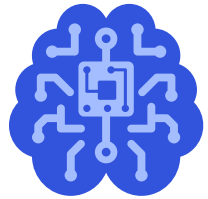
October 2019 @qualcomm_tech Amsterdam

Qualcomm

The Next Generation Convolutions

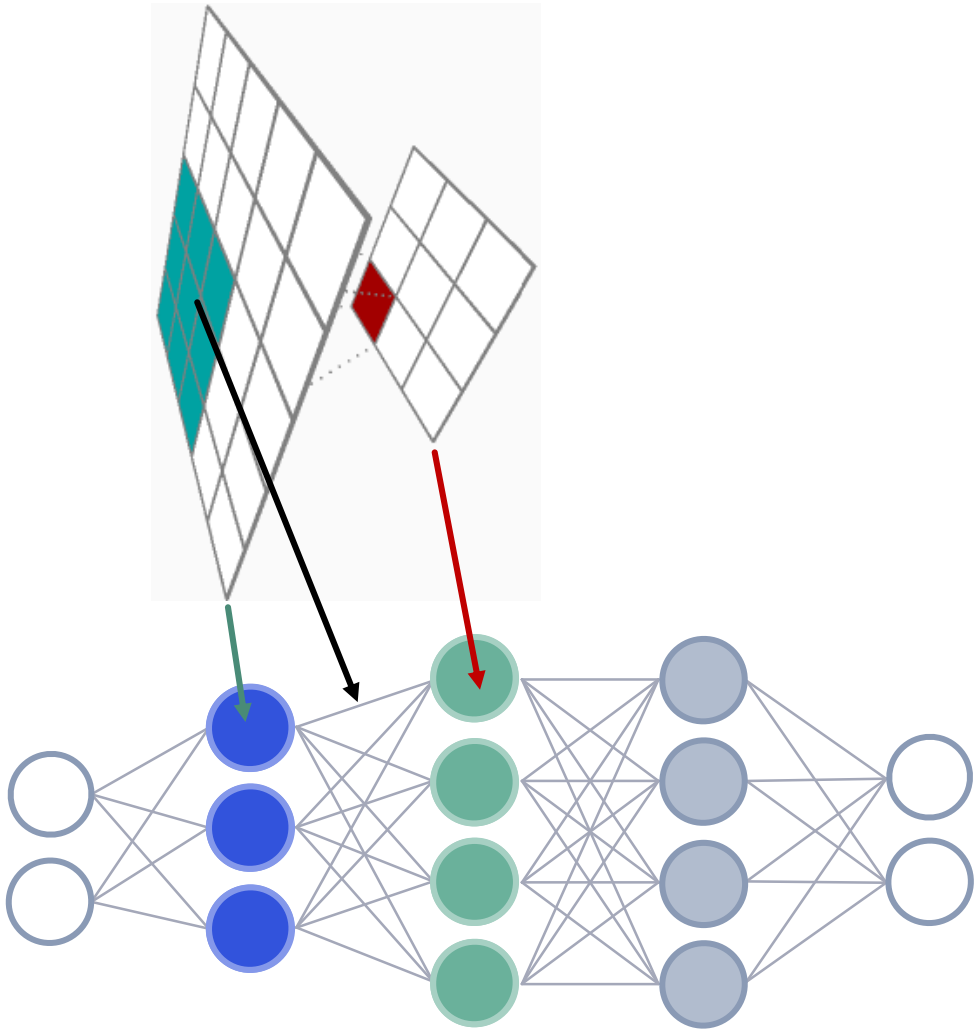
Max Welling
VP Technology
Qualcomm Technologies Netherlands



Convolutional Neural Networks

(Y. LeCun ’95)

- Each neuron is an image / feature map / channel
- Each synapse / edge is a filter
- The filter slides over the image
- At every location it multiplies the image values with the filter values and adds them up



Convolutions

- Convolution is arguably the technology that is responsible for much of the deep learning success.
- *Can we generalize convolutions?*

Generalized Convolutions

- Graph convolutions
- Group convolutions
- *Gauge convolutions*
- *Quantum convolutions*

- Convolutions are expensive to run on hardware.
- *Can we build radical new hardware to improve the efficiency of computing convolutions?*

- In memory convolutions
  - Analogue calculations
  - Brain inspired distributed memory
  - Breaks with Von Neumann architecture

Computers are consuming an increasing amount of energy

By 2025, the data center sector could be using 20% of all available electricity in the world[1]

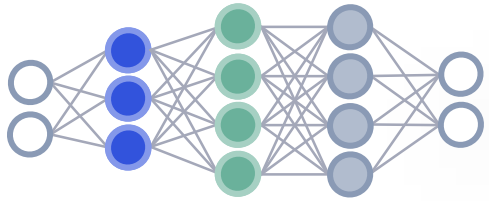
A cloud provider used the equivalent energy consumption of ~366,000 US households in 2014[2]

Bitcoin mining in 2017 used the same energy as did all of Ireland[3]

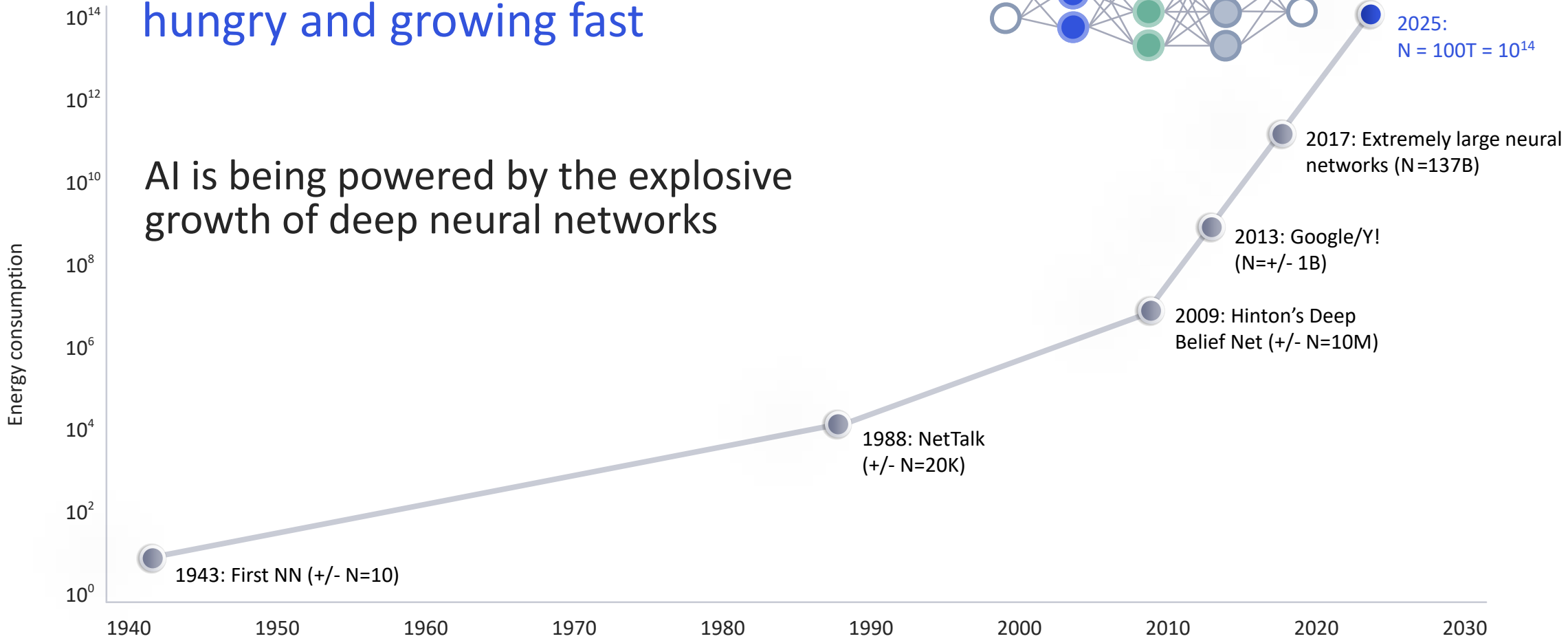
Given the economic potential of AI, these numbers will only be increasing

1. Andrae, Anders (2017) Total Consumer Power Consumption Forecast; 2. The Verge (2014); 3.The Guardian (2017)

Deep neural networks are energy hungry and growing fast



AI is being powered by the explosive growth of deep neural networks

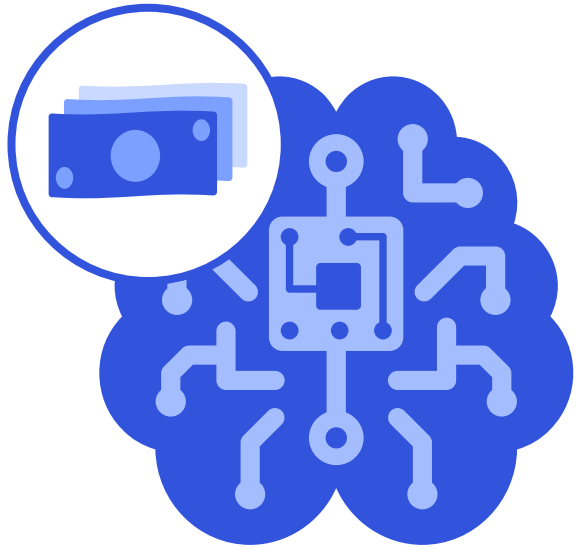


Source: Welling

2025 | Will we have reached the capacity of the human brain?

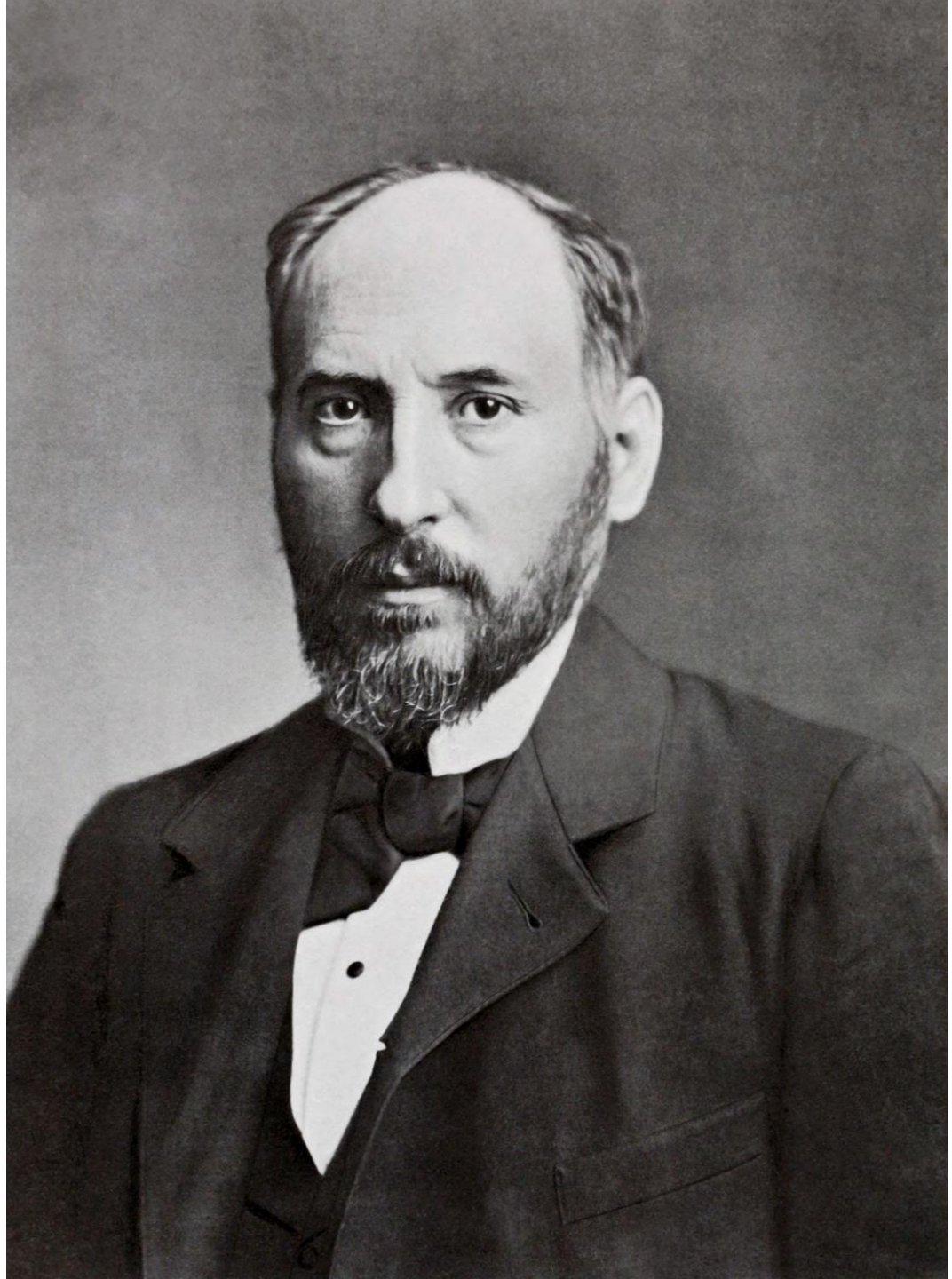
Energy efficiency of a brain is 100,000x better than current hardware

Value created by AI must exceed the cost to run the service



Economic feasibility per transaction may require cost as low as a micro-dollar (1/10,000th of a cent)

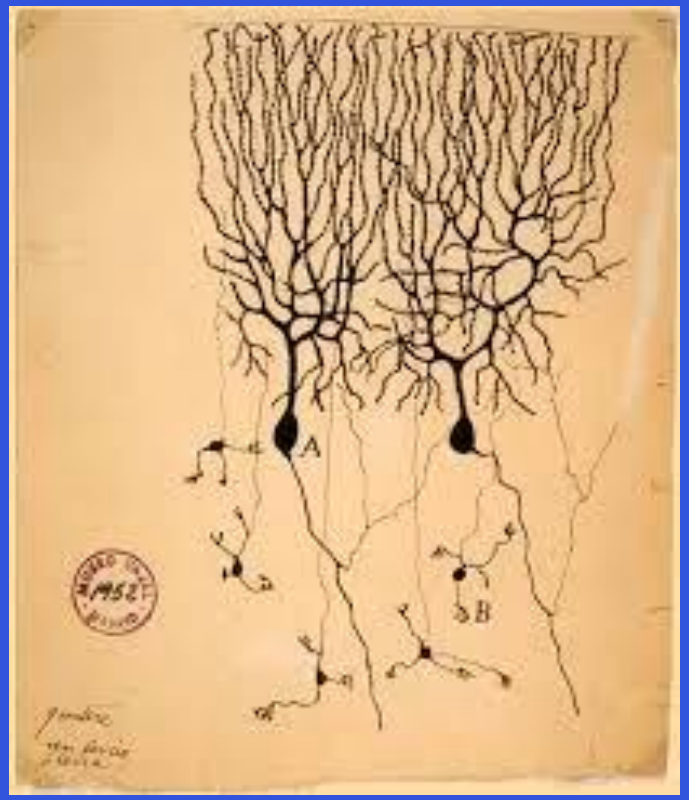
We need more AI per Joule!



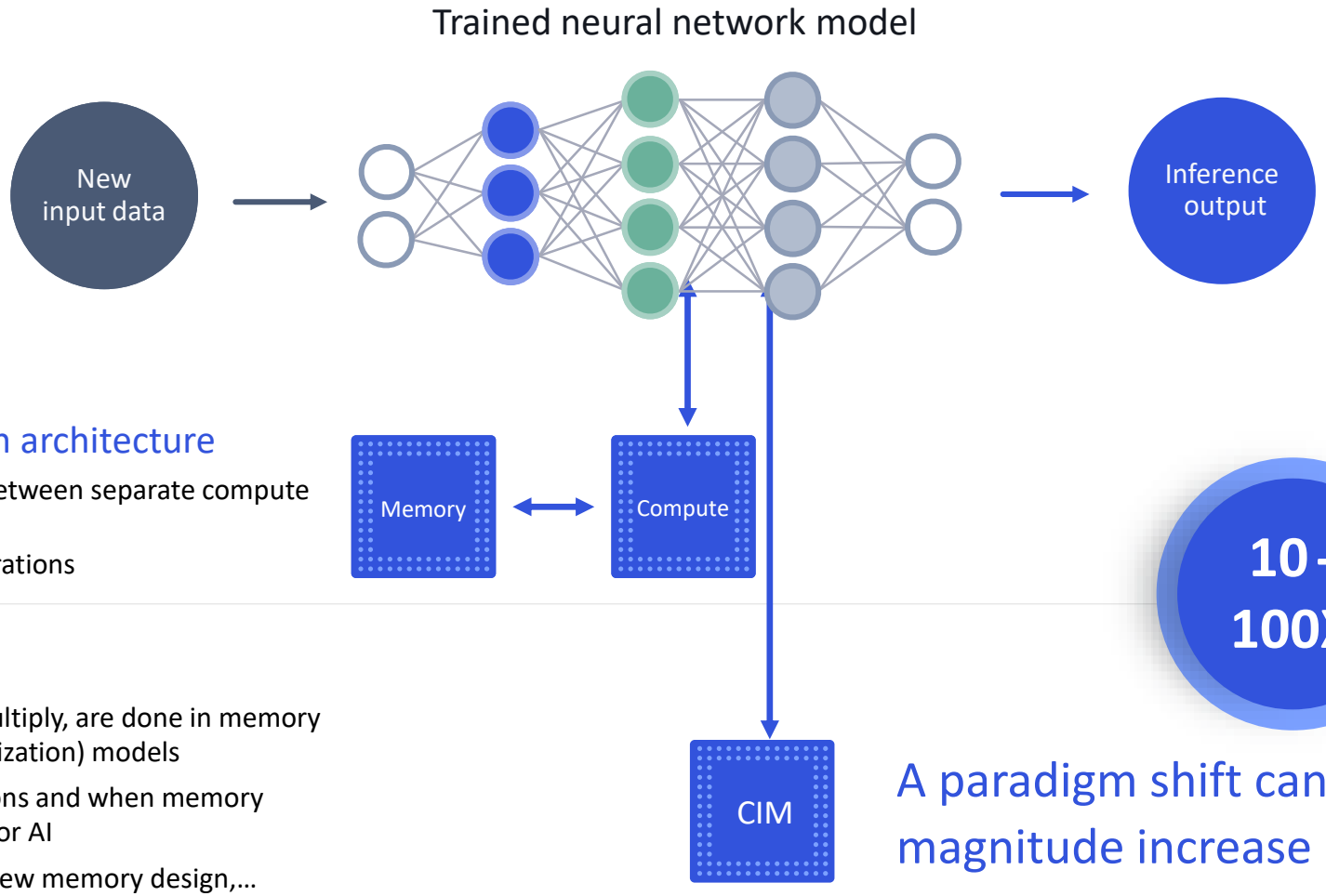
Santiago Ramón y Cajal

Father of neuroscience

- Discovered the neuron



Von Neumann architectures are not ideal for deep learning



Traditional Von Neumann architecture

- Data shuffled back and forth between separate compute and memory
- Good for general purpose operations

Compute-in-memory

- Computations, like add and multiply, are done in memory and need low-bit width (quantization) models
- Good for simple math operations and when memory becomes bottleneck—typical for AI
- Research in analog compute, new memory design,…

10 - 100X

Power efficiency improvement for 1-bit ops[1]

A paradigm shift can bring orders of magnitude increase in power efficiency

We need architectures that look more like the brain!

1 Compared to traditional Von Neumann architectures today

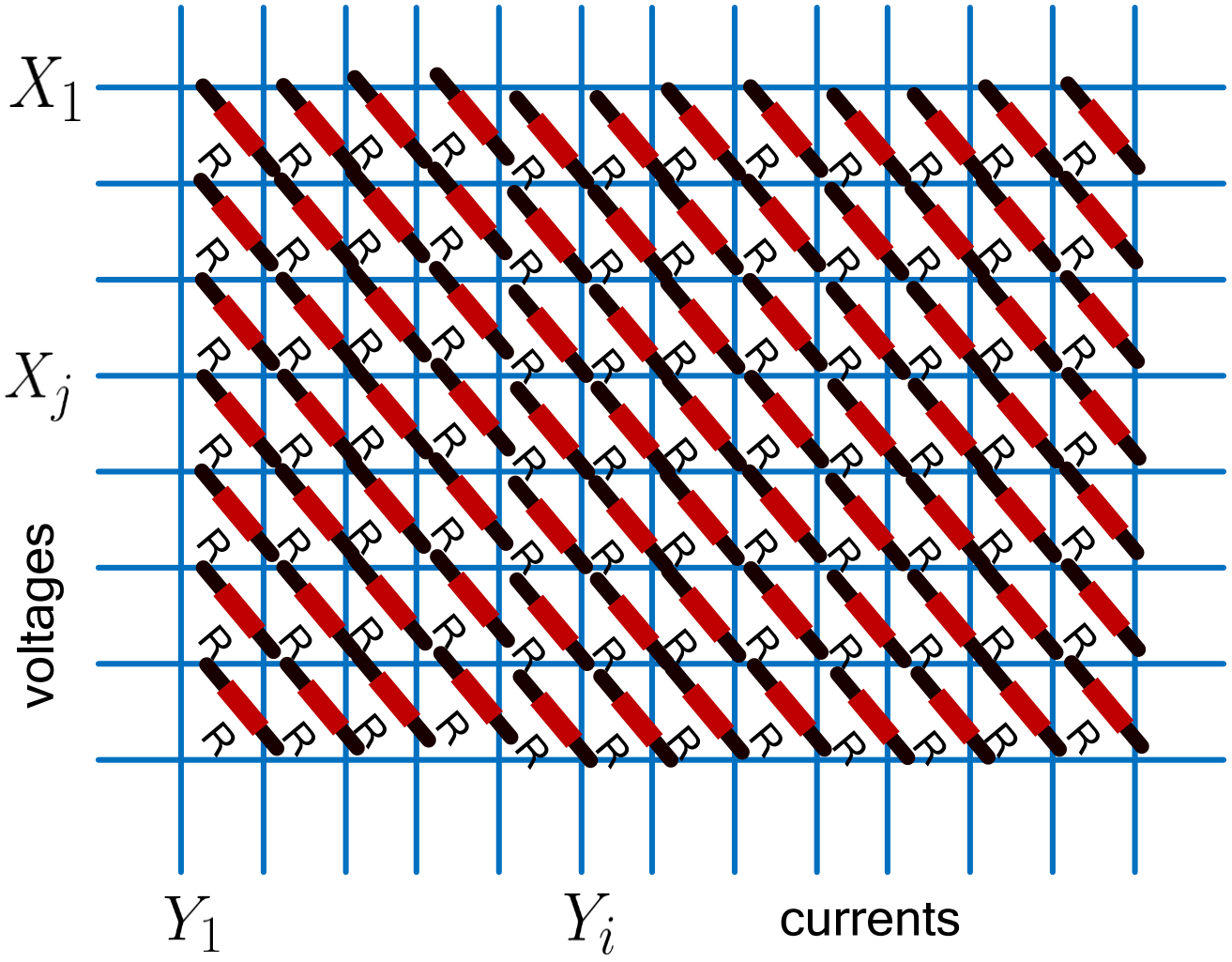
In Memory Computation

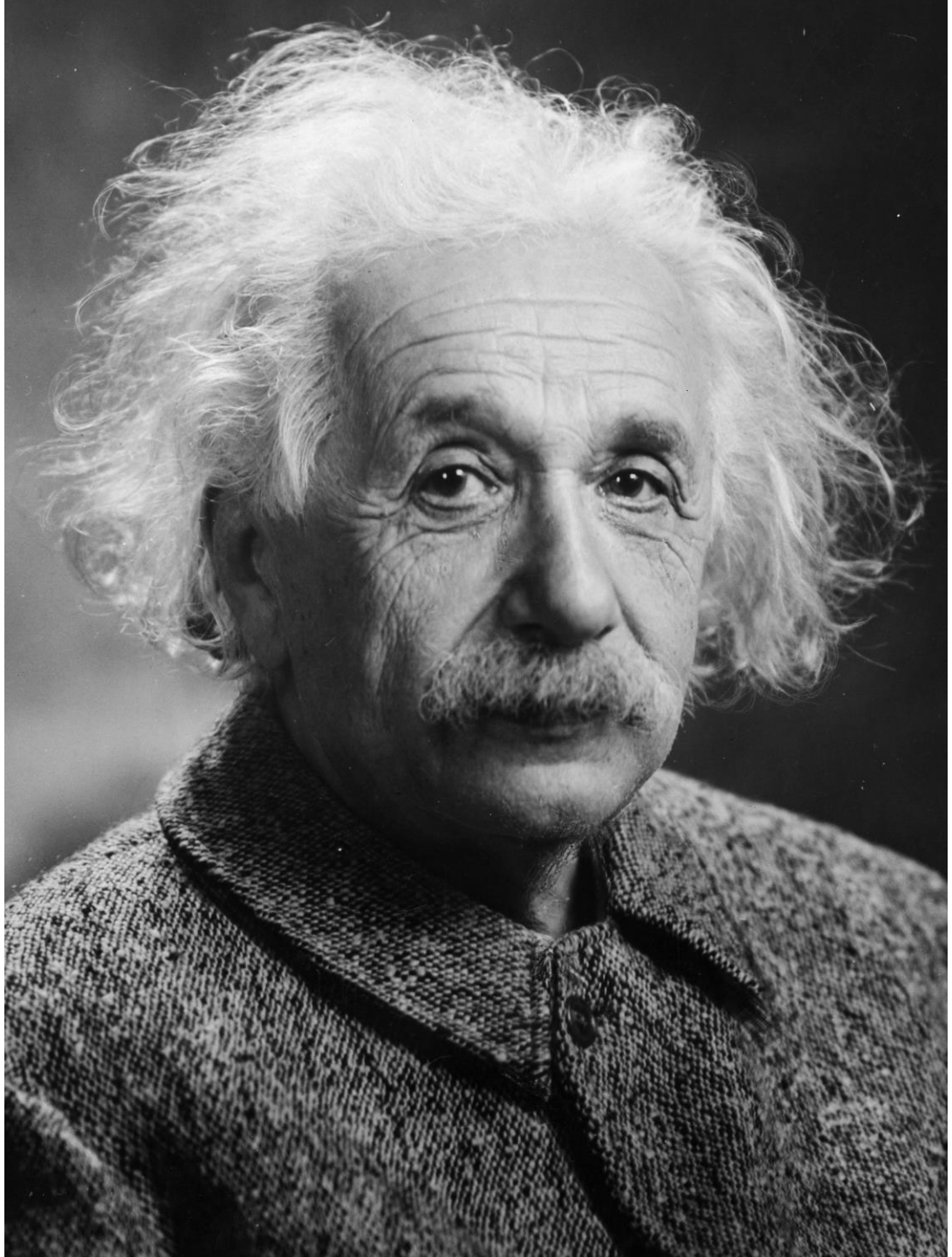
Memristor: $Y = W \cdot X \Rightarrow Y_j = \sum_i W_{ij} X_j$

Conductance: $W_{ij} = \frac{1}{R_{ij}}$

- Inputs (neural activations) are voltages (word lines)
- Weights are conductances (synapses)
- Outputs are currents and compute Y = W * X (bit lines)

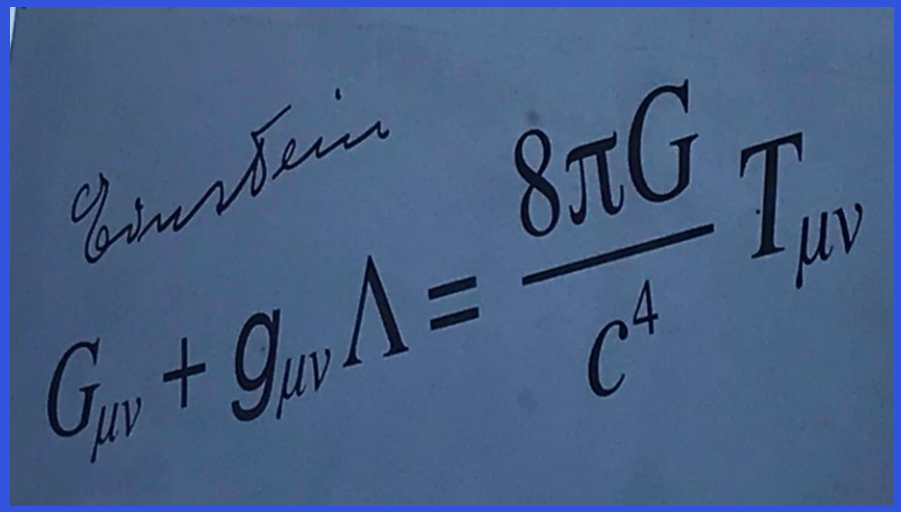
- Computation is analogue (uses less power)
- Memory is where computation is performed (like a brain)
- Deep learning can compensate for extra noise (think dropout)





Albert Einstein

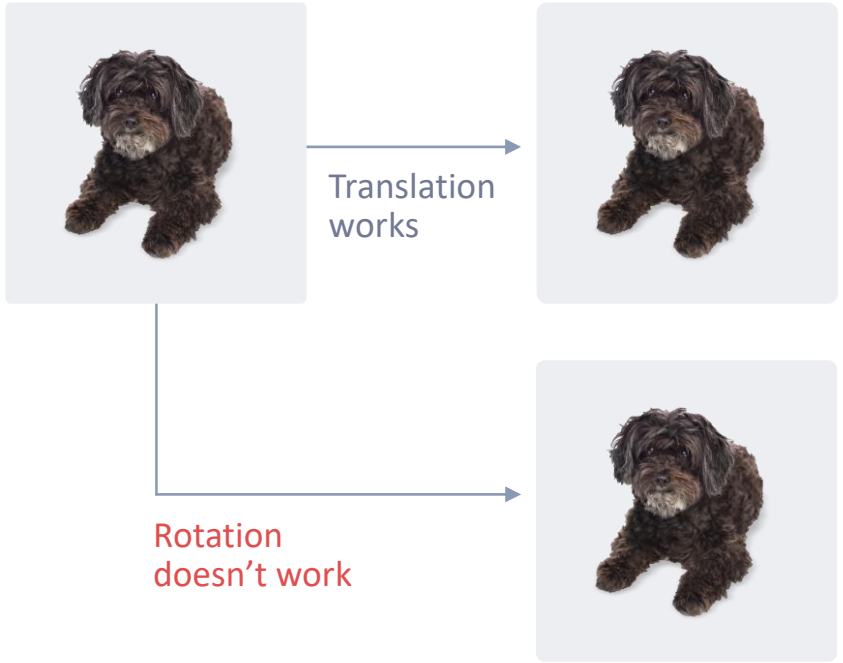
Inventor of General Relativity



Today's deep learning

Traditional CNNs

Produce state-of-the art results but…
do not generalize input like rotations



(Convolutional neural networks would need to be retrained with
new rotated images to determine new set of parameters—like filter weights)

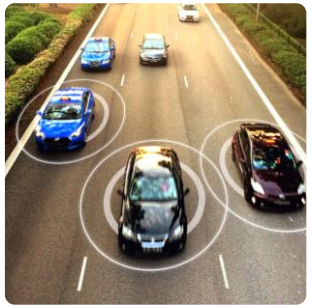
Applying foundational mathematics of physics

Like quantum field theory, to deep learning

Tomorrow's deep learning

Gauge Equivariant CNNs

No matter how you rotate or move the object,
the generalized model will still identify the object



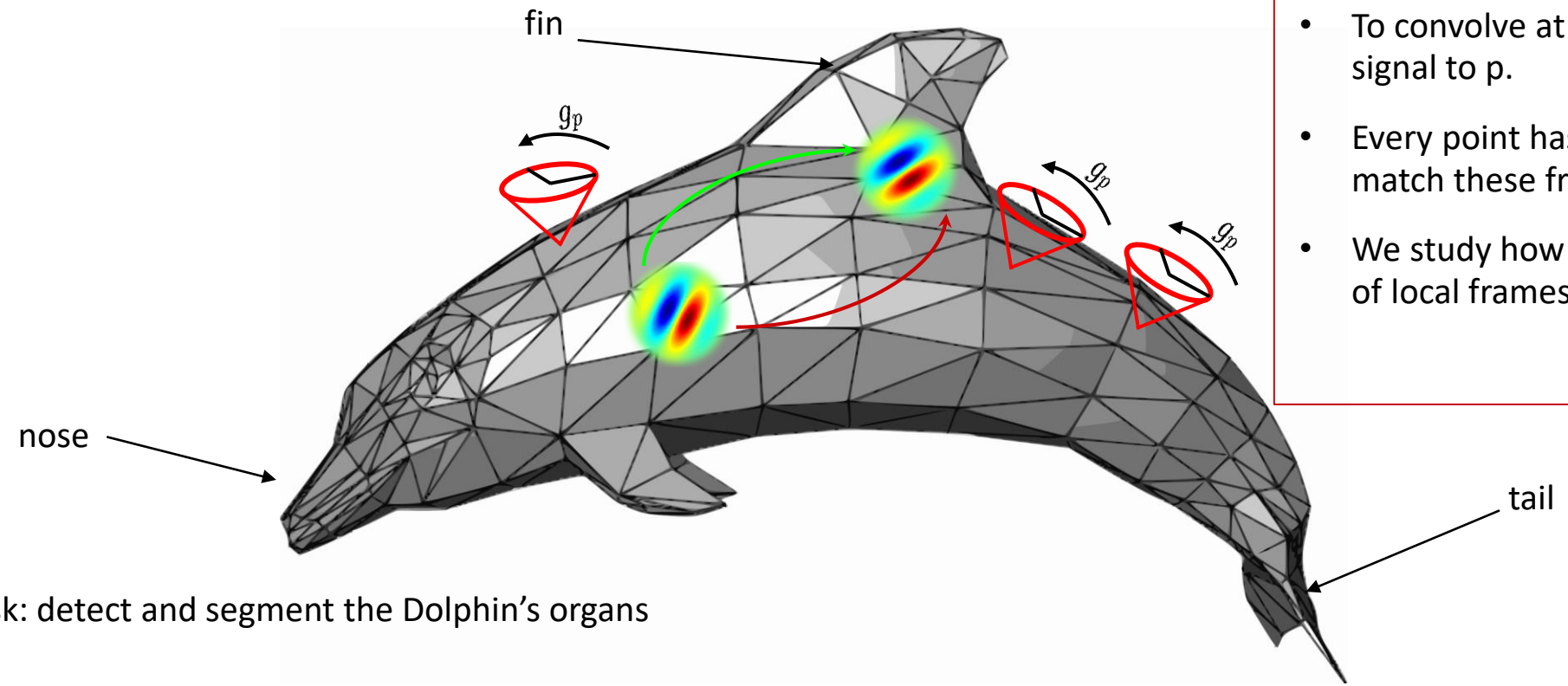
Rotated objects and images applicable to drones, robots, cars, fisheye-lens cameras. VR, AR,..

(Generalized CNNs (G-CNN): Gauge equivariant CNN, Group, and Steerable CNN pioneered by Qualcomm AI Research do not need to be retrained)

Convolutional NNs on curved surfaces

Gauge Equivariant CNNs for deep learning on curved spaces

- In the future “images” may become 3D (e.g. with depth)
- Can we do deep learning on 2D surfaces like we do on images?
- Yes, but we need to use the same theory Einstein used in General Relativity



- Kernels rotate when we transport them over a curved surface
- To convolve at p, we need to transport the signal to p.
- Every point has its own frame, so we must match these frames
- We study how results change under changes of local frames (a.k.a. gauge transformations)

Task: detect and segment the Dolphin’s organs

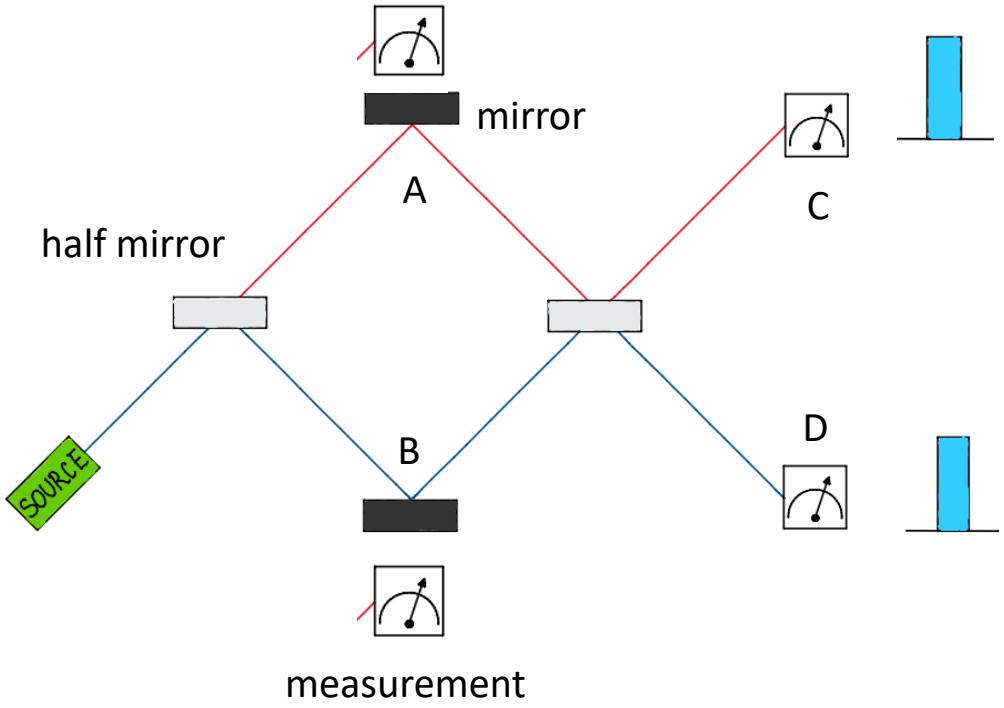
Niels Bohr, Erwin Schrödinger

Inventors quantum mechanics (among others)

$$i\hbar\frac{d}{dt}\psi(t) = H\psi(t))$$

The Crazy world of Quantum mechanics

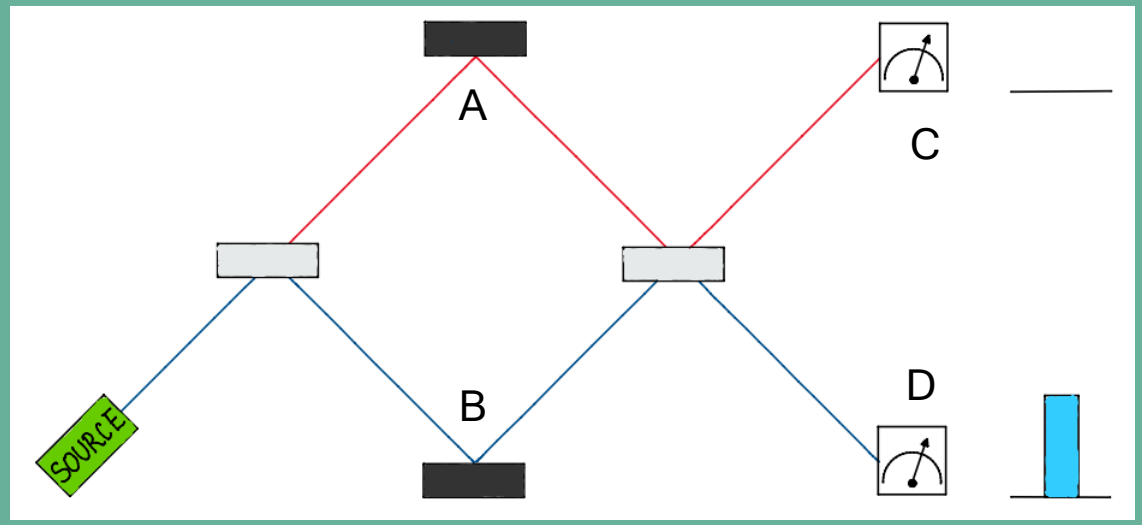
$$P(C) = P(A)P(A \to C) + P(B)P(B \to C)$$



Dutch book argument:
quantum mechanics is a consistent theory of probabilities

No Measurement at A,B

$$P(C) \neq P(A)P(A \to C) + P(B)P(B \to C)$$



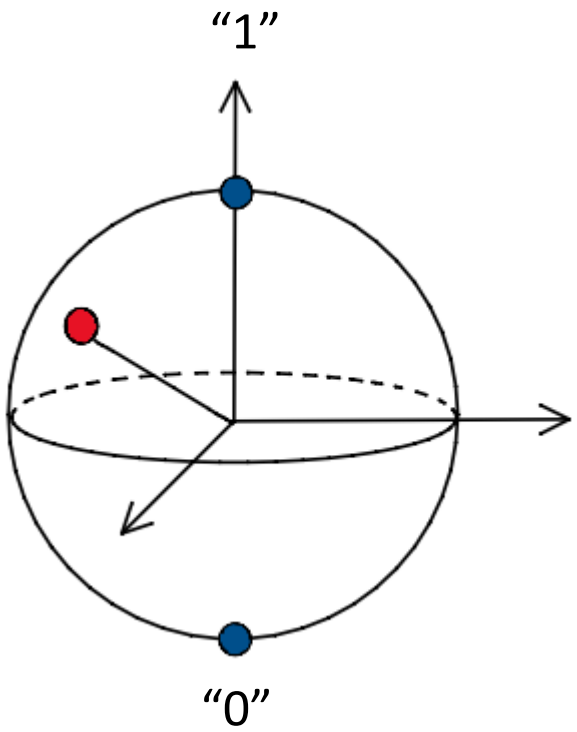
Quantum statistics is consistent but weird:
when not measuring the photon it is in both paths simultaneously!

Quantum computation

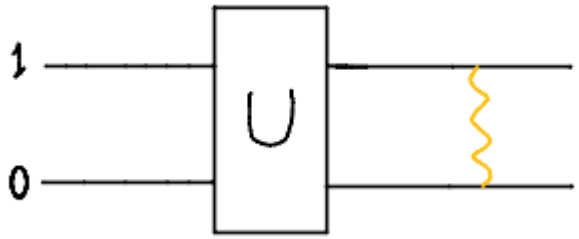
A qubit is more like a sphere than a bit.

- Qubit:

a ”0” + b “1”



- Entanglement:

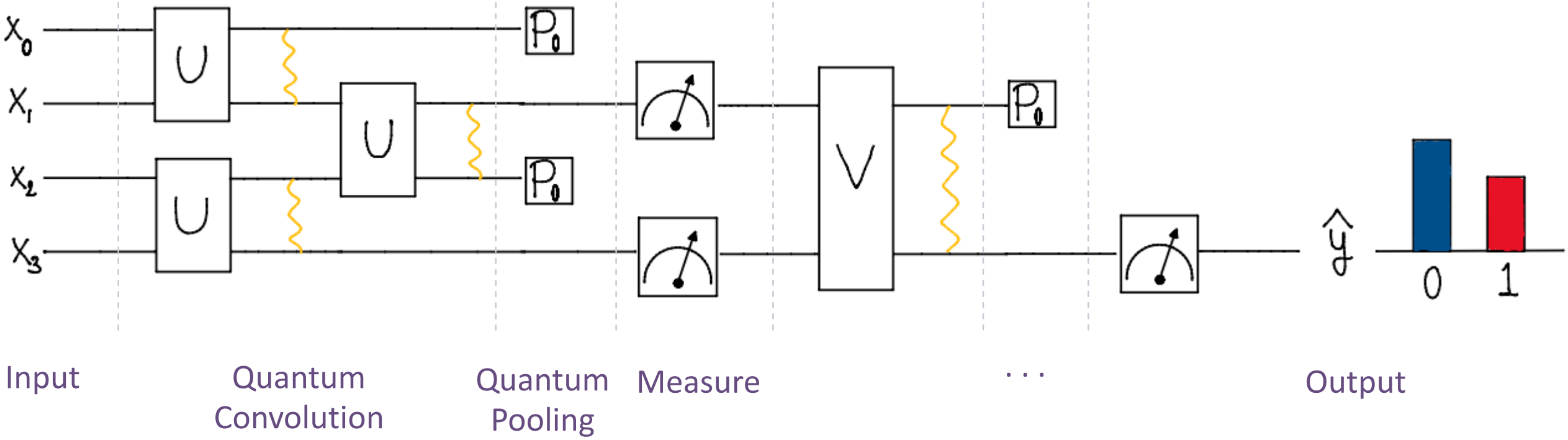


Entangled qubits span all possible combinations:

“00”, “10”, “01”, “11”

Quantum inspired convolutional kernels

Can we leverage the strange world of quantum statistics to design new quantum convolutions?



- Quantum statistics opens new doors to deep learning, even classically!
  (but with the benefit that it runs efficiently on a quantum computer)

Convolutions are the core of Deep Learning’s success

AI algorithms and hardware need to be energy efficient: perform convolutions in memory!

We can generalize convolutions to curved spaces. Applications in e.g. medical imaging.

Quantum mechanics provides a new theory of probability, leading to “quantum convolutions”.

Qualcomm

Thank you!

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog