CSCI 5922, Spring 2020

Lecture 2

https://sites.google.com/corp/colorado.edu/csci-5922-spring-2020/home

DL applications

Computer vision
Image classification
Object detection
Face detection & recognition
Pose estimation
Motion tracking
Action recognition
Image captioning
Face & scene generation
Medical image analysis
Satellite image analysis

NLP
Machine translation
PoS tagging
Syntax parsing
Named entity recognition
Question answering
Smart compose
Search ranking
Sentiment analysis
Spam filtering
Document classification

Other
Speech recognition (Google assistant, Siri)
Atari
AlphaGo, AlphaZero
Google Datacenter Cooling
Autonomous vehicles
Stock price prediction
No-limit hold’em
Weather prediction
Protein folding

MNIST

- Handwritten digits
- 28x28 grayscale
- 60,000 training examples
- 10,000 test examples

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Why do we need a test set?

- Easy to write a program to memorize the training set
- Easy to fit a model to memorize the training set
- Performance is (nearly always) slightly worse on test set than training set

MNIST

How to teach a computer

- Connectionist
  - Use neural inspired architectures
  - Train connection weights between a large number of ‘dumb’ units
  - Distributed representation
  - Black-box
- Symbolic
  - Represent input as concepts or symbols
  - Give computers facts and rules to process input symbols
  - Expert systems
  - Graph-traversal algorithms

What makes a 2?

Source: [Geoffrey Hinton's 2012 Coursera course lectures]

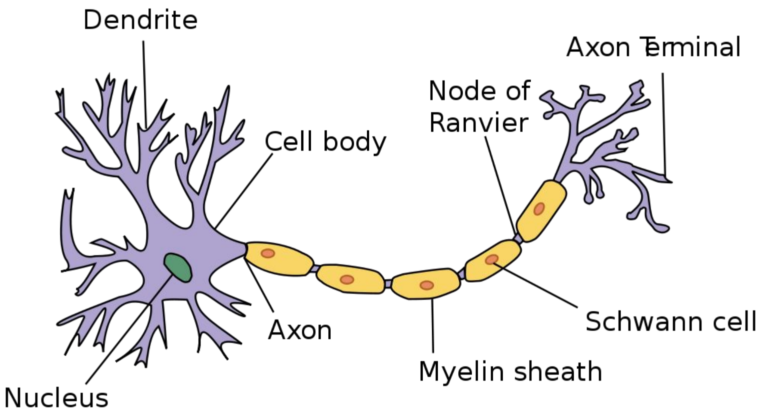
Neural information processing

- Structure of neuron first described by Ramon y Cajal

Courtesy of the Cajal Institute and the Spanish National Research Council

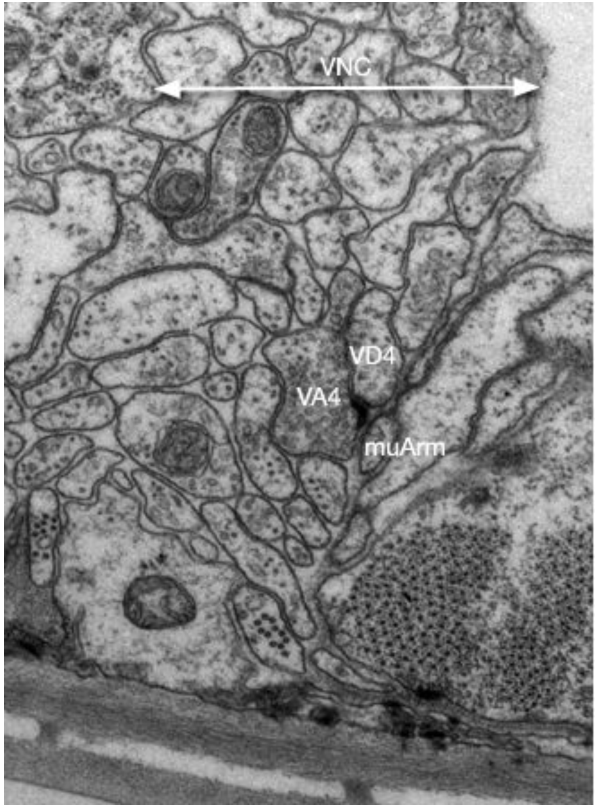
Neural information processing

- Input from other neurons received on dendrites
- Ion channels affect cell voltage
- Nonlinear feedback effects cause action potentials (depolarization)
- Action potentials propagate to axon terminal
- Synapses on axon terminals release neurotransmitters (dopamine, GABA, glutamate)
- Signal is propagated to post-synaptic neuron



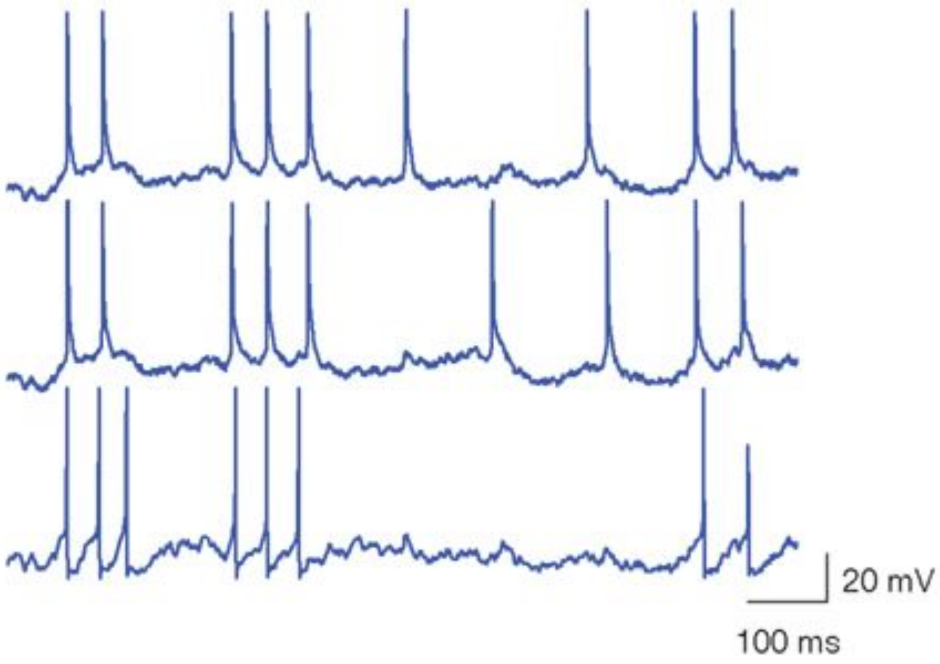
PSA: I am not a neuroscientist! This is how machine learning people think of neurons.

Synapse in *c. elegans*



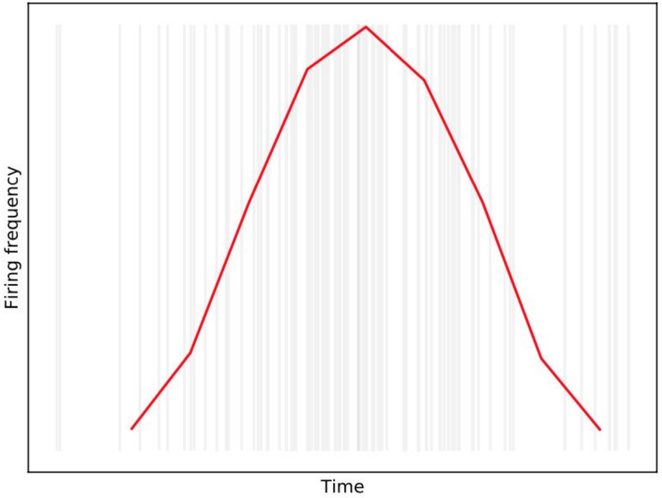
Source:
https://www.wormatlas.org/neuronalwiring.html

Spike trains



Source: Rossant, C., Goodman, D. F., Fontaine, B., Platkiewicz, J., Magnusson, A. K., & Brette, R. (2011). Fitting neuron models to spike trains. *Frontiers in neuroscience*, *5*, 9.

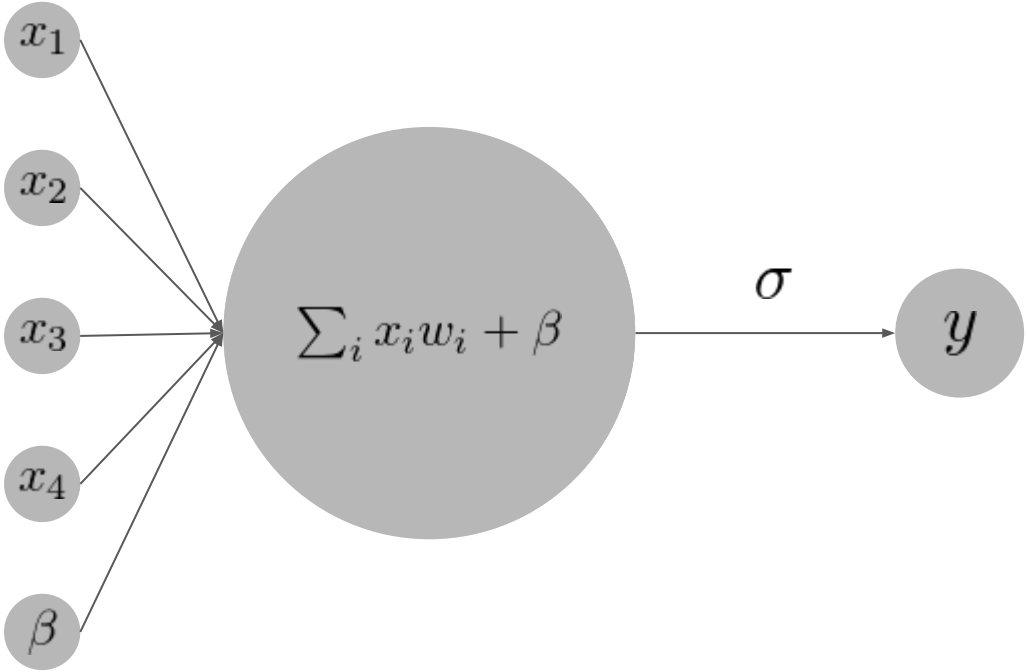
Spike rates

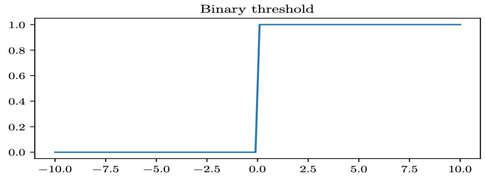


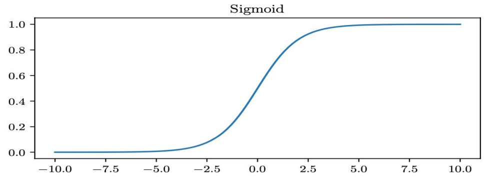
Smooth out instantaneous spikes
(e.g. with a kernel)

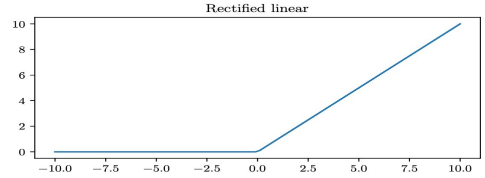
Artificial neuron

$$y = \sigma\left(\sum_i x_i w_i + \beta\right)$$

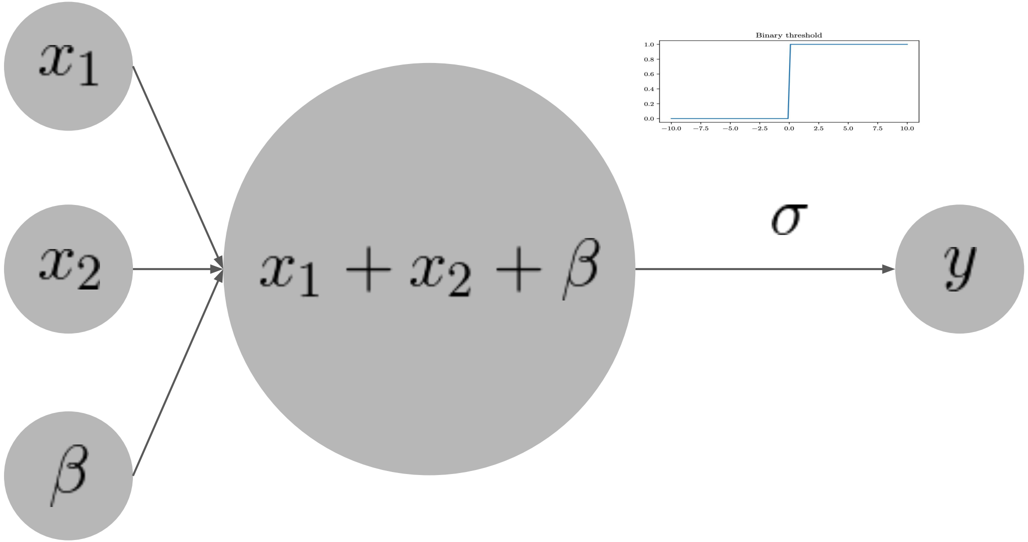








Binary threshold neuron (McCulloch-Pitts)



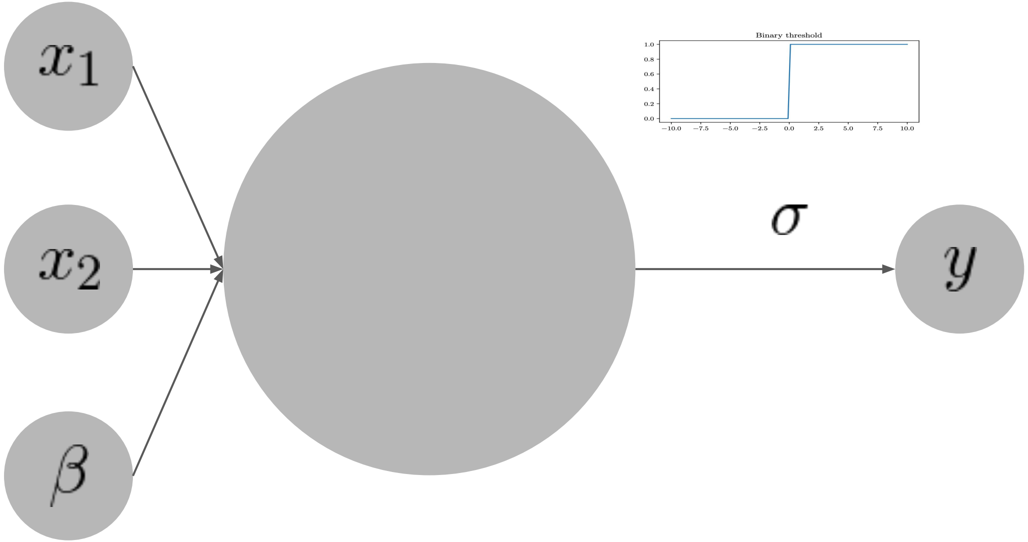
$$x_1, x_2 \in \{0, 1\}$$

$$\text{set } \beta = -1$$

$$\text{so } y = \mathbb{1}(x_1 + x_2 > 1)$$

Implements AND

Binary threshold neuron (McCulloch-Pitts)

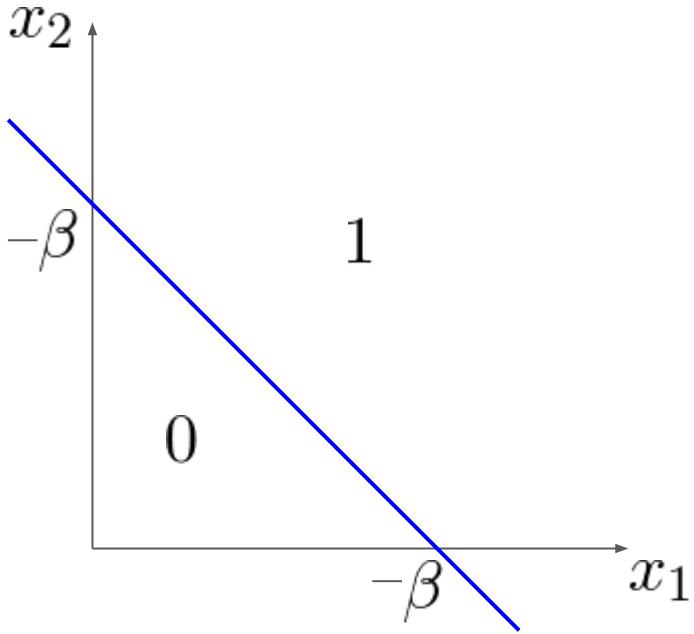


$$x_1, x_2 \in \{0, 1\}$$
$$\beta = 0$$

Implements OR

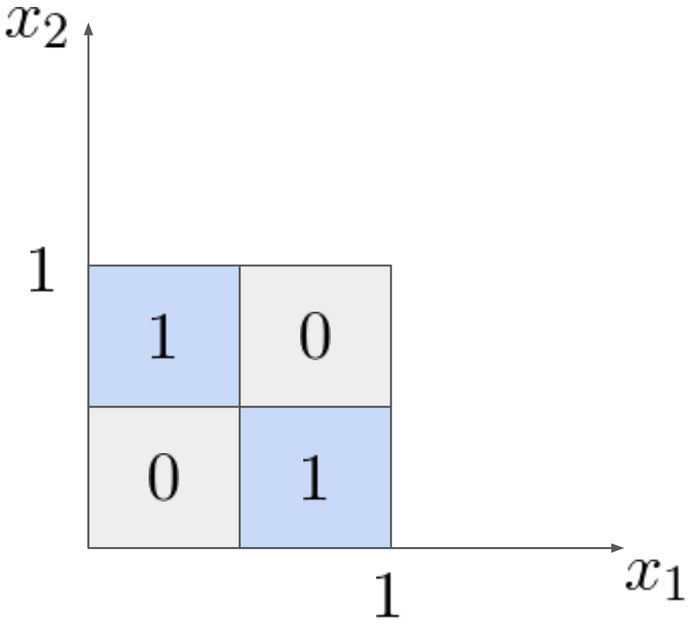
Binary threshold neuron (McCulloch-Pitts)

$$(x_1, x_2) \mapsto \mathbb{1}(x_1 + x_2 > -\beta)$$



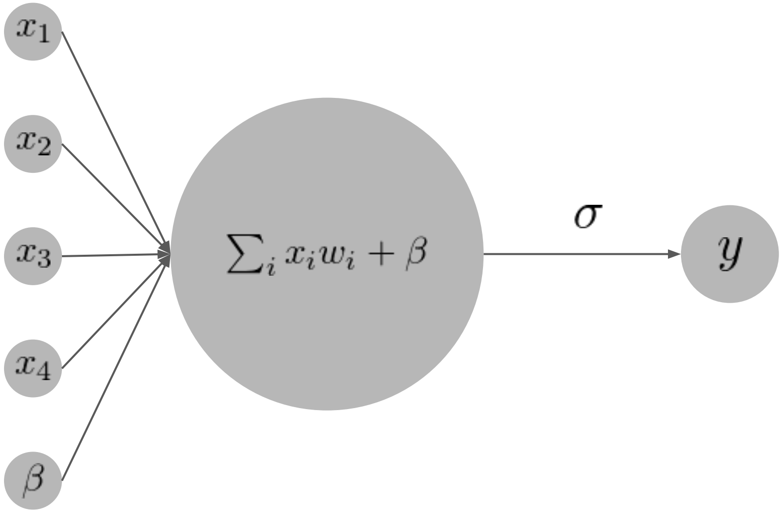
Binary threshold neuron (McCulloch-Pitts)

$$(x_1, x_2) \mapsto \mathbb{1}\{x_1 + x_2 > \beta\}$$



https://playground.tensorflow.org/

Perceptron



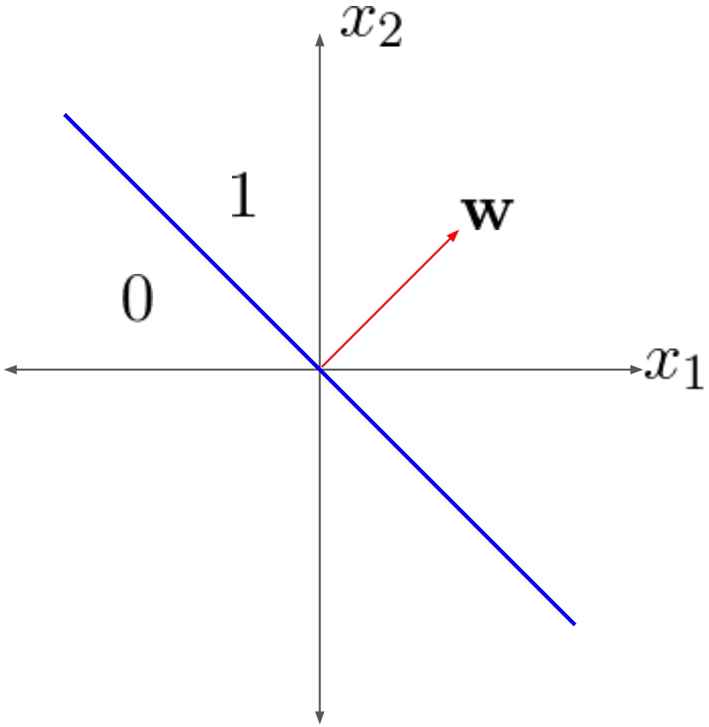
$$\mathbf{x} = (x_1, x_2, ..., x_p)$$

$$\mathbf{w} = (w_1, x_2, ..., w_p)$$

$$y = \mathbb{1}\left\{\mathbf{x}^T \mathbf{w} > 0\right\}$$

(absorb bias into **x** and **w**)

Perceptron



Perceptron

The Perceptron Algorithm: Start with the all-zeroes weight vector $\mathbf{w} = \mathbf{0}$. Then repeat the following until $\mathbf{x}^T\mathbf{w}$ has the correct sign for all $\mathbf{x} \in S$ (positive for positive examples and negative for negative examples):

1. Let $\mathbf{x} \in S$ be an example for which $\mathbf{x}^T\mathbf{w}$ does not have the correct sign.

2. Update as follows:

   (a) If $\mathbf{x}$ is a positive example, let $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{x}$.

   (b) If $\mathbf{x}$ is a negative example, let $\mathbf{w} \leftarrow \mathbf{w} - \mathbf{x}$.

Source: Hopcroft, John, and Ravi Kannan. "Foundations of data science." (2014).

Perceptron

- Training in action (video)
- It was really built! (video)
- Somewhat overhyped

“the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence ... It is expected to be finished in a year at a cost of $100,000 … Dr. Rosenblatt said Perceptrons might be fired to the planets as mechanical space explorers” - NY Times, July 8 1958

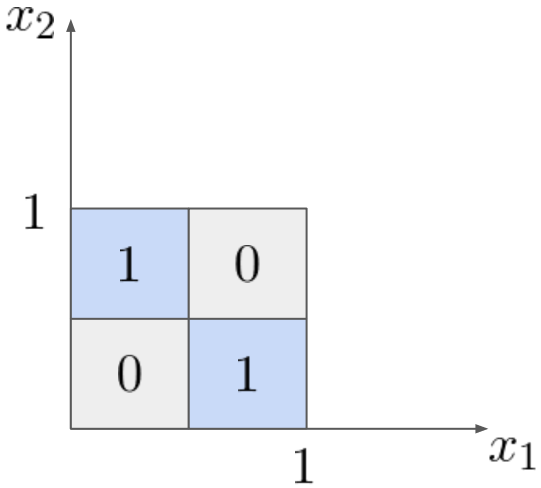
Perceptron convergence

Theorem 5.1 *If there exists a vector* $\mathbf{w}^*$ *such that* $\mathbf{x}^T\mathbf{w}^* \geq 1$ *for all positive examples* $\mathbf{x} \in S$ *and* $\mathbf{x}^T\mathbf{w}^* \leq -1$ *for all negative examples* $\mathbf{x} \in S$ *(i.e., a linear separator of margin* $\gamma = 1/|\mathbf{w}^*|$*), then the number of updates made by the Perceptron algorithm is at most* $R^2|\mathbf{w}^*|^2$*, where* $R = \max_{\mathbf{x}\in S} |\mathbf{x}|$.

Source: Hopcroft, John, and Ravi Kannan. "Foundations of data science." (2014).

Neural network history

- 1969 - Minsky & Papert, Perceptrons: An introduction to computational geometry
  - There are many things a perceptron can’t learn to do
  - Perceptrons limited to linearly separable data
  - Multilayer perceptrons?



Nonlinear processing in a single neuron

"In contrast to typical all-or-none action potentials, dCaAPs were graded; their amplitudes were maximal for threshold-level stimuli but dampened for stronger stimuli. These dCaAPs enabled the dendrites of individual human neocortical pyramidal neurons to classify linearly nonseparable inputs—a computation conventionally thought to require multilayered networks."

"Traditionally, the XOR operation has been thought to require a network solution. We found that the dCaAPs' activation function allowed them to effectively compute the XOR operation in the dendrite by suppressing the amplitude of the dCaAP when the input is above the optimal strength".

Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., ... & Larkum, M. E. (2020). Dendritic action potentials and computation in human layer 2/3 cortical neurons. Science, 367(6473), 83-87.

Neural network history

- 1970-1985
  - Attempts to discover symbolic rule discovery algorithms
  - Expert systems
- 1986
  - Backpropagation - Rumelhart, Hinton, Williams (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.
  - Overcame many objections of Minsky & Papert
  - Renewed interest in connectionism
  - Backprop in sequence models: Mozer, M. C. (1995). A focused backpropagation algorithm for temporal. Backpropagation: Theory, architectures, and applications, 137.

1990-2010

- Classification and regression trees
- Bagging
- Boosting
- High-dimensional regression (lasso, ridge)
- Wavelets
- Kernel methods
    - Mercer kernel is inner product in infinite dimensional space
- Probabilistic methods
    - Bayes nets, hidden markov models, conditional random fields
    - Fully Bayesian
        - Variational inference, MCMC

2010s

- What's hot?
  - Backpropagation
  - Multi-layer perceptrons
  - CNNs
- What changed?
  - More data
  - Moore's law
  - GPUs
  - Frameworks (Tensorflow, PyTorch, Caffe, Theano)
  - A few new ideas: ReLU, dropout, batch-norm