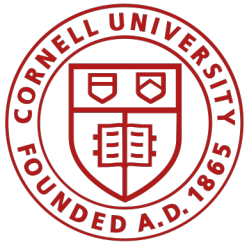


Cornell University

BOSTON UNIVERSITY

Optical Neural Networks: Neuromorphic Computing and Sensing in the Optical Domain

Tianyu Wang

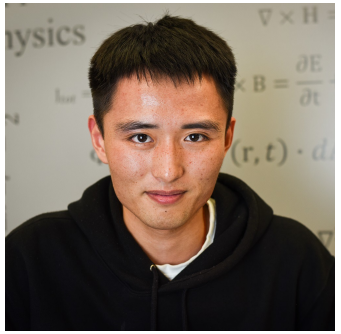
Assistant Professor in Electrical and Computer Engineering, Boston University

January 5, 2024

Acknowledgements

Prof. Peter McMahon

Cornell University



Shi-Yuan Ma

Maxwell Anderson

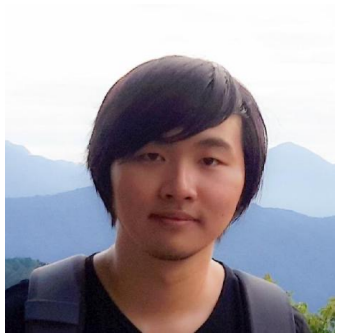
Mandar Sohoni

Martin Stein

Brian Richard

Prof. Logan Wright

Yale / NTT



Dr. Tatsuhiro Onodera

Cornell / NTT

Dr. Jérémie Laydevant

Cornell / USRA

Funding Agencies

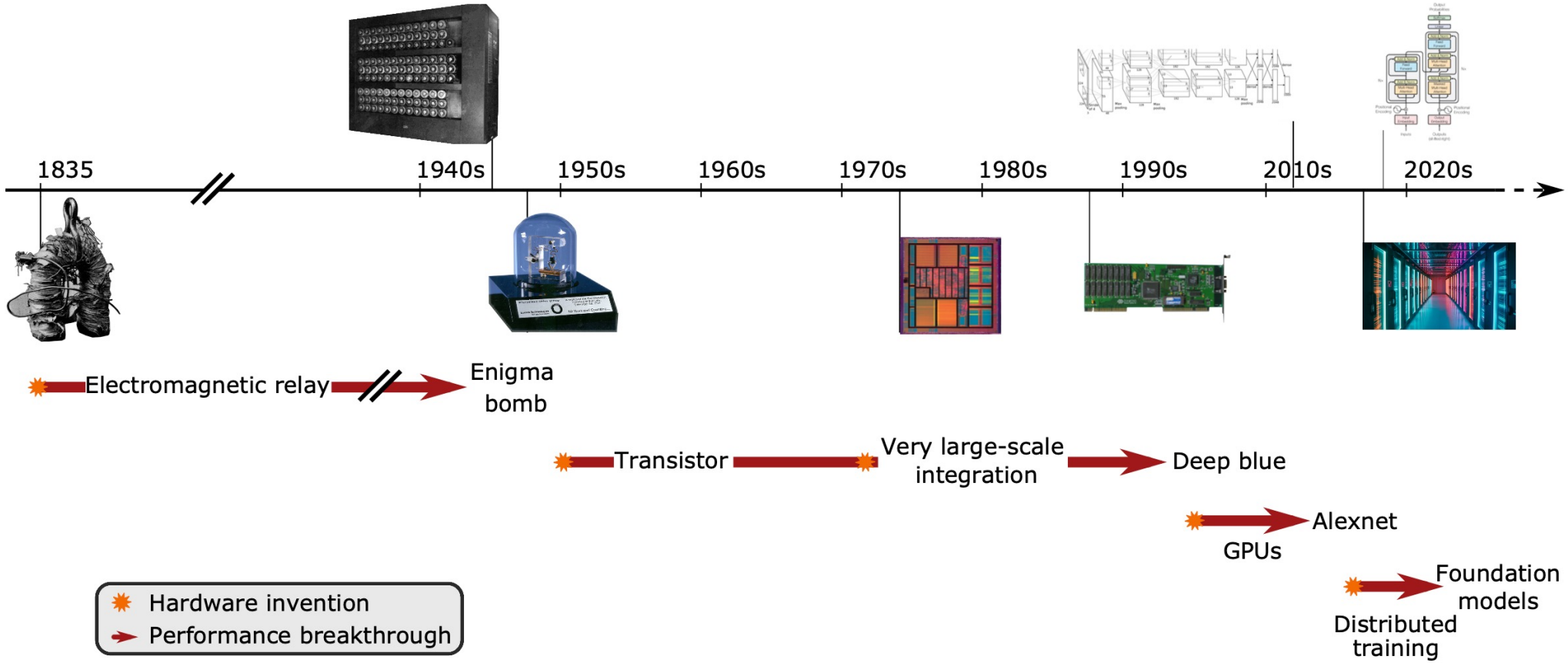
CORNELL UNIVERSITY FOUNDED A.D. 1865

NSF

KIC Kavli Institute at Cornell for Nanoscale Science

NTT NIPPON TELEGRAPH AND TELEPHONE CORPORATION

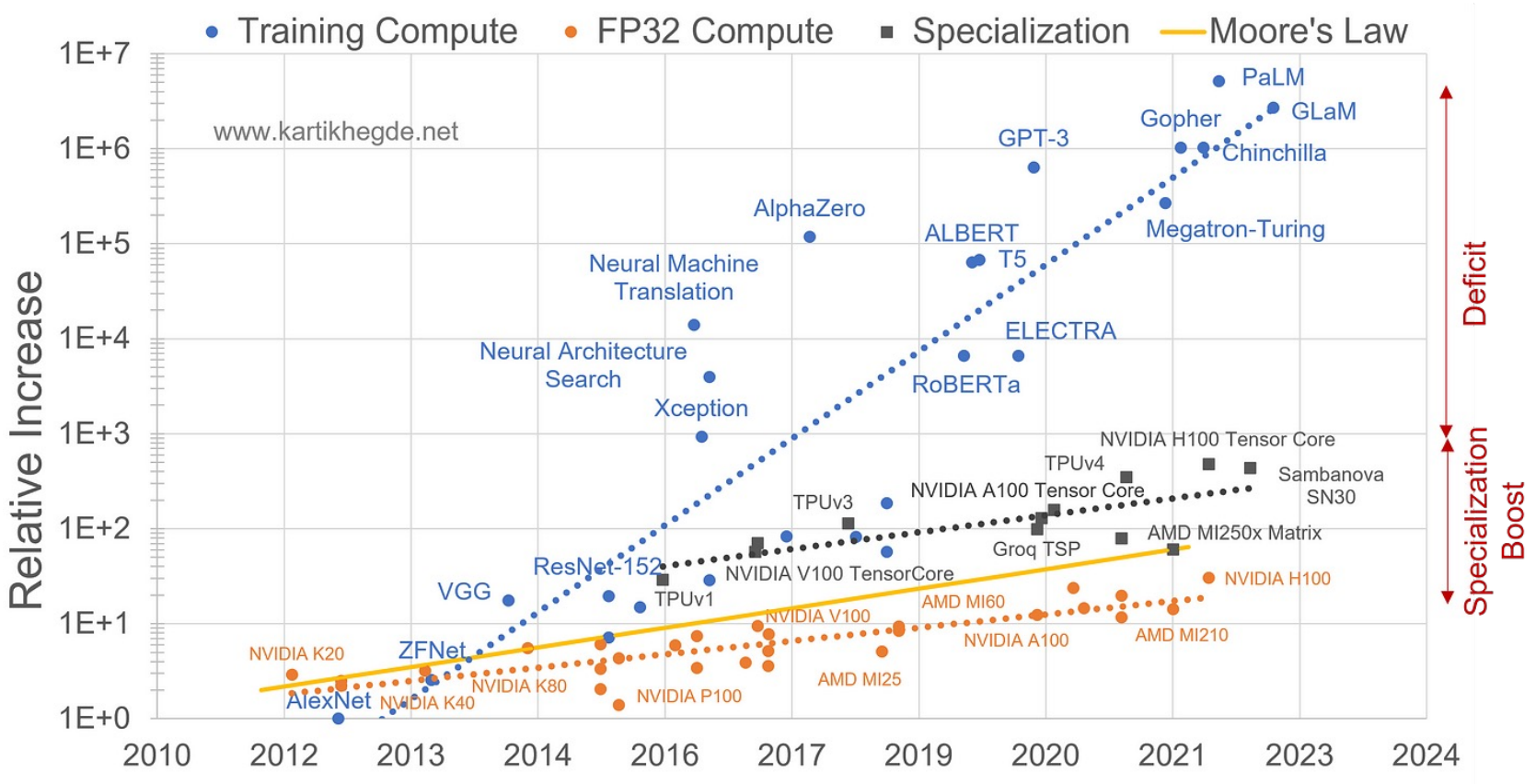
The co-evolution of computing hardware and software



The continuous scaling of hardware empowers emerging algorithms.

J. Laydevant, L. G. Wright, T. Wang, P. L. McMahon “The hardware is the software” in *Neuron*

An enlarging gap between AI hardware and software



Accelerating Deep Learning in the Post-Moore's Law World.

URL: https://kartikhegde.substack.com/p/accelerating-deep-learning-in-the

"The computing gap" motivates us to find creative solutions to hardware scaling.

Physics limits the scaling of computing hardware

At the microscopic scale, physics dominates the behavior of devices.

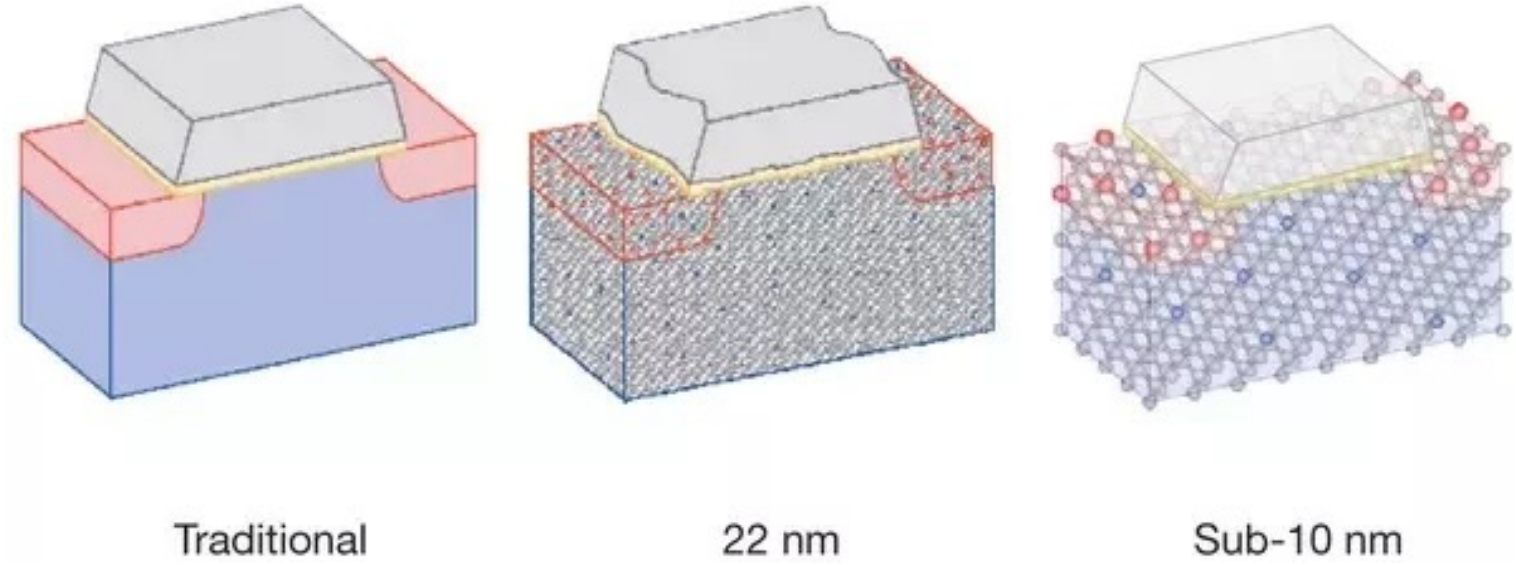


Figure credit: I. L. Markov, "Limits on fundamental limits to computation." *Nature* **512,** 147-154 (2014).

Noise: a fundamental challenge for physical computing

Thermal noise

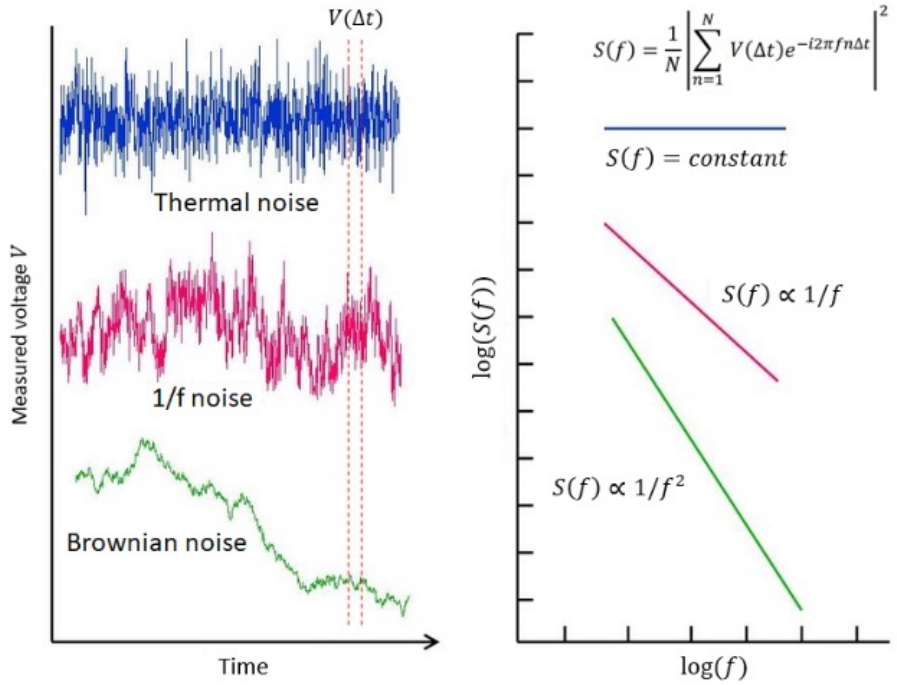


Figure credit: P. Popesso et al. arXiv:1211.4257 (2012)

Quantum noise

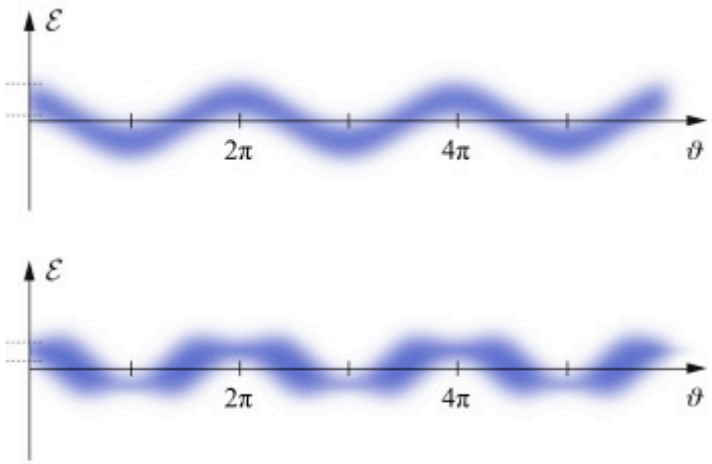


Illustration of quantum noise in coherent and squeezed light (Wikipedia)

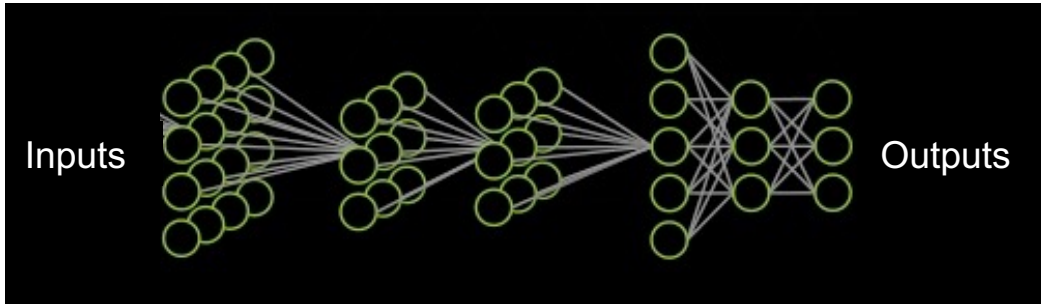
At a microscopic scale of space, time, and energy, stochasticity is ubiquitous, which prevents us from making perfect deterministic computing devices.

Instead of trying to remove unremovable noise,

can we instead dance with it by using a suitable computation model?

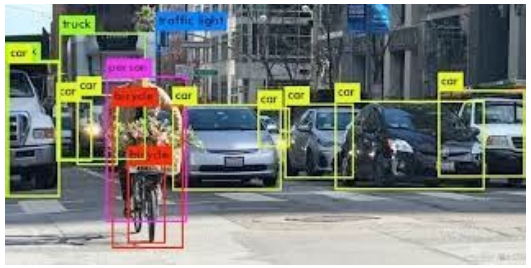
Unsurprisingly, neural networks, but why?

Neural networks are useful

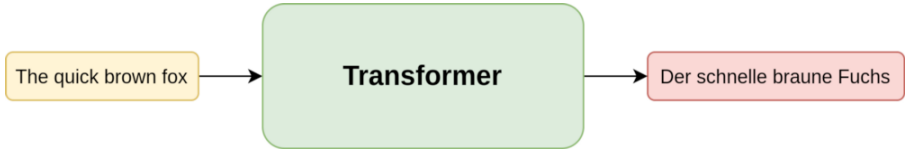


A deep neural network

Computer Vision



Natural Language Processing



Strategy Making

Illustration of AlphaZero

Physical System Simulation

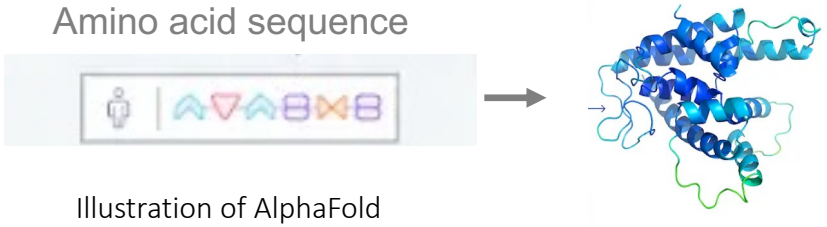


Illustration of AlphaFold

Protein 3D structure

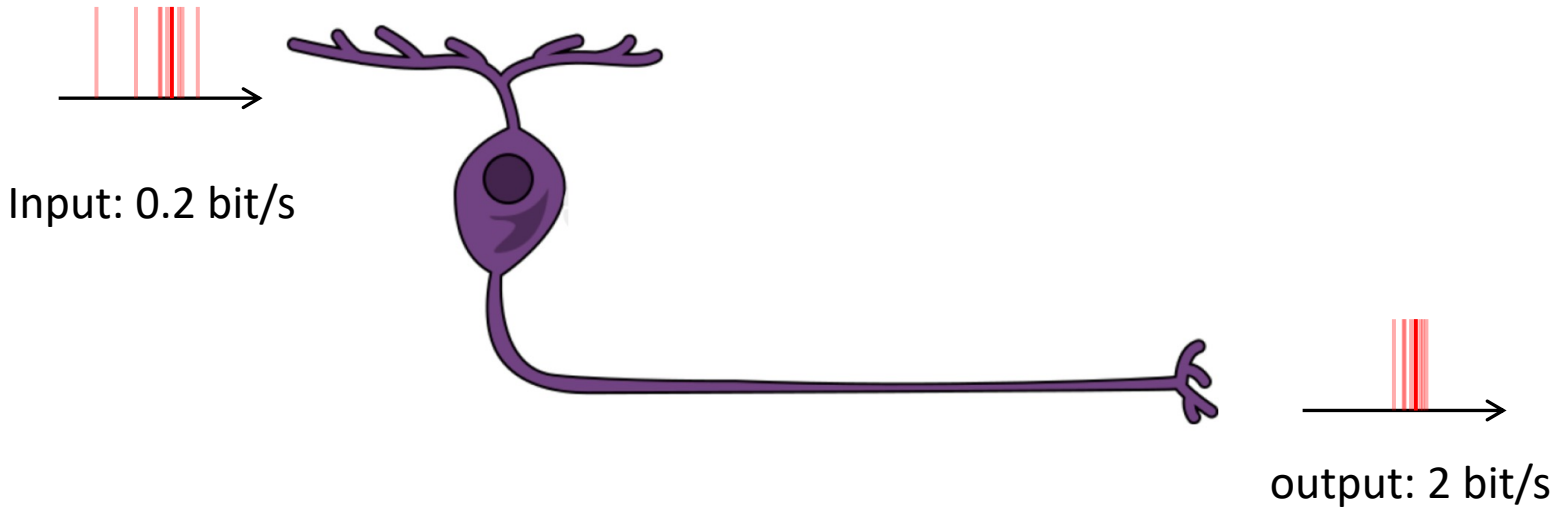
Generative Tasks

```
Prompt:
a photograph of
an astronaut
riding a horse
```

Generated by stable diffusion

Neural networks are resilient to noise

Biological neurons communicate with 2-4 bits of precision, but we are still able to perform all different tasks.



Sterling & Laughlin. *Principles of Neural Engineering* (2015)

Neural networks are resilient to noise

Deep neural networks can be trained to tolerate numerical errors (typically caused by the low bit depth of a digital hardware):

Illustration of the procedure of quantization

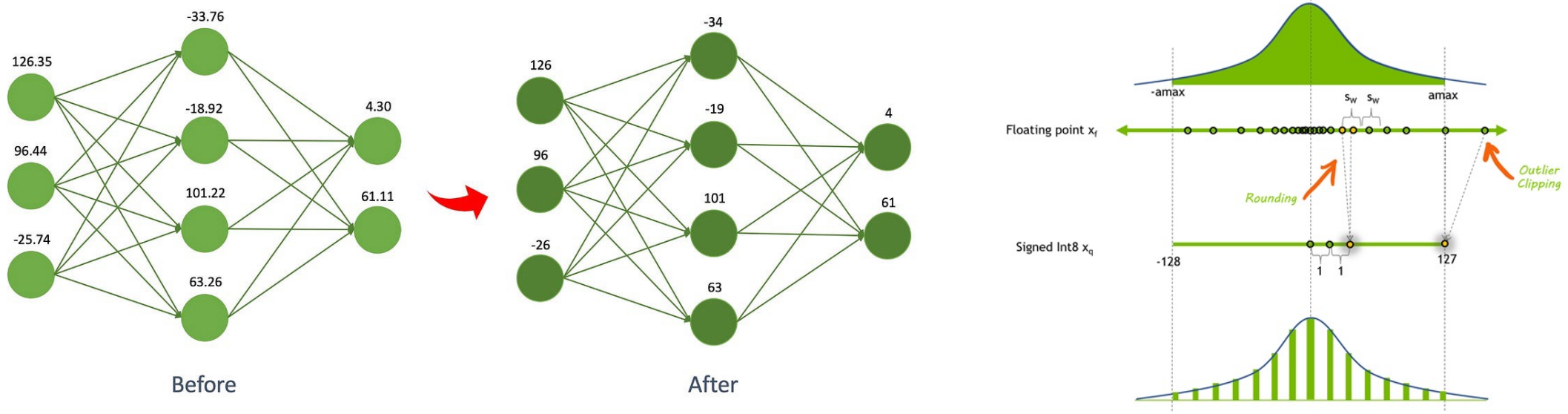


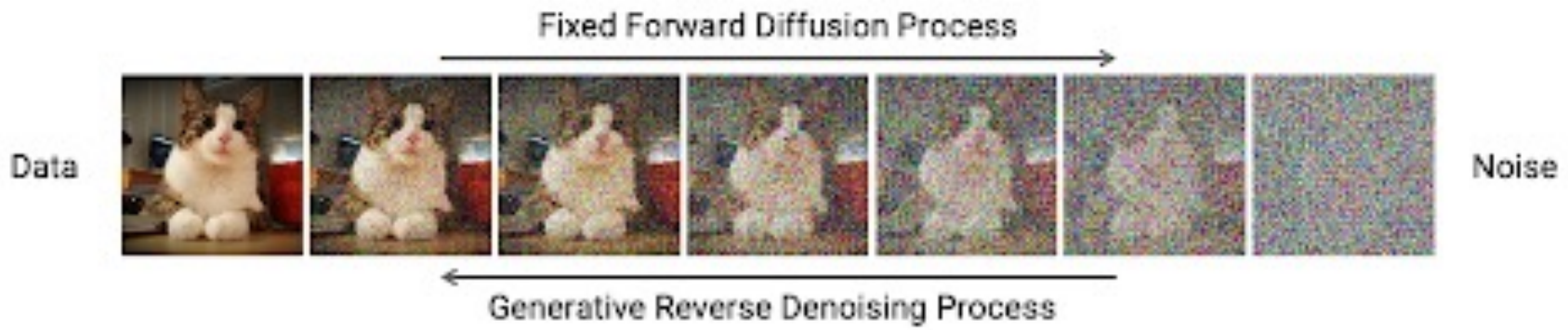
Figure credit: https://developer.nvidia.com/blog/achieving-fp32-accuracy-for-int8-inference-using-quantization-aware-training-with-tensorrt/

Quantization aware training: B. Jacob et al. CVPR 2704-2713 (2017)

4-bit transformers: T. Dettmers & L. Zettlemoyer. “The case for 4-bit precision: k-bit inference scaling laws” *ICML* (2022).

Neural networks are resilient to noise

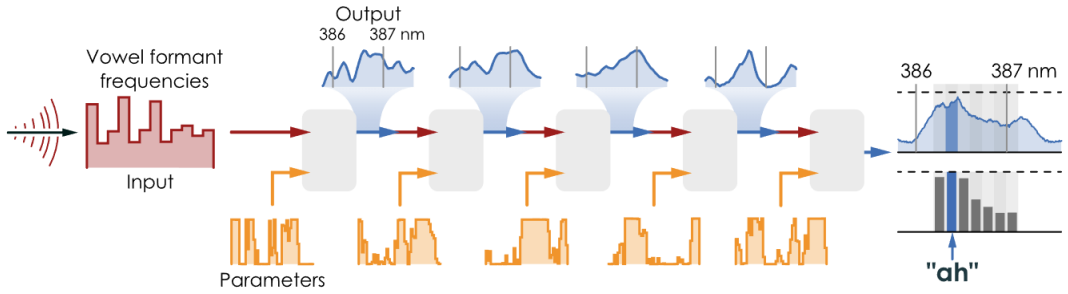
A deep neural network can even be designed to contain or remove noise.



N. Semenova, L. Larger, & D. Brunner. *Neural Networks* **146**, 151-160 (2022).

Figure credit: Z. Xiao, K. Kreis, & A. Vahdat. *arXiv:2112.07804* (2021)

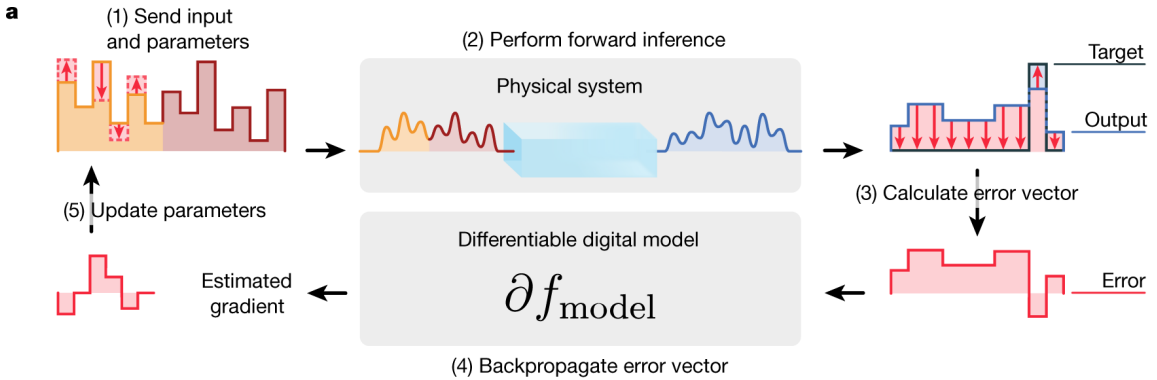
(Deep) physical neural networks



First demonstrations of PNNs: DNN-like calculations with networks of trained physical data transformations.

Potential for:
→ *Many* orders-of-magnitude better speed/efficiency
→ Learning approach to physical functionalities

Physics-aware training



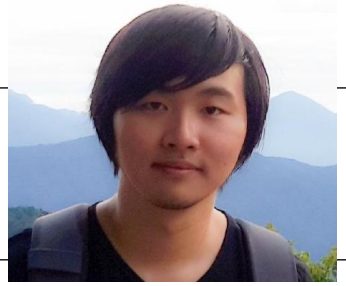
First demonstrations of backprop to train arbitrary physical systems *in situ*

→ Scales to high-dimensional parameter spaces
→ Trained PNN models inherently mitigate device imperfectio ality gap,

L.G. Wright*, T. Onodera*, M.M. Stein, T. Wang, D.T. Schachter, Z. Hu, P.L
Deep physical neural networks trained with backpropagation, *Nature* 601

https://github.com/mcmahon-lab/Physics-Aware-Training

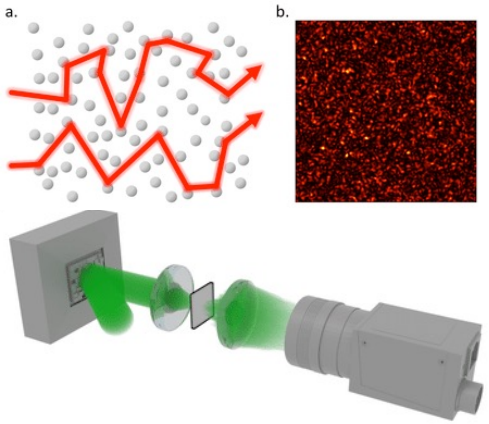
Prof. Logan Wright



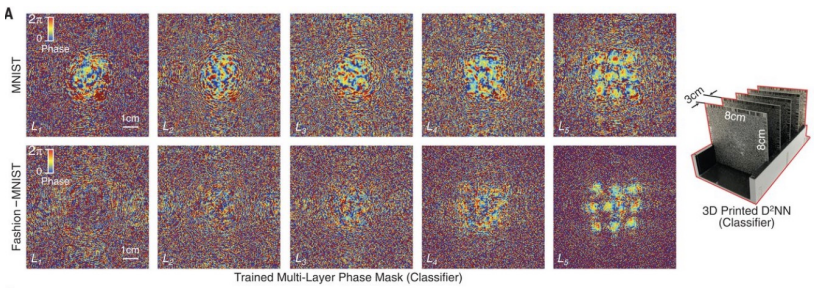
Dr. Tatsuhiro Onodera

Optical neural networks: an example of physical neural networks

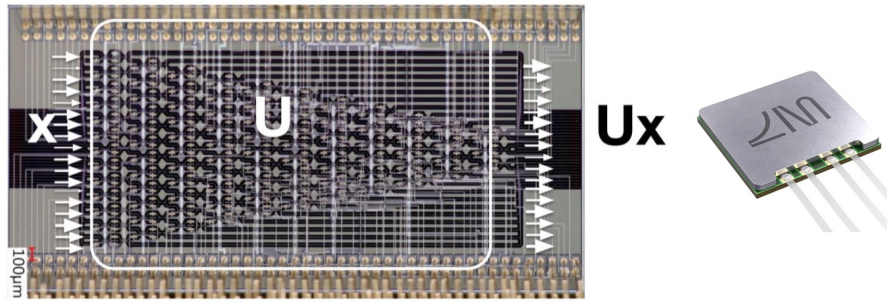
ONNs use light for analog computing with noise.



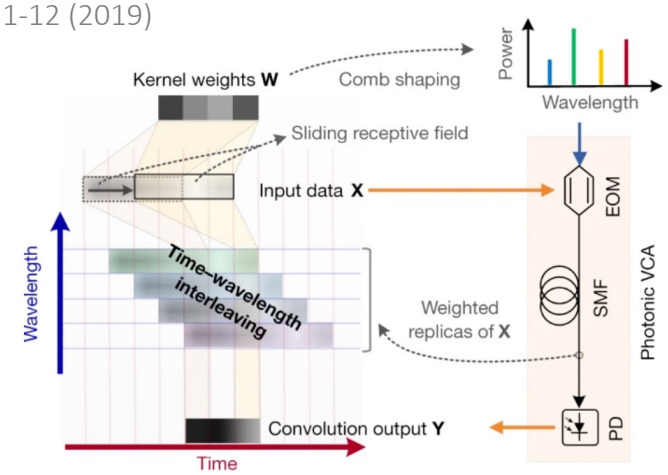
Random MVM by Multiple Light Scattering
J. Dong et al. *IEEE J. Sel. Top. Quantum Electronics*, ***26***, 1-12 (2019)



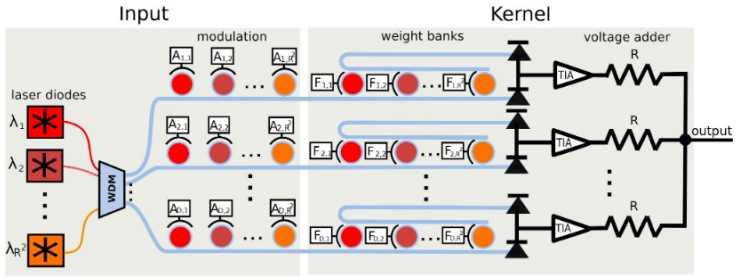
Diffractive Neural Networks
X. Lin et al. *Science*. **361**, 6406 (2018)
T. Zhou et al. *Nat. Photonics*, **15**, 367-373 (2021)



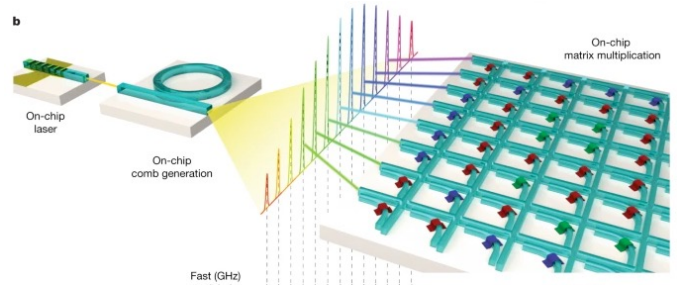
Arbitrary Unitaries with Mach-Zehnder Interferometers
Y. Shen, N. C. Harris, et al. *Nat. Photonics*. **11**, 441-446 (2017)



Time-Frequency Multiplexing Convolution
X. Xu et al. *Nature* **589**, 44-51 (2021)



Convolution with Microring Resonators
V. Bangari et al. *IEEE J. Sel. Top. Quantum Electron*. **26**, 1 (2020)

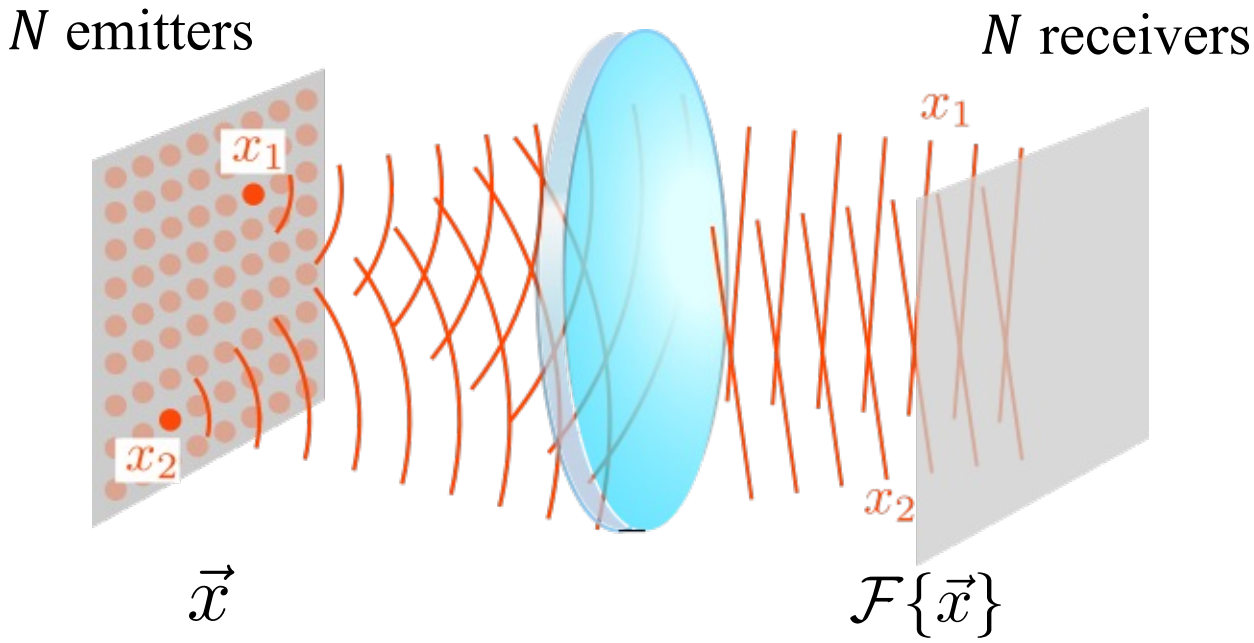


Convolution with Phase Changing Material
J. Feldmann, et al. *Nature* **589**, 52-58 (2021)

More ONN designs covered in the following reviews:
B. J. Shastri et al. *Nat. Photonics,* **15**, 102-114 (2021).
G. Wetzstein et al. *Nature*, **588**, 39-47 (2020).

Why optics for computing?

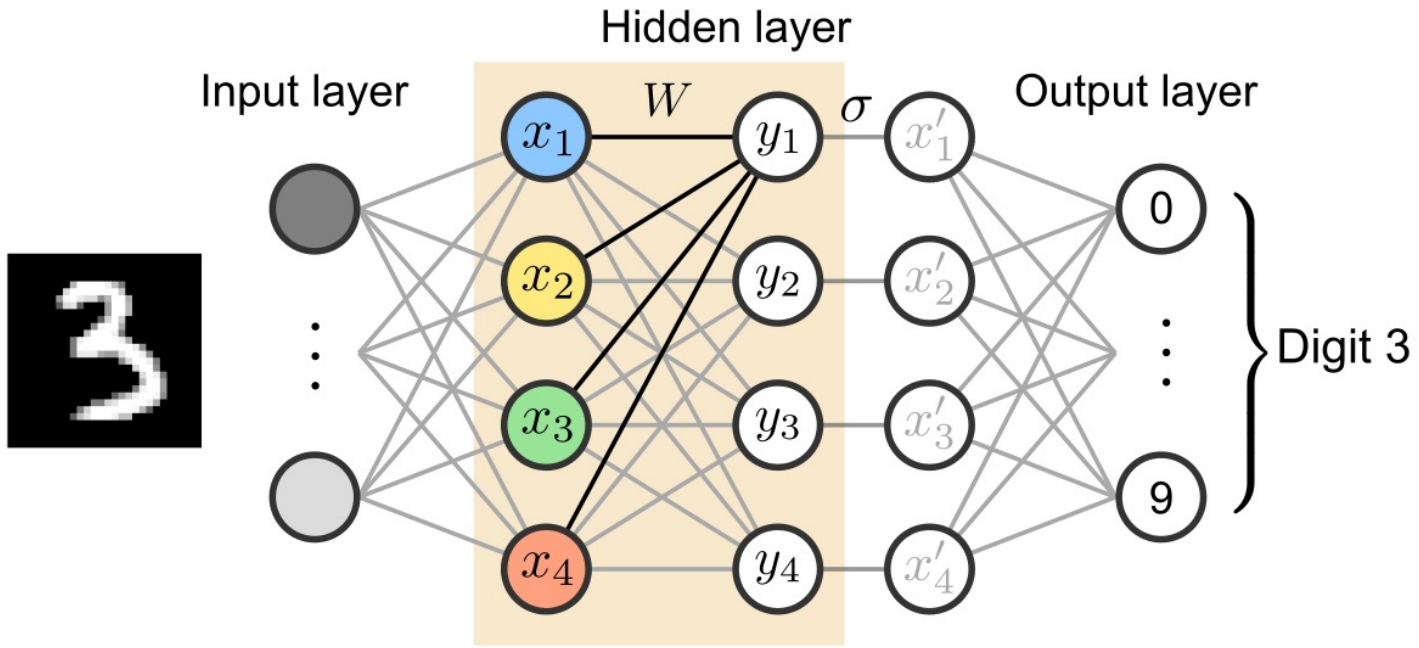
Optics is extremely efficient at large-scale linear transforms:



- Emitter / receiver energy consumption $\propto N$
- The number of scalar multiplications $\propto N^2$
- Energy per multiplication $\propto N/N^2 = 1/N$

Optical neural networks (ONNs)

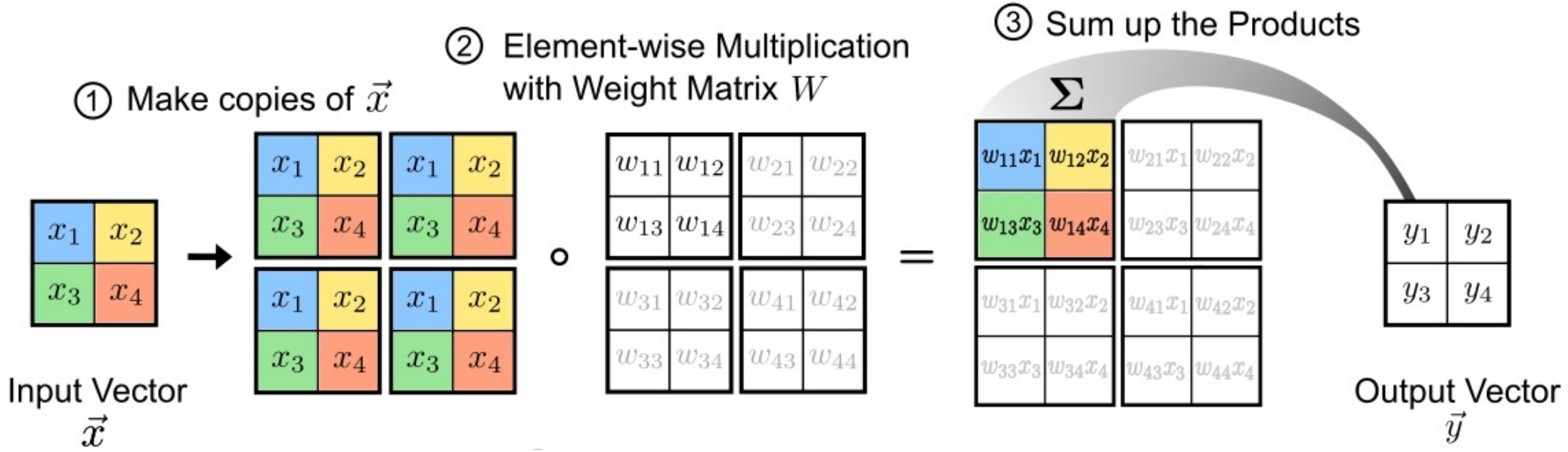
ONNs concern efficient implementation of arbitrary linear transforms:



$$x'_i = \sigma\left(\sum_j^N W_{ij} x_j + b_i\right)$$

$$O(N) \quad O(N^2) \quad O(N)$$

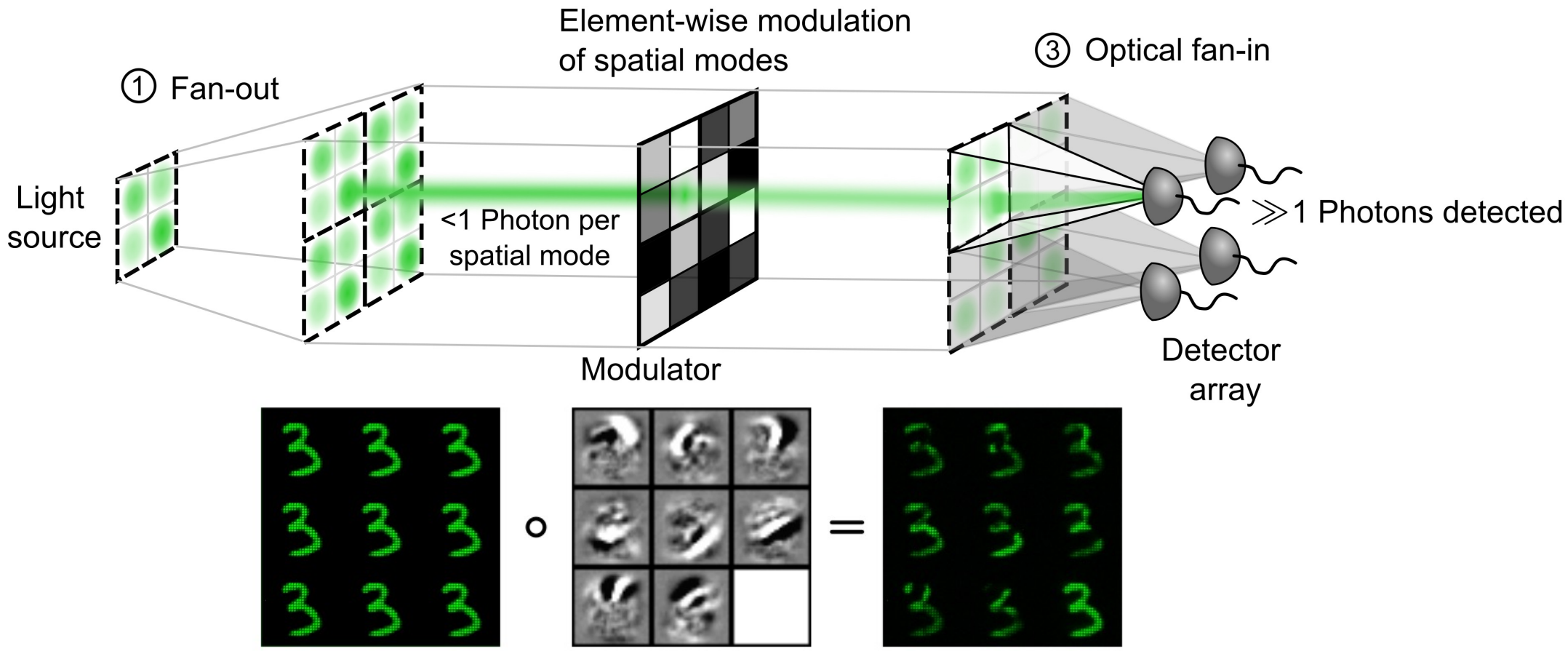
Computing optical matrix-vector multiplication for 2D inputs



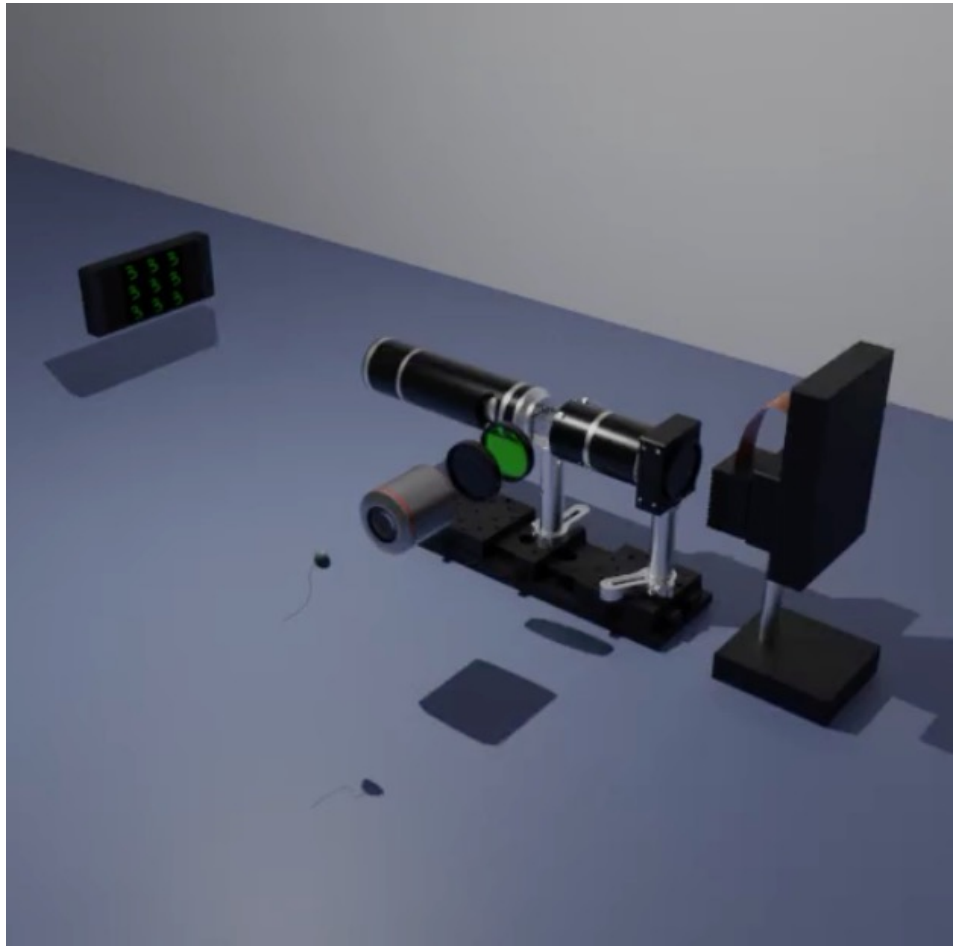
$$y_i = \sum_{j=1}^{4} w_{ij} x_j$$

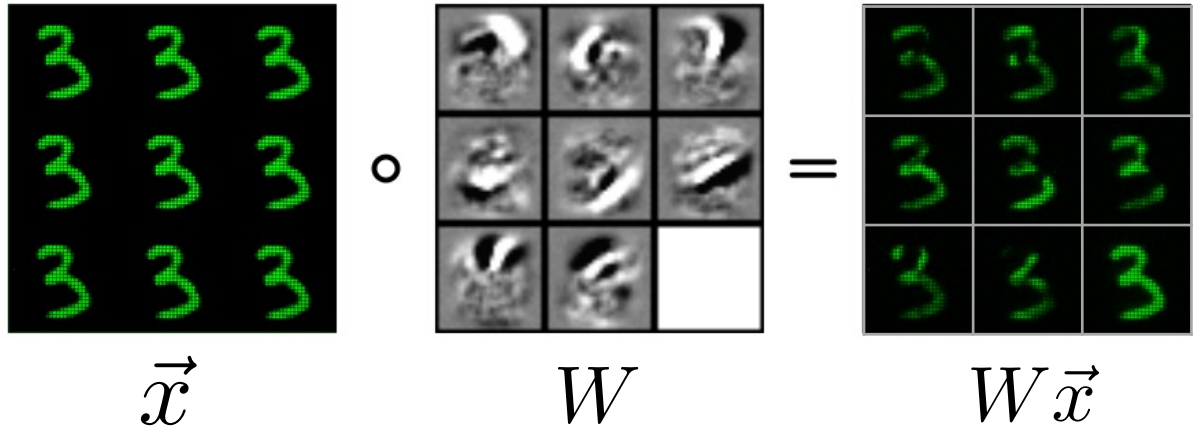
$$(i = 1, 2, 3, 4)$$

Illustration of the experimental scheme



Prototype experimental setup



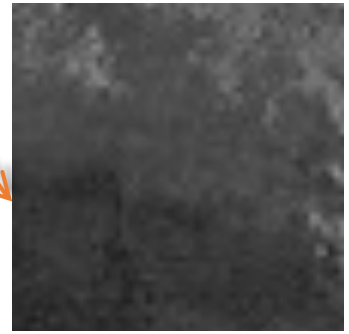


Vector-vector dot products with a vector size up to **~0.5 million** in dimension

Rendering credit: Hannah Doyle

Shot noise limits ONN energy consumption

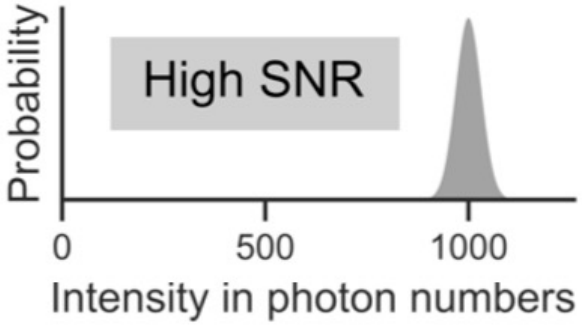
- ONNs are analog physical systems subject to photon shot noise:

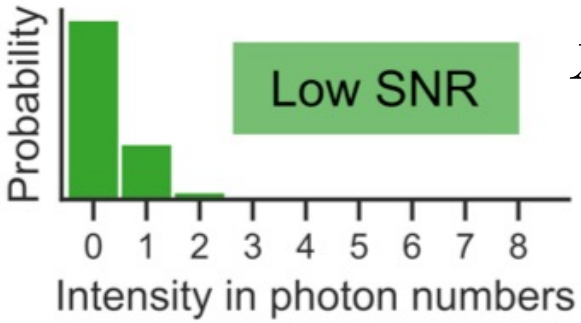


$X \sim \text{Poisson}(\mu = 1000)$

$\text{SNR} = \sqrt{\mu} = 31.6$

SNR: signal-to-noise ratio

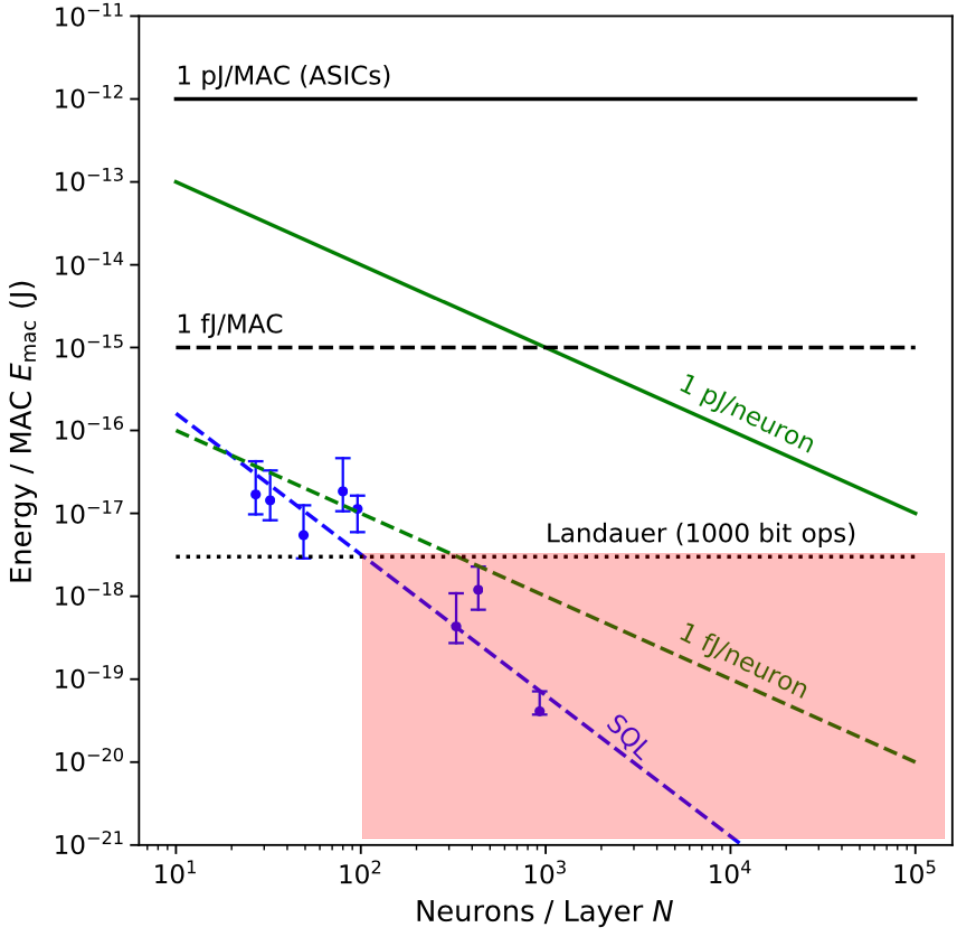




$X \sim \text{Poisson}(\mu = 0.2)$

$\text{SNR} = \sqrt{\mu} = 0.45$

Theoretical prediction: ONNs beyond the Landauer's limit



- Landauer's limit: the minimum energy required by irreversible computing in electronics
- Large-scale ONNs can do better than Landauer's limit!

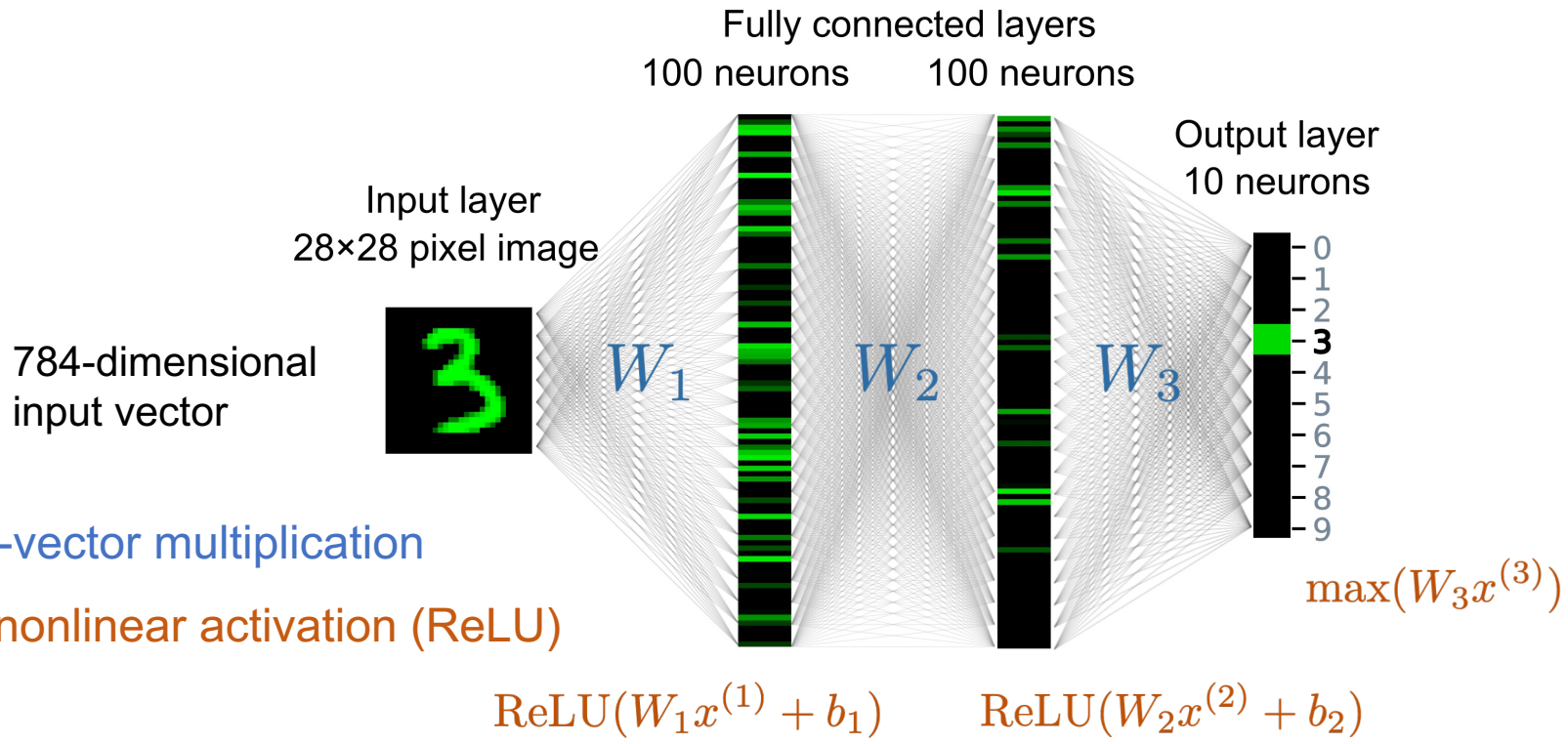
R. Hamerly, et al. *Phy. Rev.* X. **9,** 021032 (2019). (Figure credit)

M. A. Nahmias, et al. *IEEE J. Sel. Top. Quantum Electron*. **26, 1** (2019)

MAC: multiply-accumulate operation = 1 scalar multiplication and 1 addition

An ONN for image classification

We used quantization-aware training (QAT) to make the ONN noise-resilient:



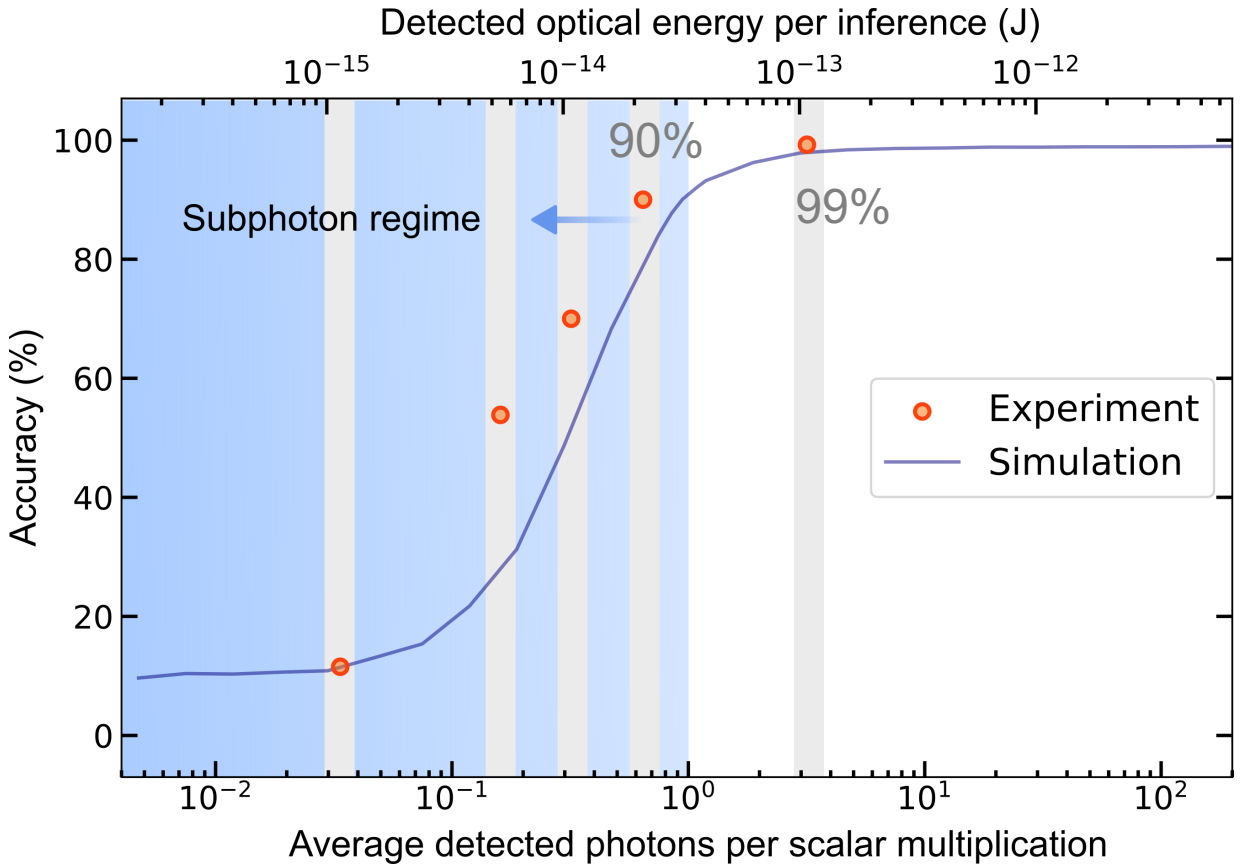
Optical matrix-vector multiplication

Digital bias + nonlinear activation (ReLU)

QAT: B. Jacob *et al.* CVPR 2704-2713 (2017)

Classification accuracy vs photon budget

High classification accuracy was obtained with even <1 photon per scalar multiplication.



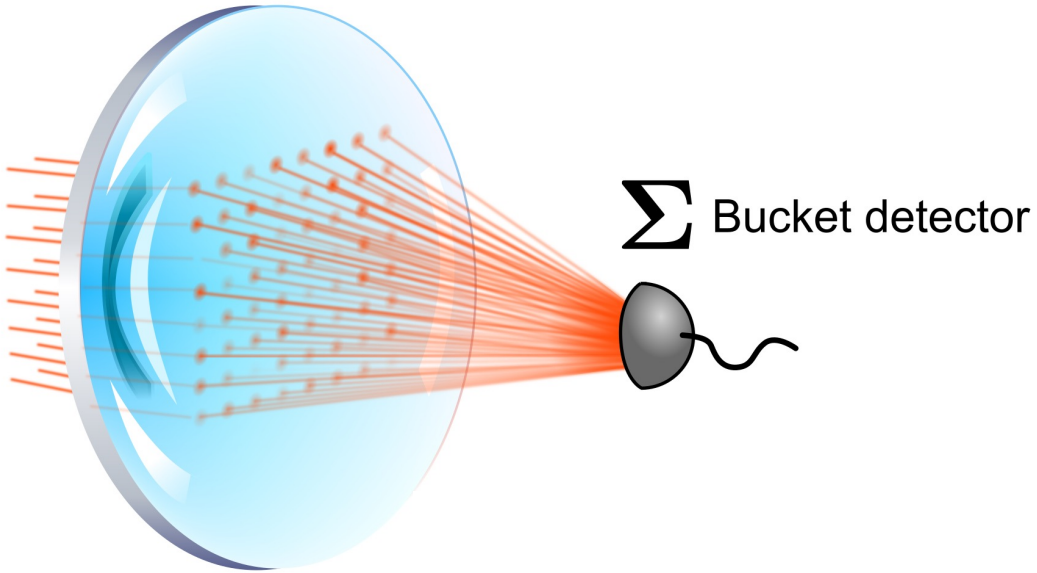
~100 fJ optical energy per image classification for ~90,000 multiply-and accumulate operations in total.

Multiplying two 8-bit integers costs ~70 fJ energy for 7-nm node electronics.

N. P. Jouppi, et al. In *ISCA* (2021)

T. Wang, S-Y, Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon.
An optical neural network using less than 1 photon per multiplication. *Nature Communications* **13**: 123 (2022)

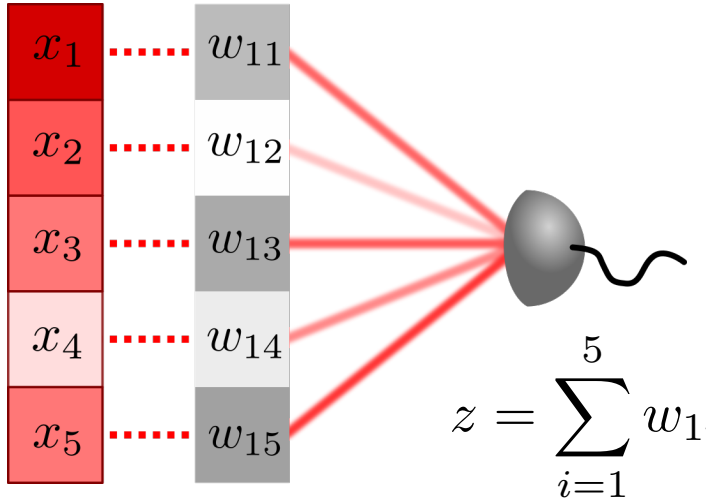
Interpretation of <1 photon per scalar multiplication



- The number of photons measured for each beam = integer
- Averaged over all the beams $\epsilon = \dfrac{\text{the total number of detected photons}}{\text{the total number of spatial modes } N}$

From analog to stochastic optical neural networks

Analog ONN



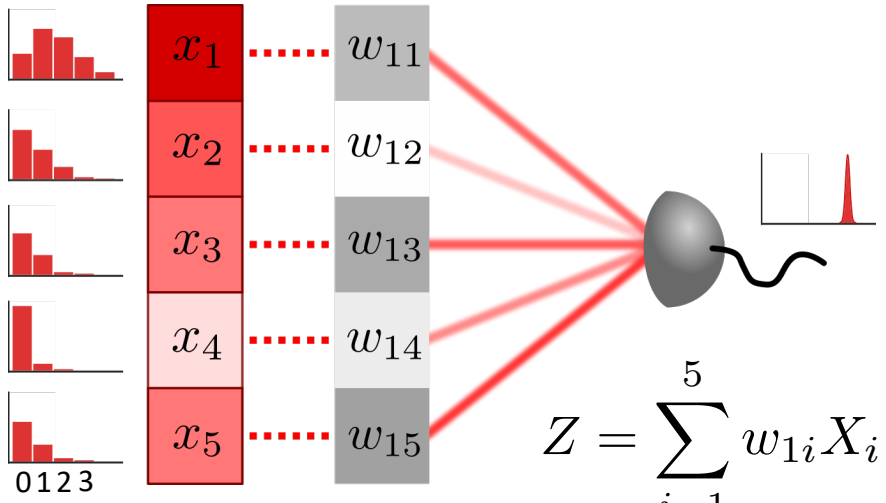
$$z = \sum_{i=1}^{5} w_{1i} x_i$$

$$X_1 \sim \text{Poisson}(\mu = x_1)$$

$$X_2 \sim \text{Poisson}(\mu = x_2)$$

$$\vdots$$

Stochastic ONN

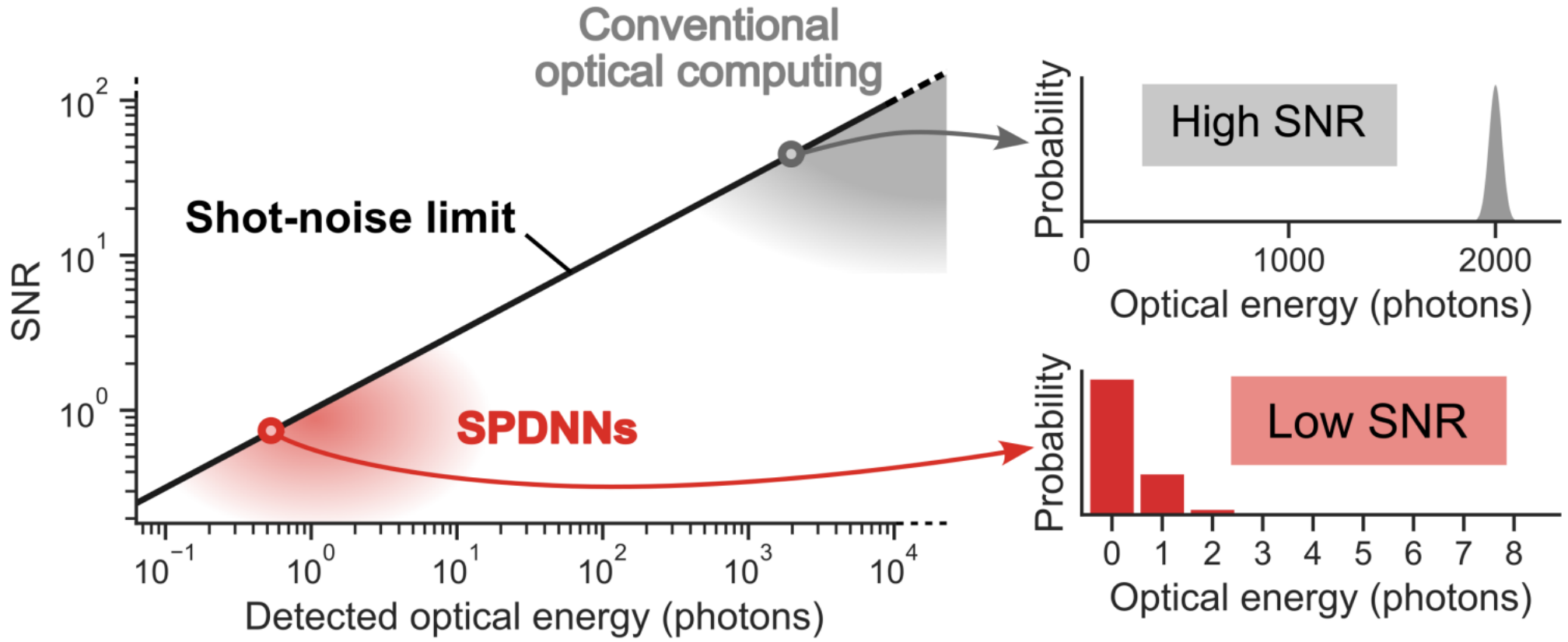


$$Z = \sum_{i=1}^{5} w_{1i} X_i$$

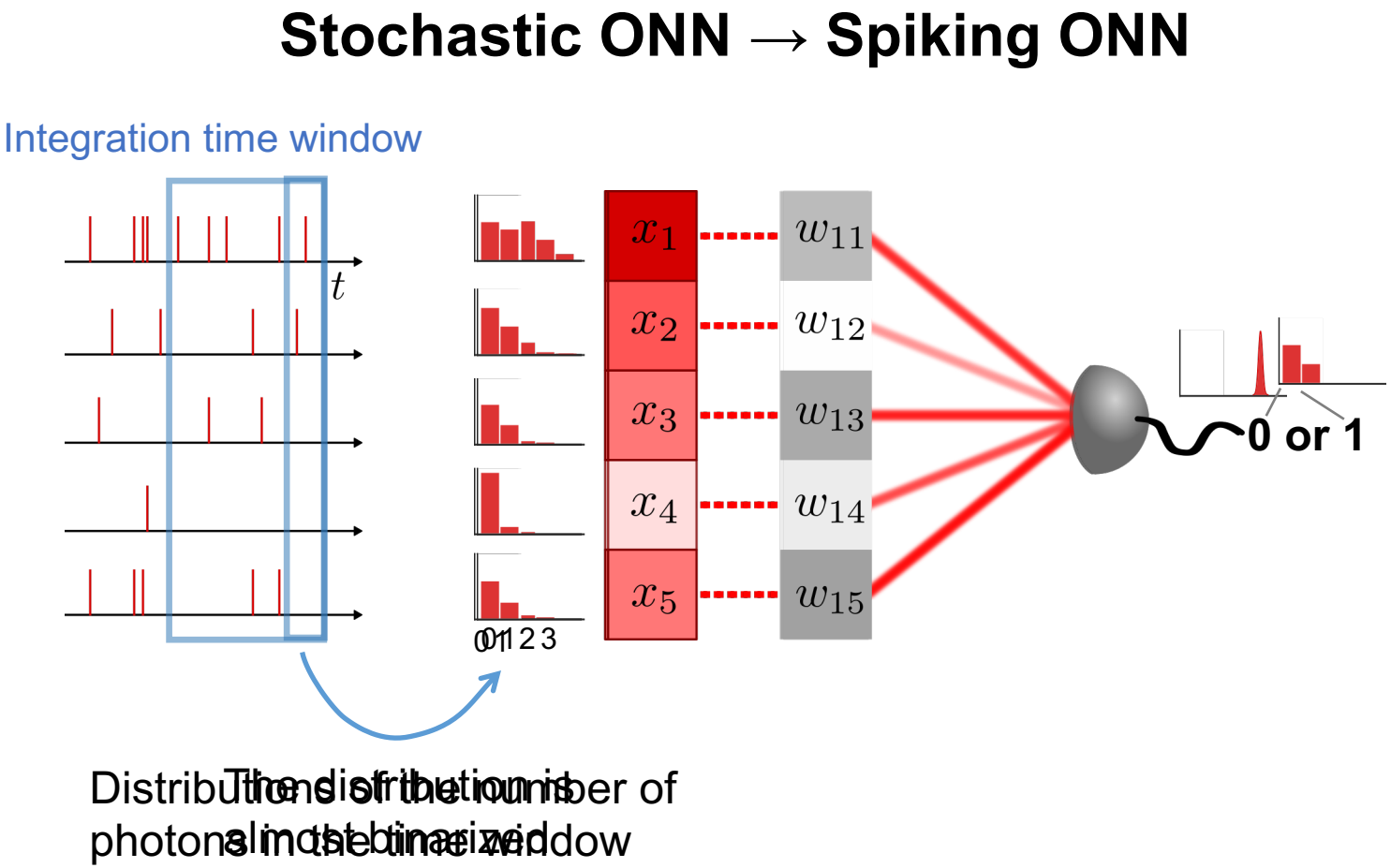
$$\sim \text{Poisson}(\mu = \sum_{i=1}^{5} w_{1i} x_i)$$

When optical intensity is reduced,
it corresponds to replacing continuous variable x_i with random variable $X_i \sim \text{Poisson}(\mu = x_i)$

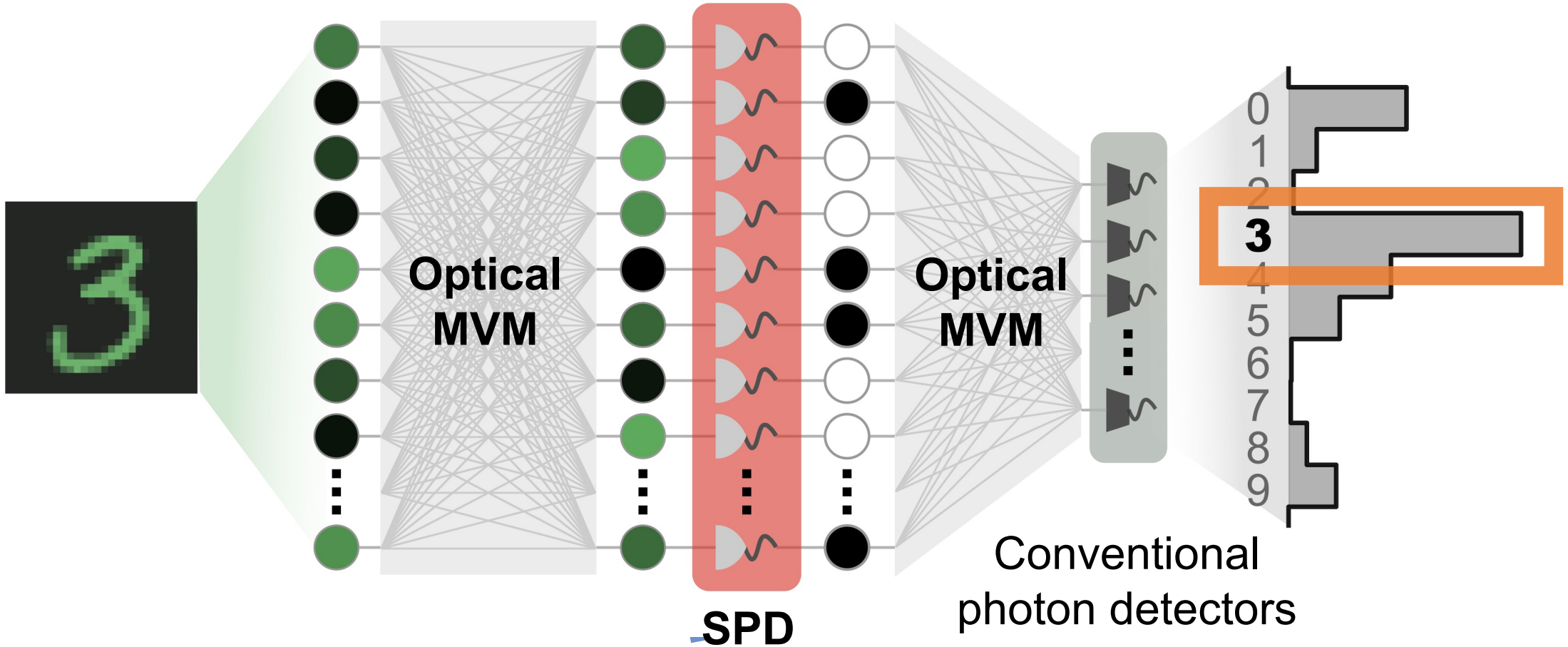
How low can it go?



Stochastic spiking ONN: a new regime with single-photon detection



Stochastic SPDNNs for deterministic classification tasks

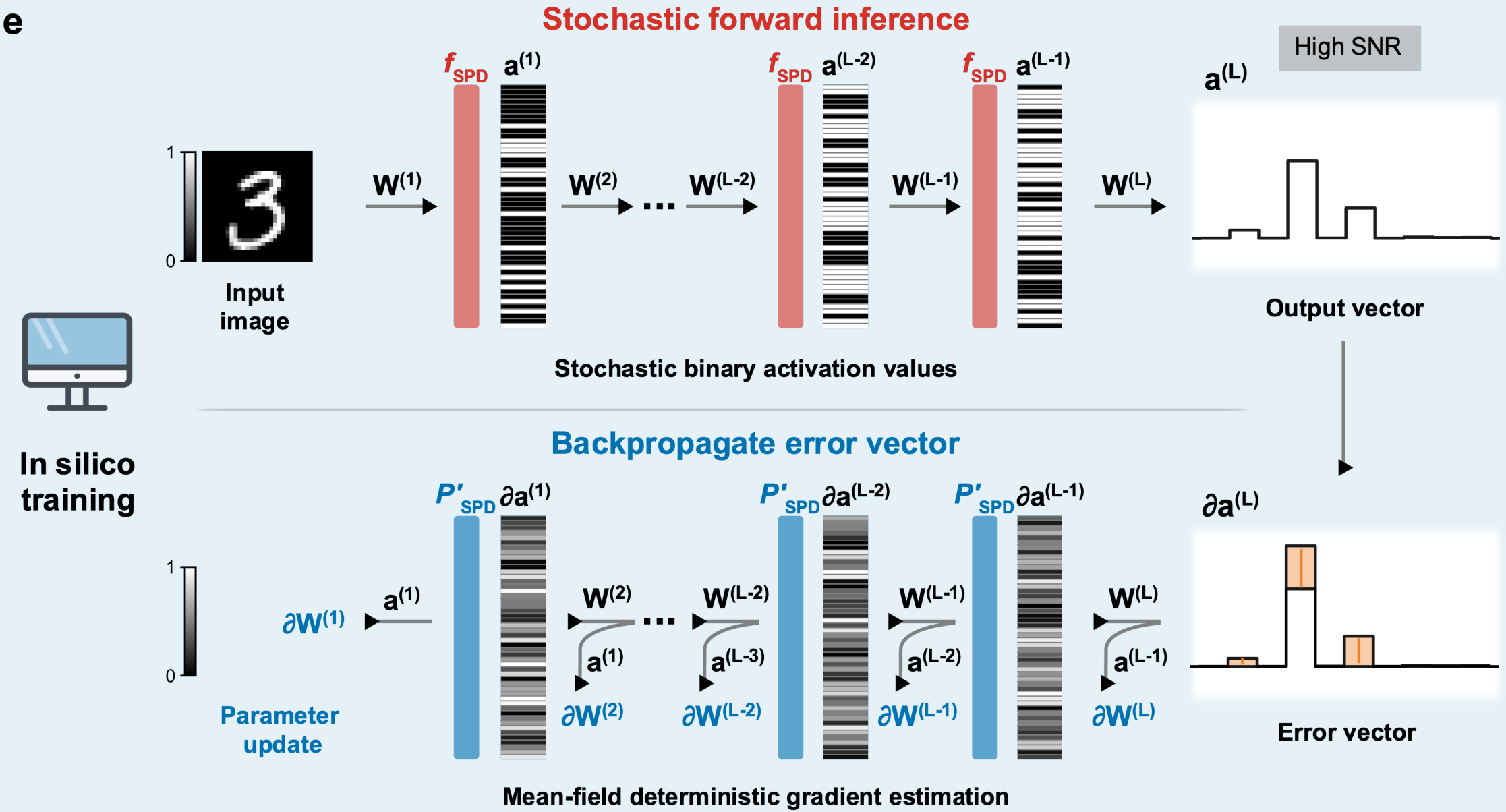


SPDNN: single-photon detection neural network

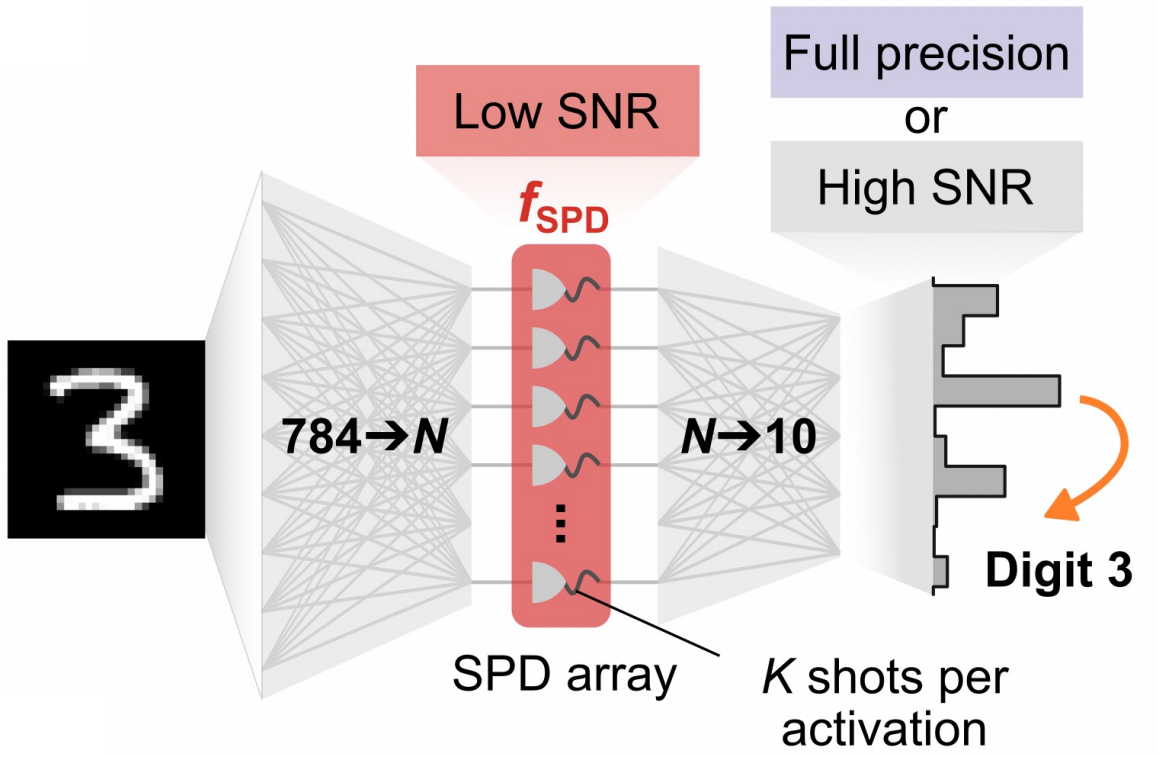
Need to control the optical energy here to be **~1 photon (SNR ~ 1)**!!

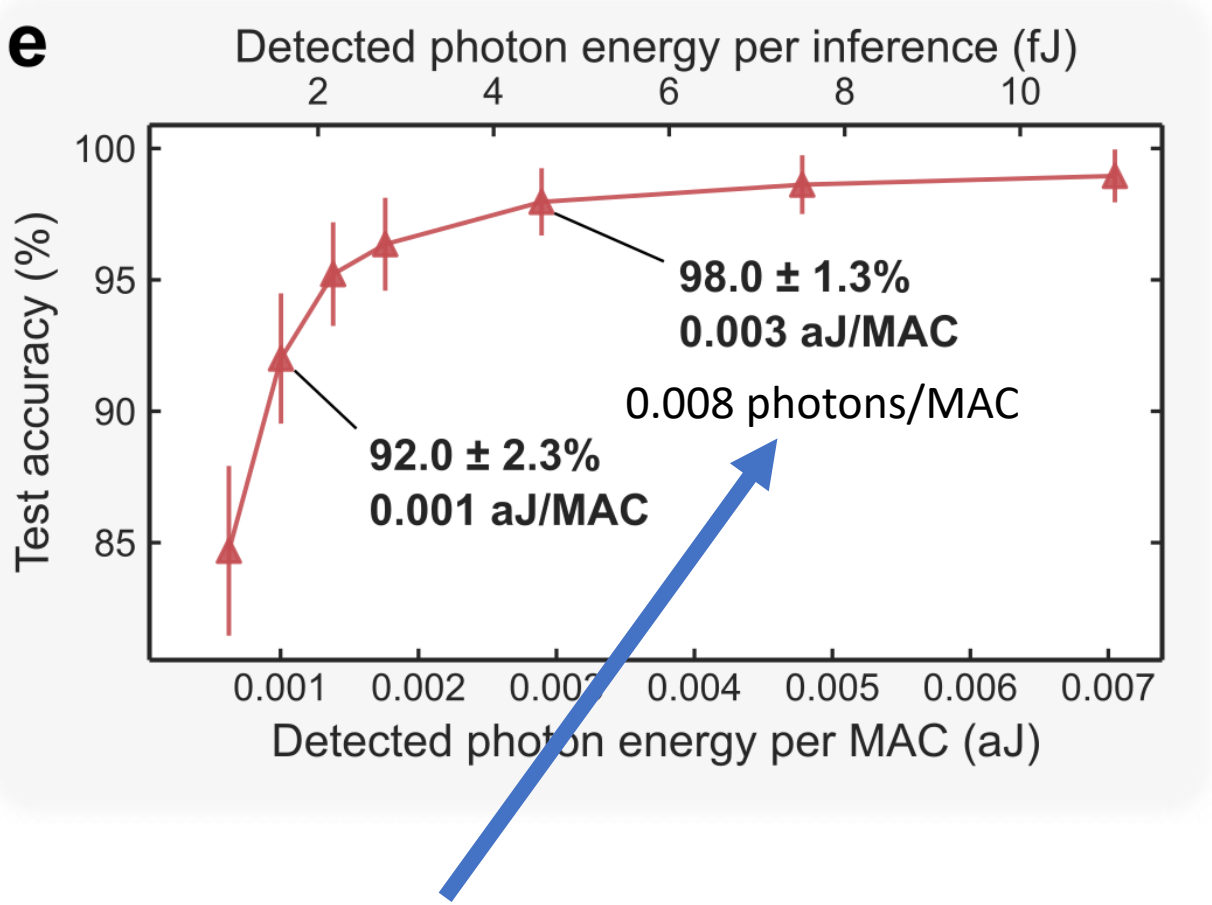
S-Y. Ma, T. Wang, J. Laydevant, L. G. Wright, and P. L. McMahon. *arXiv:2307.15712* (2023).

The key ingredient: Physics-aware, stochastic-ONN training



MNIST with < 10 fJ per inference



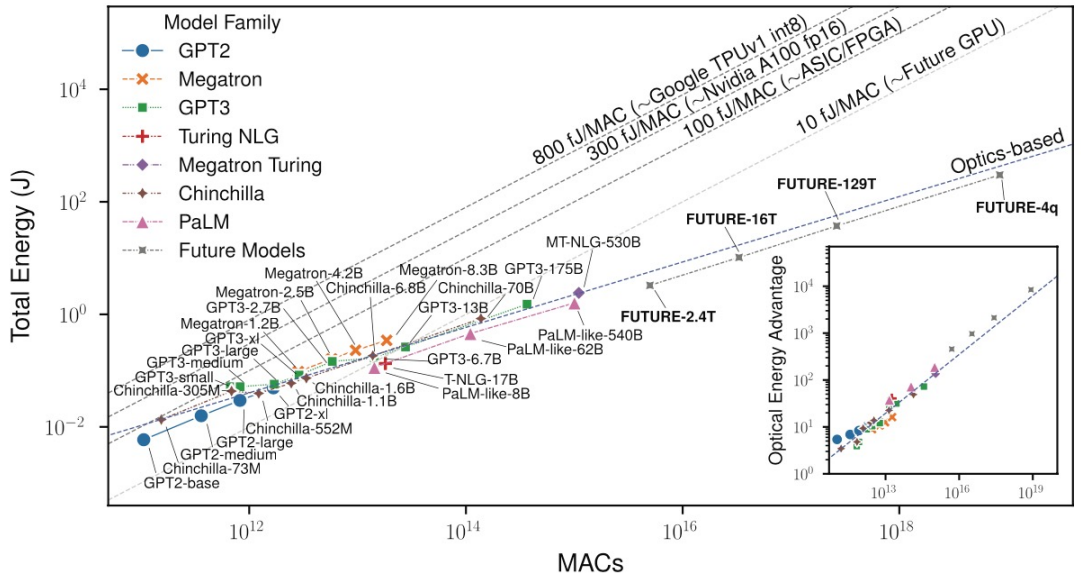


0.0008 photons/MAC
Hidden layer
(3 OoM better than before!)

S.-Y. Ma, T. Wang, J. Laydevant, L. G. Wright, & P. L. McMahon. "Quantum-noise-limited optical neural networks operating at a few quanta per activation". *arXiv:2307.15712* (2023)

ONNs for closing “the computing gap”

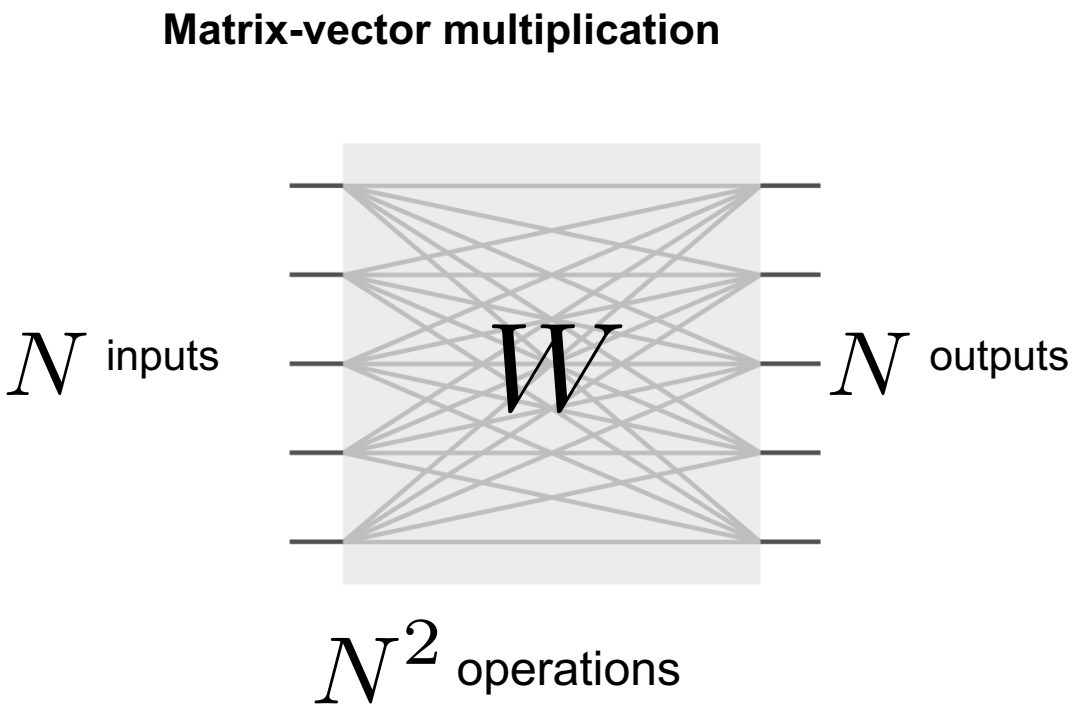
Machine-learning computation demand ≫ computing hardware capacity

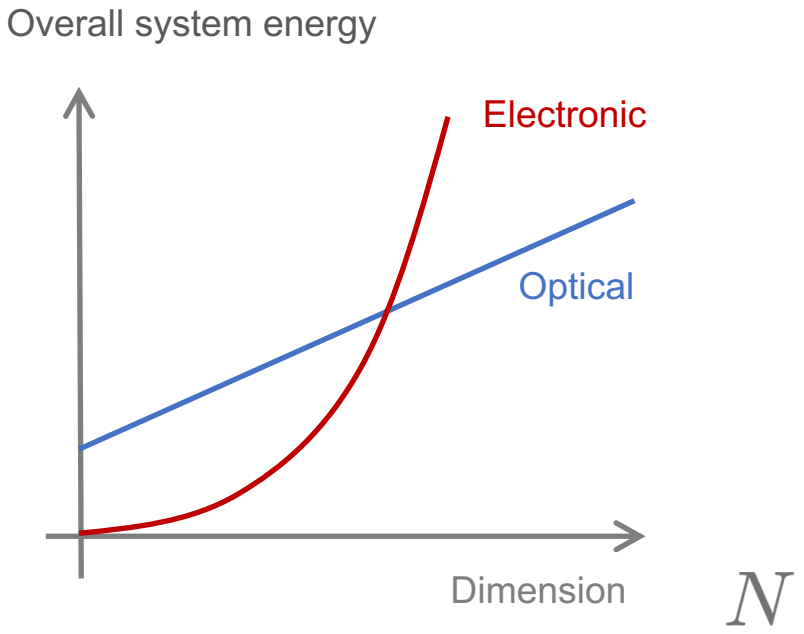


Optical processors:

10,000× higher energy efficiency than digital electronics for large machine-learning models (e.g., ChatGPT)

A favorable energy scaling law for ONNs





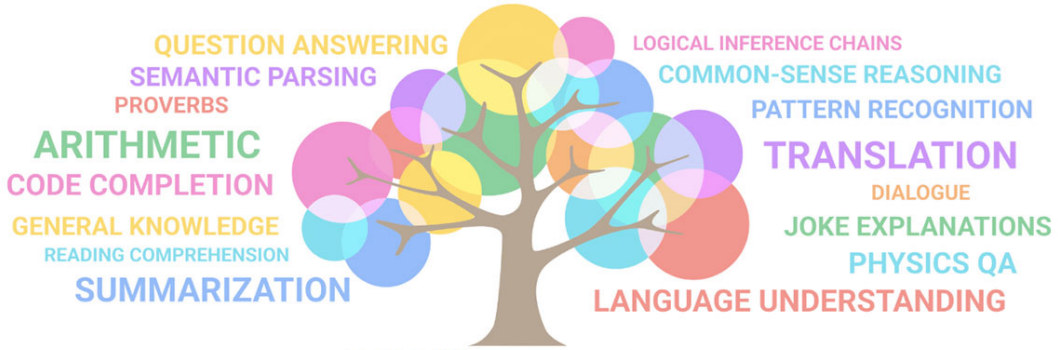
M. A. Nahmias, … P. R. Prucnal. *IEEE J. Sel. Top. Quantum Electron*. **26, 1** (2019).

Optics should become more and more advantageous as the size of the network grows

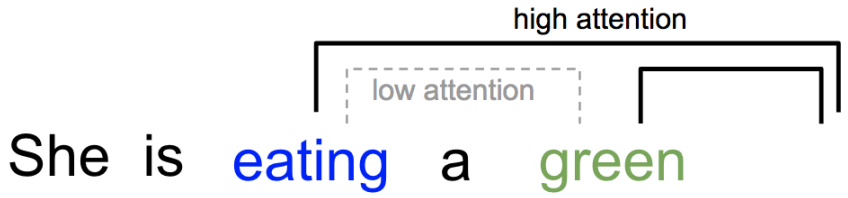
ONNs for large language models (LLMs)

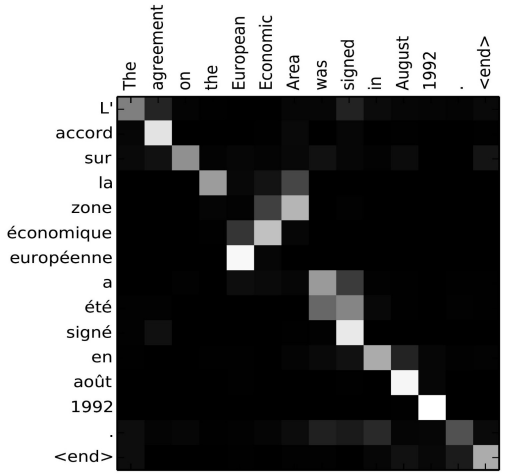
- LLMs (e.g., ChatGPT) model natural languages by capturing the statistical relation between words within context.

What can LLMs do?



Attention helps LLMs to keep track of context and relation between words



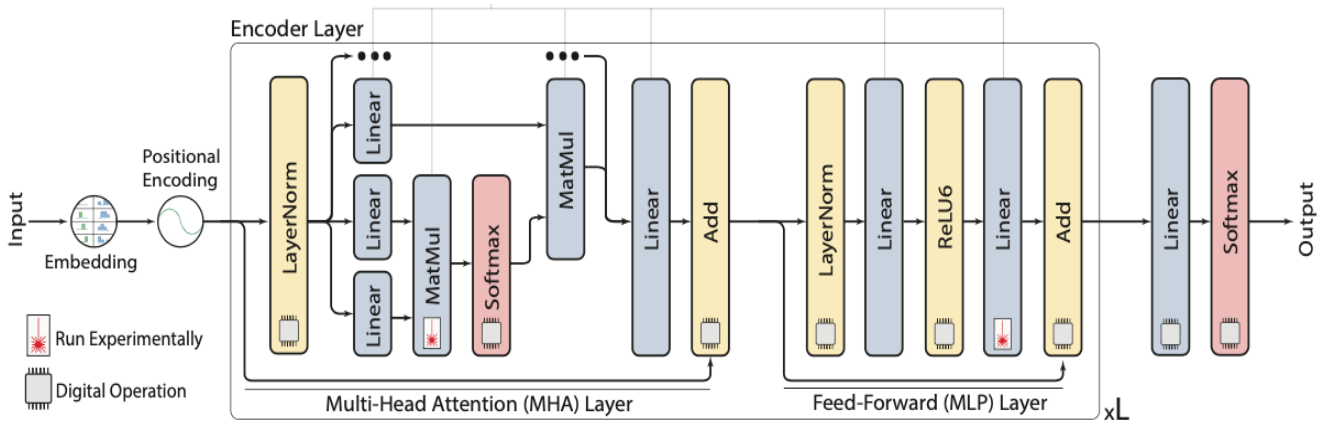


LLMs are a nice showcase of high-dimensional linear processing

Maxwell Anderson

- LLMs are primarily composed of large dot products (e.g., >1000D).

1 of 12 total attention heads and 1 of 12 total layers



A small model

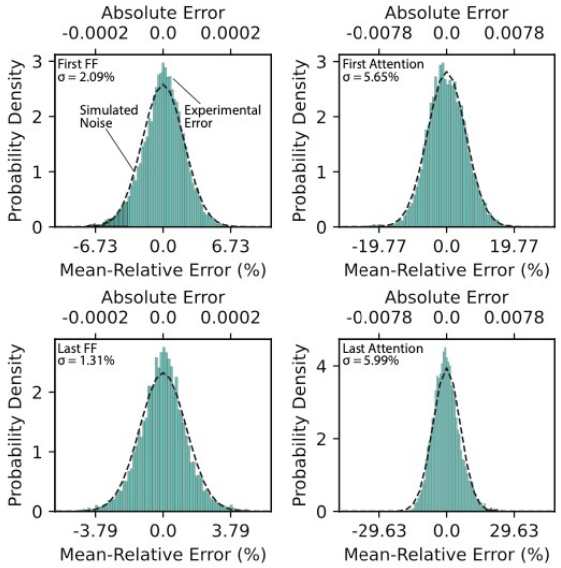
- Would ONNs work for LLMs?
- Would the optical energy scaling law still hold?

LLMs with optical precision

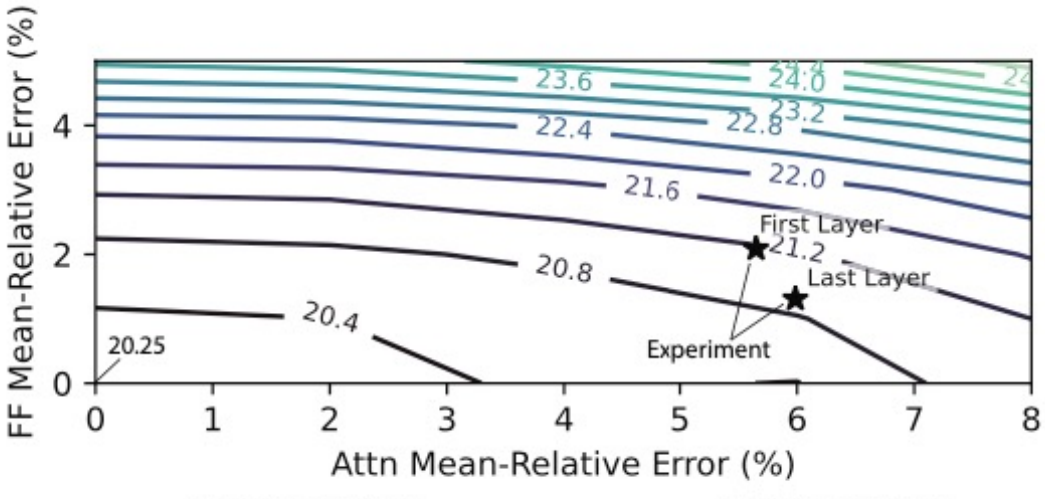
- Optical LLMs can work with moderately low precision (~5 bit input precision, or 100s photons/MAC).

Maxwell Anderson

Dot product error characteristics on our experimental setup



Dependence of the optical transformer's performance (perplexity) on the precision of each operation

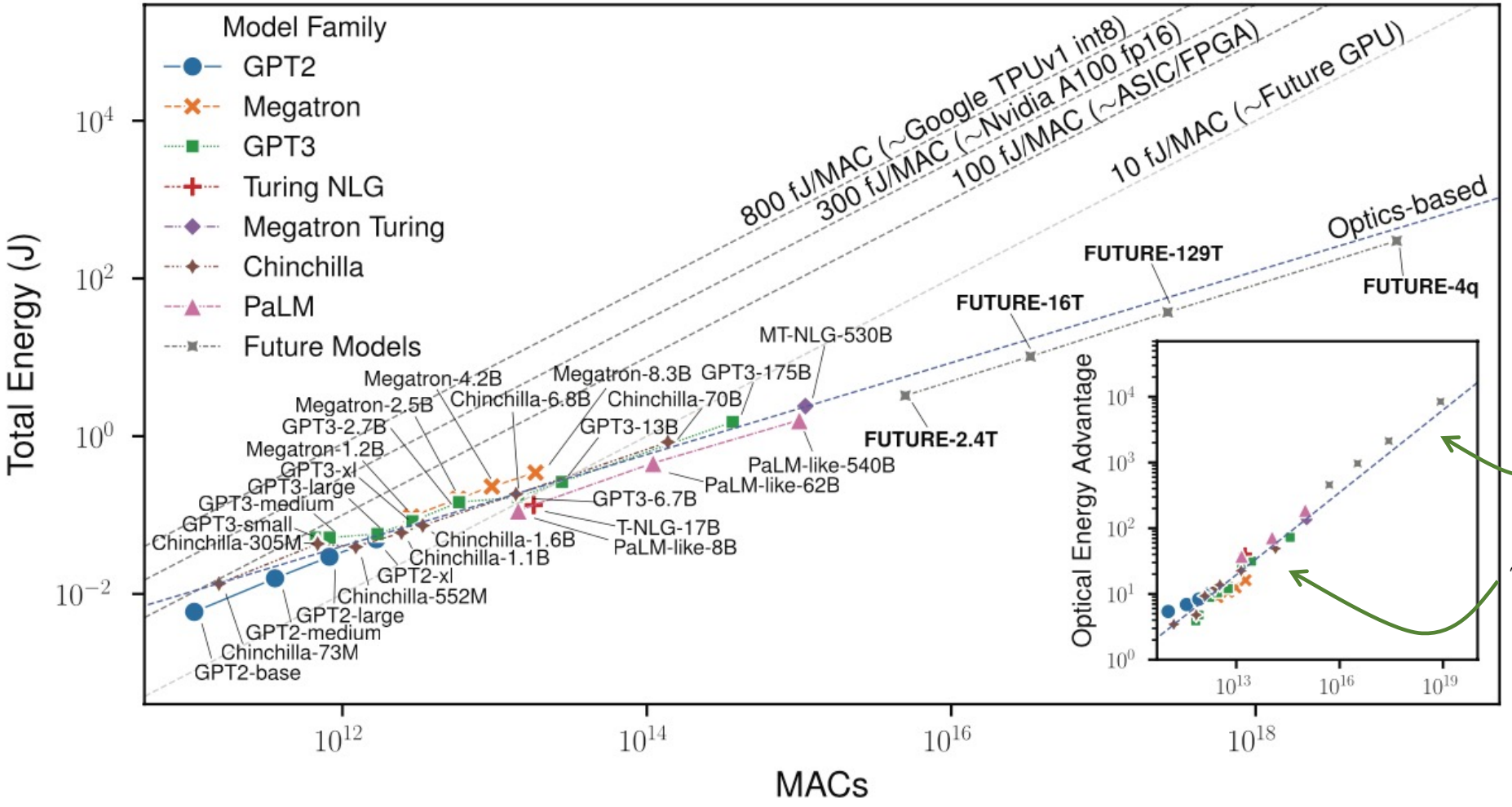


FF: Feed-forward layers, vector dimension = 4d = 3072
Attn: Attention layers, vector dimension = d = 768

M. Anderson, T. Wang, S.-Y. Ma, L.G. Wright and P.L. McMahon, *arXiv:2302.10360* (2023)

The whole-system energy scaling for LLMs

Energy advantages increase with model size, estimated based on realistic energy consumption data of each hardware component:



~8,500× for
a 23,000× larger model

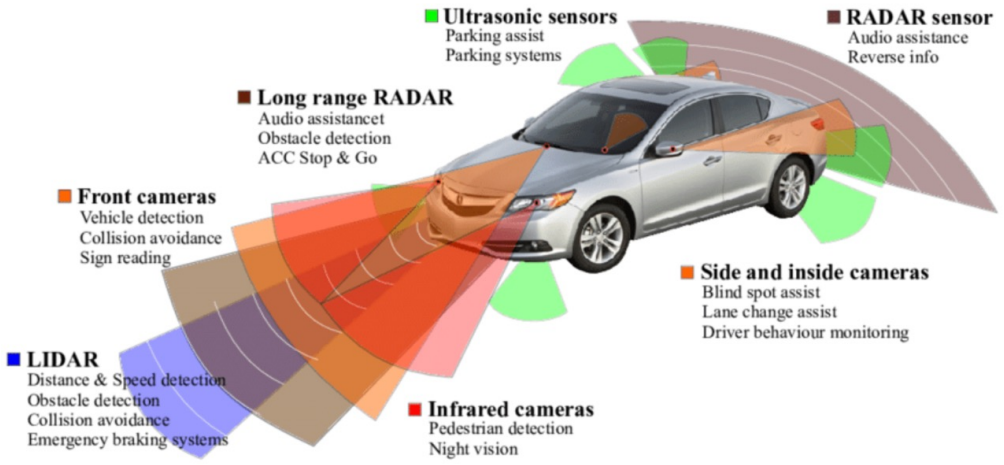
~100× for ChatGPT

M. Anderson, S.-Y. Ma, T. Wang, L.G. Wright and P.L. McMahon, *arXiv:2302.10360* (2023)

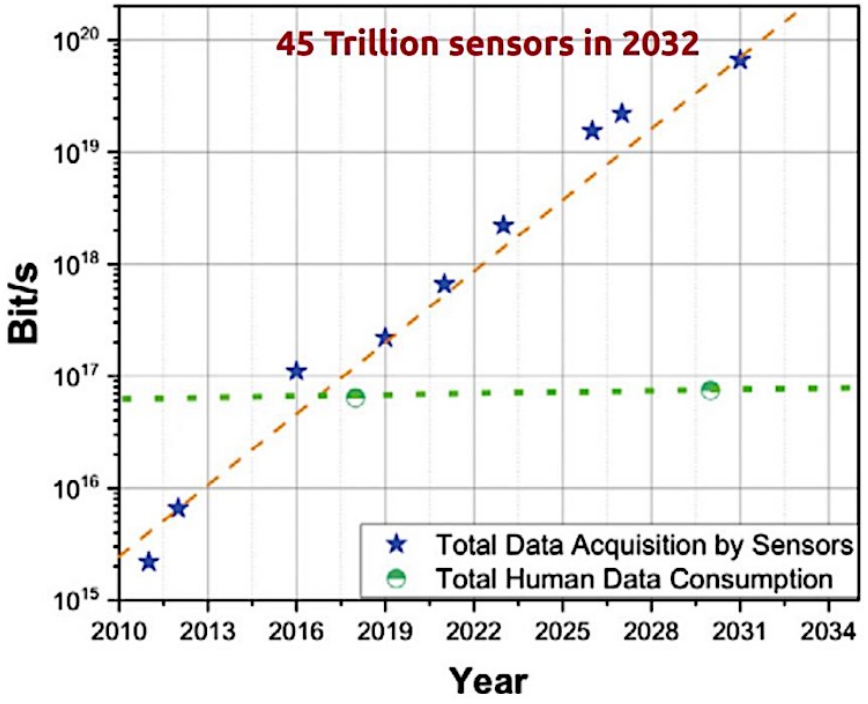
ONNs for closing “the sensor gap”

- More data is produced than that can be stored and processed.

A self-driving car with various sensors



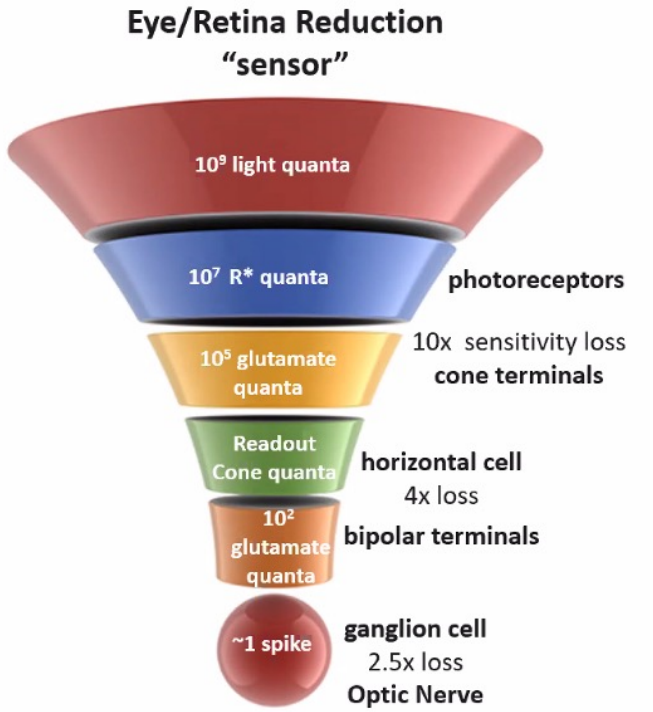
1: Typical types of sensors for ADAS.

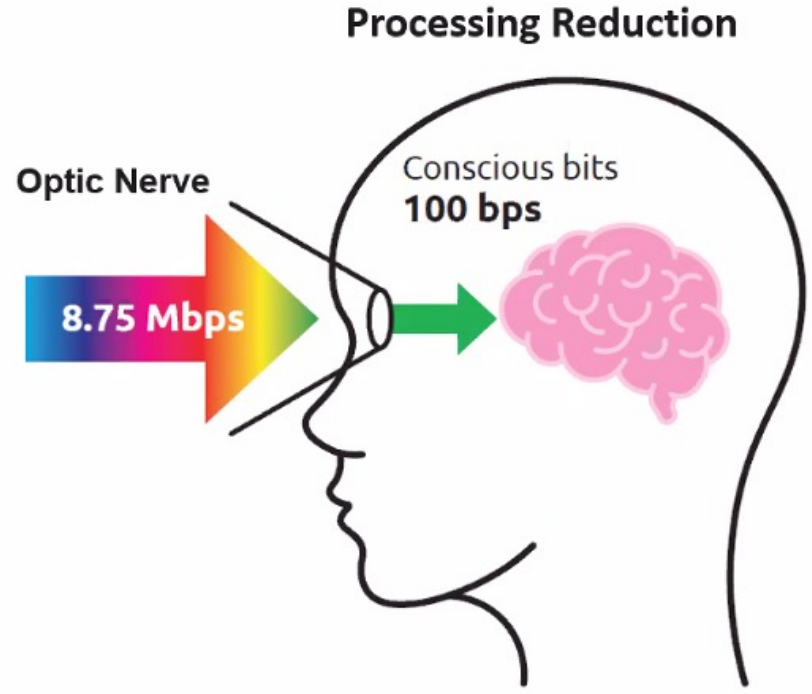


Decadal Plan for Semiconductors: New Trajectories for Analog Electronics, SIA (2021)

How does optical processing help to solve ”the sensor gap”?

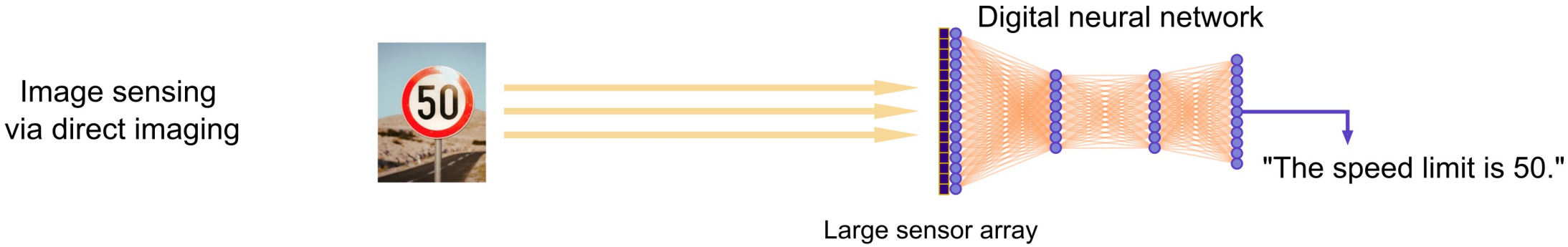
Smart sensors: a lesson from biological neural computation





Selective information reduction is an essential part of intelligence!

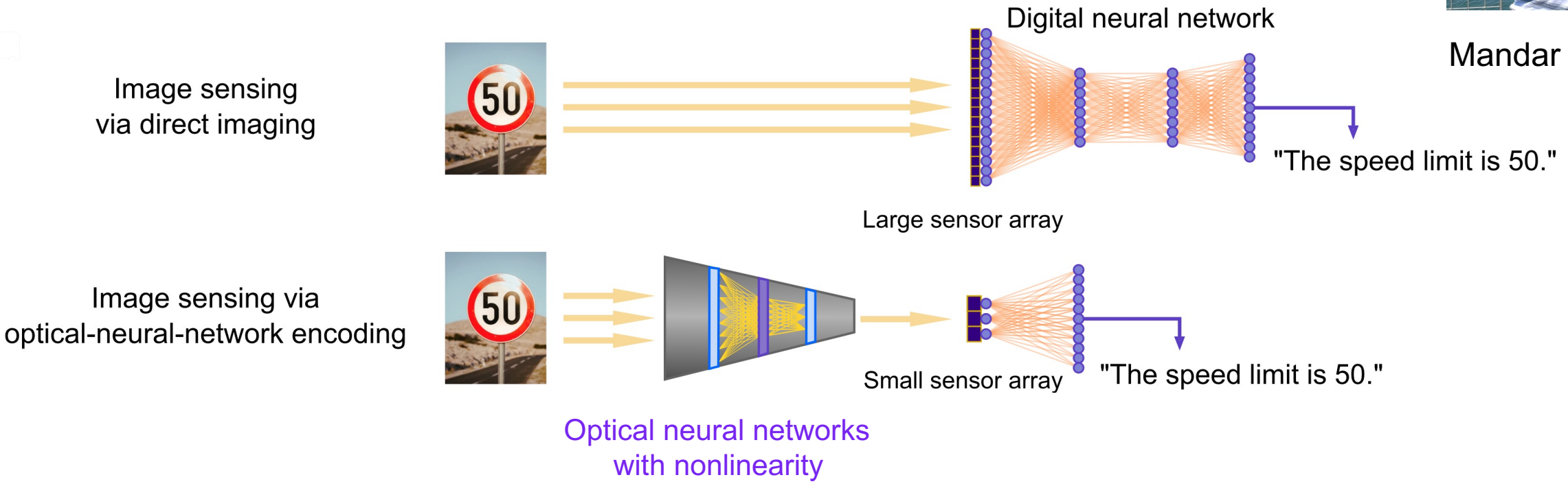
Conventional machine vision pipelines



- Relevant information is sparse in images.
- Most of data acquisition is unnecessary, and only serves to reduce response speed and energy efficiency.

Image sensing with multilayer ONNs

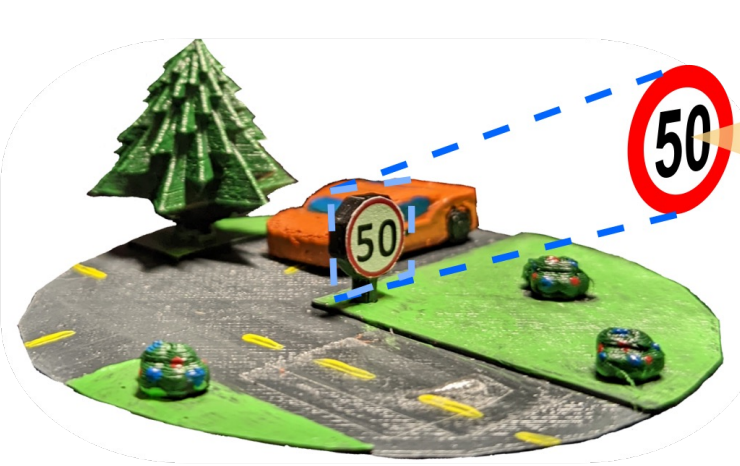
Mandar Sohoni



T. Wang*, M. M. Sohoni*, L. G. Wright, …, P. L. McMahon. "Image sensing with multilayer, nonlinear optical neural networks" *Nature Photonics* **17**, 408–415 (2023)

The diagram of our ONN image sensor

A two-layer ONN with at least 400:1 compression ratio:



Resolution equivalent to 40×40=1,600 input neurons

$\vec{x}$

6×6=36 neurons

$W_1\vec{x}$

6×6

$\sigma(W_1\vec{x})$

2×2=4 neurons

$W_2\sigma(W_1\vec{x})$

Measure and digitize

"The speed limit is 50."

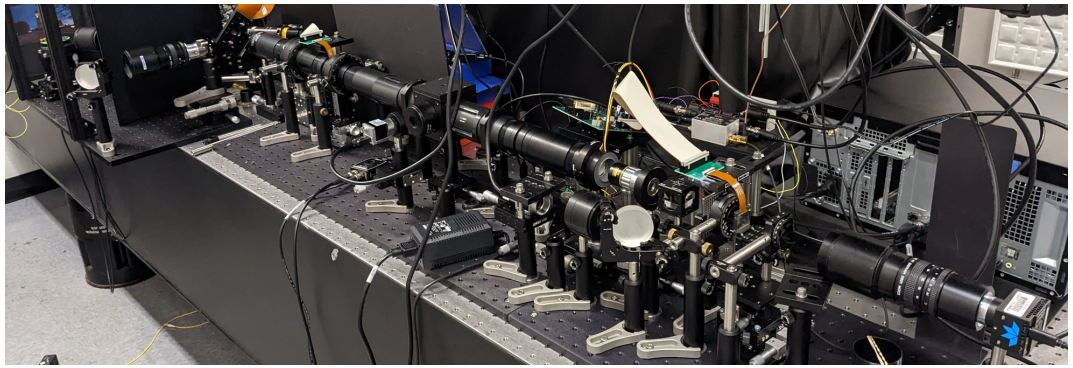
$f_{\text{digi}}(W_2\sigma(W_1\vec{x}))$

Optical fully connected layer W_1

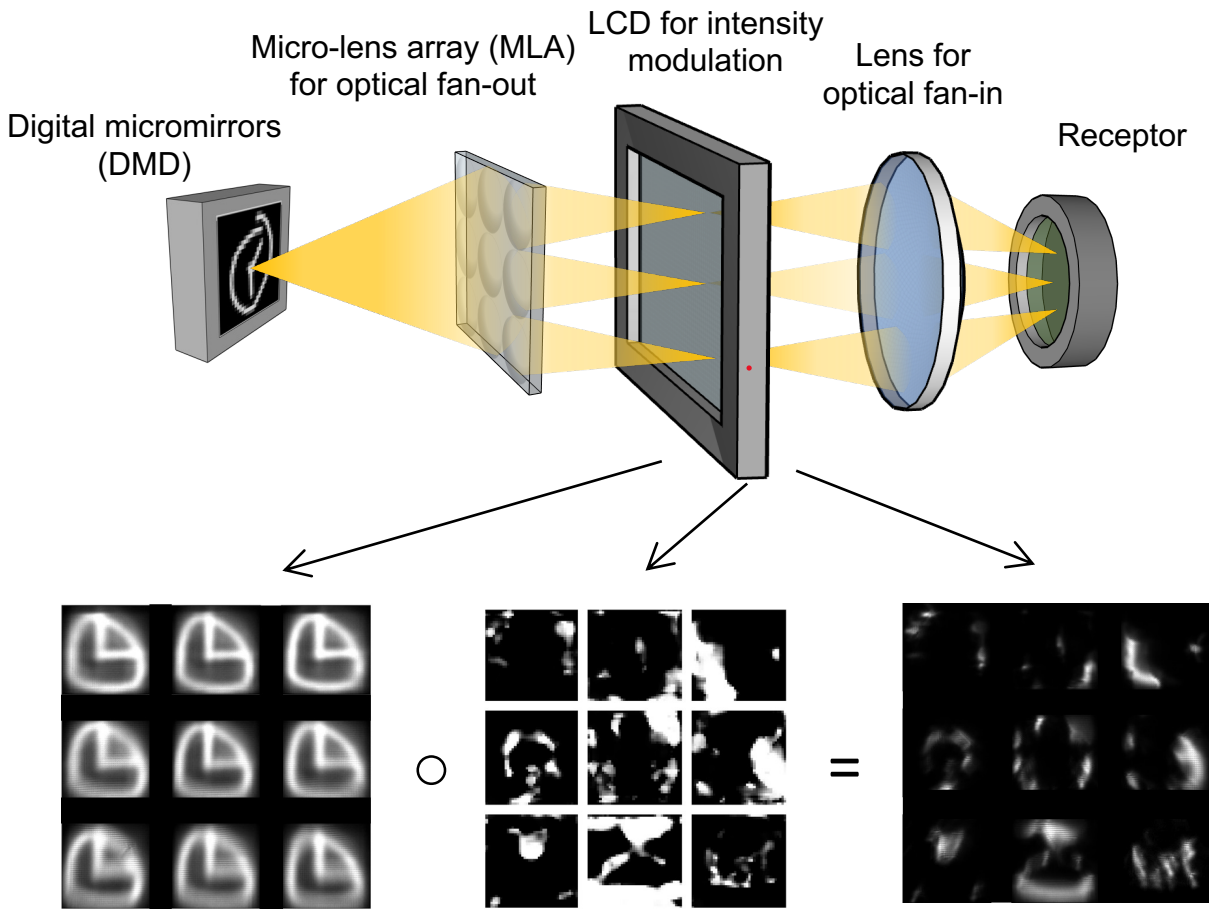
Optical-to-optical nonlinear activation function σ

Optical fully connected layer W_2

Digital single-layer backend



Implementation of 2D optical matrix-vector multiplication



Microlens array for optical fanouts

T. Georgiev, & C. Intwala. Adobe Inc.(2006)

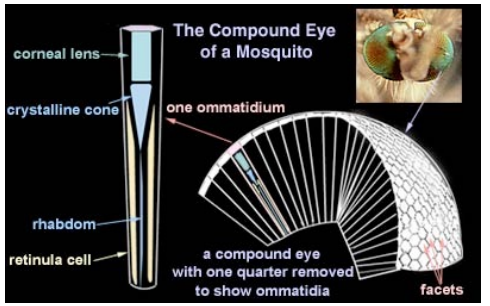


Figure credit:
https://www.ese.wustl.edu/~nehorai/research/biomim/compoundeye_b_r.html

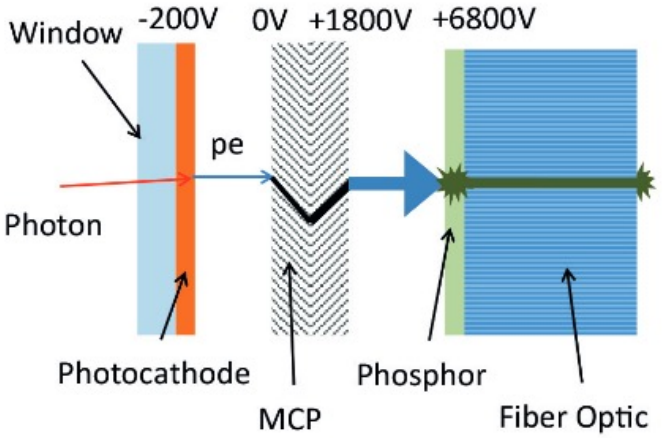
Scheme features

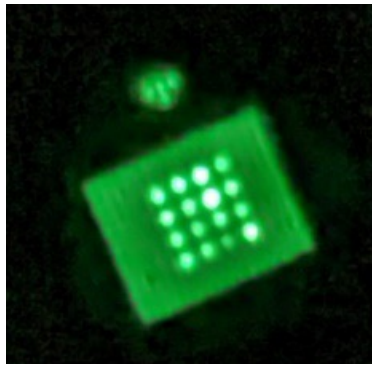
- Compatible with both coherent and incoherent light
- Directly processes 2D images
- Contains 3D object information
- Fully optical fan-out and fan-in with a minimal amount of E-O/O-E conversion.

Optical-to-optical nonlinearity via a microchannel plate

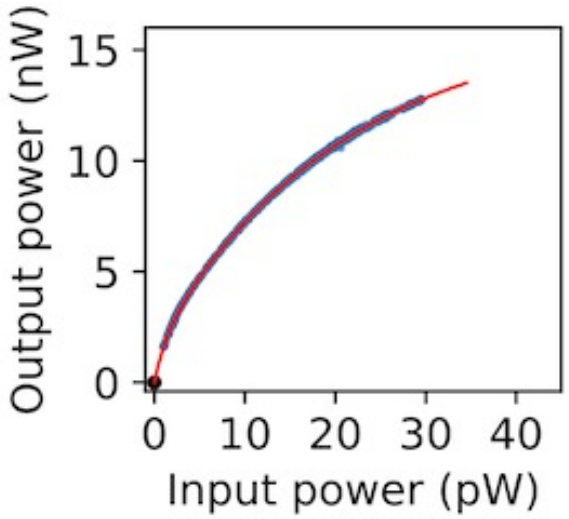
Images from Photek Inc.

The cross section view of an image intensifier





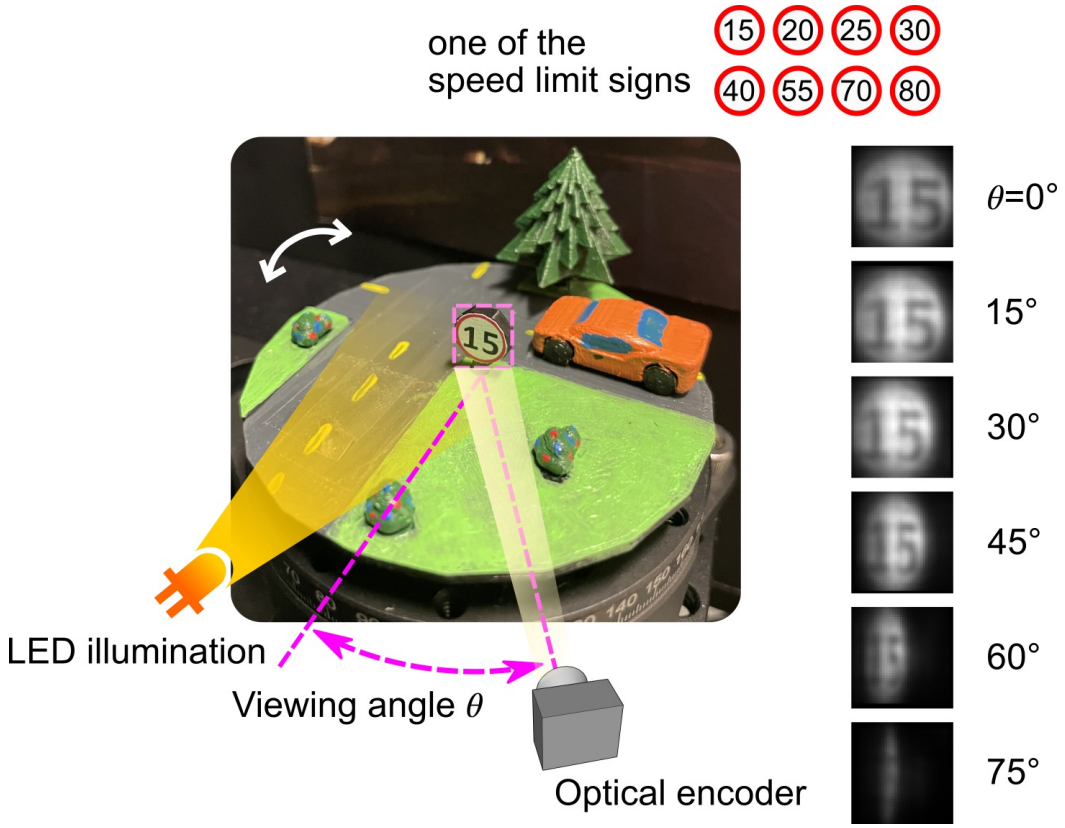
Light emission on the phosphor side



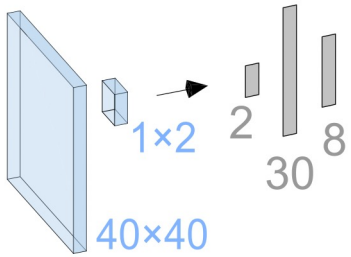
- **Cascadable**: operates under low light and provides optical amplification
- **High resolution**: independently amplifies many spatial modes.
- **Nonlinear activation**: depletion of stripe current (up to ~10 MHz in bandwidth).

Recognition of 3D physical objects

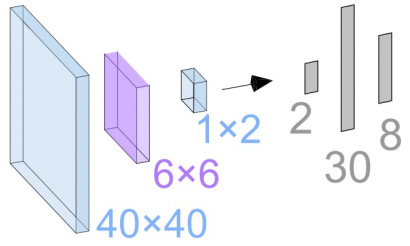
- Our ONN image sensor can directly process broadband, incoherent light reflected from 3D physical objects.

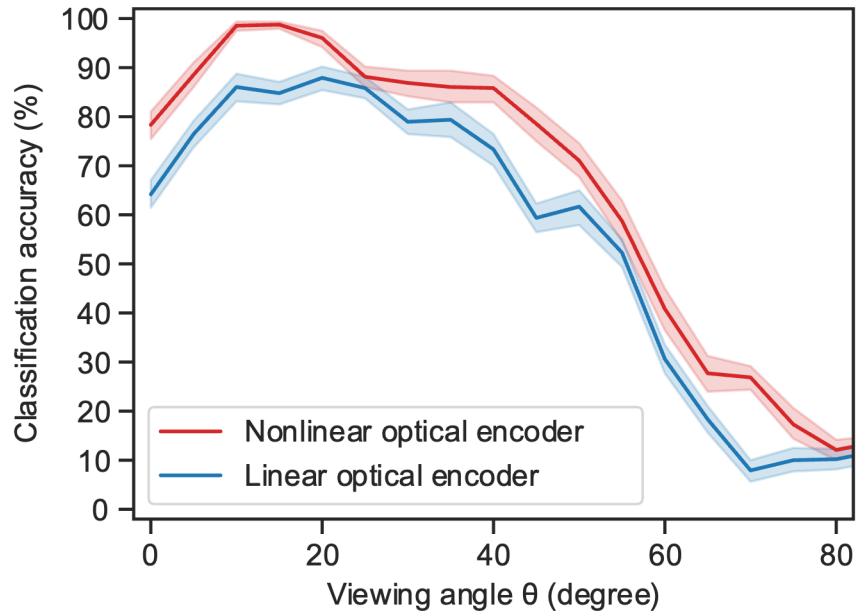


Linear optical encoder



Nonlinear optical encoder



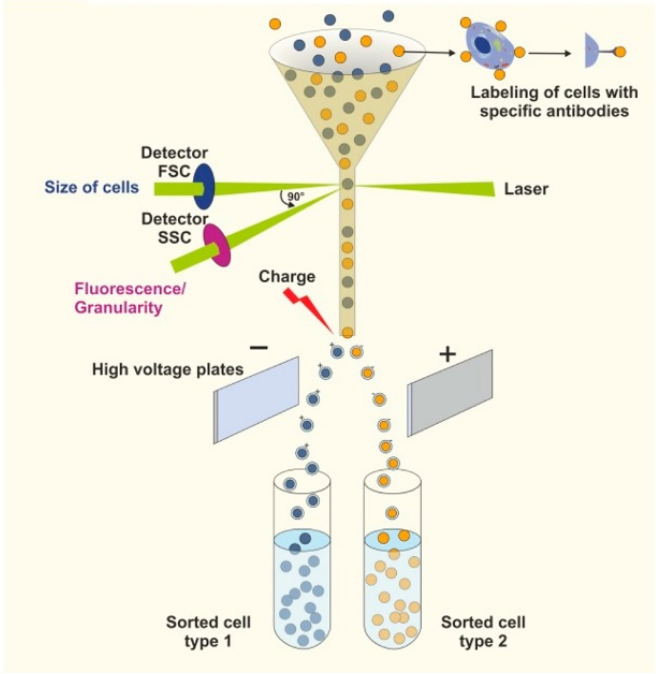


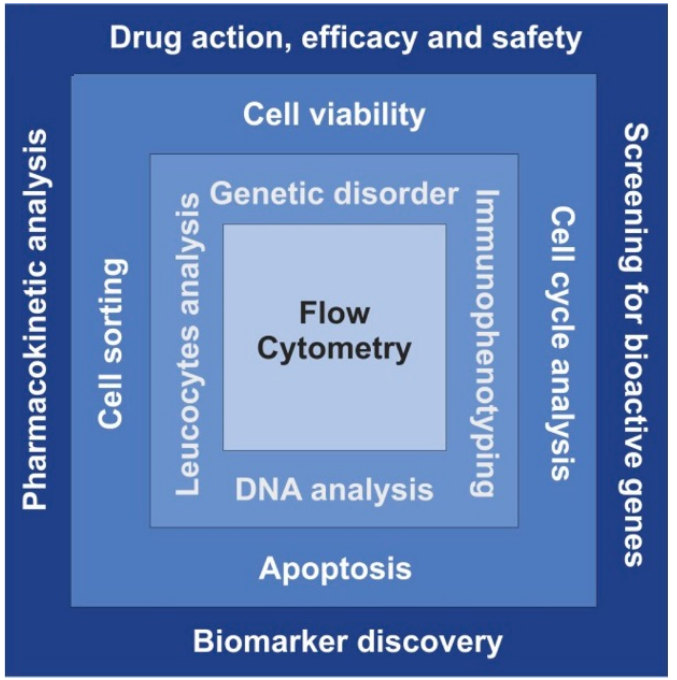
shade = standard deviation of 20 samples

Application to flow cytometry

Flow cytometry is a versatile tool for both fundamental research & clinic diagnosis:

- **High throughput rate**: 10k -100k cells per second
- **Low latency**: image classification in <1 ms
- **Low optical power**: prevent phototoxicity



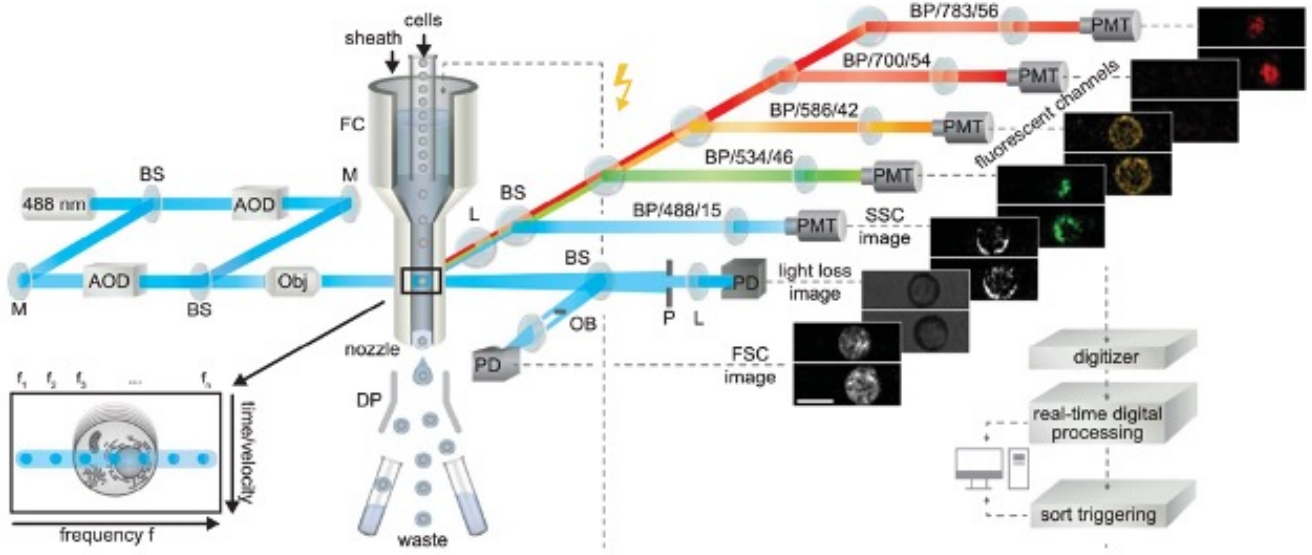


H. Drescher, S. Weiskirchen, & R. Weiskirchen. Flow cytometry: a blessing and a curse. *Biomedicines*, *9*(11), 1613 (2021).

Upgrade to image-based flow cytometry

- Image information (in addition to fluorescence) improves cell sorting accuracy.
- Image-based flow cytometry is a highly demanding machine-vision task.

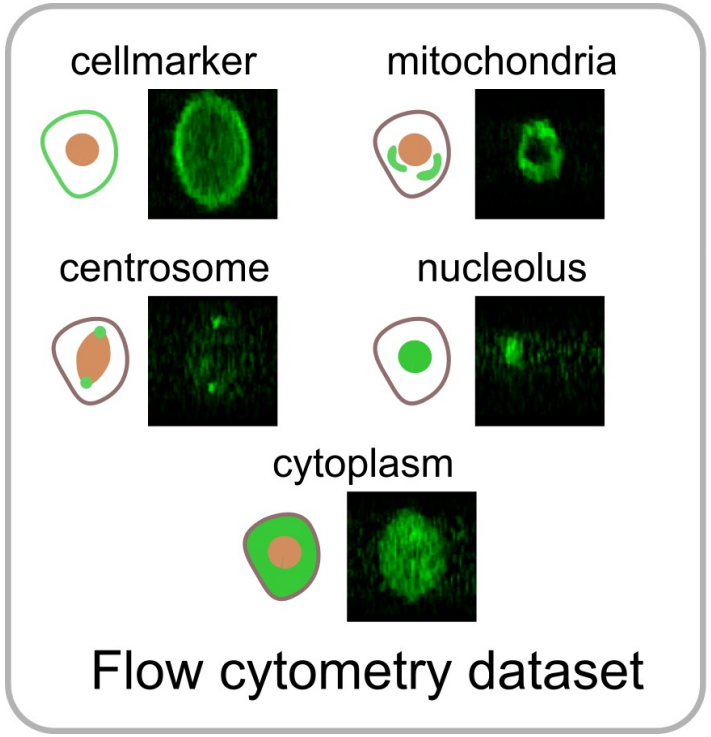
An image-based flow cytometry for cell sorting

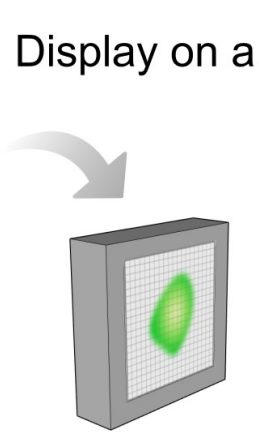


Daniel Schraivogel, … Lars M. Steinmetz "High-speed fluorescence image–enabled cell sorting." *Science* **375**.6578 (2022): 315-320.

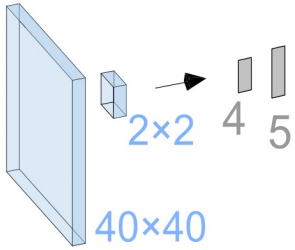
An ONN image sensor for image-based flow cytometry

Classification of cell organelles based on fluorescent images acquired in real flow cytometry experiments.



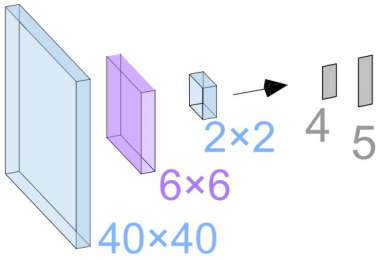


Linear optical encoder



Test accuracy: 88.1%

Nonlinear optical encoder

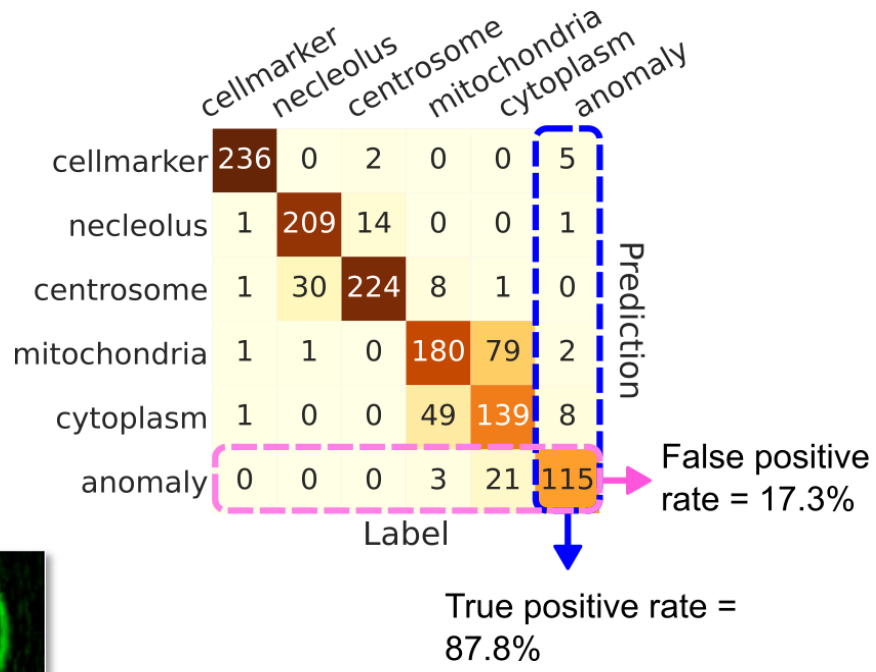


Test accuracy: 93.0%

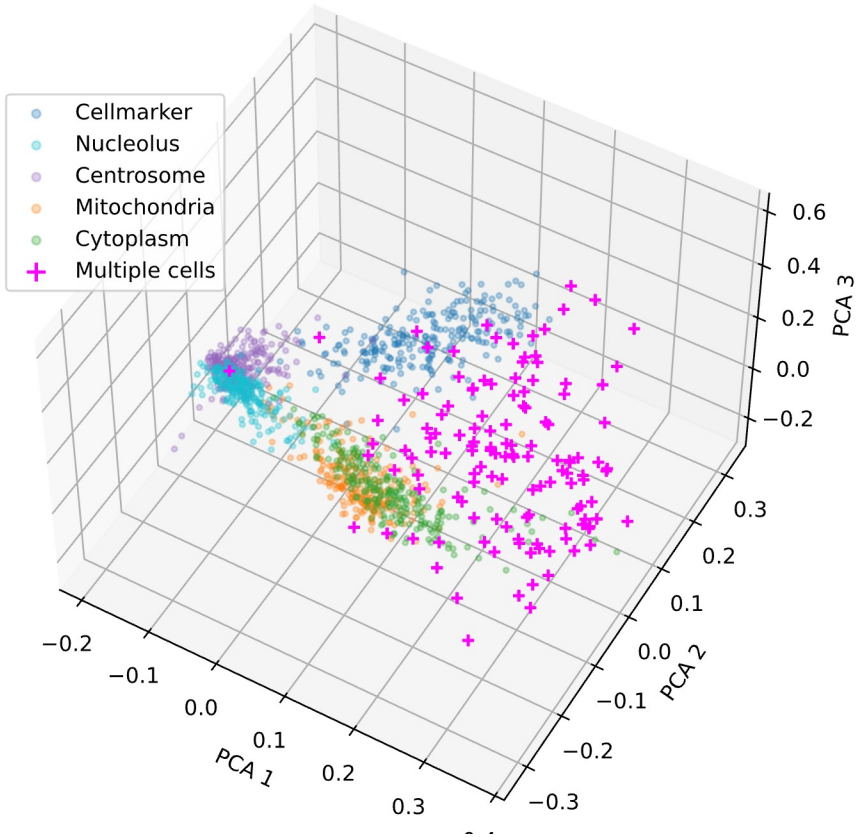
Data source: Daniel Schraivogel, … Lars M. Steinmetz "High-speed fluorescence image–enabled cell sorting." *Science* **375**.6578 (2022): 315-320.

Unsupervised learning: anomaly detection

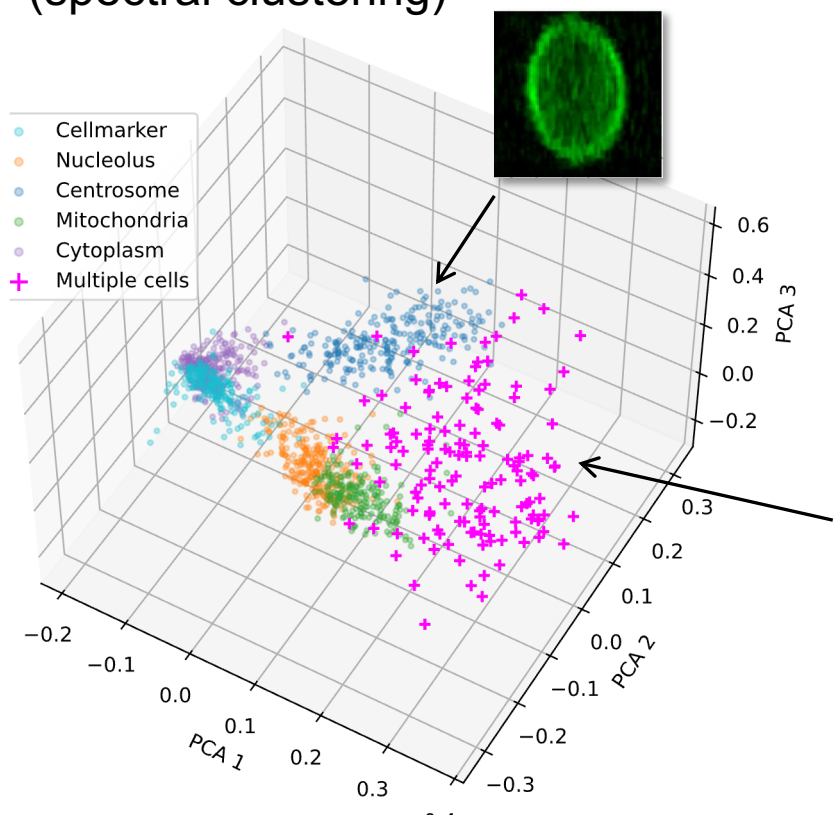
- Data structure is well preserved in the latent space.
- Anomaly data points can even be detected without training.



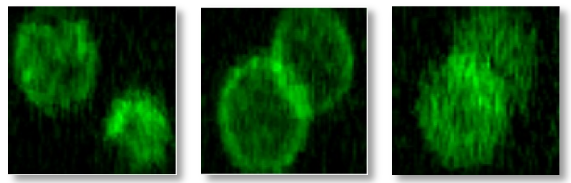
Ground truth labels



Unsupervised clustering (spectral clustering)



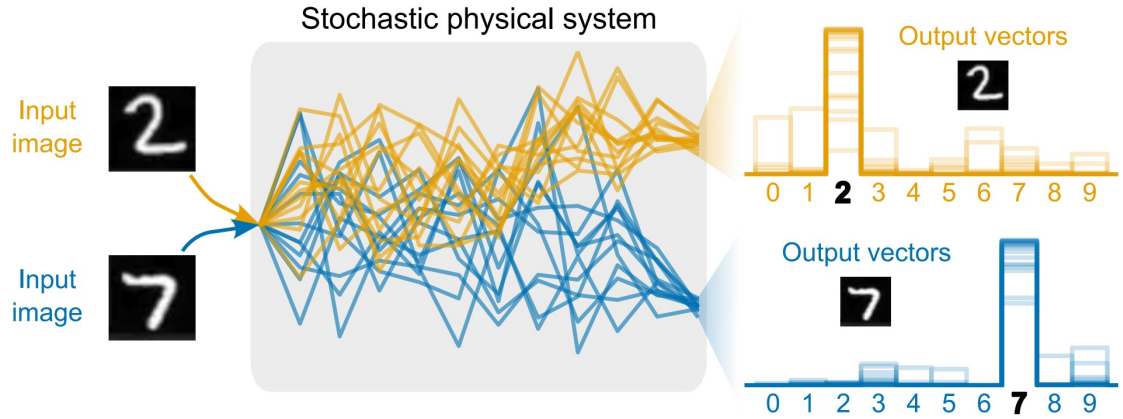
Anomaly: doublet cell clusters

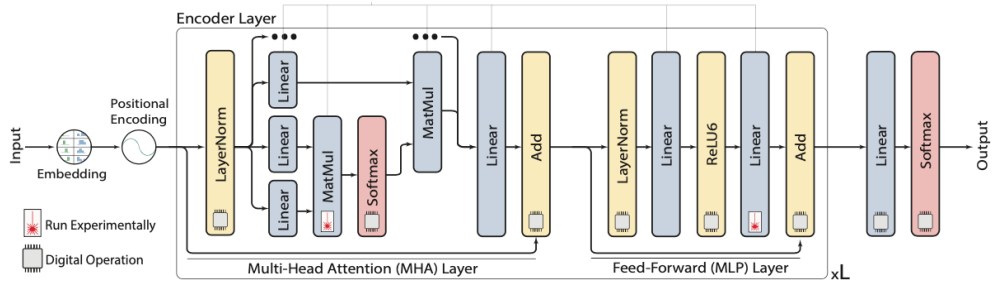


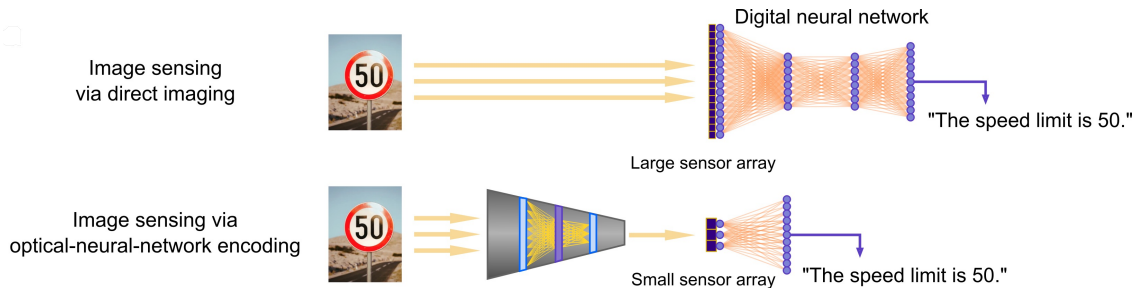
Summary

We have shown that:

- ONNs can be trained to be highly noise resilient when working in the single-photon regime.
- ONNs support an energy scaling rule favorable for large language models.
- ONNs can perform optical-domain data compression with extremely low latency.







Optical neural networks

Summary and references

Shot-noise limit of analog photonic neural networks – computing with less than one photon

T. Wang, S-Y, Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon.
An optical neural network using less than 1 photon per multiplication. *Nature Communications* **13**: 123 (2022)

Stochastic single-photon detected neural networks – computing with WAY less than one photon

S.-Y. Ma, T. Wang, J. Laydevant, L.G. Wright & P.L. McMahon. Quantum-noise-limited optical neural networks operating at a few quanta per Activation. *arXiv:2302.10360* (2023)

Optical hardware's advantage scales favorably with model size (for LLMs!)

M. Anderson, T. Wang, S.-Y. Ma, L.G. Wright and P.L. McMahon, Optical Transformers, *arXiv:2307.15712* (2023)

Optical pre-processing allows qualitatively new, quantitatively better image sensors

T. Wang*, M. M. Sohoni*, L. G. Wright, …, P. L. McMahon. Image sensing with multilayer, nonlinear optical neural networks. *Nature Photonics* **17**, 408–415 (2023)

Many opportunities in physics-aware software and breaking down hardware-software

L.G. Wright*, T. Onodera*, M.M. Stein, T. Wang, D.T. Schachter, Z. Hu, P.L. McMahon, Deep physical neural networks trained with backpropagation, *Nature* **601**, 549-555 (2022)

Joint work with J. Laydevant, T. Wang & P.L. McMahon, "The hardware is the software", *Neuron* (2023)