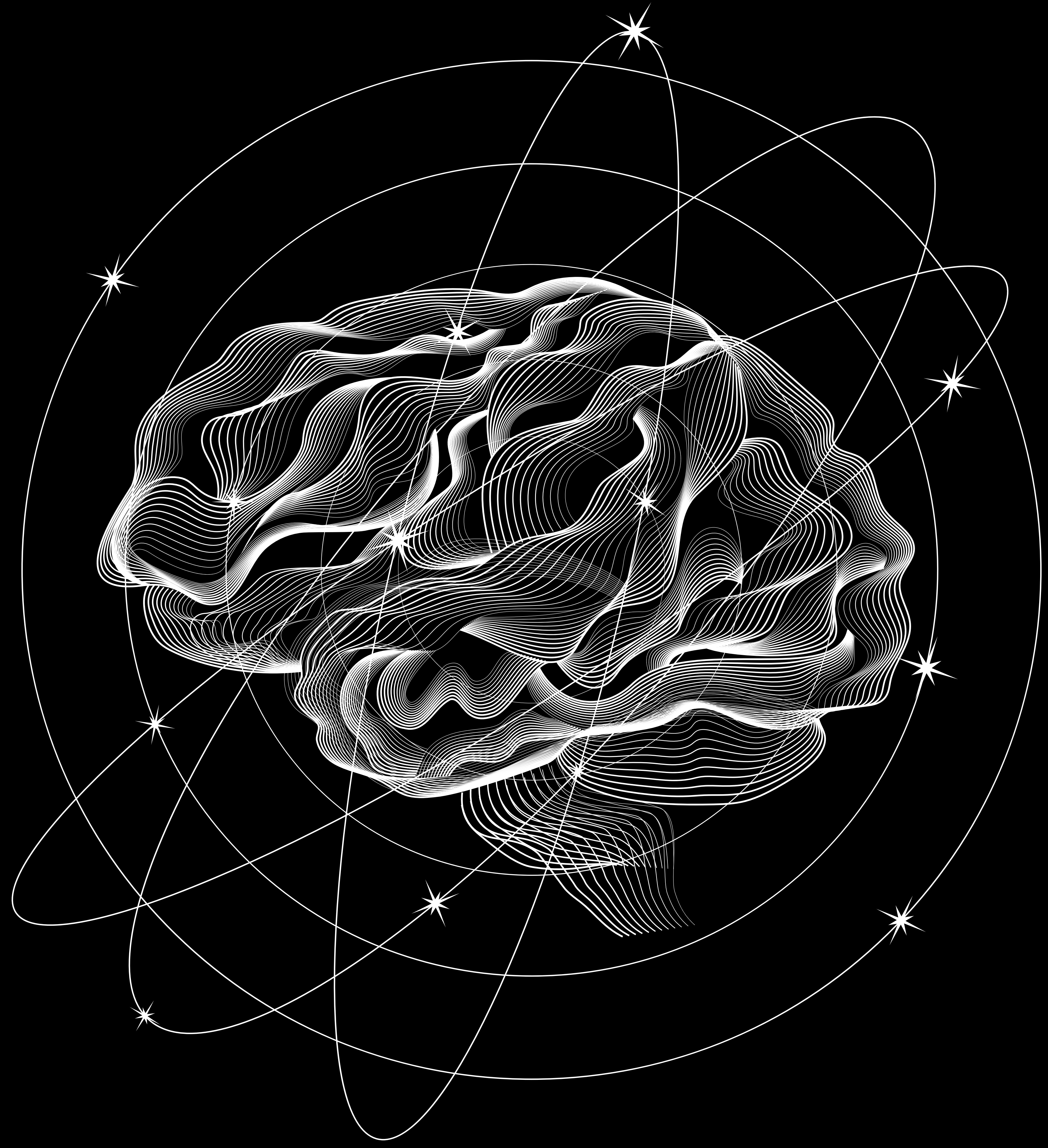


ANN

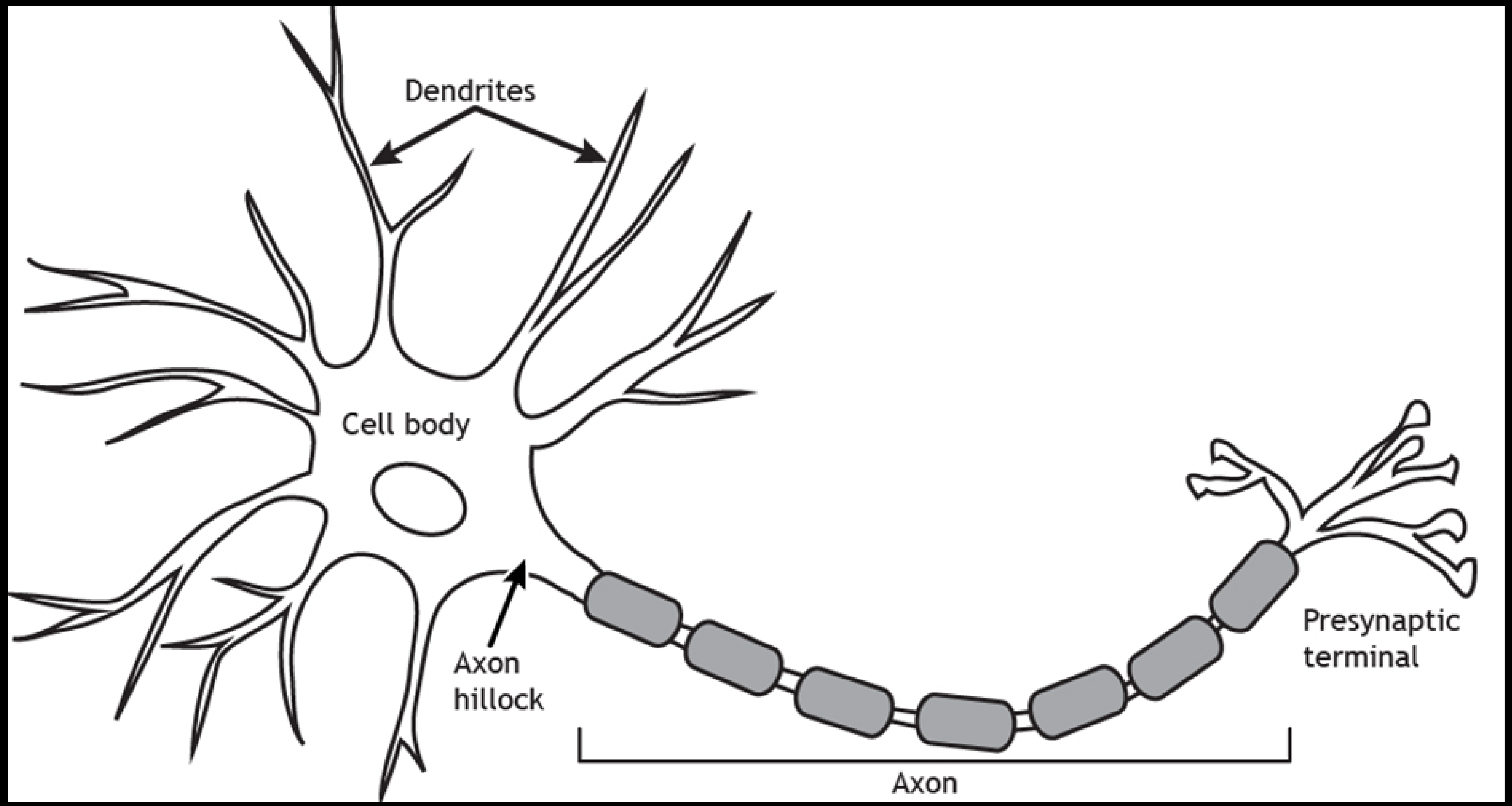
Artificial Neural Network

What is Artificial Neural Network?

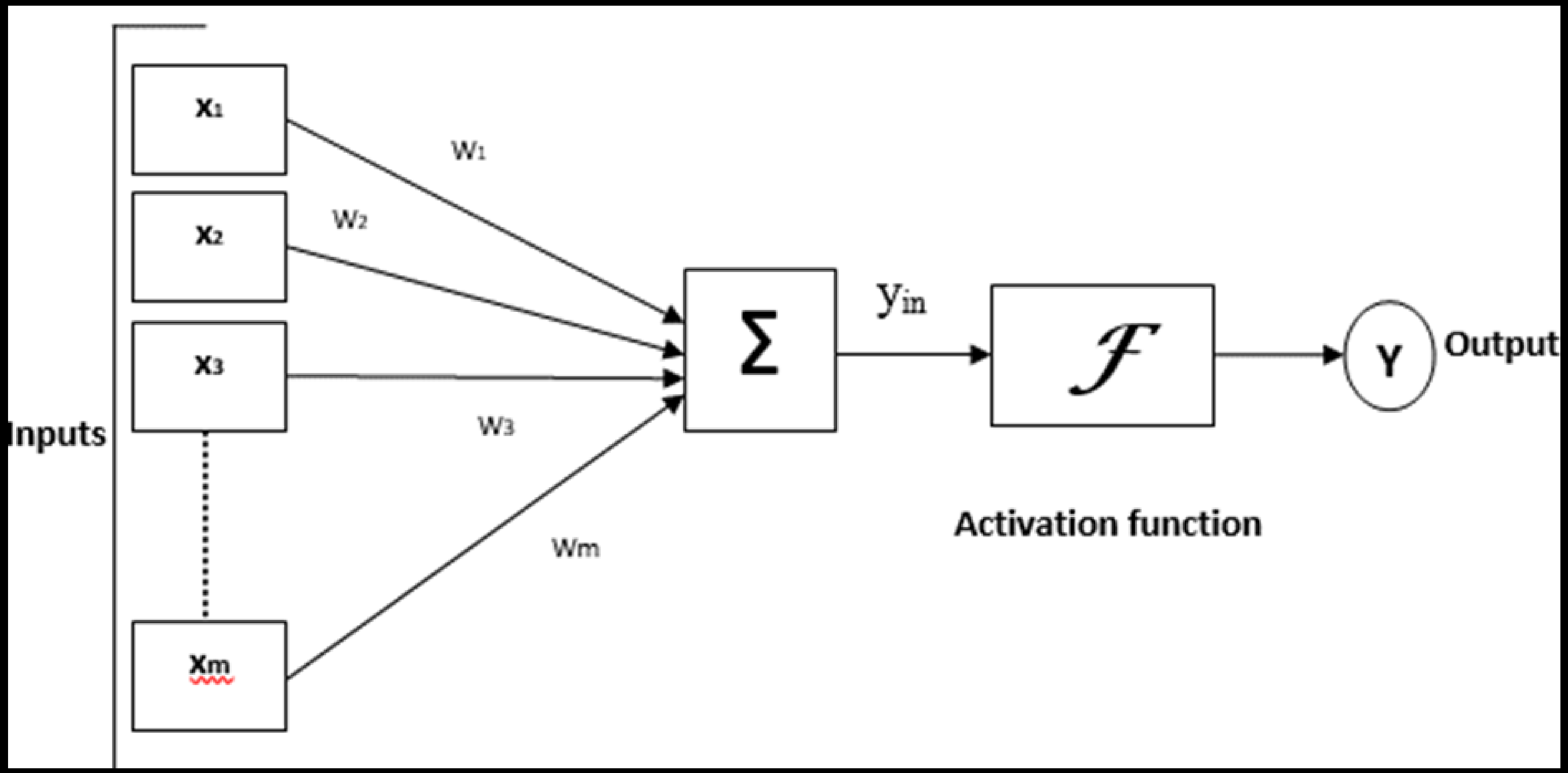
Human brain is a highly complex structure viewed as a massive, highly interconnected network of simple processing elements called neurons

Artificial neural networks (ANNs) or simply we refer it as neural network (NNs), which are simplified models (i.e. imitations) of the biological nervous system, and obviously, therefore, have been motivated by the kind of computing performed by the human brain.

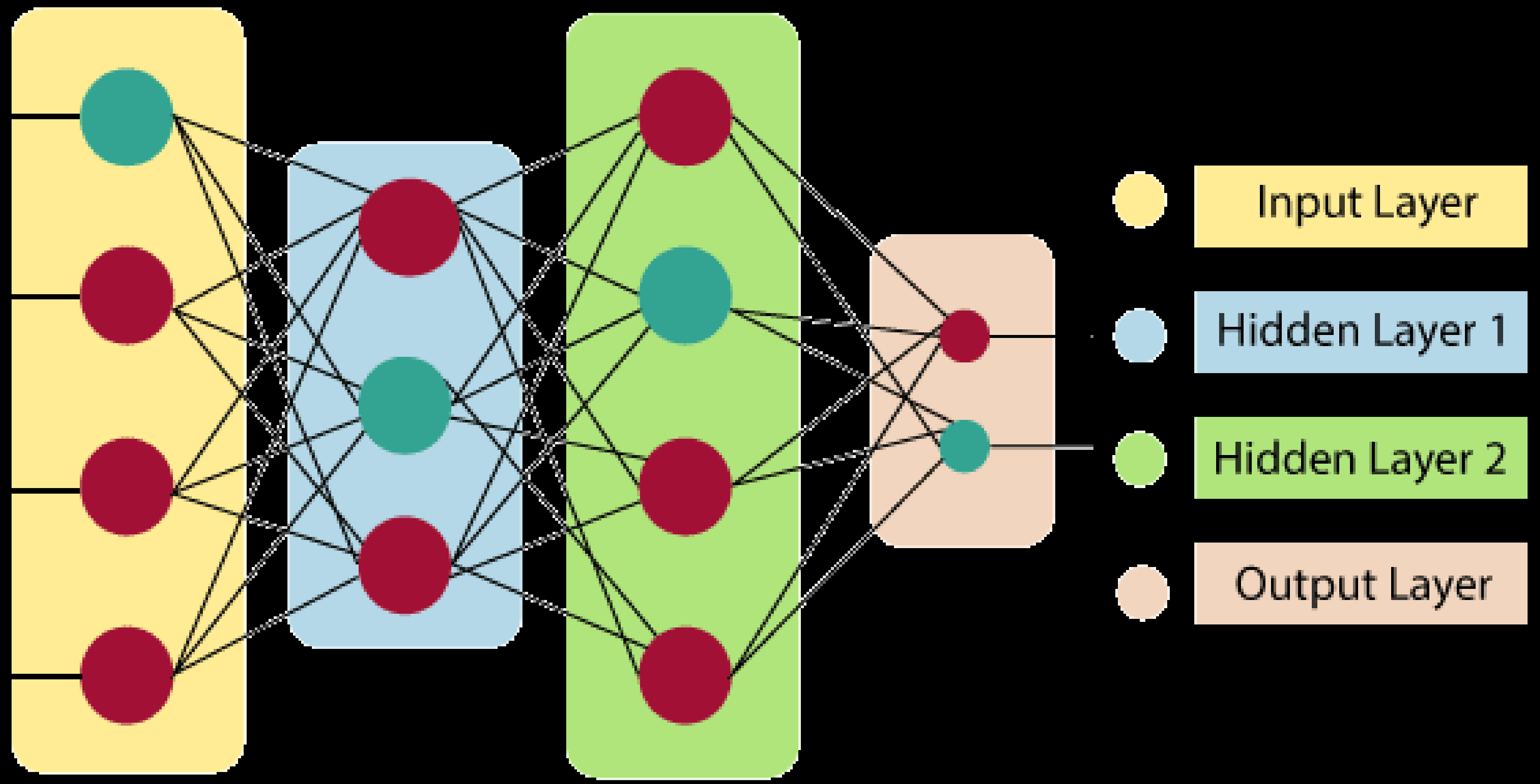
The behavior of a biological neural network can be captured by a simple model called artificial neural network.



A Biological Neuron



An Artificial Neuron



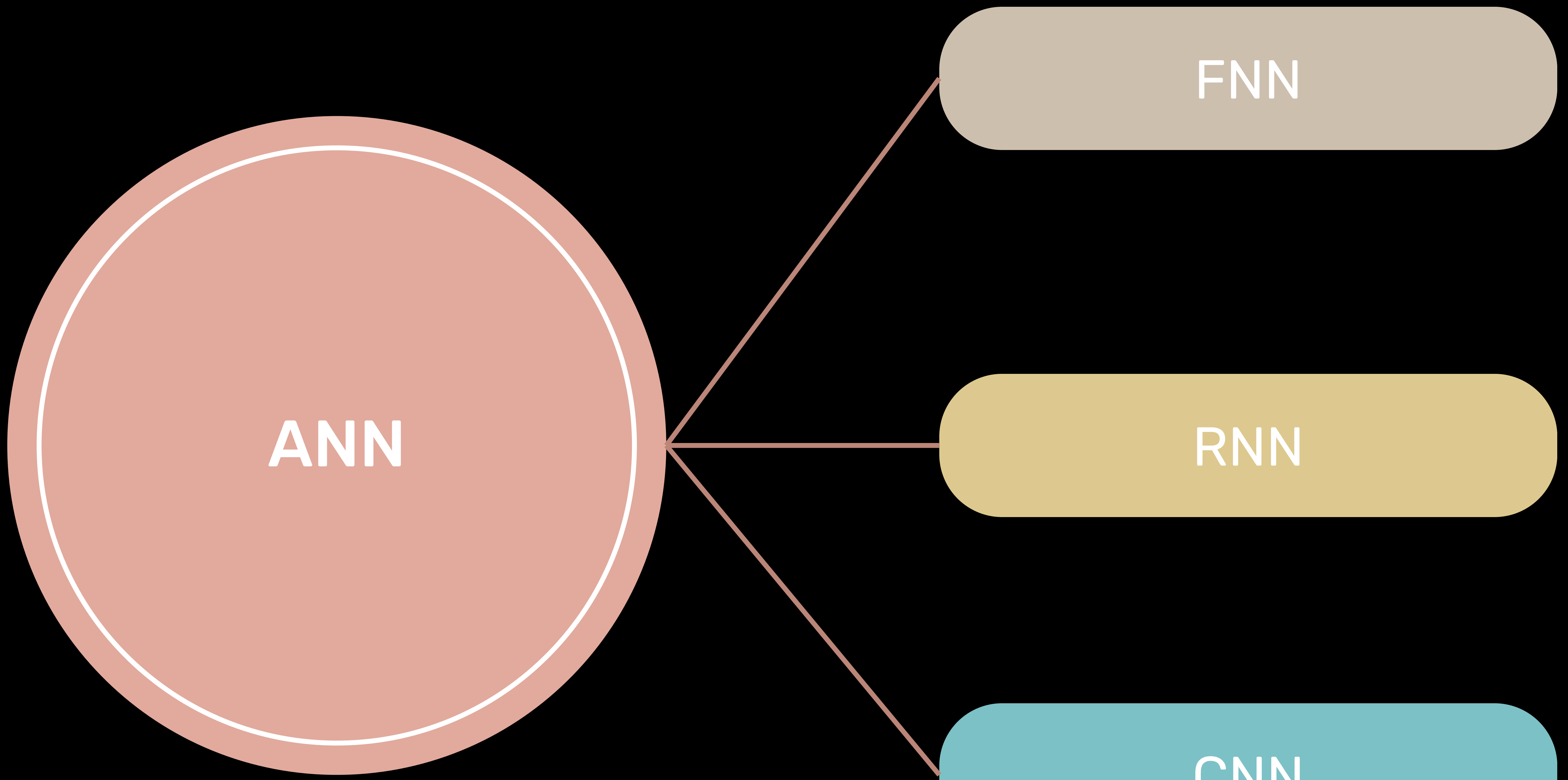
<u>ANN consists of 3 layers:</u>

Input Layer: The first layer of an ANN is the input layer, where input features are introduced into the network.

Hidden Layers: Intermediate layers, known as hidden layers, are responsible for computation and learning. Multiple hidden layers allow the network to capture very complicated relationships in the data.

Output Layer: The final layer, the output layer, produces the network's prediction or classification.

TYPES OF ANN:

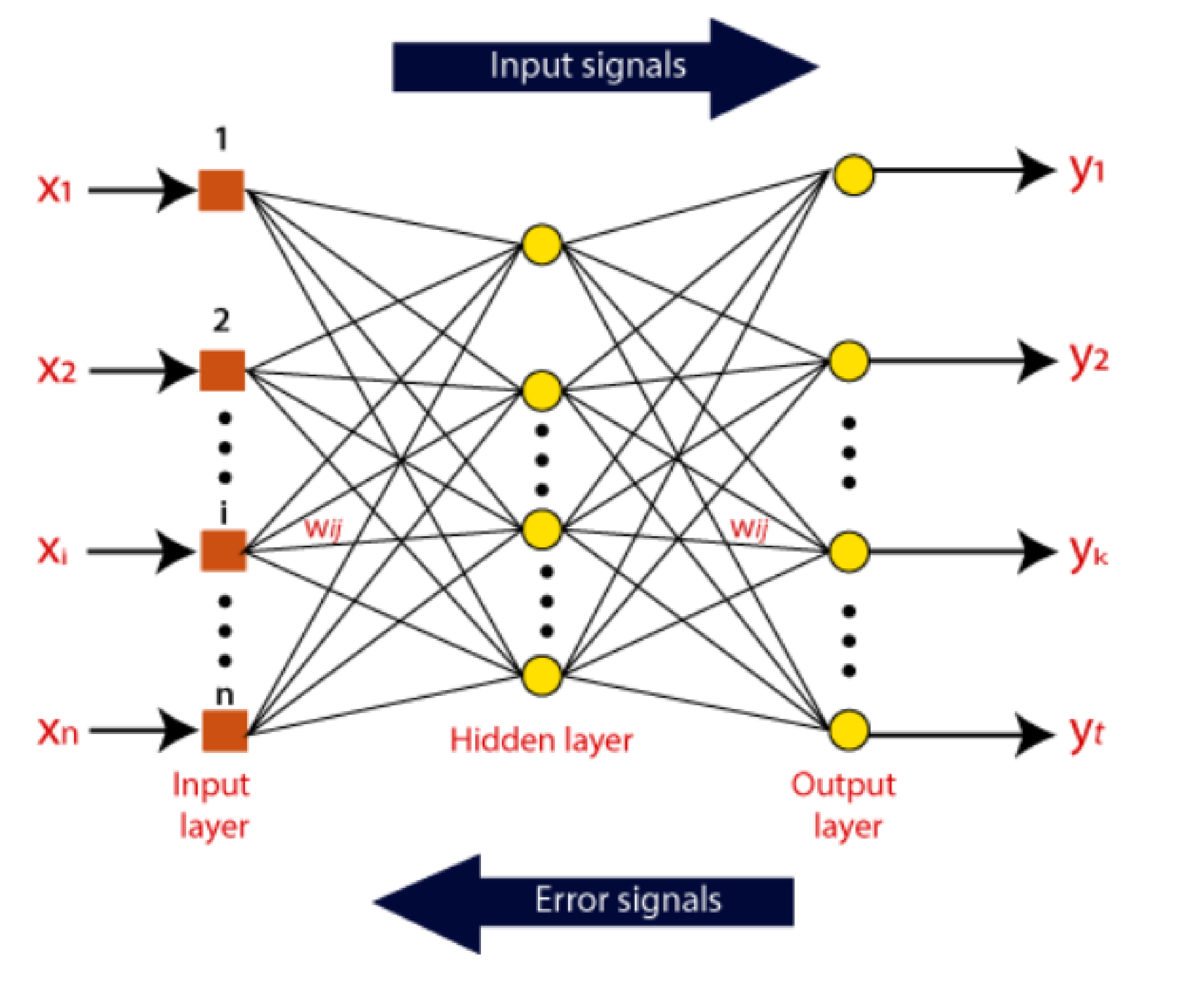


Feedforward Neural Networks (FNN): A feed-forward network is a basic neural network comprising an input layer, an output layer, and at least one layer of a neuron. Through assessment of its output by reviewing its input, the intensity of the network can be noticed based on group behavior of the associated neurons, and the output is decided. The primary advantage of this network is that it figures out how to evaluate and recognize input patterns.

Recurrent Neural Networks (RNN): a Recurrent Neural Network(RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other

Convolutional Neural Networks (CNN): A Convolutional Neural Network (CNN) is a type of Deep Learning neural network architecture commonly used in Computer Vision. Computer vision is a field of Artificial Intelligence that enables a computer to understand and interpret the image or visual data.

How do artificial neural networks work?



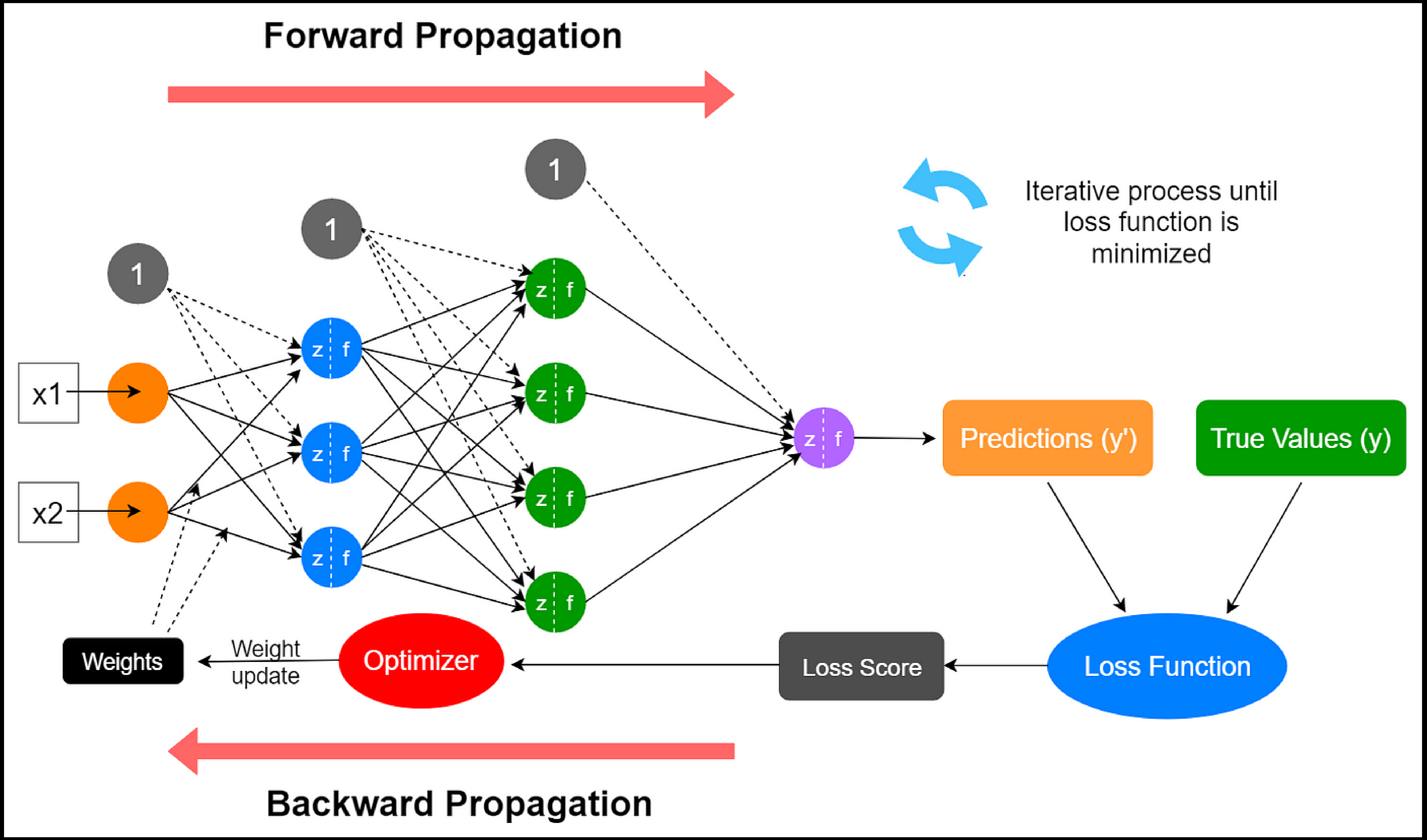
<u>Basic mechanism of the ANN:</u>

Connections between neurons are characterized by weights, determining the strength of the connection. During training, these weights are adjusted to minimize the difference between predicted and actual outputs.

Neurons also possess biases, which are added to the weighted sum before applying the activation function.

The activation process involves calculating the weighted sum of inputs, adding biases, and applying the activation function to produce an output.

Training of an ANN



Feedforward: During the feedforward process, input data propagates through the network, layer by layer, to generate an output.

Backpropagation: Backpropagation is a training process where the error in the predicted output is propagated backward through the network.

Weights are adjusted to minimize the error using optimization algorithms like gradient descent.

Loss Function: A loss function measures the disparity between predicted and actual outputs. The objective during training is to minimize this loss, enhancing the accuracy of predictions.

Training methods for ANN:

Supervised Learning: In supervised learning, the network is trained on a labeled dataset, where inputs are associated with correct outputs.

Unsupervised Learning: Unsupervised learning involves the network learning patterns and structures in unlabeled data.

Reinforcement Learning: Reinforcement learning enables the network to learn by interacting with an environment and receiving feedback in the form of rewards or penalties.

ANN & Hardware:

- ANNs can be implemented using CMOS digital circuits, leveraging existing logic elements and decades of digital circuit advances.
- Synaptic weights can be stored using digital memory cells or latches, with the number of bits critical for algorithm accuracy, especially during learning.
- Neuron state summation can be implemented using common multipliers and adder stages, but for large inputs, the element count can become significant.
- Implementing the activation function, particularly complex ones like sigmoid, requires look-up tables, slowing down computations and demanding significant power and area.

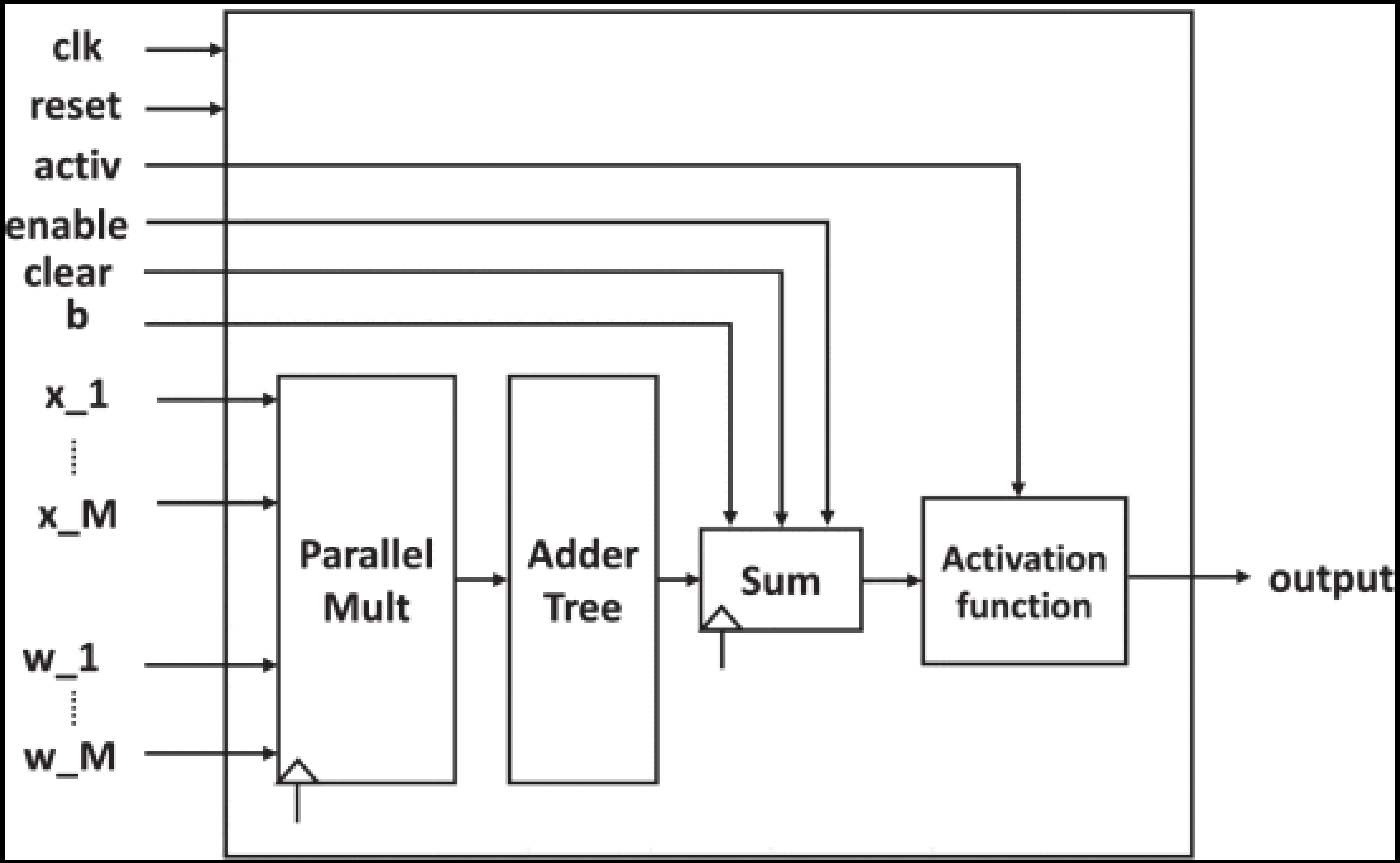
Training the hardware model:

1. **Off-chip learning** speeds up and improves accuracy by performing training calculations on software-simulated networks, but overlooks hardware manufacturing variations, potentially leading to performance disparities between simulated and real-world implementations.

2. **Chip-in-the-loop** learning combines hardware network computation with external software for algorithm execution, ensuring high precision regardless of hardware limitations, thereby facilitating accurate model updates while accounting for real-time network behavior.

3. **On-chip learning** conducts learning solely on the hardware chip, making it suitable for embedded systems with continuous learning capability, albeit at slower and less precise weight calculations, necessitating a trade-off between flexibility and hardware complexity.

A Proposed Architecture:



(Binary radix implementation)

The architecture of the binary radix design is shown Fig. The architecture uses M parallel inputs x, M weight inputs w and bias b. The choice of M is configurable. Other inputs include the control signal from the Controller block. The Parallel Mult and the Adder Tree block will calculate the sum-of-product between x and M. The output is loaded to Sum register.

The Activation function block is used to implement the activation function of each neuron. In this work, two of the most popular activation functions are implemented, which are the Sigmoid function and the ReLU function.

THANK YOU