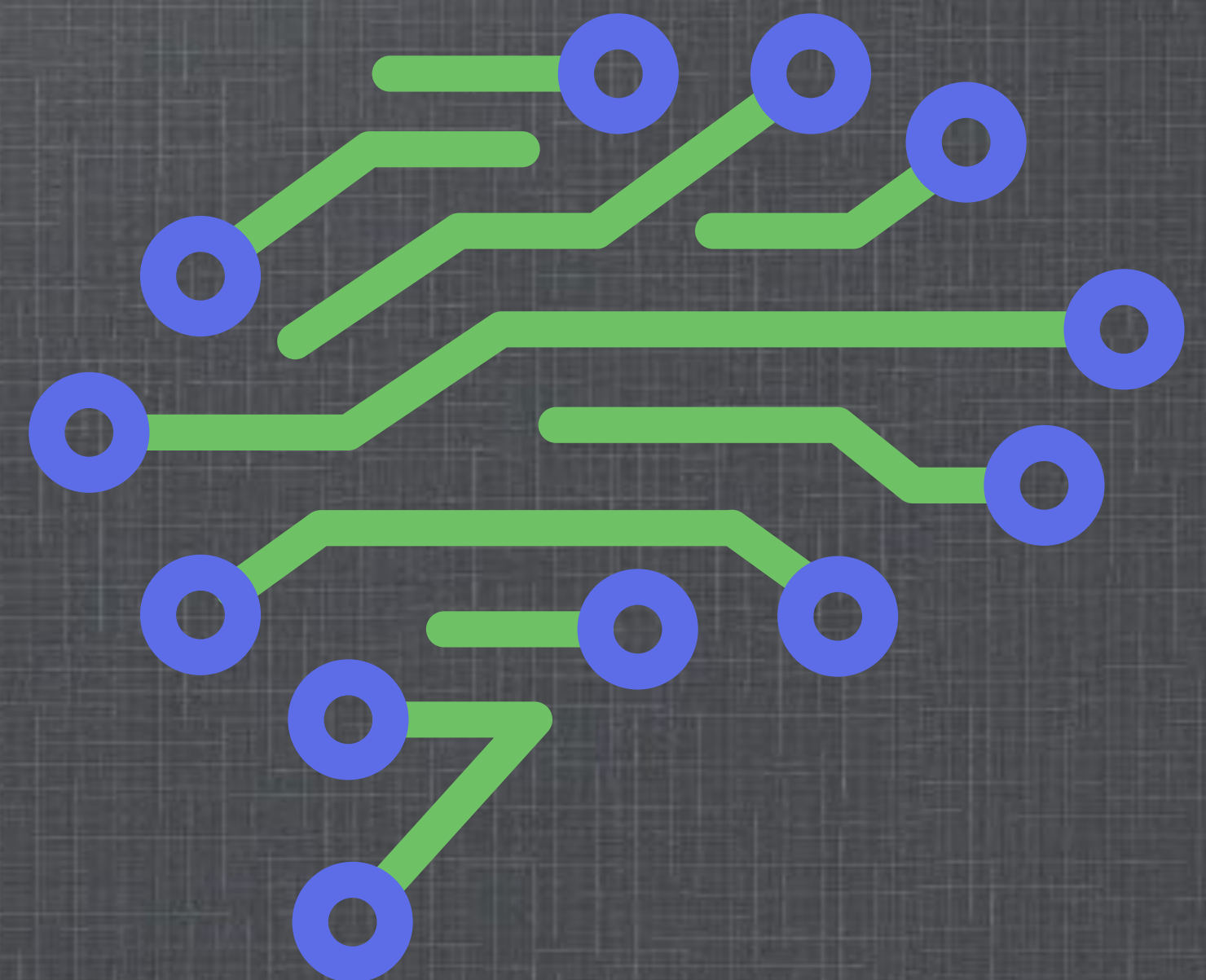
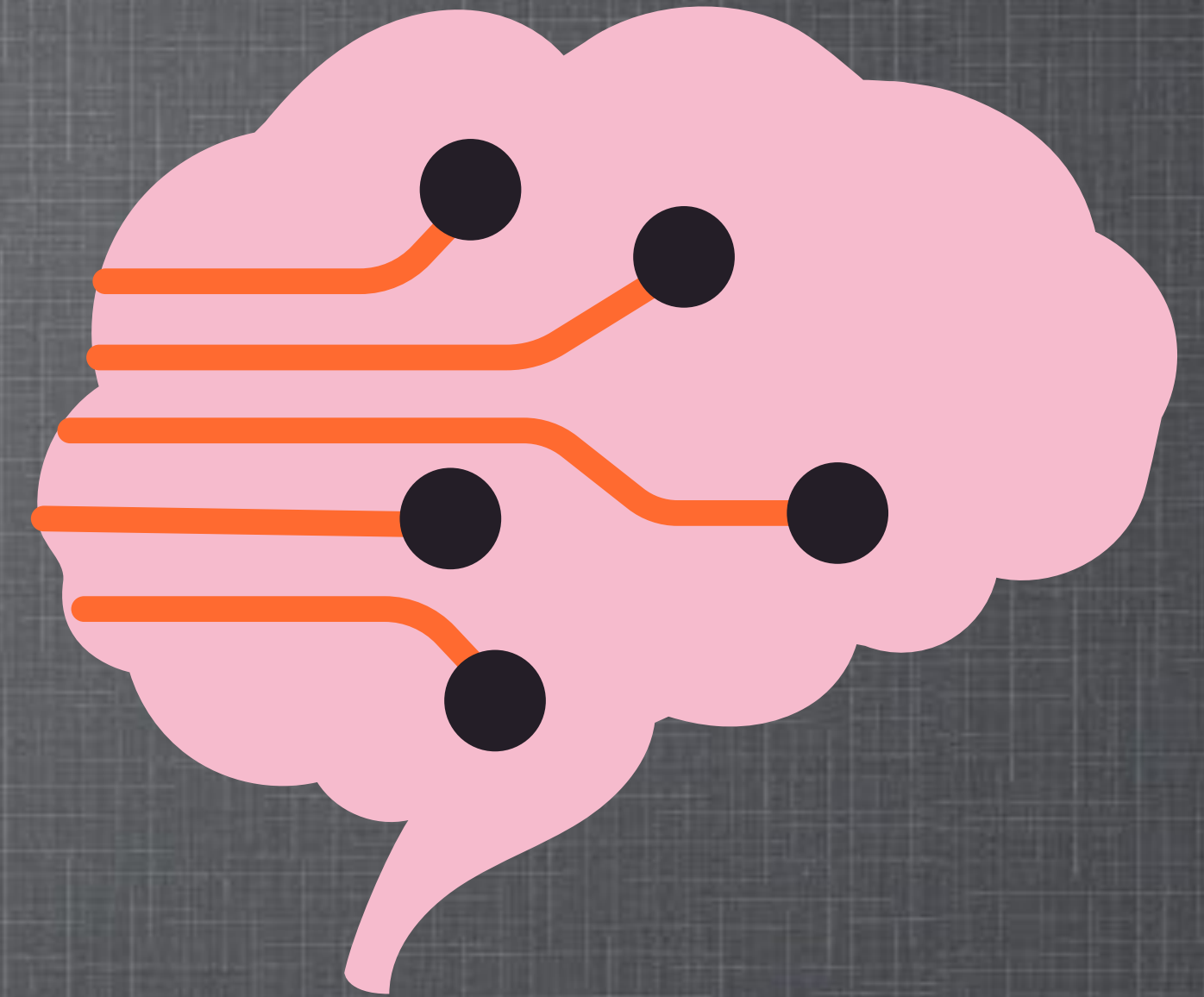


The robot has a model of itself and is able to learn about the relationship between its body parts and sensations it receives, and how its own body relates to another. These are some of the building blocks for a sense of self.

Moulin-Frier et al. (2017) *IEEE Transactions in Cognitive and Developmental Systems*.

SUMMARY

- Psychologists see intelligence as both one thing and many but fundamentally it's an aggregate capability to act purposefully and achieve goals in the world
- The human brain is a bioelectronic organ, its core role is to perform the computations that allow the existence of mind. Human cortex can be considered as a hugely interconnected network for deep neural networks. AI has co-opted this scheme
- There are two parts to meaning—sense and reference. AI knows about sense and may achieve reference through robotics



THANKS TO...



Stuart Wilson



Leah Krubitzer



Daniel Camilleri



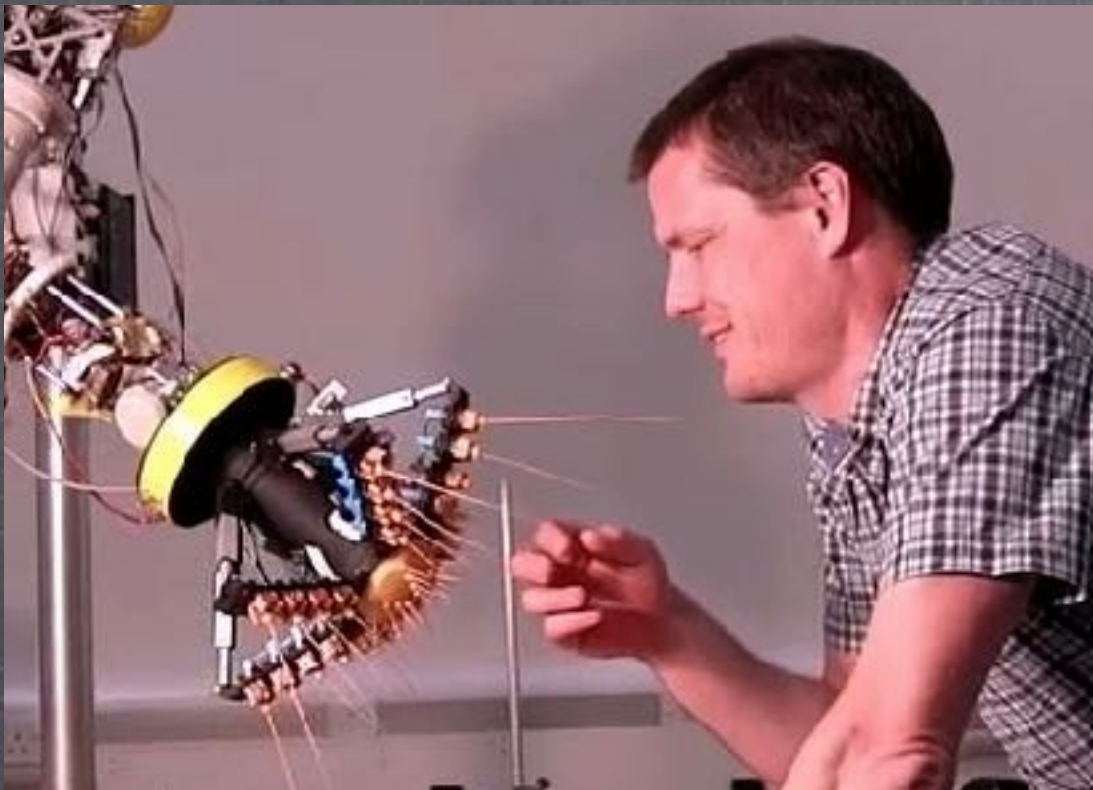
Paul Verschure



Martin Pearson



Alejandro Jimenex-Rodriguez



Ben Mitchinson

