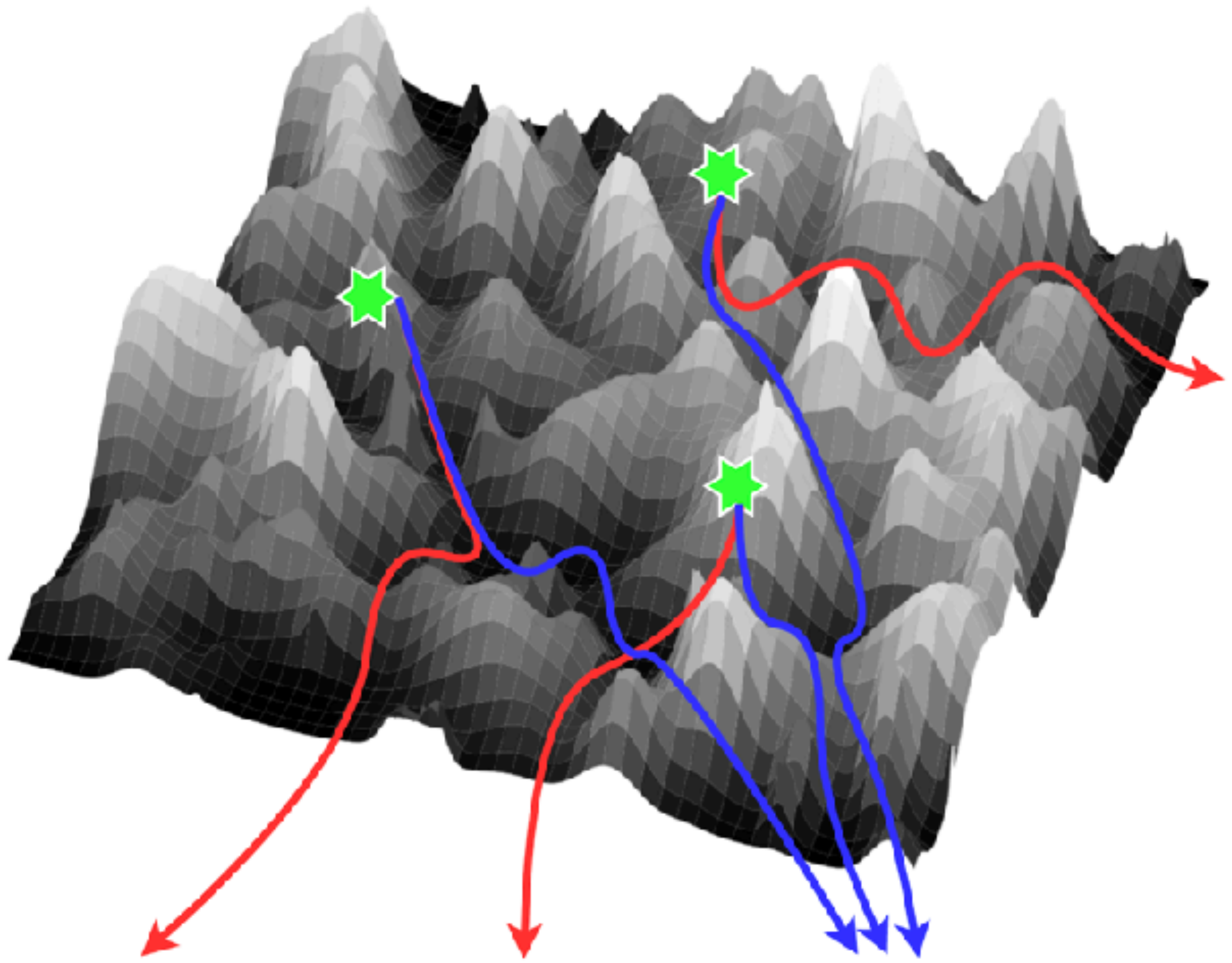
Challenges for a modern theory of neural networks

Sebastian **Goldt** (SISSA, Trieste)



Junior Math Days @ SISSA — dec 2024

Group dinner,
July 2024

New machine learning breakthroughs…

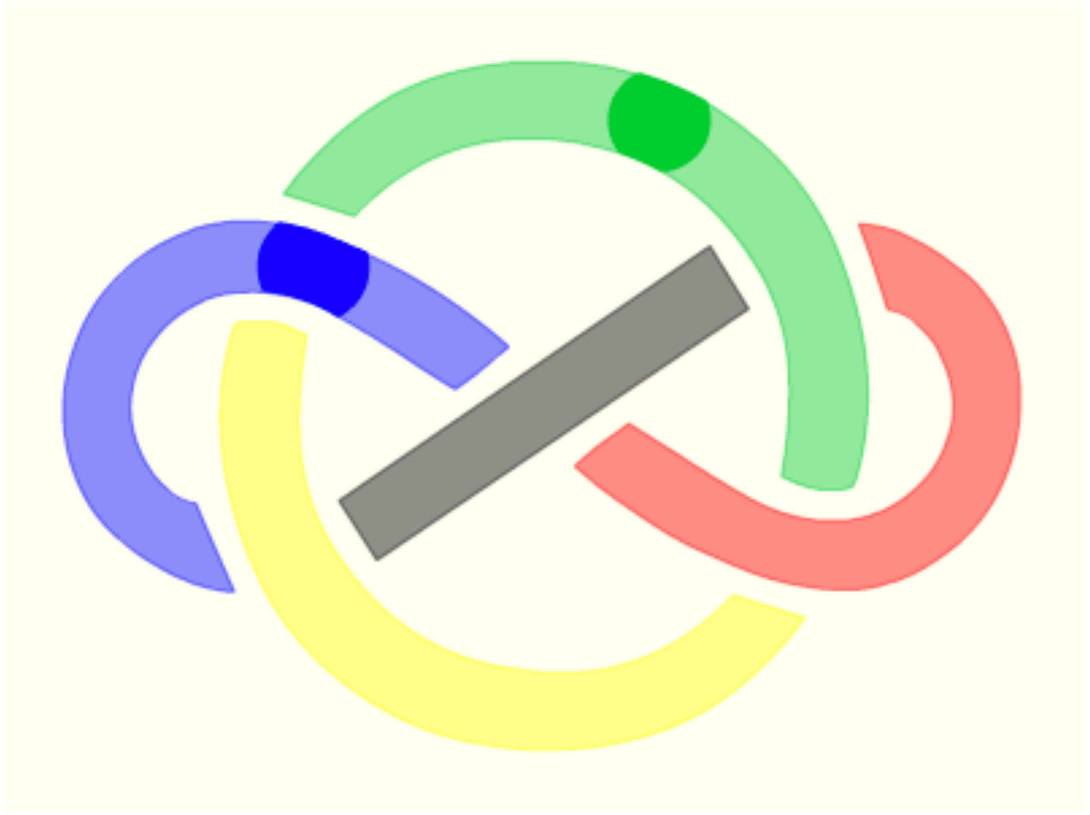
Hey ChatGP, I'm interested in the theory of neural networks. Do you know anything about that?

Yes, I'm familiar with the theory of neural networks. Neural networks are a type of machine learning algorithm that are modeled after the structure and function of the human brain. They consist of layers of interconnected "neurons," which process and transmit information. Neural networks are commonly used for tasks such as image and speech recognition, and they have been instrumental in many of the recent advances in artificial intelligence.

https://twitter.com/StelfieTT

New machine learning breakthroughs...

You heard the news



In the words of the New York Times:

“The Navy revealed the embryo of an electronic computer that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its own existence”

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

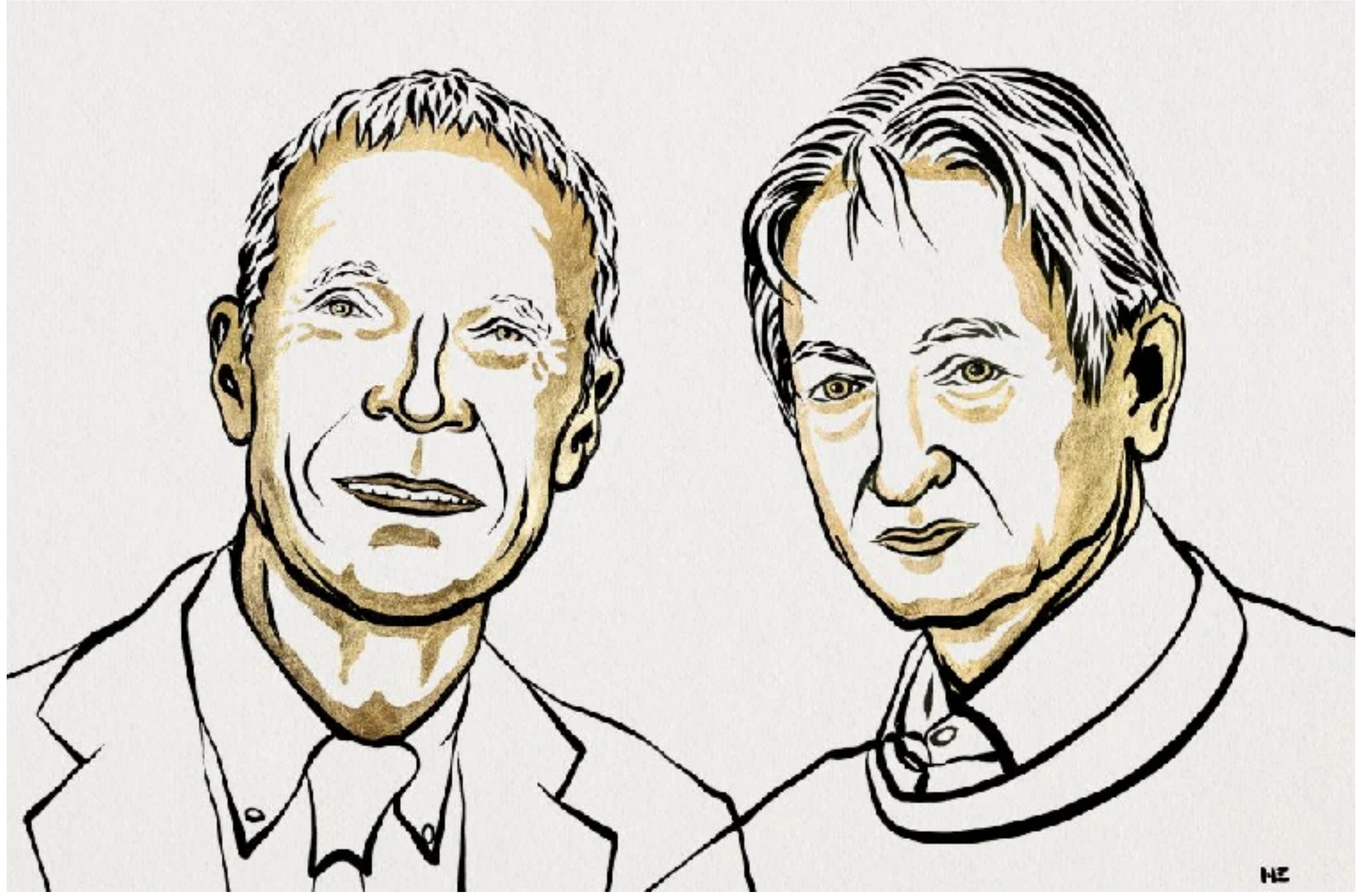
The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

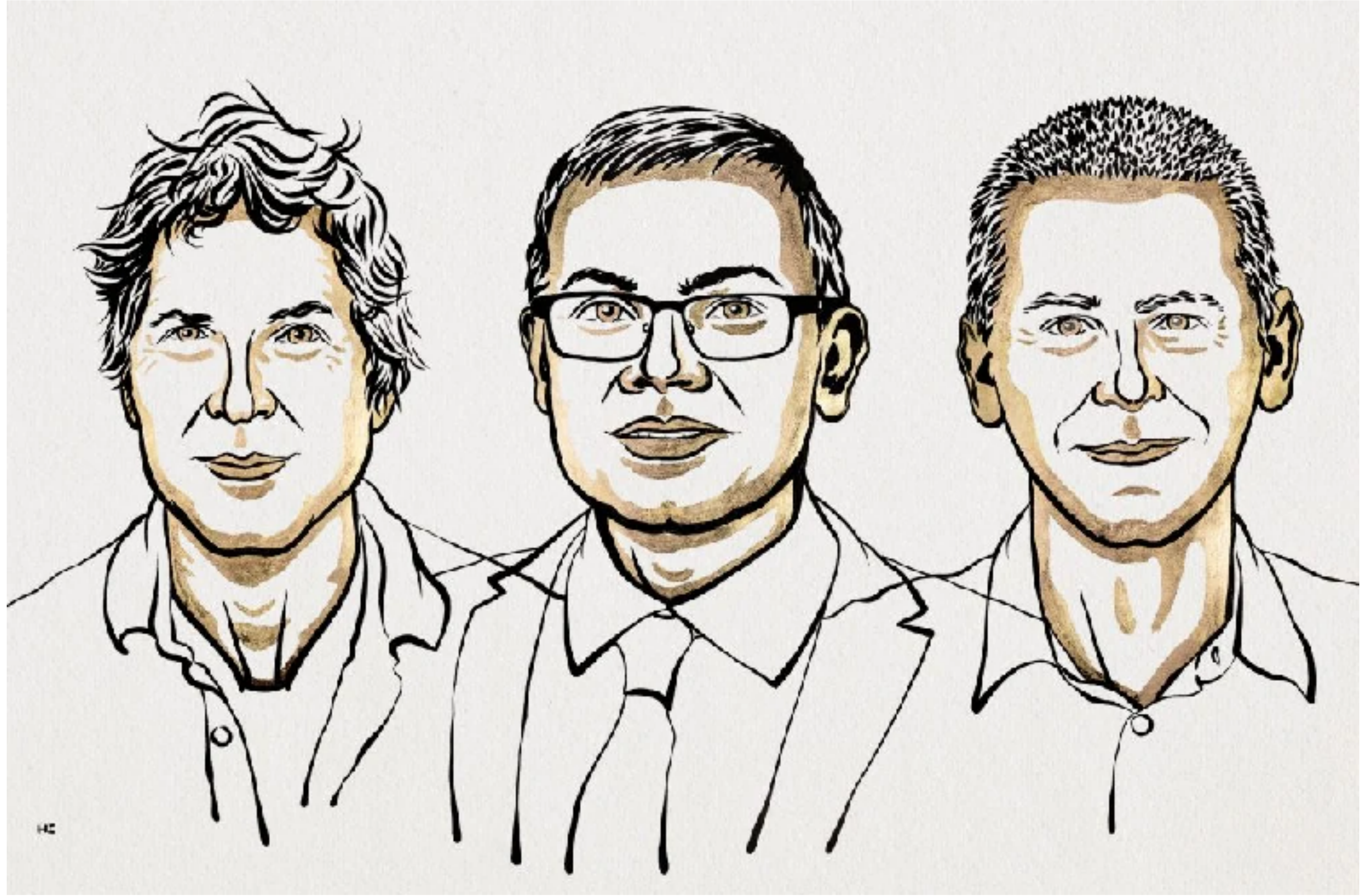
New York Times,
July 8, 1958

Some more recent attention

Talk about a crazy week



JJ Hopfield & GE Hinton



D Baker, D Hassabis and J Jumper

The plan for today

What is a neural network?

From neurons to networks.

Challenges for a modern theory

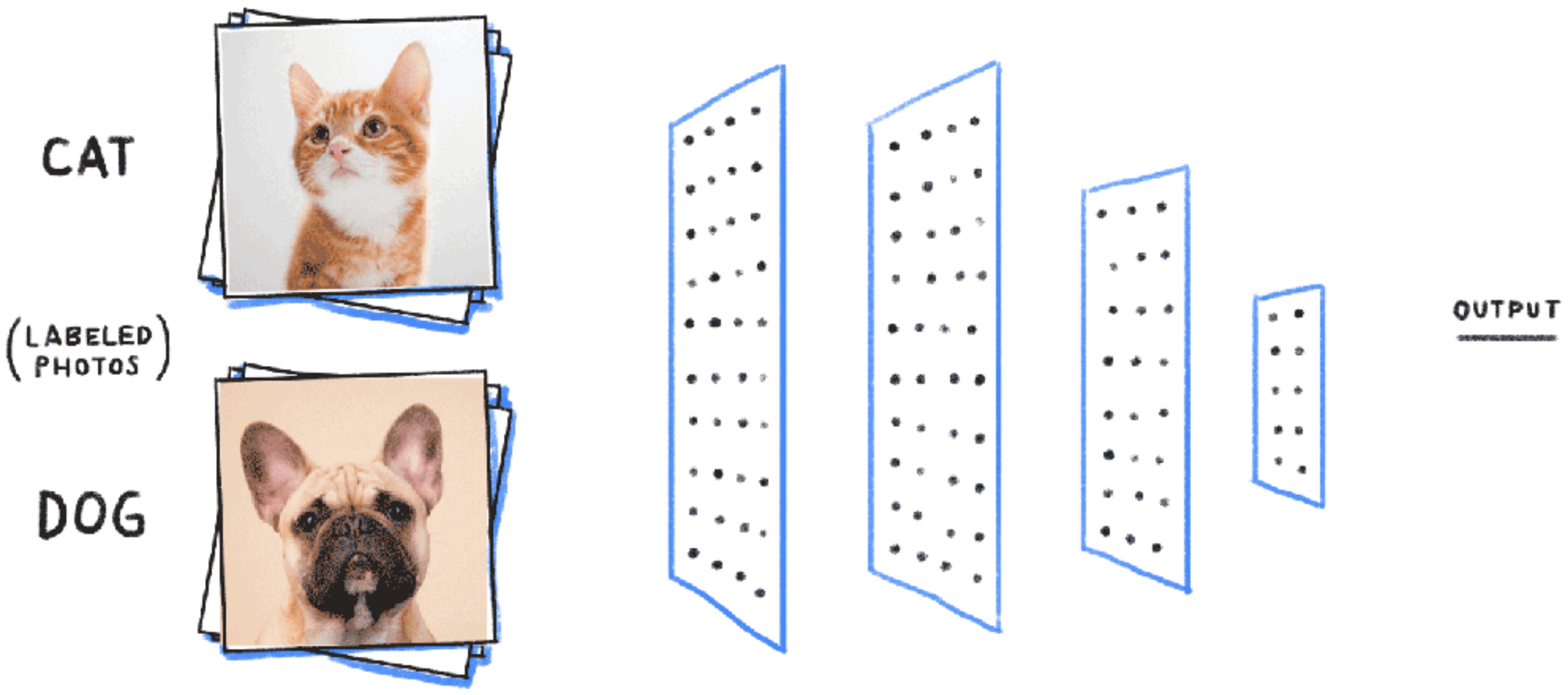
Network architecture, data structure, and learning algorithm

Part I

What is a neural network?

What is a neural network?

A neural network is a (complicated) function



Animation courtesy of Aakash Srivastava

Let's take a closer look at
one of these small black dots…

A single neuron

Many inputs, one output

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

(1943)

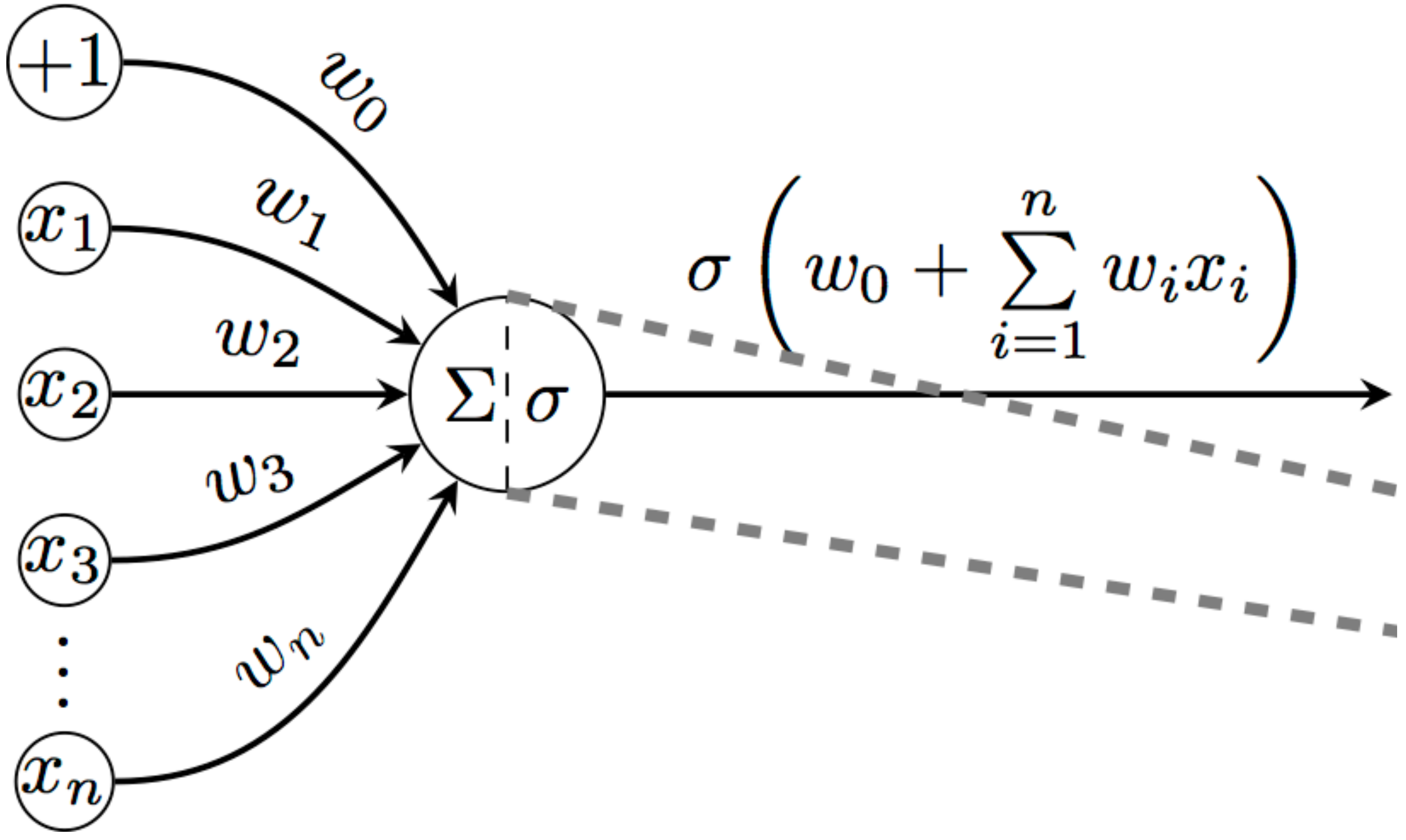
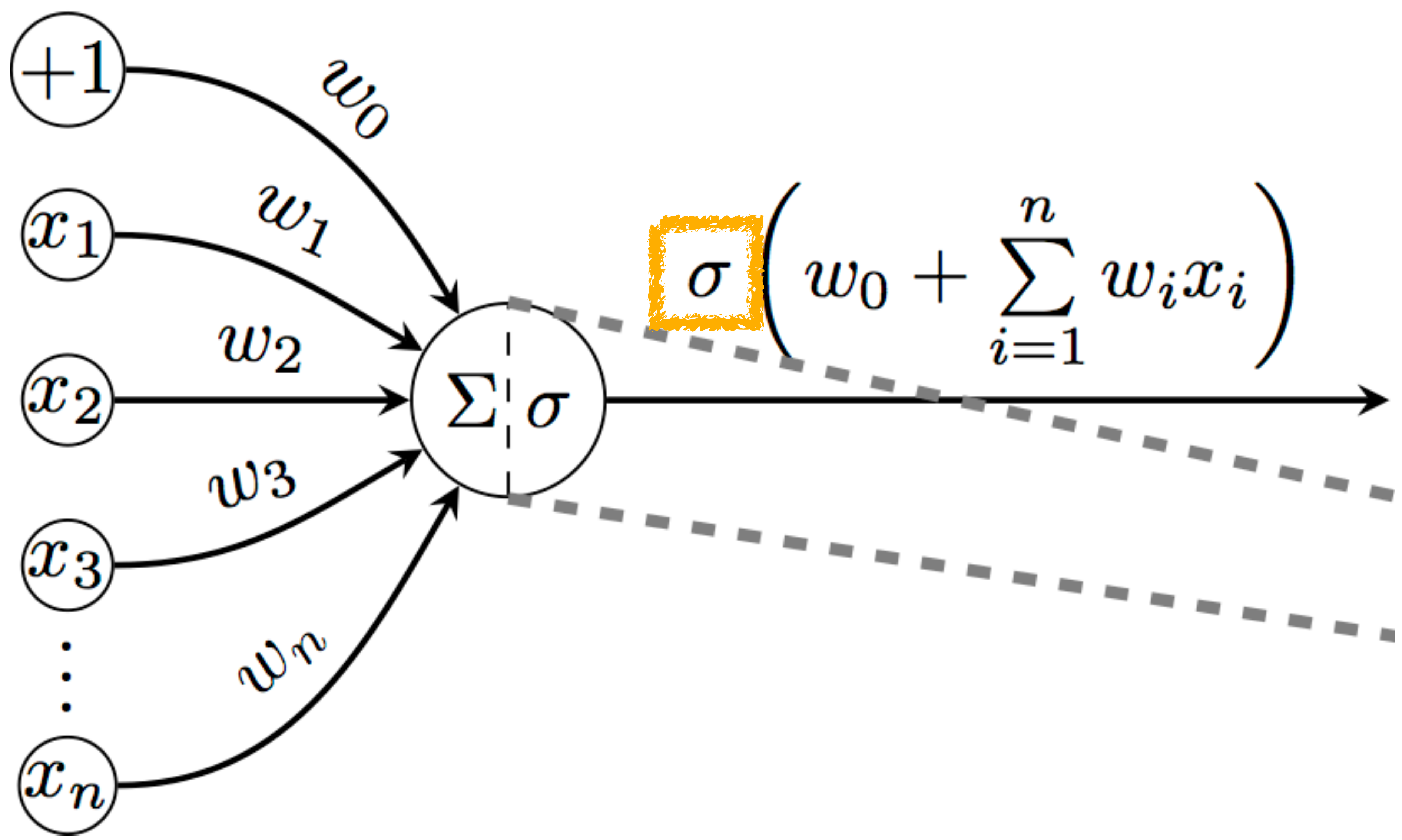


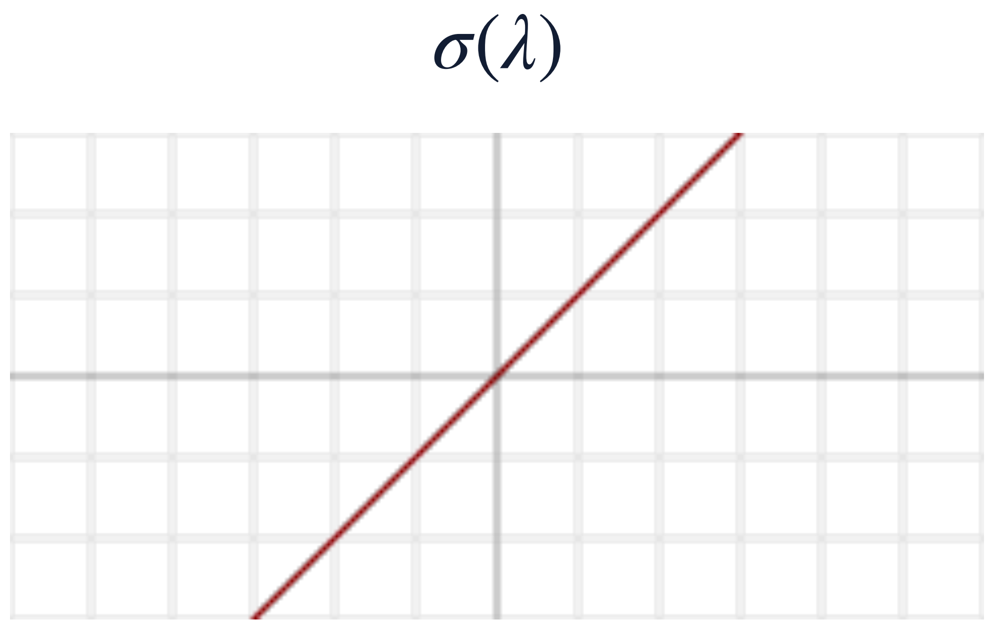
Illustration by Petar Veličković
https://github.com/PetarV-/TikZ

The activation function

More than just summing up

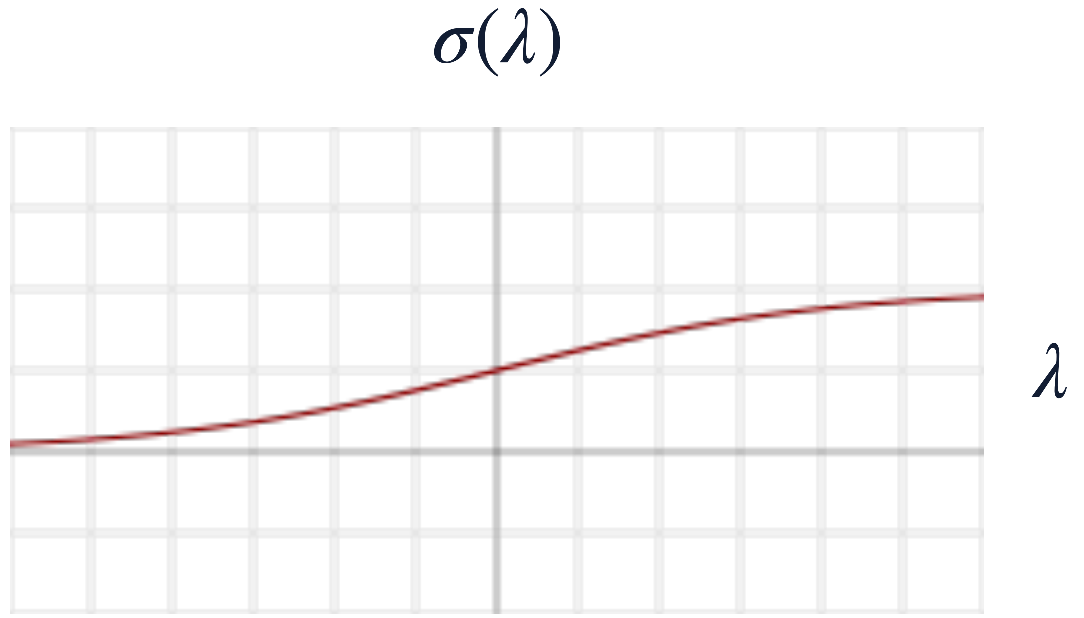


Linear

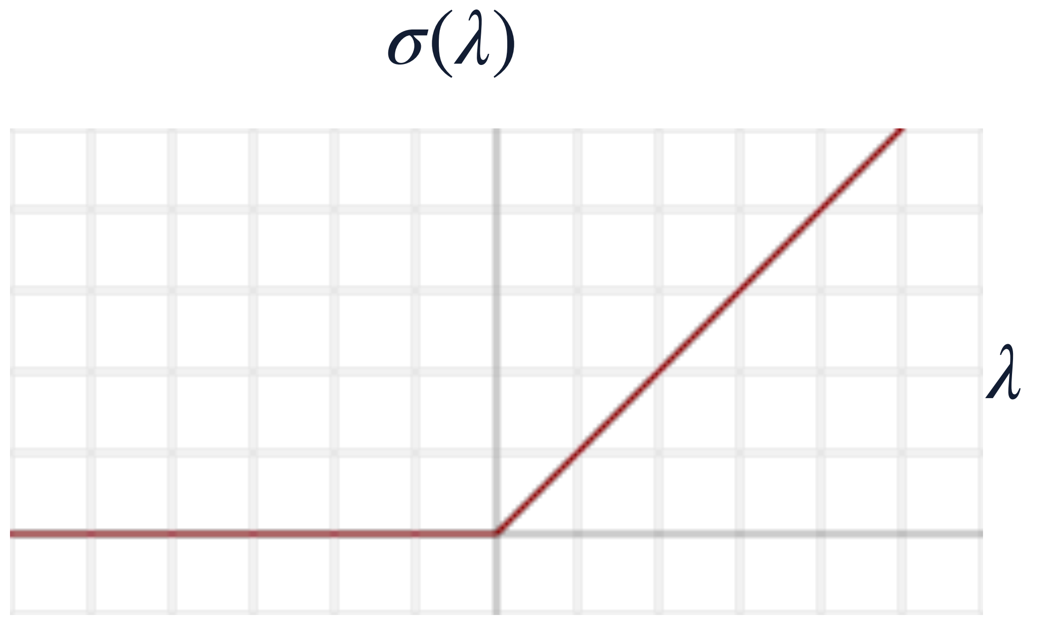


$$\lambda = w_0 + \sum_{i}^{n} w_i x_i$$

Sigmoid



ReLU



A lot of neurons

Neurons can be assembled into layers

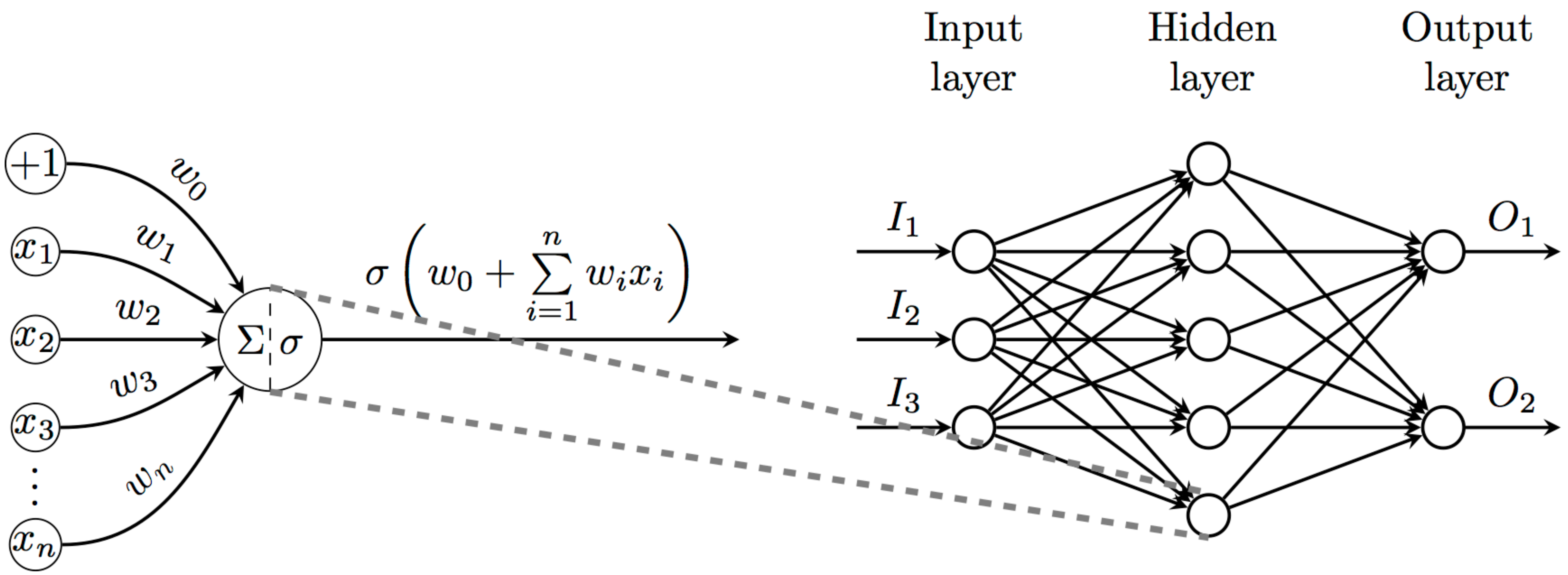
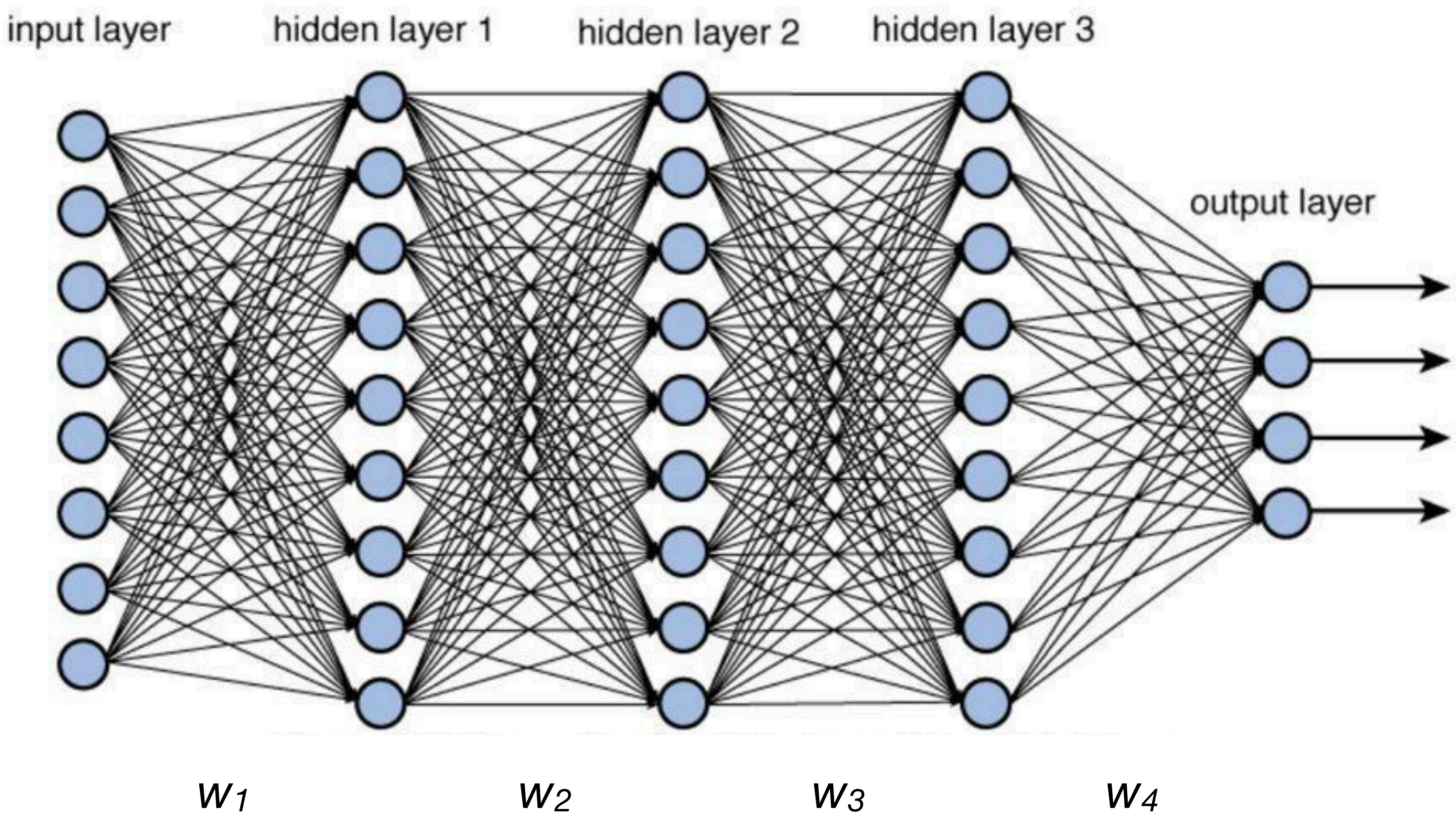


Illustration by Petar Veličković
https://github.com/PetarV-/TikZ

A lot of layers

Deep neural networks stack layers of neurons



$$y = w_4\sigma\left(w_3\sigma(w_2\sigma(w_1x))\right)$$

A mostly complete chart of neural networks

$y = w_4\sigma\left(w_3\sigma(w_2\sigma(w_1x))\right)$ is not the whole story

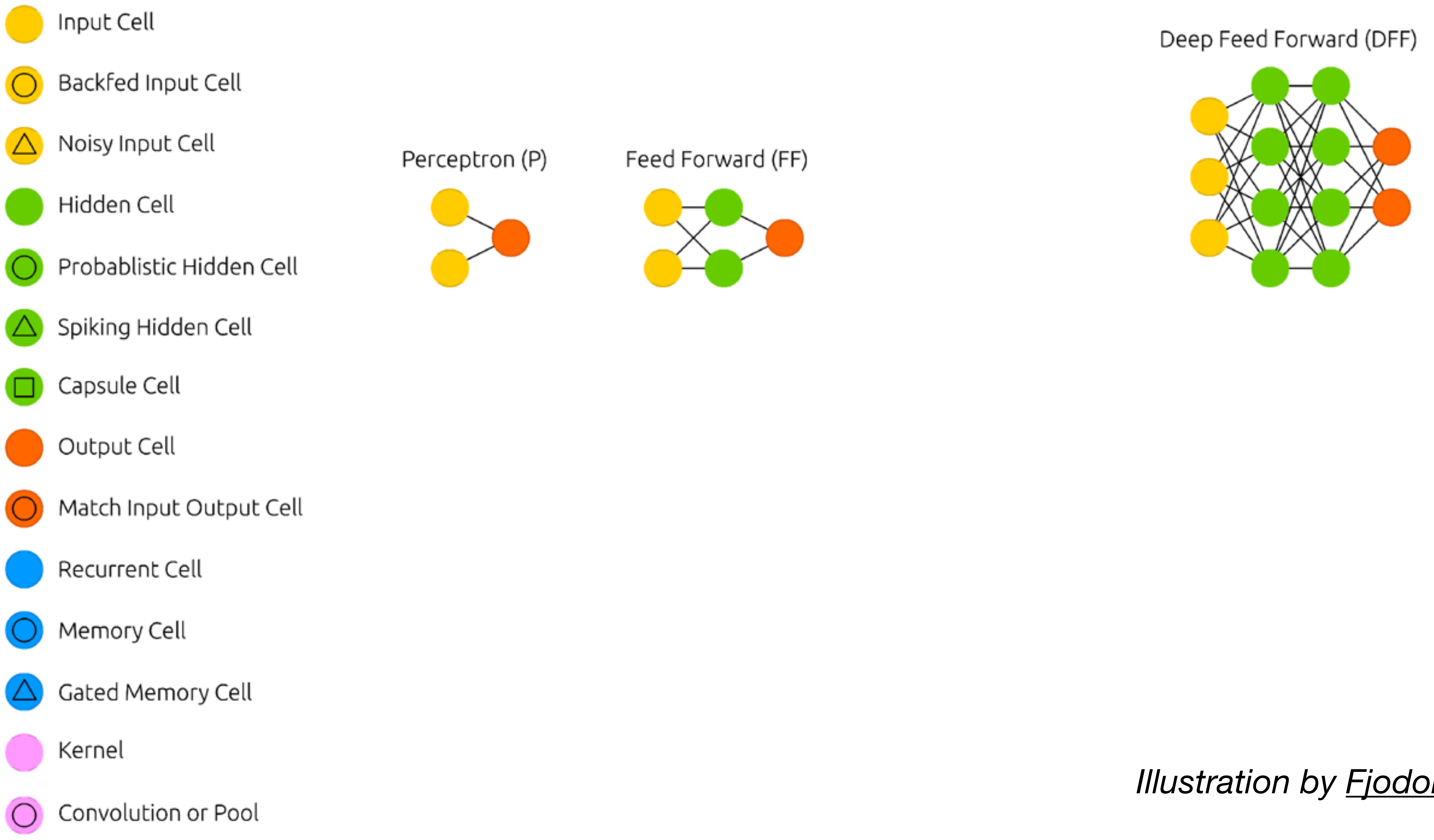


Illustration by [Fjodor van Veen]() & Stefan Leijnen

A mostly complete chart of neural networks

$y = w_4\sigma\left(w_3\sigma(w_2\sigma(w_1x))\right)$ is not the whole story

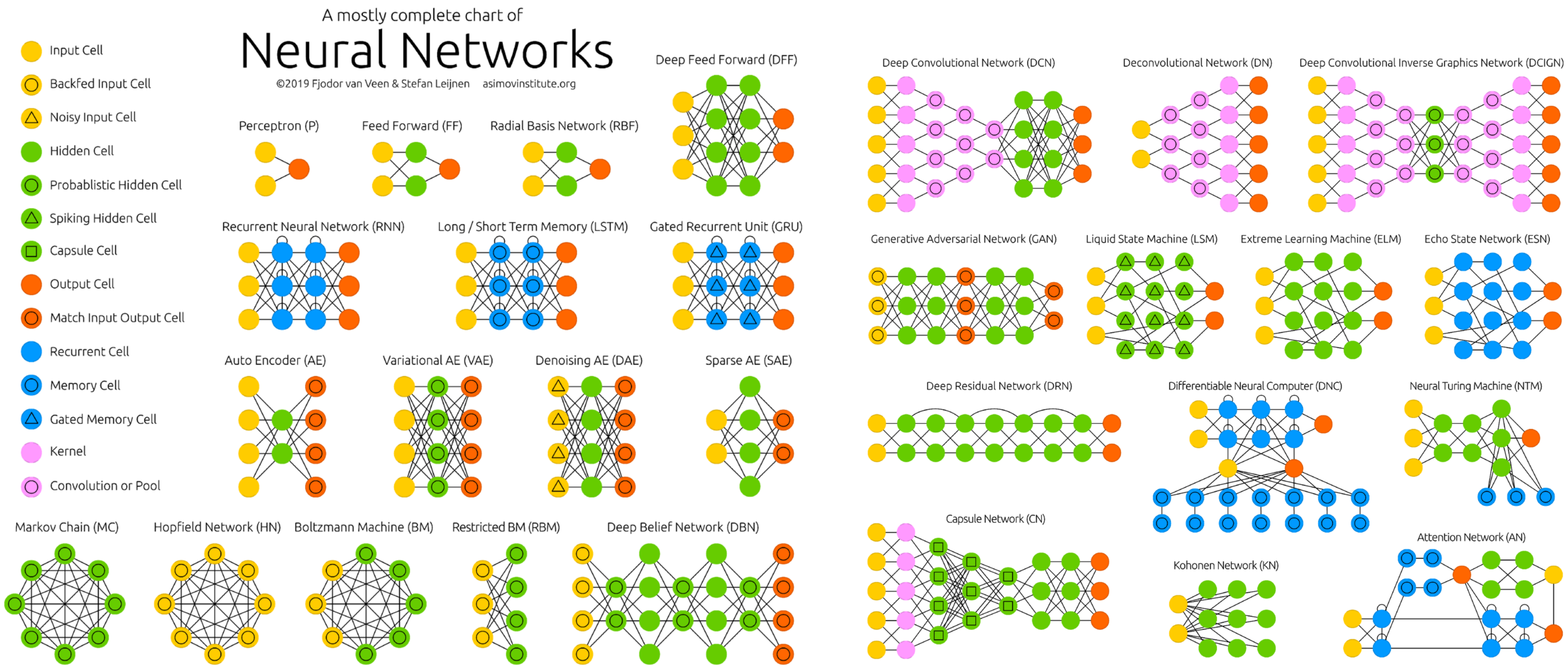
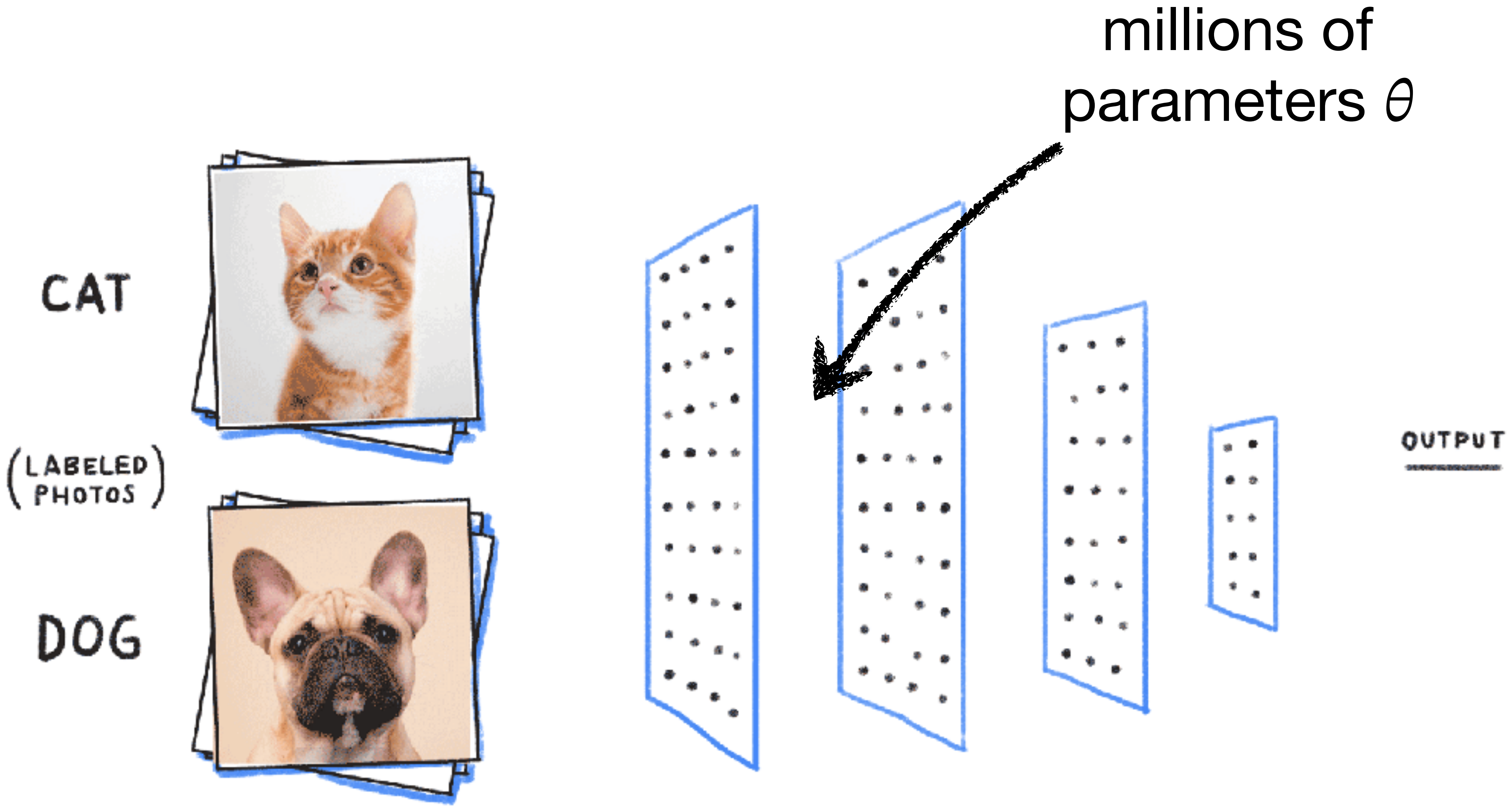


Illustration by Fjodor van Veen & Stefan Leijnen

Training a neural network

Having the architecture, how do we find the weights?



Animation courtesy of Aakash Srivastava

Training a neural network

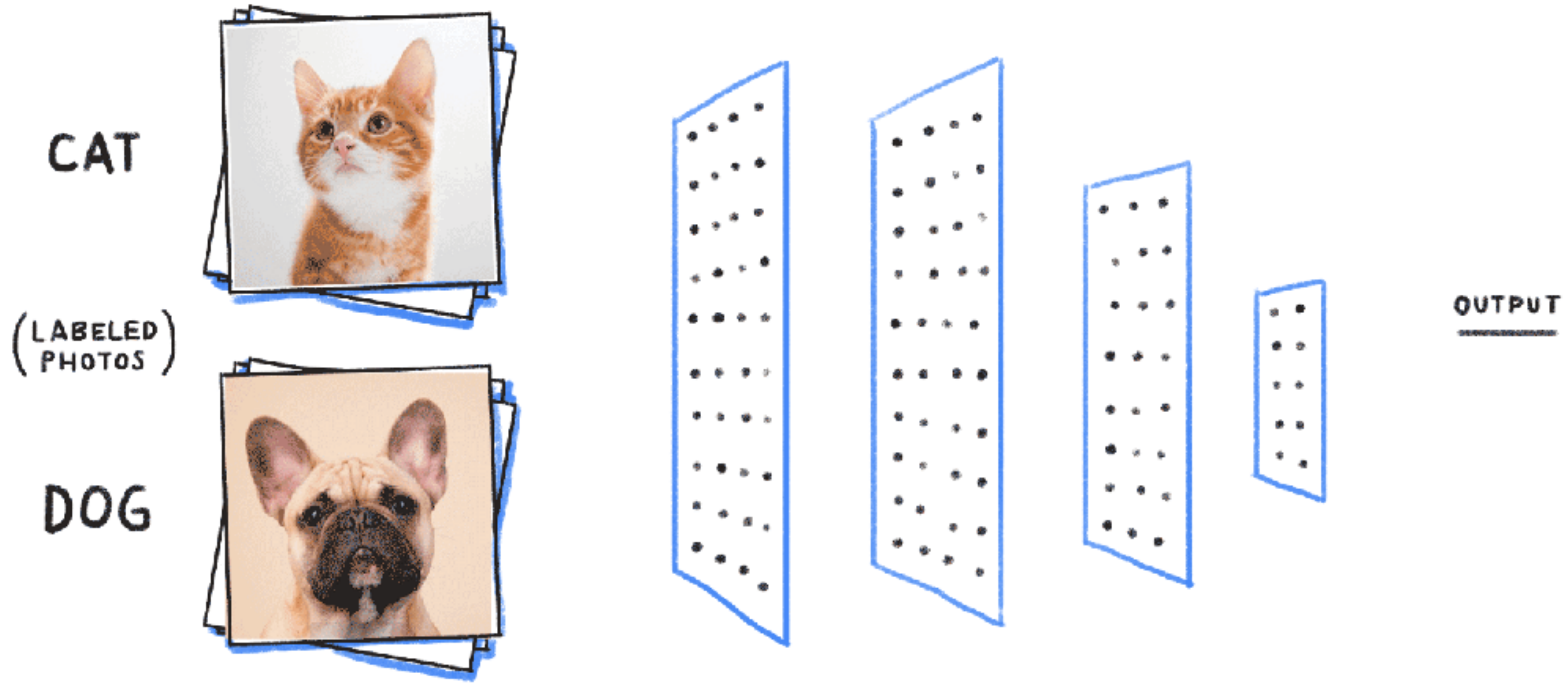
Having the architecture, how do we find the weights?

- Given a data set $\mathscr{D} = \{(x_i, y_i\}_{i=1}^{N}$

loss(θ) = # of mis-classified training images

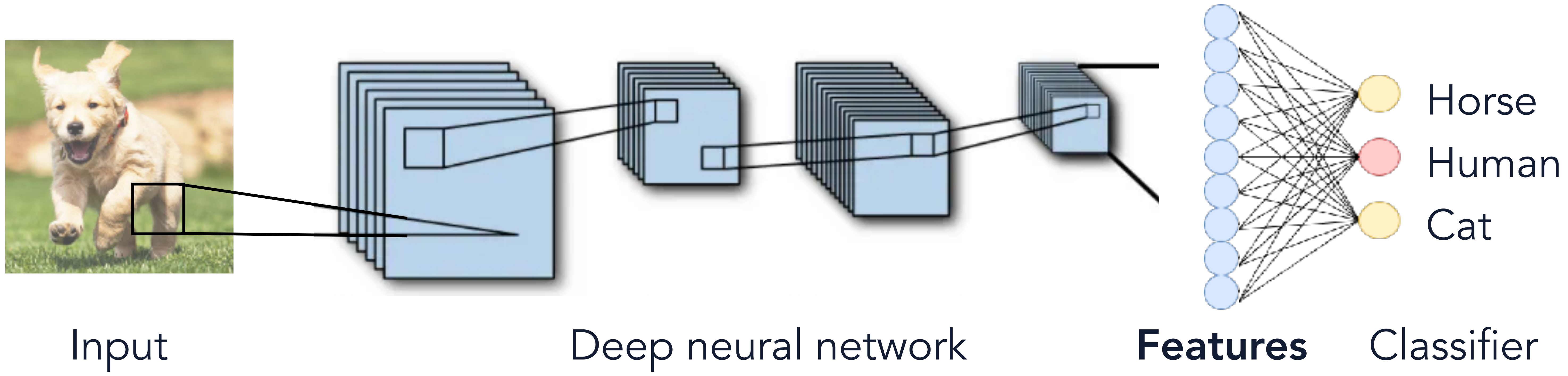
$$\theta^* = \operatorname*{argmin}_{\theta} \operatorname{loss}(\theta)$$

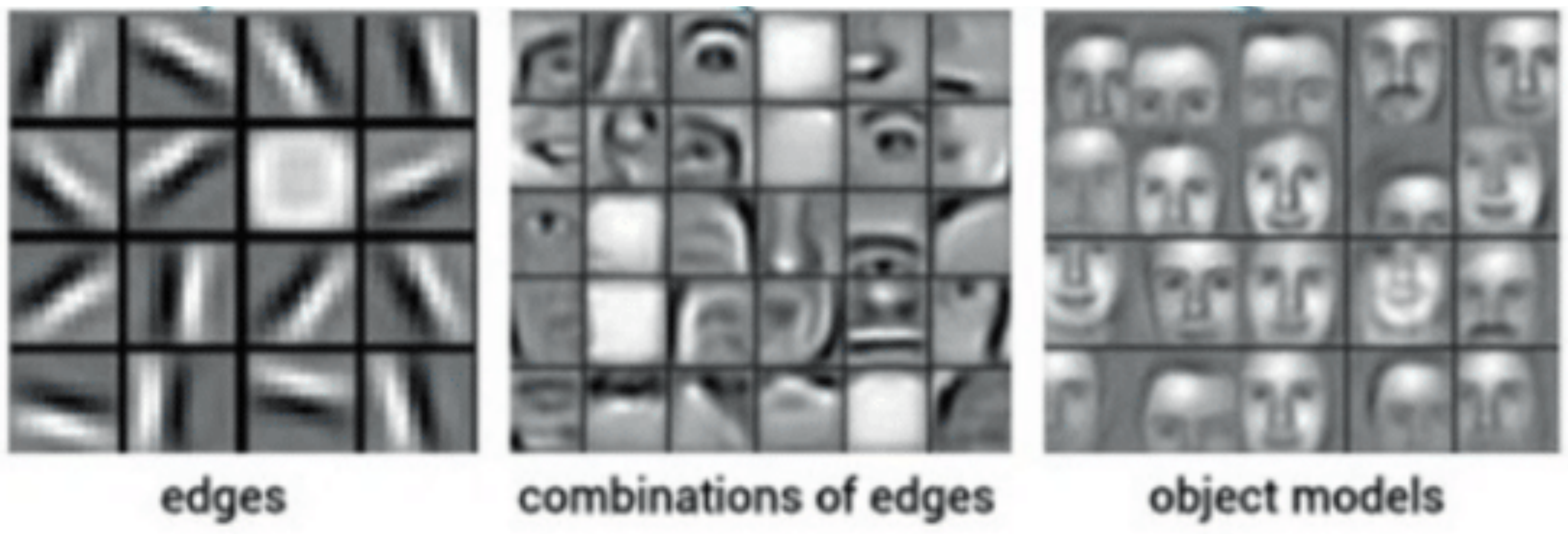
$$\epsilon_t = \operatorname{loss}(\theta^*)$$



Animation courtesy of Aakash Srivastava

How is this different from classical ML?

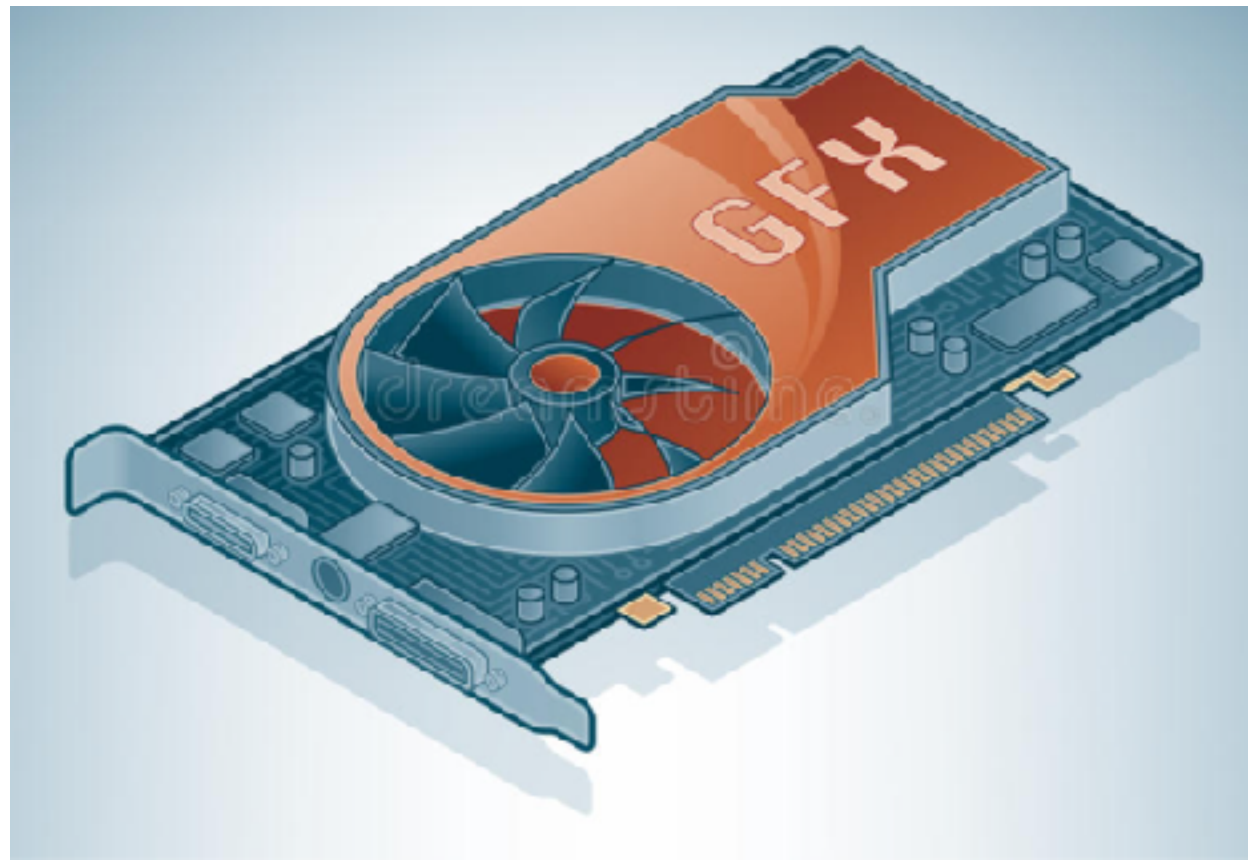




What enabled this success?

Three drivers for the success of neural networks

- It was known since the 1990s that convolutional neural networks could be trained successfully on image classification tasks.
- What was missing back then?



Computing power

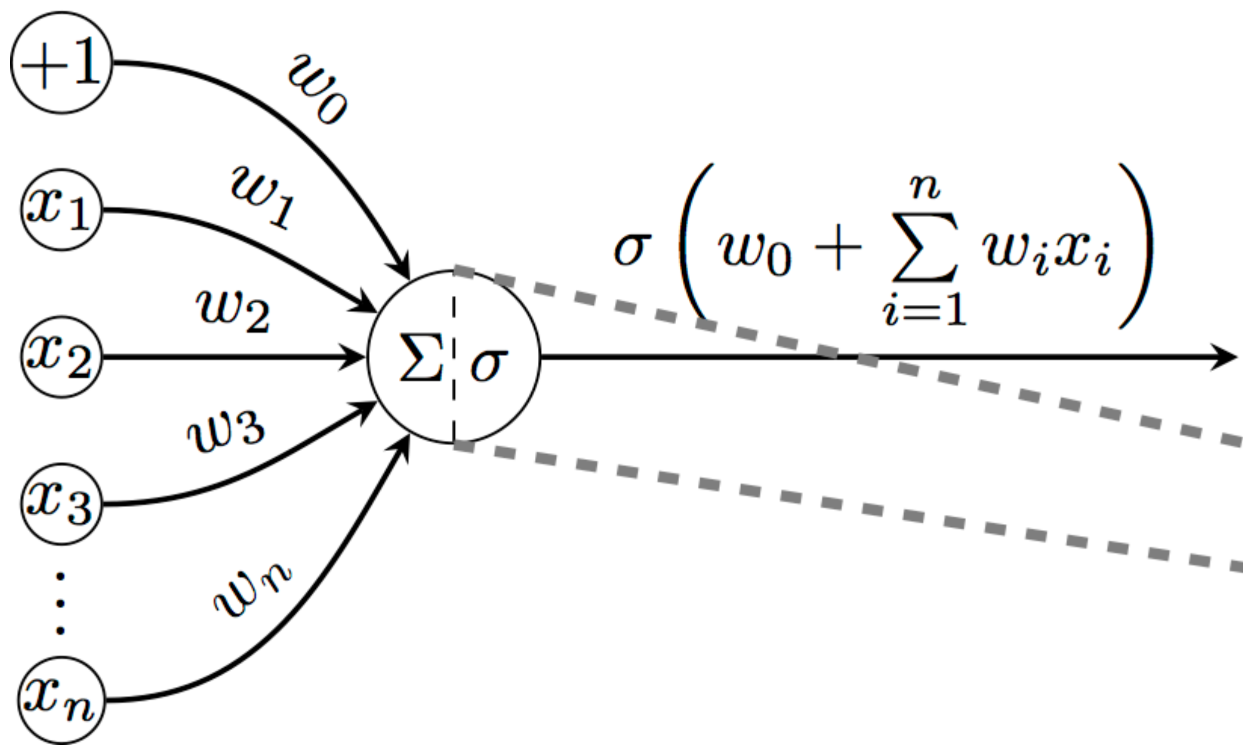
Data (internet, social media)

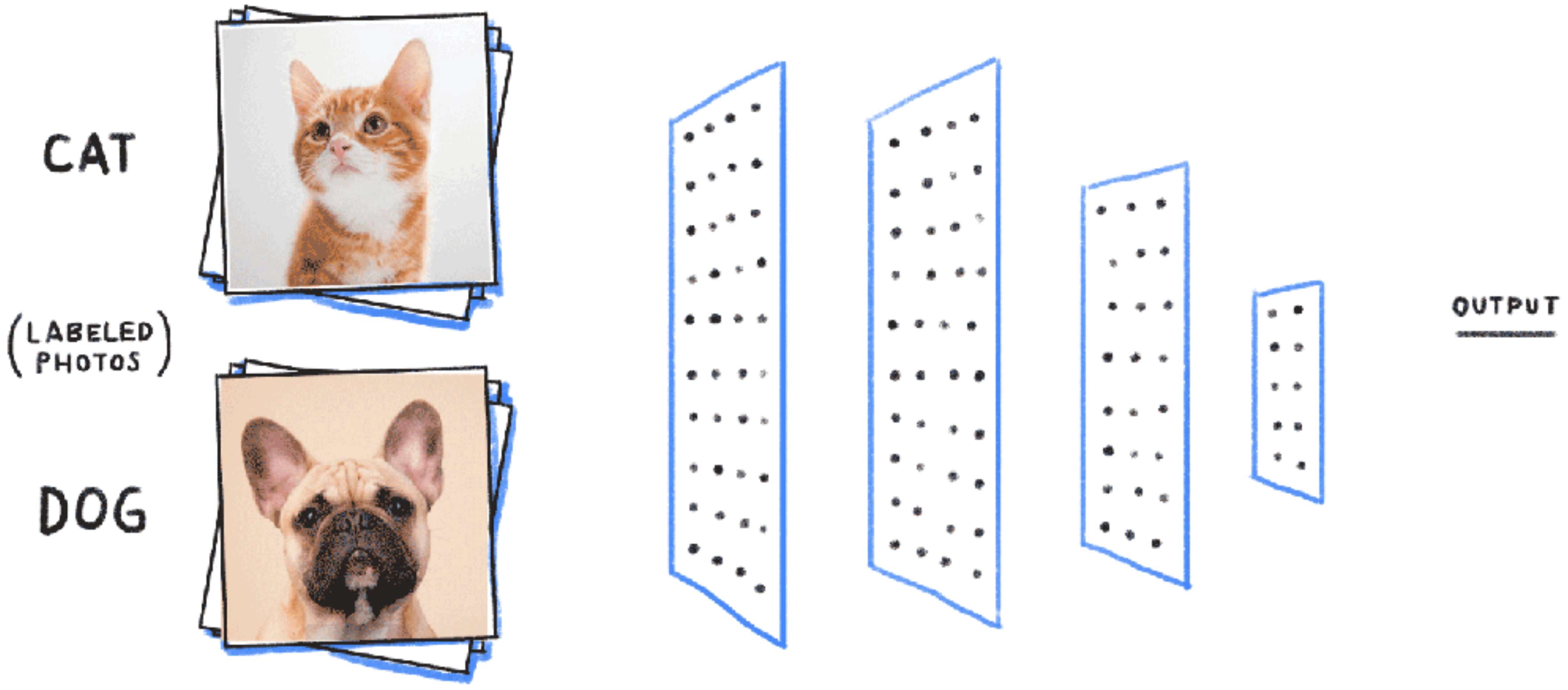
Python & Github ecosystem

A first summary

What are neural networks?

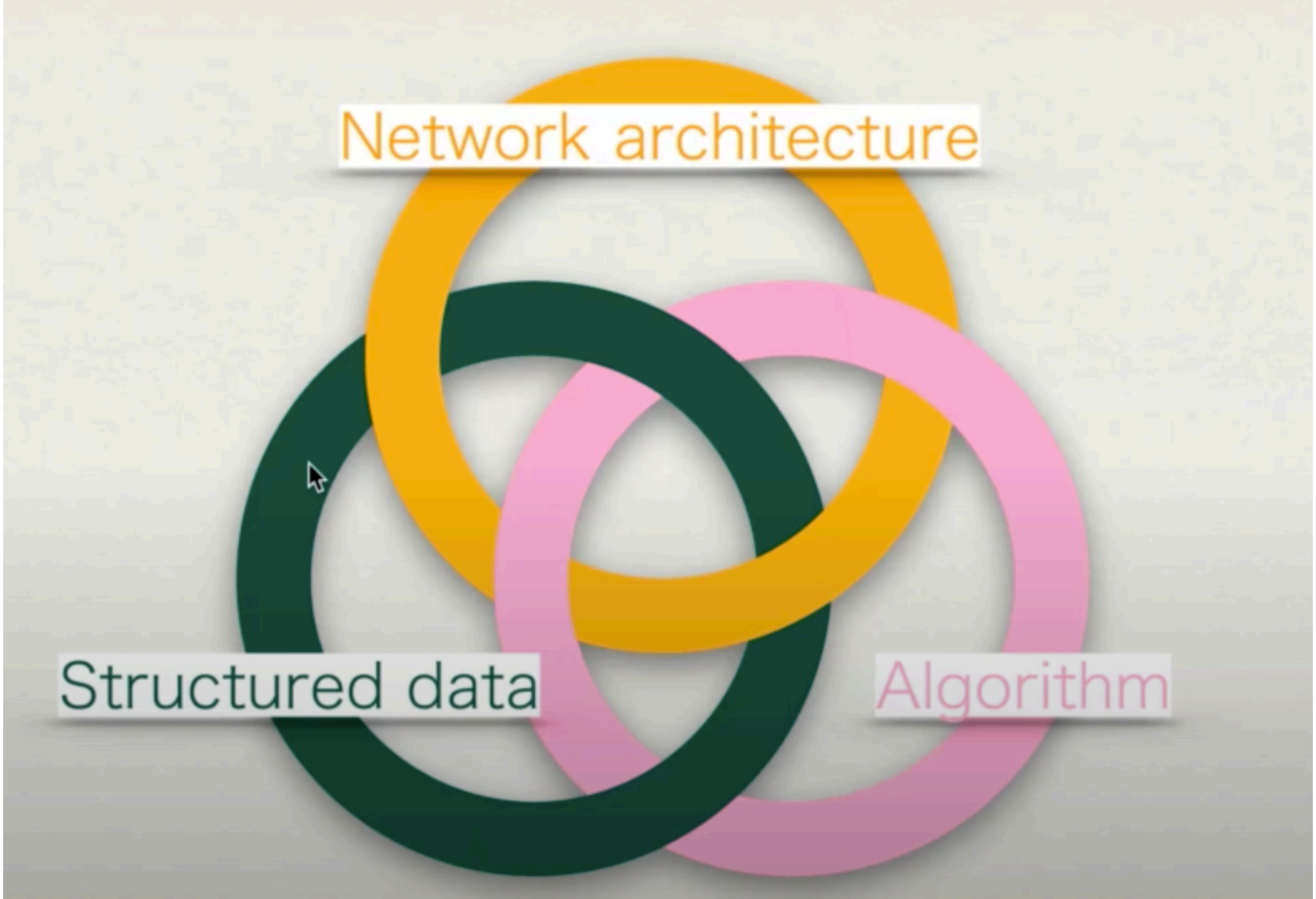
- A **Neuron** represents as a single number, nonlinearity(weighted sum of inputs).
- Many neurons yield a neural **network**.
- The ordering & wiring of neurons determines the **architecture** of the neural network.
- Training neural networks by stoch grad desc learns a **hierarchical set of features** from data.





Part II

How can we make sense of neural nets?



Network architecture
Structured data
Algorithm

Expressivity

Are neural networks rich enough as a function class?

- Cybenko et al. 1989, Barron 1993: **Universal approximation theorem.**
  - Consider two layer neural networks $f_\theta(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} a_i\sigma\left(\langle \mathbf{w}_i, \mathbf{x}\rangle + b_i\right)$

Theorem 2.2 (Barron, 1993)**.** *Assume* P *t be supported on* $\mathsf{B}(0, r)$*, and let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a function with Fourier transform* F*:* $f(\boldsymbol{x}) = \int e^{i\langle \omega, z\rangle} F(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}$. *Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be such that* $\lim_{t\to\infty} \sigma(t) = 1$, $\lim_{t\to\infty} \sigma(t) = 0$.

Define

$$N(\varepsilon) \equiv \frac{1}{\varepsilon}\left(2r\int \|\boldsymbol{\omega}\|_2\, |F(\boldsymbol{\omega})|\mathrm{d}\boldsymbol{\omega}\right)^2 . \qquad (2.14)$$

Then there exists a network of the form (2.11) with $N(\varepsilon)$ *hidden unit achieving error* $\mathbb{E}\{(\hat{f}(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x}))\} \le \varepsilon$.

Expressivity

Are neural networks rich enough as a function class?

- Cybenko et al. 1989, Barron 1993: **Universal approximation theorem.**
    - Two-layer neural networks $f_\theta(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma\left(\langle \mathbf{w}_i, \mathbf{x}\rangle + b_i\right)$

      can approximate any function arbitrarily well, provided they are wide enough.
- The problem: UAP is an **existence** theorem — (how) can we find the right weights?

Training a neural network in theory

It's not so easy!

TRAINING A 3-NODE NEURAL NETWORK IS NP-COMPLETE

Avrim Blum*
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

Ronald L. Rivest†
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

Advances in Neural Information Processing (1989)

TRAINING A 3-NODE NEURAL NETWORK IS NP-COMPLETE

Avrim Blum*
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

Ronald L. Rivest†
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

Advances in Neural Information Processing (1989)

Extends a previous result by Judd (1987)

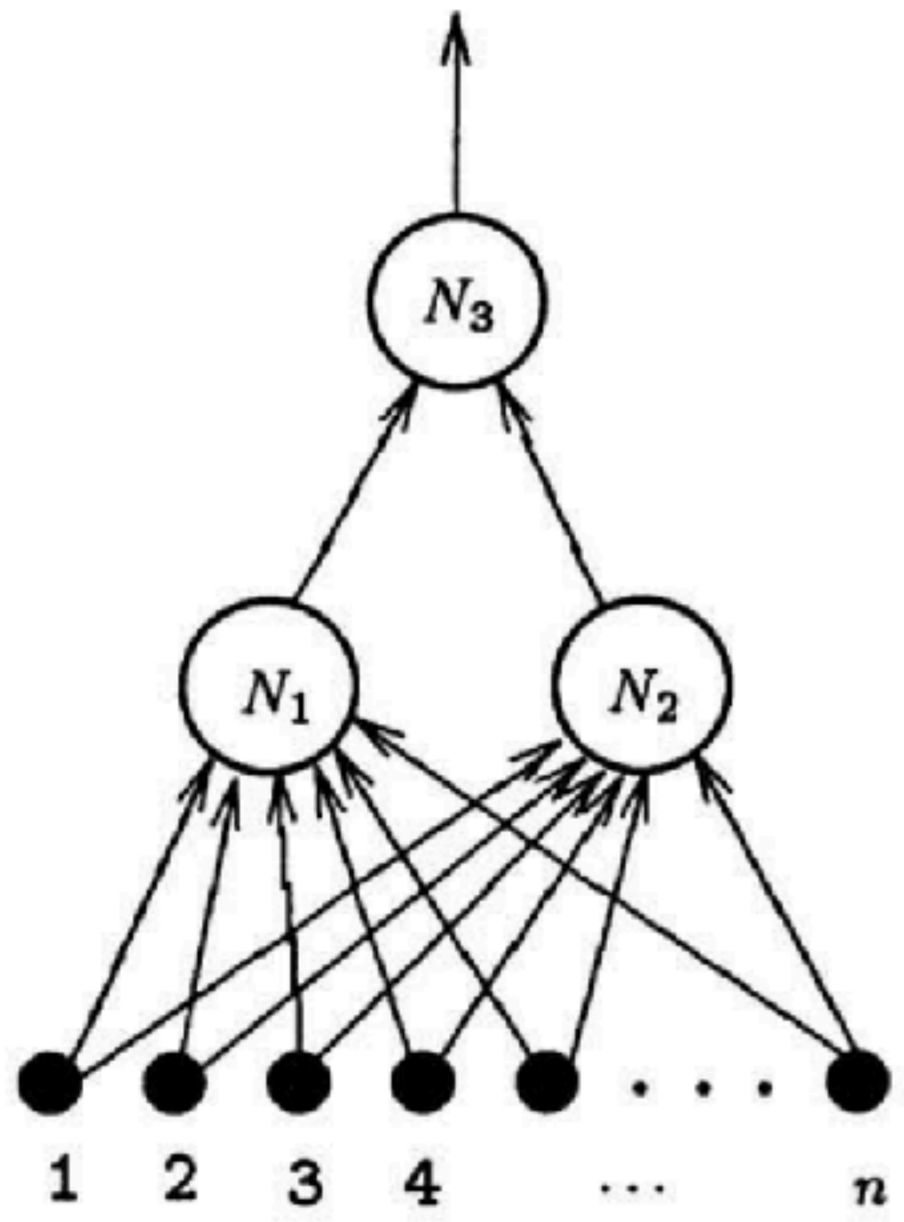


Figure 1: The three node neural network.

Given: A set of *O(n)* training examples on *n* inputs

Question: Do there exist linear threshold functions such that the three-node network fits the training set?

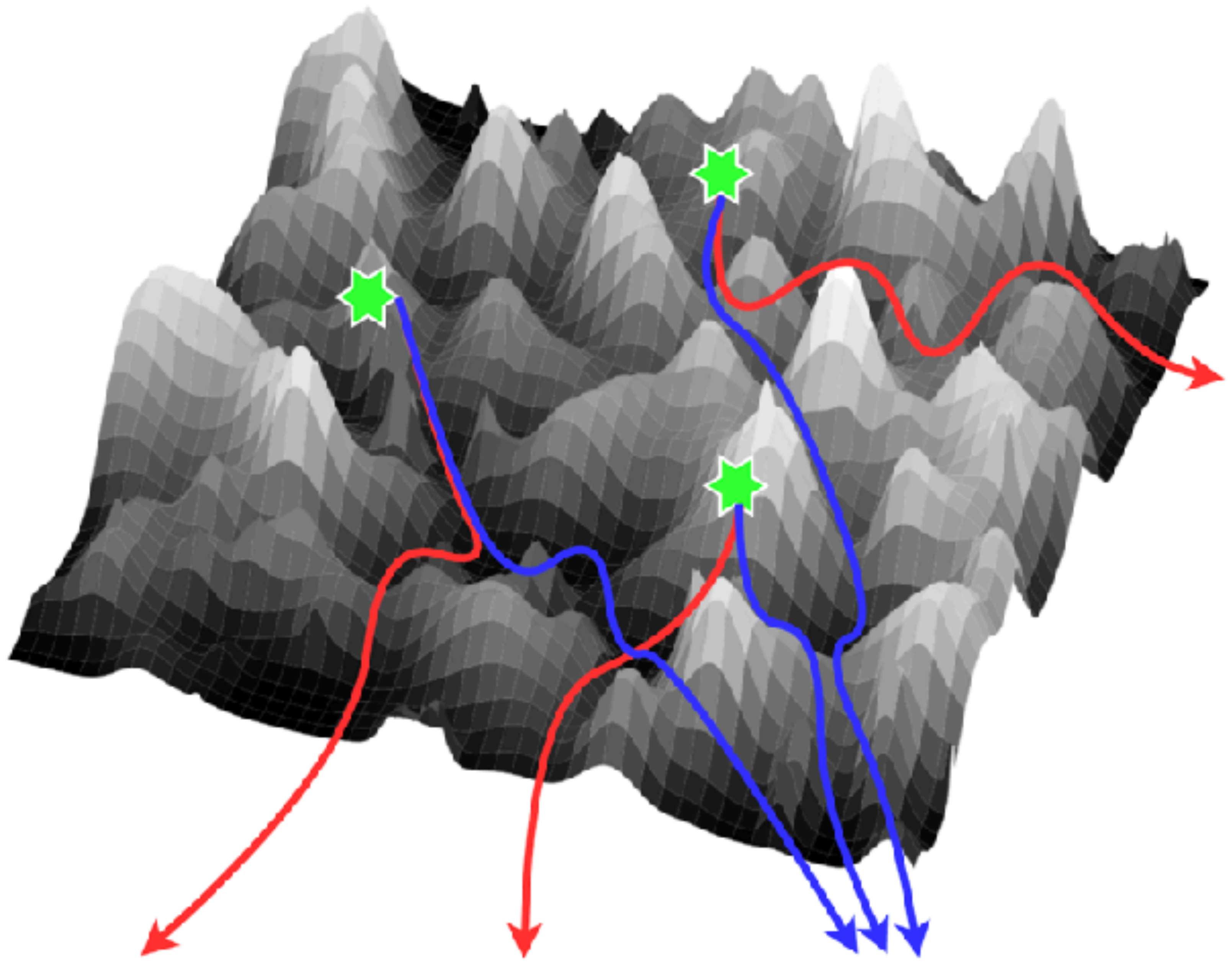
NP hard

😱😱😱

The success of SGD

Don't worry about theorems, do it anyway…

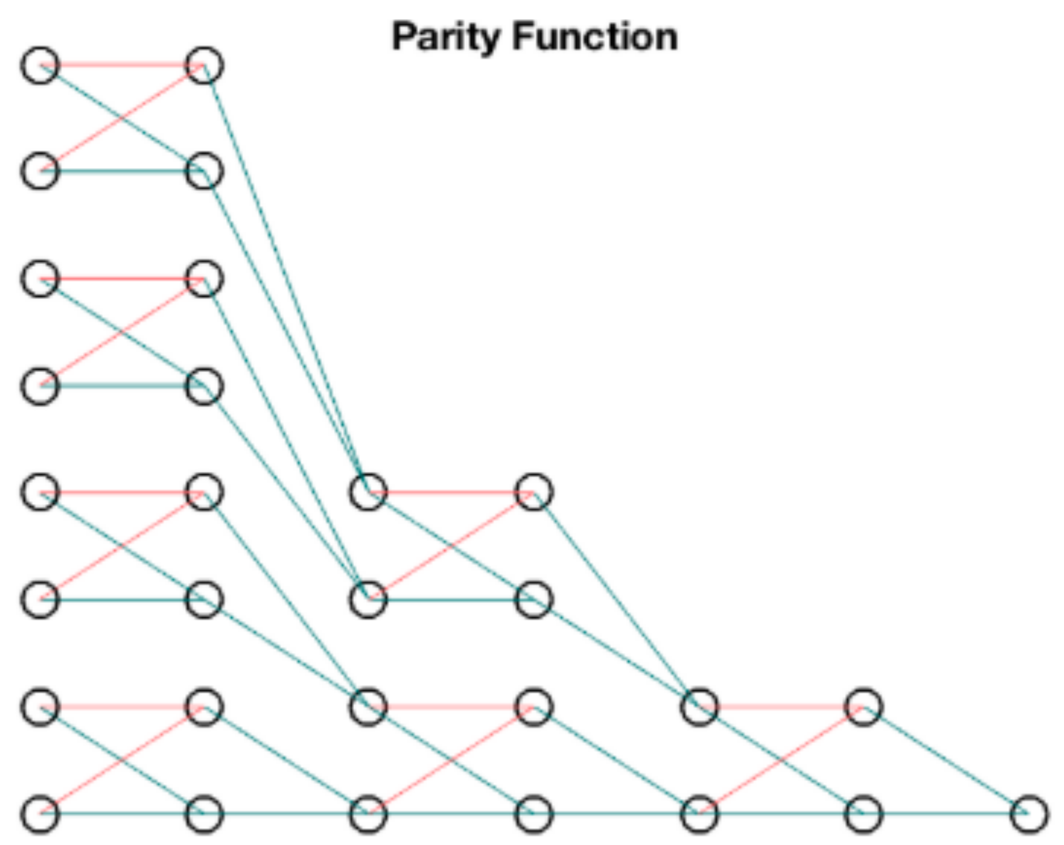
- In practice, SGD (and its variants) works extremely well to train neural networks. Why?
- We cannot understand this **statically** (by analysing the loss landscape)
  - It does have global minima which generalise poorly.
- Need to understand where the **learning dynamics** of neural networks lead us.

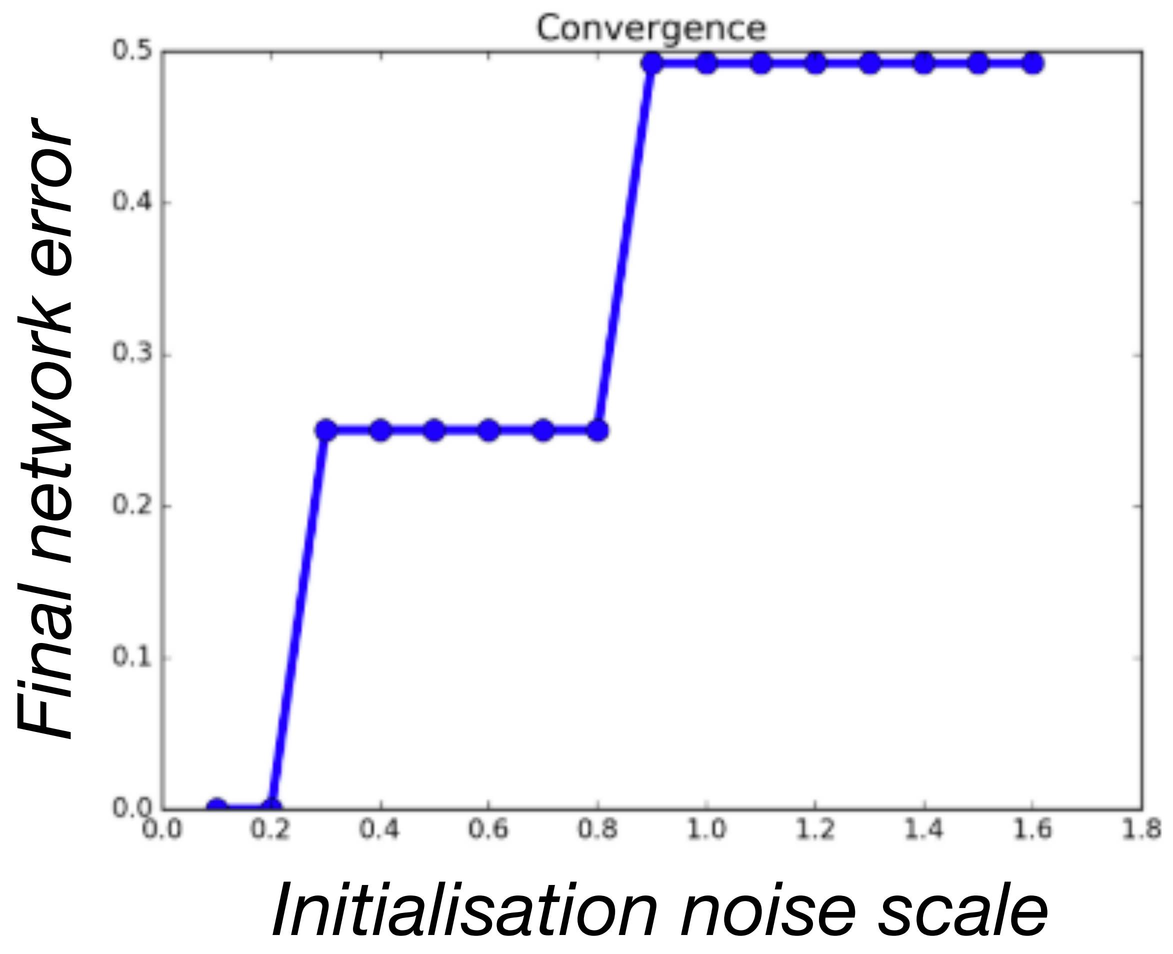


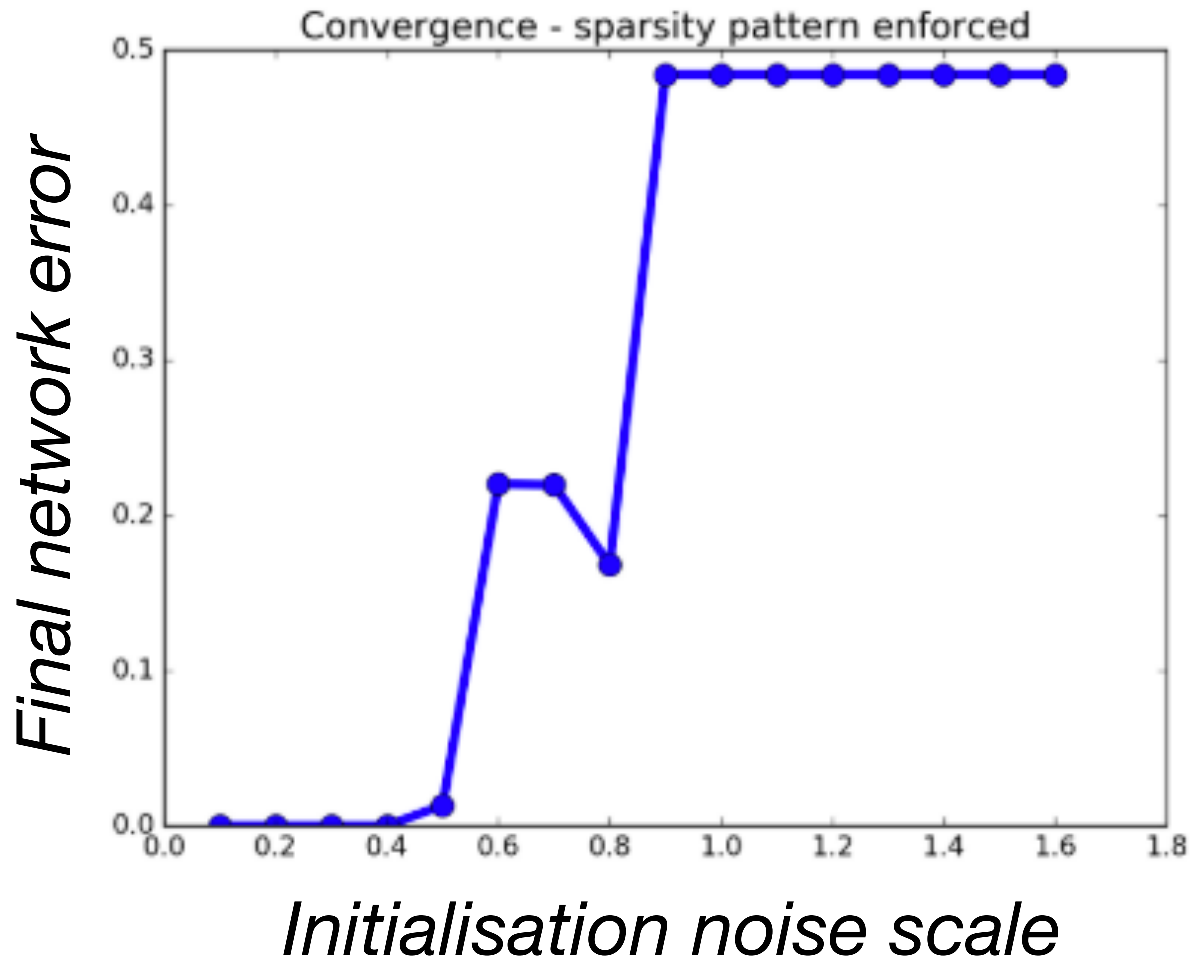
Can we learn anything with gradient descent, efficiently?

No.

- Try to learn the parity of a binary string.







The parity function can be **expressed exactly** by this neural network

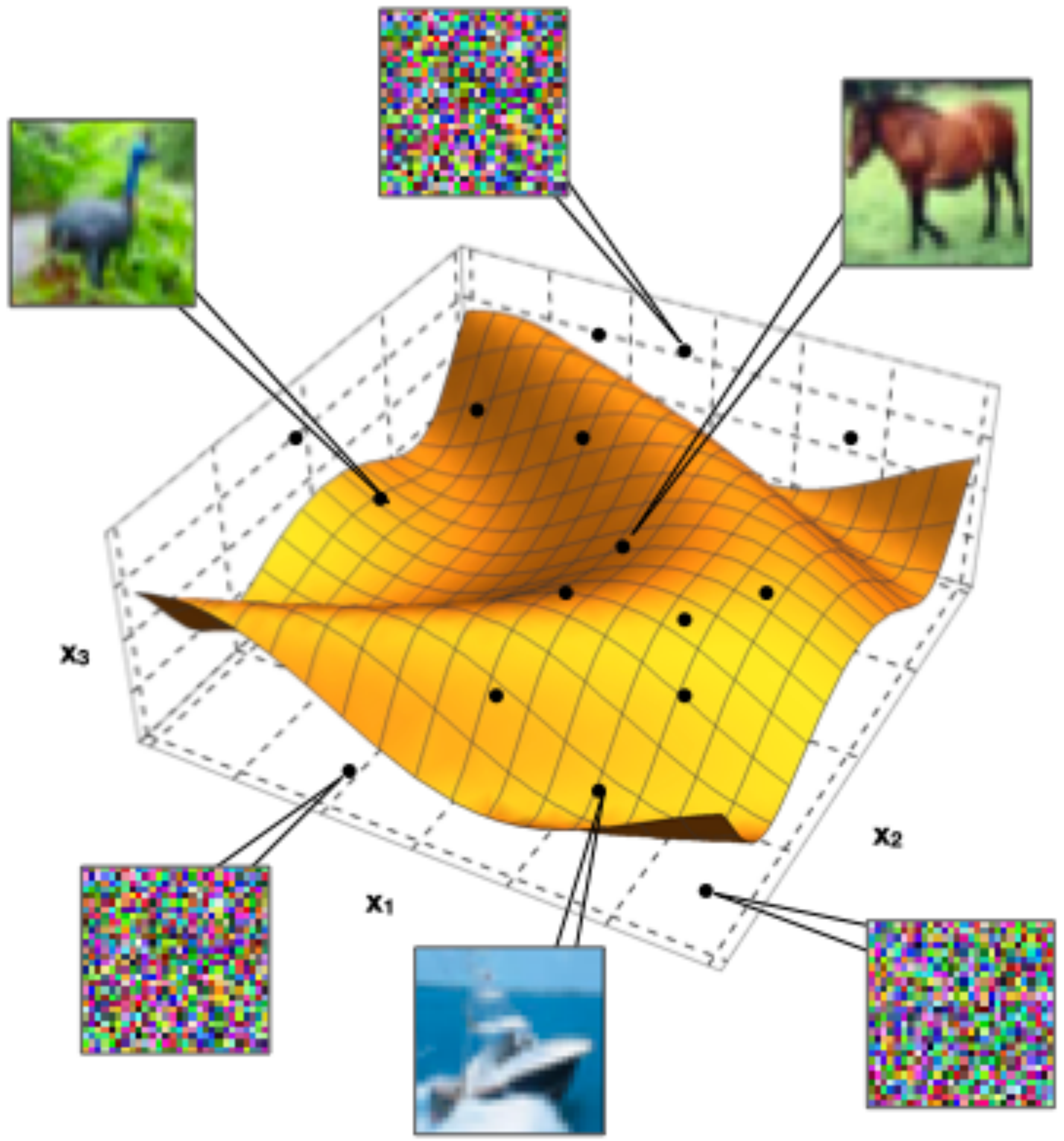
Gradient descent struggles to learn this function…

… even if we explicitly impose the sparsity of the target network.

The structure of the data

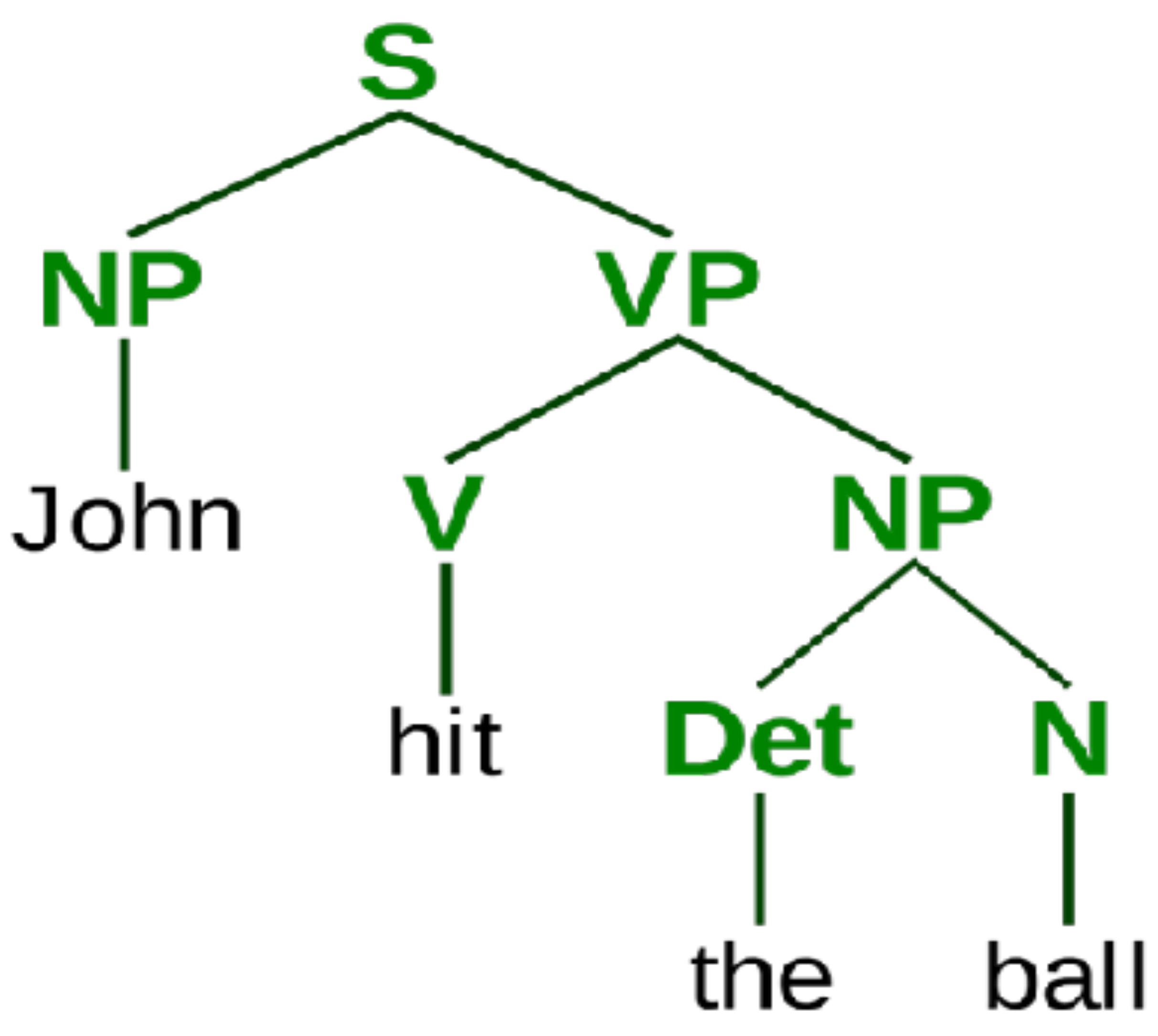
The final ingredient in the success of neural networks

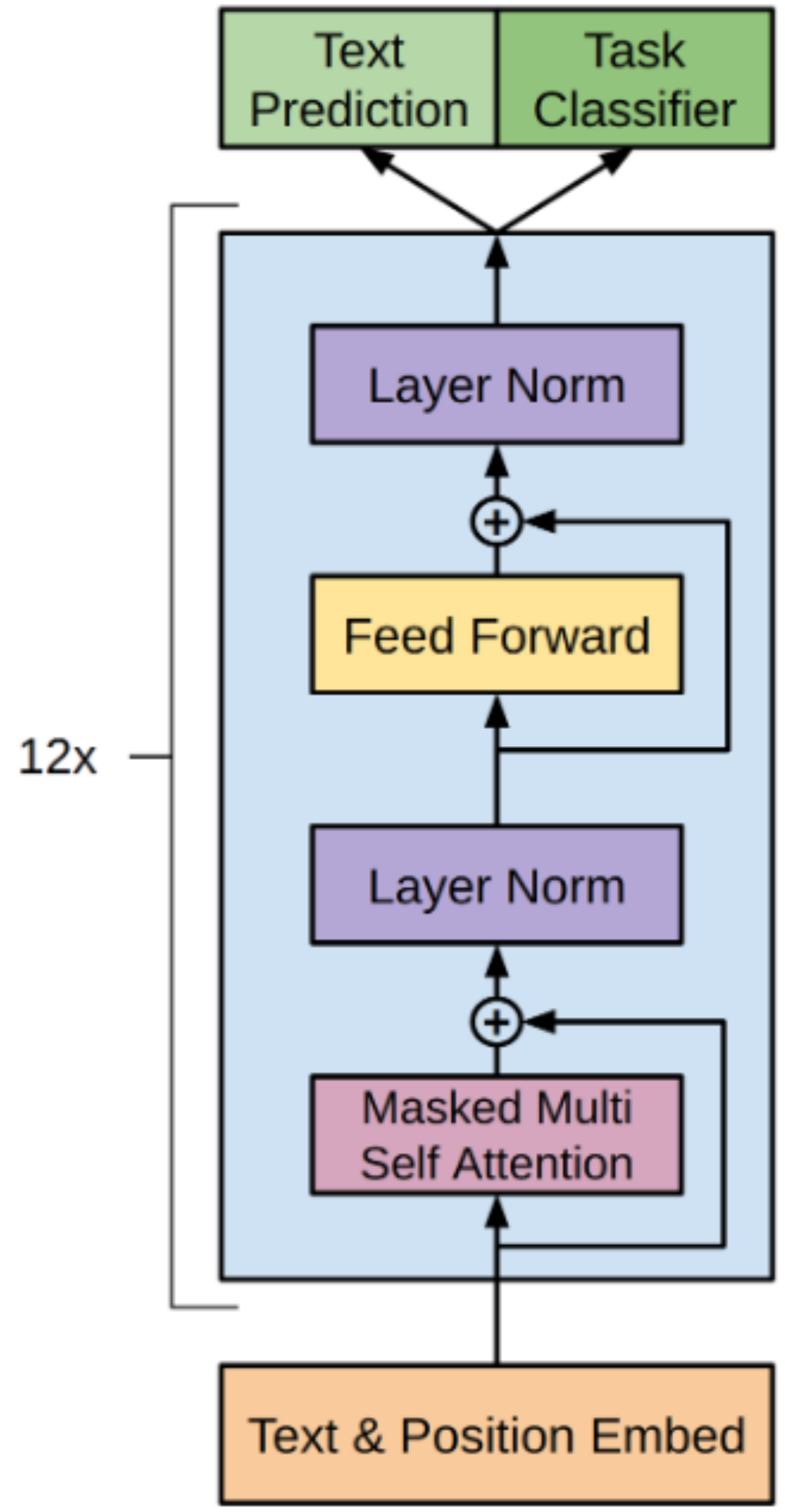
- **Curse of dimensionality:** if we only assume that the optimal target function $f^*(\mathbf{x})$ is Lipschitz-continuous, i.e. $|f^*(\mathbf{x}') - f^*(\mathbf{x})| \leq L||\mathbf{x}' - \mathbf{x}||$,
  - you need $n \sim (1/\epsilon)^{d/2+1}$ samples to learn the function within an error ε…
- Neural networks are therefore able to exploit the structure in the data.
- We know a lot about the structure of images, but much less about the structure of language…



From hierarchical data models to deep networks

Reverse engineering images and text





The challenge for a modern theory of neural networks

A challenge for mathematics, theoretical physics, and computer science

- Neural networks are made of simple, elementary building blocks.
- How learning **emerges** from the interaction of billions of neurons is an enormous theoretical challenge.
- A modern theory for deep learning needs to account for the interplay of
  - learning dynamics,
  - data structure,
  - and network architecture.