UNIVERSITY OF
SOUTH CAROLINA

# Neuro-Edge:
# Neuromorphic-Enhanced Edge Computing

**Ramtin Zand, Ph.D.**

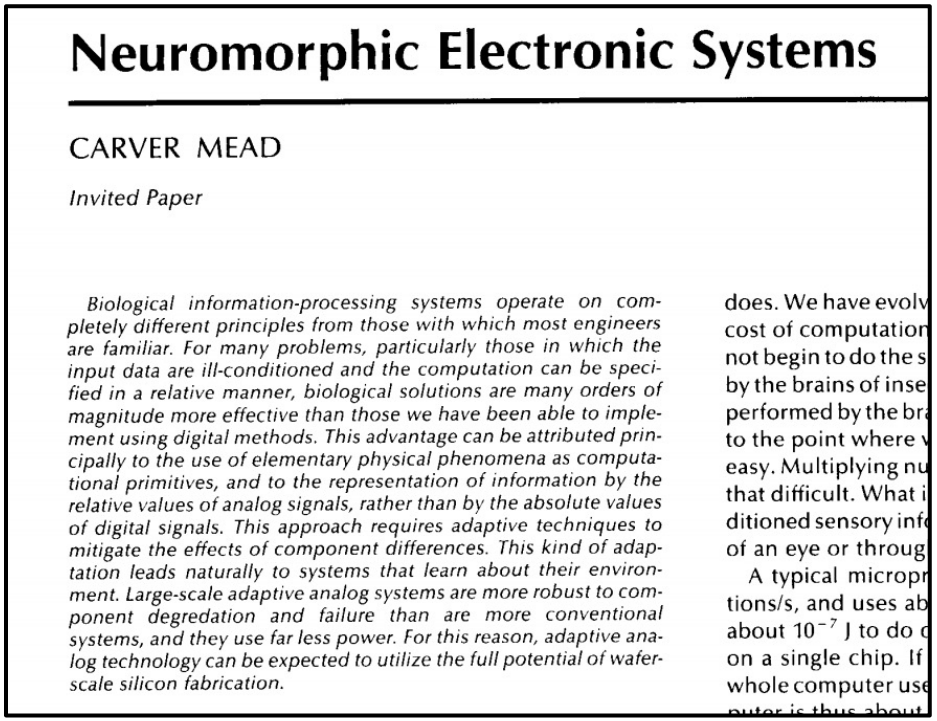
Director of Intelligent Circuits, Architectures, and Systems (iCAS) Lab
Assistant Professor of Computer Science and Engineering Department

# What is Neuromorphic Computing?

## Neuromorphic Computing:

- A concept developed (or at least popularized!) by Carver Mead in the late 1980s

- The main objective of neuromorphic computing is embodying the physical processes that underlie the computations of biological neural networks (NNs) within the physics of the very large-scale integration (VLSI) circuits



Neuromorphic Electronic Systems

CARVER MEAD

*Invited Paper*

*Biological information-processing systems operate on completely different principles from those with which most engineers are familiar. For many problems, particularly those in which the input data are ill-conditioned and the computation can be specified in a relative manner, biological solutions are many orders of magnitude more effective than those we have been able to implement using digital methods. This advantage can be attributed principally to the use of elementary physical phenomena as computational primitives, and to the representation of information by the relative values of analog signals, rather than by the absolute values of digital signals. This approach requires adaptive techniques to mitigate the effects of component differences. This kind of adaptation leads naturally to systems that learn about their environment. Large-scale adaptive analog systems are more robust to component degredation and failure than are more conventional systems, and they use far less power. For this reason, adaptive analog technology can be expected to utilize the full potential of wafer-scale silicon fabrication.*

does. We have evolv
cost of computation
not begin to do the s
by the brains of inse
performed by the br
to the point where
easy. Multiplying nu
that difficult. What i
ditioned sensory inf
of an eye or throug
A typical micropr
tions/s, and uses ab
about $10^{-7}$ J to do
on a single chip. If
whole computer use
puter is thus about

C. Mead, "Neuromorphic electronic systems," in *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629-1636, Oct. 1990.

# Why Neuromorphic Computing?

## What is the main motivation for neuromorphic computing?

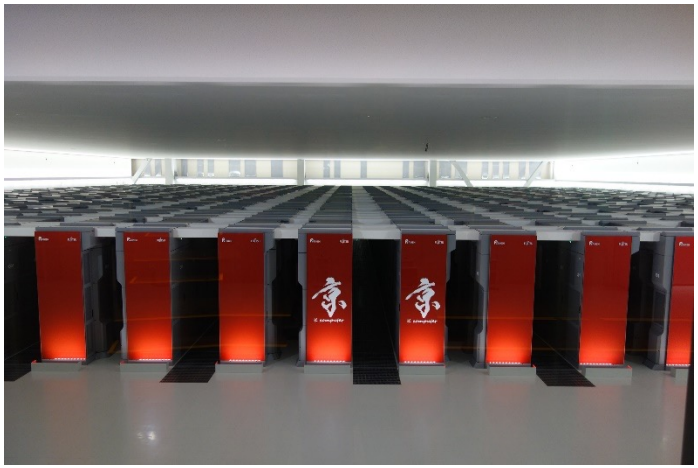
The extreme efficiency of brain for learning and pattern recognition tasks compared to existing computing platforms.

*How Efficient?!*

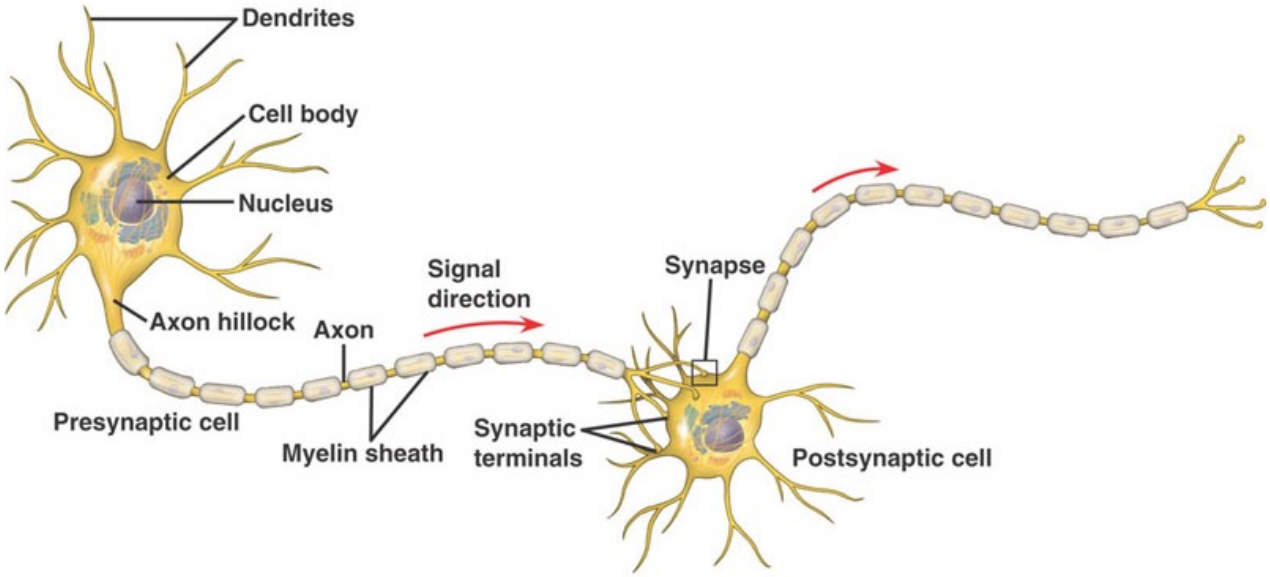
Picture Credit: Stanford University

- It takes 4.86 years for the Japan's K-computer to simulate 1-day of brain activity → 1700X slower than brain
- *K-computer* Power consumption = ~12 MW/h versus ~20W/h for brain

https://www.youtube.com/watch?v=c-stmgiXCZA



# Biological Neurons




**How brain works:** https://www.youtube.com/watch?v=o9K6GDBnByk

## Key Points?

- 100 Billion Neurons ➔ Highly Parallel
- Responsive in Nature ➔ Asynchronous Operation
- In-Memory Processes ➔ Avoid processor-memory bottleneck using beyond Von-Neumann architectures

# Spiking Neural Networks (SNN)

## Motivation and Vision:

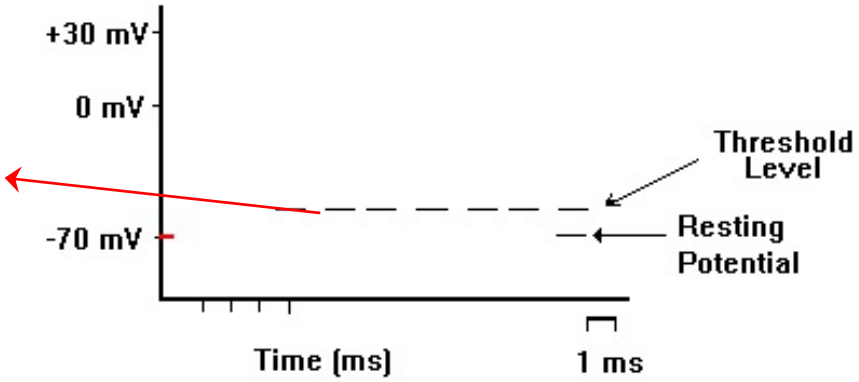
- Deep Neural Networks do not actually mimic the behavior of Brain's neurons
- Spiking Neural Networks aim to bridge the gap between neuroscience and machine learning as the third generation of ML algorithms

# Spiking Neural Networks (SNN)

## Fundamentals of SNN:

- SNNs operate based on ***spikes***, which are discrete events happening at different points in time, as opposed to continuous activation functions in conventional DNNs
- The occurrence of the spike depends on the membrane potential of the neuron that can be defined by a differential equation:

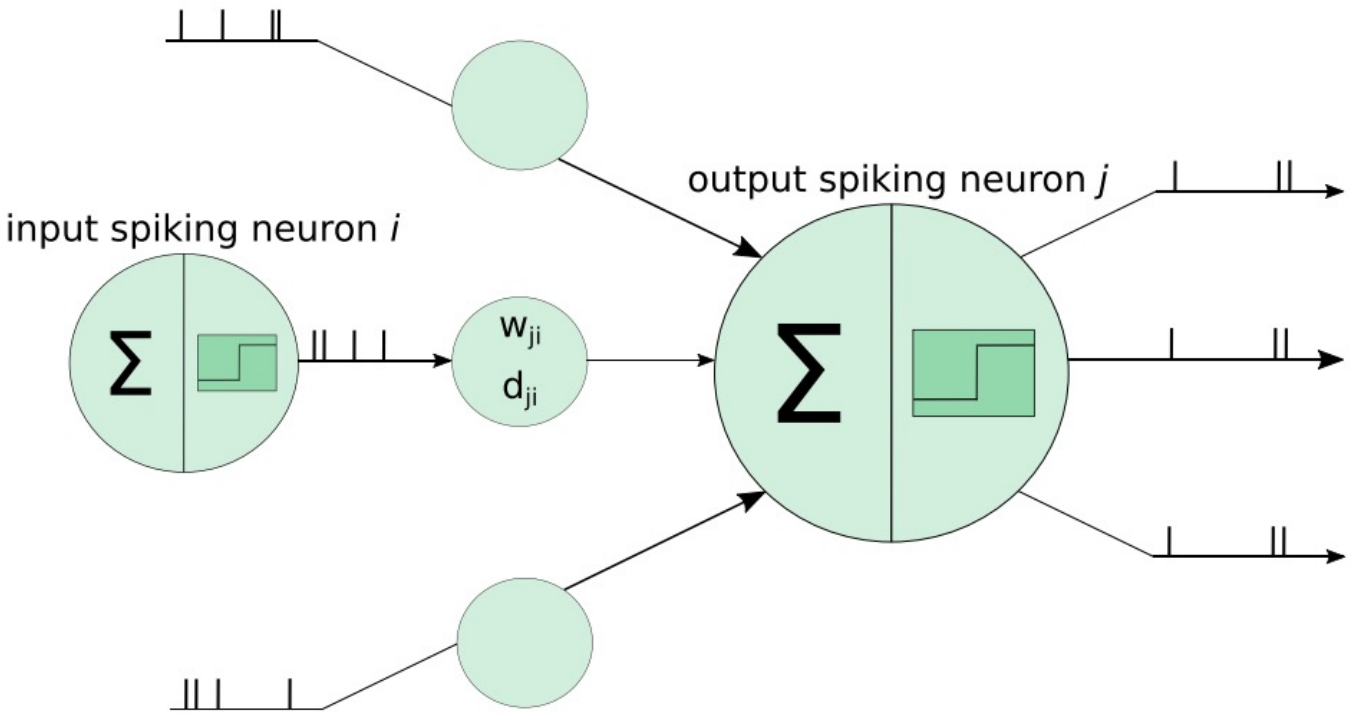
*Once a neuron reaches a certain potential, it spikes!*




**Source:** https://towardsdatascience.com/spiking-neural-networks-the-next-generation-of-machine-learning-84e167f4eb2b

# Spiking Neural Networks (SNN)

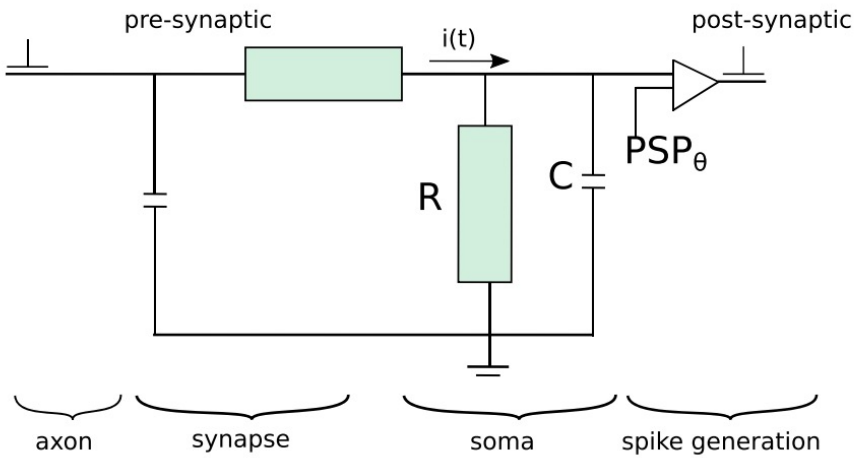
## SNN Architecture:



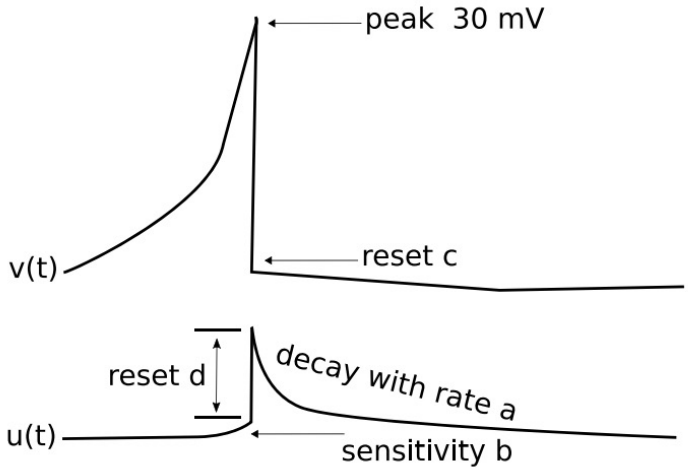

**Source:** Lobo, Jesus L., et al. "Spiking neural networks and online learning: An overview and perspectives." *Neural Networks* 121 (2020): 88-100.

# SPIKING NEURON MODELS

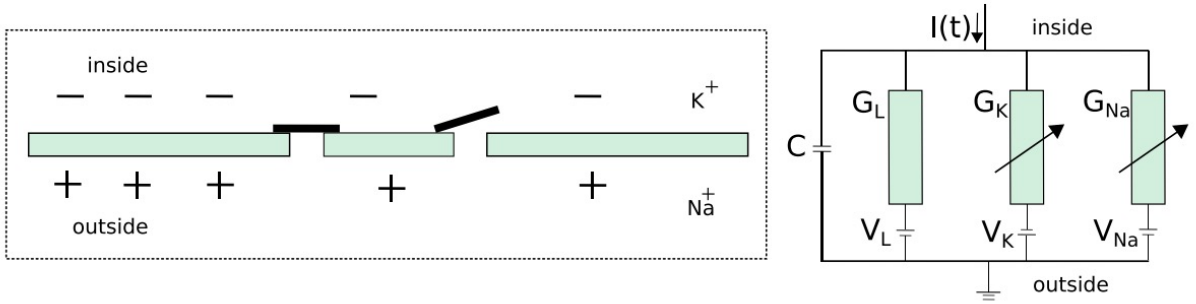
# Spiking Neuron Models




(a) LIF model




(b) Izhikevich model

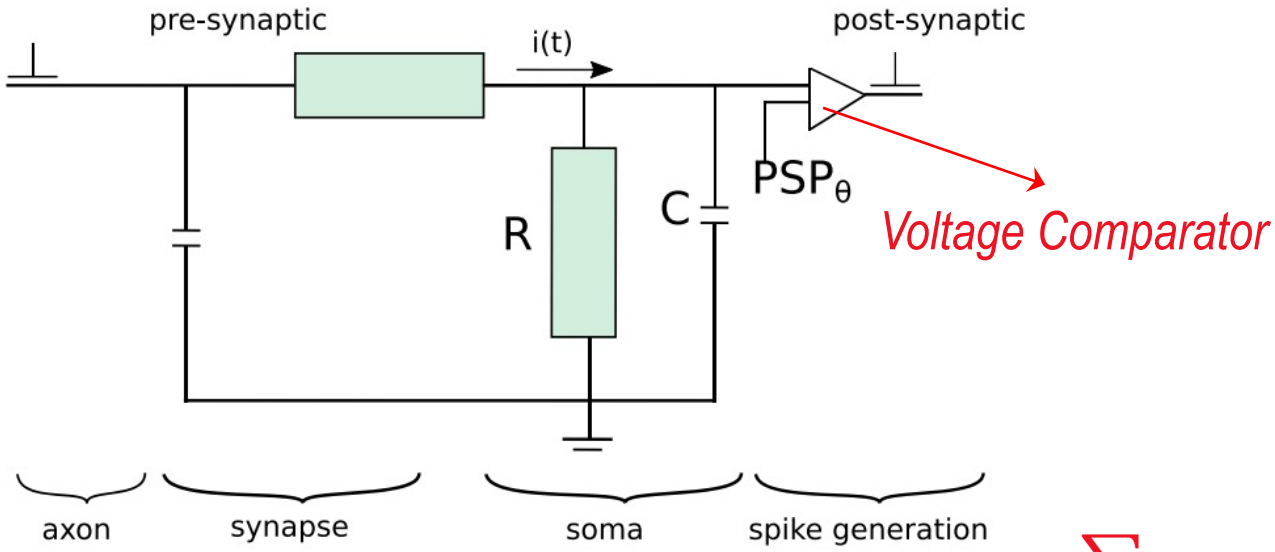



(c) Hodgkin-Huxley model

**Source:** Lobo, Jesus L., et al. "Spiking neural networks and online learning: An overview and perspectives." *Neural Networks* 121 (2020): 88-100.

# Spiking Neuron Models

## Leaky Integrate-and-Fire (LIF)



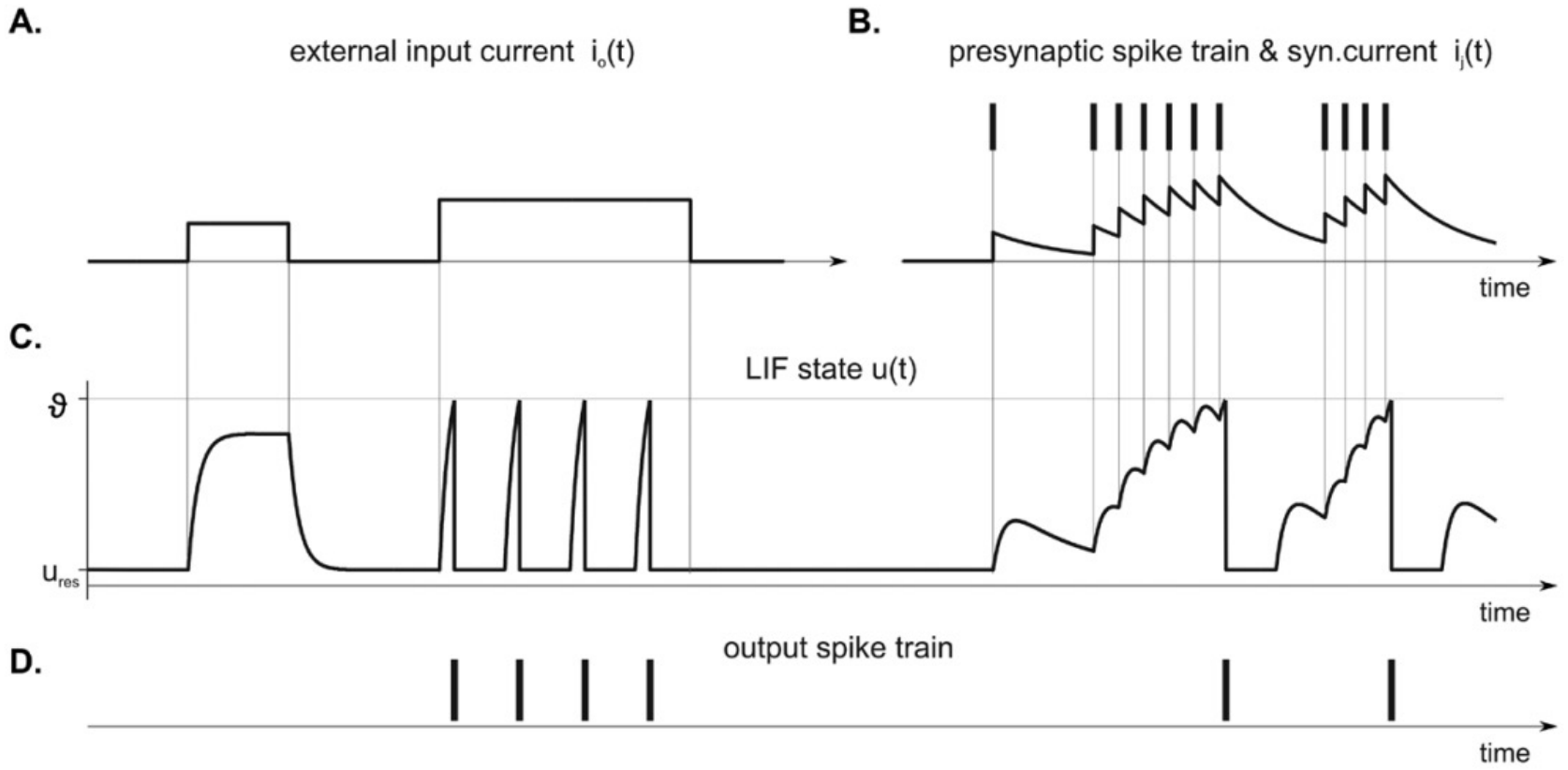

$$I(t) = I_R + I_{cap} \Rightarrow t_m \frac{du}{dt} = -u(t) + RI(t)$$

$\sum w_j i_j$

$t_m = RC$, i.e. time constant of membrane

**Source:** Lobo, Jesus L., et al. "Spiking neural networks and online learning: An overview and perspectives." *Neural Networks* 121 (2020): 88-100.

# LIF Neuron




**Source:** Ponulak, Filip, and Andrzej Kasinski. "Introduction to spiking neural networks: Information processing, learning and applications." *Acta neurobiologiae experimentalis* 71.4 (2011): 409-433.

# DATA REPRESENTATION IN SNNs

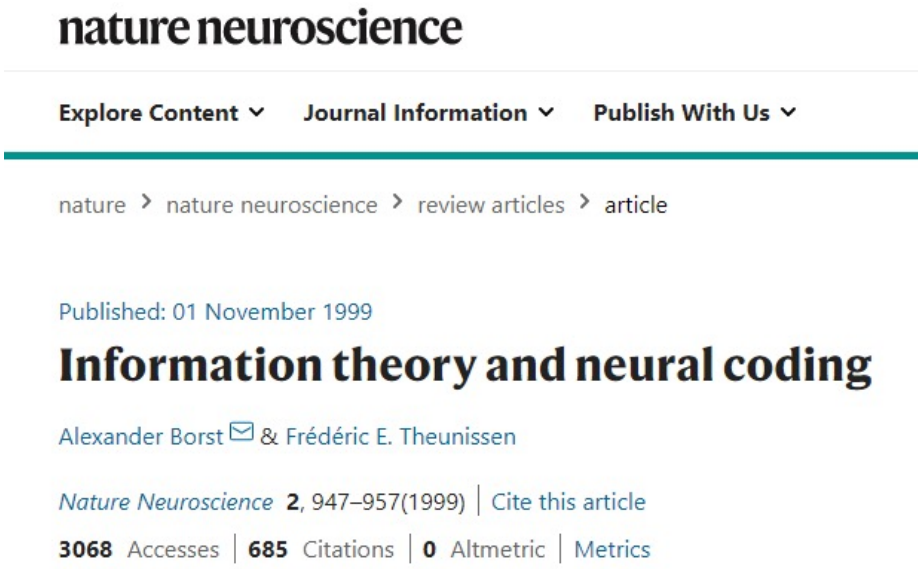
# Neural Coding

## Neural Coding:

- A field in neuroscience focusing on the relationship between a stimulus and neuron responses

## Fundamentals Questions:

- What is being encoded?
- How is it being encoded?
- With what precision?




**Source:** [Information theory and neural coding | Nature Neuroscience]

# Spike Trains

There are two main approaches for generating spike trains:

## 1) Event-driven neuromorphic sensors

Frontiers | Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades | Neuroscience (frontiersin.org)

## 2) Encoding

- **Rate-based Encoding**
- **Temporal Encoding**

# Rate-Based Coding

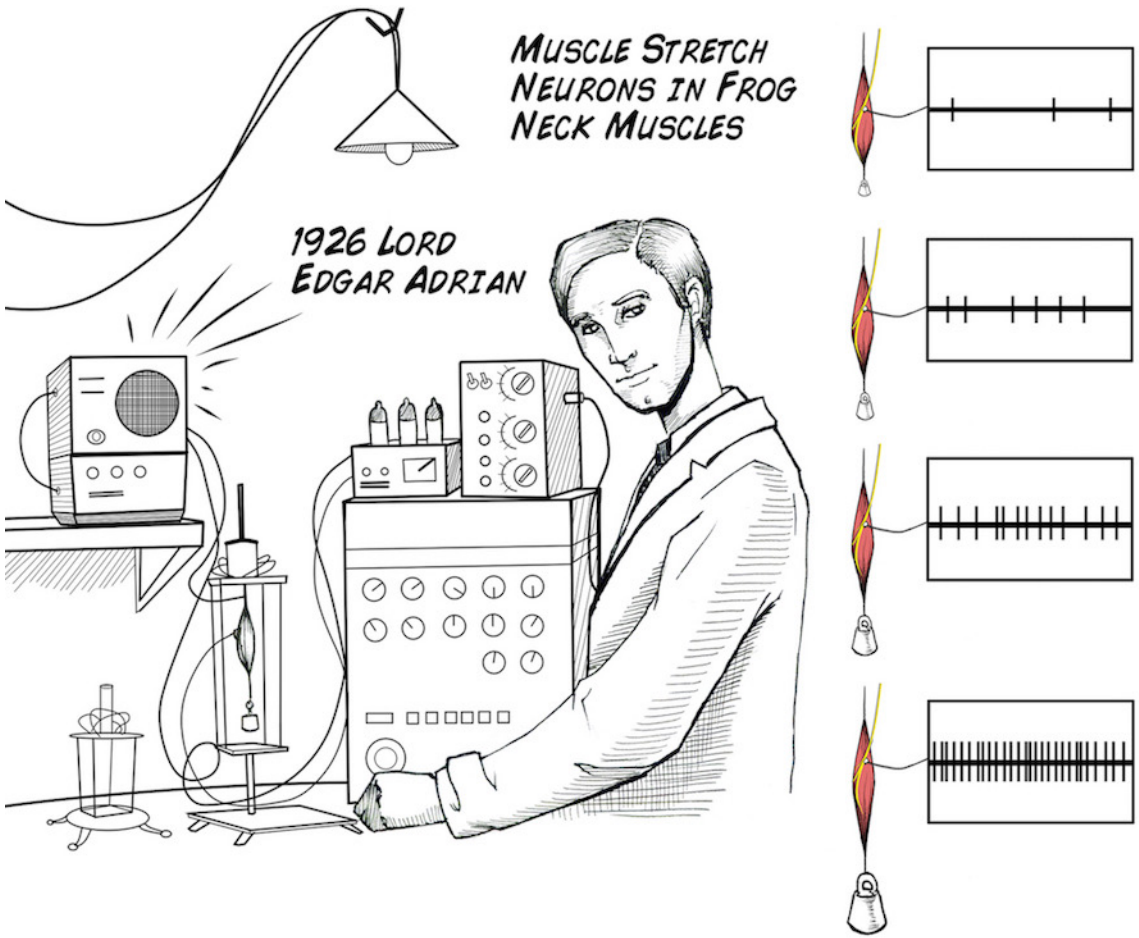



**Image Credit:** [Experiment: Rate Coding (backyardbrains.com)](). 

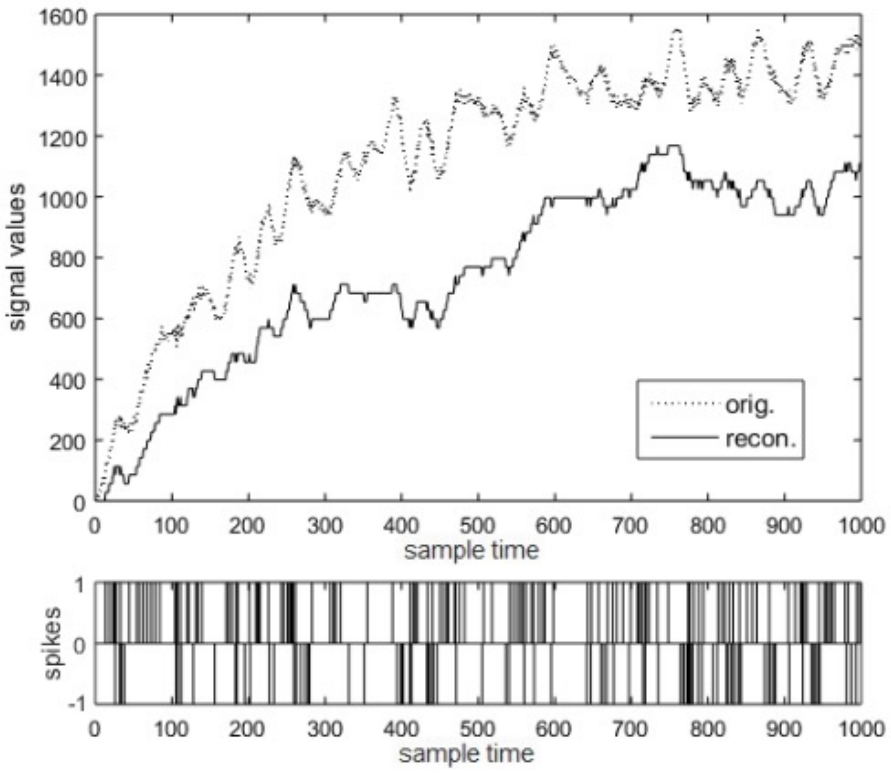
**Original paper:** [The impulses produced by sensory nerve-endings - Adrian - 1926 - The Journal of Physiology - Wiley Online Library]()

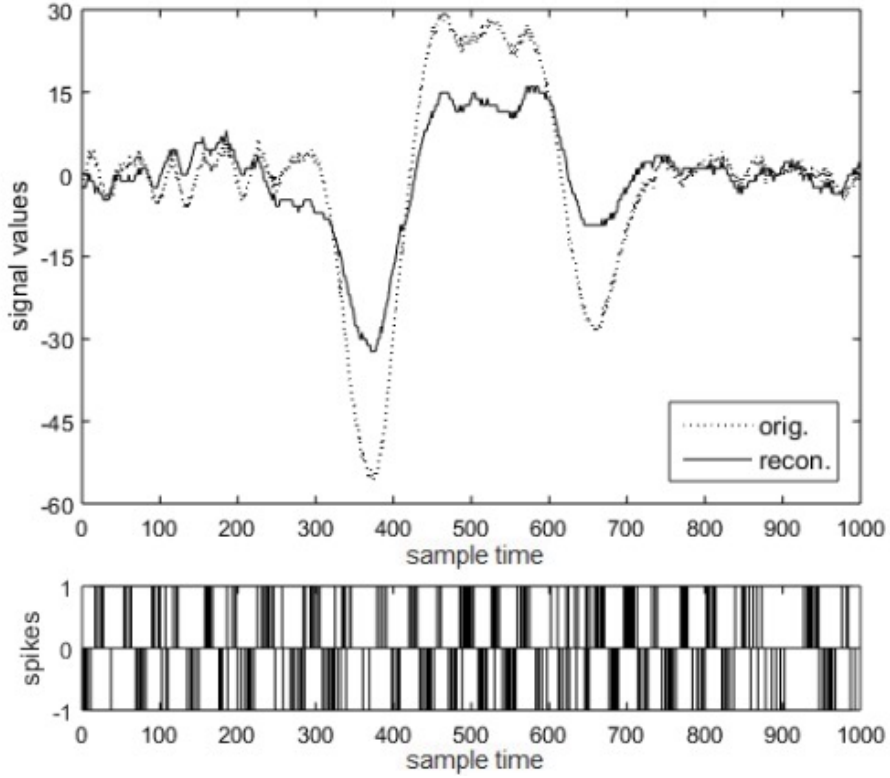
# Temporal Coding

## Temporal Encoding Strategies

- Rank Order Coding
- Threshold-Based Coding
- Step-Forward
- Moving-Window
- Ben's Spiker Algorithm
- Gaussian Receptive Field population Coding
- Local Optimum

# Threshold-based (TB) Coding




**Source:** Selection and Optimization of Temporal Spike Encoding Methods for Spiking Neural Networks - IEEE Journals & Magazine

# LEARNING IN SNNs

# Learning in SNNs

**Synaptic Plasticity:** The capability of synapses to change their strength (weight)

**Hebbian Learning:** Cells that fire together wire together

$$\Delta w_{ij} \propto v_i v_j$$

*change in the synaptic weight between the presynaptic neuron i and postsynaptic neuron j*

*Neuron i firing rate*

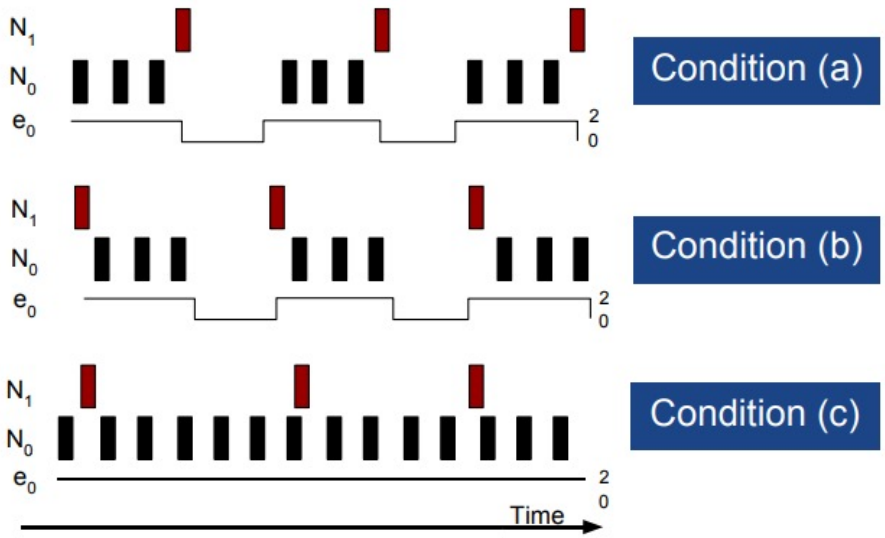
*Neuron j firing rate*

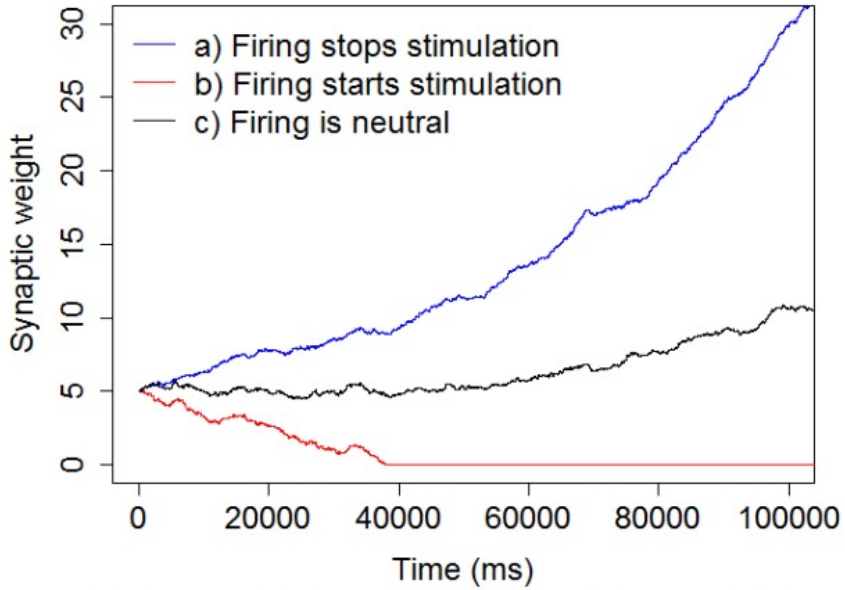
**Note:** The formula did not account for the synaptic depression!

# Learning in SNNs

## Unsupervised Learning:

- Spike-Timing-Dependent-Plasticity (STDP)
  - Presynaptic spikes preceding postsynaptic spikes ➔ Potentiation
  - Reverse order ➔ Depression



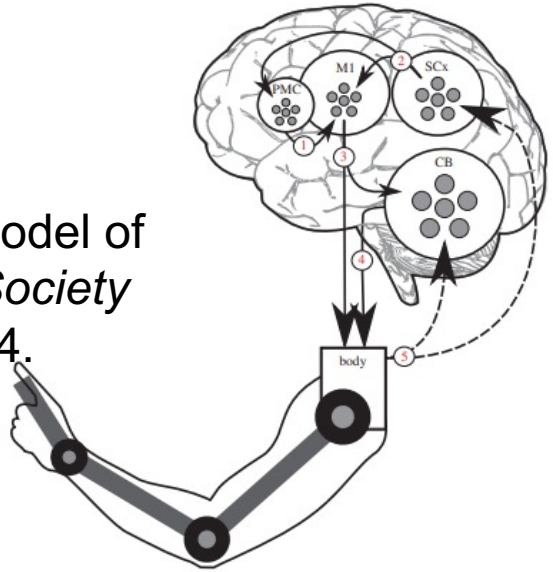

**Source:** Sinapayen, Lana, et al. "Learning by Stimulation Avoidance as a primary principle of spiking neural networks dynamics." *Artificial Life Conference Proceedings 13*. One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press, 2015.

# Learning in SNNs

STDP-like processes have played an important role in many applications of SNNs such as:

## ➢ Robotics Control

**Source:** DeWolf, Travis, et al. "A spiking neural model of adaptive arm control." *Proceedings of the Royal Society B: Biological Sciences* 283.1843 (2016): 20162134.




## ➢ Pattern Recognition

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

# Learning in SNNs
## Supervised Learning

### Supervised Hebbian Learning (SHL)

- Probably the most biologically-realistic and straightforward method to realize supervision in a SNN.
- The Hebbian process is supervised by a teaching signal (usually in form of a current) that reinforces the postsynaptic neuron to fire at target times and remain silent at other times.

- **Challenges:**

1) No mechanism to weaken the synaptic weights that caused the neuron to fire at undesired times since the teacher currents suppress all undesired firing during training.
2) Synapses continue to change even if the neurons fire exactly at the desired time

### Examples of other Supervised Learning approaches for SNN:

1) Remote Supervised Method (ReSuMe)

Ponulak, F. (2005). *Resume-new supervised learning method for spiking neural networks, (vol. 42)*. Institute of Control and Information Engineering, Poznan University of Technology.

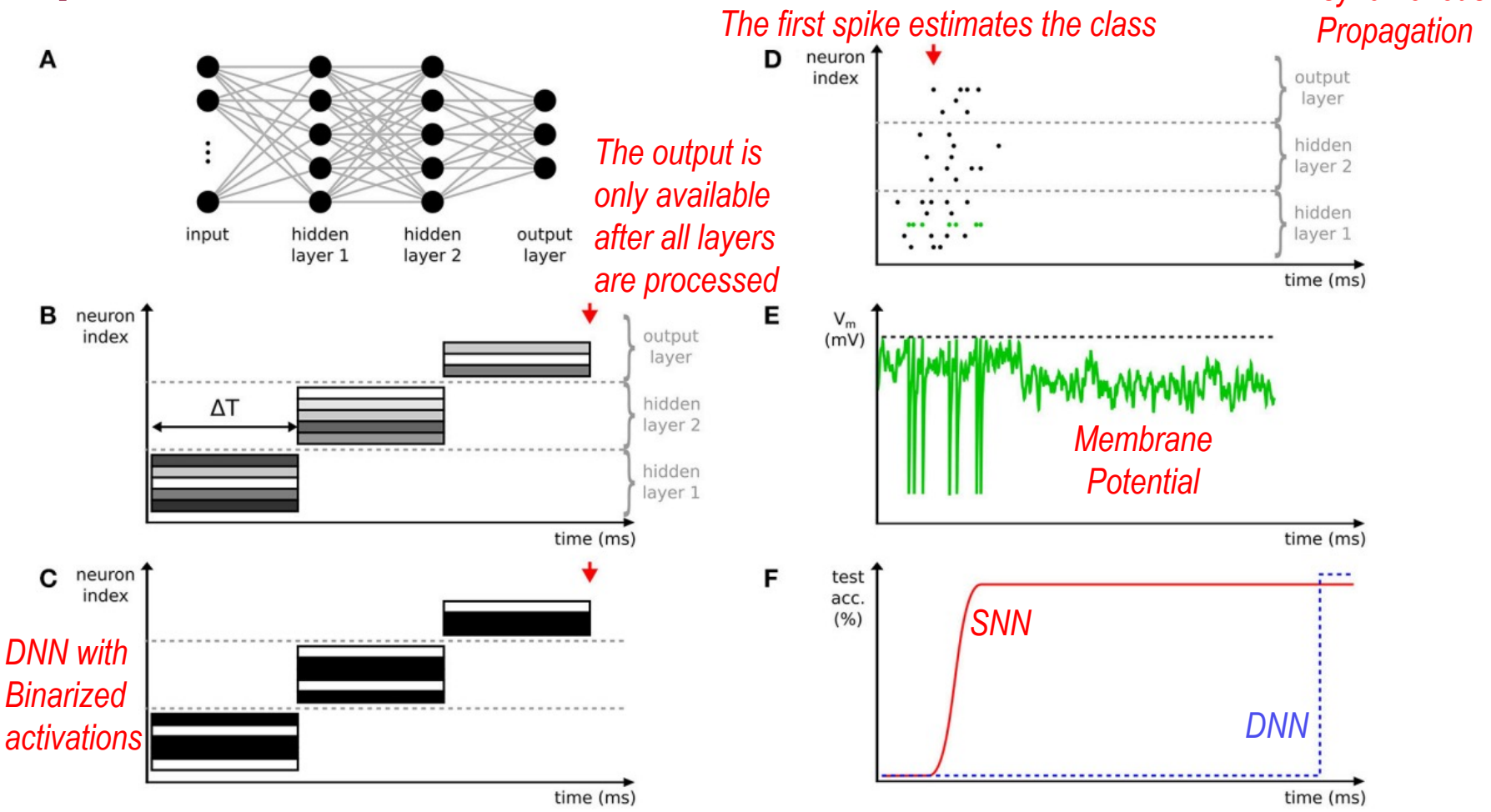
2) SpikeProp

Bohte, Sander M., Joost N. Kok, and Han La Poutre. "Error-backpropagation in temporally encoded networks of spiking neurons." *Neurocomputing* 48.1-4 (2002): 17-37.

# Deep SNN

# Deep Spiking Neural Networks (SNNs)

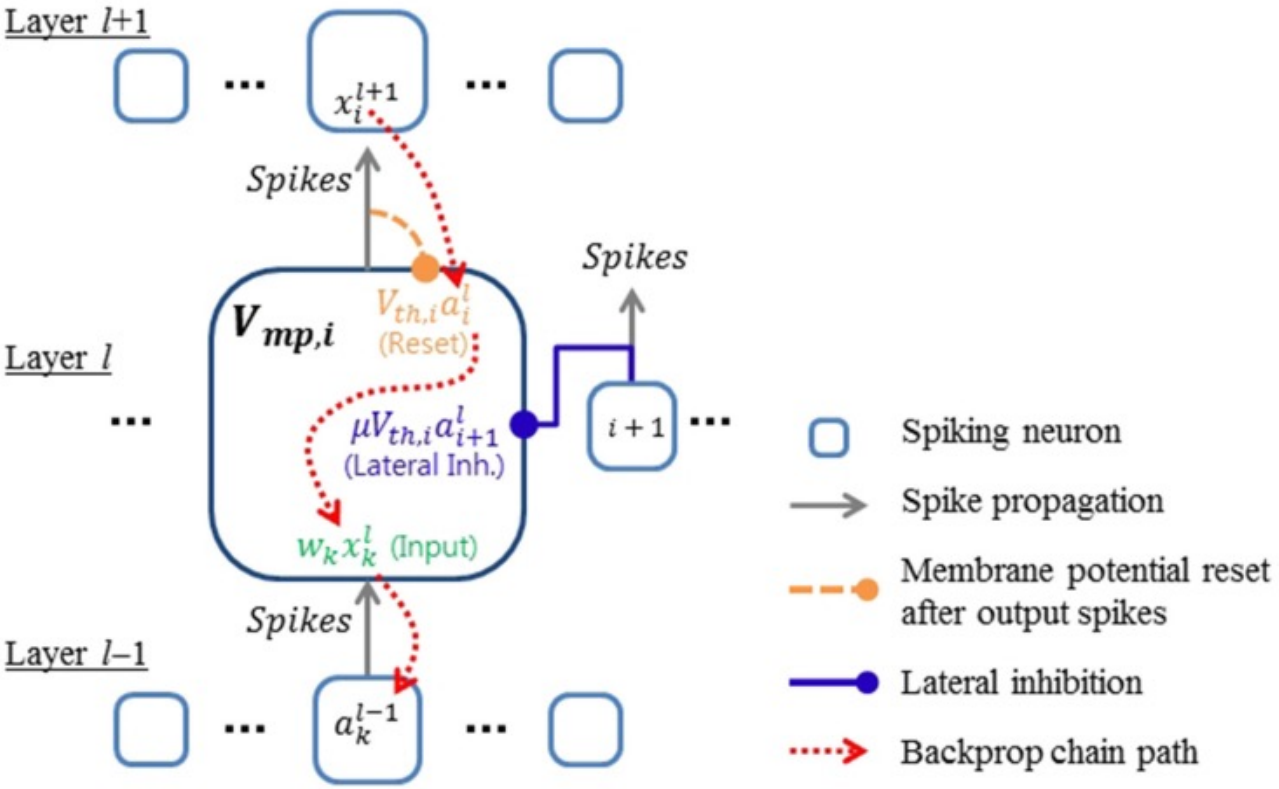
## Comparison of SNNs with conventional DNNs




**Source:** Pfeiffer, M., & Pfeil, T. (2018). Deep learning with spiking neurons: opportunities and challenges. *Frontiers in neuroscience*, *12*, 774.

# Deep Spiking Neural Networks (SNNs)
## Training: Supervised Learning with Spikes

➔ *Lee et al. (2016)* performed the gradient descent on real valued *membrane potentials*.




**Source:** Lee, Jun Haeng, Tobi Delbruck, and Michael Pfeiffer. "Training deep spiking neural networks using backpropagation." *Frontiers in neuroscience* 10 (2016): 508.

# Deep Spiking Neural Networks (SNNs)
Training

## Conversion of Deep Neural Networks

**Objective:** Map a trained DNN into a Deep SNN to address the gradient descent challenge in spiking networks. The mapping includes:

- Weights and structure of the DNN
- Input-output encoding

**Advantages:**

- State-of-the-art DNNs can be readily mapped into SNNs
- The DNNs can be trained without considering that it should be later converted to SNN
- Once the DNN parameters are obtained, the conversion into SNN only involves parsing and straightforward transformation

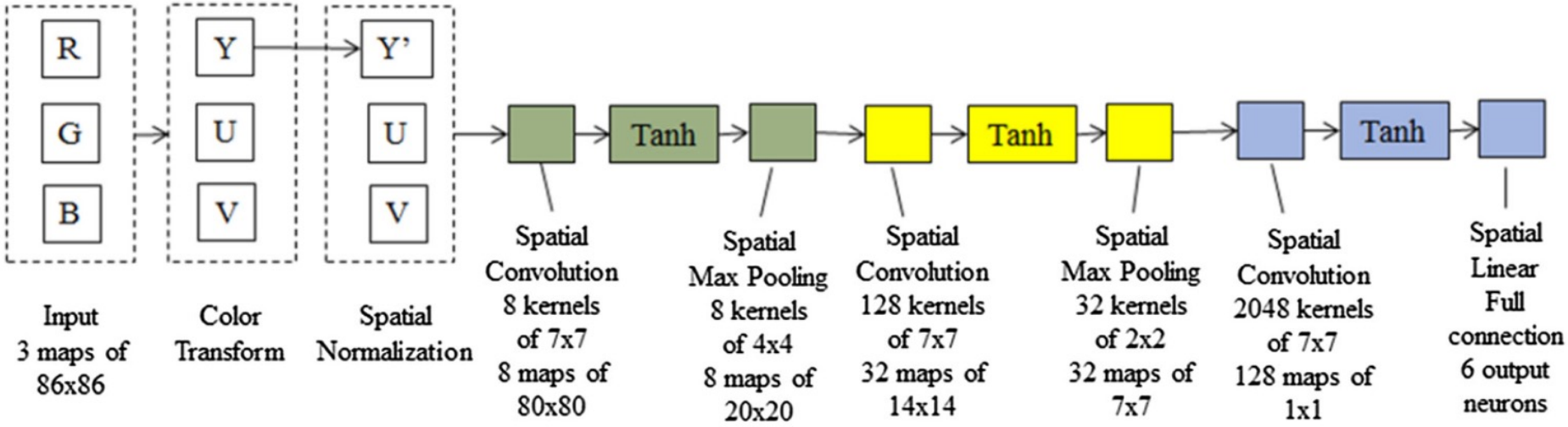
**Spiking Neural Network Conversion Toolbox:**

https://snntoolbox.readthedocs.io/en/latest/guide/intro.html

# Spiking CNN Architecture

## Flow of converting a CNN into SNN:


## Step I: CNN

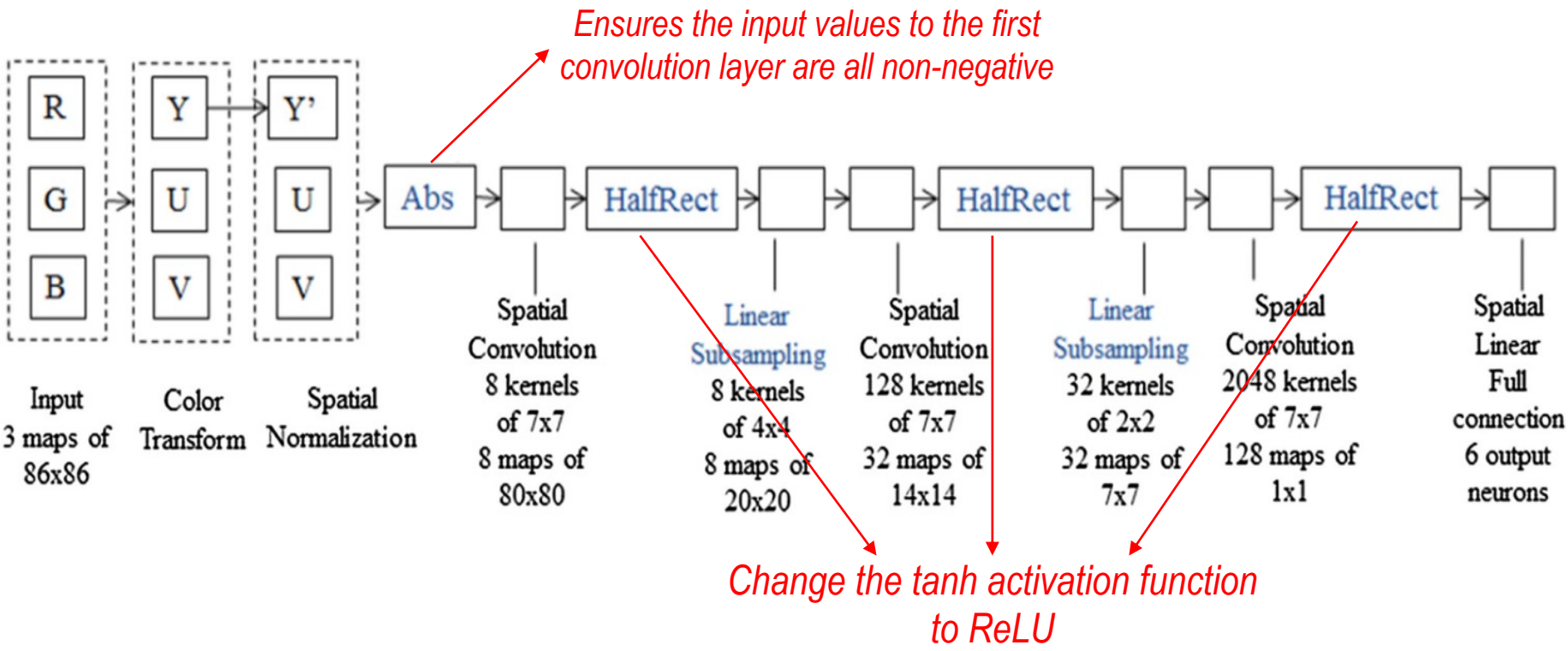



**Source:** Cao, Yongqiang, Yang Chen, and Deepak Khosla. "Spiking deep convolutional neural networks for energy-efficient object recognition." *International Journal of Computer Vision* 113.1 (2015): 54-66.

# Spiking CNN Architecture

## Step II: Tailored CNN

A. Make sure that output values in all layers are positive:

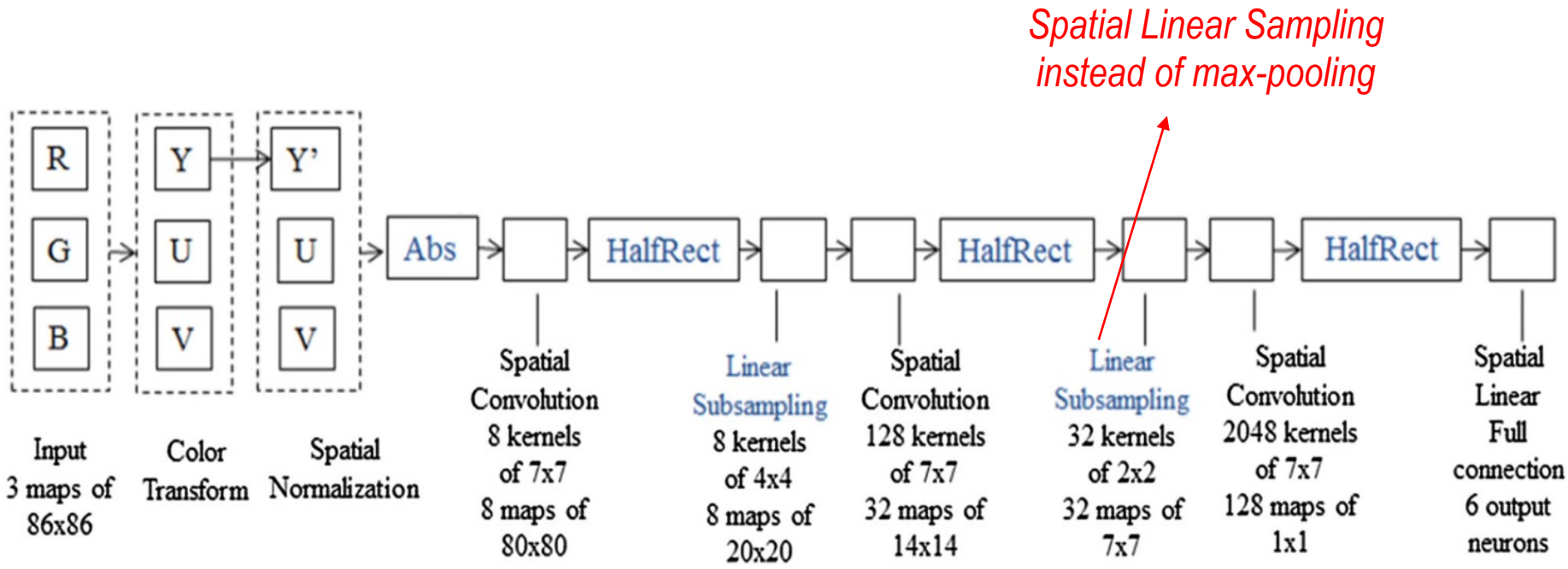



**Source:** Cao, Yongqiang, Yang Chen, and Deepak Khosla. "Spiking deep convolutional neural networks for energy-efficient object recognition." *International Journal of Computer Vision* 113.1 (2015): 54-66.

# Spiking CNN Architecture

## Step II: Tailored CNN

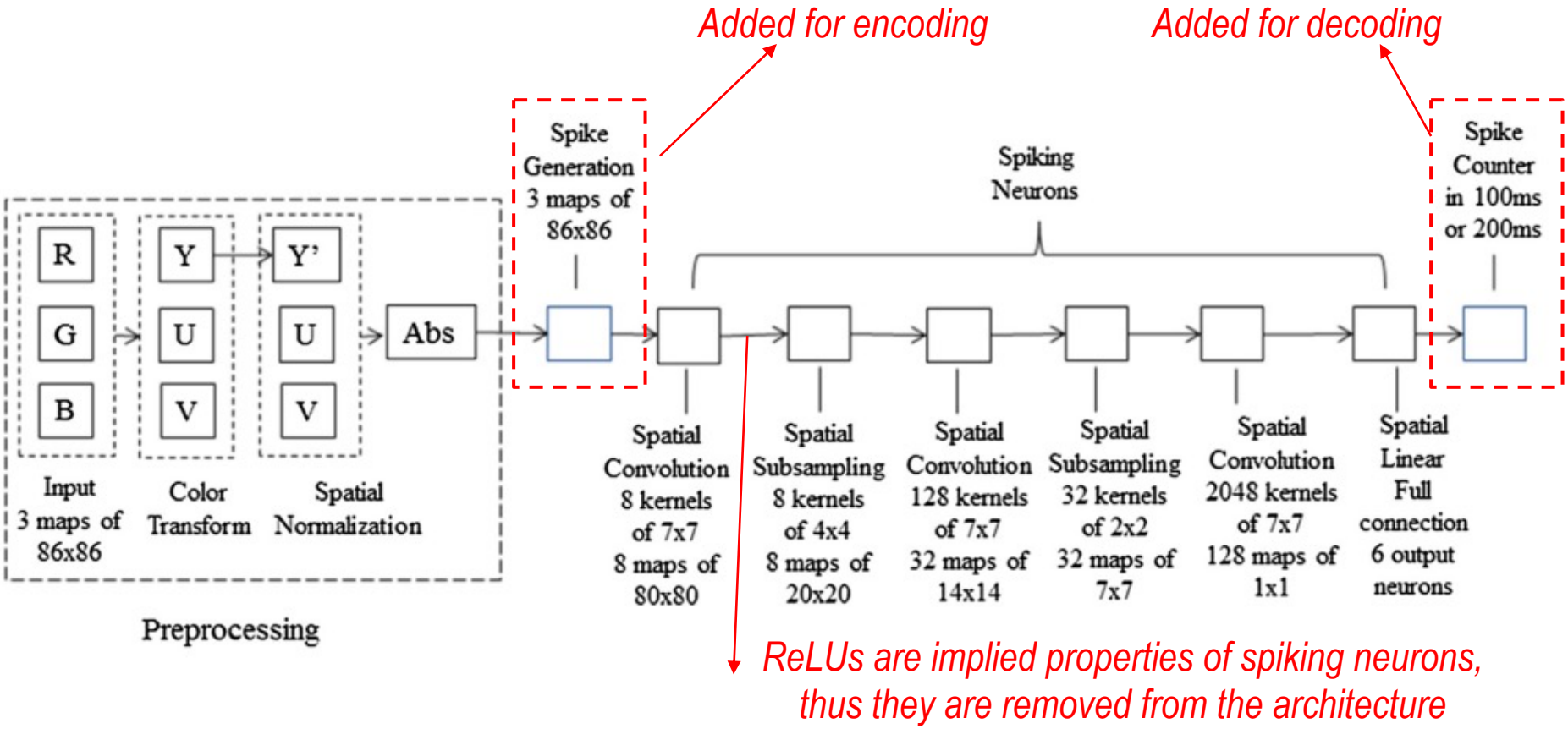
B. Use spatial linear subsampling instead of nonlinear max pooling. Spatial linear subsampling simply adds all pixels over a small image neighborhood.




**Source:** Cao, Yongqiang, Yang Chen, and Deepak Khosla. "Spiking deep convolutional neural networks for energy-efficient object recognition." *International Journal of Computer Vision* 113.1 (2015): 54-66.

# Spiking CNN Architecture

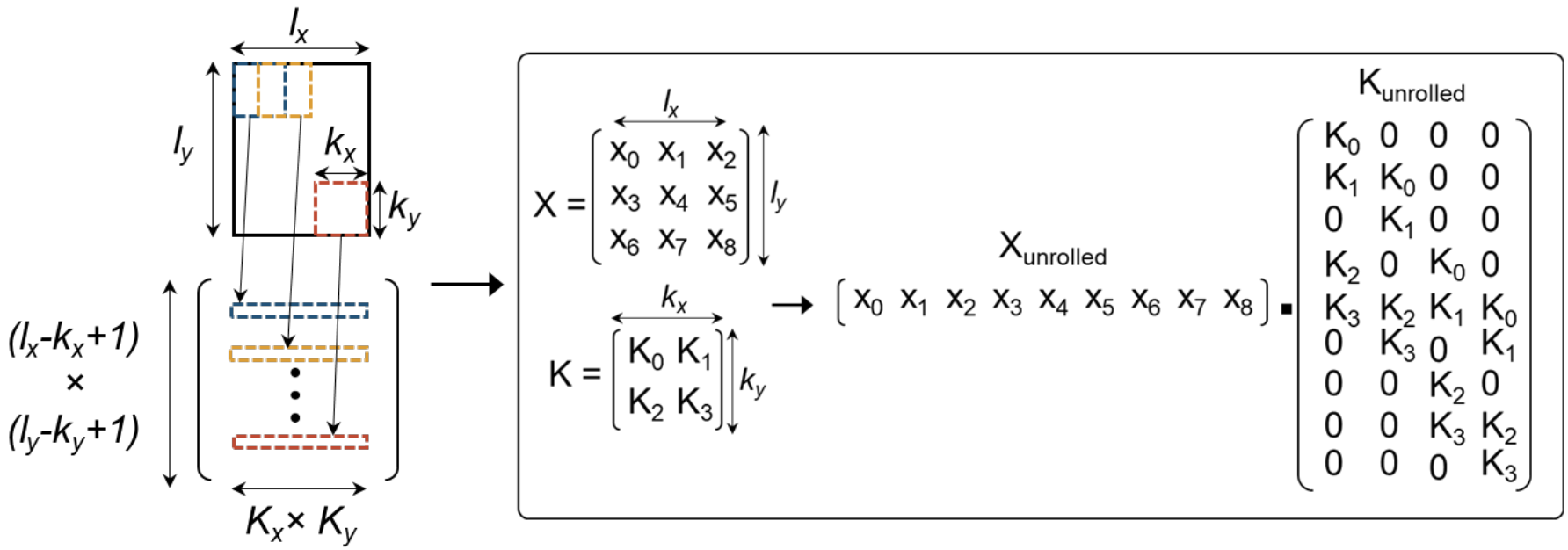
## Step III: Spiking CNN




**Source:** Cao, Yongqiang, Yang Chen, and Deepak Khosla. "Spiking deep convolutional neural networks for energy-efficient object recognition." *International Journal of Computer Vision* 113.1 (2015): 54-66.

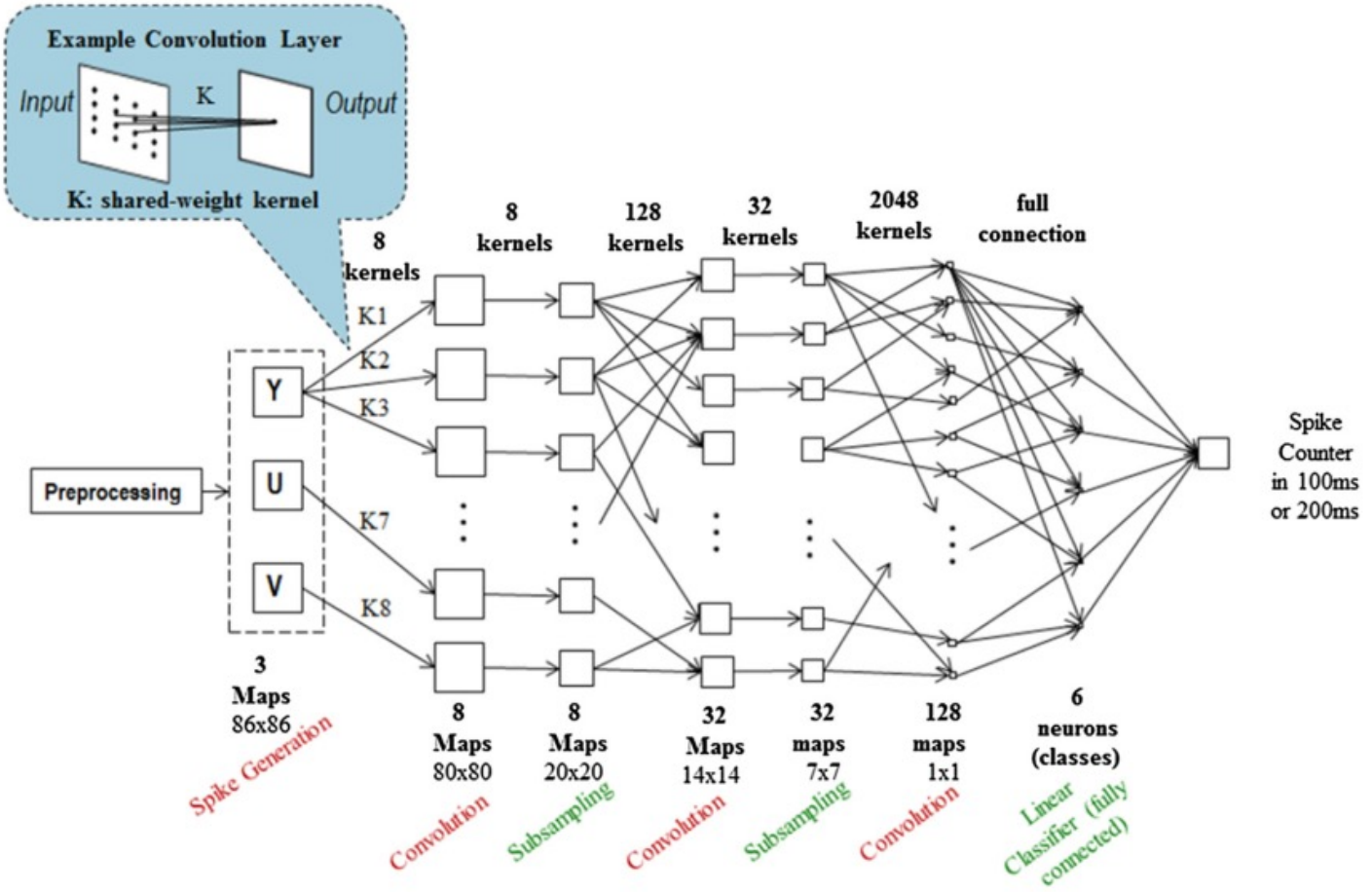
# Spiking CNN Architecture

## Step III: Spiking CNN (Convolution Unrolling)



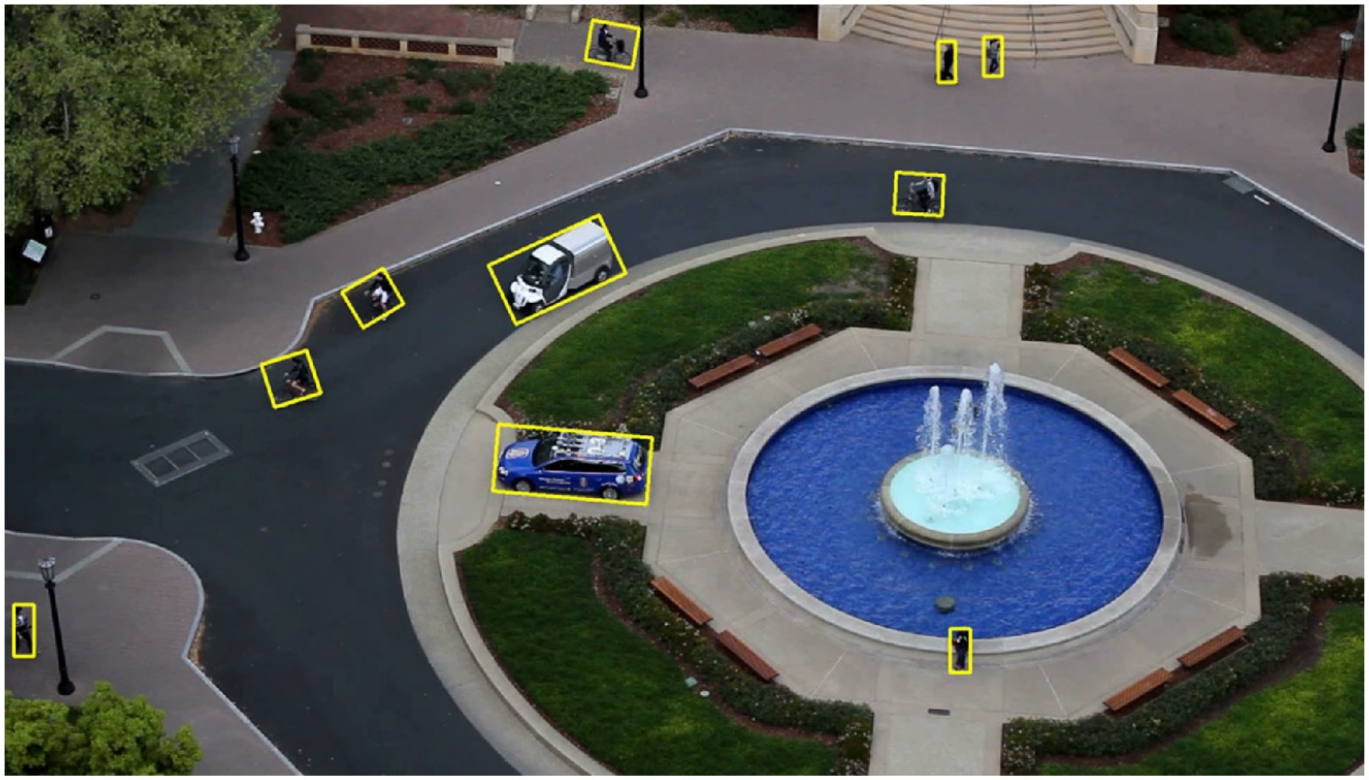
# Spiking CNN Architecture

**Convolution Unrolling:** Converts the CNN architecture to DNN that can be readily implemented by SNN architecture.




**Source:** Cao, Yongqiang, Yang Chen, and Deepak Khosla. "Spiking deep convolutional neural networks for energy-efficient object recognition." *International Journal of Computer Vision* 113.1 (2015): 54-66.

# Spiking CNN Sample Application



DARPA Neovision2 Tower data set

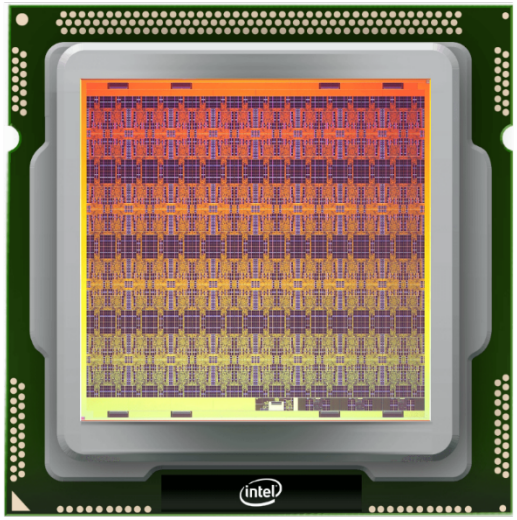
**Source:** Cao, Yongqiang, Yang Chen, and Deepak Khosla. "Spiking deep convolutional neural networks for energy-efficient object recognition." *International Journal of Computer Vision* 113.1 (2015): 54-66.

# SNN HARDWARE

# Commercial Neuromorphic Chips?

## Intel's Loihi Chip:

- Features a manycore mesh including 128 neuromorphic cores designed based on a specialized architecture optimized for SNN algorithms and fabricated on 14nm process technology
- Each Neuromorphic Core implements 1,024 spiking neural units that are grouped into sets of trees constituting neurons.
- 130,000 neurons optimized for SNNs
- The mesh protocol supports scaling to 4096 on-chip cores, as well as 16,384 chips through hierarchical addressing




**Loihi: A Neuromorphic Manycore Processor with On-Chip Learning:**

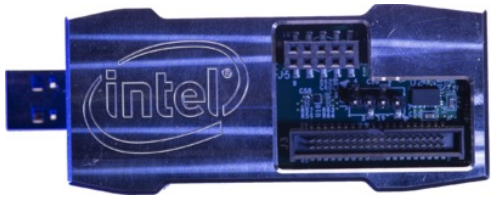
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8259423

**Intel's 'Loihi' Neuromorphic Chip in the Lab:**

https://www.youtube.com/watch?v=cDKnt9ldXv0

# Loihi-based Neuromorphic Systems

- **Kapoho Bay:** A USB stick that includes 1 or 2 Loihi chips




- **Wolf Mountain:** 4 Loihi chips, 524K neurons

- **Nahuku:** 8-32 Loihi chips, 1M-4M neurons

# SNN APPLICATION

# Online Learning

## Online Learning vs. Batch Learning:

- **Batch learning:**
  - An entire group of samples is accessible
  - Learning algorithm may scan the entire batch before updating the model
- **Online Learning:**
  - Only a single sample is provided at every time instant
  - Learning algorithm incrementally updates the model upon the arrival of every new sample

**Source:** Lobo, Jesus L., et al. "Spiking neural networks and online learning: An overview and perspectives." *Neural Networks* 121 (2020): 88-100.

# Online Learning Applications

## Internet of Things (IoT):

- **Conventional Definition:** Sensors and Actuators (*edge*) connected to computing systems (*cloud*) through networks.
- Requires online learning due to the generation of huge amount of data continuously in real time.

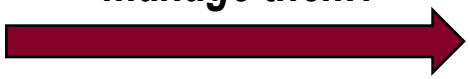
https://it.toolbox.com/blogs/alenibric/is-iot-a-danger-to-the-future-071719

# SNN for Online Learning Applications

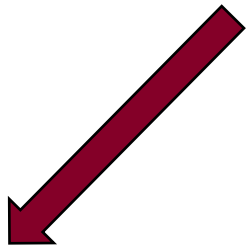
## Streams of Data in IoT Applications

- *Sensor networks*
- *Traffic management*
- *Click-streams in web surfing*
- *Manufacturing processes*
- *Call detail record*
- *Emails*
- *Blogging*
- *Twitter posts*

***How can we process and manage them?***



**Online Learning**



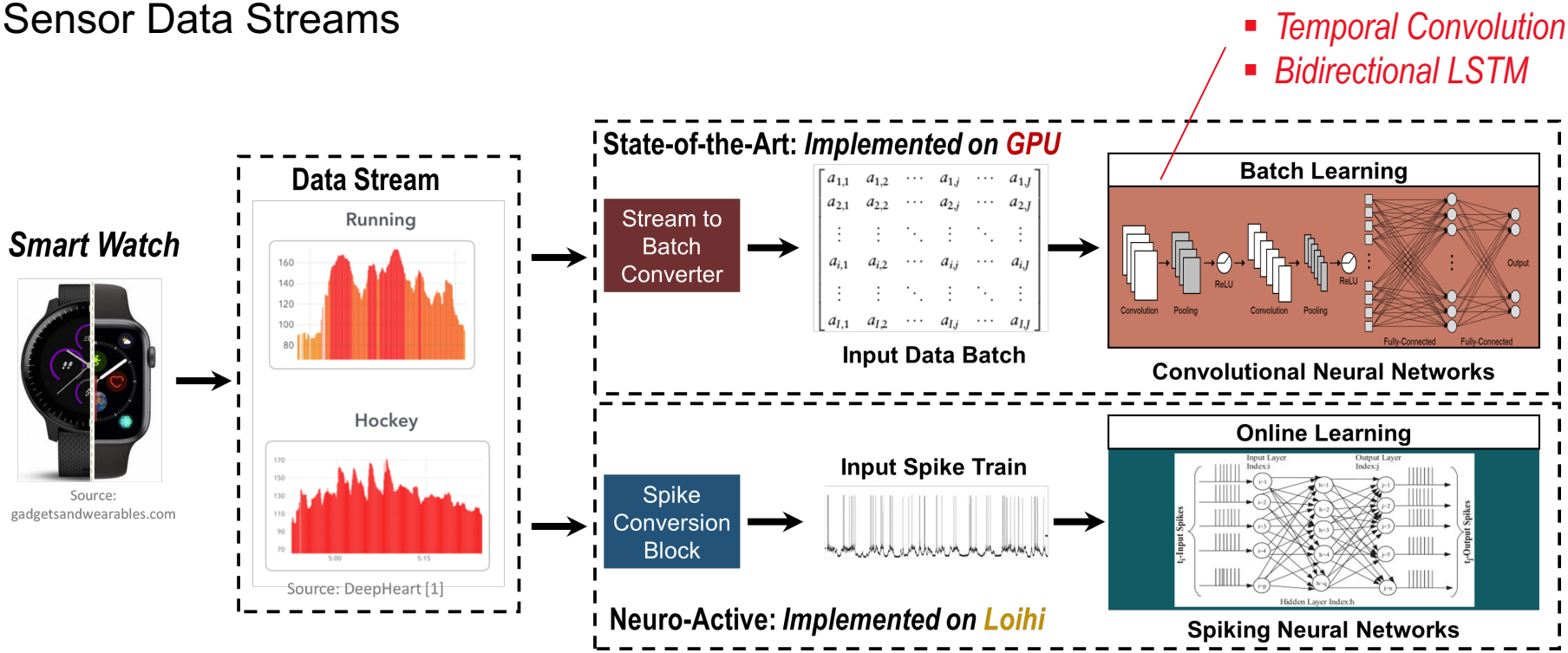
## Spiking Neural Networks

- *Ability to capture temporal associations*
- *Local learning rules, e.g. STDP*
- *Real-time learning with reduced computational complexity*
- *Adaptation to the Drift: Most of the conventional classification models require to be retrained if the environment is changed, while some SNNs can overcome this drawback.*

**Source:** Lobo, Jesus L., et al. "Spiking neural networks and online learning: An overview and perspectives." *Neural Networks* 121 (2020): 88-100.

# Neuro-Edge: Neuromorphic-Enhanced Edge Computing

**1. Neuro-Active:** A Neuromorphic Activity Recognition System Using Smartwatch Sensor Data Streams




**2. Neuro-Heart:** Stream learning for detecting multiple medical conditions such as diabetes, high blood pressure, cardiac arrhythmia

**3. Neuro-Speech:** Real-time recognition of auditory objects using neuromorphic architectures