SAL
SILICON AUSTRIA LABS

VLSI FOR MACHINE LEARNING

PEDRO JULIAN, DIEGO GIGENA, NICOLÁS RODRIGUEZ

CAE 2023, CORDOBA

ORGANIZATION

SAL
SILICON AUSTRIA LABS

- 1. Introduction
- 2. Basic elements of DNN
  - Neurons, layers, Convolution, Relu, MaxPool, etc.
- 3. Introduction to Pytorch
  - Tutorial: Mnist with 2 layers.
- 4. Case study: Tutorial and example
  - Task 1.1: You will have to design a multilayer NN using the tools (dataset provided, but your design)
- 5. VLSI realizations
  - Architectures, quantization
  - Basic elements (multipliers, adders, registers, etc)
  - Task 1.2: quantify size of designed multilayer NN
- 6. ML architectures (depth wise, Resnet, LSTM, transformers)
- 7. Review of case study
  - Task 1.3: Given an available area, e.g. 2mm^2, estimate resources you can build and the time to execute the multilayer NN

	Lunes	Martes	Mier.
9-10:40	Introduction - Pedro	Práctica	Práctica
11-13	Basic blocks of DNN -Nico	Caso de studio - Diego	Práctica
14-15:40	Intro Pytorch - Diego	Implementaciones en VLSI - Nico	Revisión de proyectos - ALL
16-18	Práctica	Práctica	Advanced architectures - ALL

SAL
SILICON AUSTRIA LABS

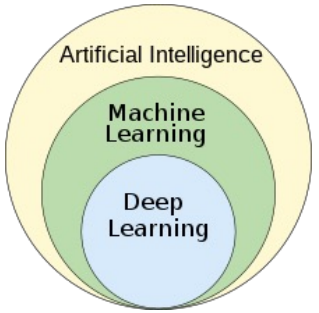
1. INTRO

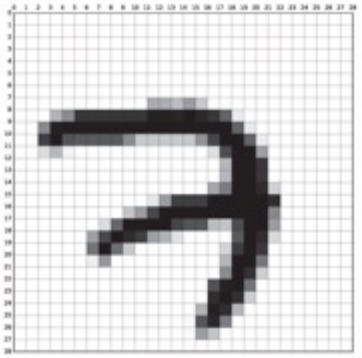
SAL
SILICON AUSTRIA LABS

MACHINE LEARNING

SAL
SILICON AUSTRIA LABS

- **Machine learning** (**ML**) is a field of inquiry devoted to understanding and building methods that "learn" – that is, methods that leverage data to improve performance on some set of tasks.[1] It is seen as a part of artificial intelligence.





Input

Model

Output

7

prediction

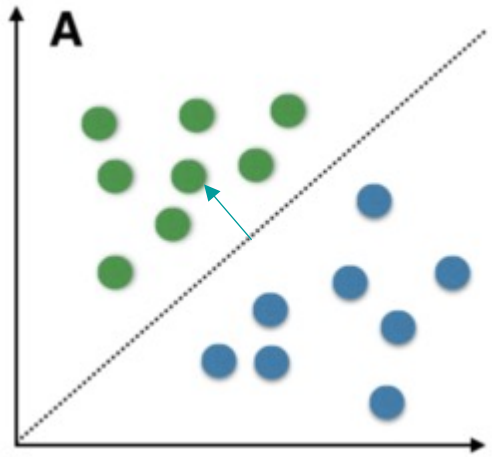
Samples

METHODS

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

SUPERVISED LEARNING

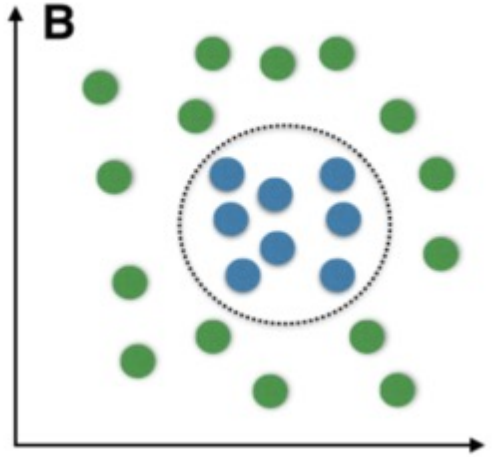
- Training set has a set of input data and the desired output
- Example: Support vector machine



$Y=ax+b$

Model

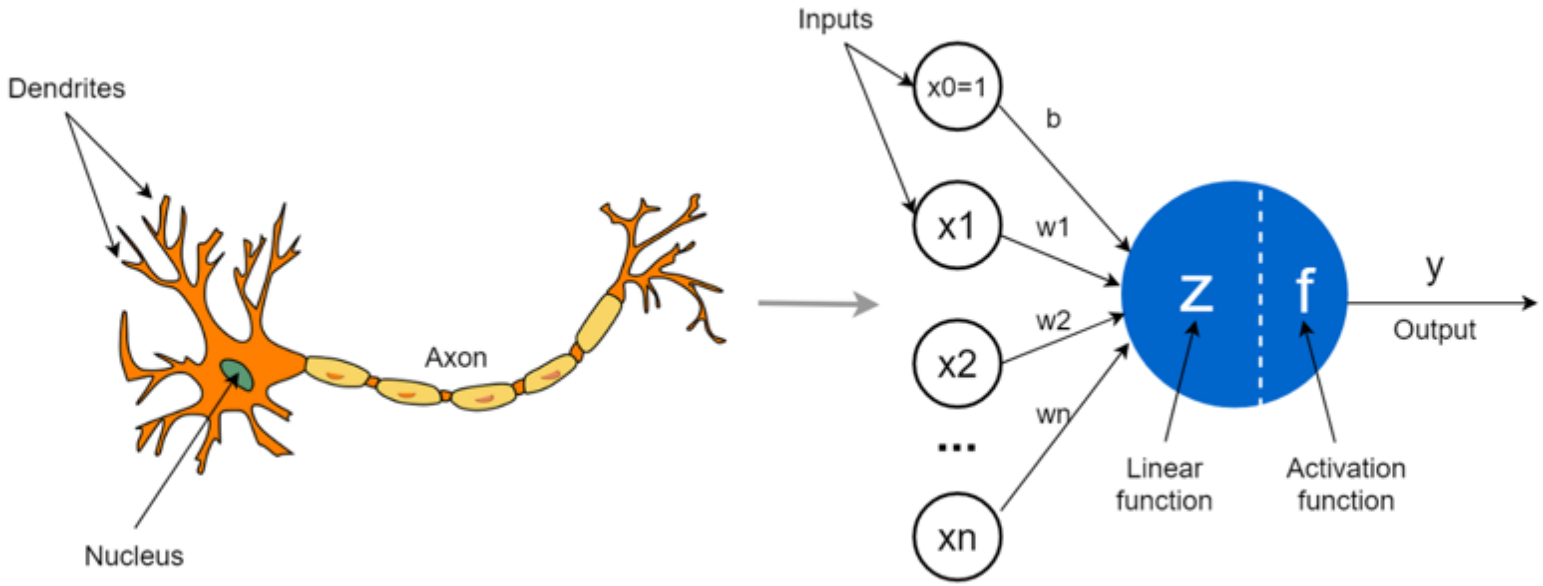
- Change a and b, so that for all input data, the output classes are on the right side of the line



Different models

SUPERVISED LEARNING: ANN

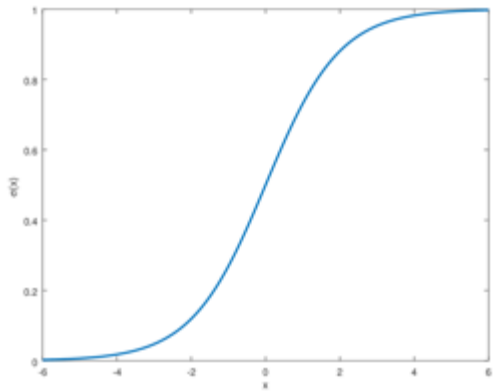
- Artificial Neural Networks:
  - Single Neuron



Single Neuron:

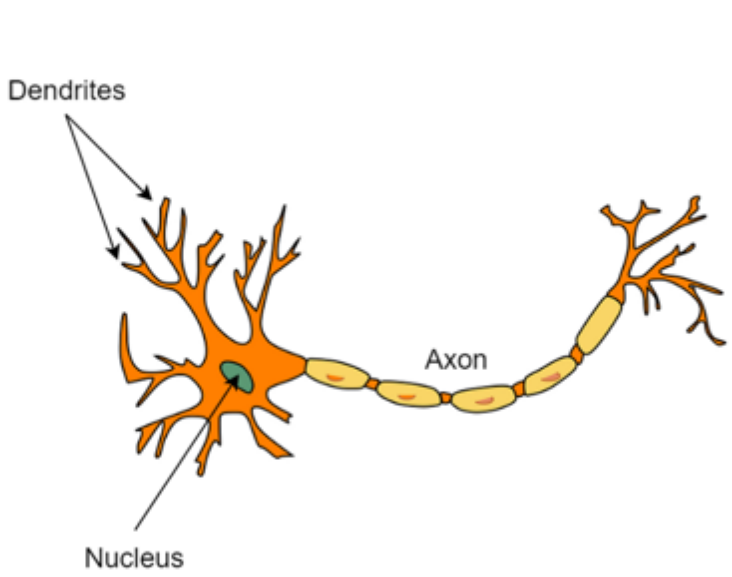
Model

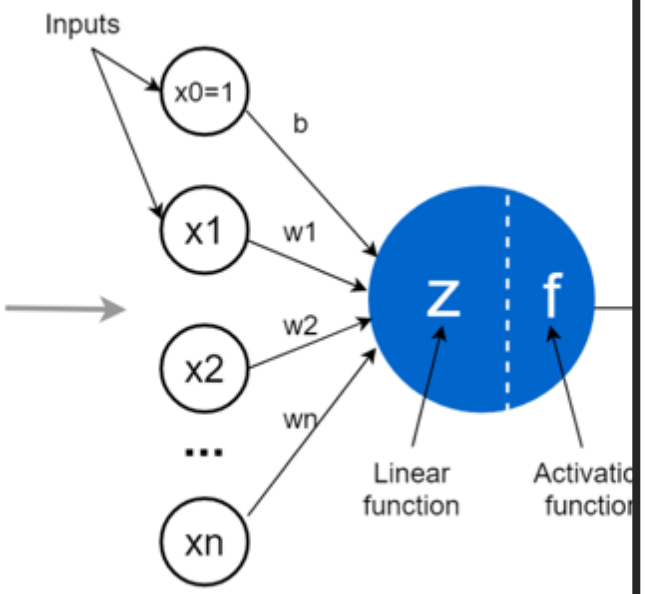
$$y = f(\sum_i w_i x_i + b)$$

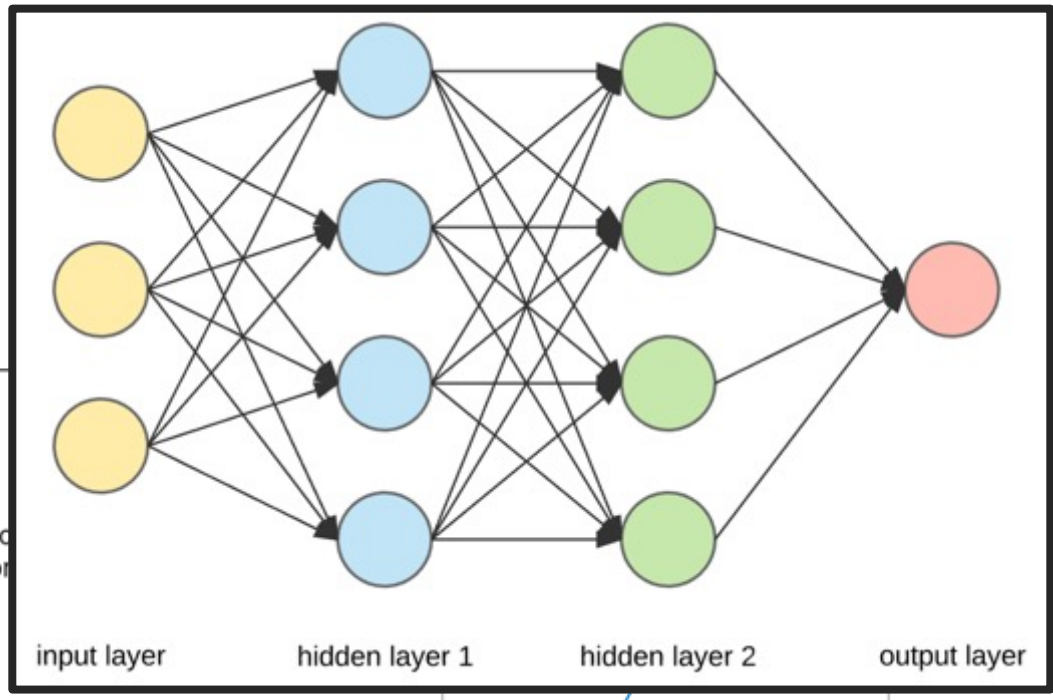


SUPERVISED LEARNING: ANN

- Artificial Neural Networks:
  - Single Neuron





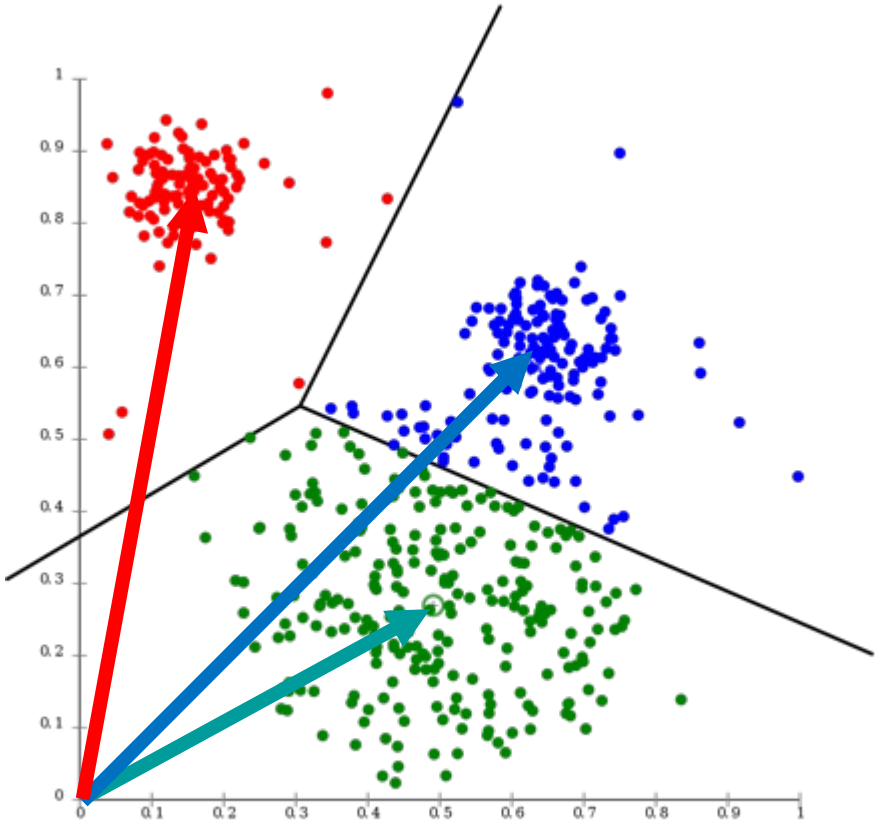


UNSUPERVISED LEARNING

- No tagged desired output:
- Model extracts patterns
- Example: Clustering

K-means clustering

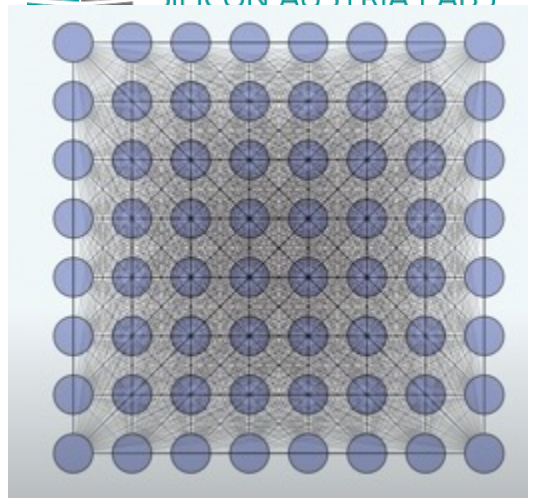
Model:
- Assign k centroids
- Cluster elements so the distance is minimized



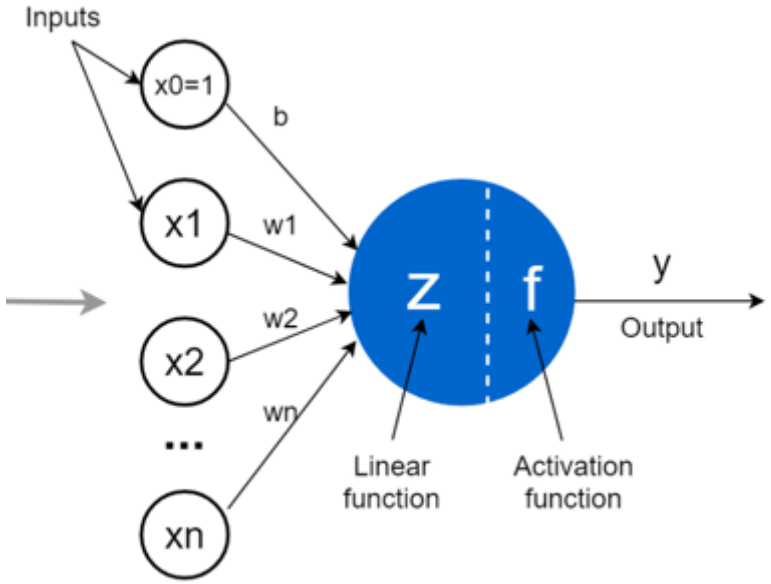
UNSUPERVISED LEARNING

- No tagged desired output:
- Model extracts patterns
- Example: Hebbian Learning

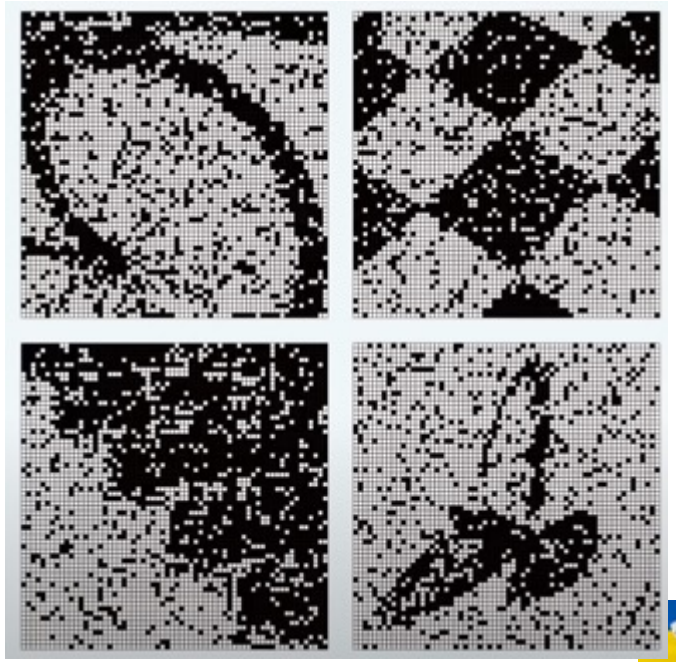
Example: Hopfield Network



Hebbian Learning



Model:
- ANN
- Weight wi is increased if input xi and output are “active” together

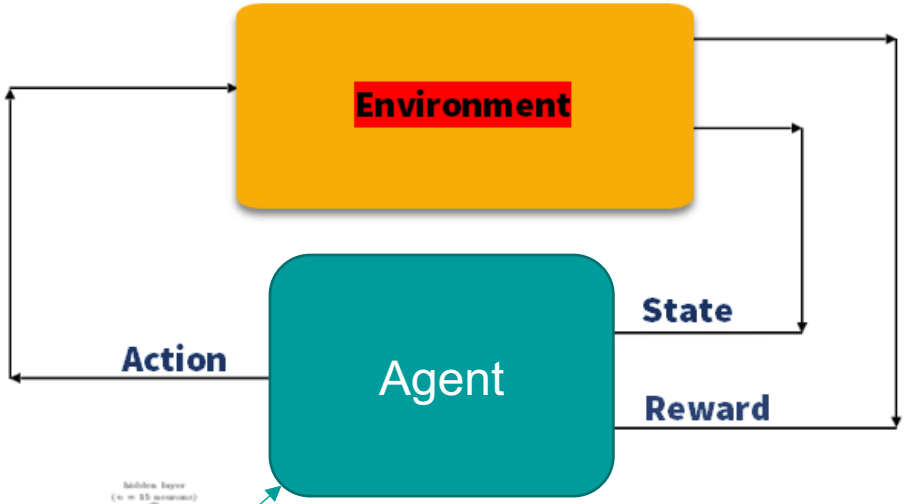


https://youtu.be/piF6D6CQxUw?t=831

REINFORCEMENT LEARNING

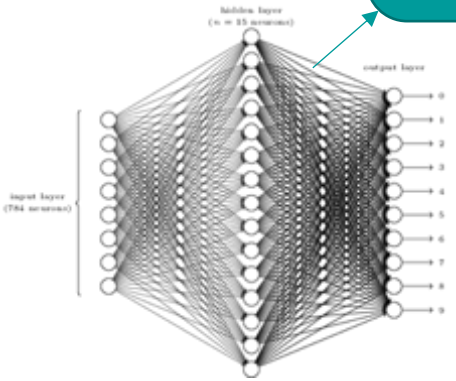
- Agent interacts with environment
- Long term reward

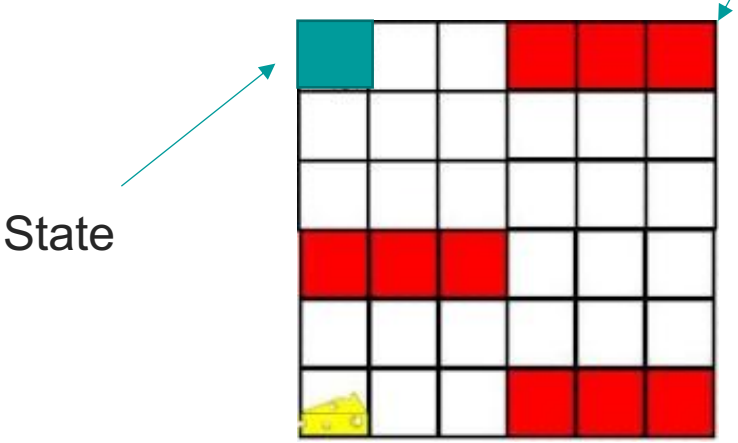
Weak supervision



Model:

- agent learns to perform a task through repeated trial and error interactions with a dynamic environment
    - Another example: chess



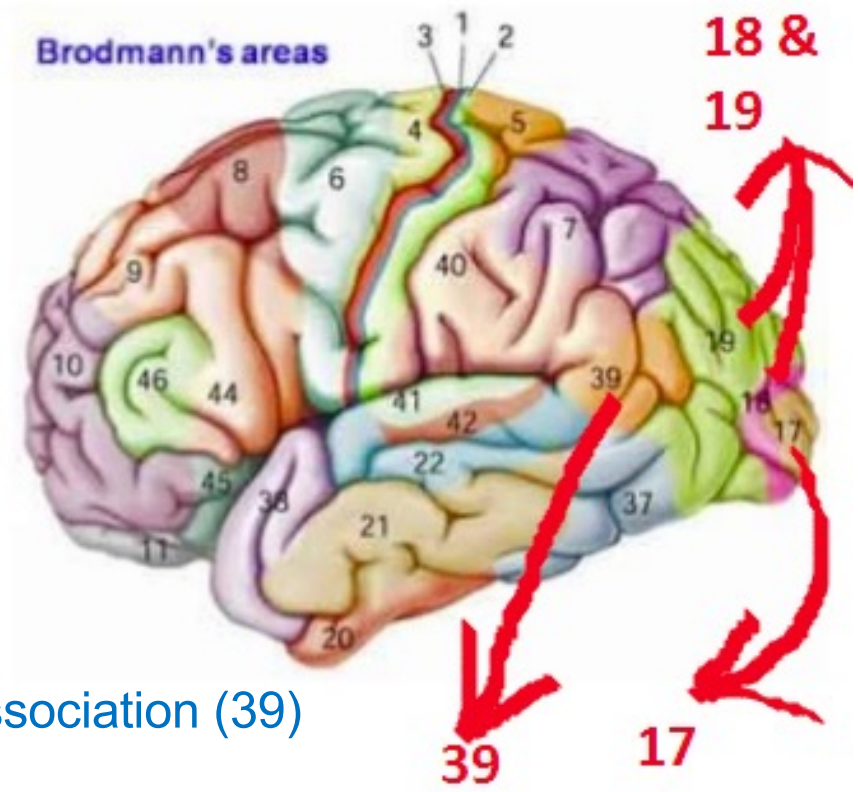


NEURONS

- How nature operates?

LESSONS FROM BIOLOGY

Visual association (18&19)



Higher association (39)

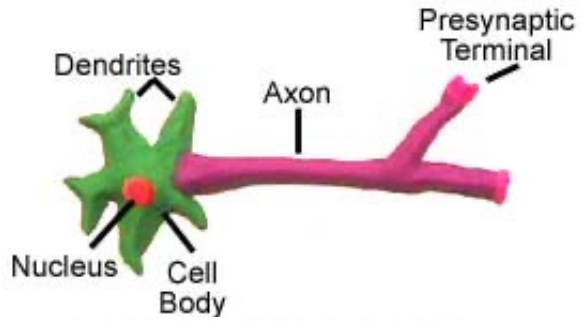
Primary area (17)

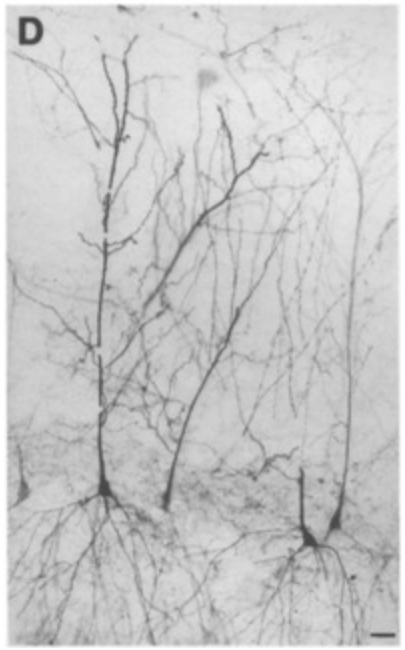
- 100 x 10^9 neurons
- 1000 x 10^{12} connections
- Frequency: 10Hz
- Power: 20W
- Efficiency: 16mW/cm^3

CELLULAR BIO-PROCESSING

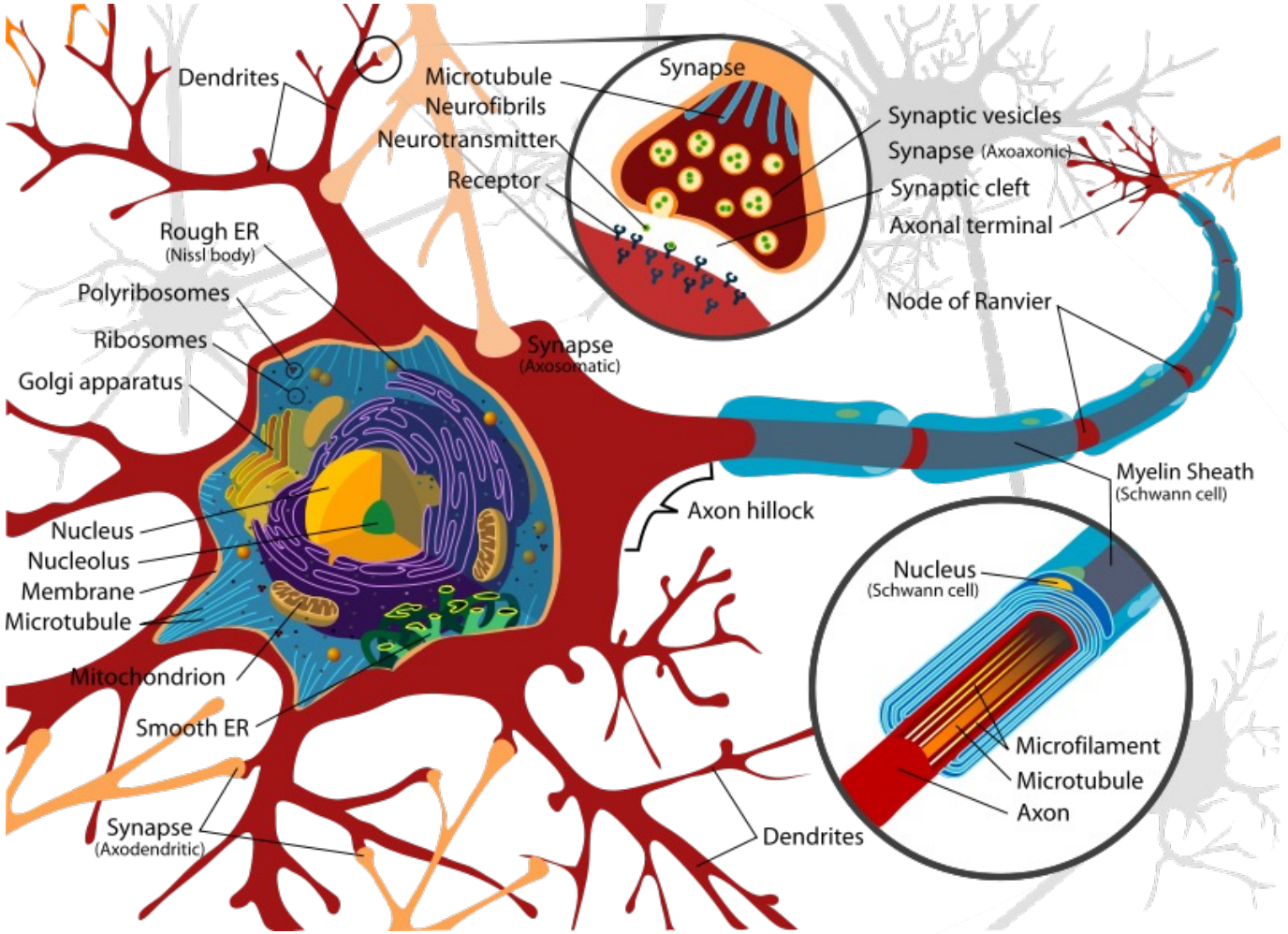
- Local communication
- Large collection of cells (Massive parallel processing)
- Low speed
- Balance between complexity vs. area (volumen)
  - 3D ...

BIOLOGICAL NEURONS



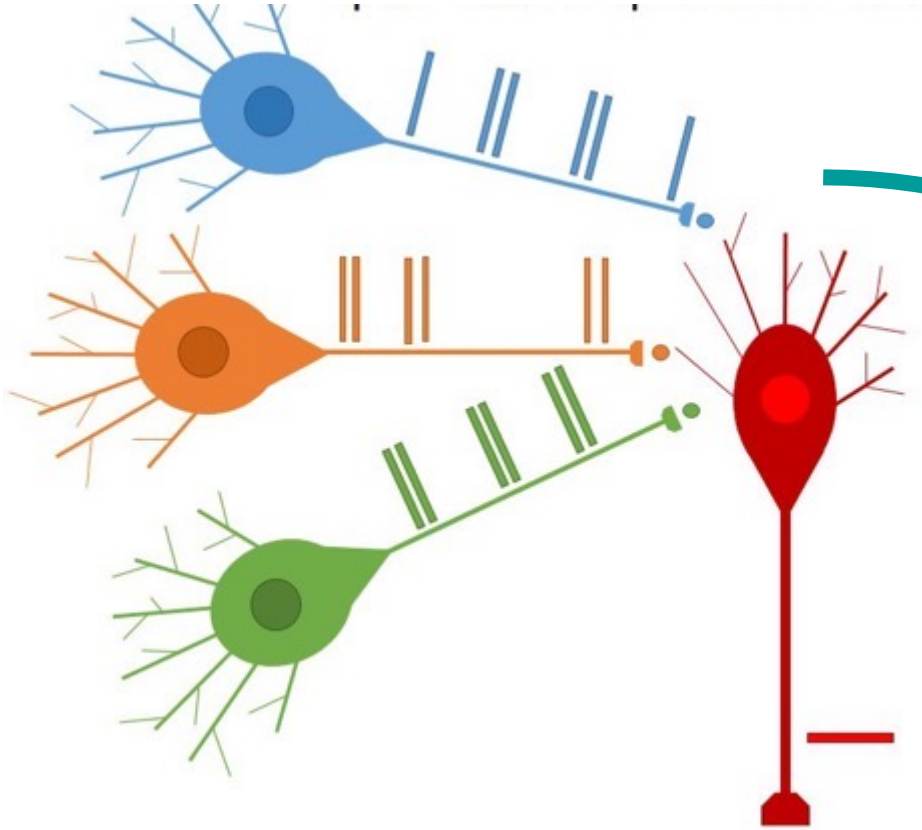


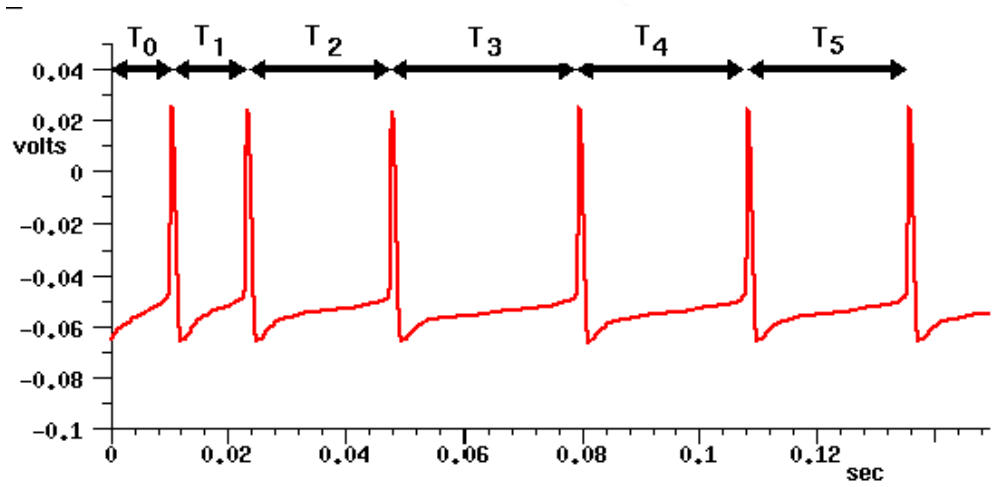
10um



Megias et al. Neuroscience, 2001

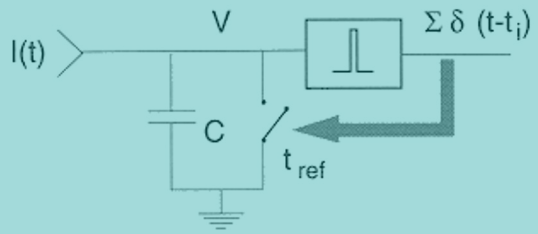
NEURONS




T0
T1
T2
T3
T4
T5
0.04
0.02
volts
0
-0.02
-0.04
-0.06
-0.08
-0.1
0
0.02
0.04
0.06
0.08
0.1
0.12
sec

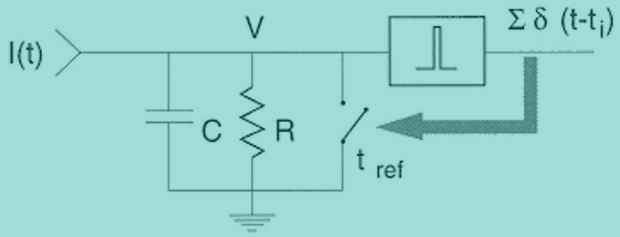
MODELS

Perfect Integrate-and-Fire Unit



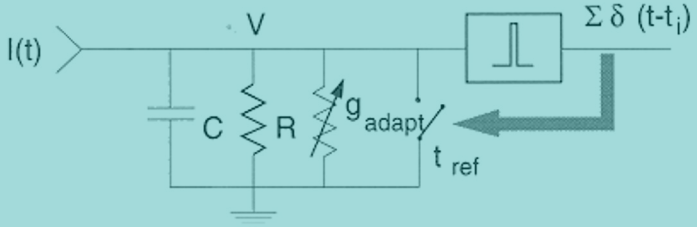
B)

Leaky Integrate-and-Fire Unit



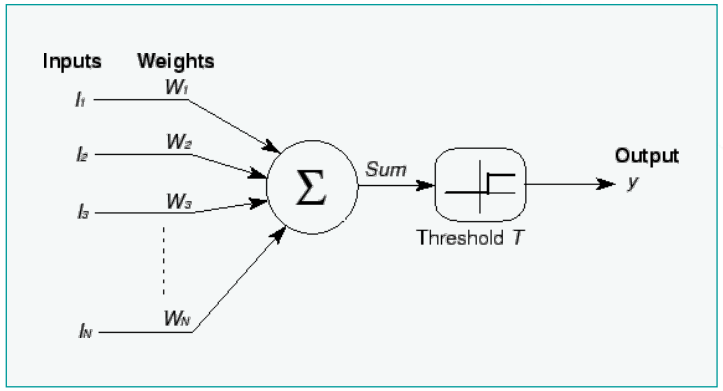
C)

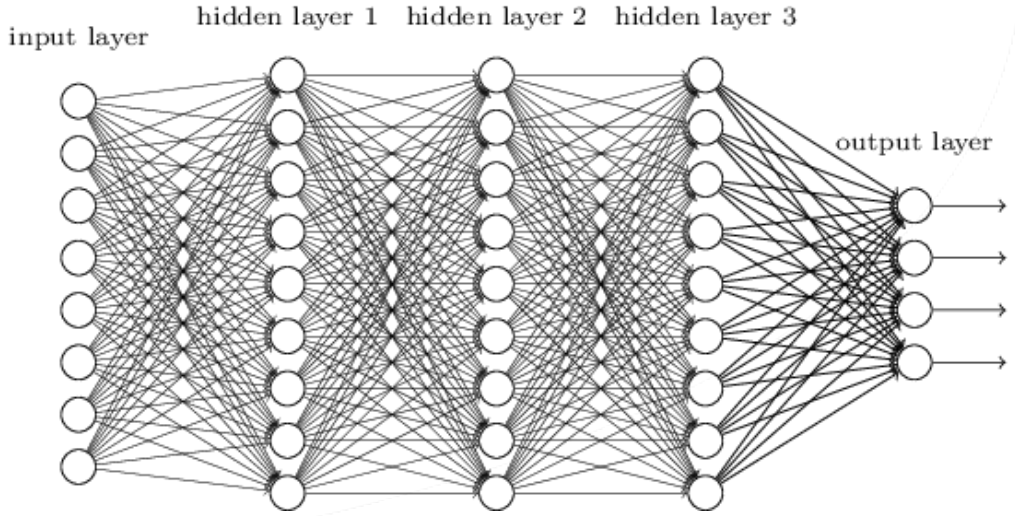
Adapting Integrate-and-Fire Unit



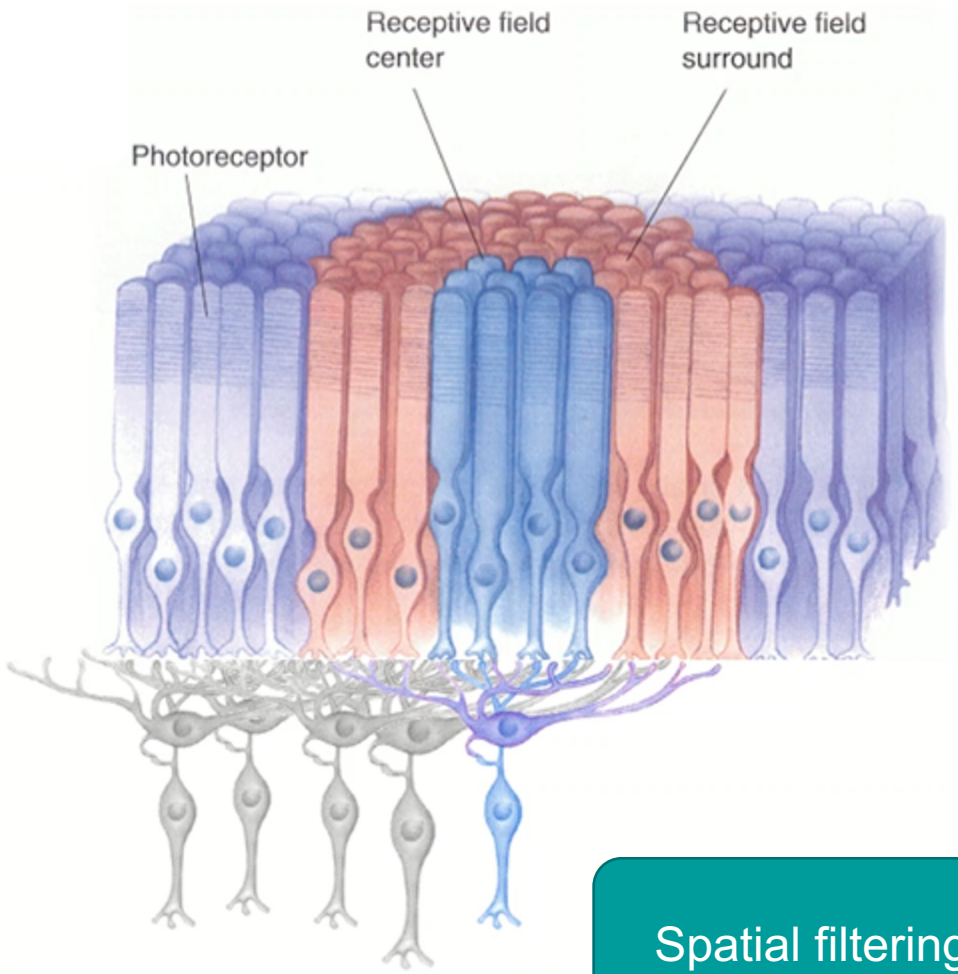
Spiking NN

Artificial NN

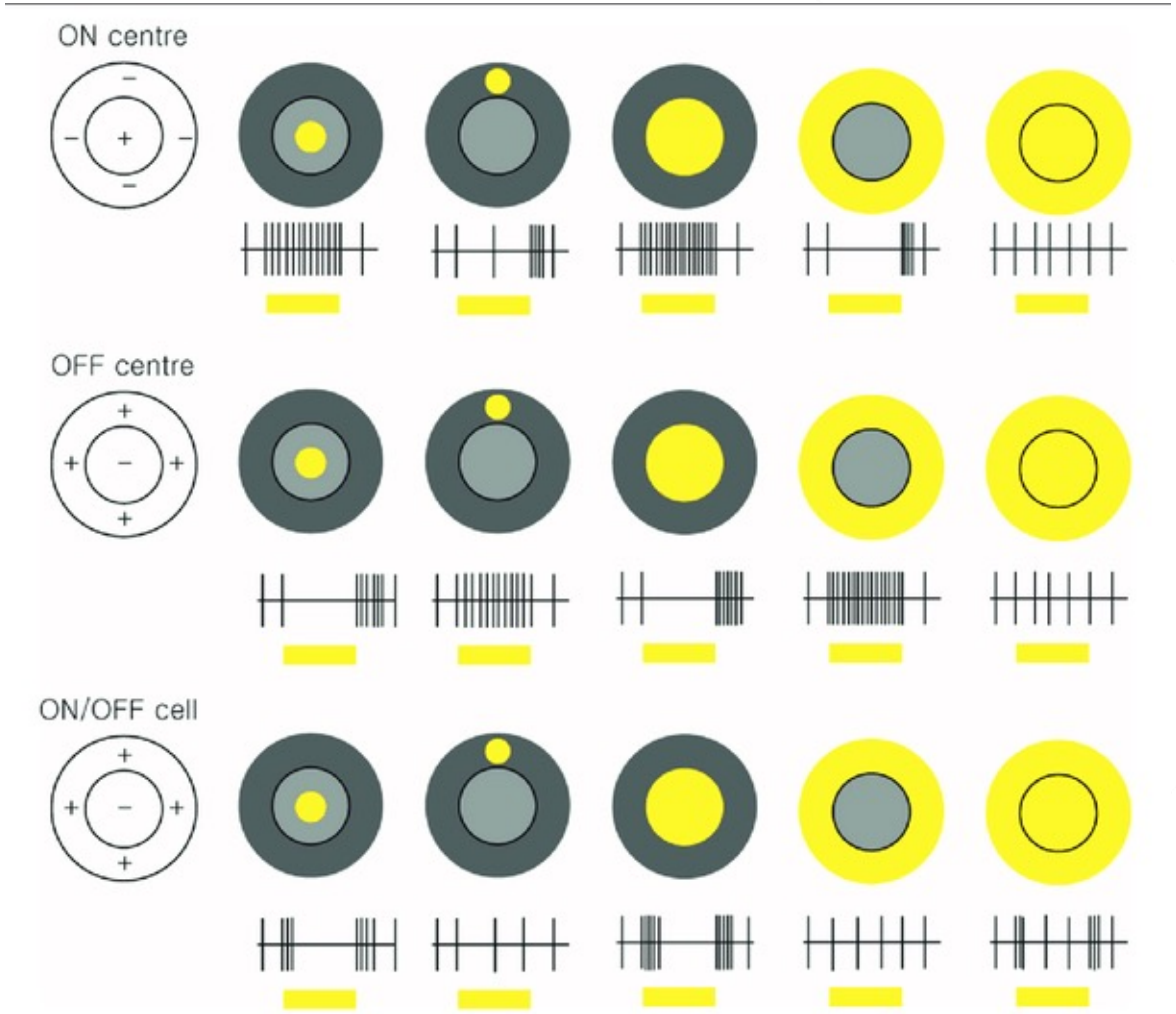


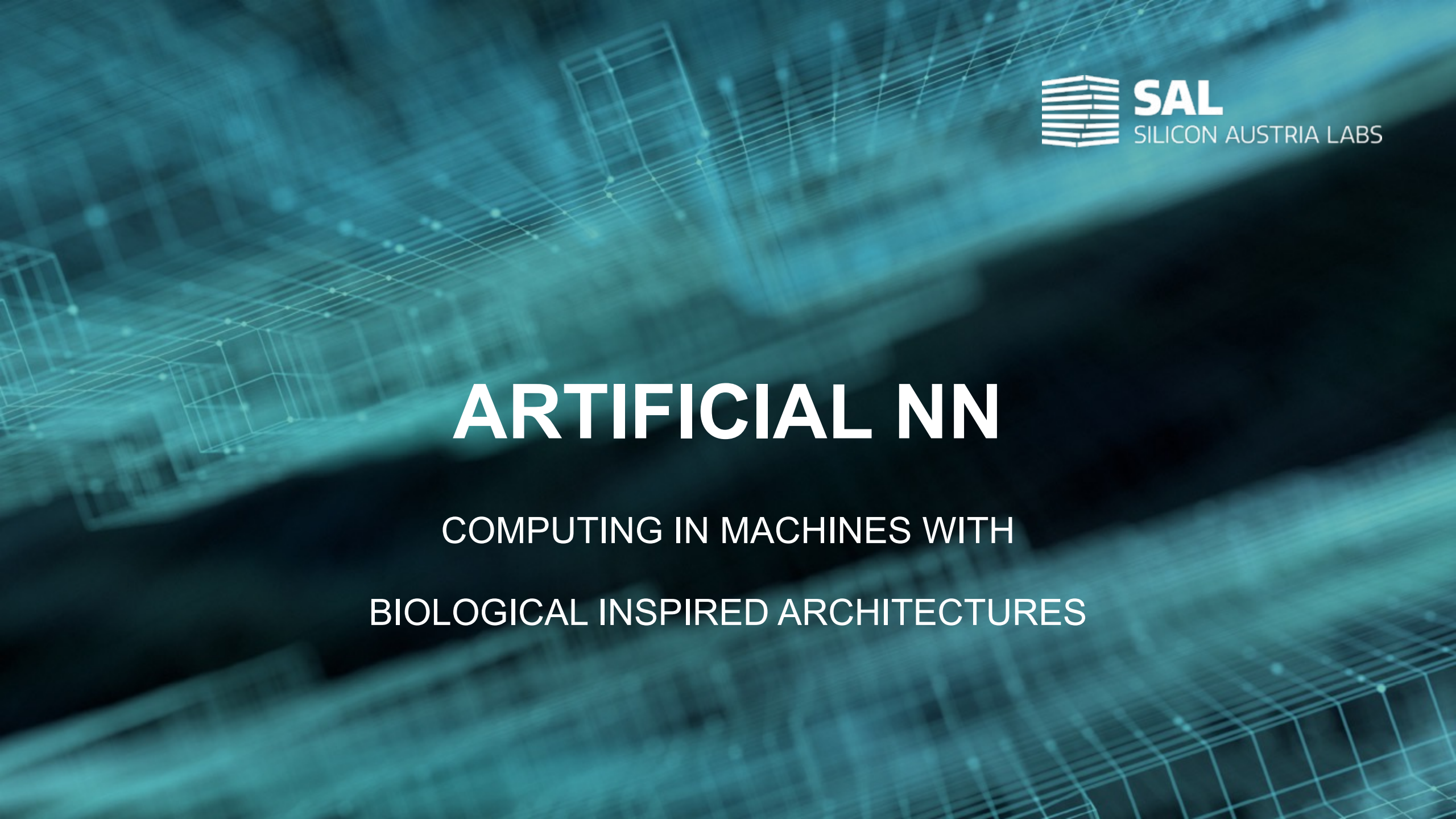


RETINAL IMAGE PROCESSING



Spatial filtering



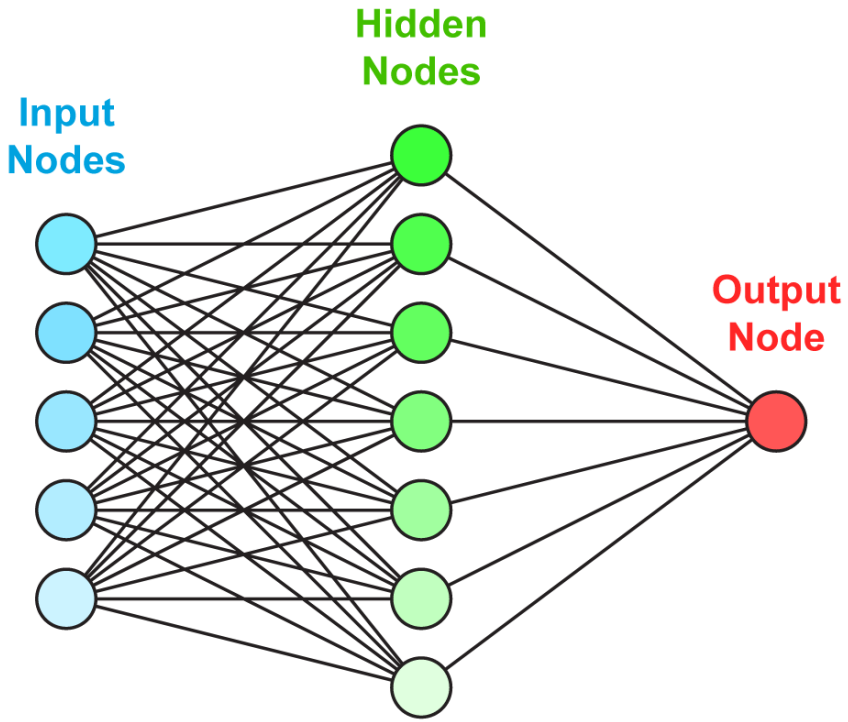


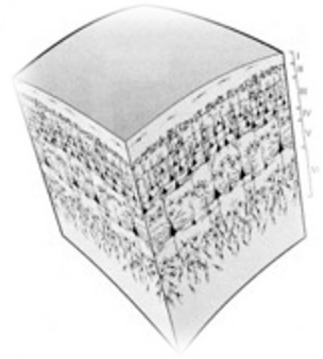
ARTIFICIAL NN

COMPUTING IN MACHINES WITH

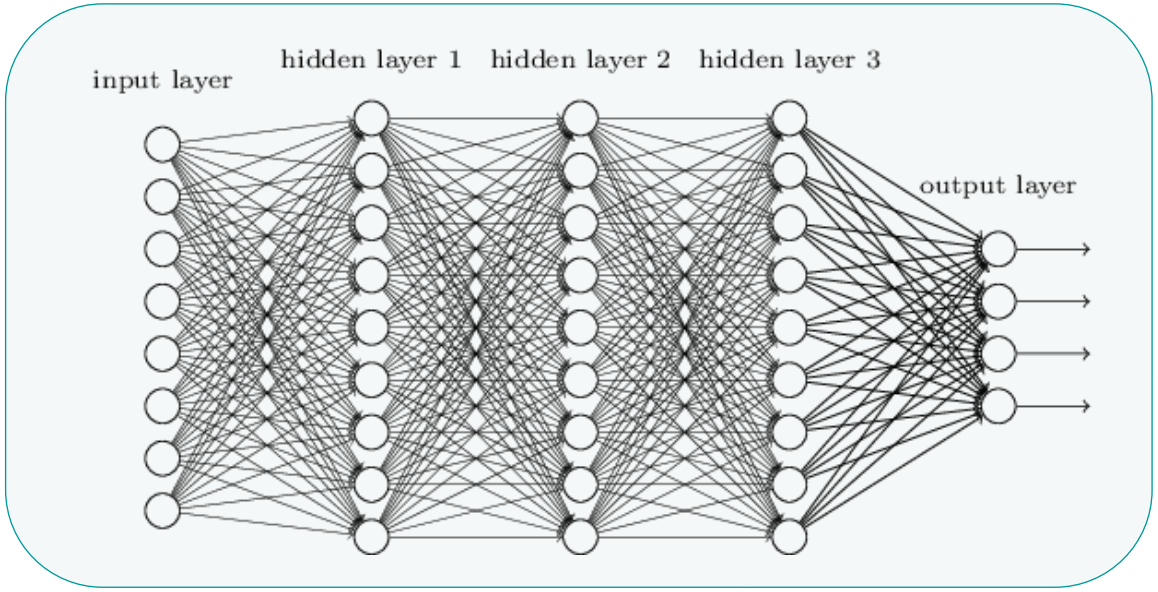
BIOLOGICAL INSPIRED ARCHITECTURES

FULLY CONNECTED NN



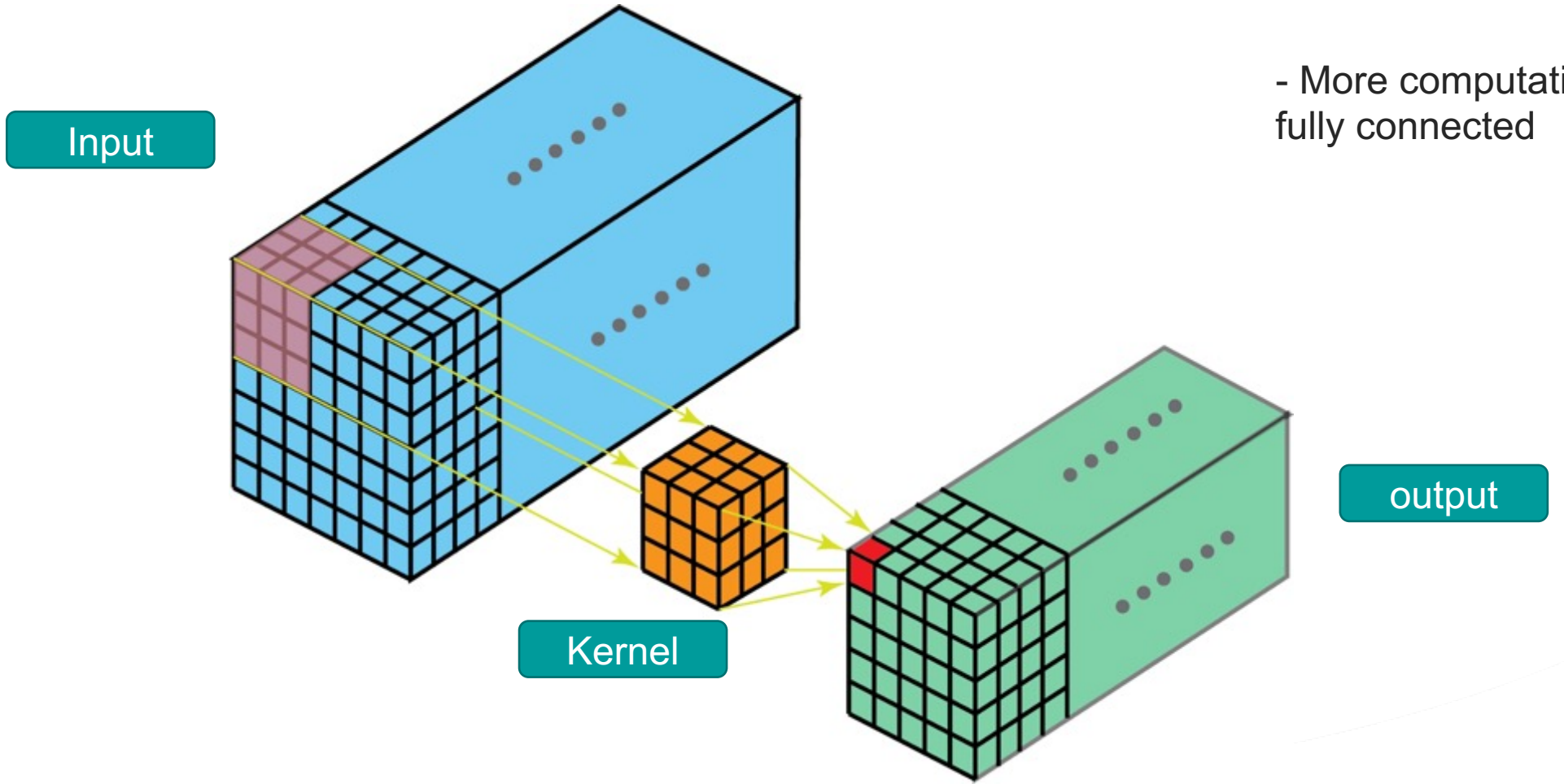


- All neurons from one layer connect to each neuron of the next layer



CONVOLUTIONAL NN

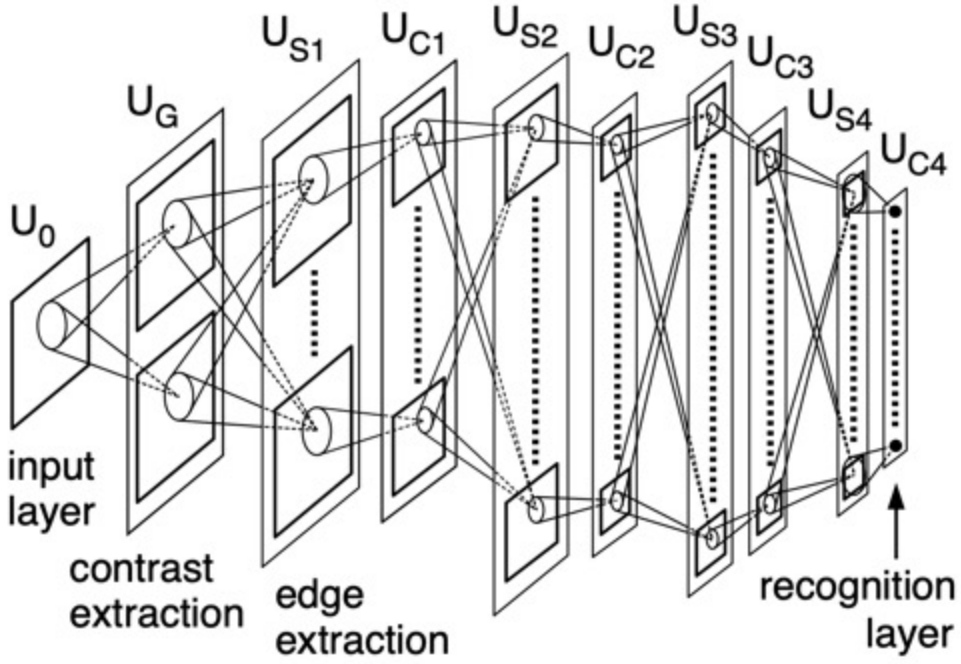
- A Convolutional layer performs a spatial (3D in general) convolution (spatial multiplication) with a kernel (filter)
  - Resembles a more elaborate ON/OFF center cell

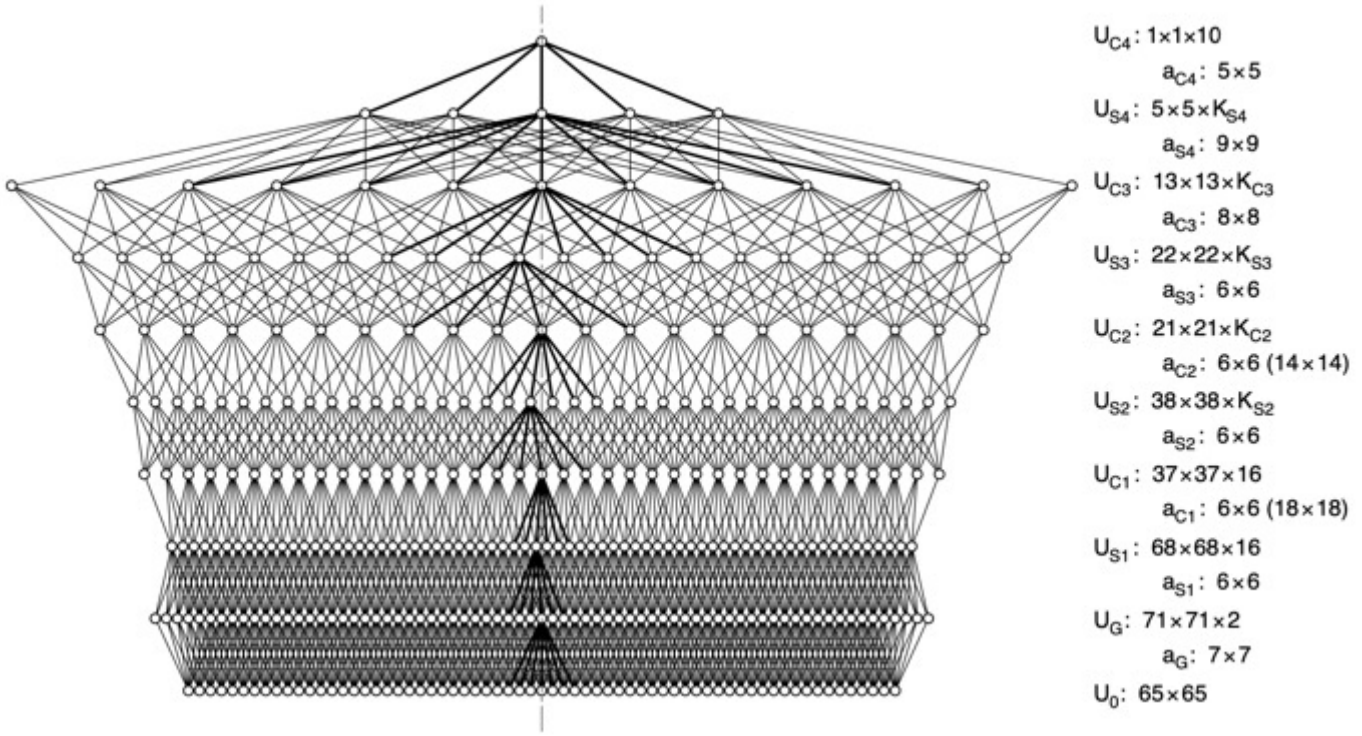


- More computationally intensive than fully connected

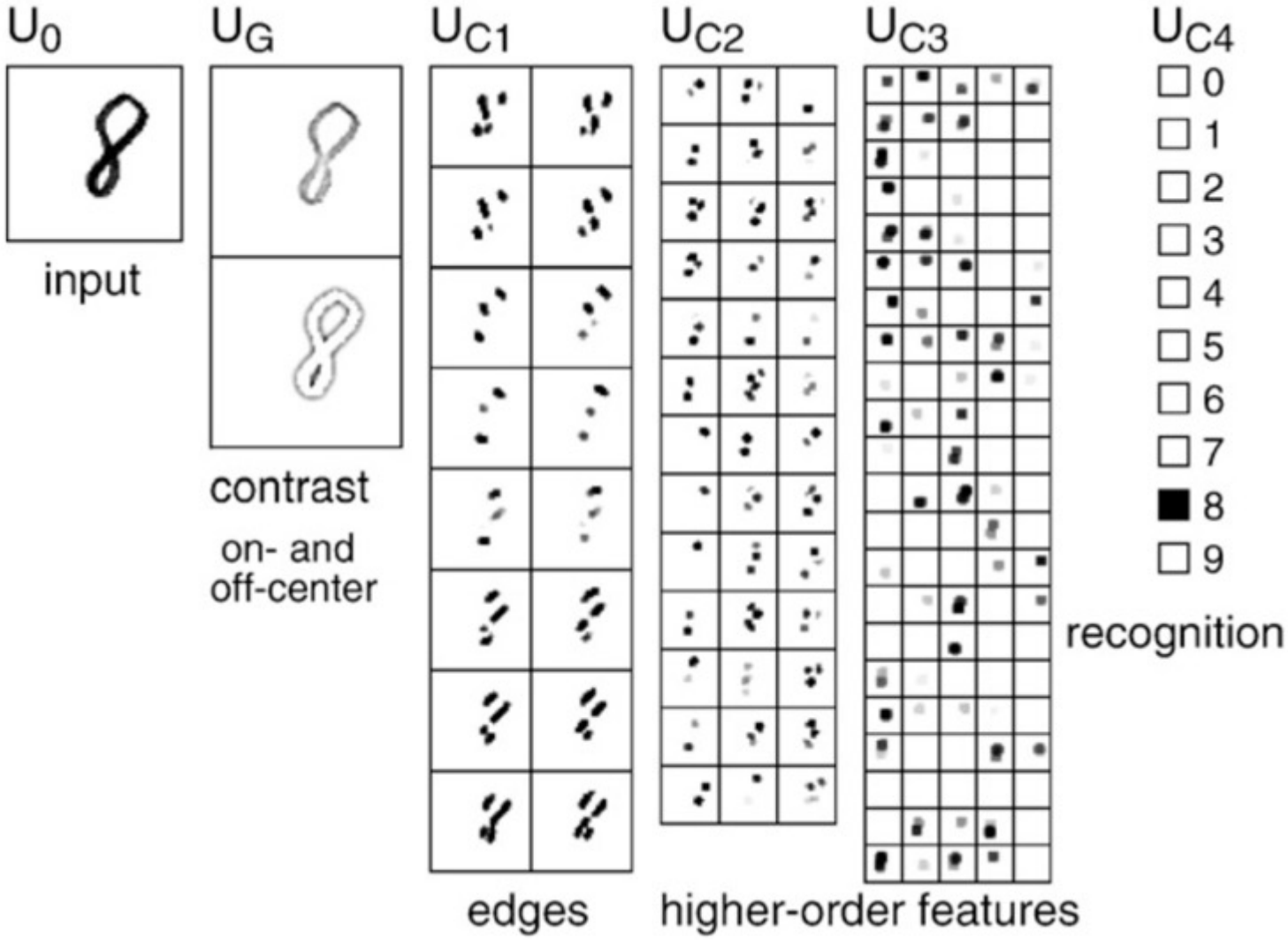
FROM SMALL TO HUGE

- 1979: The first "convolutional neural networks" were used by Kunihiko Fukushima (Neocognitron). Fukushima designed neural networks with multiple pooling and convolutional layers.



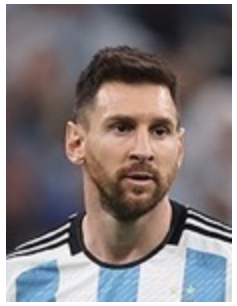


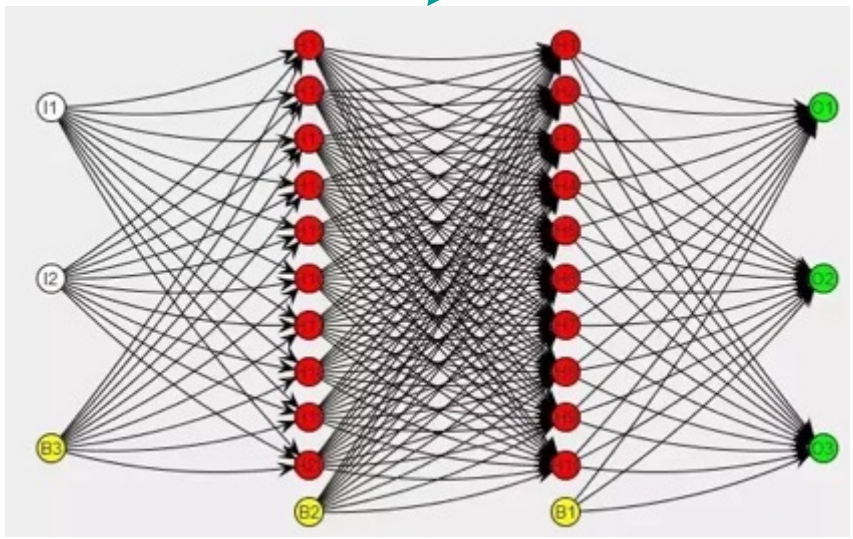
FROM SMALL TO HUGE



FROM SMALL TO HUGE

- 1985: Rumelhart, Williams, and Hinton demonstrated back propagation in a neural network could provide “interesting” distribution representations



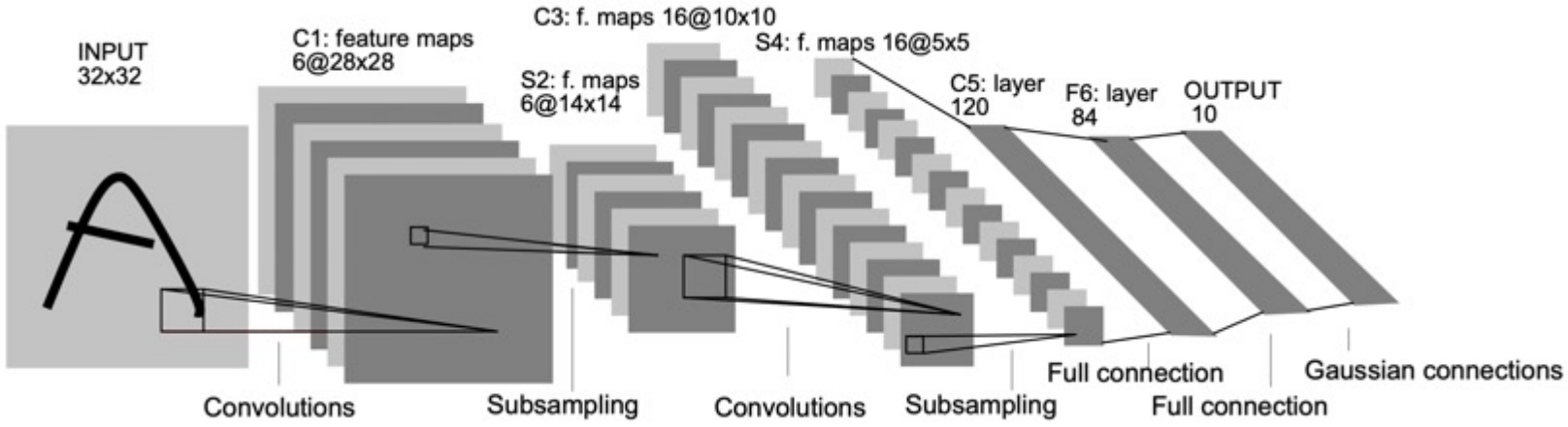


	Out	Truth	Error
Messi	2	10	-8
Mbappe	8	0	8
Dibu	4	0	4

Backpropagation propagates the error back, using the gradients to adjust each parameter

FROM SMALL TO HUGE

- In 1989, Yann LeCun provided the first practical demonstration of backpropagation. He combined convolutional neural networks with back propagation onto read “handwritten” digits.



- LeNet 5 would have take several weeks of training back then

- LeNet 1 (2600 param.) ~ 100.000 multiply/add operations

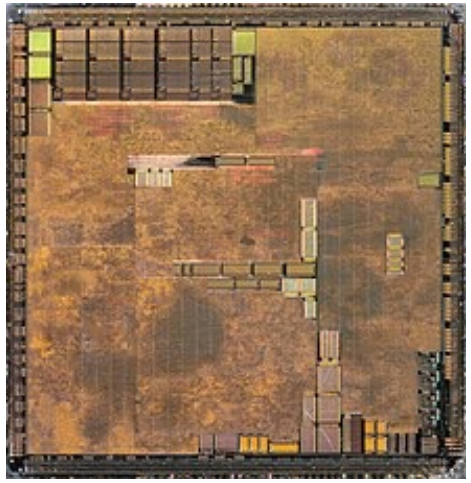
- MNIST 60.000 element dataset

FROM SMALL TO HUGE

SAL SILICON AUSTRIA LABS

- 1999: computers started becoming faster and GPU (graphics processing units) available.

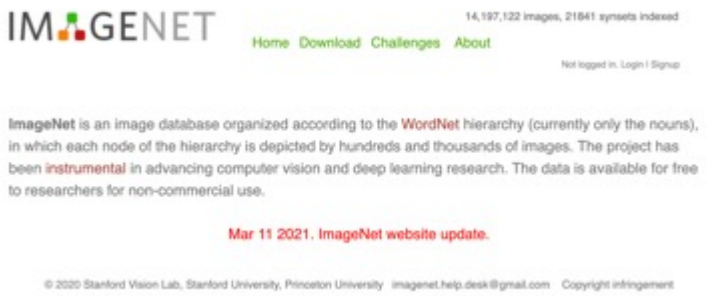
1999 Nvidia GeForce 256



Overview	
Manufacturer	NVIDIA
Original Series	GeForce 256
Release Date	October 11th, 1999
Launch Price	$179 USD
PCB Code	180-P0003-0100
Board Model	NVIDIA P3
Graphics Processing Unit	
GPU Model	NV10
Architecture	NV10
Fabrication Process	220 nm
Die Size	111 mm^2
Transistors Count	17M
Transistors Density	153.2K $^{TRAN}/_{mm^2}$
Pixel Pipelines	4
TMUs	4
ROPs	4
Clocks	
Base Clock	120 MHz
Boost Clock	TBC MHz
Memory Clock	143 MHz
Effective Memory Clock	143 Mbps
Memory Configuration	
Memory Size	32 MB
Memory Type	SDR
Memory Bus Width	64-bit
Memory Bandwidth	1.1 GB/s

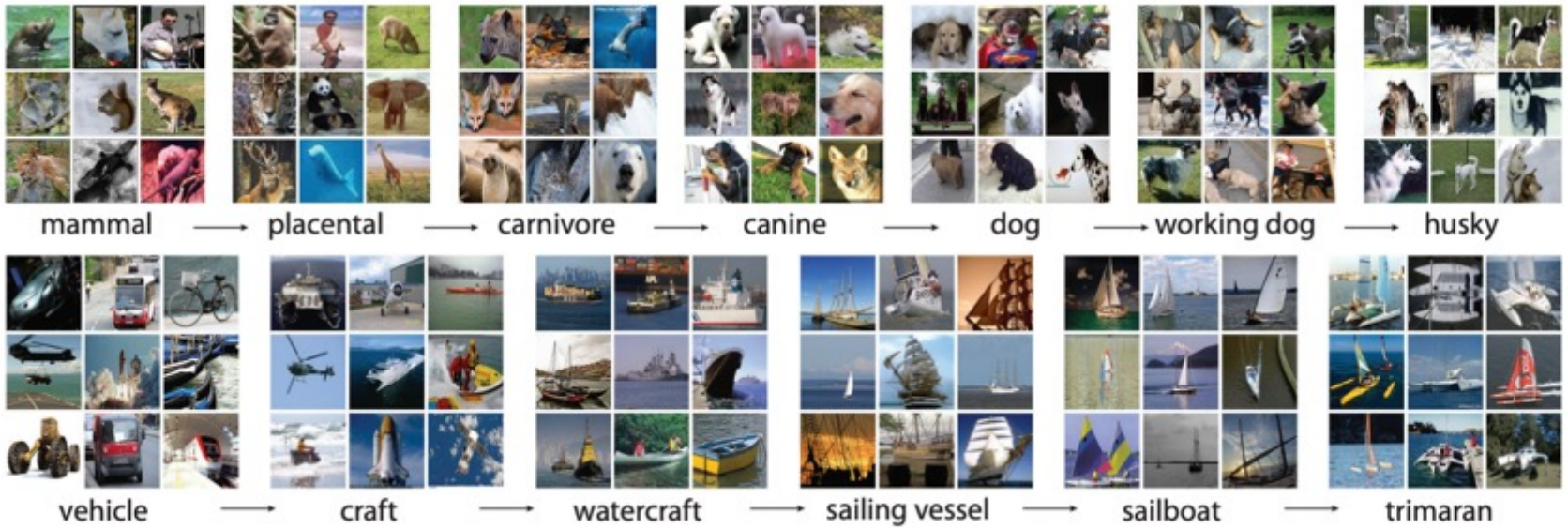
FROM SMALL TO HUGE

- 2009: Fei-Fei Li from Stanford launched ImageNet, a free database of more than 14 million labeled images.



ImageNet: A Large-Scale Hierarchical Image Database

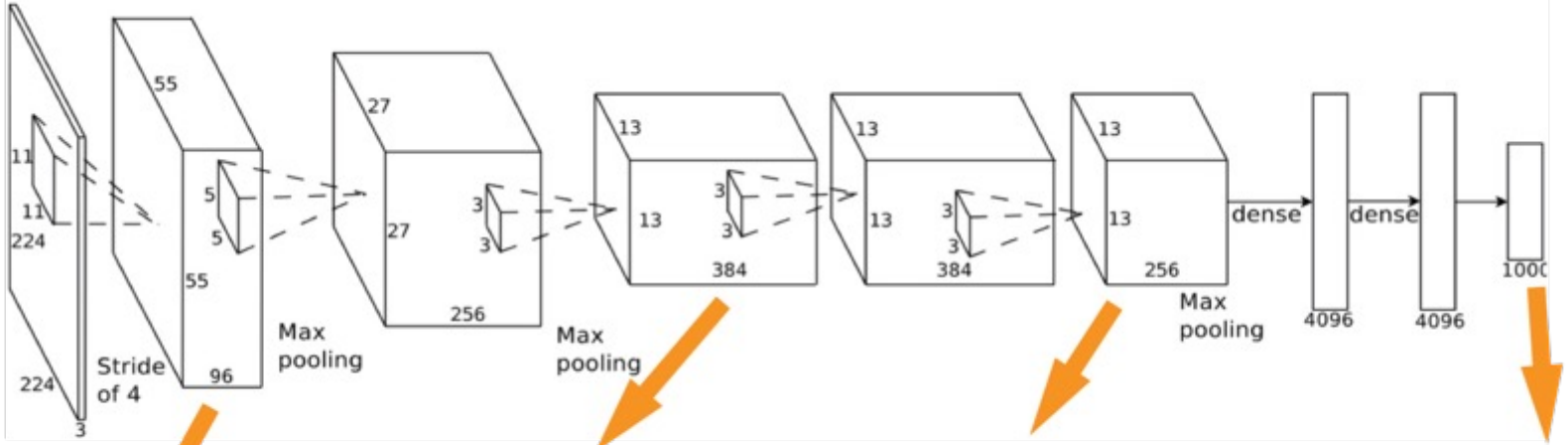
Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei
Dept. of Computer Science, Princeton University, USA
{jiadeng, wdong, rsocher, jial, li, feifeili}@cs.princeton.edu

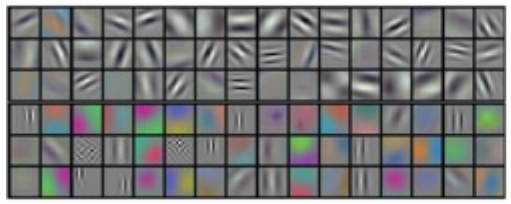


3.2 millions

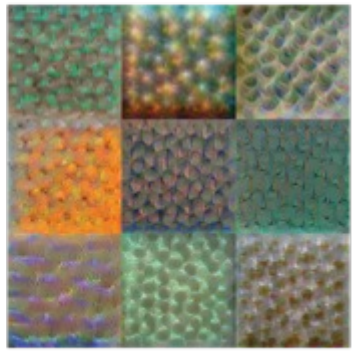
FROM SMALL TO HUGE

- 2011-2012: AlexNet, won several international competitions

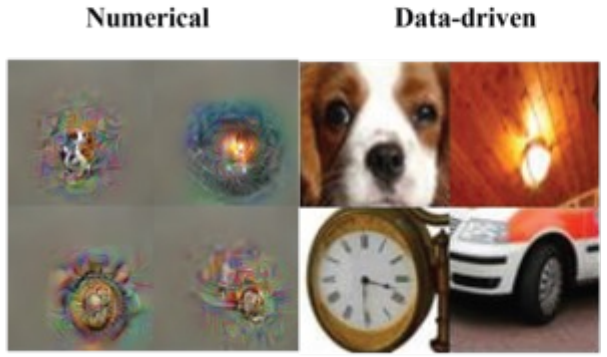




Conv 1: Edge+Blob



Conv 3: Texture



Conv 5: Object Parts

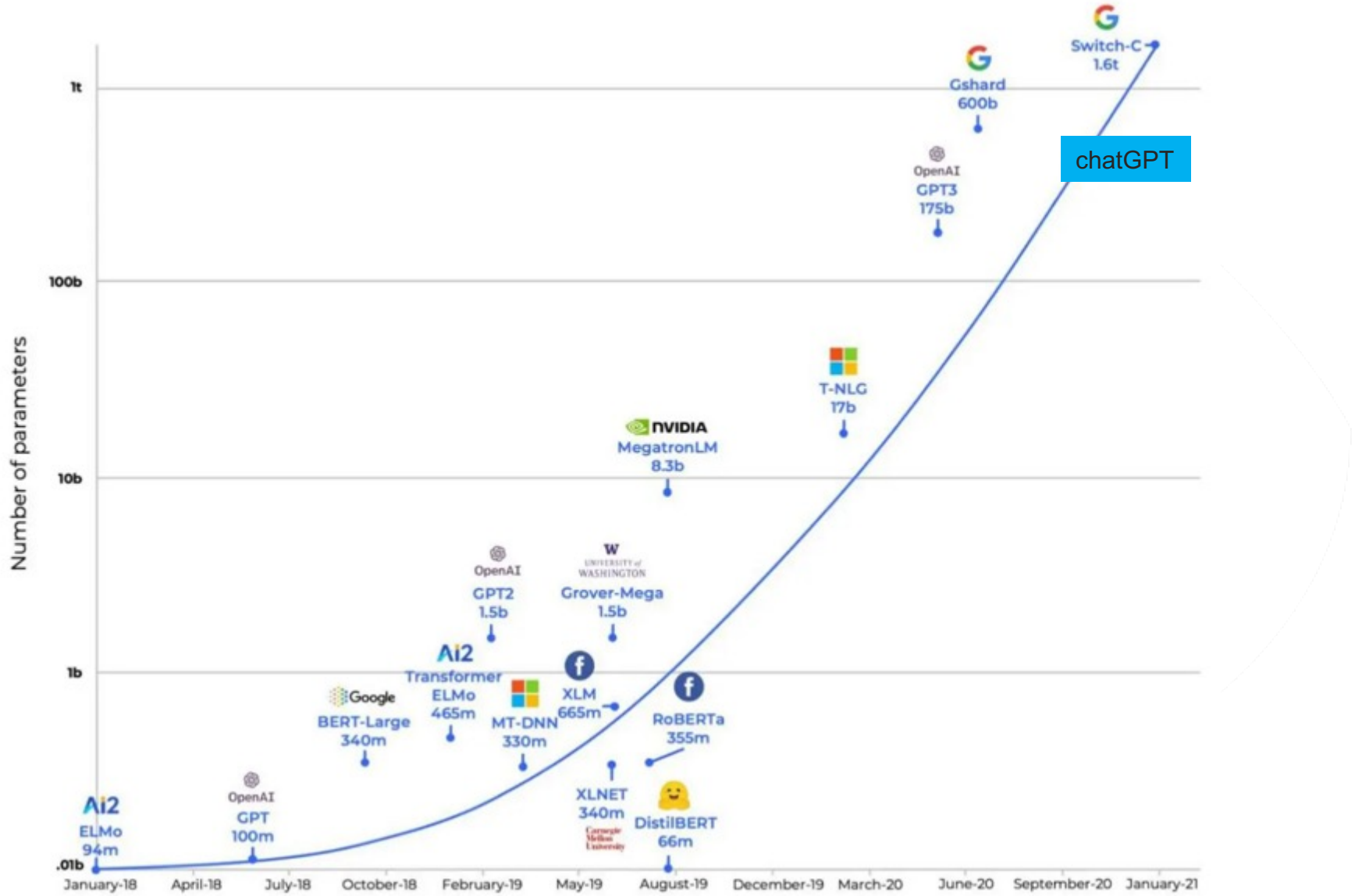
Fc8: Object Classes

- **5 convolutional kernels:**
  - 96 11x11, 256 5x5, 384 3x3, 384 3x3, 256 3x3
- **3 fully connected:**
  - 4096, 4096, 1000.

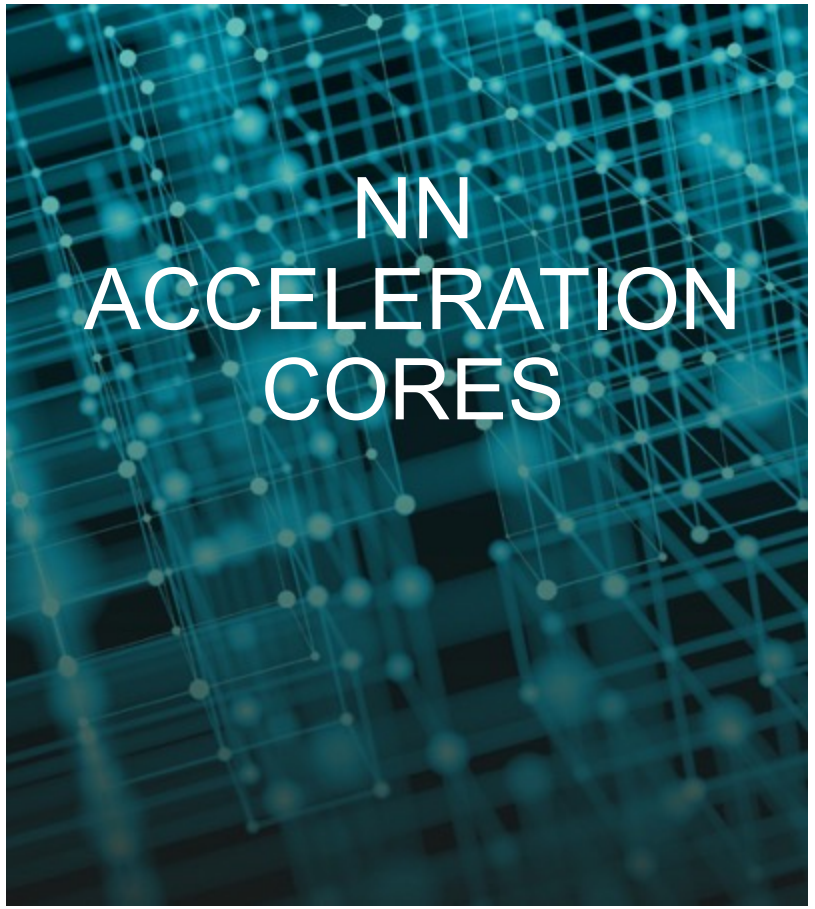
650K neuronas, 60M parámetros, 630M conexiones

FROM SMALL TO HUGE

- 2018 till today
- Natural language processing engines
- Models with hundred billion parameters

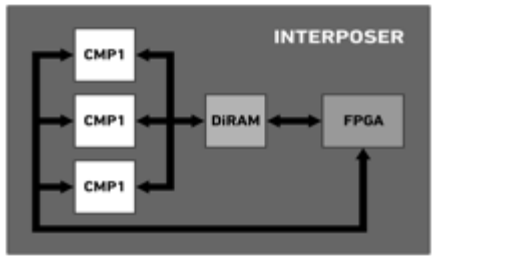


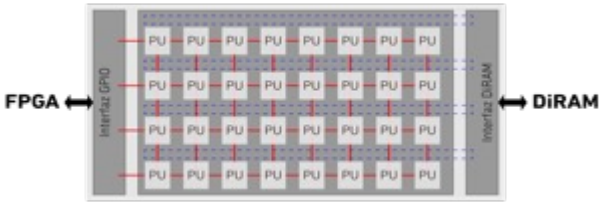
CORES



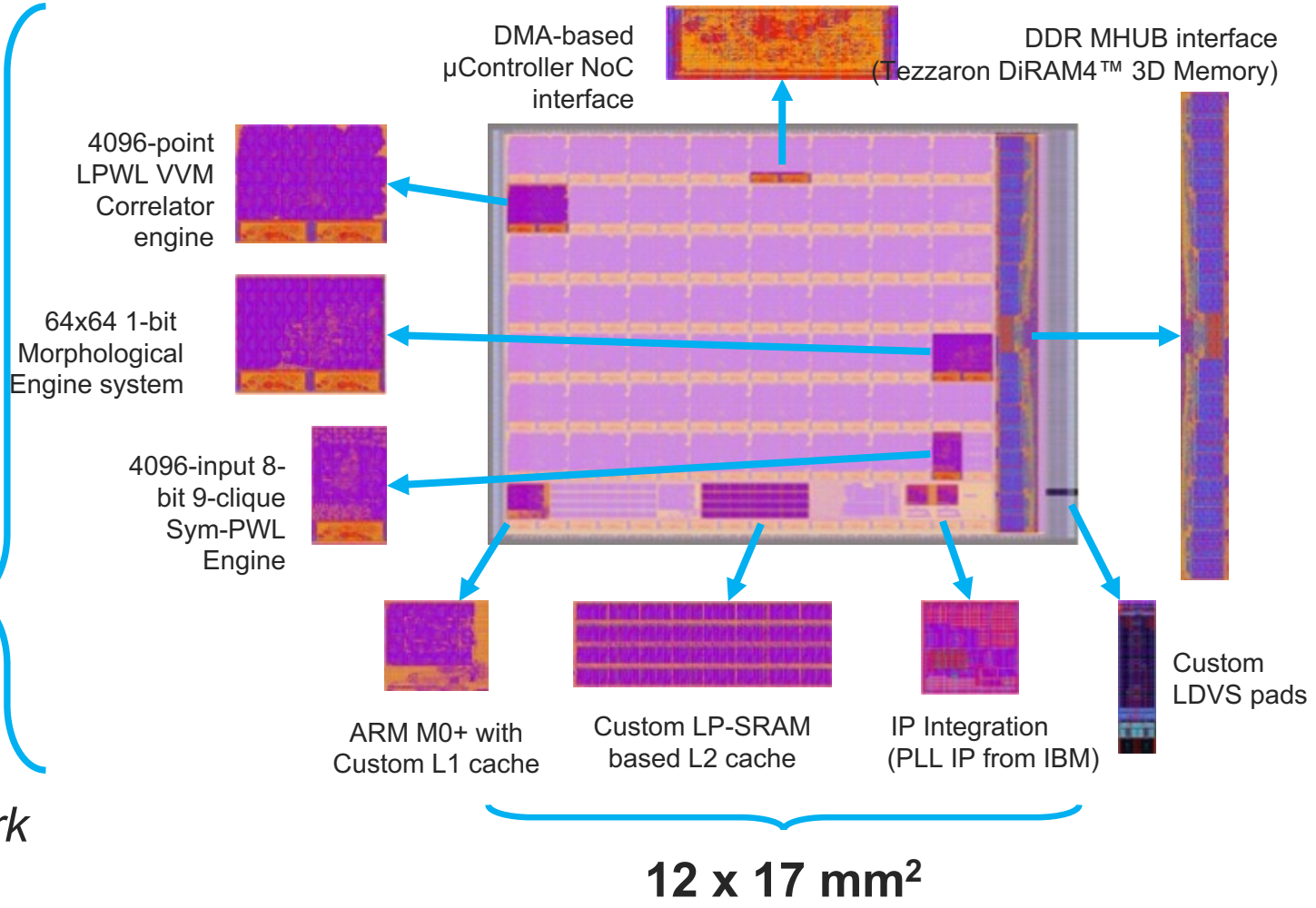
- A collection of computational cores with improved energy efficient

HETEROGENOUS *SALAMI* CHIP





L1-NoC → *token ring*
L2-NoC → *mesh network*



SAL
SILICON AUSTRIA LABS

END