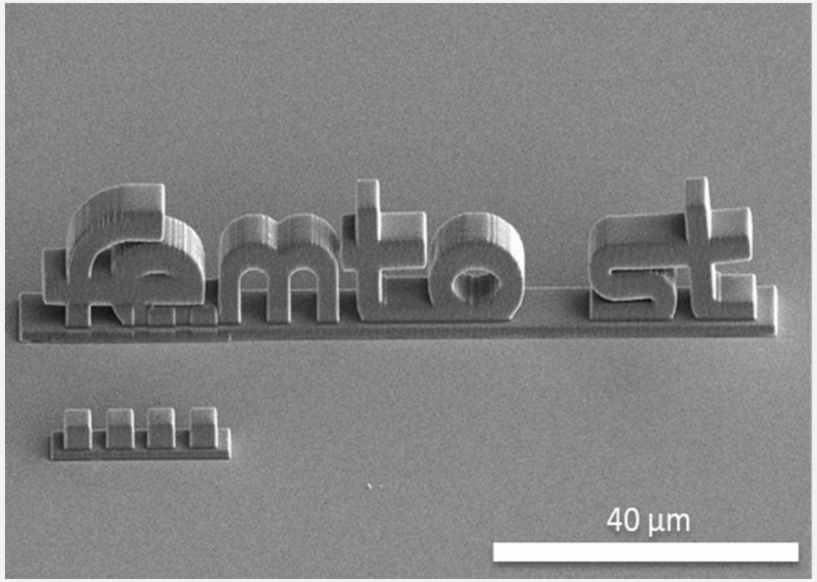
Optical computing

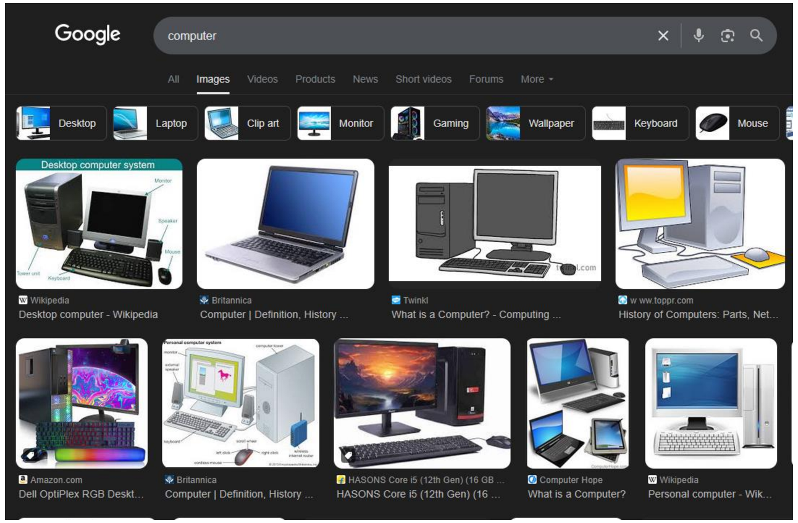
Daniel Brunner
FEMTO-ST, CNRS/UFC

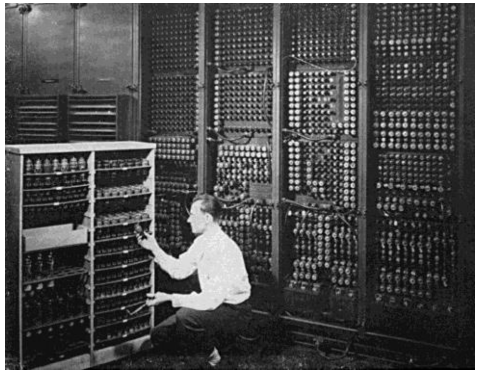
SMYLE Summerschool
Neuchâtel, Switzerland

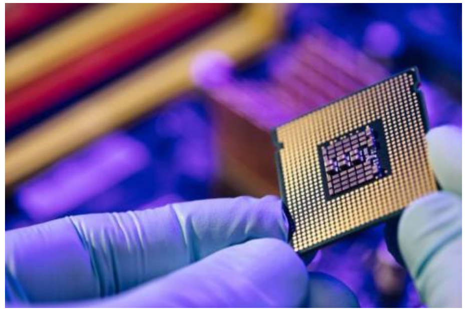


Computing

What is a computer ? / What is computing?


Google
computer
All
Images
Videos
Products
News
Short videos
Forums
More
Desktop
Laptop
Clip art
Monitor
Gaming
Wallpaper
Keyboard
Mouse
Desktop computer system
Wikipedia
Desktop computer - Wikipedia
Britannica
Computer | Definition, History ...
Twinkl
What is a Computer? - Computing ...
w ww.toppr.com
History of Computers: Parts, Net...
Amazon.com
Dell OptiPlex RGB Deskt...
Britannica
Computer | Definition, History ...
HASONS Core i5 (12th Gen) (16 GB ...
HASONS Core i5 (12th Gen) (16 ...
Computer Hope
What is a Computer?
Wikipedia
Personal computer - Wik...

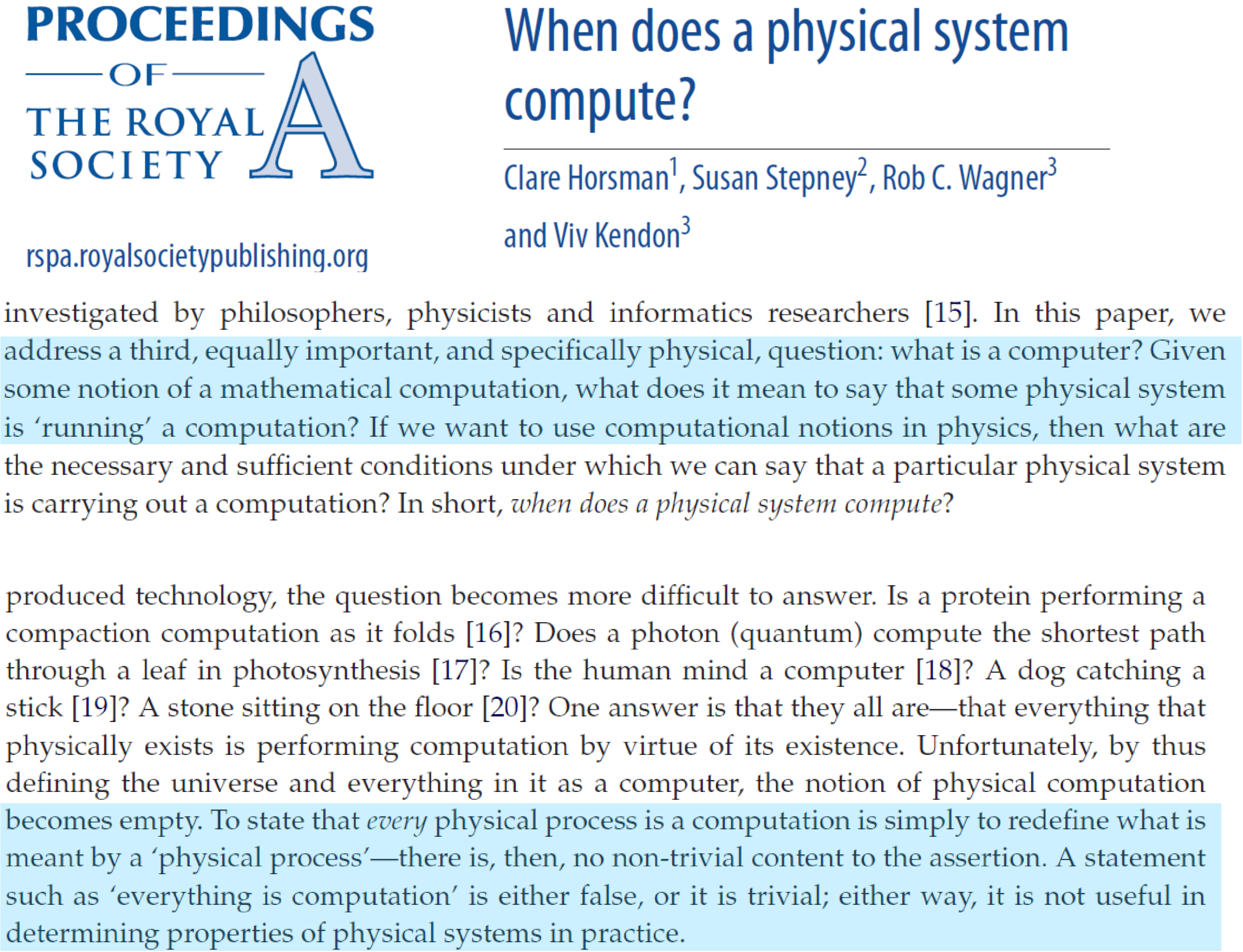




BROAD DEFINITIONS

Computing is the process of **transforming information** according to well-defined rules.

A **computer** is a physical system that performs computation that is, it takes **inputs**, **processes** them according to **rules**, and produces **outputs**.



PROCEEDINGS OF THE ROYAL SOCIETY A

rspa.royalsocietypublishing.org

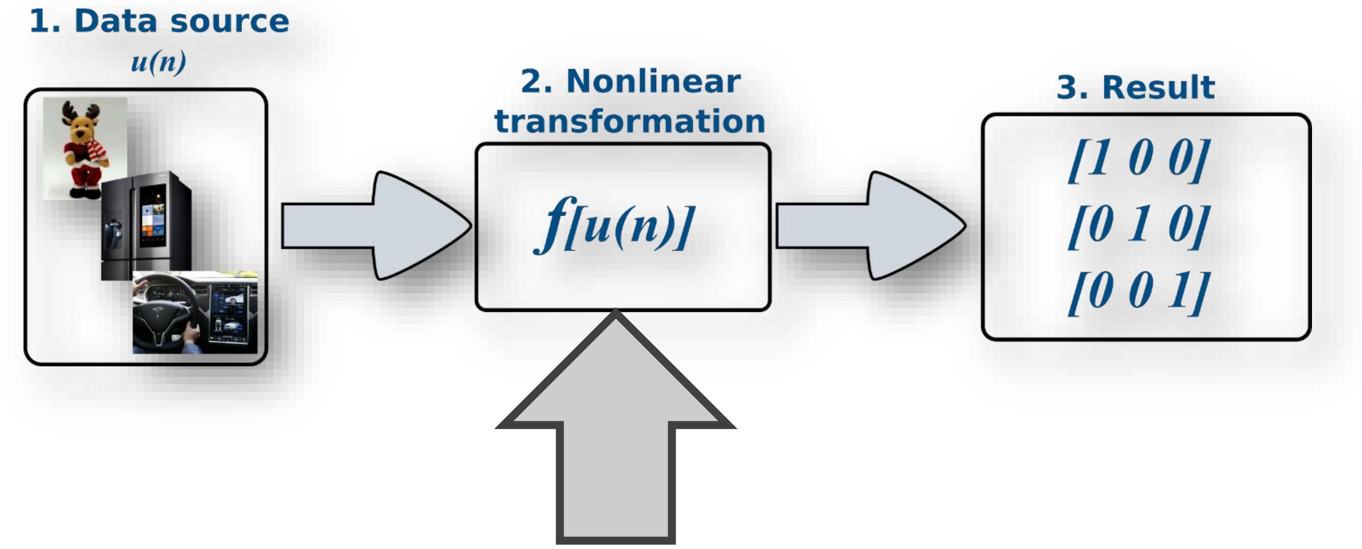
When does a physical system compute?

Clare Horsman[1], Susan Stepney[2], Rob C. Wagner[3] and Viv Kendon[3]

investigated by philosophers, physicists and informatics researchers [15]. In this paper, we address a third, equally important, and specifically physical, question: what is a computer? Given some notion of a mathematical computation, what does it mean to say that some physical system is 'running' a computation? If we want to use computational notions in physics, then what are the necessary and sufficient conditions under which we can say that a particular physical system is carrying out a computation? In short, *when does a physical system compute*?

produced technology, the question becomes more difficult to answer. Is a protein performing a compaction computation as it folds [16]? Does a photon (quantum) compute the shortest path through a leaf in photosynthesis [17]? Is the human mind a computer [18]? A dog catching a stick [19]? A stone sitting on the floor [20]? One answer is that they all are—that everything that physically exists is performing computation by virtue of its existence. Unfortunately, by thus defining the universe and everything in it as a computer, the notion of physical computation becomes empty. To state that *every* physical process is a computation is simply to redefine what is meant by a 'physical process'—there is, then, no non-trivial content to the assertion. A statement such as 'everything is computation' is either false, or it is trivial; either way, it is not useful in determining properties of physical systems in practice.

Computing: a nonlinear function



Programming: **identify** $f[u(n)]$

Computation: **calculate** $f[u(n)]$

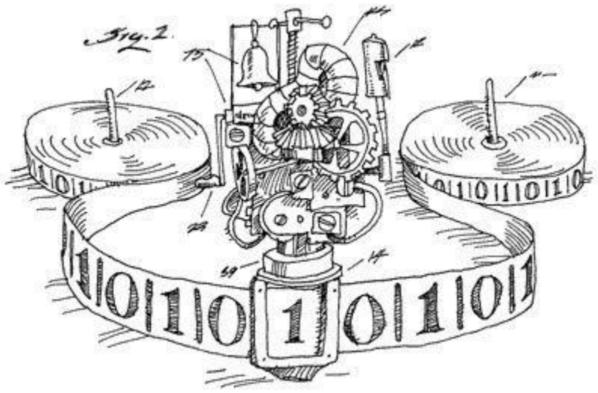
1. Programming strategies

- Algorithmic / analytical
- Regression to model input -> result

2. Computation strategies

- Abstract: Turing / von Neumann machine
- Substrate: System is function

Turing concept and von Neumann architecture

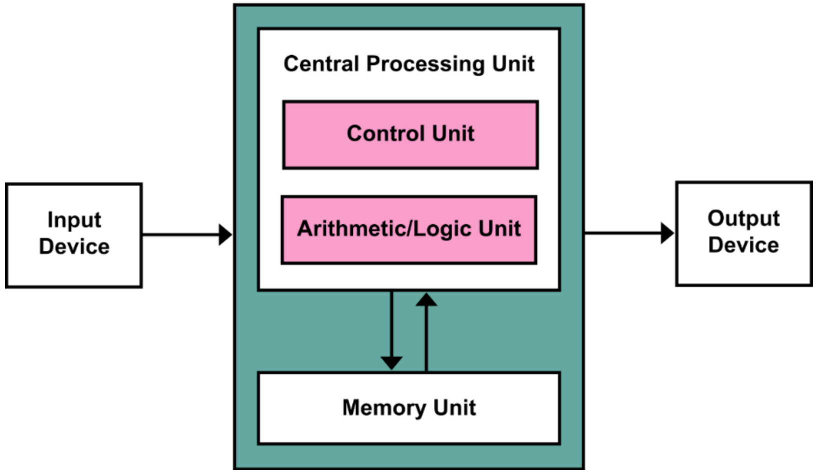


© Hussein Nur

1. Single 'point' operation, only local action
2. 'Unlimited', nonvolatile memory

➢ Serial operation according to symbolic operations

➢ Conceptual separation 'operation' and memory



https://en.wikipedia.org/wiki/Von_Neumann_architecture

1. **Essentially a 'technical' fix**
   1. Operation: in-silico (transistor)
   2. Memory: punch cards / magnetic

➢ **Combine different materials by spatially separating operation and memory**

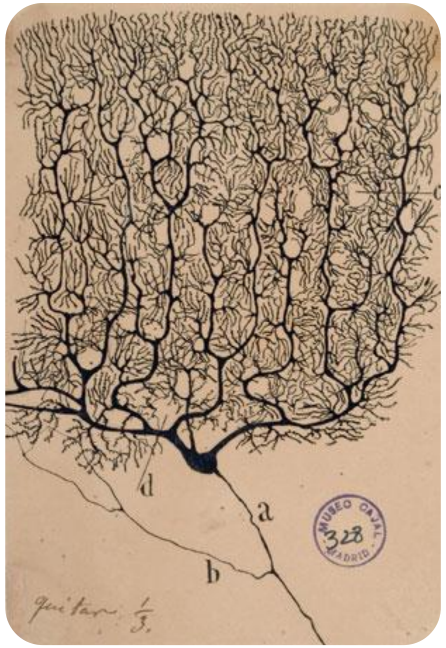
Motivation behind binary electronics

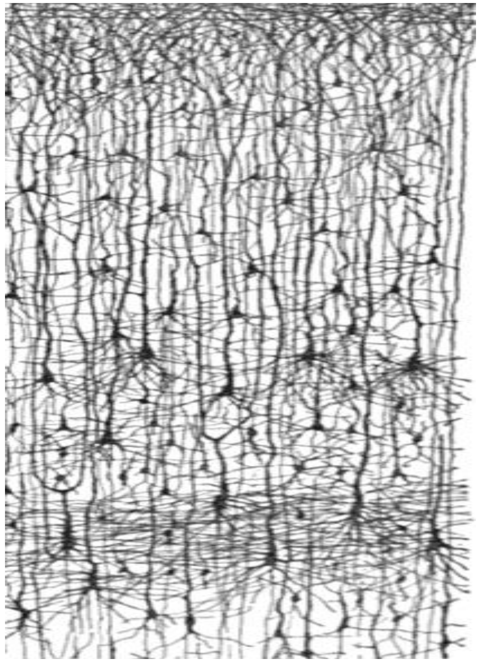
> “The binary scale seems particularly well suited for electronic computation because of its simplicity and the fact that **valve equipment** can very **easily produce and distinguish two sizes of pulse.**”
>
> Alan Turing

⇨ Nature of algorithm defines ideal system

Neural Networks (1852-1943): anatomical inspiration

© Ramón y Cajal



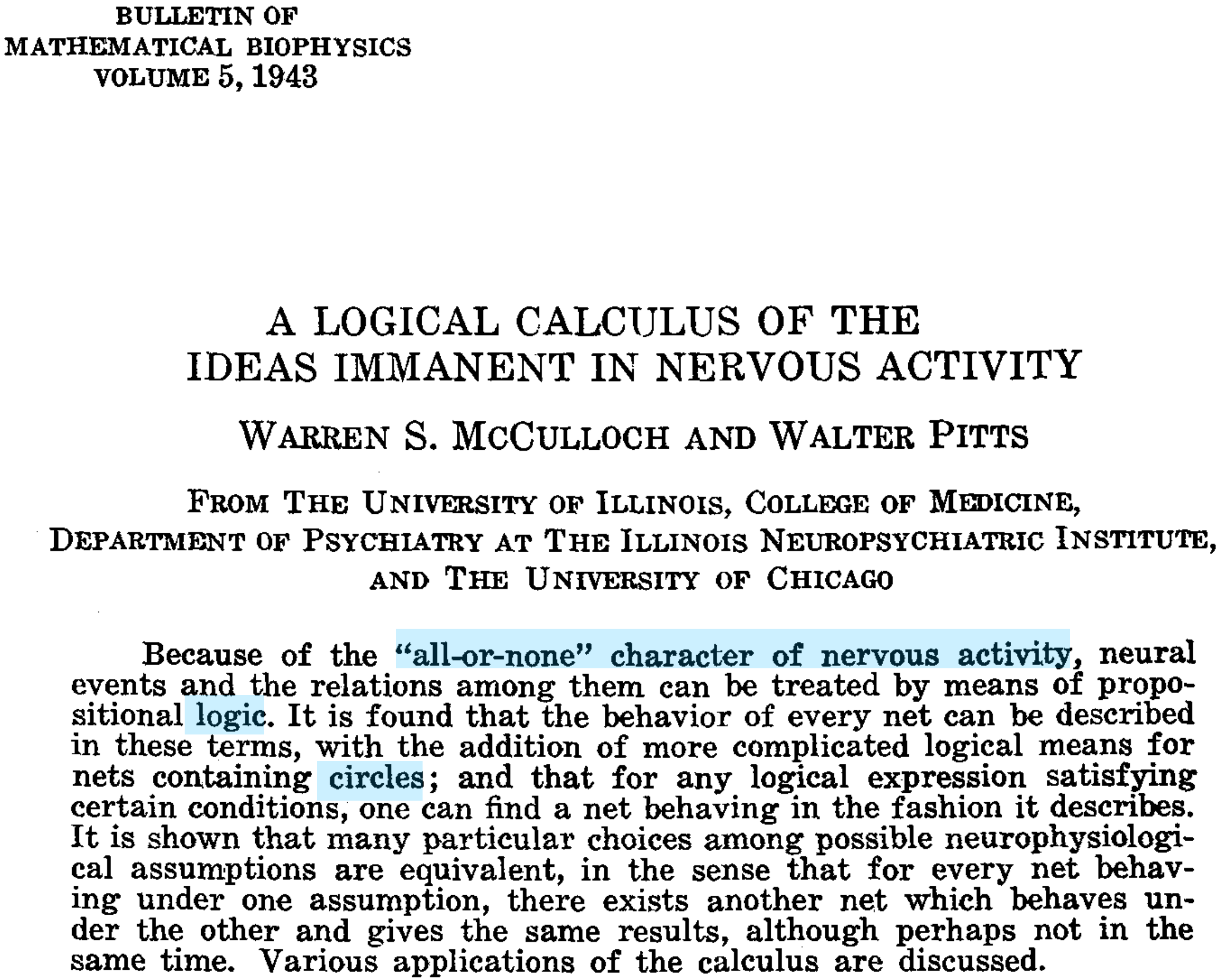


- Anatomical ‘analysis’ of human brain by Ramon y Cajal.
- Brain comprises discrete elements densely connected to a gigantic network.

- 1894: “The ability of neurons to grow in an adult and their power to create new connections can explain learning.” This statement is considered to be the origin of the synaptic theory of memory.

- Ramon y Cajal awarded Nobel prize in 1906.

Neural Networks: computational basis



BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

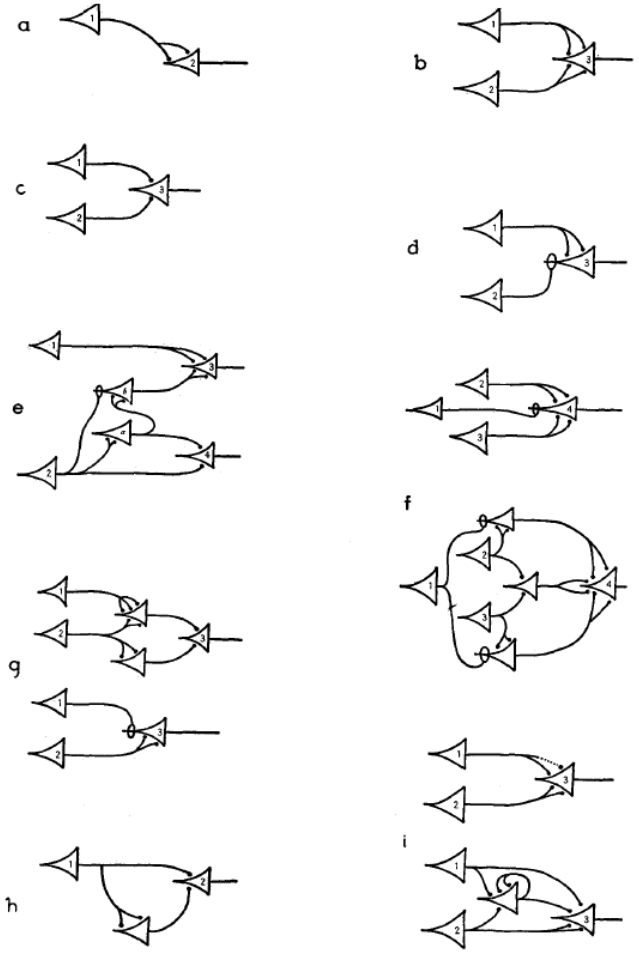
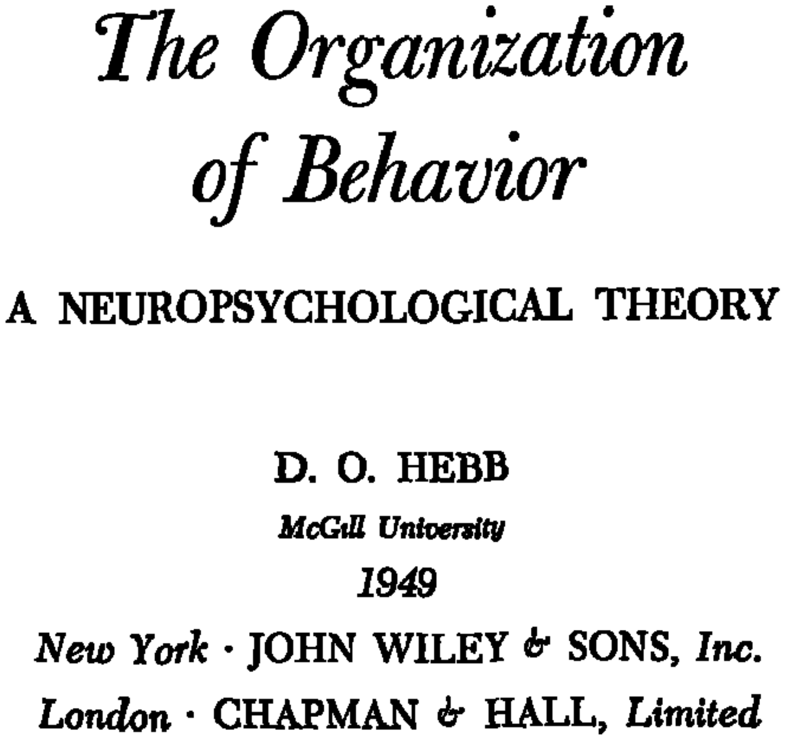


FIGURE 1

Neural Networks: ‘programming’



The Organization of Behavior

A NEUROPSYCHOLOGICAL THEORY

D. O. HEBB

McGill University

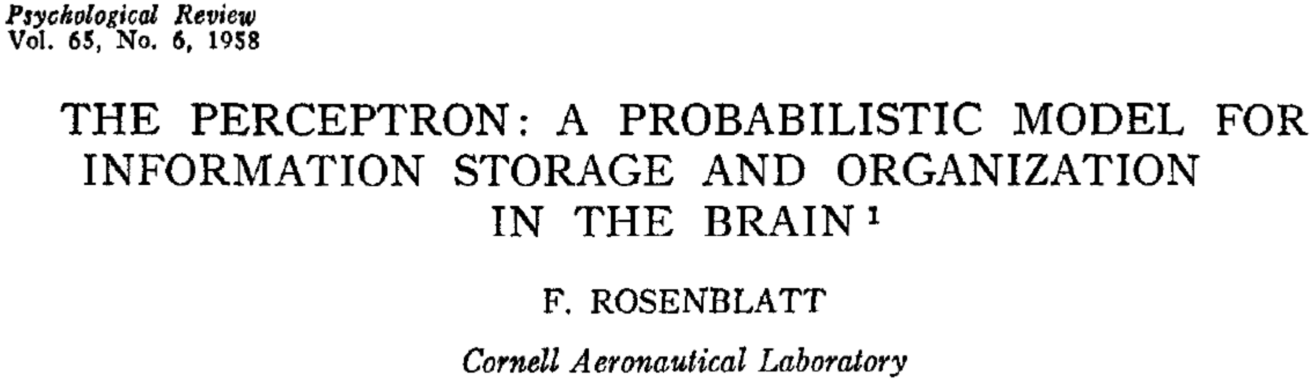
1949

New York · JOHN WILEY & SONS, *Inc.*
London · CHAPMAN & HALL, *Limited*

- “The general idea is an old one, that any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated' so that activity in one facilitates activity in the other.”
- “When one cell repeatedly assists in firing another, the axon of the first cell develops synaptic knobs (or enlarges them if they already exist) in contact with the soma of the second cell.”
- ‘Fire together: wire together’

- Recipe for ‘continuous’ programming: incrementally modify topology instead of writing a premeditated program.

First attempt: Perceptron

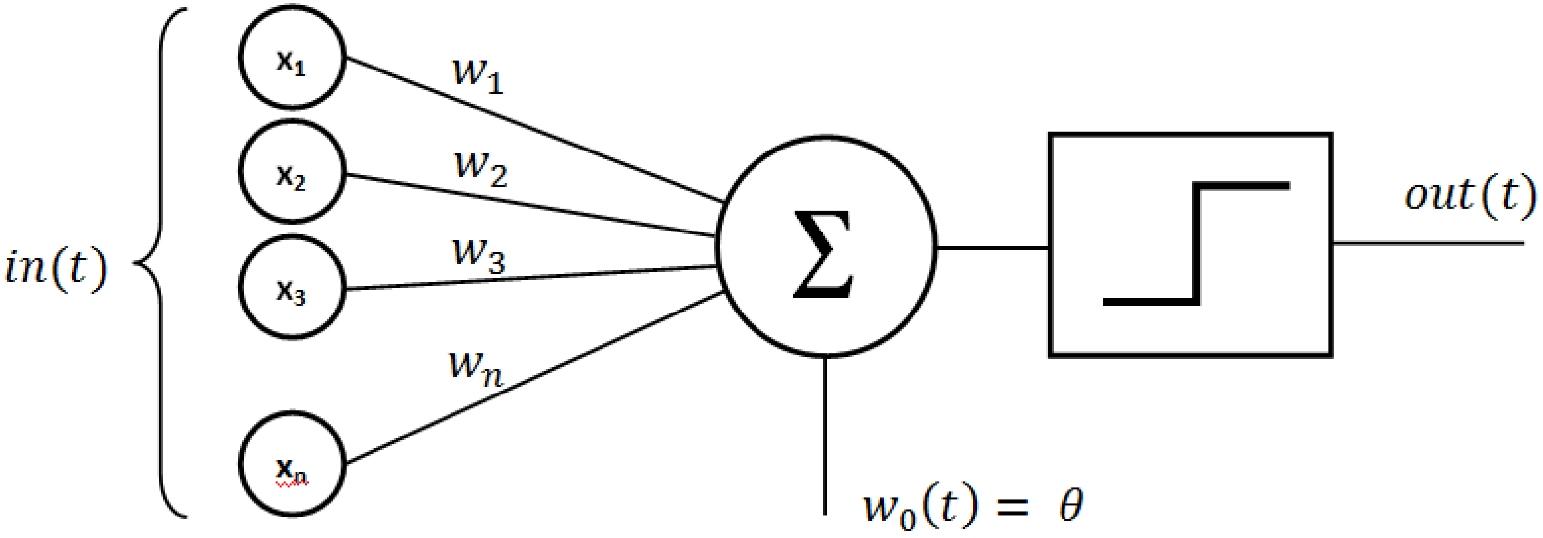


Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN[1]

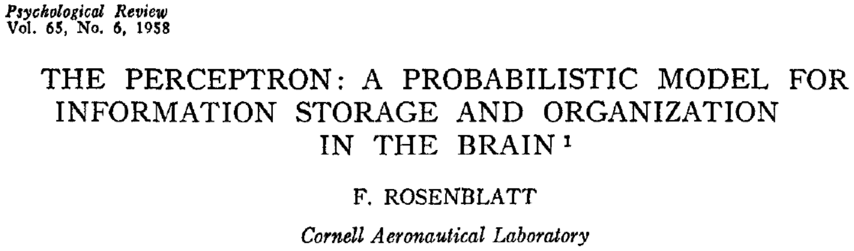
F. ROSENBLATT

Cornell Aeronautical Laboratory



Wikipedia

Initial enthusiasm

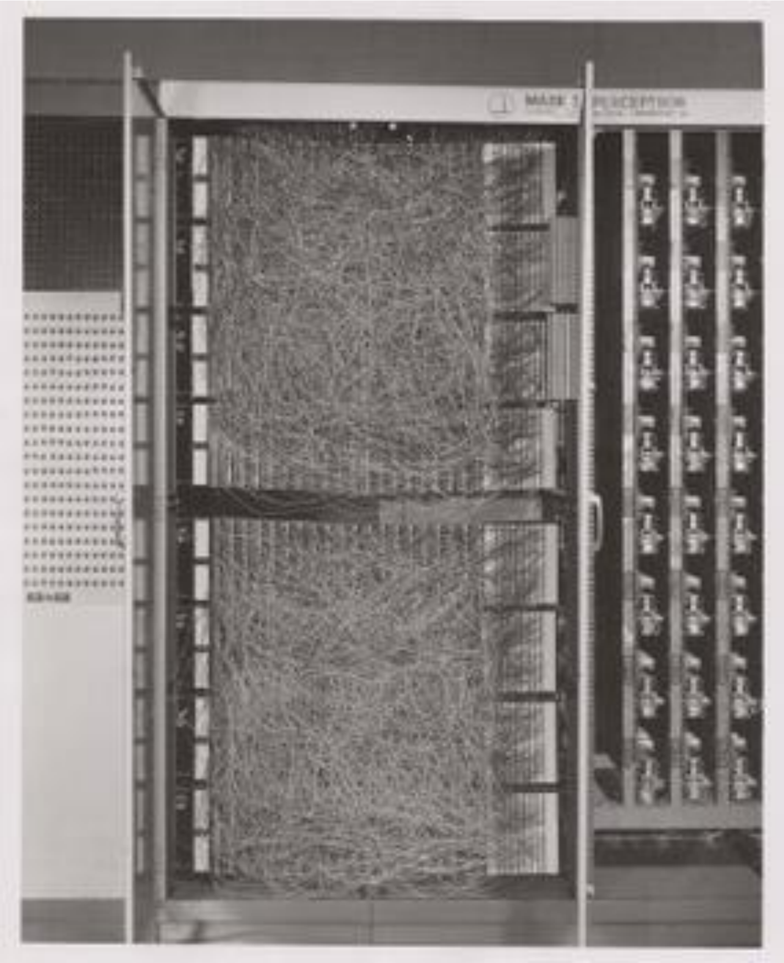


Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN[1]

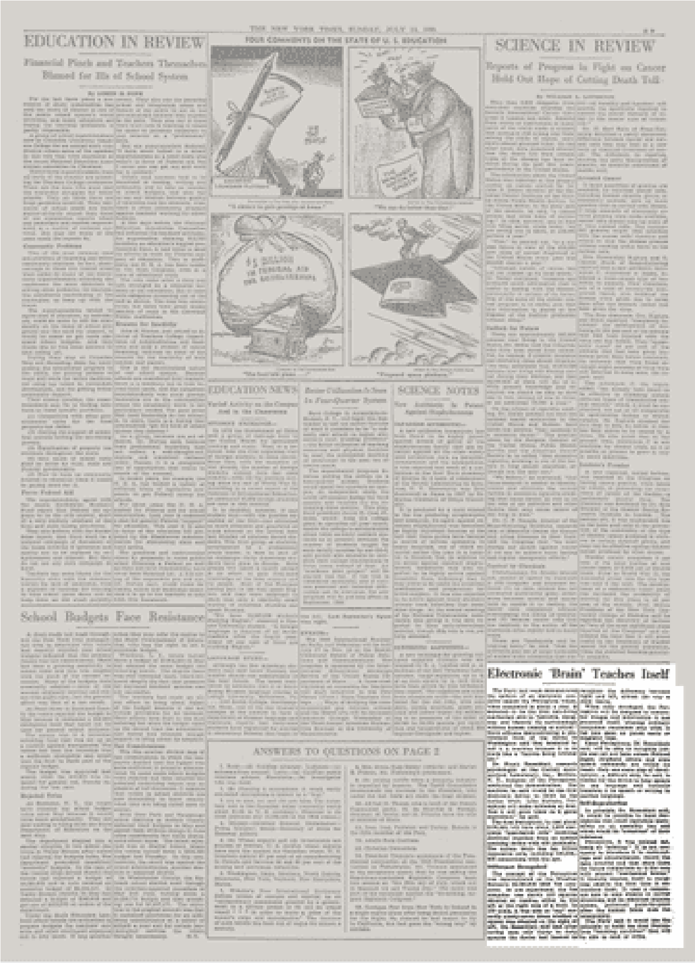
F. ROSENBLATT

Cornell Aeronautical Laboratory



The New York Times 13/07/1958:

"The Navy last week demonstrated the embryo of an electronic computer named the Perceptron which, when completed in about a year, is expected to be the first non-living mechanism able to "perceive, recognize and identify its surroundings without human training or control."



EDUCATION IN REVIEW

SCIENCE IN REVIEW

EDUCATION NEWS

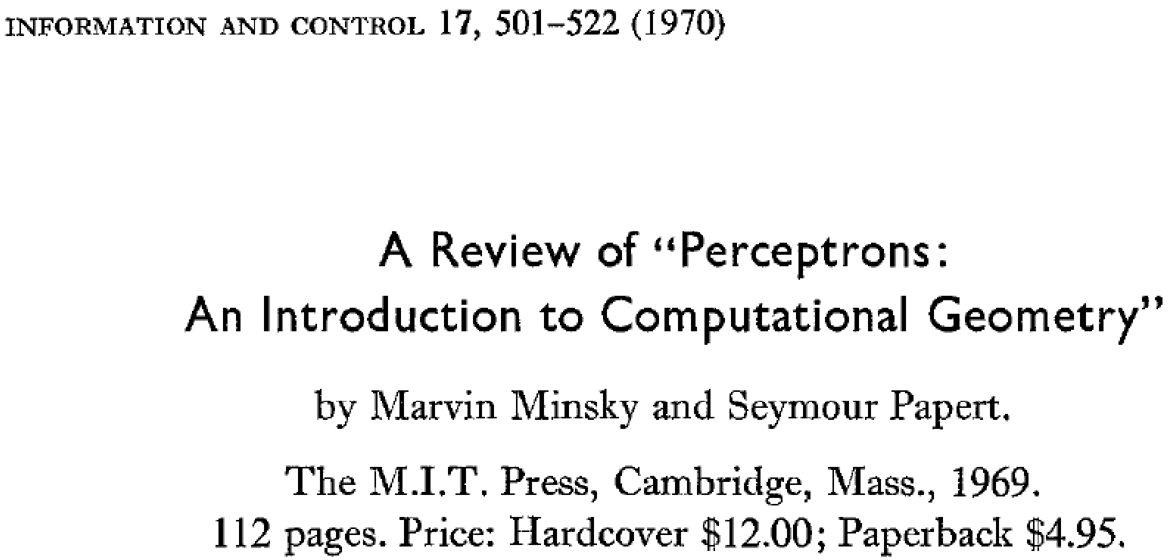
SCIENCE NOTES

School Budgets Face Resistance

ANSWERS TO QUESTIONS ON PAGE 2

Electronic 'Brain' Teaches Itself

Problem: linear separation only



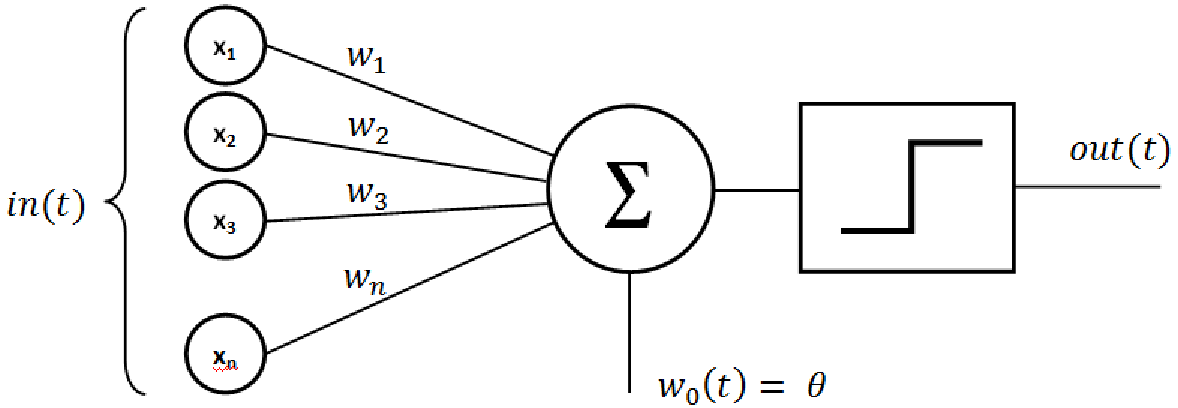
INFORMATION AND CONTROL 17, 501–522 (1970)

A Review of "Perceptrons:
An Introduction to Computational Geometry"

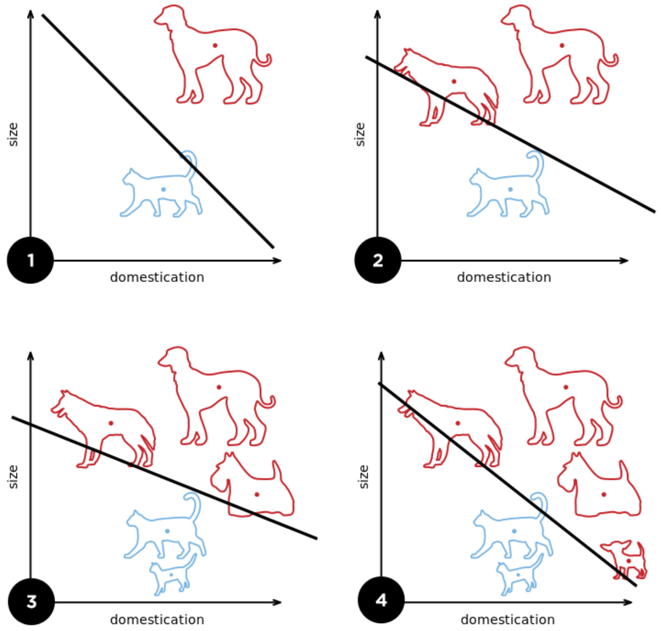
by Marvin Minsky and Seymour Papert.

The M.I.T. Press, Cambridge, Mass., 1969.
112 pages. Price: Hardcover $12.00; Paperback $4.95.

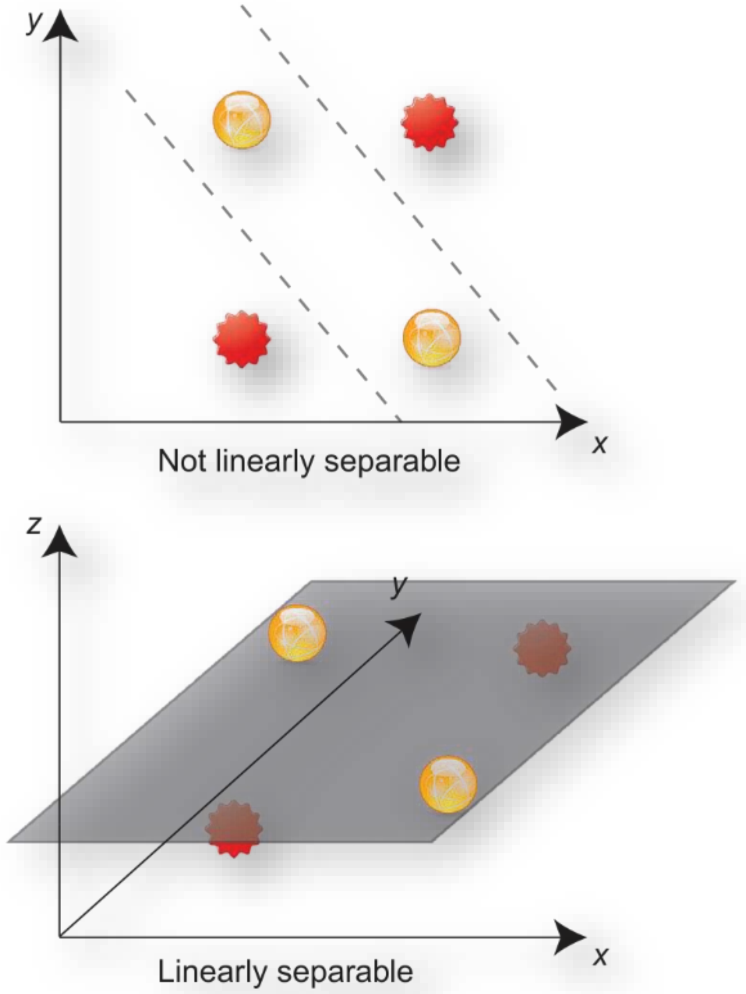
Single Perceptrons cannot solve the XOR problem or separate linearly-non separable classes



Wikipedia



The basics: why the perceptron failed the XOR



Appeltant et al., Nat. Commun. **2**, 468 (2011).

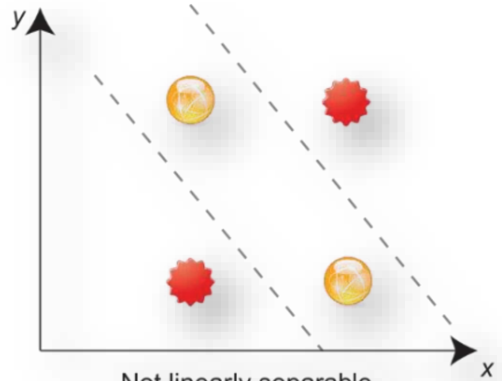
Covers theorem:

“A complex (pattern-)classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated.”

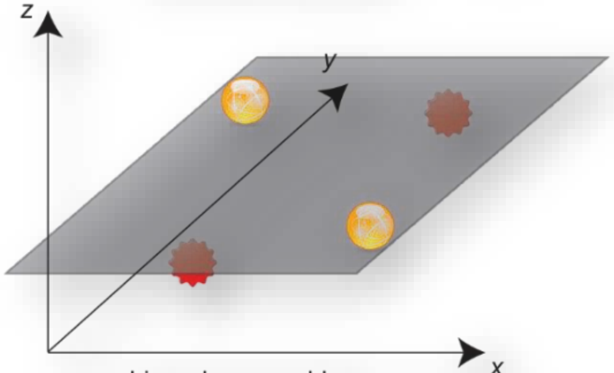
Dimensionality and nonlinearity

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{2R_1+R_2\to R_2} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{-3R_1+R_3\to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & -1 & -3 \end{bmatrix}$$

$$\xrightarrow{R_2+R_3\to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{-2R_2+R_1\to R_1} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$



Not linearly separable



Linearly separable

$$\boldsymbol{y}(n) = f[W\boldsymbol{x}(n) + b]$$

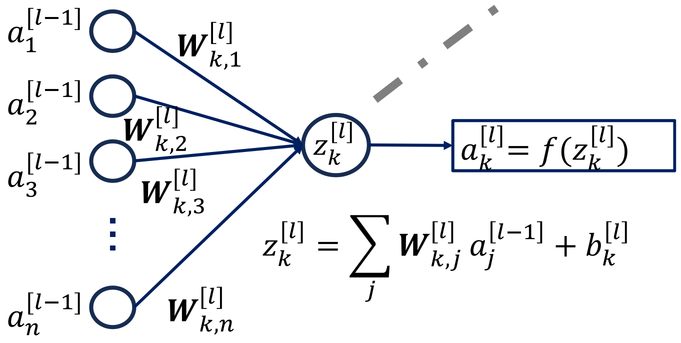
- Collect set $\boldsymbol{y}(n)$ for $\boldsymbol{x}(n), n \in [1 \dots N]$, where N is number of examples
  - $\boldsymbol{x}(n) \in \mathbb{R}^{1\times L}$
  - $\boldsymbol{y}(n) \in \mathbb{R}^{1\times M}$
  - Lets assume : $L \ll M$
  - Append all $\boldsymbol{x}(n)$ and $\boldsymbol{y}(n)$ to $\boldsymbol{X} \in \mathbb{R}^{N\times L}$ and $\boldsymbol{Y} \in \mathbb{R}^{N\times M}$

- Dimensions in vector space
  - scalar product between vectors is zero: orthogonal
  - Linear independent vectors are not parallel
  - Of a matrix: rank of matrix
  - Without nonlinearity: rank of $\boldsymbol{X}$ is rank of $\boldsymbol{Y}$ is L.

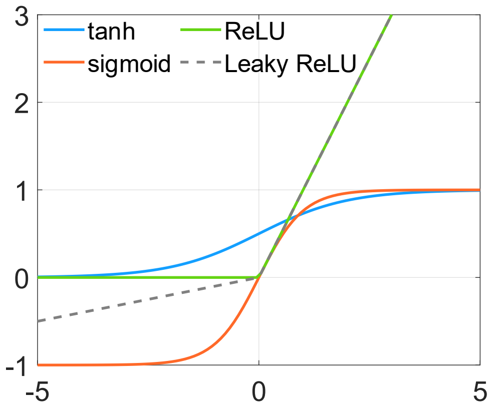
➢ **Why nonlinearity is essential: dimensionality expansion**

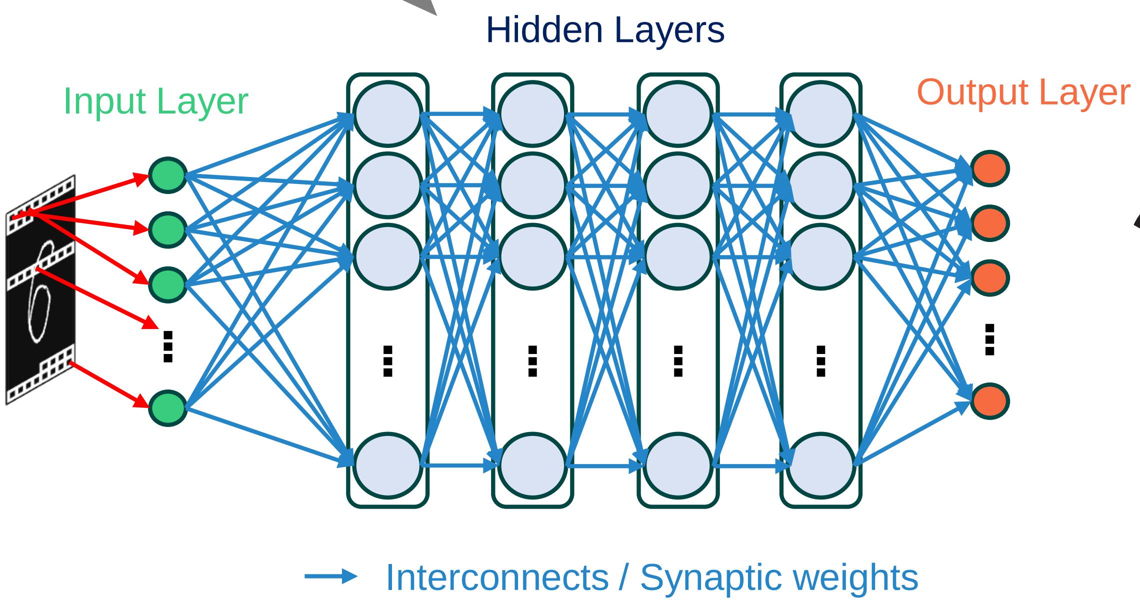
WHAT IS A NEURAL NETWORK?

Why The interest in Neural Networks in the first place ?



$$z_k^{[l]} = \sum_j W_{k,j}^{[l]}\, a_j^{[l-1]} + b_k^{[l]}$$





∼

Ramon y Cajal

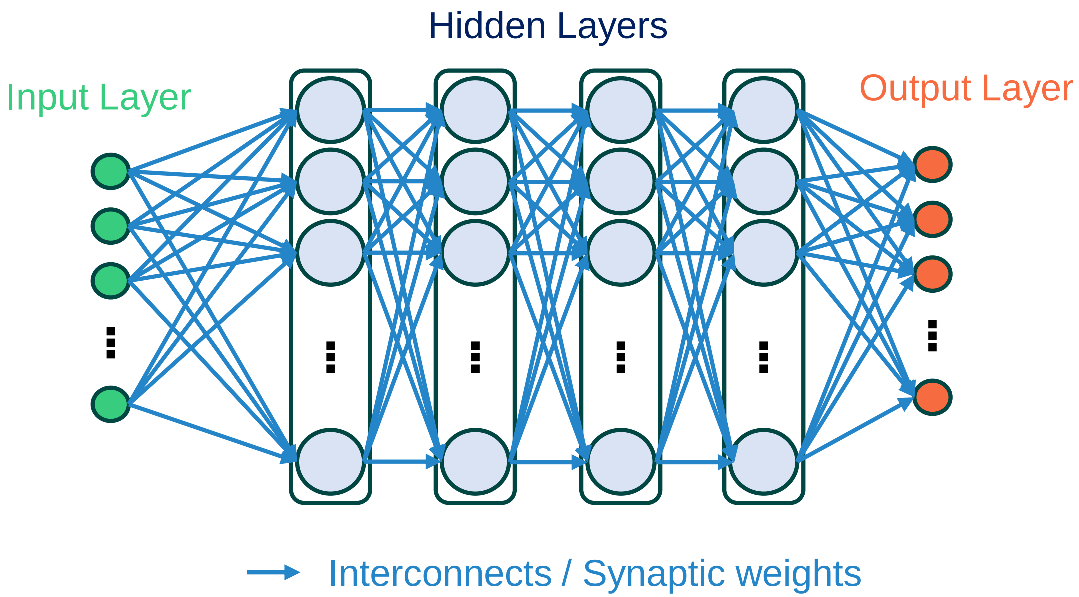
Nonlinearity and depth allow for more complex information processing.

HOW DOES A NEURAL NETWORK LEARN?

Example : MNIST Dataset

Input images Labels / Target

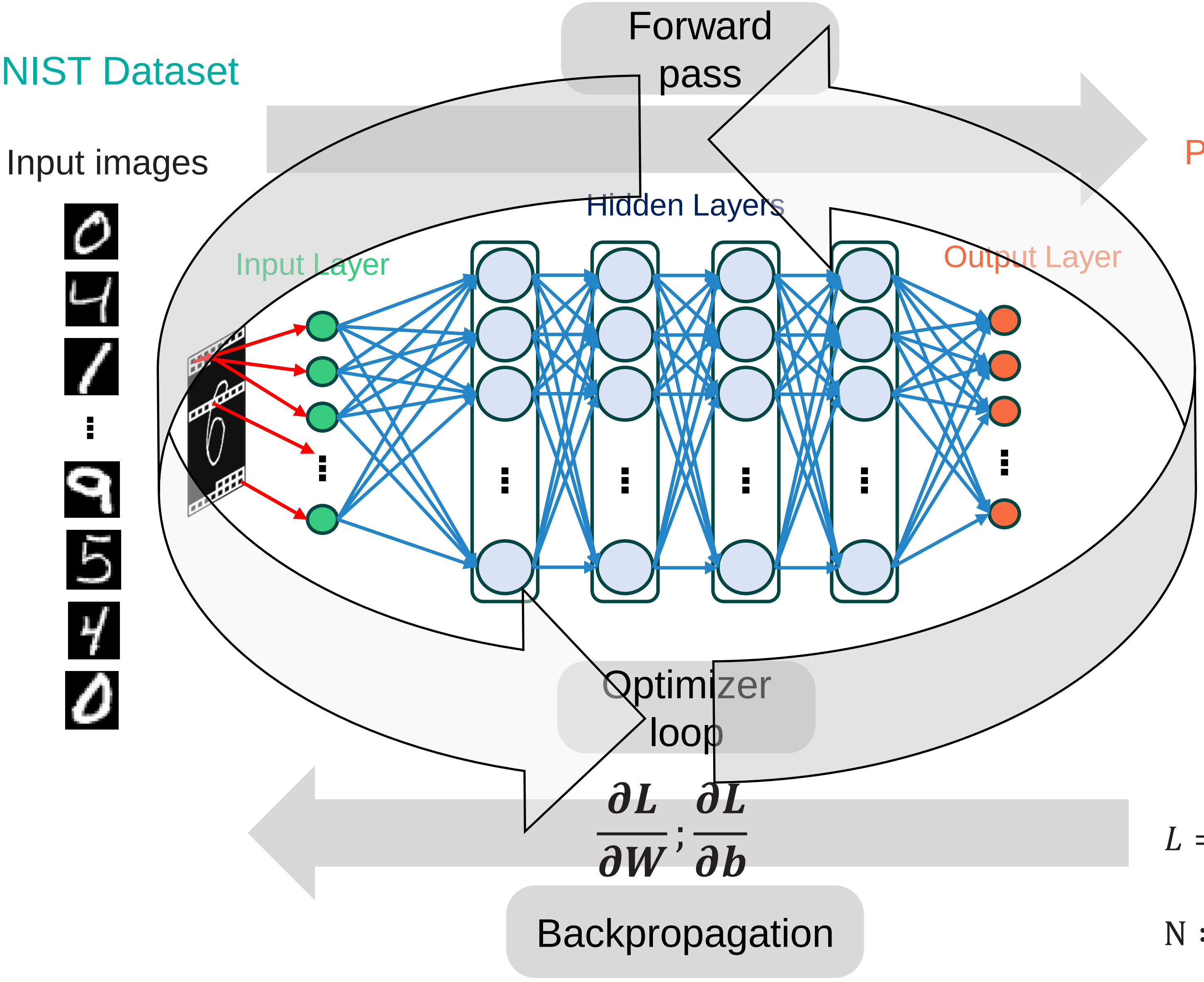
“0”
“4”
“1”
⋮
“9”
“5”
“4”
"0”



60 000 images with corresponding labels: supervised learning.

HOW DOES A NEURAL NETWORK LEARN?

Example : MNIST Dataset



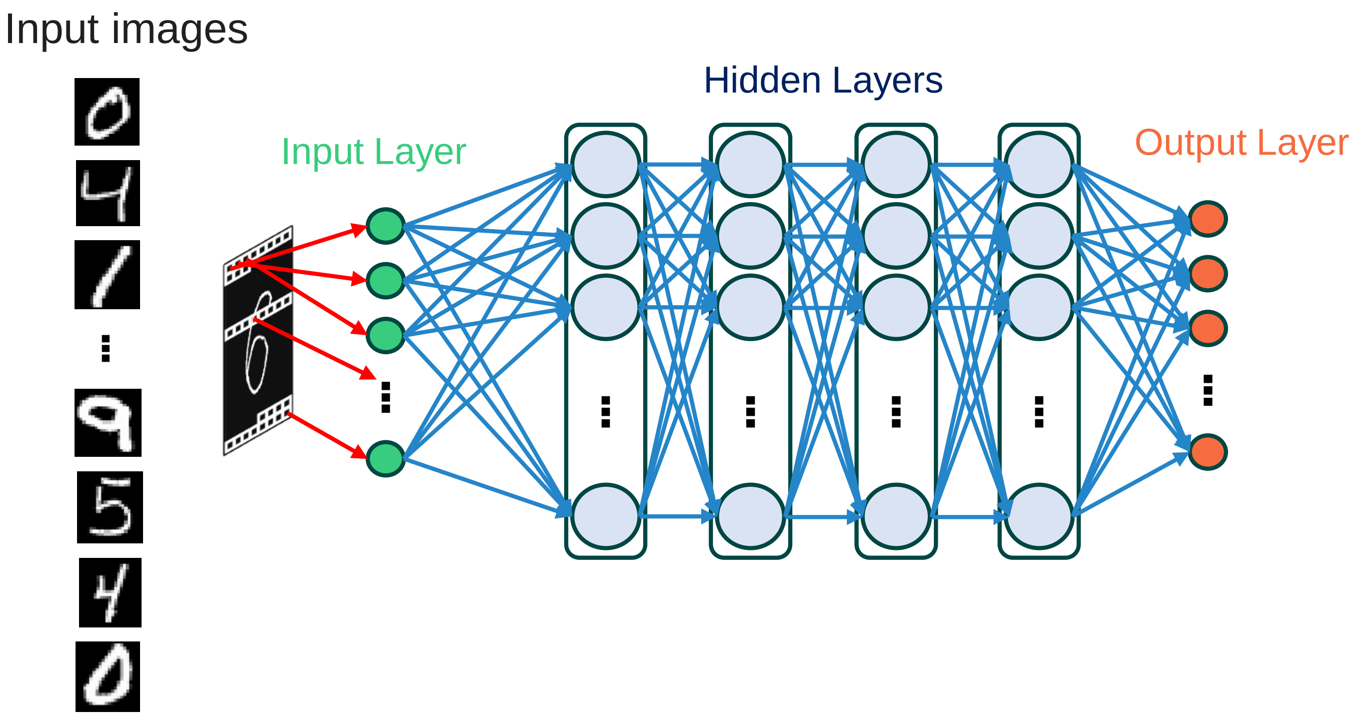
Prediction	Target
“9”	“0”
“0”	“4”
“1”	“1”
⋮	⋮
“8”	“9”
“6”	“5”
“3”	“4”
“2”	"0”

$$L = \frac{1}{N}\sum_{i=1}^{N}\left(y_{\text{pred}} - y_{\text{target}}\right)^2$$

N : number of elements in each vector

HOW DOES A NEURAL NETWORK LEARN?

Example : MNIST Dataset



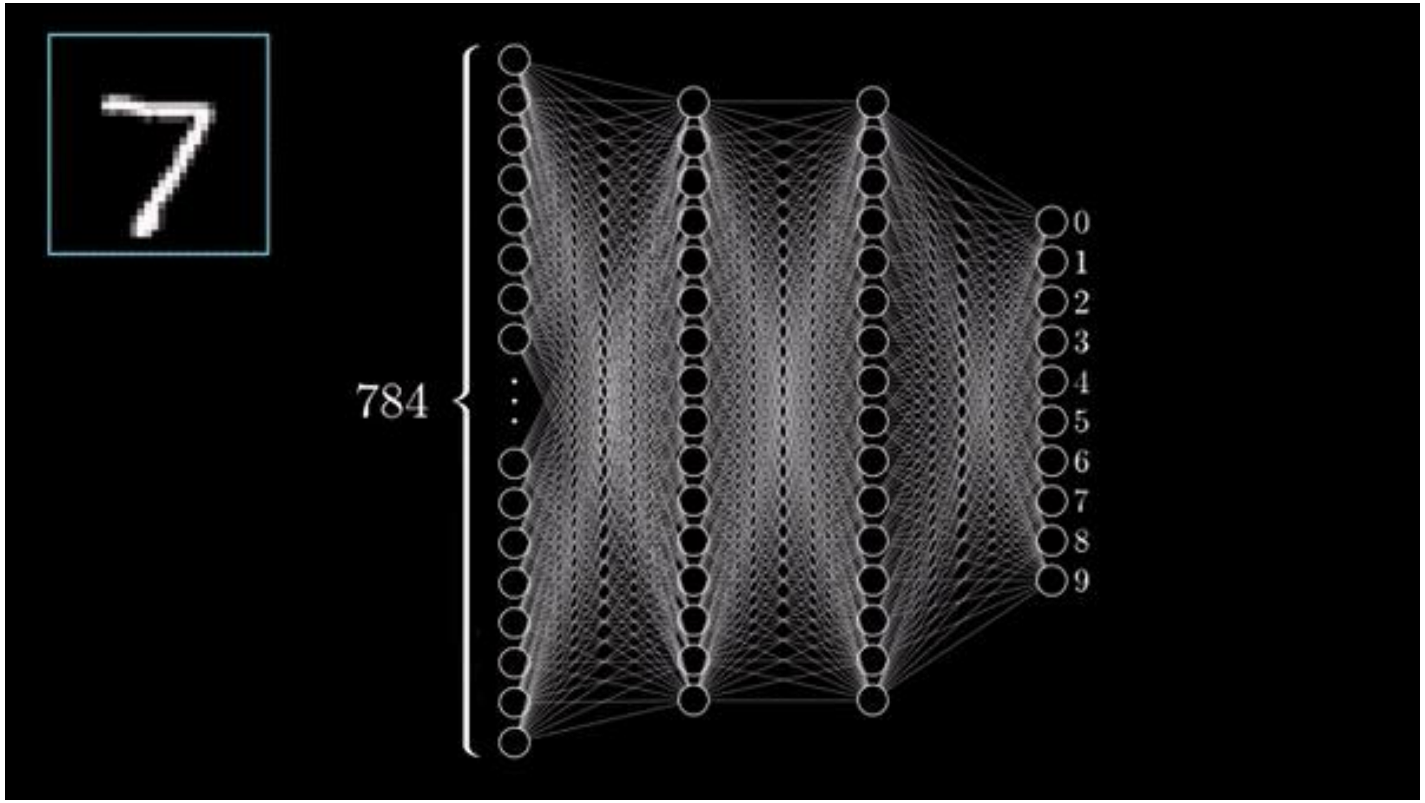
Prediction	Labels
“0”	“0”
“4”	“4”
“1”	“1”
⋮	⋮
“9”	“9”
“6”	“5”
“4”	“4”
“0”	"0”

To ensure generalization the NN is tested on unseen data.

NN structure embodies the task we want to perform.

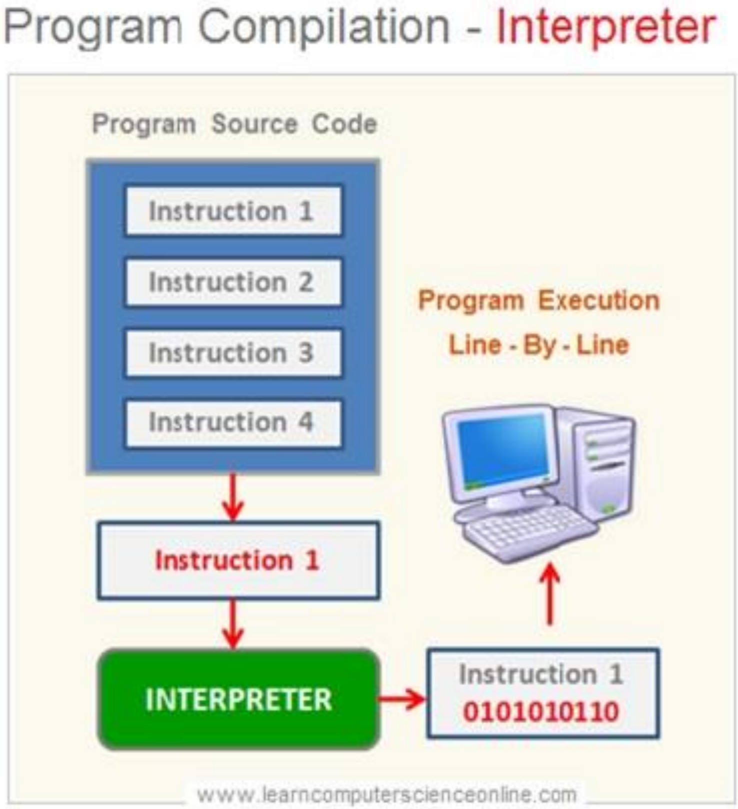
Parallel Computing = propagation of information through the NN.

NEURAL NETWORKS : PARALLEL COMPUTING



3Blue1Brown: https://www.youtube.com/watch?v=IHZwWFHWa-w&t=1087s

≠



NNs are an inherently **connectionist approach to computing.**

NNs are **universal function approximators**, with a NN that is **big enough and lots of data you can fit anything.**

Analog computing

EARLY ANALOG COMPUTING DEVICES

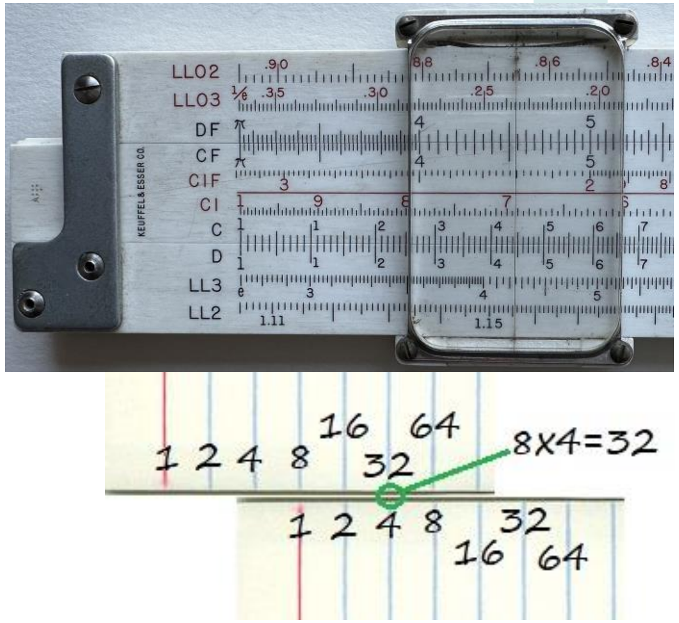
Analog

Process **continuous** values using physical quantities (voltages, currents, rotations, etc.)

Tracks Sun and Moon, predicts solar and lunar eclipses, shows the phases of the Moon, follows the calendar (including leap years).

Astrolabe

Telling time, finding position, position of stars at any given time.



Slide rule

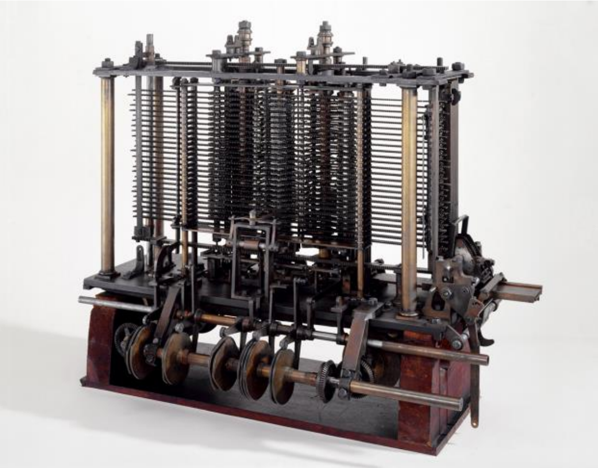
EARLY DIGITAL COMPUTING DEVICES

Digital

Process **discrete** values typically binary (0s and 1s).

Abaccus

Pascaline



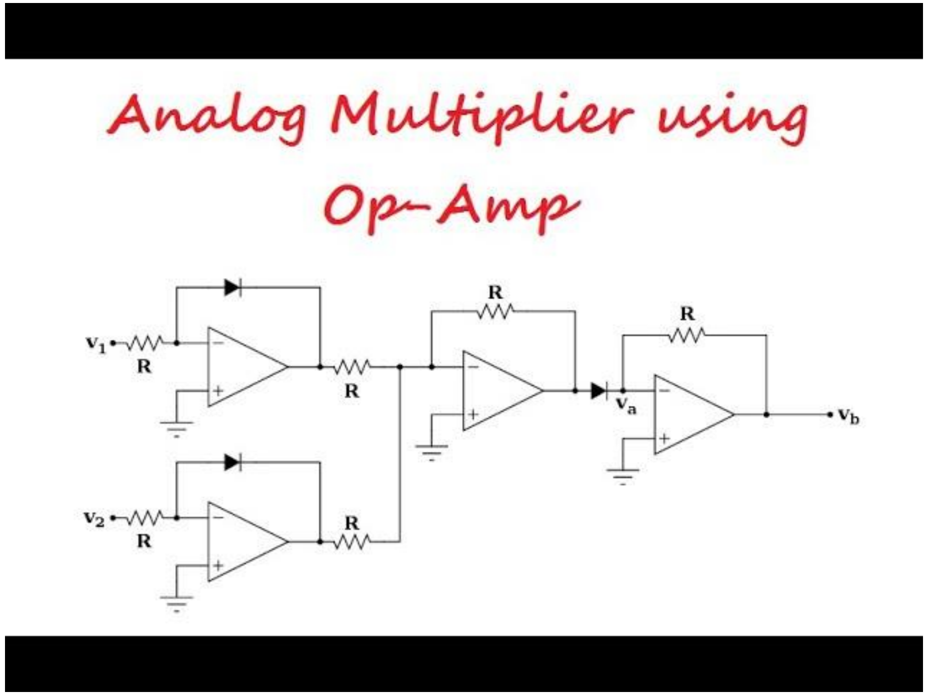
Early digital computers are mostly calculators.

Babbage analyzer

Analog

Process **continuous** values using physical quantities (voltages, currents, rotations, etc.)

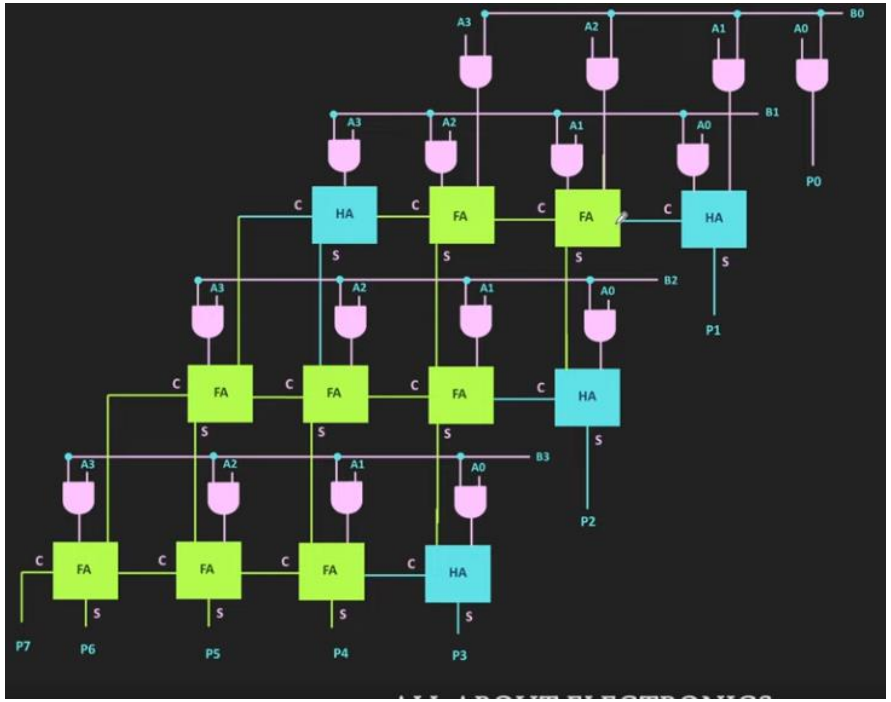
Analog:



Elevate electronics with Neha @youtube

Roughly 100 transistors

Digital (ex 4 bit):



All about electronics @youtube

Roughly 2000 transistors for 8-bit only

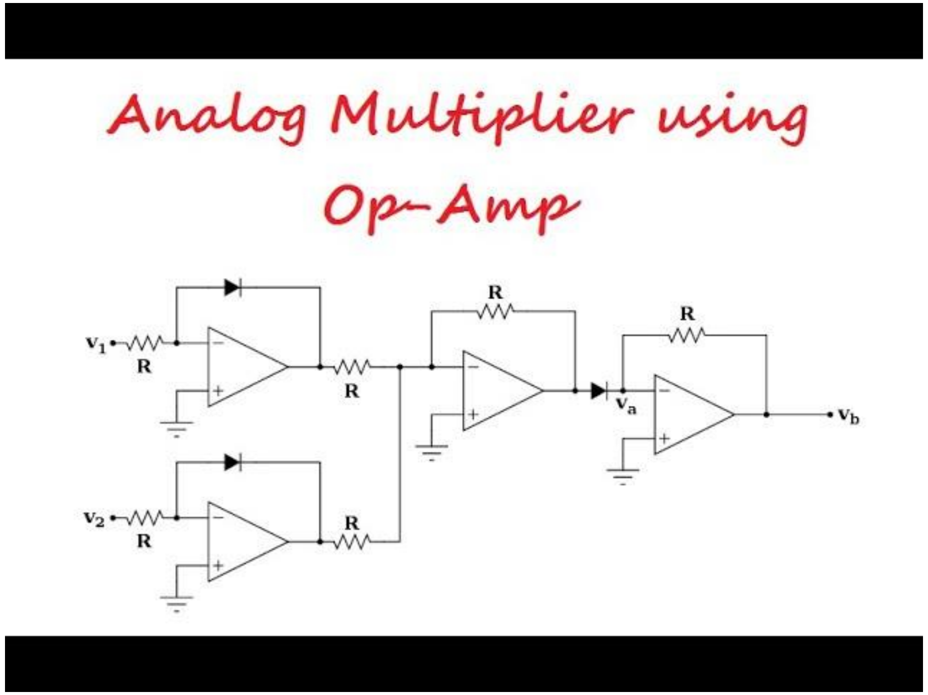
ANALOG SUPREMACY?

How do you multiply 2 numbers in a simpler way using electronics?

Analog

Process **continuous** values using physical quantities (voltages, currents, rotations, etc.)

Analog:

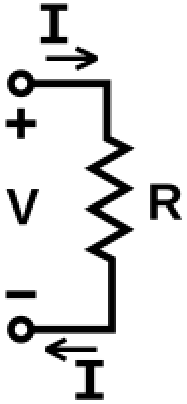


Elevate electronics with Neha @youtube

Roughly 100 transistors

Much simpler way of multiplying 2 numbers (we don’t especially care to encode them specifically in voltages):

Ohm’s Law:

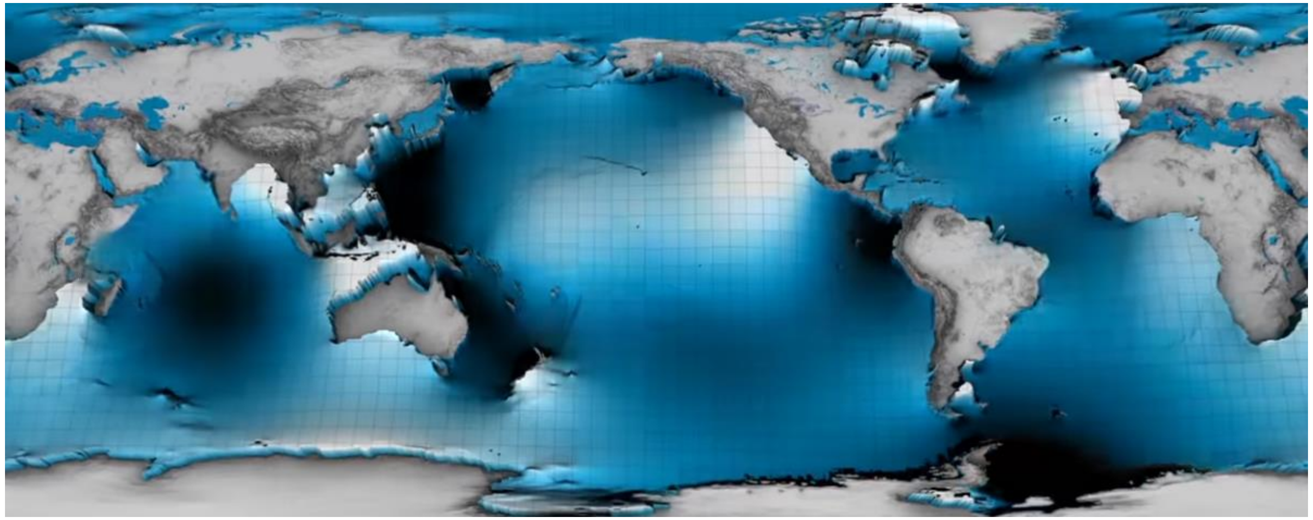


@Wikipedia

1. Choose a resistor
2. Apply a voltage
3. Measure current is I = V/R = V*G
4. Variable resistor and power supply!

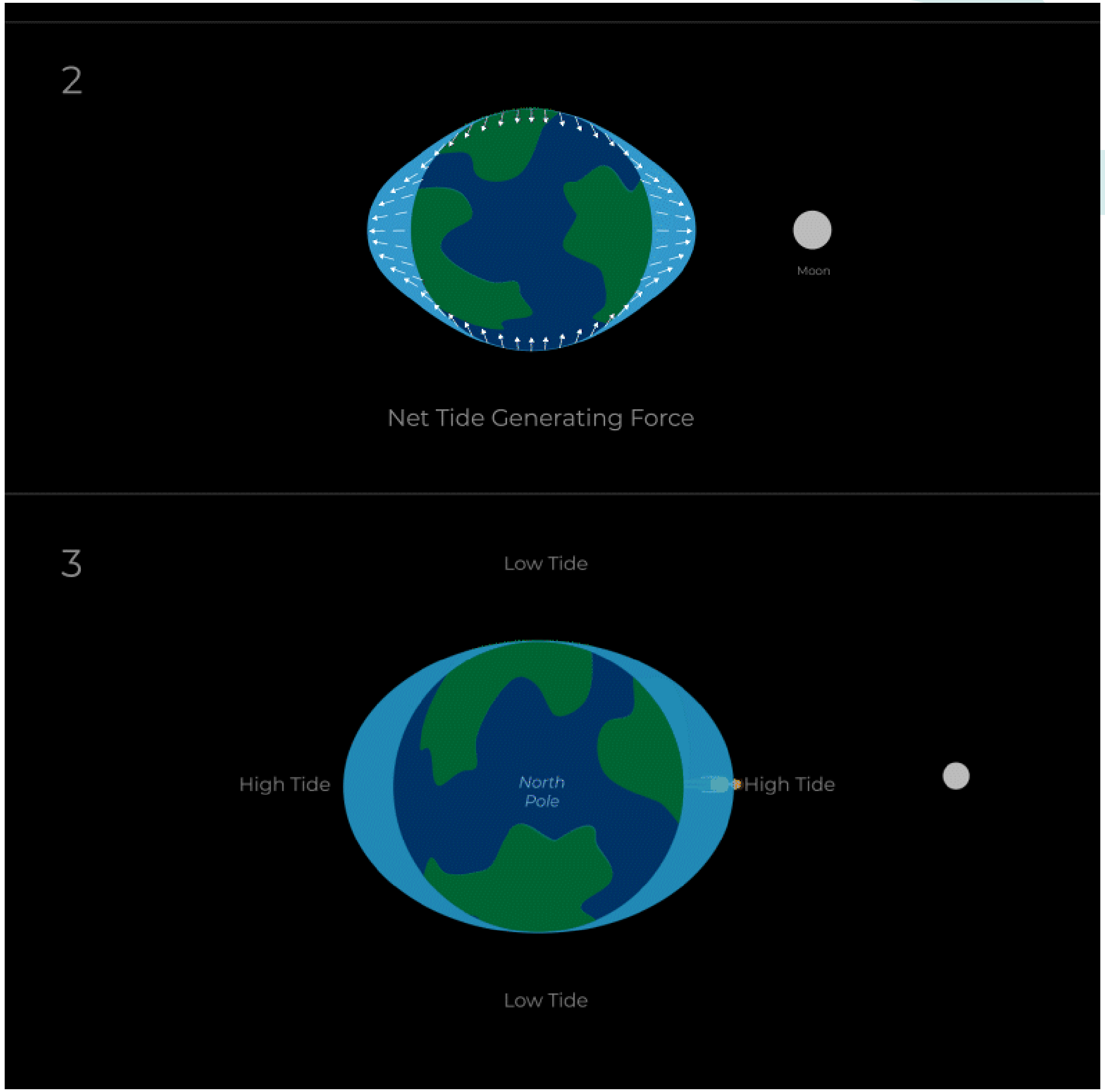
Problem : Predicting the tides

Timing as well as magnitude variation



@Veritasium youtube

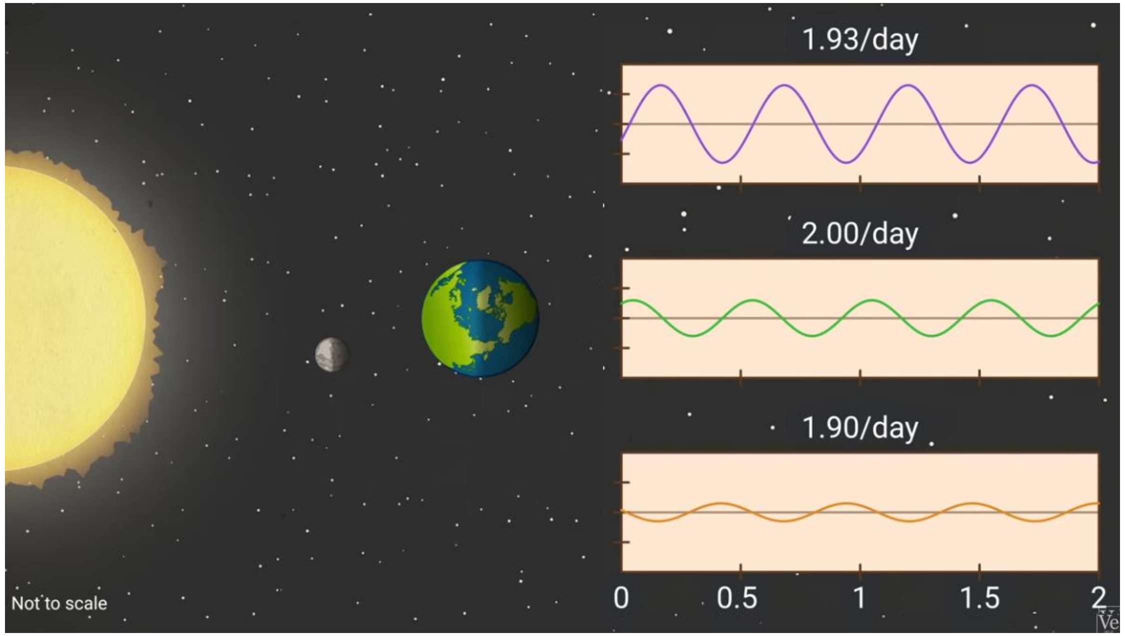
Problem : Predicting the tides

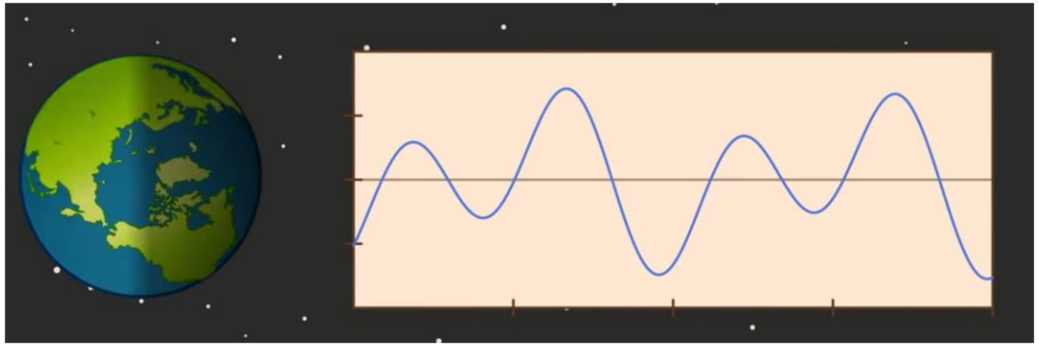


Laplace found that tides are driven by a few specific astronomical frequencies

Problem : Predicting the tides

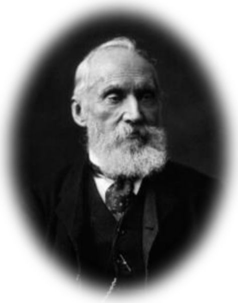
Laplace found that tides are driven by a few specific astronomical frequencies:



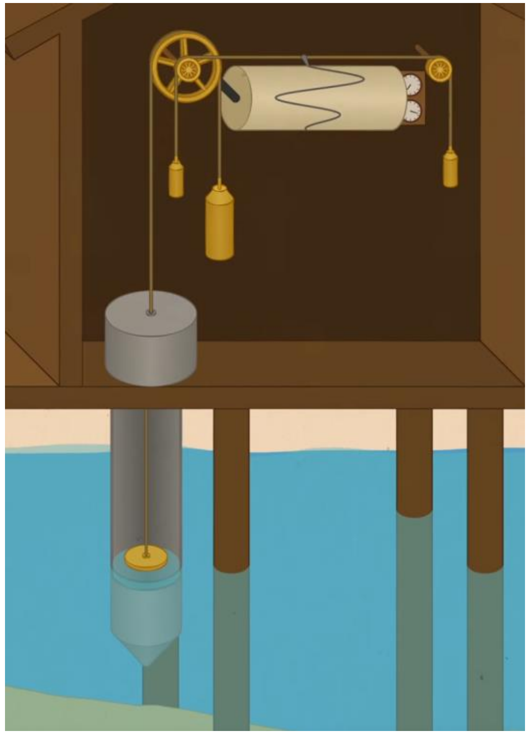


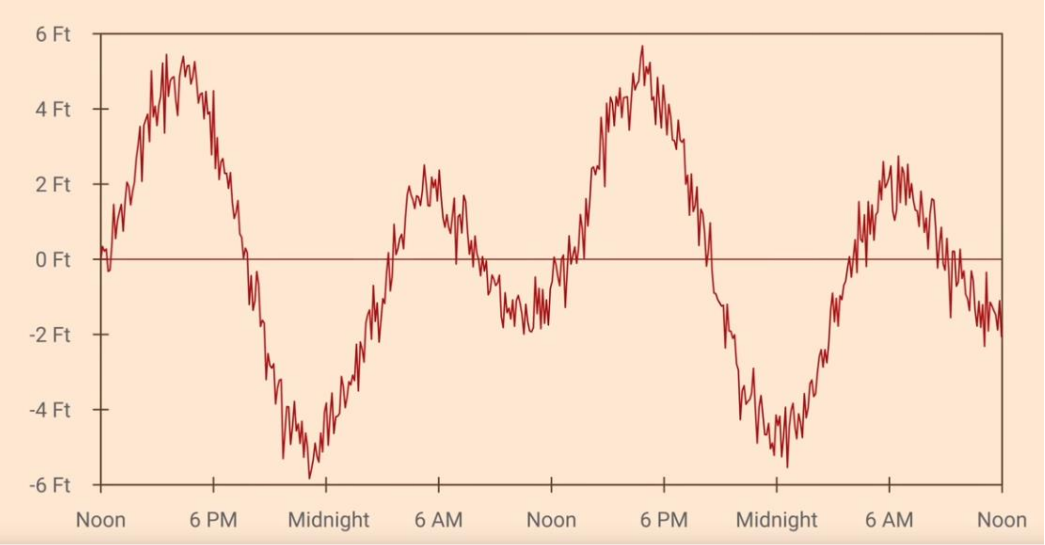
@Veritasium youtube

Problem : Predicting the tides



First: record the tide using a tide gauge.



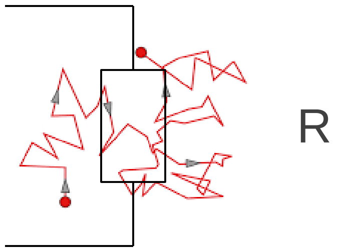


@Veritasium youtube

But, why digital was successful: physical limit to precision

Switching an electronic signal line:

- Charging wire: $\mathrm{E} = \mathrm{CV}^2$
  - C/m = 1.5 pF/cm
- Mean square thermal noise: $\mathrm{V}_N = \sqrt{4k_bTBR}$
  - $k_b = 1.38 \times 10^{-23}$ J/K, B=1 GHz, R=1 MΩ, T=300 K
  - ➢ V_N=4 mV

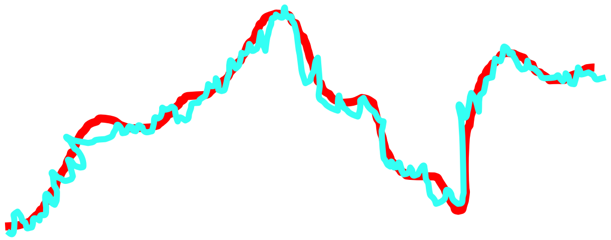


Noise in digital signal:

- Digital sequence:
  - $\mathrm{D_N}\mathrm{D}_{N-1}\cdots\mathrm{D}_1\mathrm{D}_0$
  - Noise will corrupt each of the digits with equal probability
    - $\overline{D}_{\mathrm{N}}\mathrm{D}_{N-1}\cdots\mathrm{D}_1\mathrm{D}_0$
  - ➢ Noise exceeding digitization threshold is unacceptable

Noise in analog signal:

- Signal comprises noise less amplitude
- Noise is additive or multiplicative
- Always of scale unity for an normalized signal strength



Physical limit to precision: Energy / Power

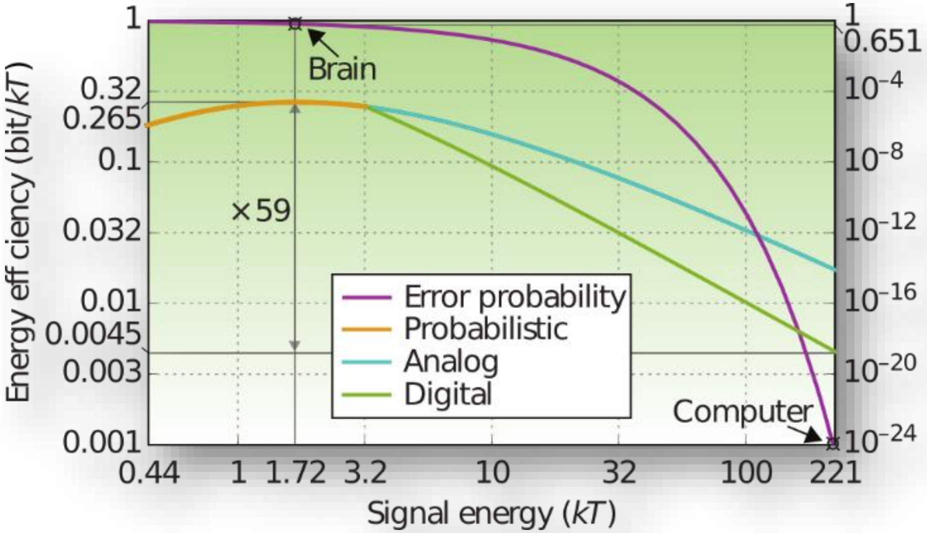
Thermodynamic limits:

- Low pass circuit: $\mathrm{SNR} = \sqrt{E/kT}$
- $kT(@\mathrm{RT}) = 4 \cdot 10^{-21}\ \mathrm{J}$
- For 8 bit: $\mathrm{E} = 255^2 \cdot kT \approx 0.3\ \mathrm{fJ}$

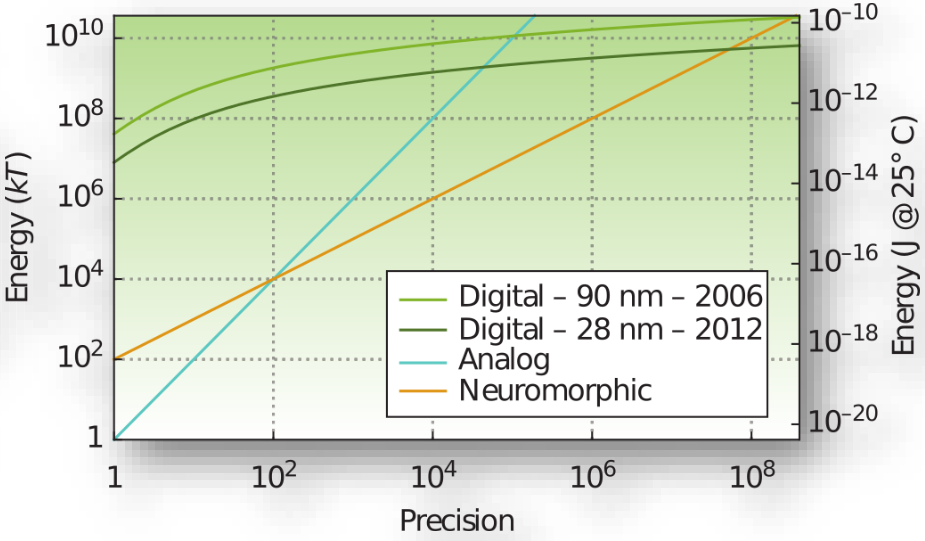
Linked to devices:

- Limit for signal @1 GHz: $\mathrm{P} \approx 0.3\ \mu\mathrm{W}$
- Detection: $\mathrm{NEP} \approx 0.4 \frac{pW}{\sqrt{Hz}}, @1\ \mathrm{GHz} \approx 10\ \mathrm{nW}$
- ➢ For 1 GHz: $\mathrm{PEN} \cdot 255 \approx 3\ \mu\mathrm{W}$

➢ **ONLY x10**

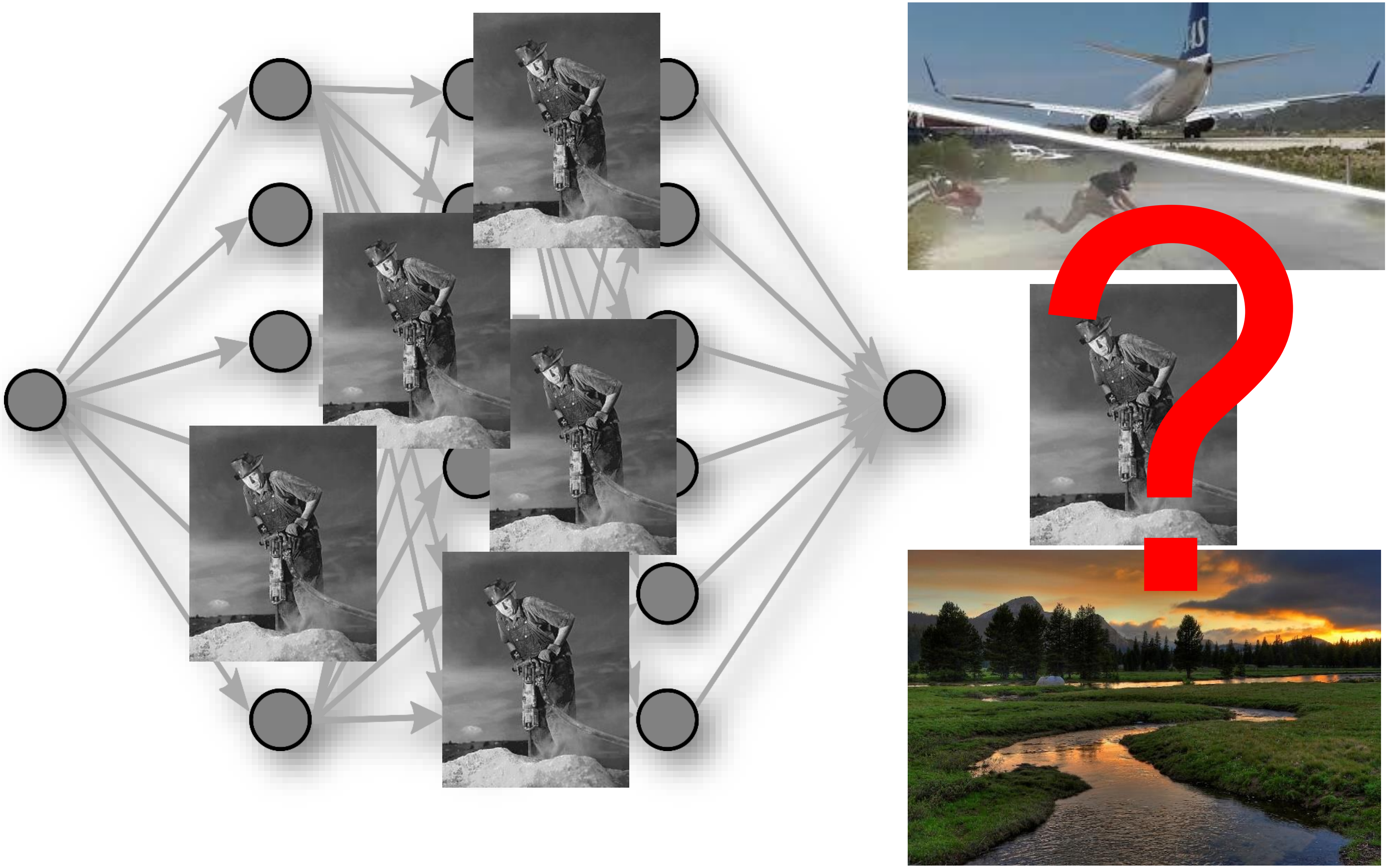


Boahen, Computing in Science & Engineering 2017.

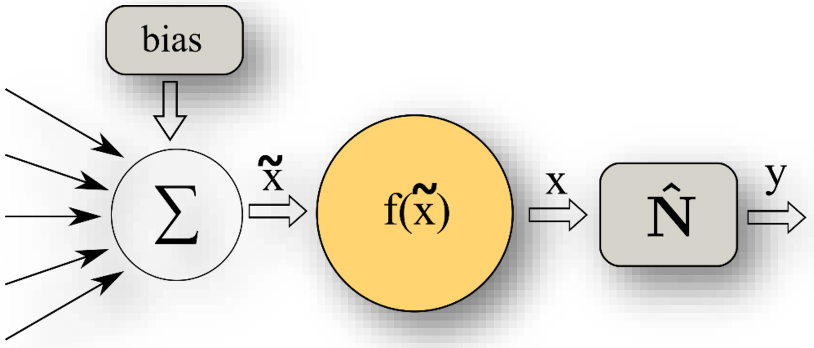


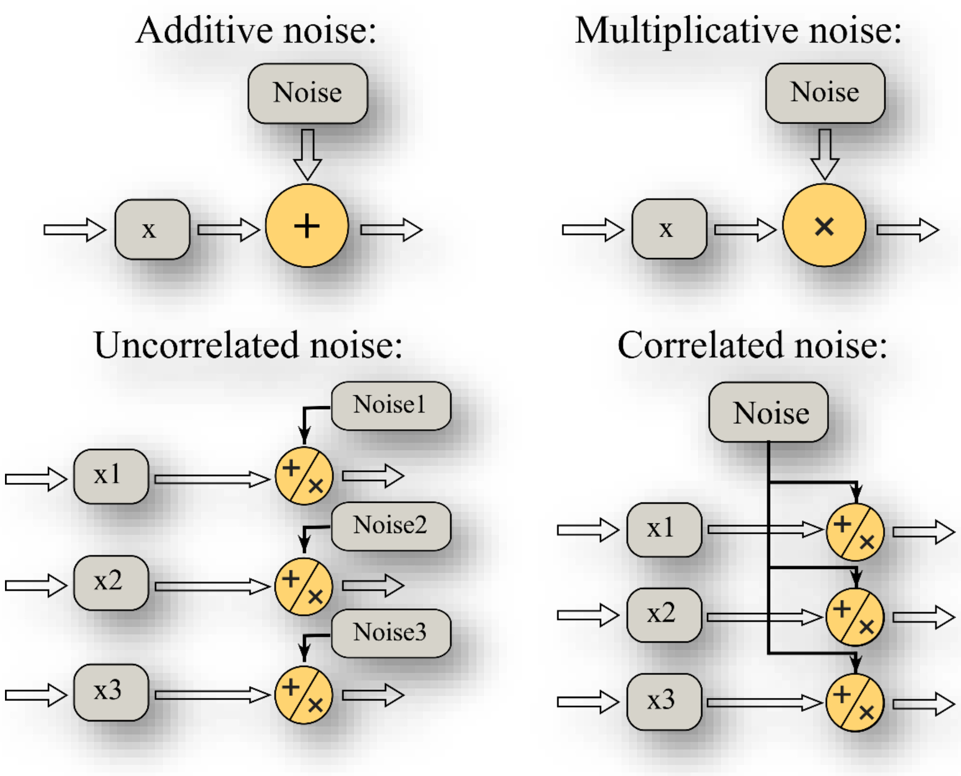
- ○ Approach SNR limit within an order of magnitude
- ○ For low precision analog is clearly superior
- ○ Digital NN hardware: approaching few-bit precision(!!!???)

Analog NNs: amplify, unitary or damp noise?



Noisy neurons inside networks



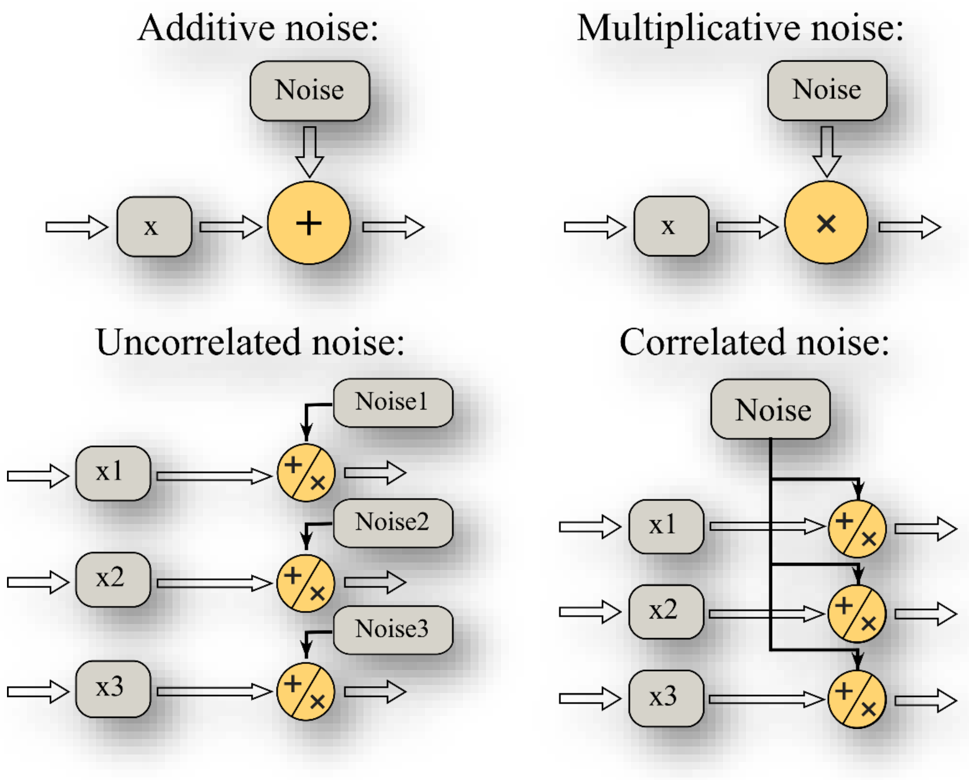


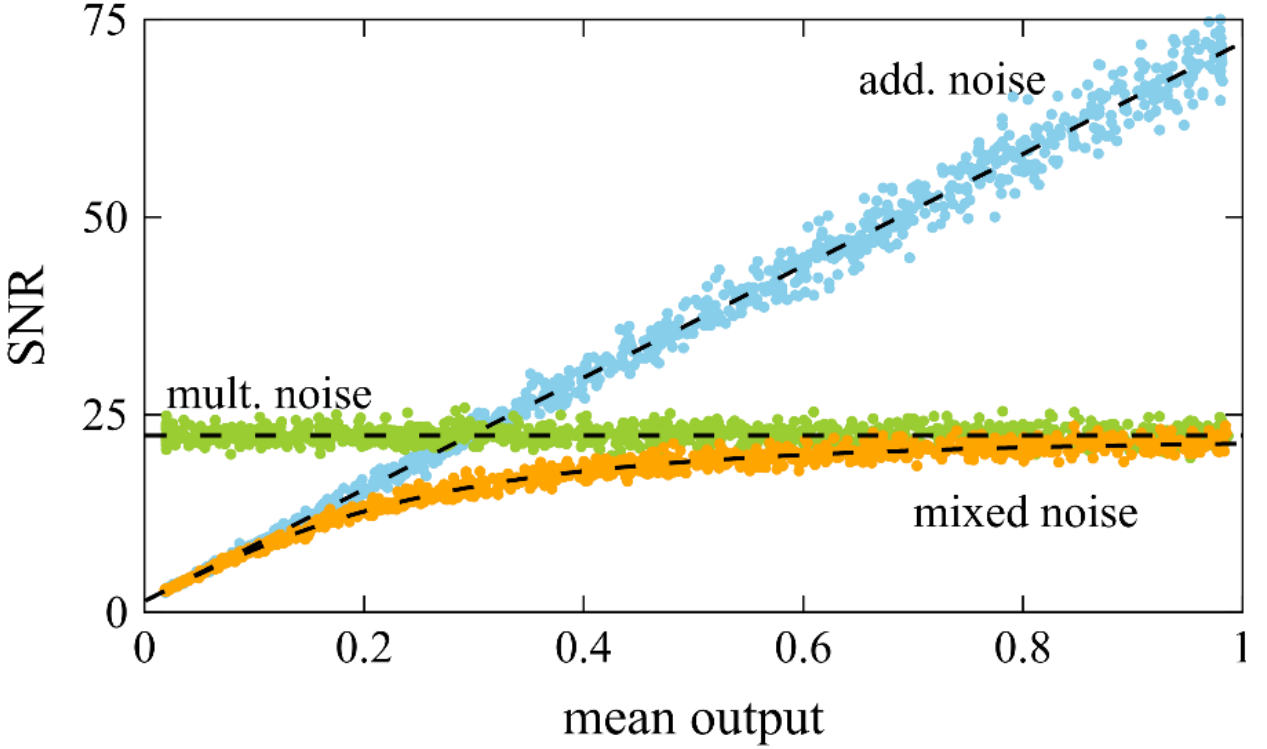
Signal to noise ratio $\mathrm{SNR} = \dfrac{\mathrm{E}(y_{n,i}^{t})}{\left(\mathrm{Var}(y_{n,i}^{t})\right)^{1/2}}$

Semenova , et al., Chaos 29, 103128 (2019).

Semenova, Larger, Brunner, Neural Networks 146, 151 (2022).

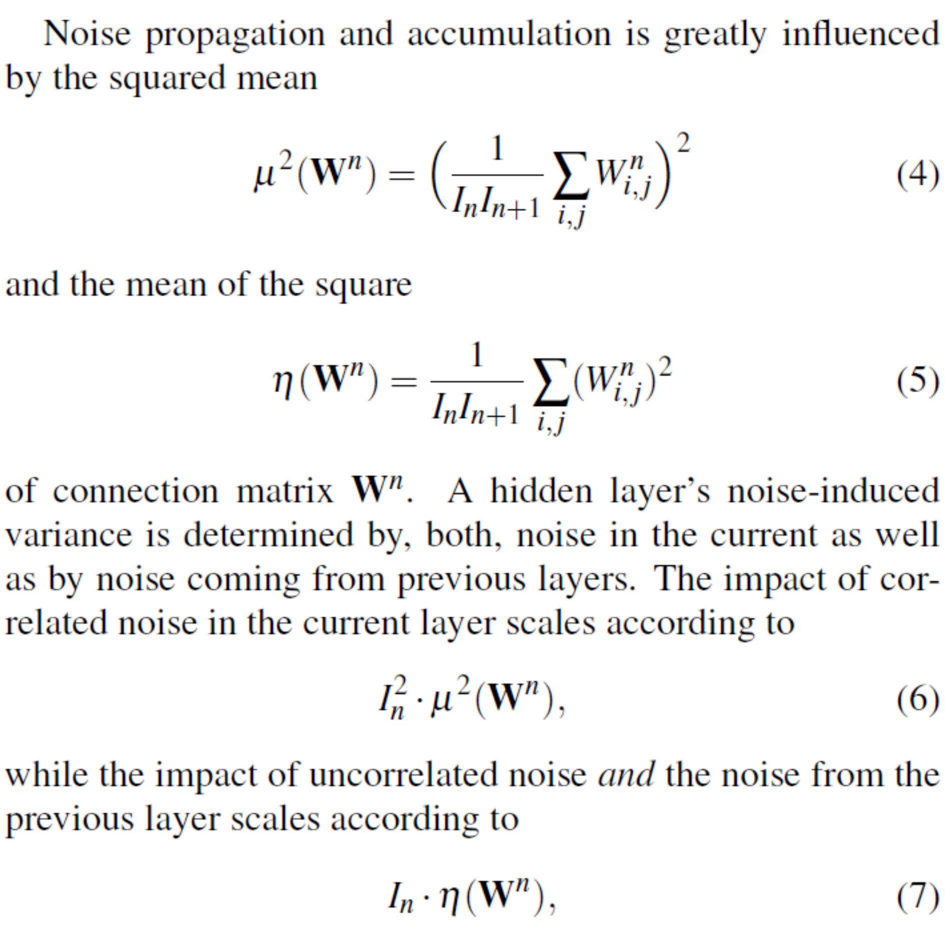
Single neuron: linear AND nonlinear





Semenova , et al., Chaos 29, 103128 (2019).

What influences noise propagation?



Noise propagation and accumulation is greatly influenced by the squared mean

$$\mu^2(\mathbf{W}^n) = \left(\frac{1}{I_n I_{n+1}} \sum_{i,j} W^n_{i,j}\right)^2 \qquad (4)$$

and the mean of the square

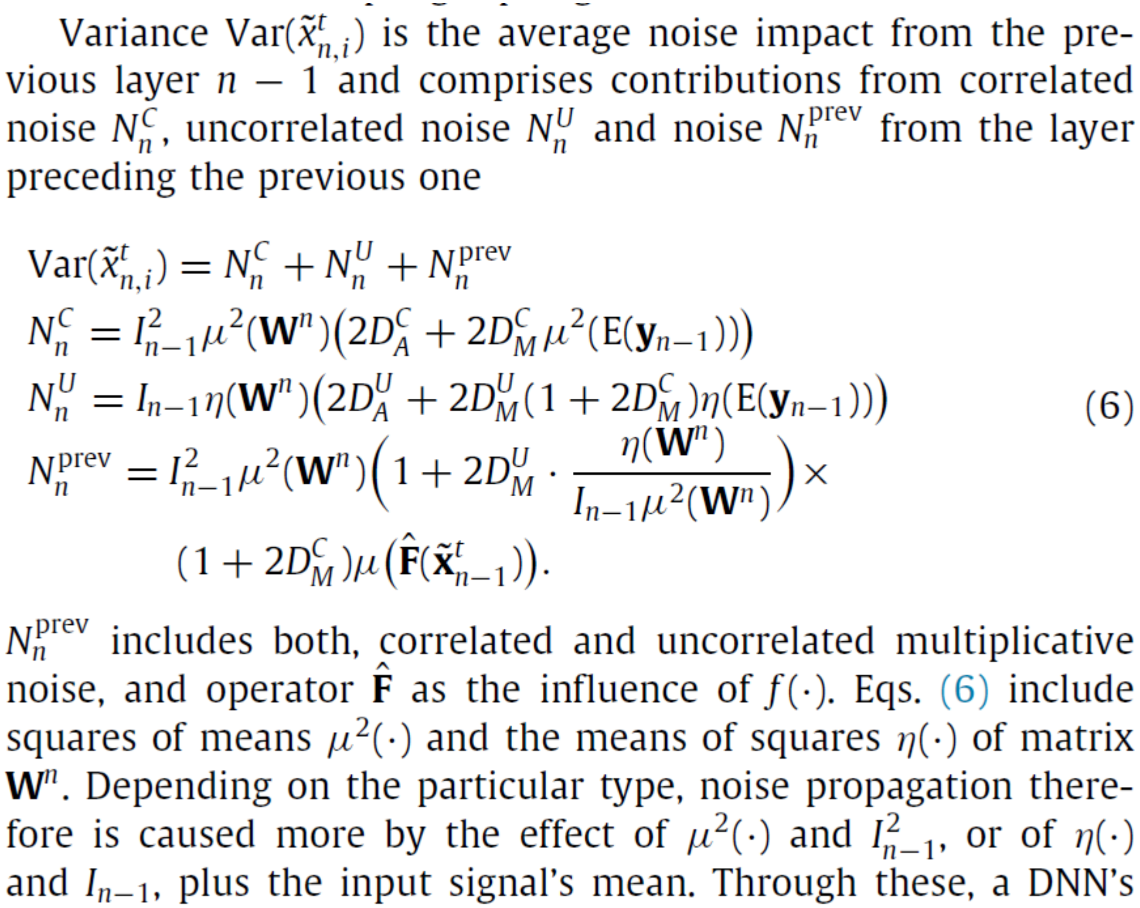
$$\eta(\mathbf{W}^n) = \frac{1}{I_n I_{n+1}} \sum_{i,j} (W^n_{i,j})^2 \qquad (5)$$

of connection matrix $\mathbf{W}^n$. A hidden layer's noise-induced variance is determined by, both, noise in the current as well as by noise coming from previous layers. The impact of correlated noise in the current layer scales according to

$$I_n^2 \cdot \mu^2(\mathbf{W}^n), \qquad (6)$$

while the impact of uncorrelated noise *and* the noise from the previous layer scales according to

$$I_n \cdot \eta(\mathbf{W}^n), \qquad (7)$$



Variance $\mathrm{Var}(\tilde{x}^t_{n,i})$ is the average noise impact from the previous layer $n-1$ and comprises contributions from correlated noise N^C_n, uncorrelated noise N^U_n and noise N^{prev}_n from the layer preceding the previous one

$$\begin{aligned}
\mathrm{Var}(\tilde{x}^t_{n,i}) &= N^C_n + N^U_n + N^{\mathrm{prev}}_n \\
N^C_n &= I^2_{n-1}\mu^2(\mathbf{W}^n)\big(2D^C_A + 2D^C_M\mu^2(\mathrm{E}(\mathbf{y}_{n-1}))\big) \\
N^U_n &= I_{n-1}\eta(\mathbf{W}^n)\big(2D^U_A + 2D^U_M(1+2D^C_M)\eta(\mathrm{E}(\mathbf{y}_{n-1}))\big) \\
N^{\mathrm{prev}}_n &= I^2_{n-1}\mu^2(\mathbf{W}^n)\left(1 + 2D^U_M \cdot \frac{\eta(\mathbf{W}^n)}{I_{n-1}\mu^2(\mathbf{W}^n)}\right) \times \\
&\qquad (1+2D^C_M)\mu\big(\hat{\mathbf{F}}(\tilde{\mathbf{x}}^t_{n-1})\big).
\end{aligned} \qquad (6)$$

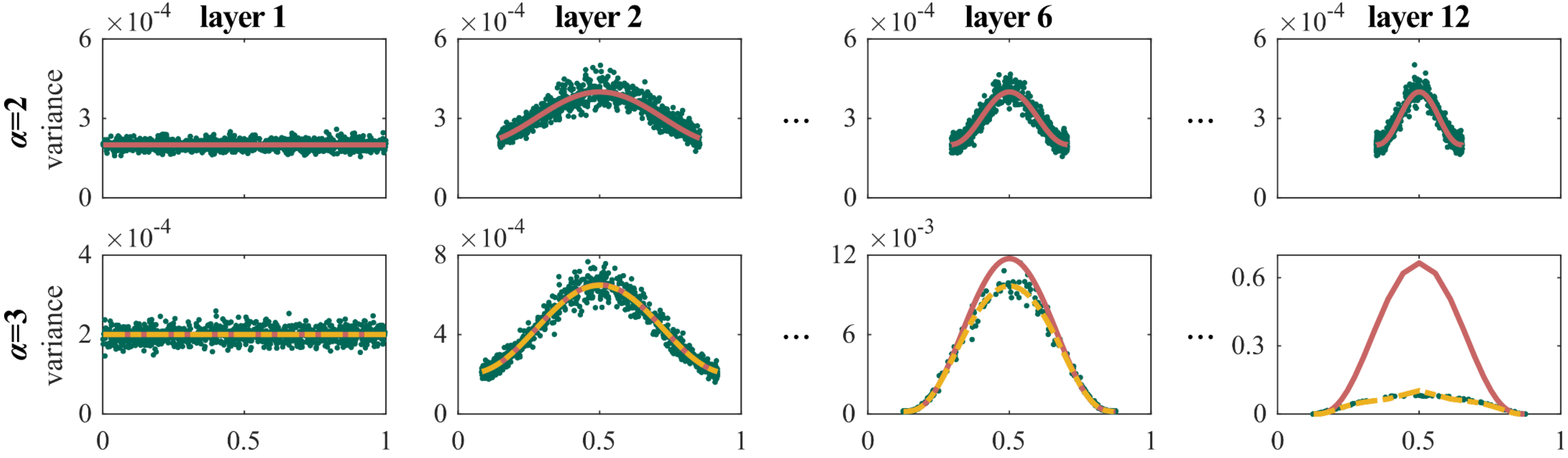
N^{prev}_n includes both, correlated and uncorrelated multiplicative noise, and operator $\hat{\mathbf{F}}$ as the influence of $f(\cdot)$. Eqs. (6) include squares of means $\mu^2(\cdot)$ and the means of squares $\eta(\cdot)$ of matrix $\mathbf{W}^n$. Depending on the particular type, noise propagation therefore is caused more by the effect of $\mu^2(\cdot)$ and I^2_{n-1}, or of $\eta(\cdot)$ and I_{n-1}, plus the input signal's mean. Through these, a DNN's

Noise operator for cascaded layers n :

$$S_n = f\left(E\left(\boxed{f^{-1}(y^t_{n-1})}\right)\right) S_{n-1}$$

Noise accumulation

$$f(x) = 0.5\frac{\alpha(x-0.5)}{\sqrt{1+\alpha^2(x-0.5)^2}}$$

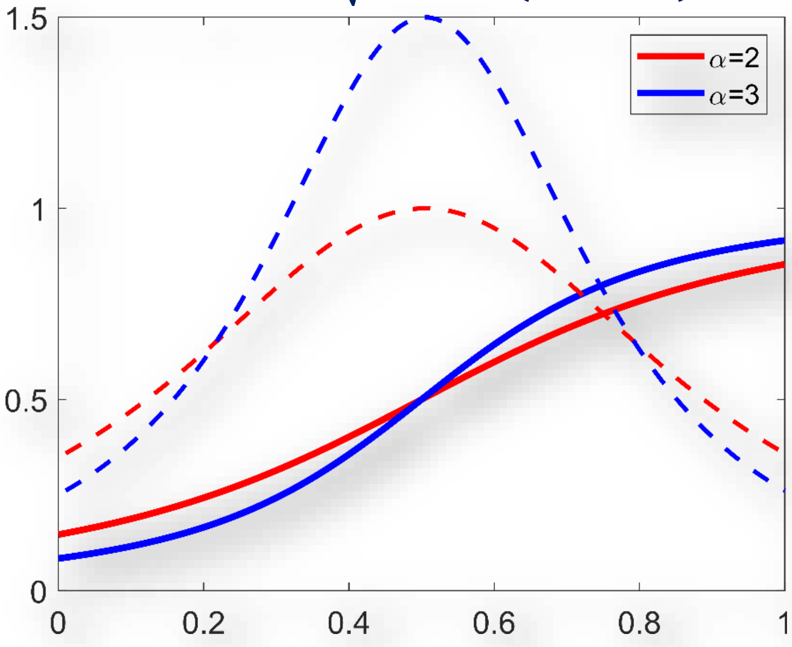


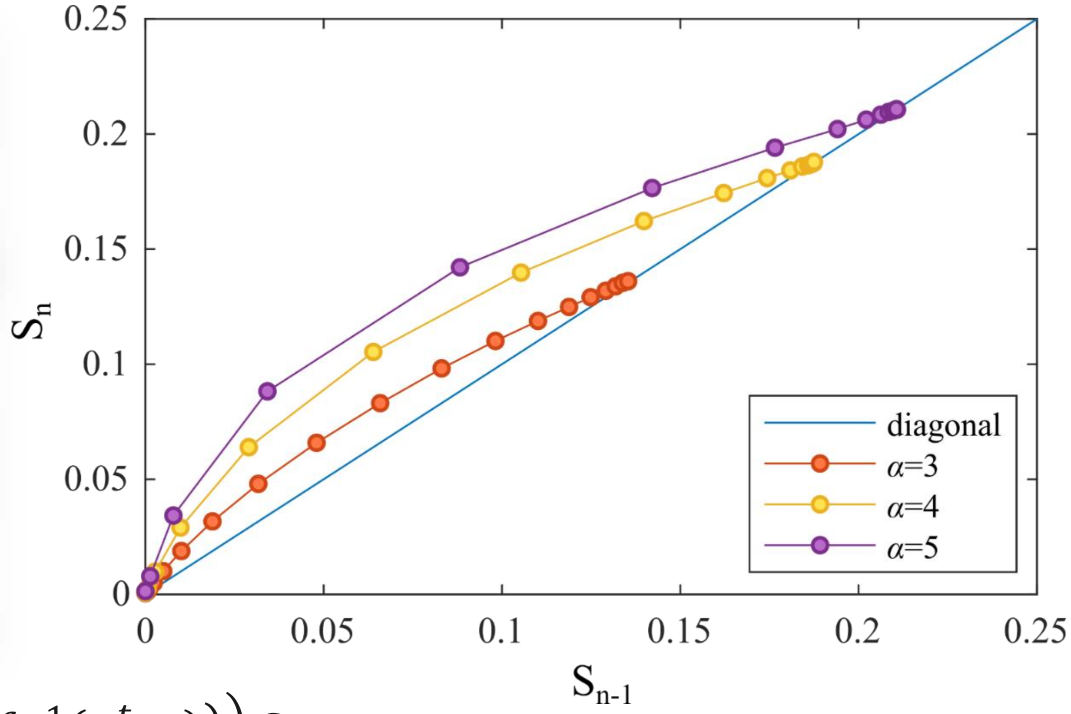
- ○ Noise accumulates only for $\alpha > 2$
- ○ For $\alpha \leq 2$ output noise on level of individual neurons
- ○ For strong nonlinearity: first order approximation fails

Semenova, Larger, Brunner, Neural Networks 146, 151 (2022).

Noise accumulation

$$f(x) = 0.5\frac{\alpha(x-0.5)}{\sqrt{1+\alpha^2(x-0.5)^2}}$$





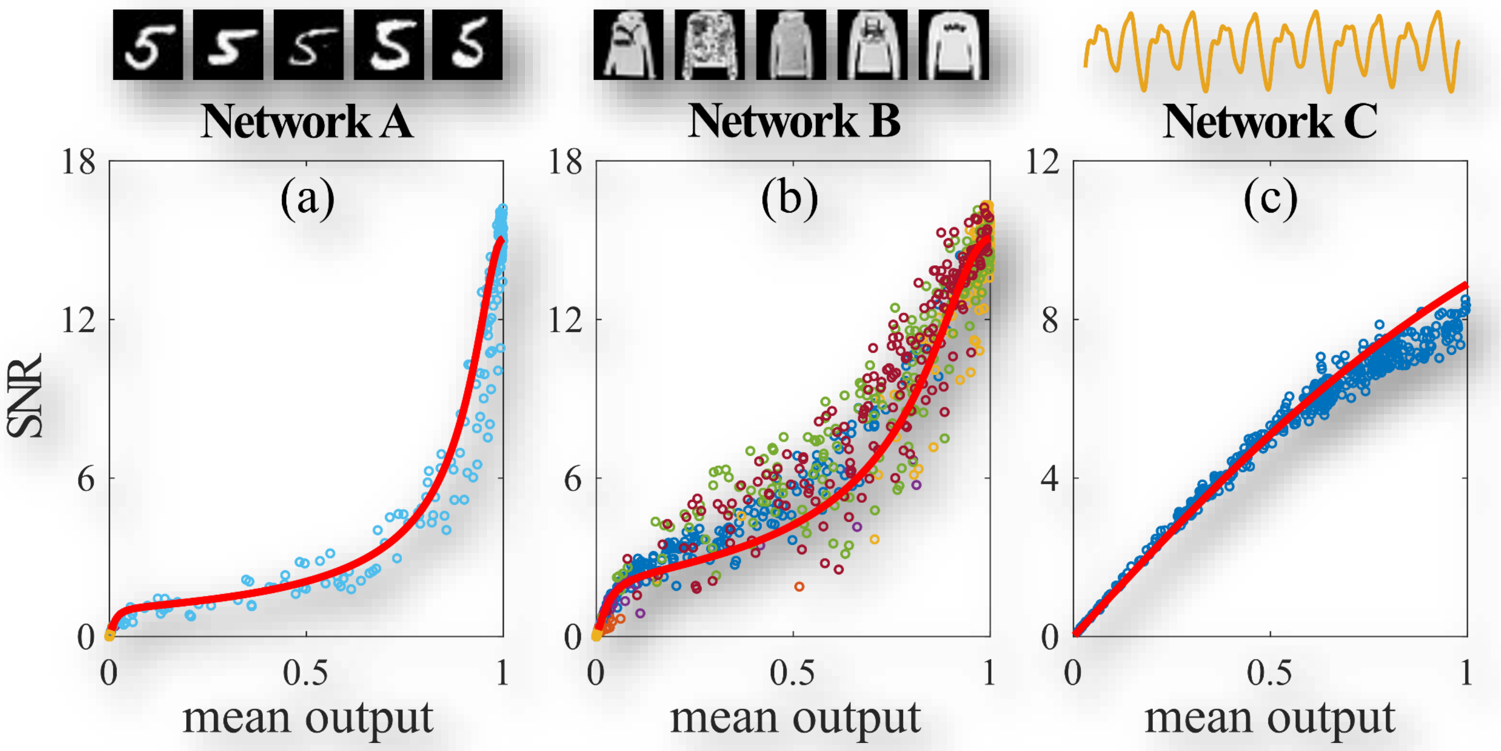
$$S_n = f\left(E\left(f^{-1}(y_{n-1}^t)\right)\right)S_{n-1}$$

This is a geometric progression: **decreasing for** $f^{-1}(\cdot) < 1$

- Noise accumulates **generally** bound
- Can be **frozen** for$f^{-1}(\cdot) < 1$

Semenova, Larger, Brunner, Neural Networks 146, 151 (2022).

And what with fully trained DNN?



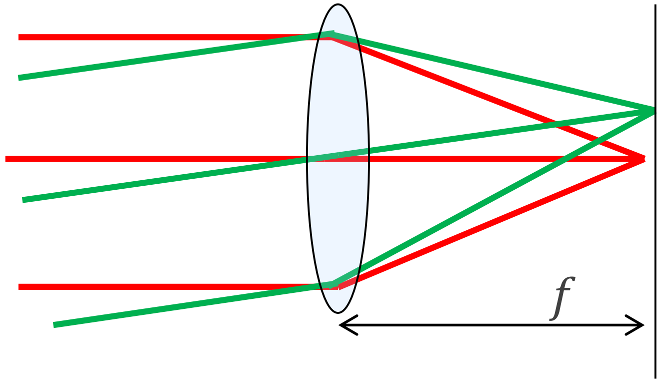
- Trained networks: weight AND state statistics required
- Approximation state statistics: $g(x) = \exp(c_4 x^4 + c_3 x^3 + c_2 x^2 + c_1 x)$
- Works very well through all layers

Semenova, Larger, Brunner, Neural Networks 146, 151 (2022).

Optical computing

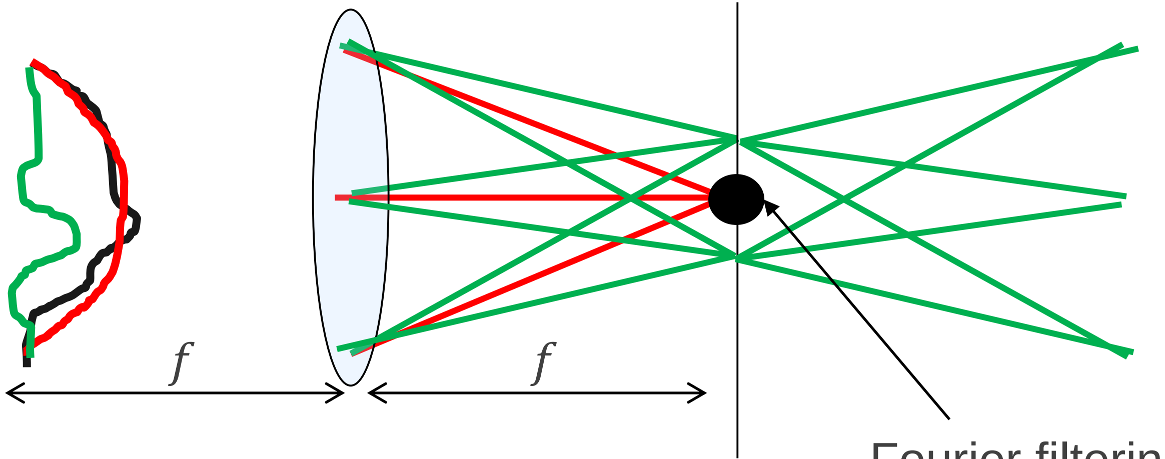
Classical (linear) optical computing

- Lens: focusing



- More precisely: Fourier transform
  - $E(x,y,f) = \frac{-if}{\lambda}\int_0^{\pi}\sin\theta\int_0^{2\pi}E(\theta,\phi)e^{i(k_z z-k_x x-k_y y)}d\theta d\phi$
  - $E(x,y,f) = \frac{-if}{\lambda}\mathrm{FT}\{\frac{E(\theta,\phi)e^{i(k_z z)}}{\cos\theta}\}$

Fourier (Spatial-frequency) plane

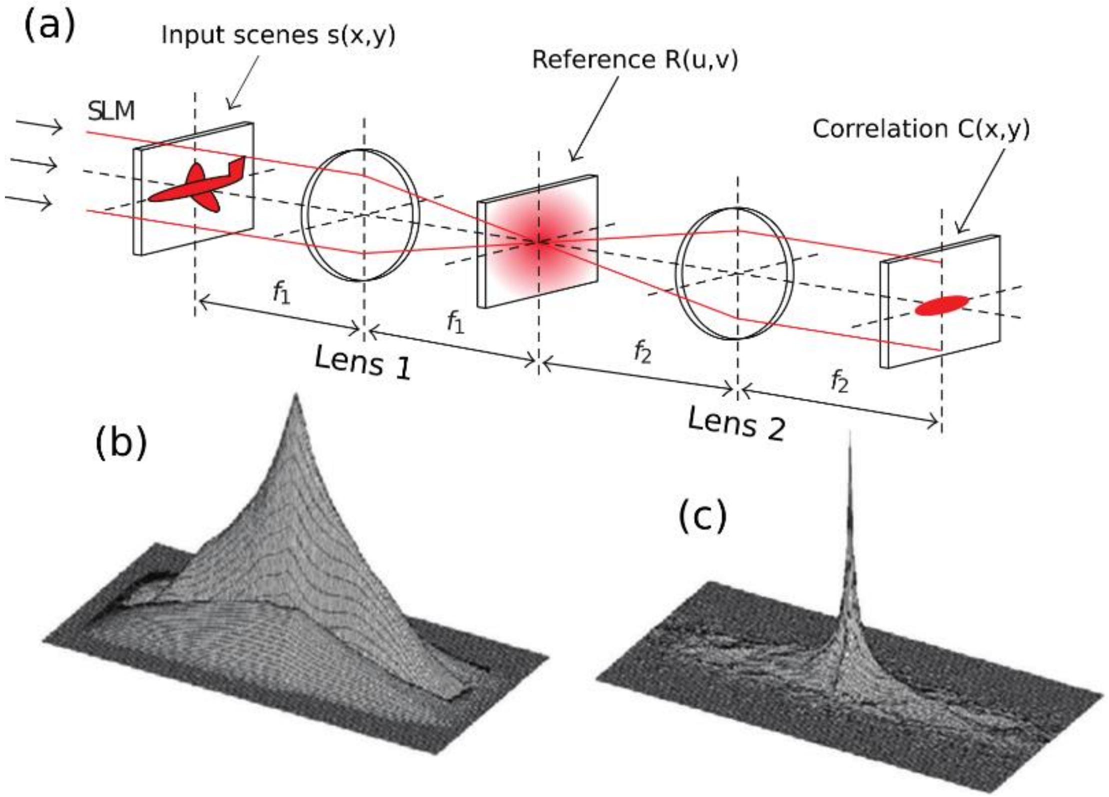


Object plane

Fourier filtering

P. Ambs, “*Optical Computing: a 60-year adventure*”, *Advances in Optical technologies (2010)*.

Classical (linear) optical computing



- Long history, well established field
- Passive, fully parallel
  - Ultra-high space/bandwidth product (PetaBit/s/cm^2)
- But: only linear
- Bulk optics: computer with a microscope objective?

P. Ambs, “*Optical Computing: a 60-year adventure”, Advances in Optical technologies (2010)*.

Experimental performance of a binary phase-only optical correlator using visual and infrared imagery

S.P. Kozaitis

Florida Institute of Technology, Department of Electrical and Computer Engineering
150 W. University Blvd., Melbourne, FL 32901-6988

S. Halby and W. Foor

Rome Air Development Center, Photonics Laboratory
Griffiss AFB, NY 13441

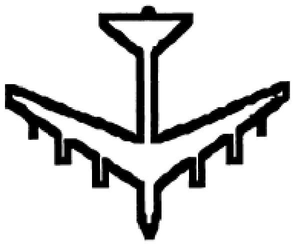


Fig.1 Sample of image in database

⊗

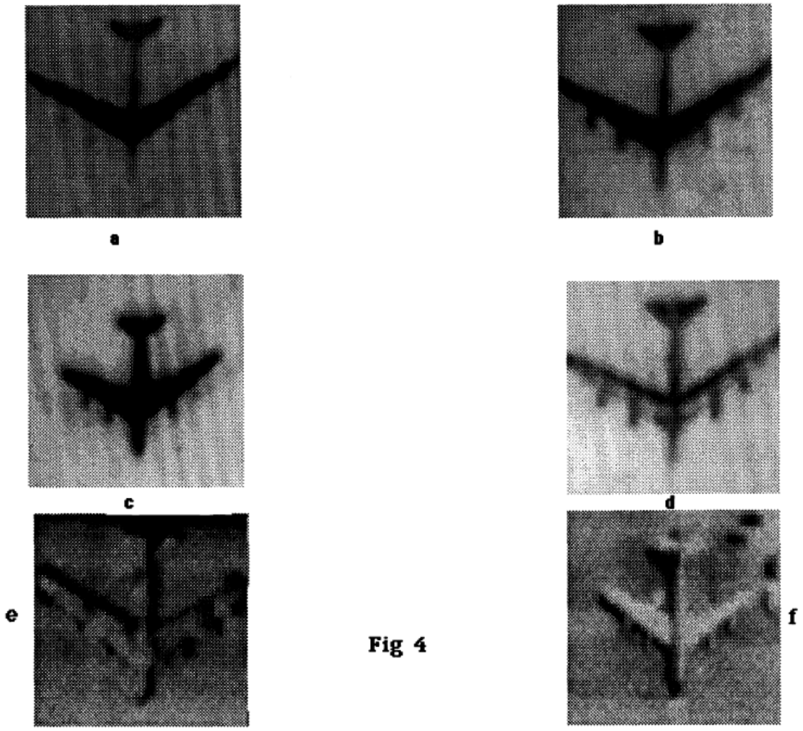


Fig 4

=

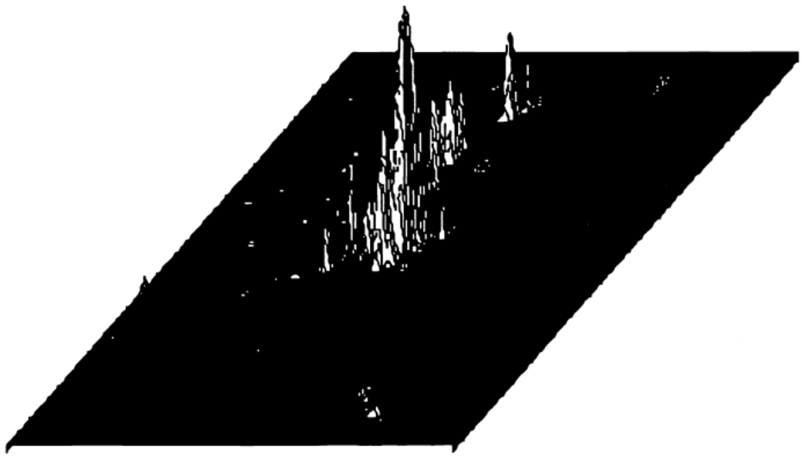


Fig. 5 Experimental correlation results of image in Fig. 4a and filter of Fig. 1

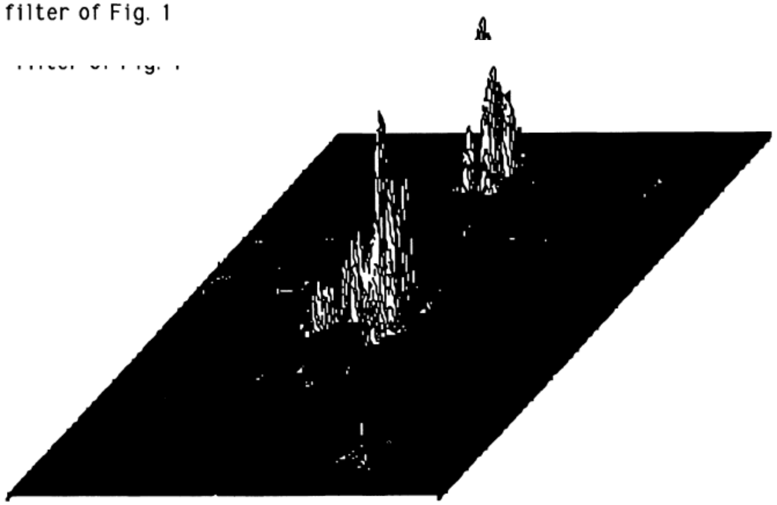
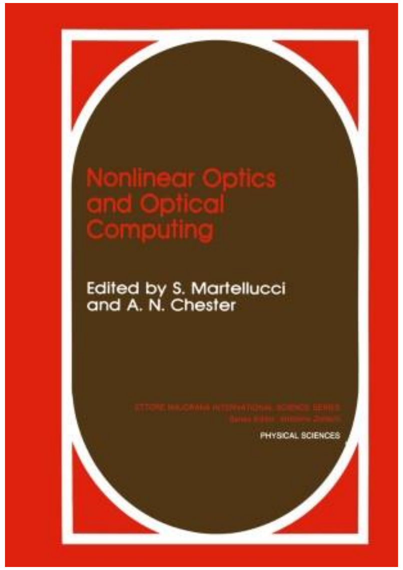


Fig. 6 Experimental correlation results of image in Fig. 4b and filter of Fig. 1

Principle and long-known motivation

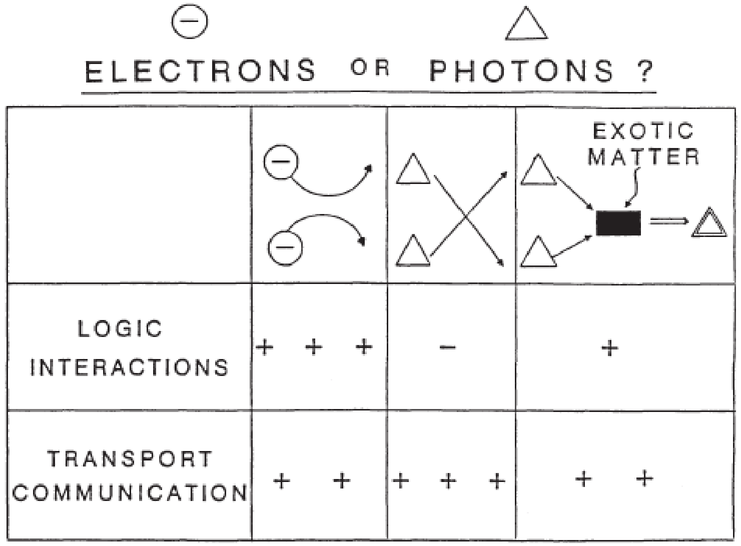

Nonlinear Optics and Optical Computing
Edited by S. Martellucci and A. N. Chester
PHYSICAL SCIENCES

S. Martellucci, A. N. Chester (1990).

PRINCIPLES OF OPTICAL COMPUTING

A.W. Lohmann

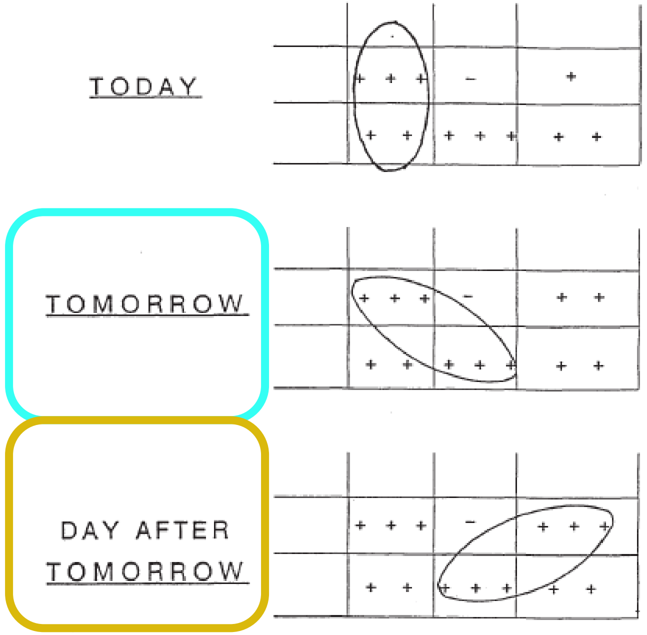
University of Erlangen



ELECTRONS OR PHOTONS ?

			EXOTIC MATTER
LOGIC INTERACTIONS	+ + +	−	+
TRANSPORT COMMUNICATION	+ +	+ + +	+ +

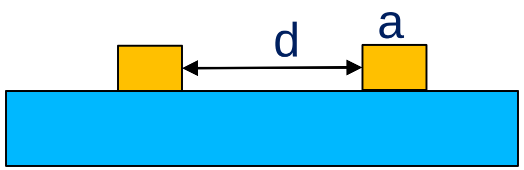
Fig. 1. Electrons or photons? The advantages (+) and handicaps (-) of electrons and photons and the role of exotic matter for enabling photons to perform logic.



TODAY	+ + +	−	+
	+ +	+ + +	+ +
TOMORROW	+ + +	−	+ +
	+ +	+ + +	+ +
DAY AFTER TOMORROW	+ + +	−	+ + +
	+ +	+ + +	+ +

Fig. 2. From today's all-electronic computer towards an all-optical computer via an intermediate hybrid system.

Electrical connections: wires

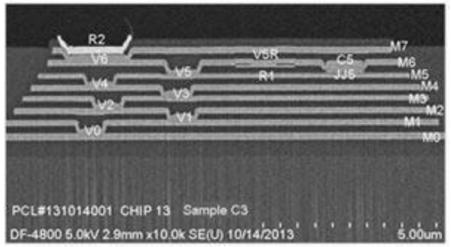


Switching an electronic signal line:

- $\mathrm{E} = \mathrm{CV}^2$
- Scaling: $\mathrm{C} \propto \frac{\pi \epsilon l}{\ln\left(\frac{d}{2a}+\sqrt{\frac{d^2}{4a^2}+1}\right)}$ → only log. Ratio (oh my god)
- Voltage: limited to 0.1 due to thermodynamics of semiconductors
- Charging wire: $\mathrm{E} = \mathrm{CV}^2 \simeq 600 \cdot 10^{-15}\ J$

1.5 pF/cm

2 pF/cm

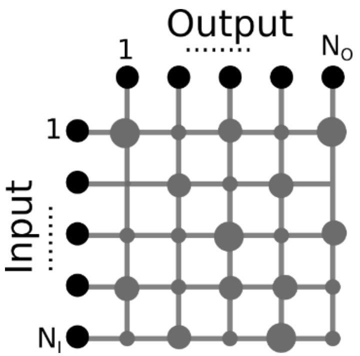

R2
V6
V5R
C5
JJ5
R1
M7
M6
M5
M4
M3
M2
M1
M0
V4
V5
V3
V2
V1
V0
PCL#131014001 CHIP 13 Sample C3
DF-4800 5.0kV 2.9mm x10.0k SE(U) 10/14/2013
5.00um

- Problem: neural networks ARE wires

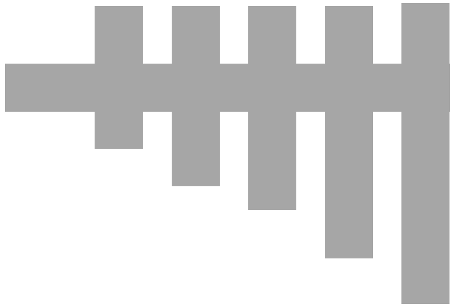
Magen, *et al.*, Proceedings of the 2004 international workshop on System level interconnect prediction.

Miller, Journal of Lightwave Technology **35**, 346 (2017).

Optical communication



Energy: switching of a capacity

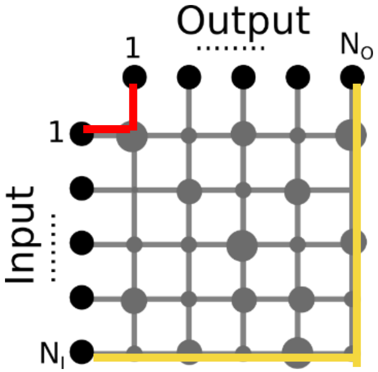


$$E \propto \; C \propto A$$

$$A = w^2\left(N_o + \frac{1}{2}\mathrm{N}_i N_o\right) + C$$

$$A \approx w^2(N + \frac{1}{2}N^2)$$

Energy: propagation losses need to be restored



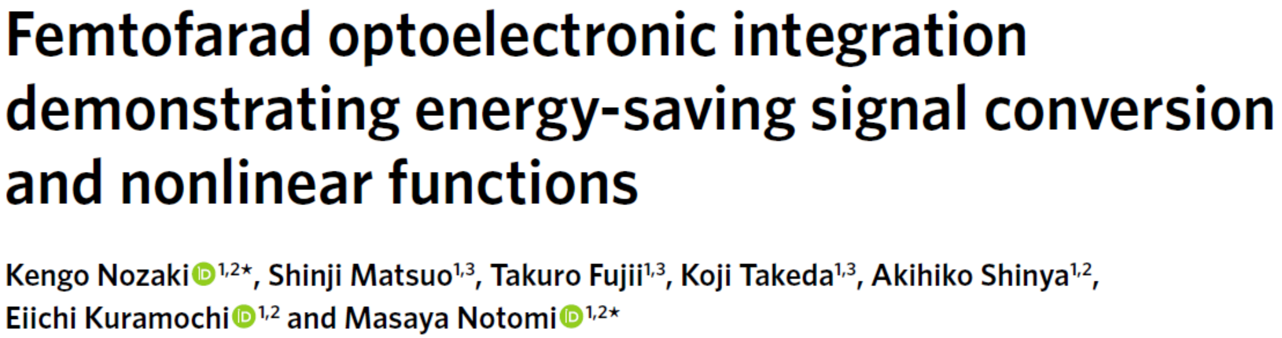
Transmission = $T \propto e^{-\gamma l}$

$L_{\min} \approx 2w,\, L_{\max} \approx \; 2Nw$

$\hat{L} \approx Nw$

This results in a cross-over point: at which lengths optical communication becomes more efficient? The determining factors are $\gamma \left[\mathbf{\frac{1}{m}}\right]$ and electro <-> optical conversion.

Electro-optical and opto-electronic conversion



Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions

Kengo Nozaki[1,2*], Shinji Matsuo[1,3], Takuro Fujii[1,3], Koji Takeda[1,3], Akihiko Shinya[1,2], Eiichi Kuramochi[1,2] and Masaya Notomi[1,2*]

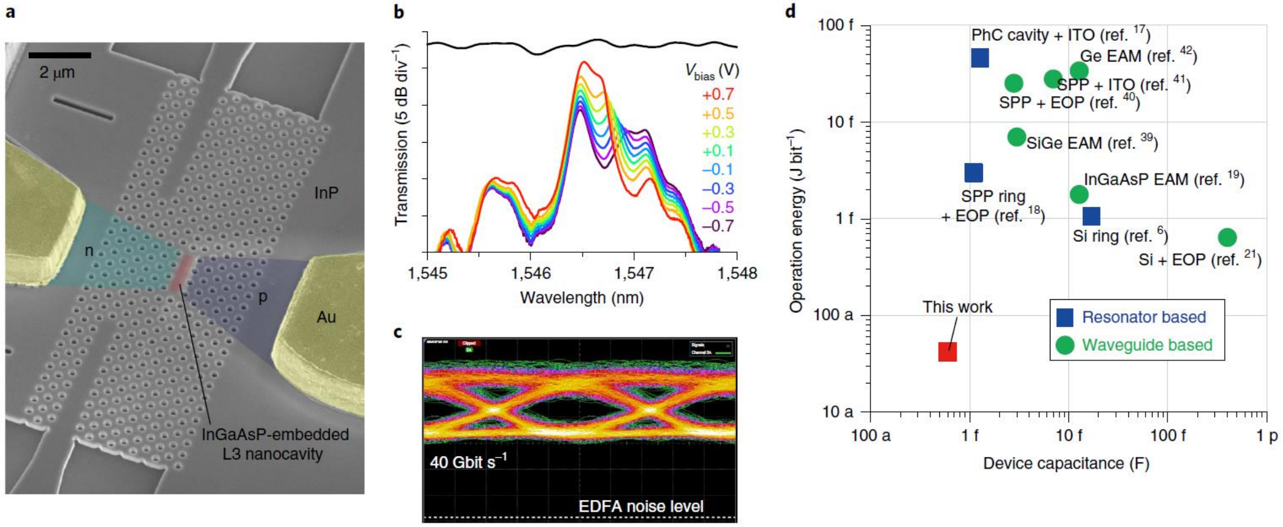
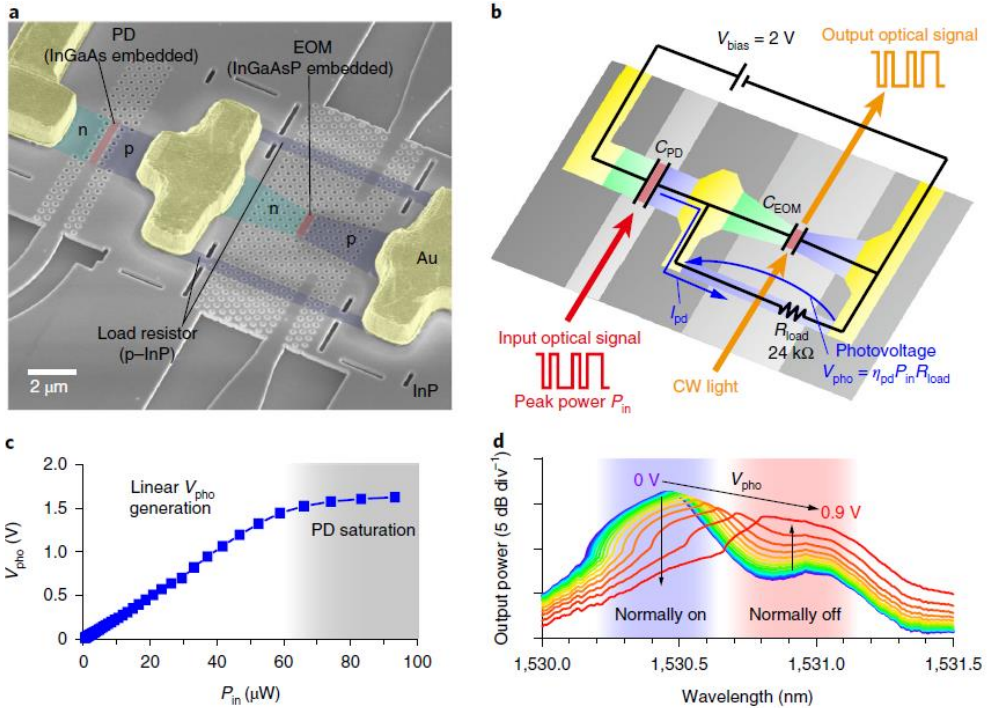
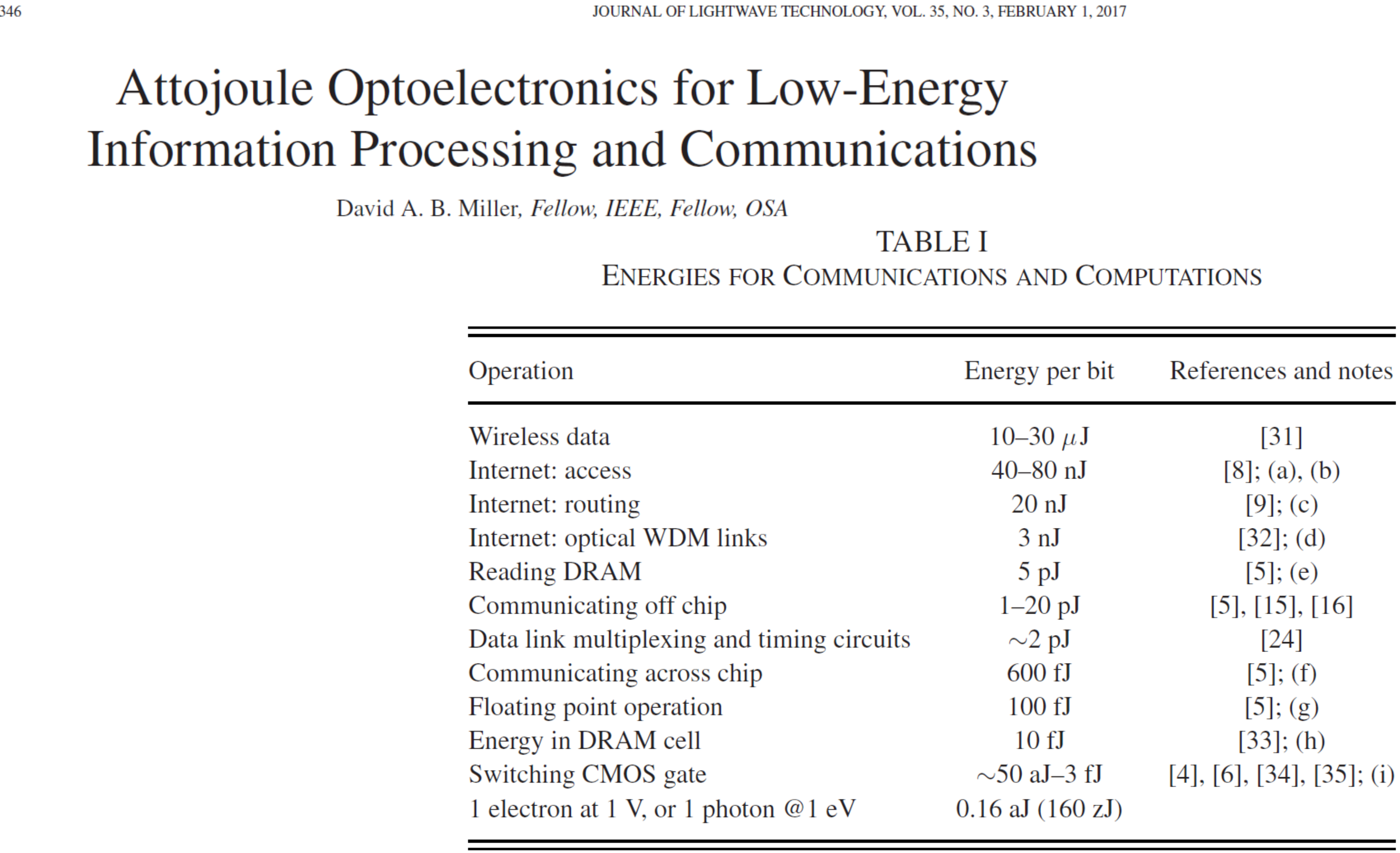


Fig. 1 | PhC-nanocavity EOM. a, Scanning electron microscope (SEM) image of the EOM. **b**, Transmission spectra for different d.c. bias voltages. The black



Electrical connections: hitting a wall



346 JOURNAL OF LIGHTWAVE TECHNOLOGY, VOL. 35, NO. 3, FEBRUARY 1, 2017

Attojoule Optoelectronics for Low-Energy Information Processing and Communications

David A. B. Miller, *Fellow, IEEE, Fellow, OSA*

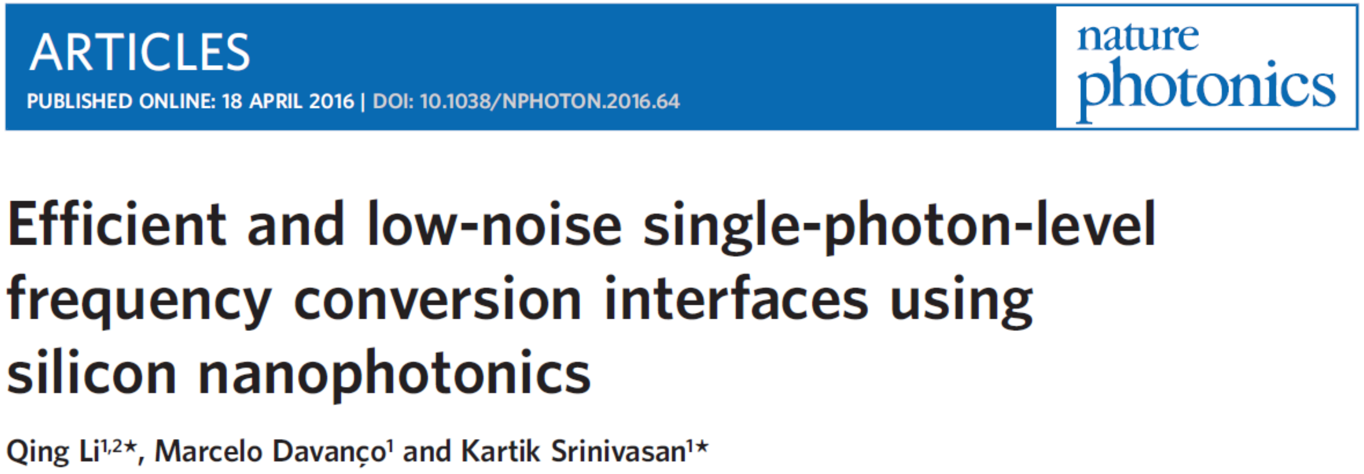
TABLE I
ENERGIES FOR COMMUNICATIONS AND COMPUTATIONS

Operation	Energy per bit	References and notes
Wireless data	10–30 μJ	[31]
Internet: access	40–80 nJ	[8]; (a), (b)
Internet: routing	20 nJ	[9]; (c)
Internet: optical WDM links	3 nJ	[32]; (d)
Reading DRAM	5 pJ	[5]; (e)
Communicating off chip	1–20 pJ	[5], [15], [16]
Data link multiplexing and timing circuits	∼2 pJ	[24]
Communicating across chip	600 fJ	[5]; (f)
Floating point operation	100 fJ	[5]; (g)
Energy in DRAM cell	10 fJ	[33]; (h)
Switching CMOS gate	∼50 aJ–3 fJ	[4], [6], [34], [35]; (i)
1 electron at 1 V, or 1 photon @1 eV	0.16 aJ (160 zJ)	

Miller, Journal of Lightwave Technology **35**, 346 (2017).

Unconventional optical computing

Two schools of ‘philosophy’

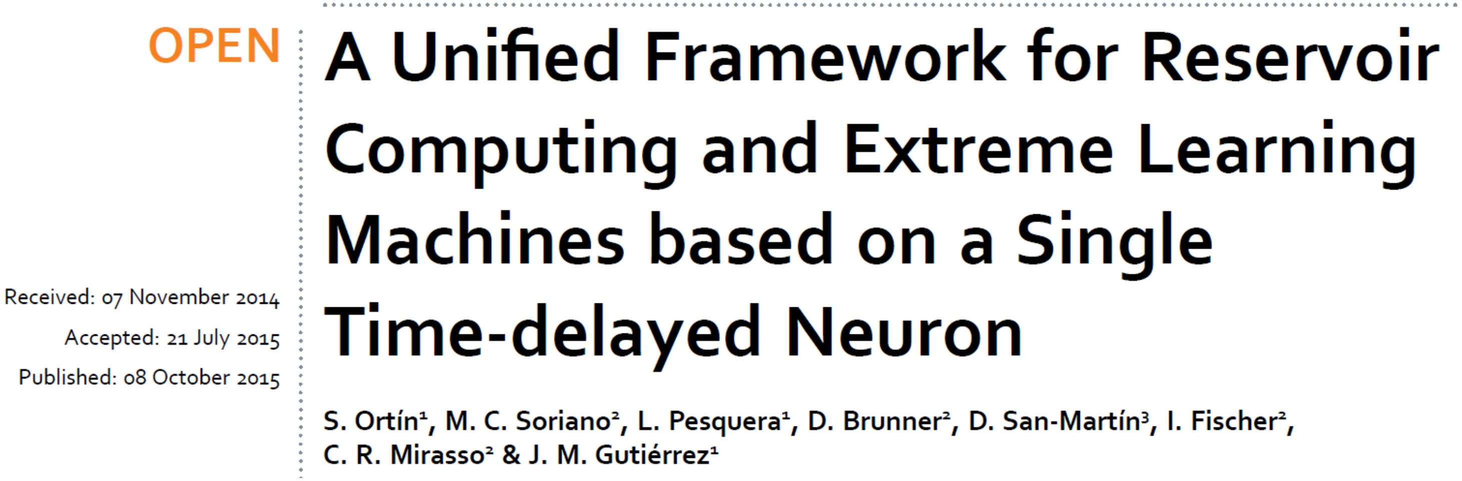


ARTICLES
PUBLISHED ONLINE: 18 APRIL 2016 | DOI: 10.1038/NPHOTON.2016.64
nature photonics

Efficient and low-noise single-photon-level frequency conversion interfaces using silicon nanophotonics

Qing Li[1,2*], Marcelo Davanço[1] and Kartik Srinivasan[1*]

top down vs. bottom up



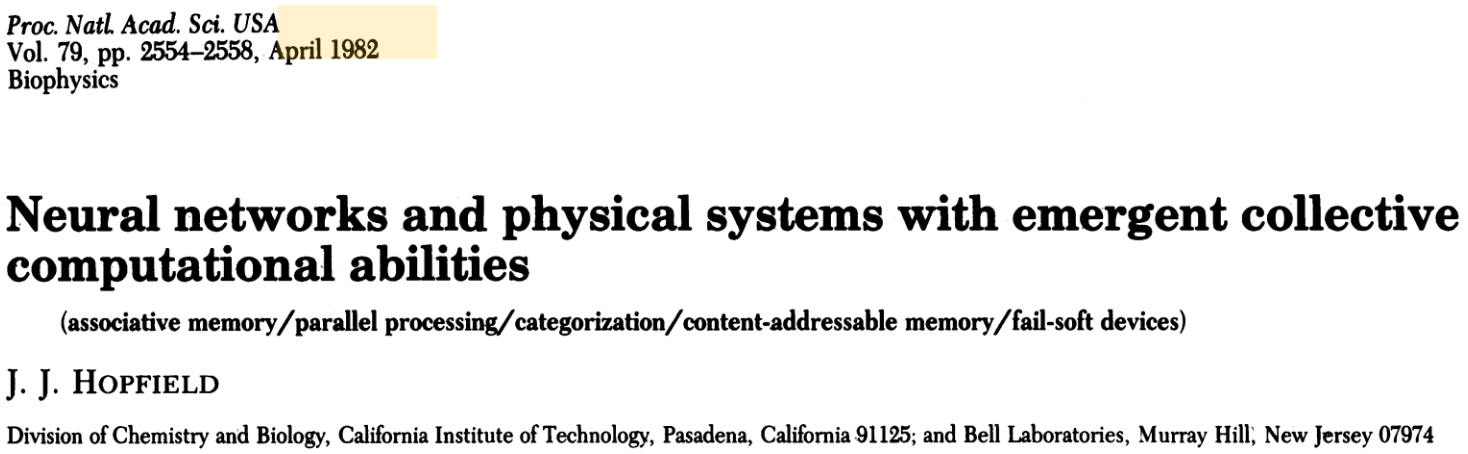
OPEN

A Unified Framework for Reservoir Computing and Extreme Learning Machines based on a Single Time-delayed Neuron

Received: 07 November 2014
Accepted: 21 July 2015
Published: 08 October 2015

S. Ortín[1], M. C. Soriano[2], L. Pesquera[1], D. Brunner[2], D. San-Martín[3], I. Fischer[2], C. R. Mirasso[2] & J. M. Gutiérrez[1]

Hopfield networks



Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2554–2558, April 1982
Biophysics

Neural networks and physical systems with emergent collective computational abilities

(associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)

J. J. Hopfield

Division of Chemistry and Biology, California Institute of Technology, Pasadena, California 91125; and Bell Laboratories, Murray Hill, New Jersey 07974

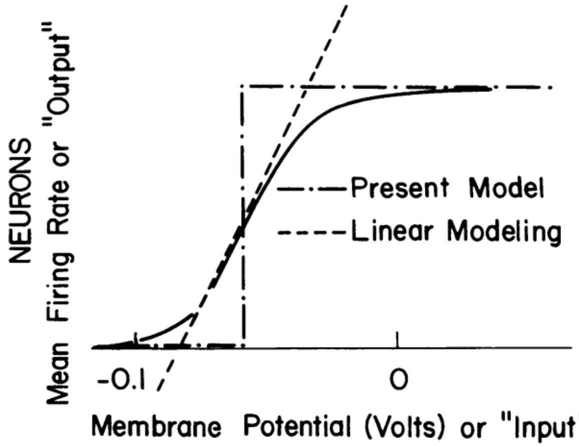
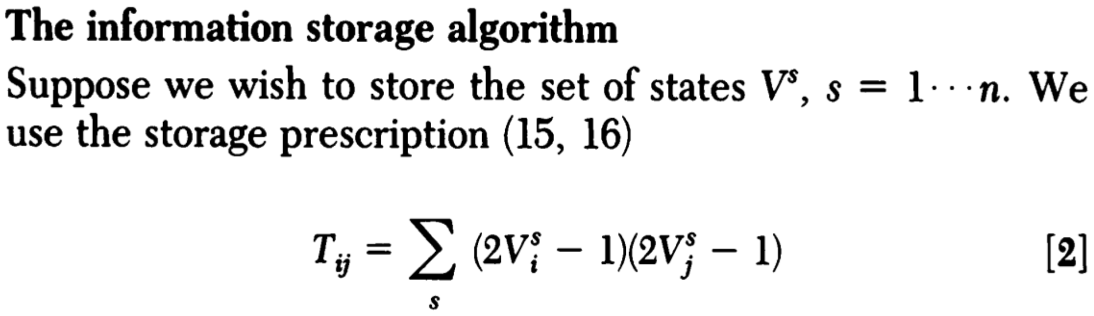


FIG. 1. Firing rate versus membrane voltage for a typical neuron (solid line), dropping to 0 for large negative potentials and saturating for positive potentials. The broken lines show approximations used in modeling.



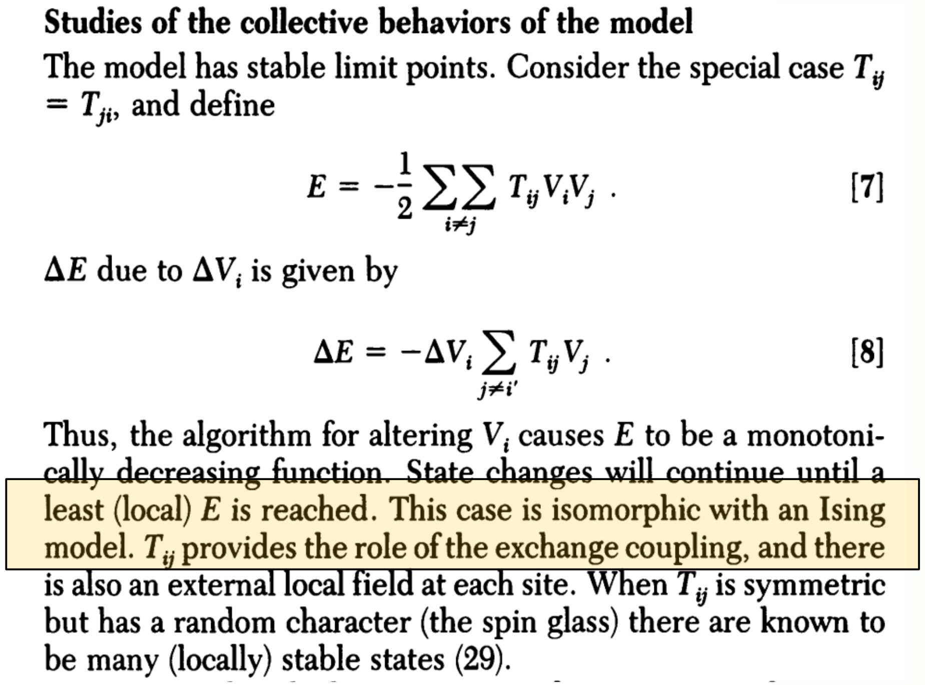
The information storage algorithm

Suppose we wish to store the set of states V^s, $s = 1 \cdots n$. We use the storage prescription (15, 16)

$$T_{ij} = \sum_s (2V_i^s - 1)(2V_j^s - 1) \qquad [2]$$

- V^s are memory entries
- Stable if V_i^s is V_j^s -> correcting memory
- Connections T_{ij} analytically defined

Hopfield, why its so nice for physicists



Studies of the collective behaviors of the model

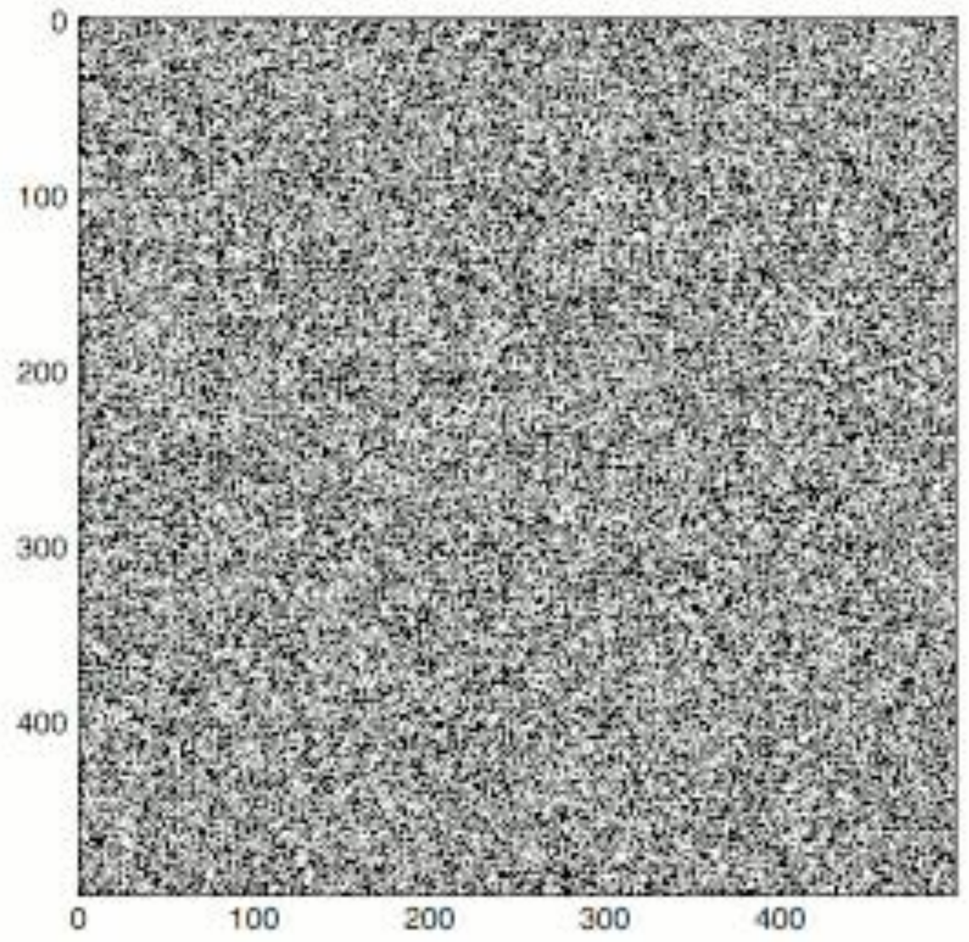
The model has stable limit points. Consider the special case $T_{ij} = T_{ji}$, and define

$$E = -\frac{1}{2} \sum\sum_{i \neq j} T_{ij} V_i V_j \ . \qquad [7]$$

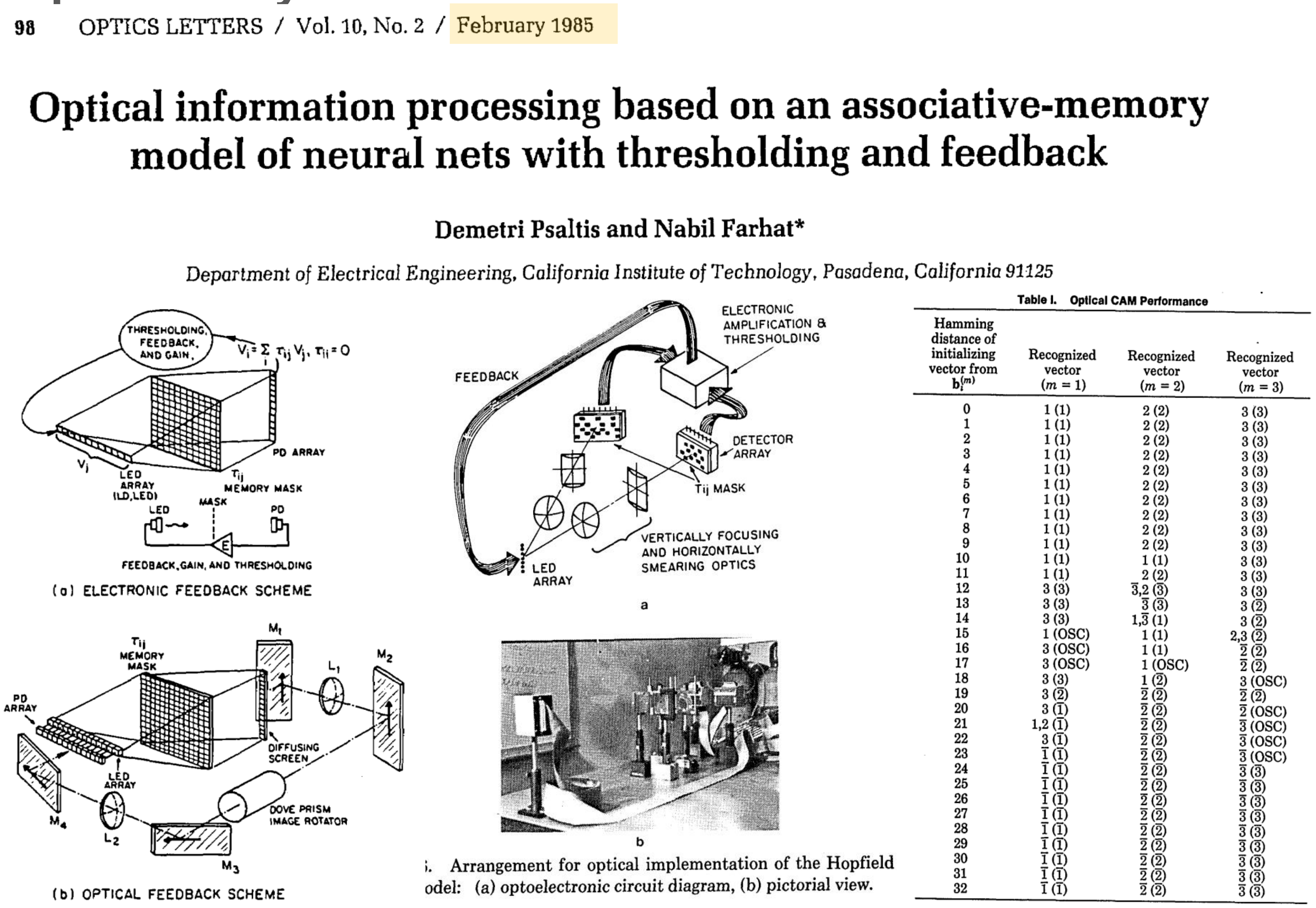
ΔE due to ΔV_i is given by

$$\Delta E = -\Delta V_i \sum_{j \neq i'} T_{ij} V_j \ . \qquad [8]$$

Thus, the algorithm for altering V_i causes E to be a monotonically decreasing function. State changes will continue until a least (local) E is reached. This case is isomorphic with an Ising model. T_{ij} provides the role of the exchange coupling, and there is also an external local field at each site. When T_{ij} is symmetric but has a random character (the spin glass) there are known to be many (locally) stable states (29).



Hopfield: why its nice II



98 OPTICS LETTERS / Vol. 10, No. 2 / February 1985

Optical information processing based on an associative-memory model of neural nets with thresholding and feedback

Demetri Psaltis and Nabil Farhat*

Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125

THRESHOLDING, FEEDBACK, AND GAIN,
$V_i = \sum_j T_{ij} V_j, \; T_{ii} = 0$
PD ARRAY
V_j
LED ARRAY (LD, LED)
T_{ij} MEMORY MASK
LED
MASK
PD
E
FEEDBACK, GAIN, AND THRESHOLDING
(a) ELECTRONIC FEEDBACK SCHEME
T_{ij} MEMORY MASK
M_1
M_2
L_1
PD ARRAY
DIFFUSING SCREEN
LED ARRAY
M_4
L_2
DOVE PRISM IMAGE ROTATOR
M_3
(b) OPTICAL FEEDBACK SCHEME
ELECTRONIC AMPLIFICATION & THRESHOLDING
FEEDBACK
DETECTOR ARRAY
T_{ij} MASK
VERTICALLY FOCUSING AND HORIZONTALLY SMEARING OPTICS
LED ARRAY
a
b

Arrangement for optical implementation of the Hopfield odel: (a) optoelectronic circuit diagram, (b) pictorial view.

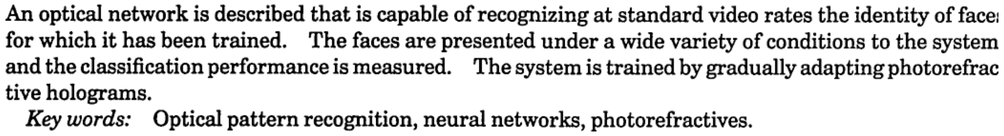
Table I. Optical CAM Performance

Hamming distance of initializing vector from $\mathbf{b}_i^{(m)}$	Recognized vector ($m = 1$)	Recognized vector ($m = 2$)	Recognized vector ($m = 3$)
0	1 (1)	2 (2)	3 (3)
1	1 (1)	2 (2)	3 (3)
2	1 (1)	2 (2)	3 (3)
3	1 (1)	2 (2)	3 (3)
4	1 (1)	2 (2)	3 (3)
5	1 (1)	2 (2)	3 (3)
6	1 (1)	2 (2)	3 (3)
7	1 (1)	2 (2)	3 (3)
8	1 (1)	2 (2)	3 (3)
9	1 (1)	2 (2)	3 (3)
10	1 (1)	1 (1)	3 (3)
11	1 (1)	2 (2)	3 (3)
12	3 (3)	$\bar{3}$,2 ($\bar{3}$)	3 (3)
13	3 (3)	$\bar{3}$ ($\bar{3}$)	3 ($\bar{2}$)
14	3 (3)	1,$\bar{3}$ (1)	3 ($\bar{2}$)
15	1 (OSC)	1 (1)	2,3 ($\bar{2}$)
16	3 (OSC)	1 (1)	$\bar{2}$ ($\bar{2}$)
17	3 (OSC)	1 (OSC)	$\bar{2}$ ($\bar{2}$)
18	3 (3)	1 ($\bar{2}$)	3 (OSC)
19	3 ($\bar{2}$)	$\bar{2}$ ($\bar{2}$)	$\bar{2}$ ($\bar{2}$)
20	3 ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{2}$ (OSC)
21	1,2 ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ (OSC)
22	3 ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ (OSC)
23	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ (OSC)
24	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
25	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
26	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
27	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
28	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
29	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
30	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
31	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)
32	$\bar{1}$ ($\bar{1}$)	$\bar{2}$ ($\bar{2}$)	$\bar{3}$ ($\bar{3}$)

And finally: multilayer networks

Optical network for real-time face recognition

Hsin-Yu Sidney Li, Yong Qiao, and Demetri Psaltis



An optical network is described that is capable of recognizing at standard video rates the identity of faces for which it has been trained. The faces are presented under a wide variety of conditions to the system and the classification performance is measured. The system is trained by gradually adapting photorefractive holograms.

Key words: Optical pattern recognition, neural networks, photorefractives.

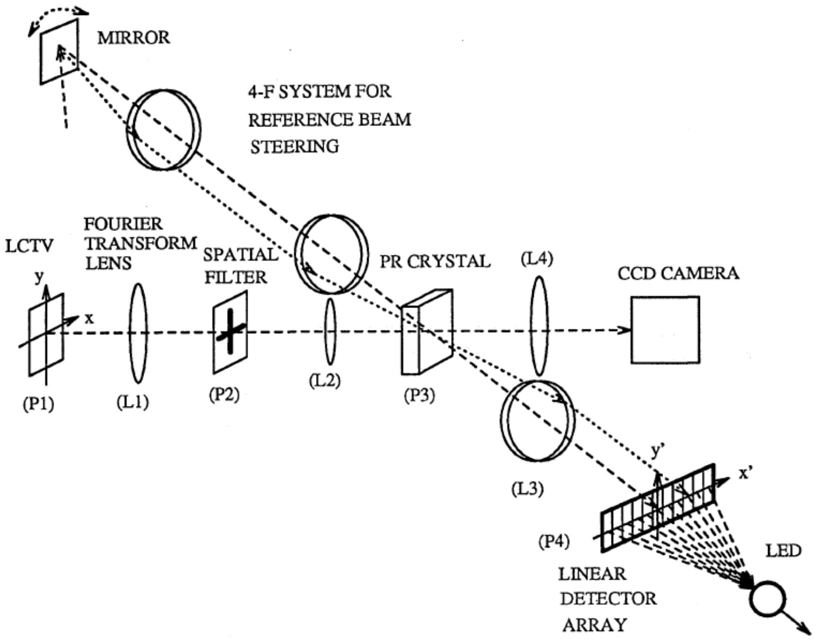


Fig. 1. Optical setup of the face-recognition system; PR, photorefractive.

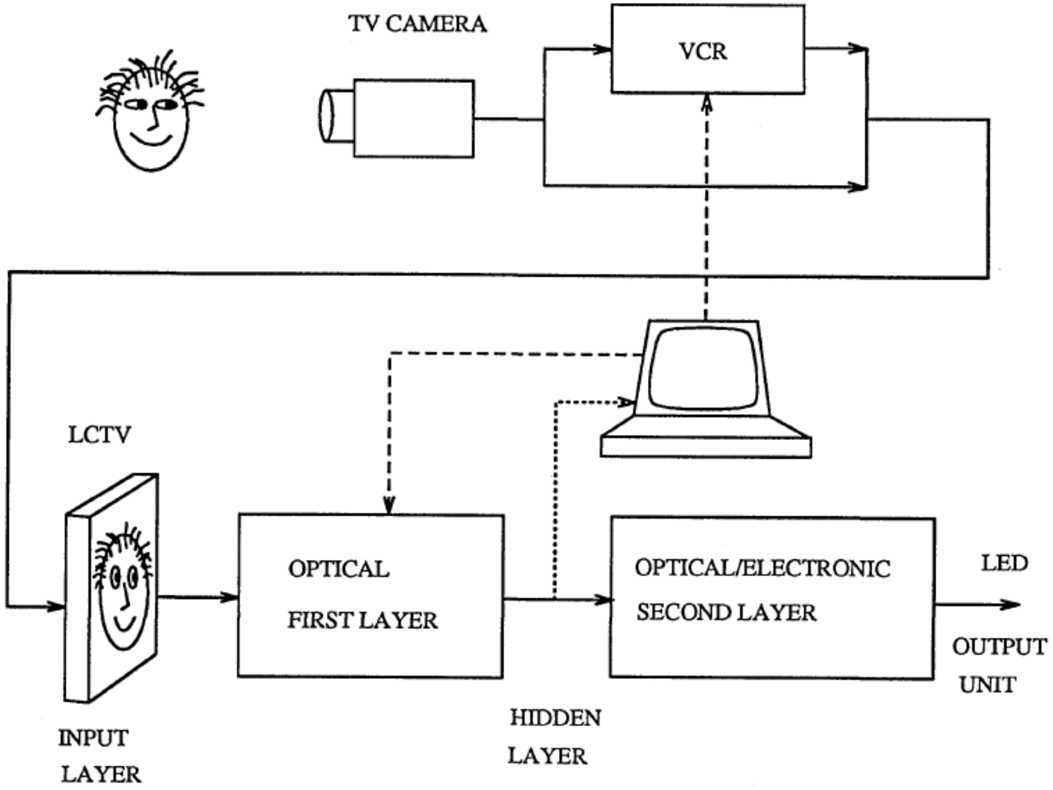


Fig. 5. Schematic diagram of overall system.

Implementation

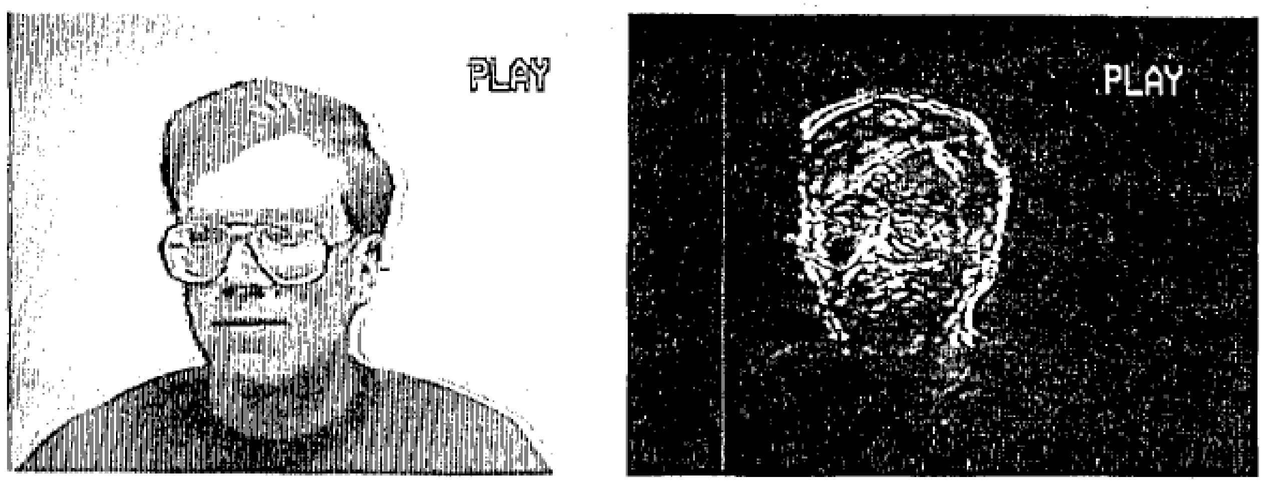
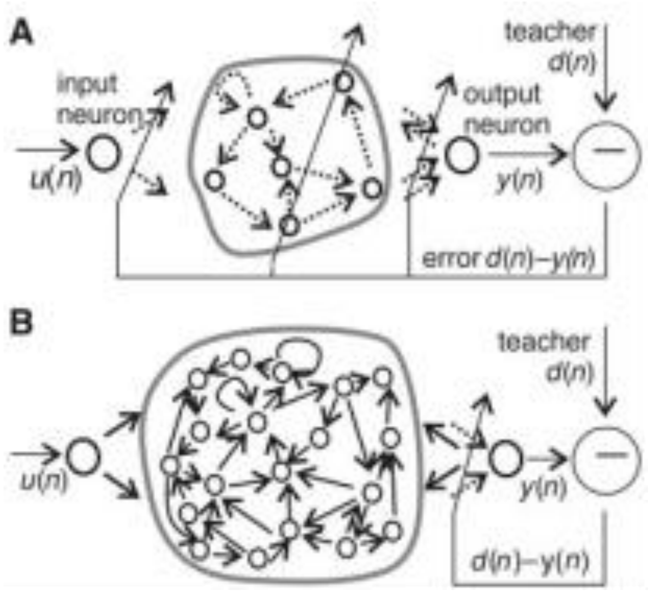


Fig. 3. Edge-enhanced image and original face.

Fig. 7. Photographs showing part of the training session.

Another great starting concept: Reservoir computing / Extreme learning machines



Jaeger, Science **304**, 78 (2004).

A: fully adjusted network

- Optimal performance (maybe)
- Exploding/vanishing gradients
- None-converging beyond stability

B: only output weights

- Random input/internal weights
- Training is simple matrix inversion

<u>Echo state networks</u>

- Nonlinear maps with discrete time
- Origin in field of control system
- ELM: no recurrent connections

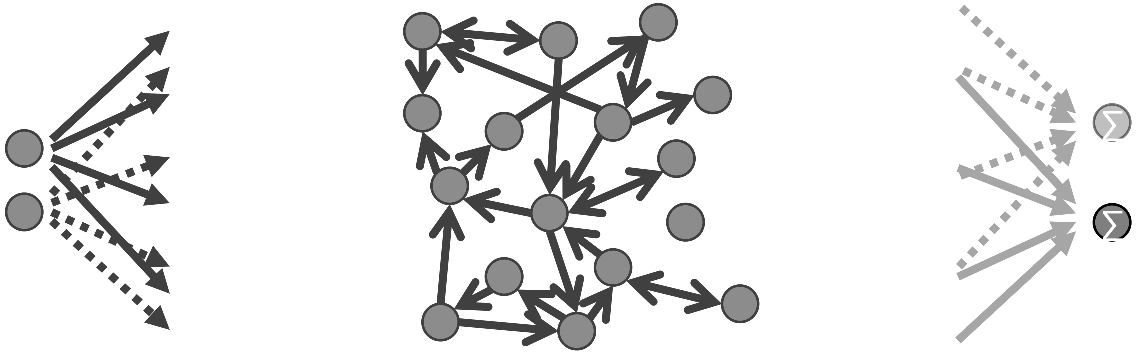
*Jaeger and Haas, Science **304,** 5667 (2004).*

<u>Liquid state machines</u>

- Excitable (spiking) neurons
- Origin in Computational neuro science

*Buonomano, Maass, Nature reviews in Neuroscience **10,** 113 (2009).*

Experiment type A: network state attainable

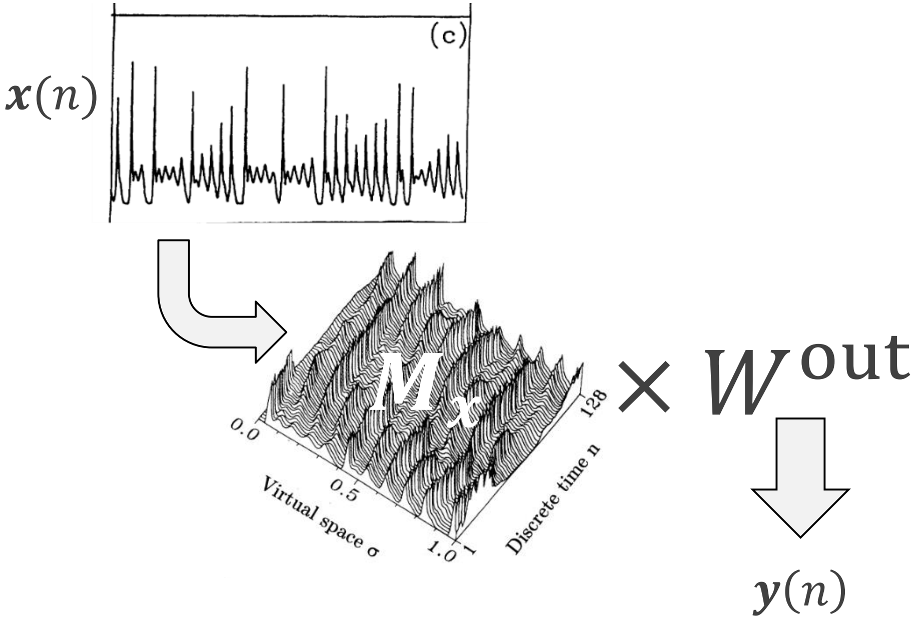


Input:
$\boldsymbol{u}(n+1)$

Network state:
$\boldsymbol{x}(n+1) =$
$\boldsymbol{f}[W\boldsymbol{x}(n) + W^{\text{inj}}\boldsymbol{u}(n+1) + \boldsymbol{b}]$

Output:
$\boldsymbol{y}(n+1) =$
$W^{\text{out}}\boldsymbol{x}(n+1)$

Arecchi *et al.*, Phys. Rev. A **45**, 4225 (1992).



1. Training data: set of inputs $\boldsymbol{u}(n)$ for which $\boldsymbol{y}(n)$ is known
2. Collect $\boldsymbol{x}(n)$ for $n \,\epsilon [1, \dots, T]$
3. L -fold cross-validation: randomly label input data by $l \,\epsilon [1, \dots, L]$
4. Select one $l_c \,\epsilon [1, \dots, L]$
   - M_x: concatenated $\boldsymbol{x}(n)$ for $l \neq l_c$
   - T^T: concatenated matrix of $\boldsymbol{y}(n)$ for $l \neq l_c$
   - Obtain $W^{\text{out}} = (M_x M_x^T + \lambda I)^{-1}(M_x T^T)$
   - $\boldsymbol{y}(n)$, measure error for l_c
   - Repeat for all l_c
5. Measure error for unseen data

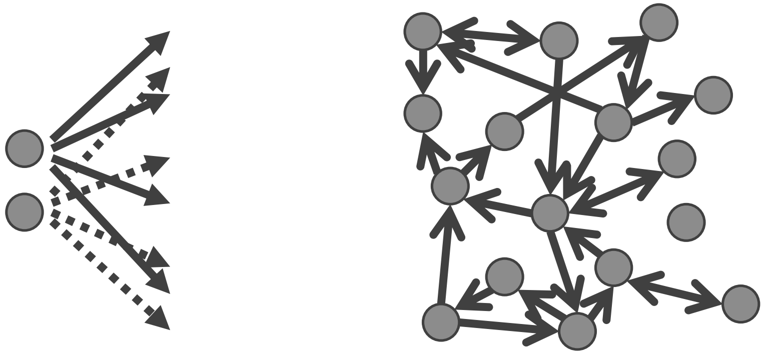
Why is this great?

- ○ Training offline
- ○ ‘Short’ experiment

Why is this not great?

- ○ Training offline
- ○ Technological relevance?

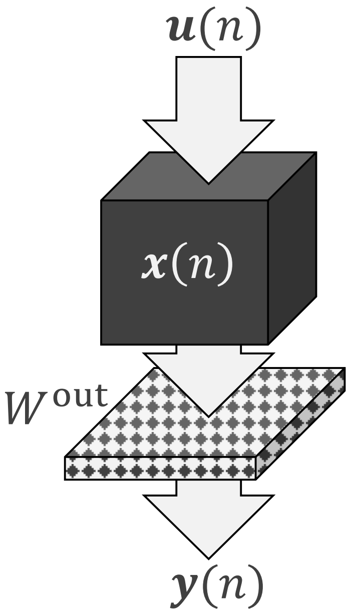
Experiment type B: network state not attainable



Input:
$\boldsymbol{u}(n+1)$

Network state:
$\boldsymbol{x}(n+1) =$
$\boldsymbol{f}[W\boldsymbol{x}(n) + W^{\text{inj}}\boldsymbol{u}(n+1) + \boldsymbol{b}]$

Output:
$\boldsymbol{y}(n+1) =$
$W^{\text{out}}\boldsymbol{x}(n+1)$



1. Training data: set of inputs $\boldsymbol{u}(n)$ for which $\boldsymbol{y}(n)$ is known
2. Define starting W^{out}
   - Collect $\boldsymbol{y}(n)$ for $n \,\epsilon [1, \dots, T]$
   - Measure error
3. Learn
   - Modify W^{out}
   - Collect $\boldsymbol{y}(n)$ for $n \,\epsilon [1, \dots, T]$
   - Measure error
   - Evaluate modification and repeat
4. Measure error for unseen examples

Why is this great?

- Training online
- In-situ (vs. in silico)
- ‘Autonomous’ system

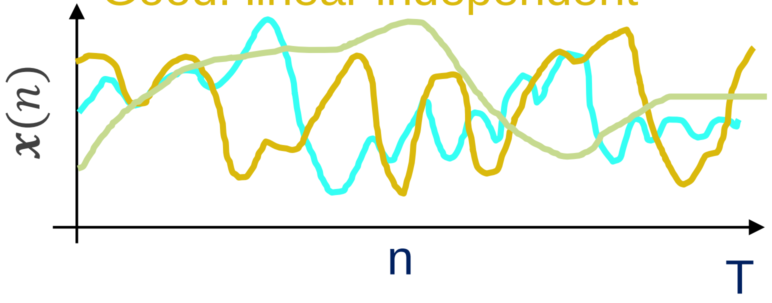
Why is this not great?

- Training online
- Blind to network state
- Comparison tricky

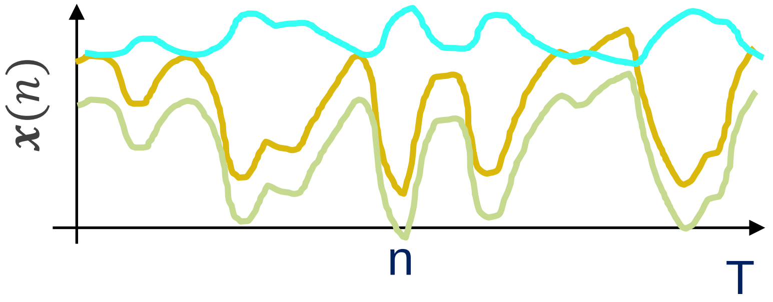
A good (well behaved) reservoir

Diversity (dimensionality)

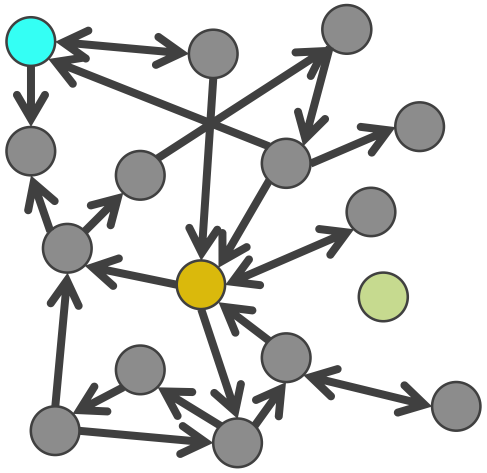
Good: linear independent



Bad: linear dependent

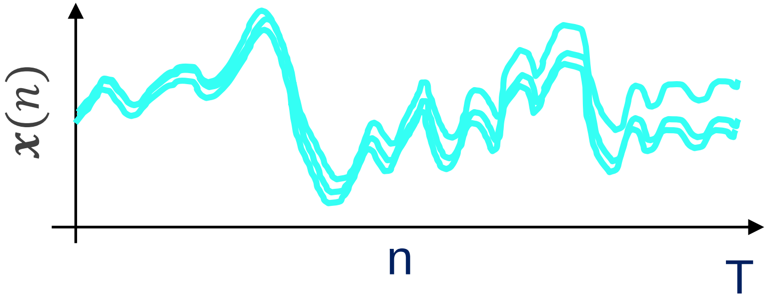


Dambre *et al.*, Sci. Rep. **2**, 514 (2012).

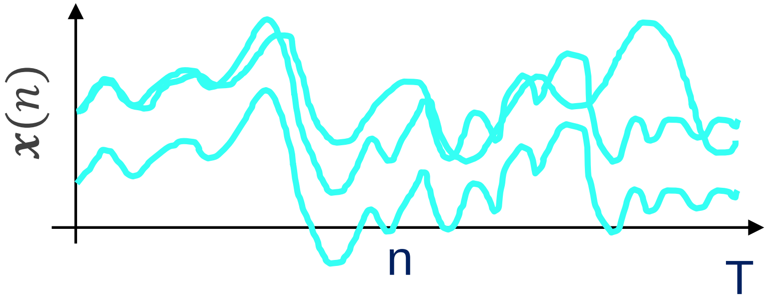


Reproducibility (consistency)

Good: small variation



Bad: big or exploding variation

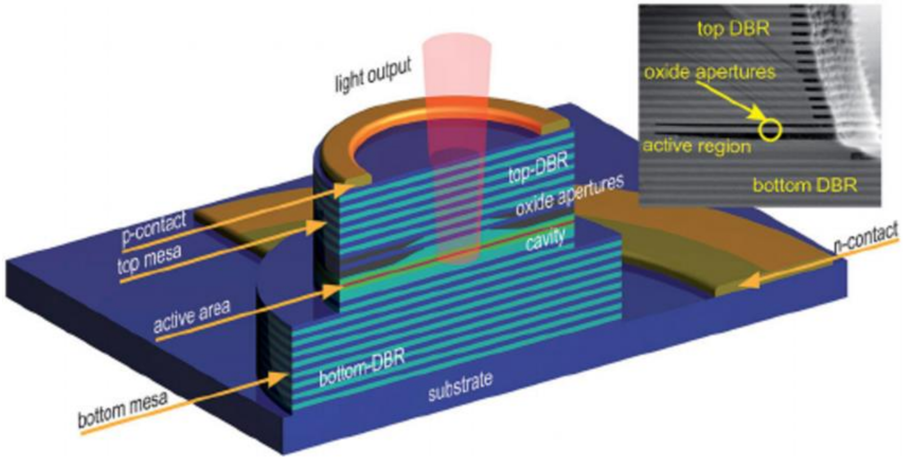


Uchida *et al.*, PRL **93**, 244102 (2004).

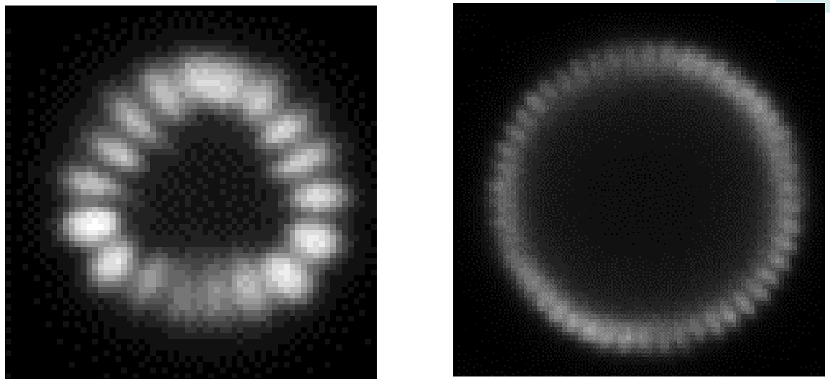
Autonomous computing with a laser

SCALING DIMENSION ENERGY EFFICIENCY / AUTONOMY

- **Large Area Multimode Vertical-Cavity Surface-Emitting Lasers** (LA-VCSELs) are **quantum well** semiconductor lasers. They differ from conventional VCSELs by their large (> 6 microns) diameter.



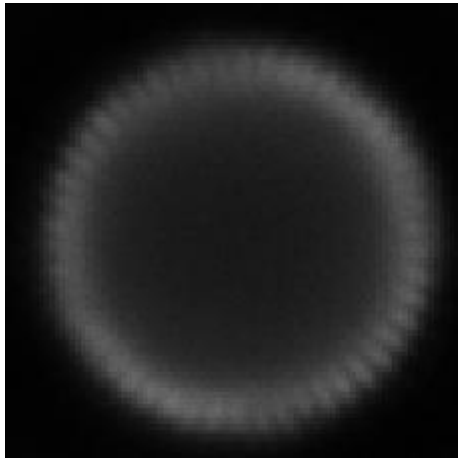
Single mode



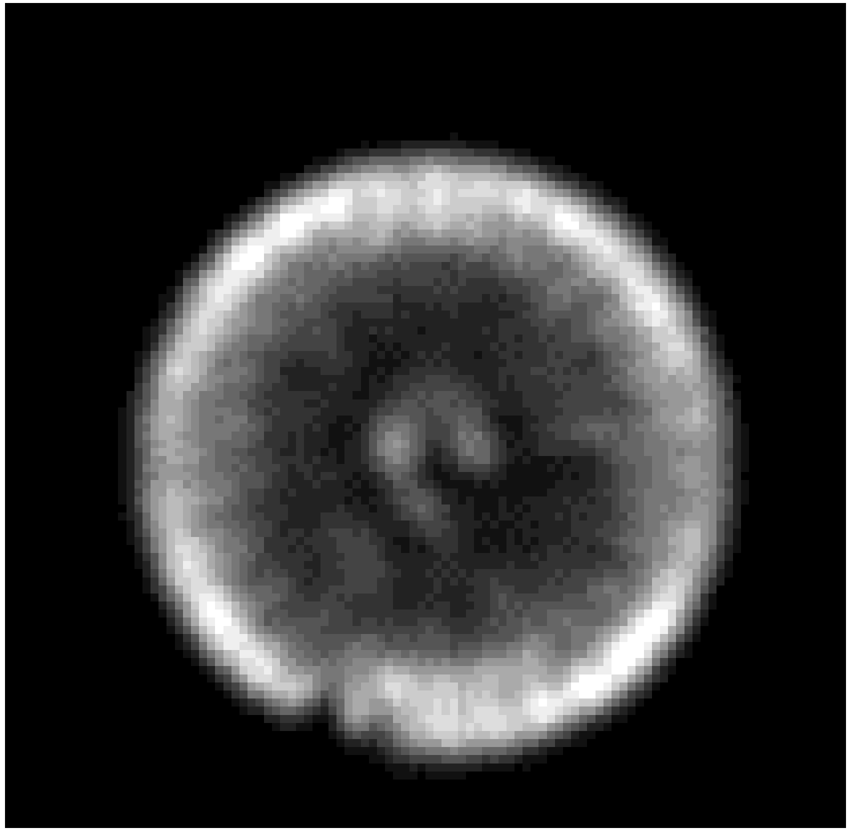
Multimode

- These are **highly multimode, highly non-linear** devices.
- The **VCSEL dynamics** change with **optical injection**.
- Different injection conditions yield **very different mode profiles.**
- Can be modulated at **high bandwidths (20 GHz)**

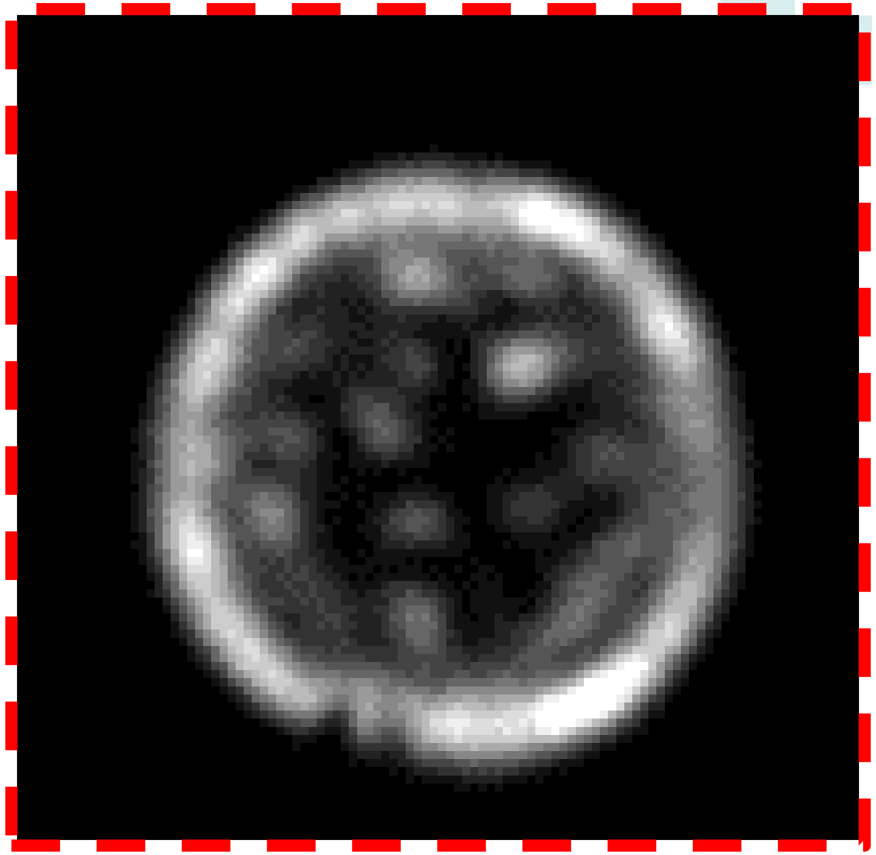
CONCEPT ILLUSTRATION



Free running

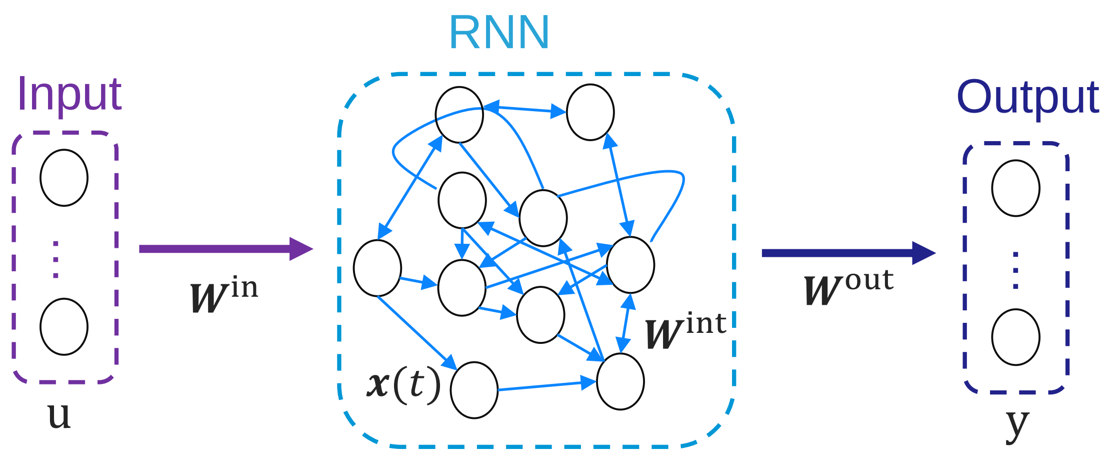


Off resonant injection

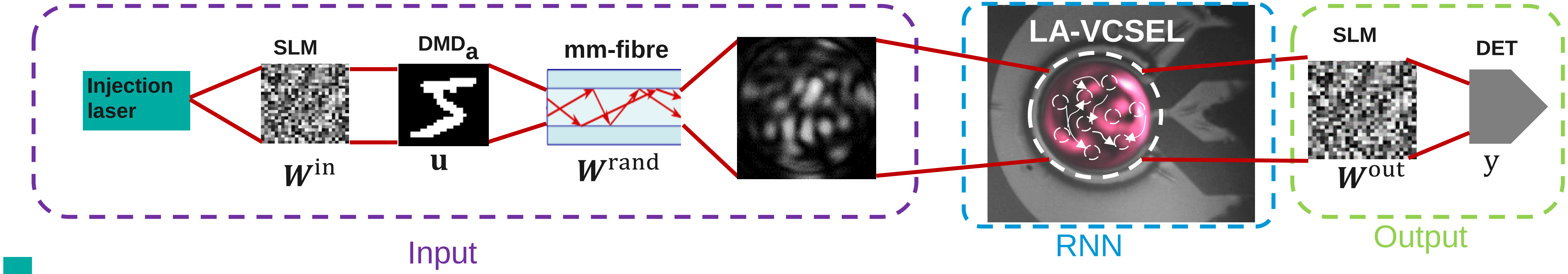


Resonant injection

Concept



Working principle

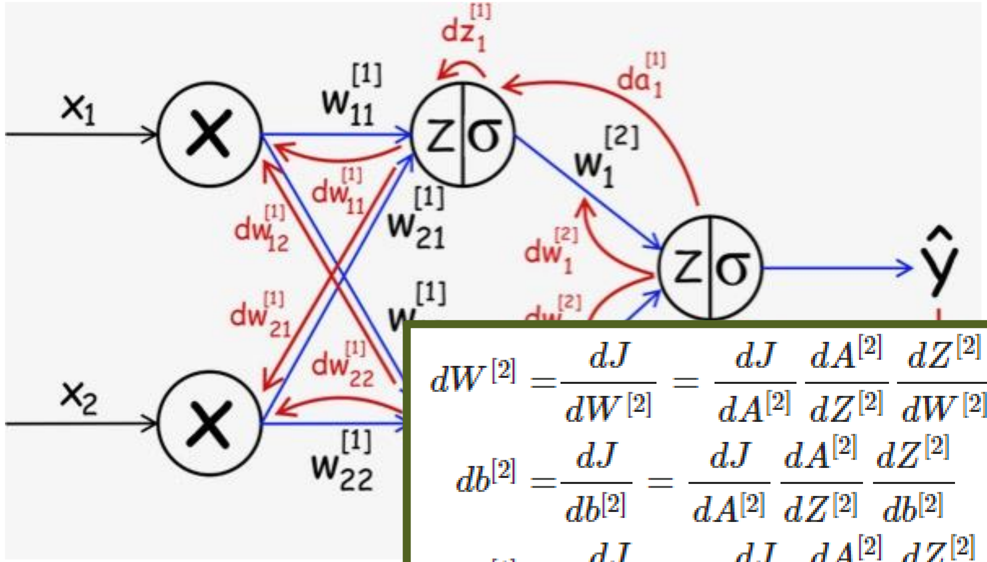


Porte, et al., *J. Phys. Photonics 3, 024017 (2021).*

SCALING DIMENSION: LEARNING

Nature

Deep learning



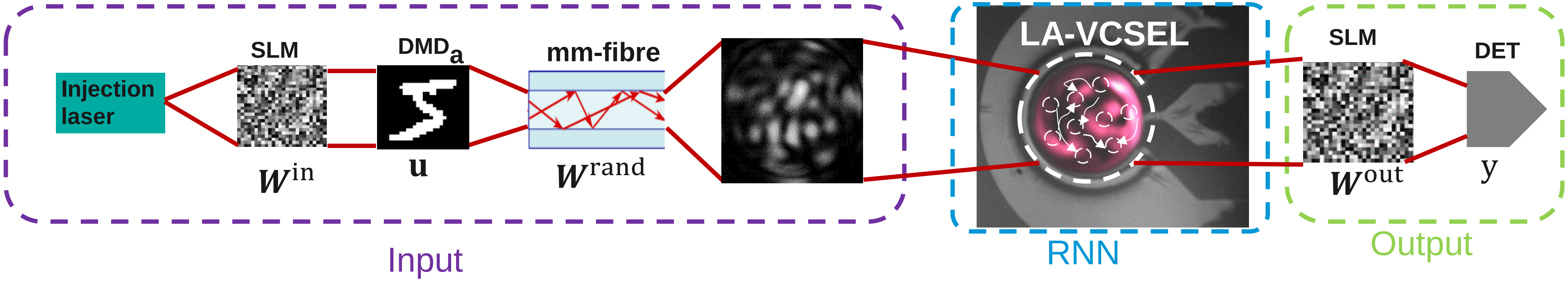
$$dW^{[2]} = \frac{dJ}{dW^{[2]}} = \frac{dJ}{dA^{[2]}} \frac{dA^{[2]}}{dZ^{[2]}} \frac{dZ^{[2]}}{dW^{[2]}}$$

$$db^{[2]} = \frac{dJ}{db^{[2]}} = \frac{dJ}{dA^{[2]}} \frac{dA^{[2]}}{dZ^{[2]}} \frac{dZ^{[2]}}{db^{[2]}}$$

$$dW^{[1]} = \frac{dJ}{dW^{[2]}} = \frac{dJ}{dA^{[2]}} \frac{dA^{[2]}}{dZ^{[2]}} \frac{dZ^{[2]}}{dA^{[1]}} \frac{dA^{[1]}}{dZ^{[1]}} \frac{dZ^{[1]}}{dW^{[1]}}$$

$$db^{[1]} = \frac{dJ}{dW^{[2]}} = \frac{dJ}{dA^{[2]}} \frac{dA^{[2]}}{dZ^{[2]}} \frac{dZ^{[2]}}{dA^{[1]}} \frac{dA^{[1]}}{dZ^{[1]}} \frac{dZ^{[1]}}{db^{[1]}}$$

COMPATIBLE TRAINING STRATEGIES



Given the constrains we have : ***Black box optimization***

Evolutionary Strategies / reinforcement learning : adaptation of distribution (mean, std. dev.) based on a sampled population.

Measured Gradient estimation: physically measuring the gradient with perturbations and performing gradient descent

1.Hansen, Nikolaus. "The CMA evolution strategy: A tutorial." *arXiv preprint arXiv:1604.00772* (2016).
2.Spall, James C. "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation." *IEEE transactions on automatic control* 37.3 (1992): 332-341.

GRADIENT-BASED STRATEGIES

Low hanging fruit Gradient Descent:

```
• Initialize weights : $\boldsymbol{W}$
• $\epsilon = small\ value$
• $k_{weights}$
• $for\ i = 1: epochs$
    $s = k_{weights}\ randomly\ selected\ among\ W$
    $for\ k = 1: s$
        $\frac{\boldsymbol{\partial L}}{\boldsymbol{\partial W}}(\boldsymbol{k}) \approx \frac{L(\boldsymbol{W}(\boldsymbol{k}) + \epsilon) - L(\boldsymbol{W}(\boldsymbol{k}) - \epsilon)}{2\epsilon}$
    $\boldsymbol{W} = \boldsymbol{W} - \alpha * \frac{\boldsymbol{\partial L}}{\boldsymbol{\partial w}}$
```

Perturbs each weights one by one

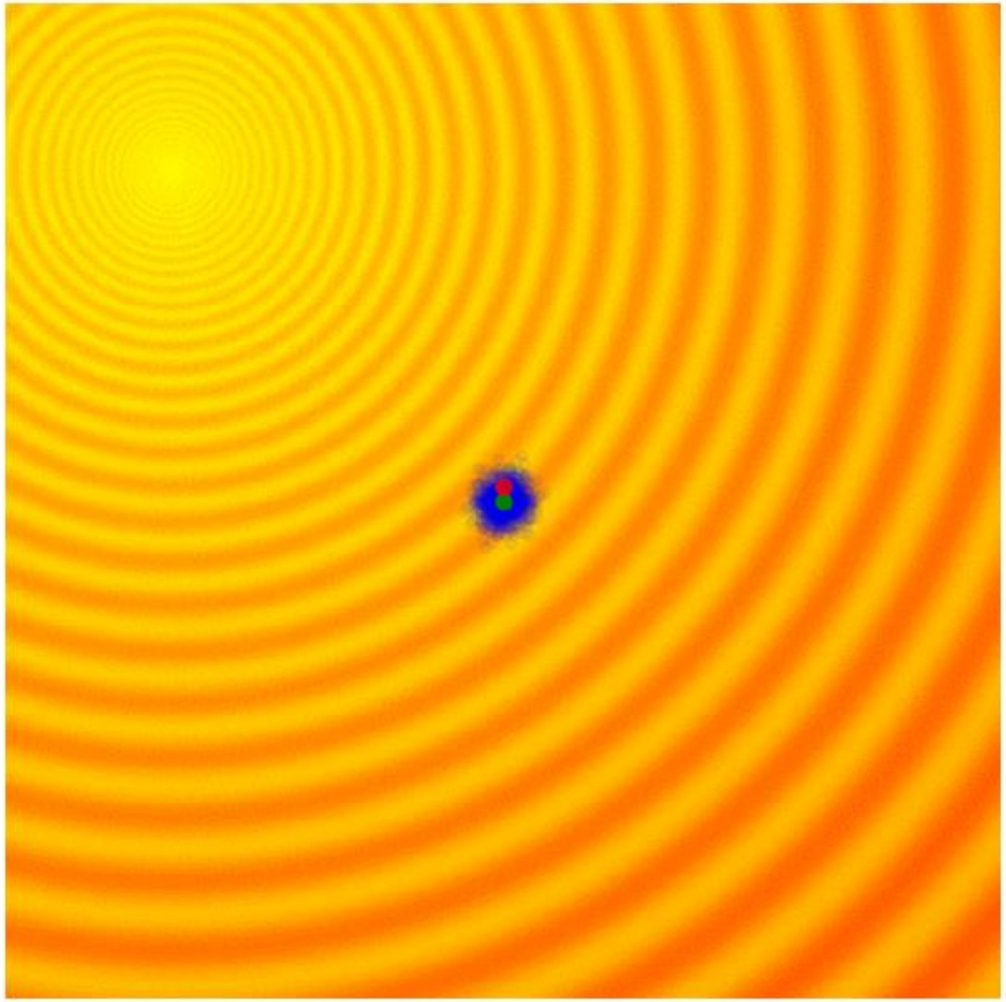
SPSA (Simultaneous Perturbation Stochastic approximation):

```
• Initialize weights : $\boldsymbol{W}$
• $\epsilon = \text{small value}$
• $for\ i = 1: epochs$
    $\boldsymbol{\Delta} = \text{rand}\{-1,1\}\ \text{of size}(\text{W})$
    $\frac{\boldsymbol{\partial L}}{\boldsymbol{\partial W}} \approx \frac{L(\boldsymbol{W} + \epsilon * \boldsymbol{\Delta}) - L(\boldsymbol{W} - \epsilon * \boldsymbol{\Delta})}{2\epsilon} * \boldsymbol{\Delta}$
    $\boldsymbol{W} = \boldsymbol{W} - \alpha * \frac{\boldsymbol{\partial L}}{\boldsymbol{\partial w}}$
```

Perturbs all weights according to one direction

Spall, James C. "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation." *IEEE transactions on automatic control* 37.3 (1992): 332-341.

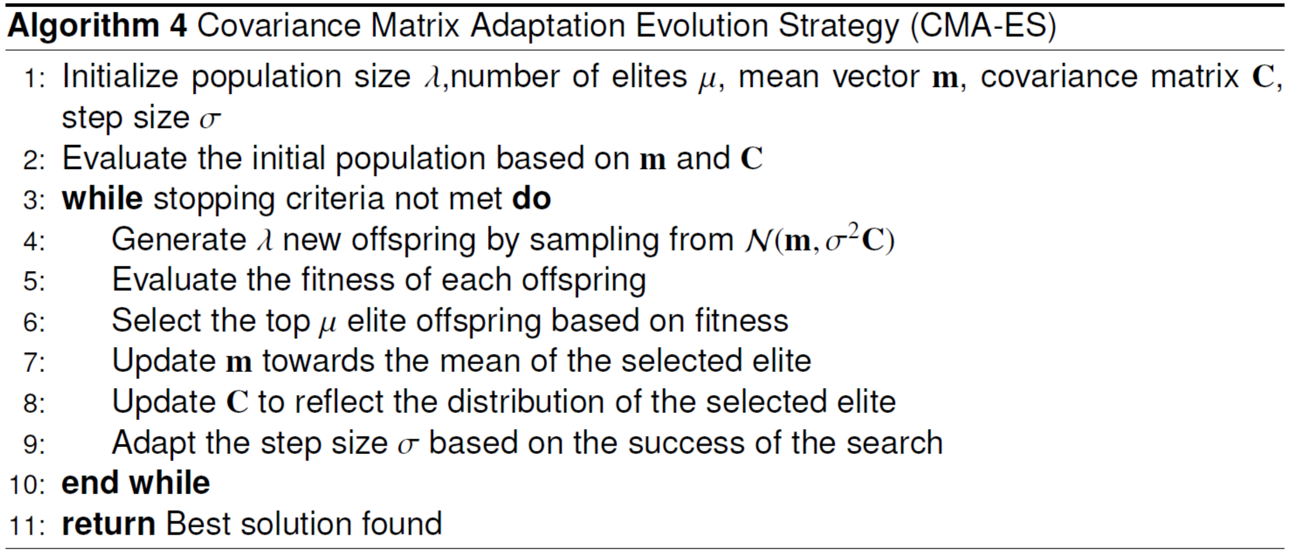
SIMPLEST EVOLUTIONARY STRATEGY



```
While (termination criteria not met) do:
    • Generate a population P from a normal distribution
    • Select the best point from the population.
    • Set the new mean of the distribution to this best point
    • Generate a new population P centered around the new mean
```

CMA-ES AND PEPG

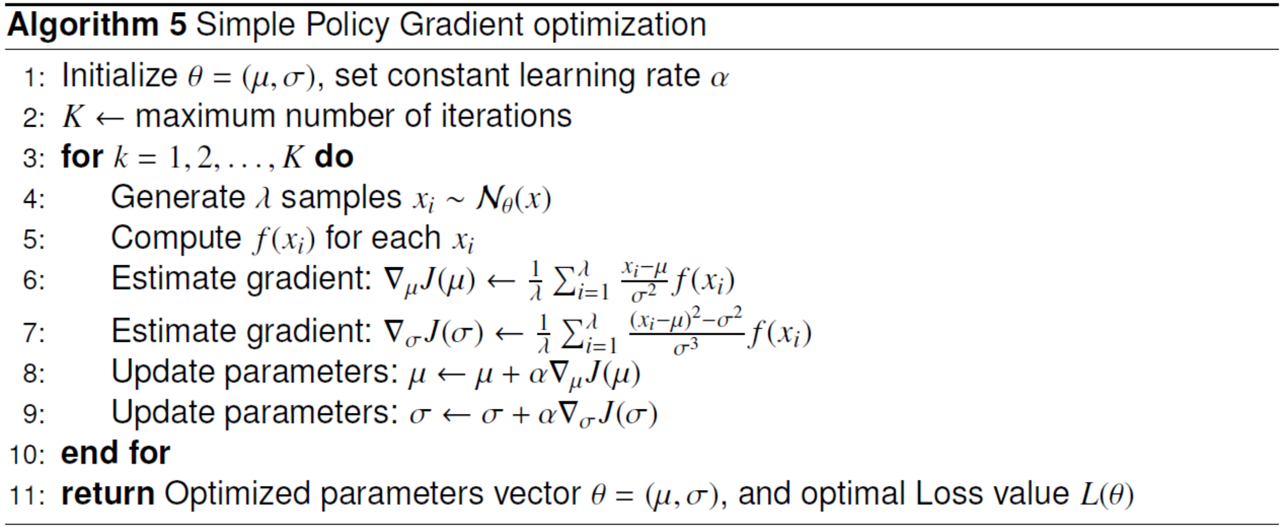
- **CMA**: Adapts the shape of a multivariate gaussian distribution, by estimating and mutating the covariance matrix, and mean according to some elites in the population.



Algorithm 4 Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

```
Initialize population size $\lambda$,number of elites $\mu$, mean vector $\mathbf{m}$, covariance matrix $\mathbf{C}$, step size $\sigma$
Evaluate the initial population based on $\mathbf{m}$ and $\mathbf{C}$
while stopping criteria not met do
    Generate $\lambda$ new offspring by sampling from $\mathcal{N}(\mathbf{m}, \sigma^2\mathbf{C})$
    Evaluate the fitness of each offspring
    Select the top $\mu$ elite offspring based on fitness
    Update $\mathbf{m}$ towards the mean of the selected elite
    Update $\mathbf{C}$ to reflect the distribution of the selected elite
    Adapt the step size $\sigma$ based on the success of the search
end while
return Best solution found
```

- **PEPG**: Estimates the gradient of the Loss function with respect to the parameters of the distribution, and updates the parameters of this distribution.



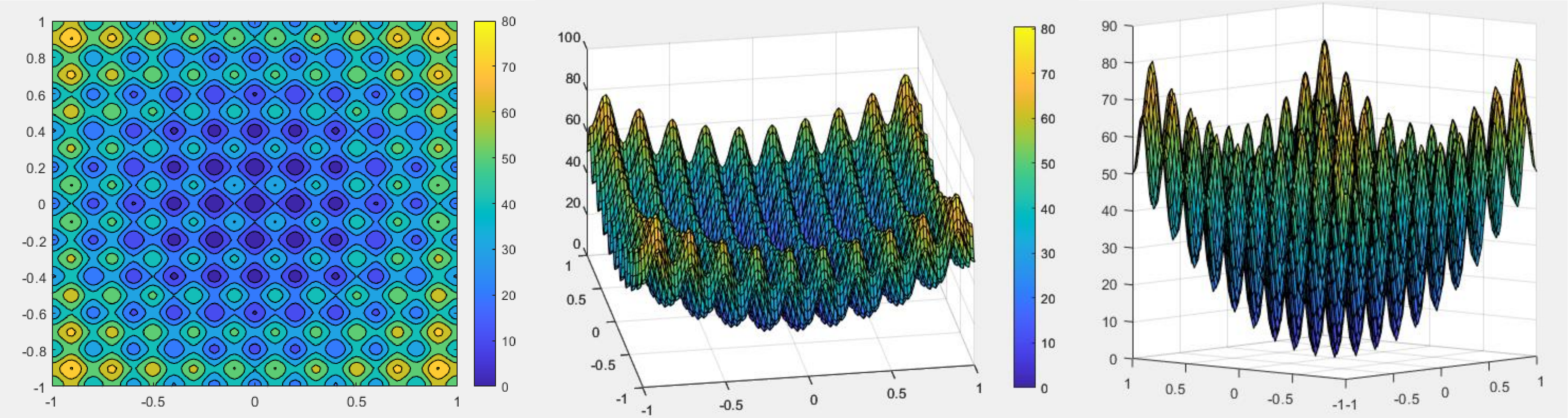
Algorithm 5 Simple Policy Gradient optimization

```
Initialize $\theta = (\mu, \sigma)$, set constant learning rate $\alpha$
$K \leftarrow$ maximum number of iterations
for $k = 1, 2, \ldots, K$ do
    Generate $\lambda$ samples $x_i \sim \mathcal{N}_\theta(x)$
    Compute $f(x_i)$ for each $x_i$
    Estimate gradient: $\nabla_\mu J(\mu) \leftarrow \frac{1}{\lambda}\sum_{i=1}^{\lambda} \frac{x_i-\mu}{\sigma^2} f(x_i)$
    Estimate gradient: $\nabla_\sigma J(\sigma) \leftarrow \frac{1}{\lambda}\sum_{i=1}^{\lambda} \frac{(x_i-\mu)^2-\sigma^2}{\sigma^3} f(x_i)$
    Update parameters: $\mu \leftarrow \mu + \alpha\nabla_\mu J(\mu)$
    Update parameters: $\sigma \leftarrow \sigma + \alpha\nabla_\sigma J(\sigma)$
end for
return Optimized parameters vector $\theta = (\mu, \sigma)$, and optimal Loss value $L(\theta)$
```

https://en.wikipedia.org/wiki/CMA-ES

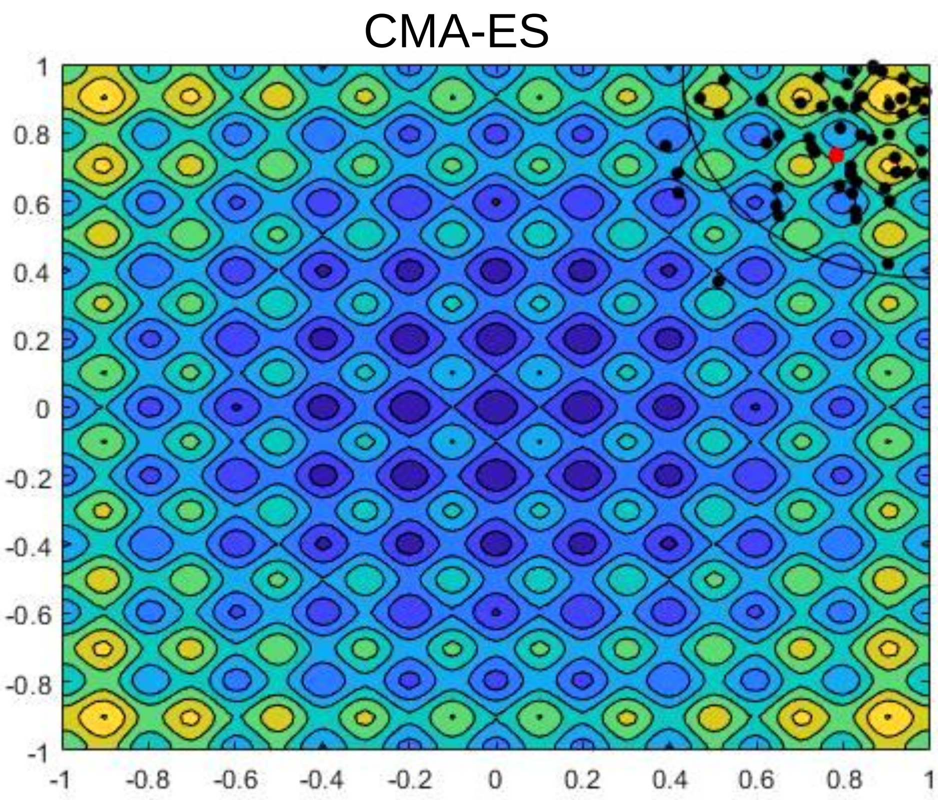
CMA-ES AND PEPG

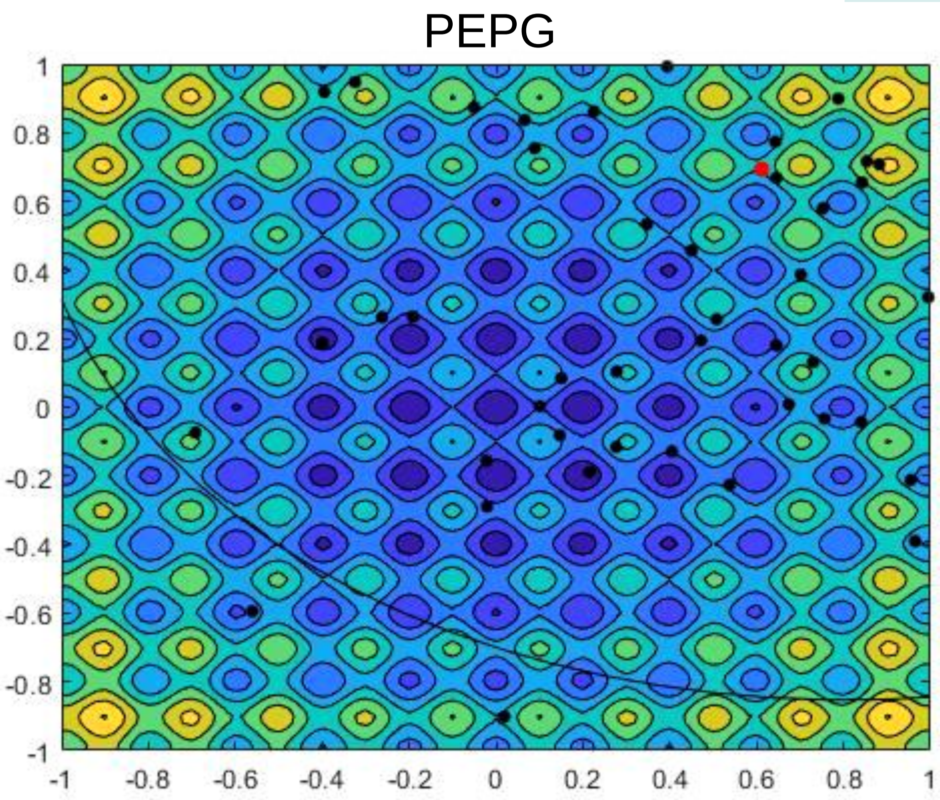
Toy optimization problem : Rastrigin function



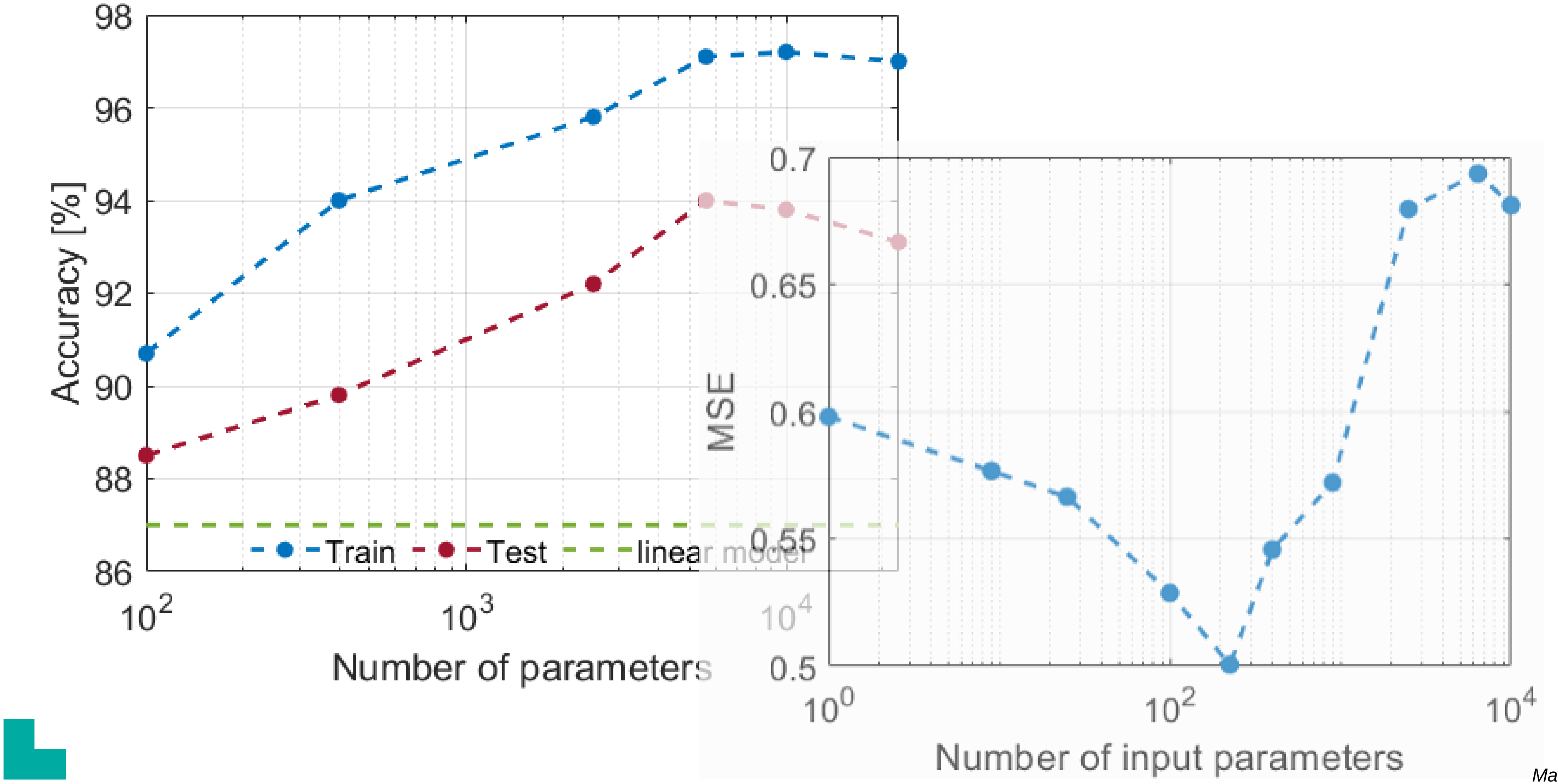
https://en.wikipedia.org/wiki/CMA-ES

CMA-ES AND PEPG

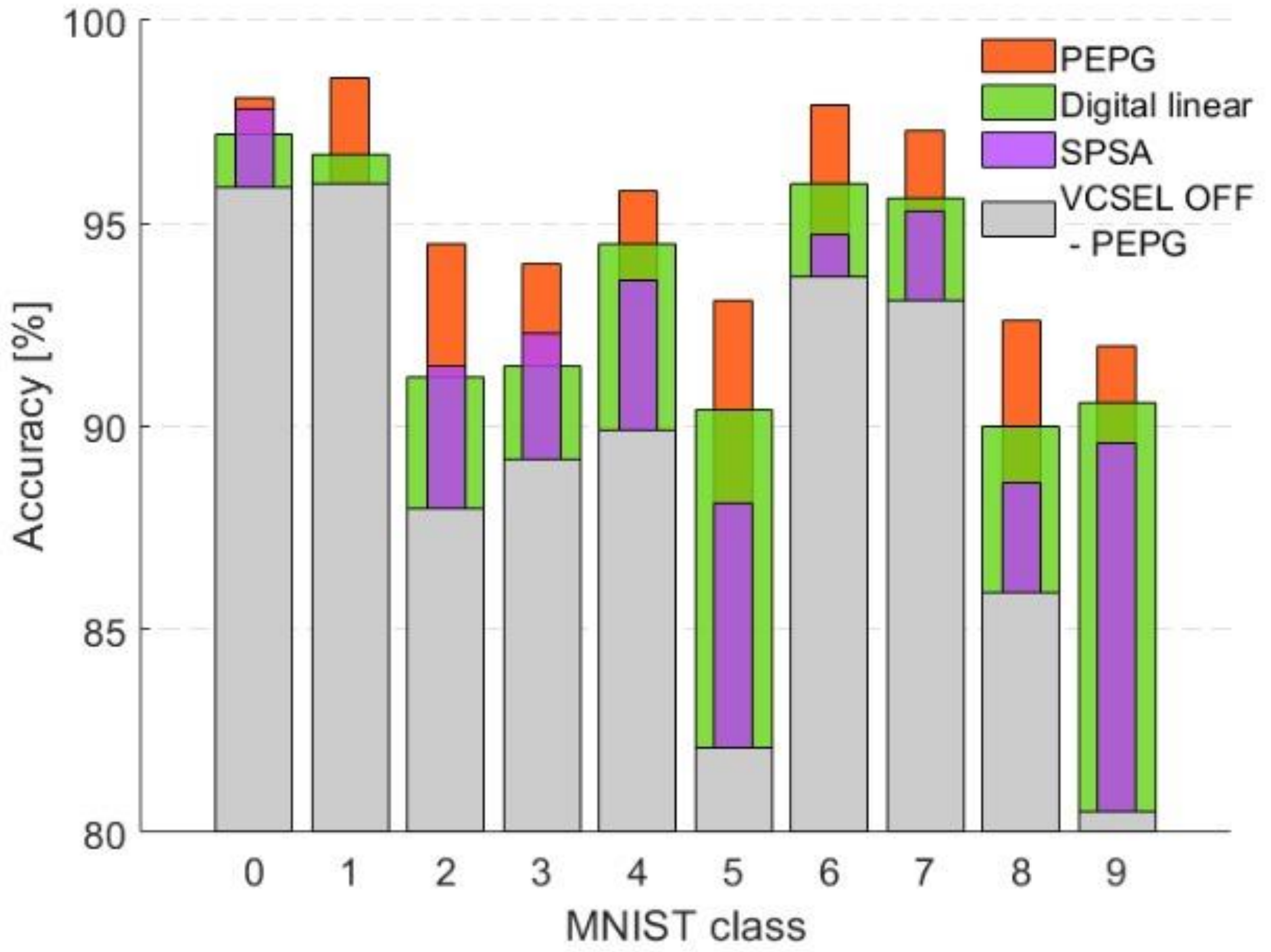




TRAINING THE WEIGHTS SPSA



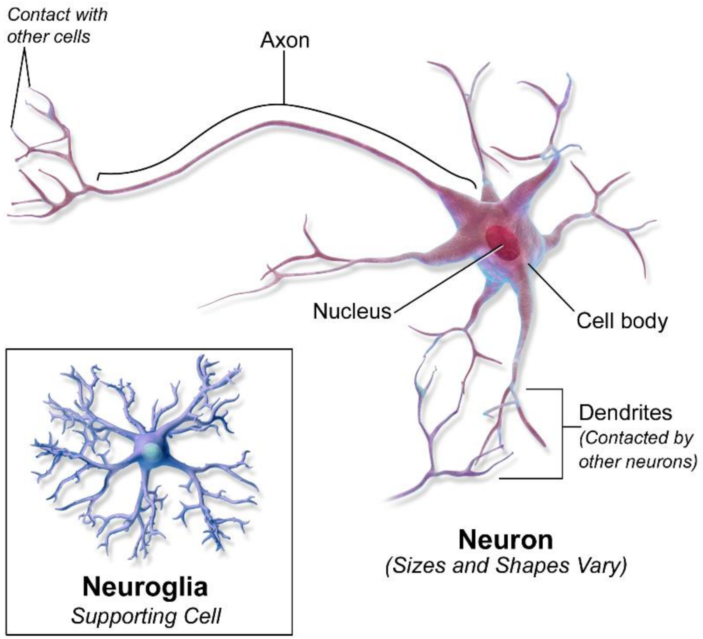
Manuscript in preparation



- 5625 Neurons
- Linear limit: ~93% - computational benefit only above that.
- Linear experiment shows how difficult in-situ learning is.
- SPSA: not sufficient to surpass linear limit.
- PEPG clearly goes beyond.
- First fully analogue neural network surpassing linear limit.

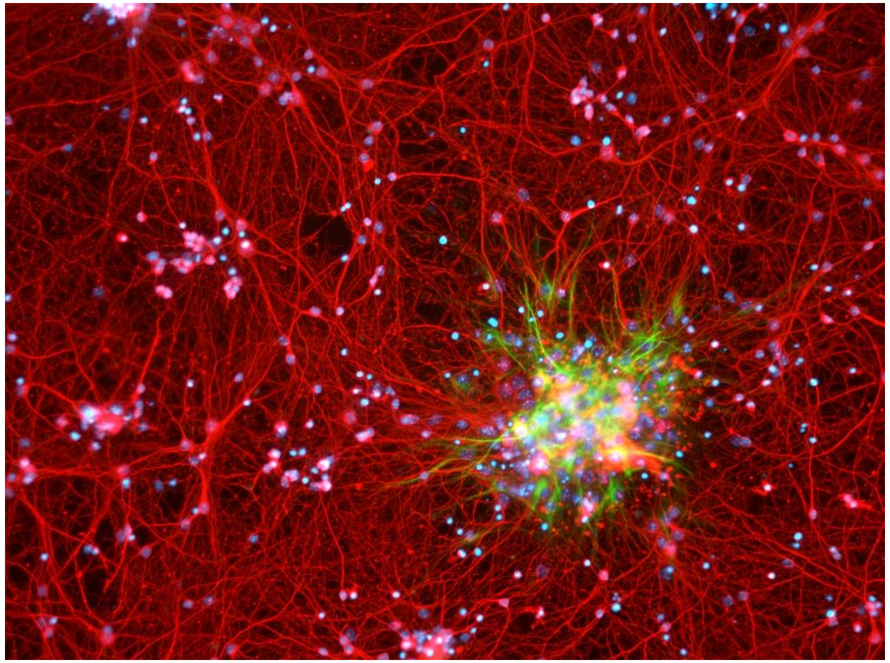
Manuscript in preparation

SCALING DIMENSION: COMPLEXITY OF OUR ‘NEURAL’ NETWORK



"Lab 1 Neurohistology - Neurons". *vanat.cvm.umn.edu.*

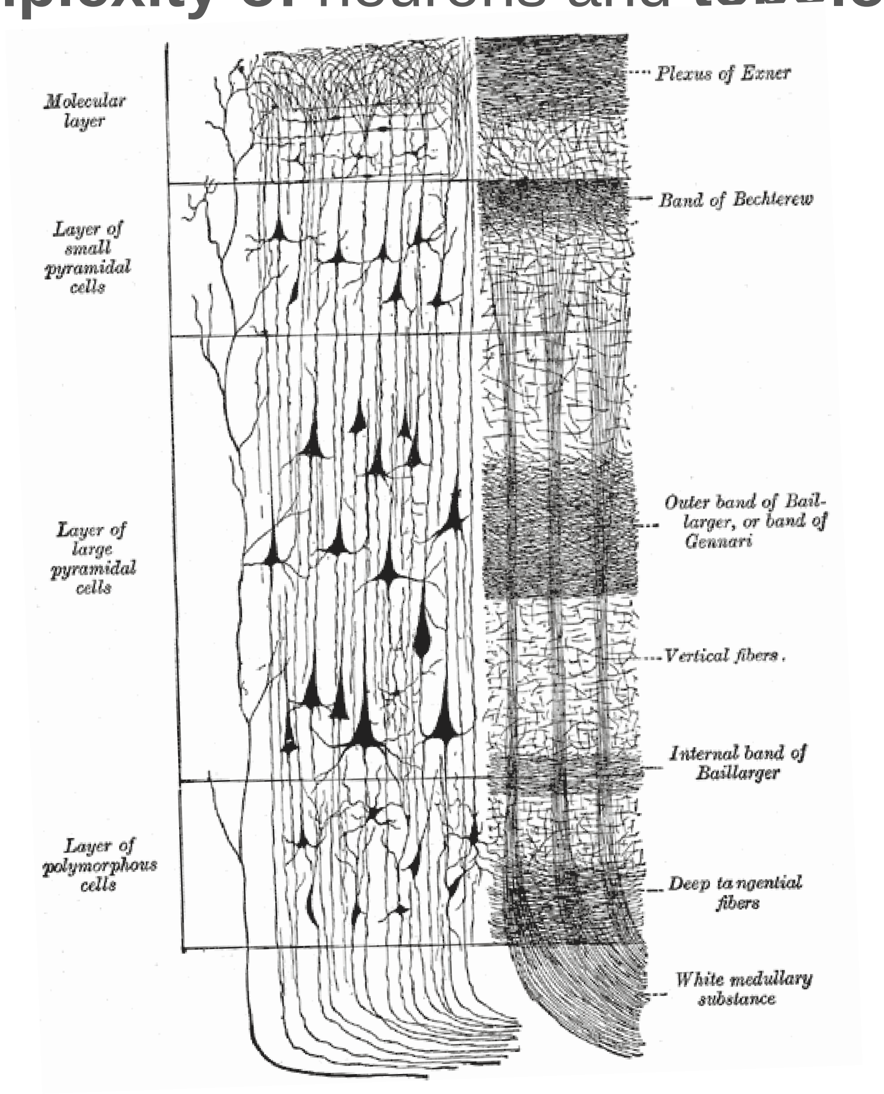
- Spiking activity
- Dendrites are nonlinear
- That makes Nonlinearity change from O(N) -> O(N^2)
  - Will break back-propagation of errors algorithm



By Dchordpdx - Own work, CC BY 4.0,

- Neurons not the only ‘elements’
- 2nd type: Glia cells, potentially more than neurons
  - Astrocites
    - Wrap around up to 1000…2000000 synapses and modulate their response

Complexity of neurons and **topologies**

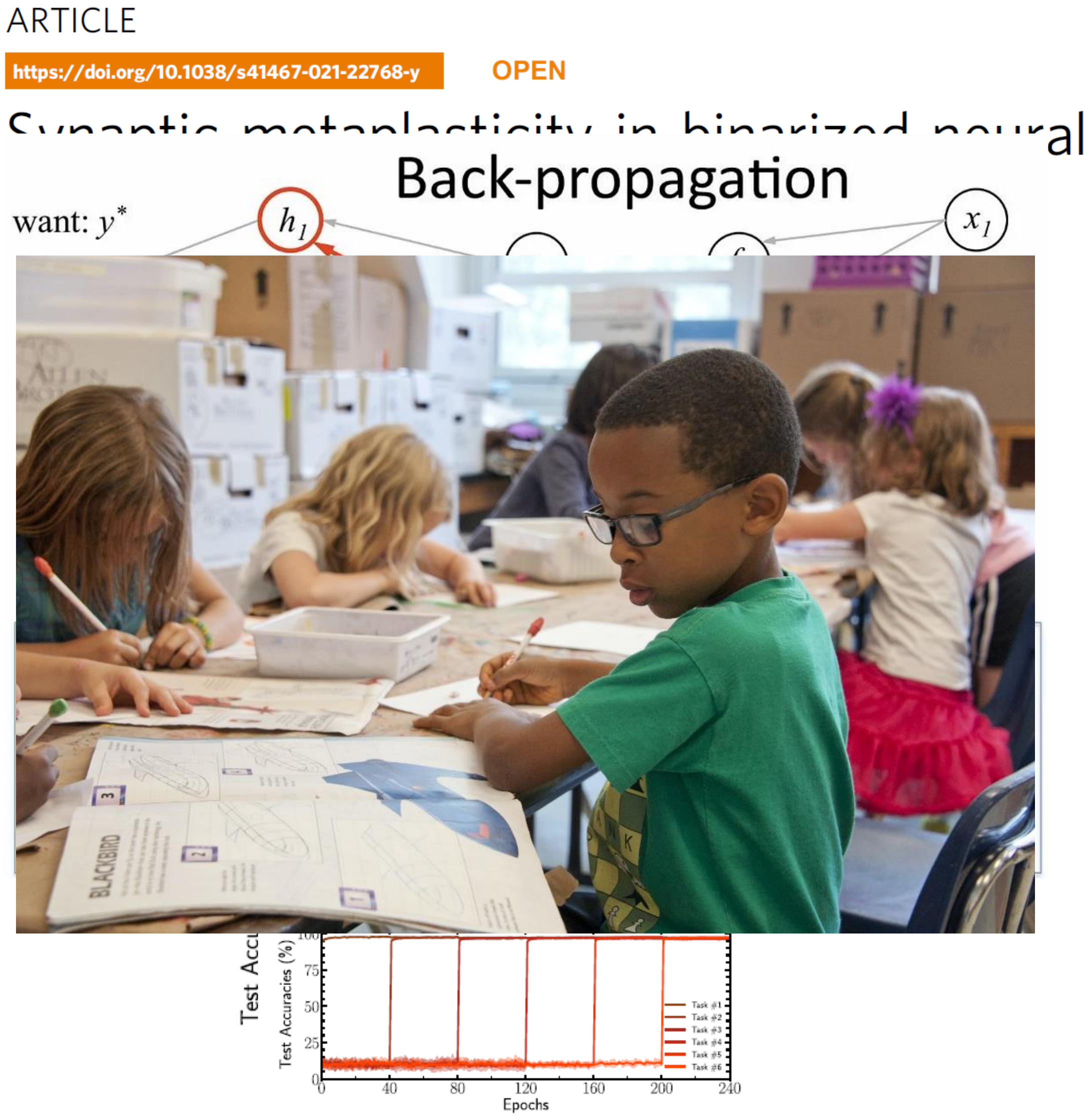

Molecular layer
Layer of small pyramidal cells
Layer of large pyramidal cells
Layer of polymorphous cells
Plexus of Exner
Band of Bechterew
Outer band of Baillarger, or band of Gennari
Vertical fibers.
Internal band of Baillarger
Deep tangential fibers
White medullary substance

Henry Vandyke Carter - Henry Gray (1918) Anatomy of the Human Body

- Neocortex comprises of 6 layers
- Layers can host various types of neurons
- Some simplified canonical topologies:
  - Layer 2: locally connects to other layers of Neocortex
  - Layer 3: connects up and down and receives most of the external input (sensory stimuli)
  - Layer 4: connect other areas out and inside of and neocortex
  - Layer 5: connects outside of cortex, i.e. motor control
  - Layer 6: connects to thalamus

Primer for beyond today: algorithm and functionality

- Learning without back prop
  - Back propagation requires storing activations and derivatives
  - How about NL weights? Can’t handle it
- Learning without forgetting
  - Train output to identify ‘3’ – now train the same to identify ‘4’ -> it will forget the ‘3’
- Tokens of a GPT vs. human learning
  - Transformers required an astonishing amount of tokens to learn, most likely orders of magnitudes more than humans
- Compositionality and systematicity
- Non-digital: NN hardware school??
- +++++++



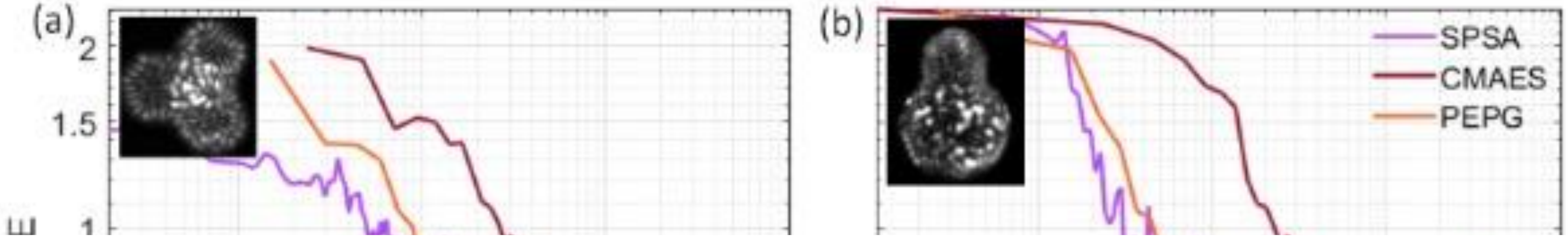
ARTICLE

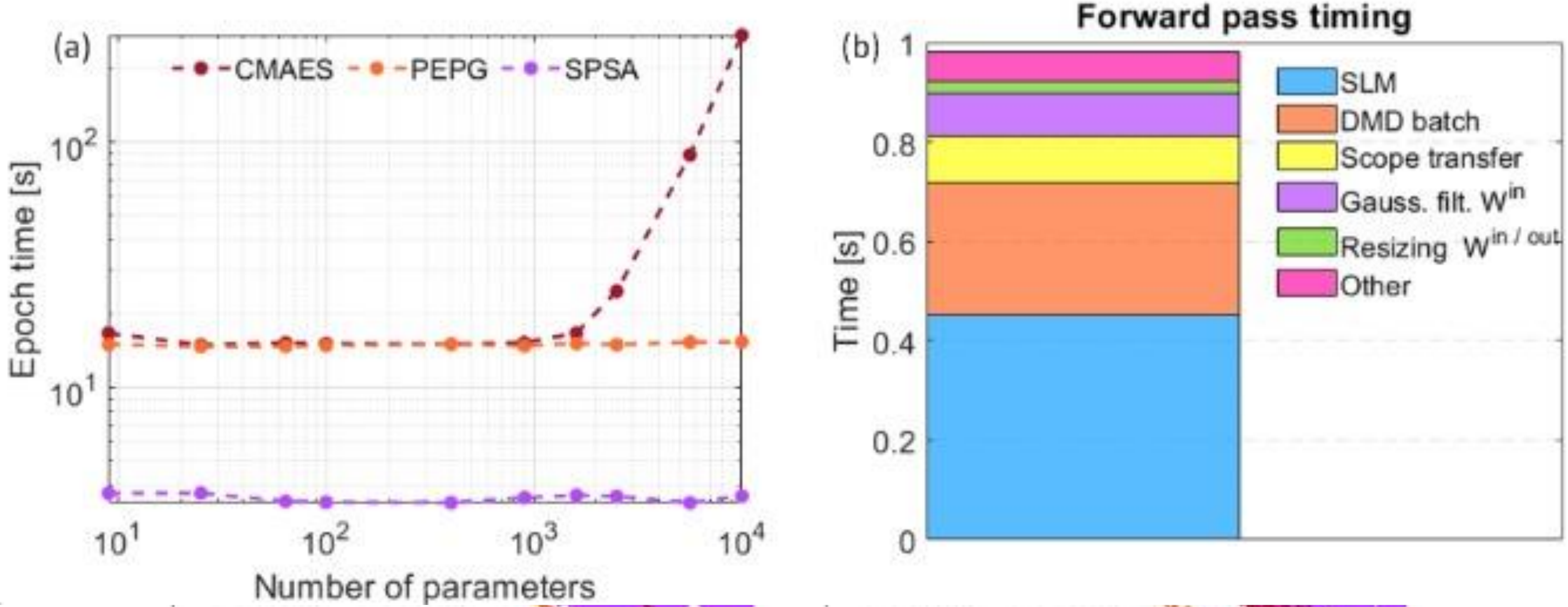
https://doi.org/10.1038/s41467-021-22768-y OPEN

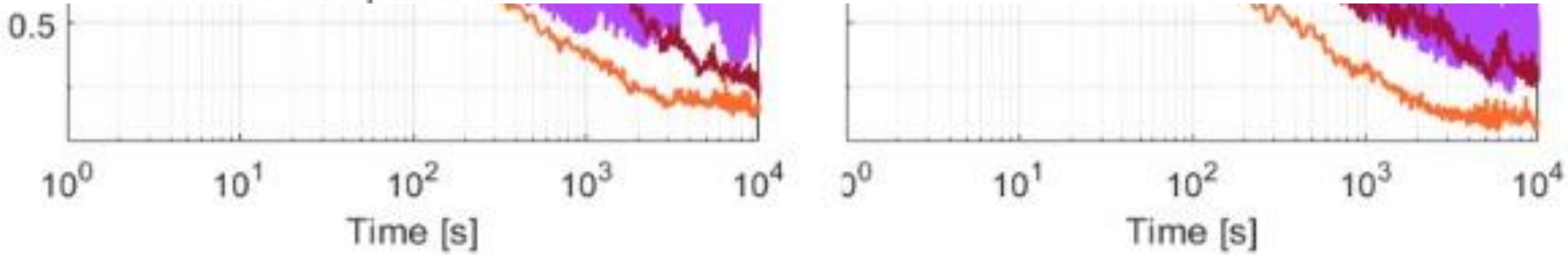
Synaptic metaplasticity in binarized neural

Summary

- **Long term field and undertaking (for me)**
  - New birds-eye view is required: systems, not components
  - (Almost) as many fundamental questions as there are questions
  - Concepts and systems need to converge
  - Most importantly: how to learn conceptually more efficient / less complex
- **Are epochs, #/o parameters and MAC/J really the right dimension?**
- **Good news**
  - Fascinating research and possibilities
  - **Personal** point of view:
    we are only scratching the surface
- **Most importantly**: NNs are here to stay – unlike quantum comp. with already demonstrated clear computational relevance.

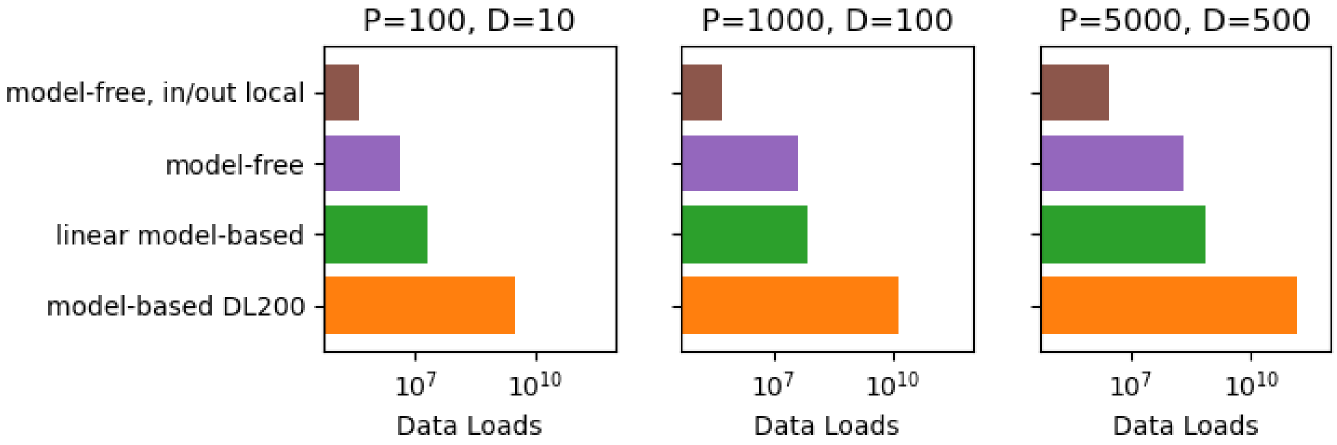






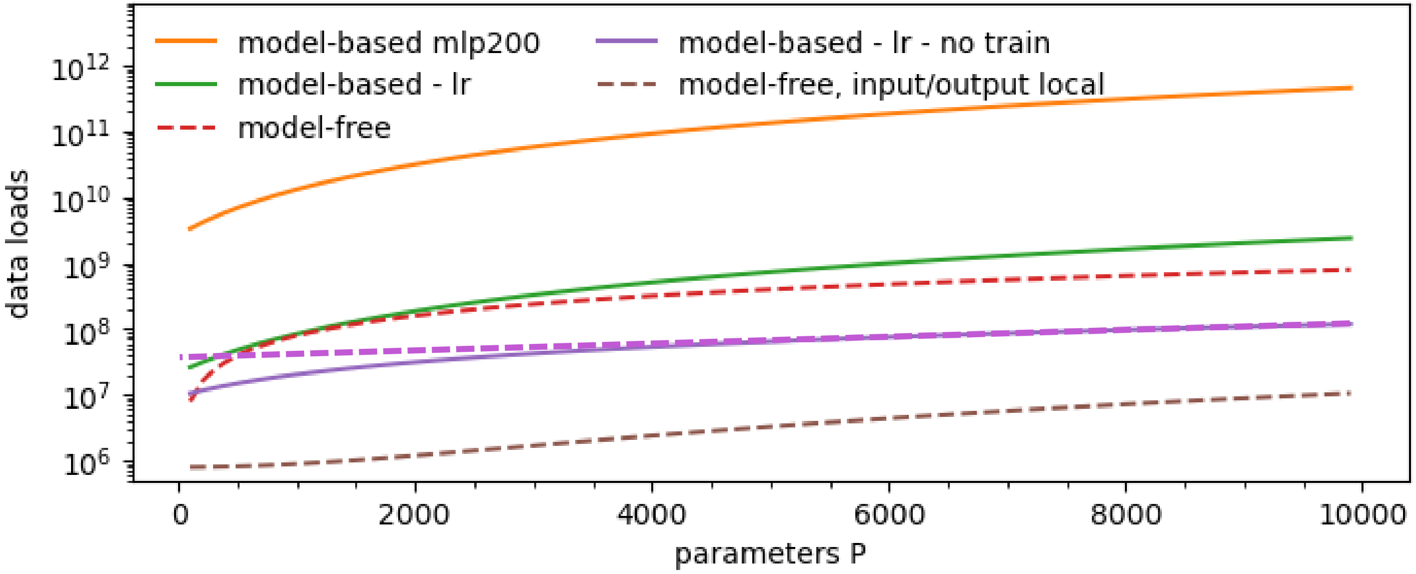
Manuscript in preparation

WHAT'S THE PROBLEM FOR MODEL BASED: VON NEUMANN AGAIN



D: population size
P: parameters neuromorphic system
K=1000 examples / batch

Training digital twin (model-based DL200) NOT INCLUDED

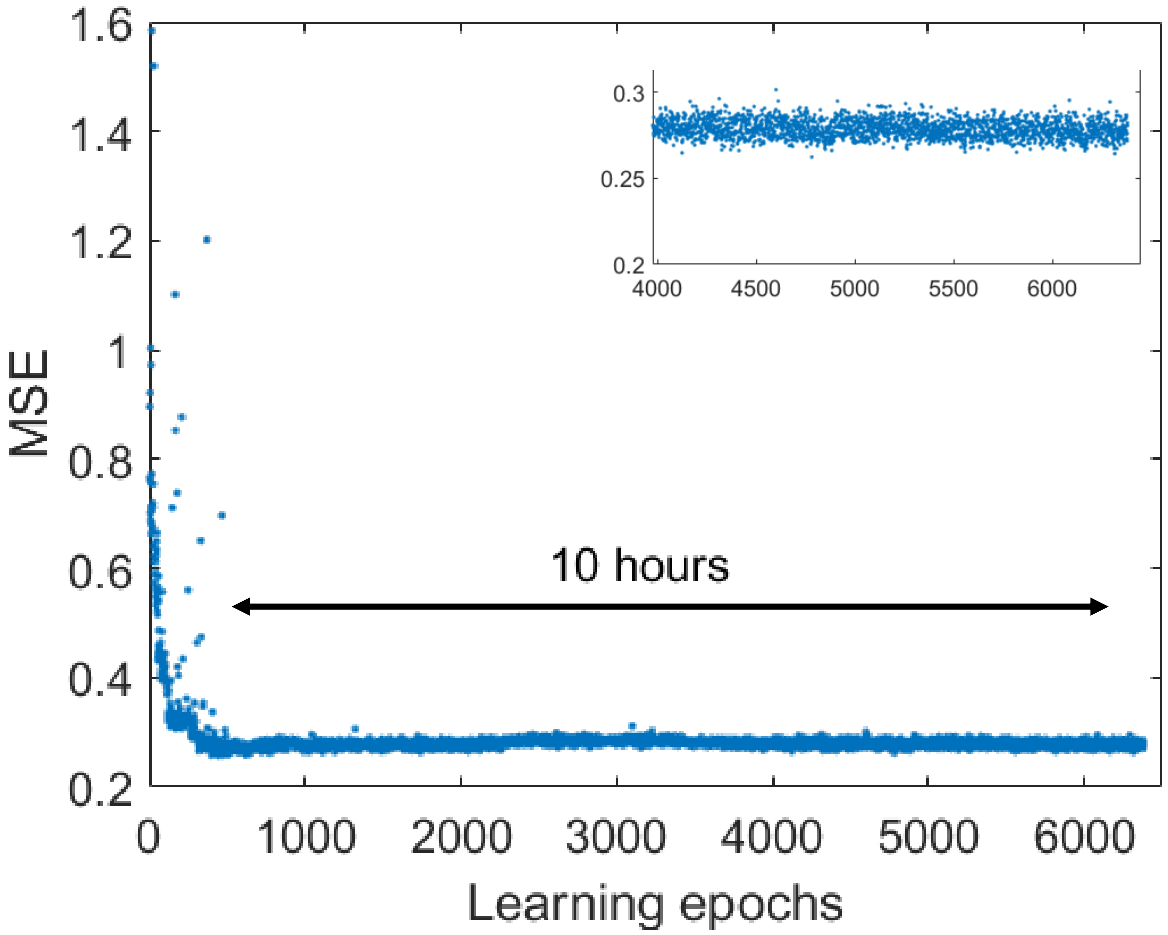


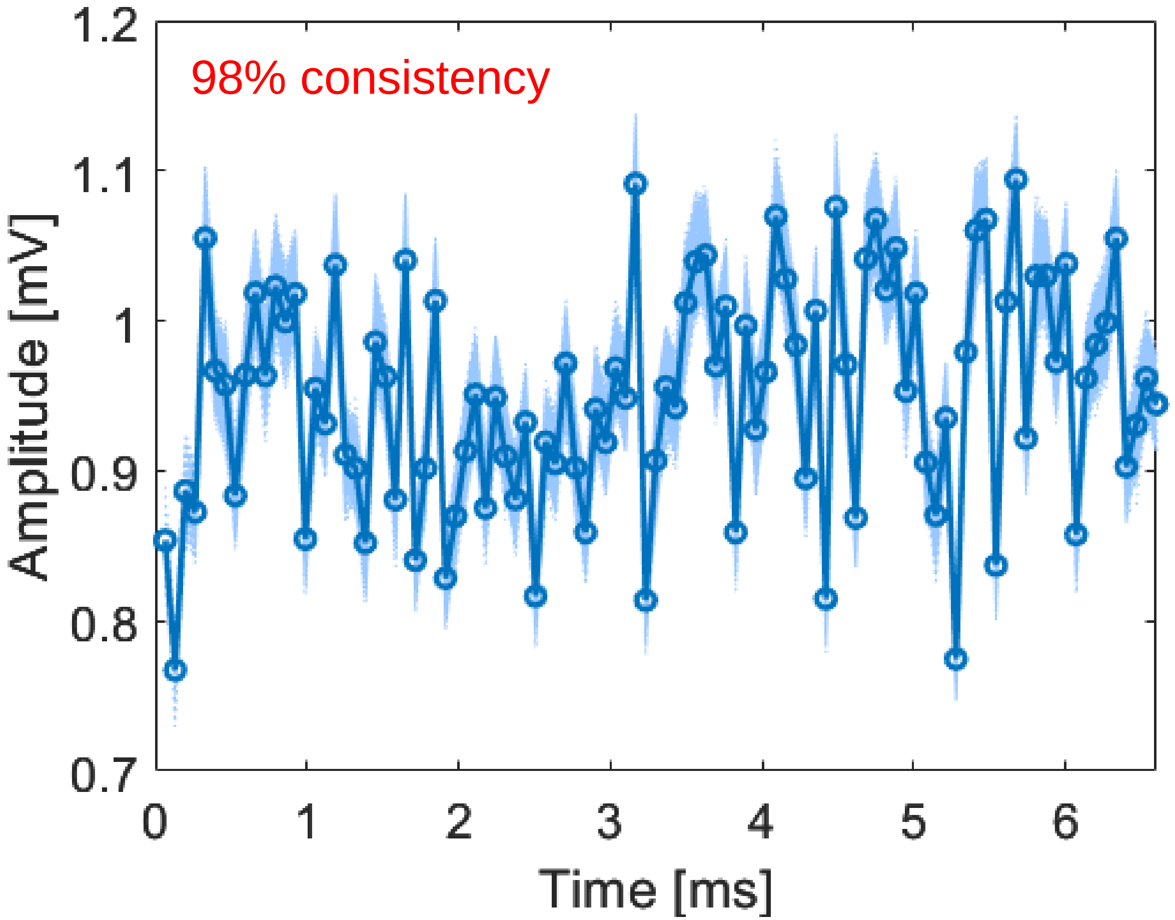
1. From model-free to model-based: 10^6 increased memory cost (at least)
2. If we assume realistic drift, then training will happen frequently

➢ **Training will most likely consume all benefits we get from photonics (or unconventional) hardware**

Manuscript in preparation

LONG TERM STABILITY





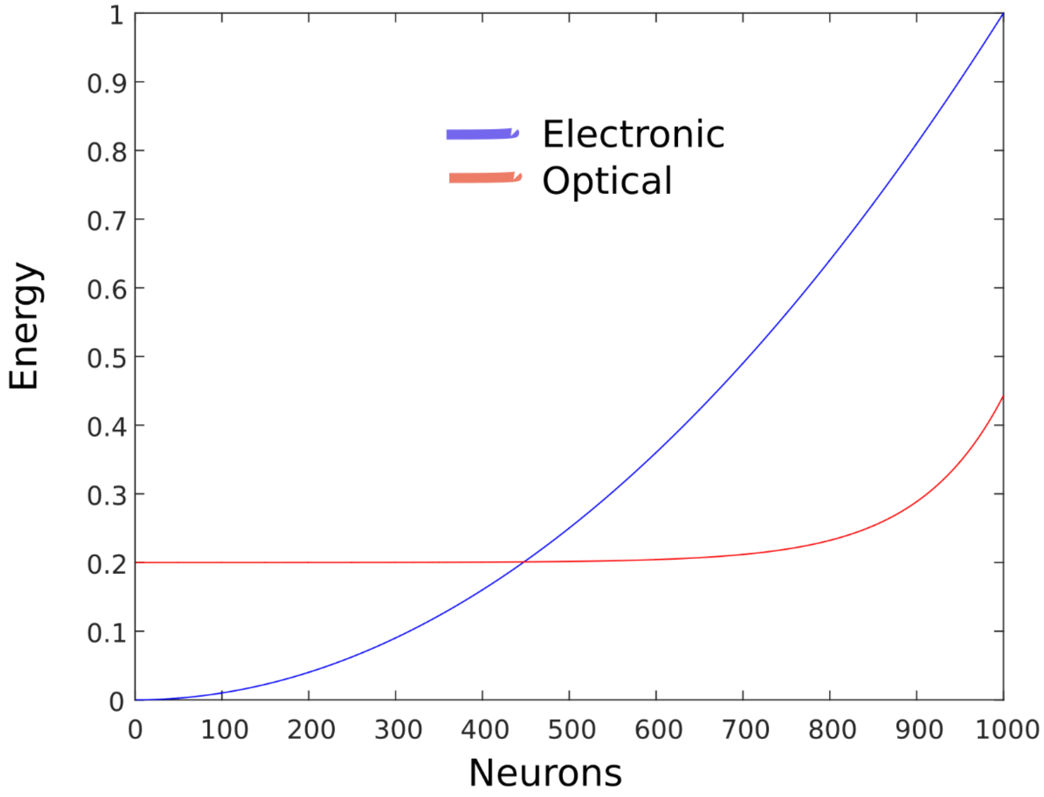
Optical vs. electronic NN communication

Cross-over point

- At which number of neurons N optical communication becomes more efficient?
- Factors: losses $\gamma\ \left[\frac{1}{\mathrm{m}}\right]$ and electro <-> optical conversion E^{EO}.
  - Electronic: $E \propto w^2(N + \frac{1}{2}N^2)$
  - Optical: $E \propto e^{Nw} + E^{\mathrm{EO}}$

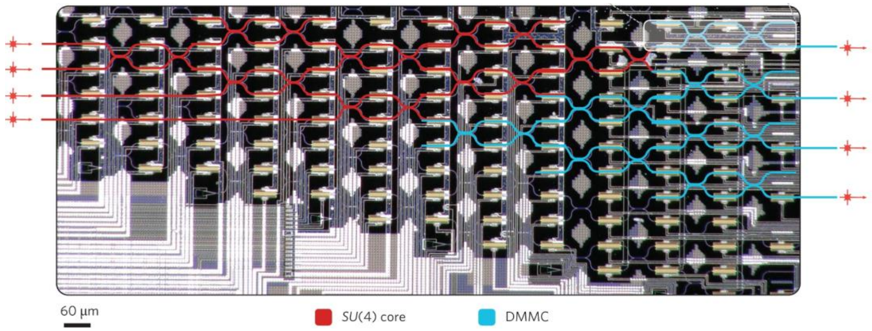
Of fundamental importance:

- Optical losses are not a natural constant -> engineering problem! (no absolut lower limit)
- EO conversion is a ‘point’ problem -> **not dependent on N**

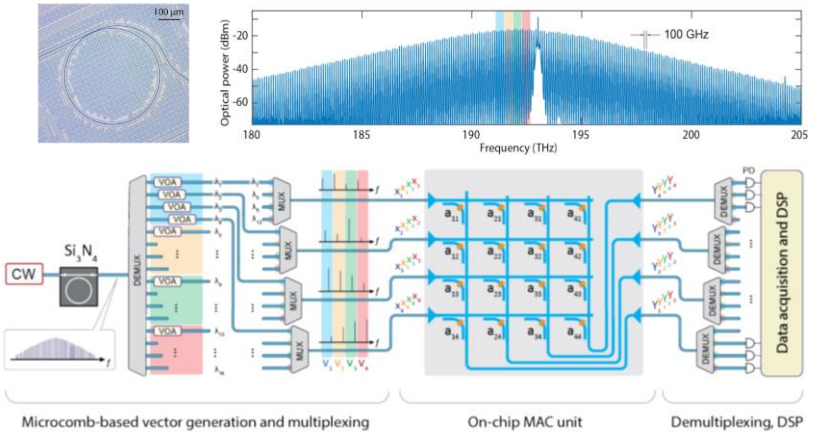


SCIENTIFIC MOTIVATION: OPTICAL NEURAL NETWORKS STATE-OF-THE-ART

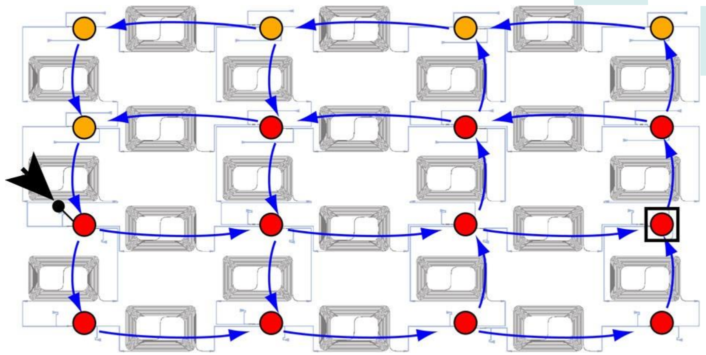
Matrix-vector multiplier with coherent photonics [1]



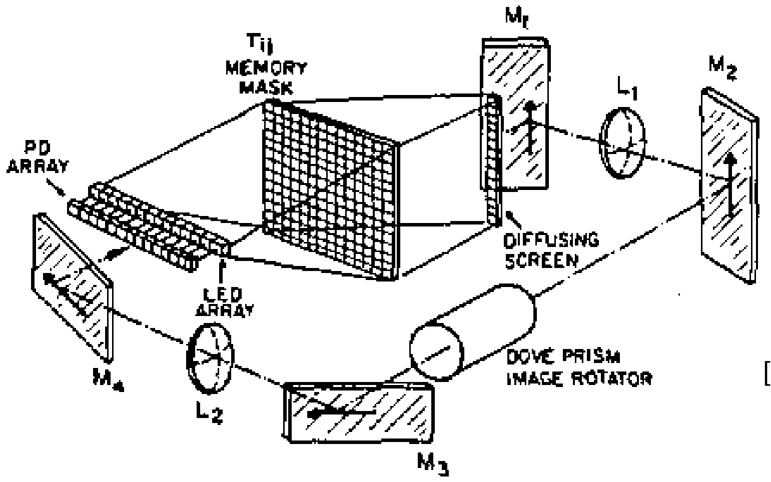
Parallel convolutions using WDM [2]



Reservoir computing on silicon chips [3]



Optical vector-matrix multiplier with optical feedback [4]



[1] Y. Shen et al., *Nature Photonics* **93**, 441-446 (2017)
[2] Feldmann, J. et al. *Nature* **589**, 52–58 (2021)
[3] Vandoorne, K. et al. *Nat Commun* **5**, 3541 (2014)

[4] Psaltis, Demetri et al *Optics letters* **10** 2 ,98-100 (1985)

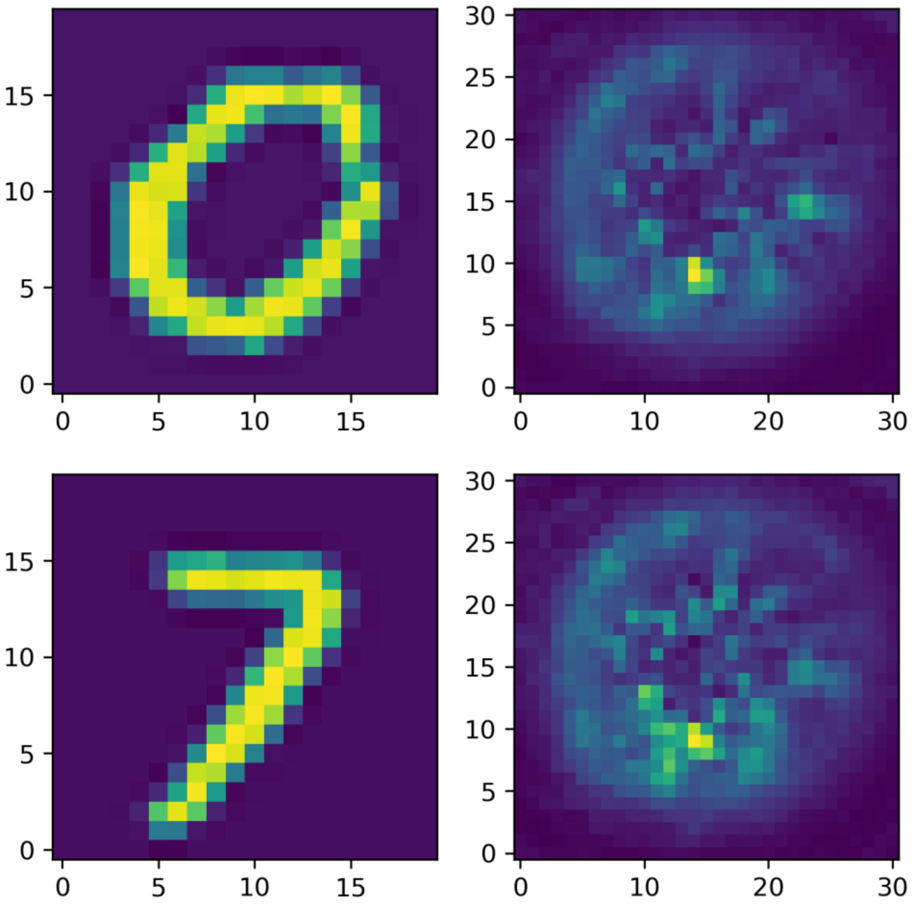
COMMON TRAINING METHODS

Training Mechanism	**Advantages**	**Disadvantages**
Reservoir Computing	- easy to implement	- does not exploit the full potential of the hardware
BP with backward pass in a digital twin [1]	- allows to train the full system	- needs a precise model or digital twin, big overhead
Augmented Direct Feedback Alignment [2]	- allows to train the full system	- partial knowledge of the model is needed - gradients have a random direction, reduced performance
Forward-Forward training [3]	- allows to train deep models layerwise	- needs a precise model or digital twin
Equilibrium propagation [4]	- performs backprop based on system dynamics	- energy function of the system needs to match certain criteria

Having a potentially resource heavy model is **counterproductive, setup is not autonomous.**

1. Wright, L. G., et al.. (2022). Deep physical neural networks trained with backpropagation. *Nature*, *601*(7894), 549-555.
2. Nakajima et al. (2022). Physical deep learning with biologically inspired training method: gradient-free approach for physical hardware. *Nature Communications*, *13*(1), 7847.
3. Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*.
4. Scellier, B., & Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, *11*, 24.

OFFLINE TRAINING



System	Performance
NN 100 training Wout +	60%
NN 100 training Wout +/-	87%
NN 100 training Wout and Win	97%
VCSEL offline Wout training full res	88%