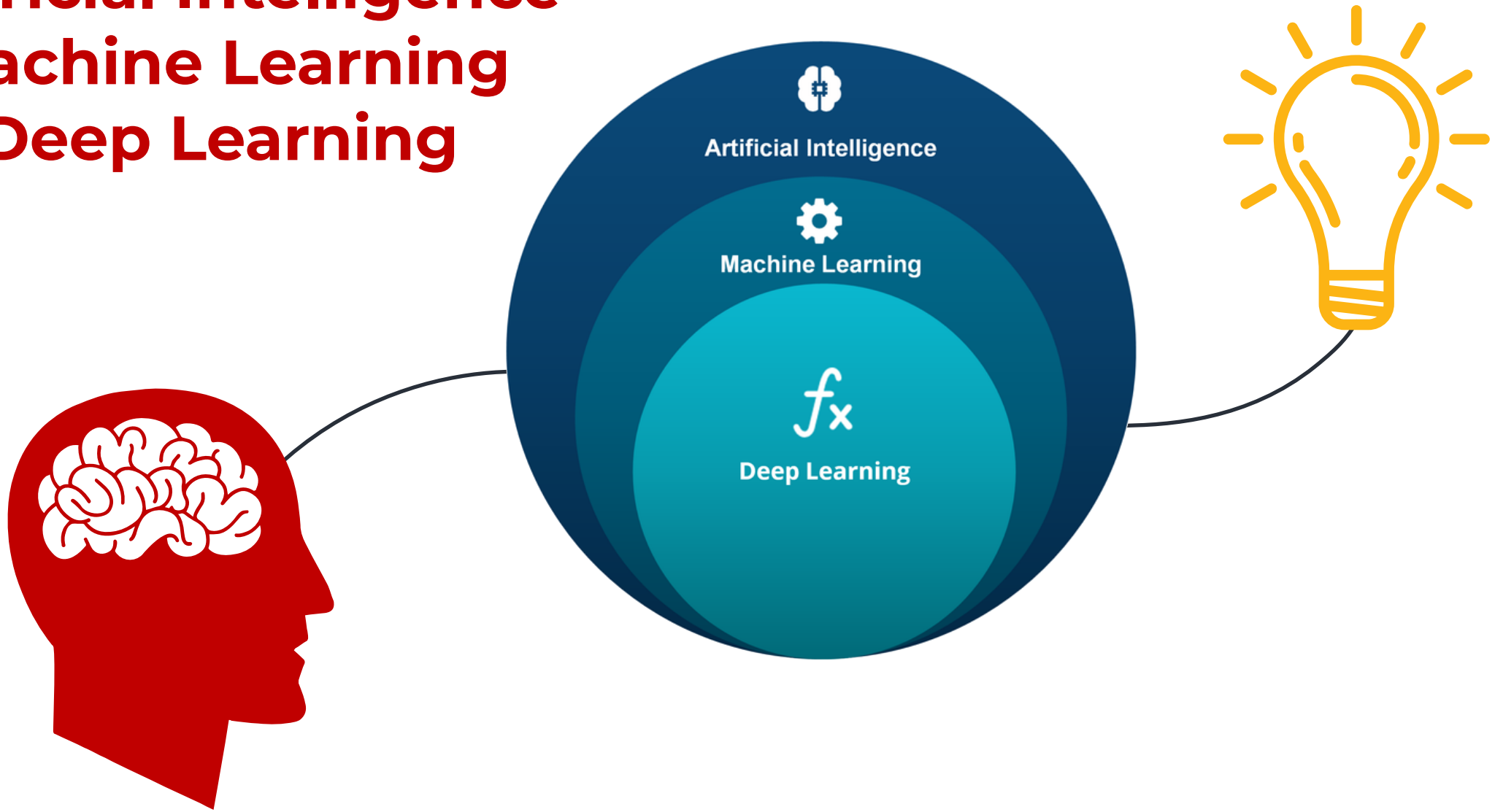
Welcome To

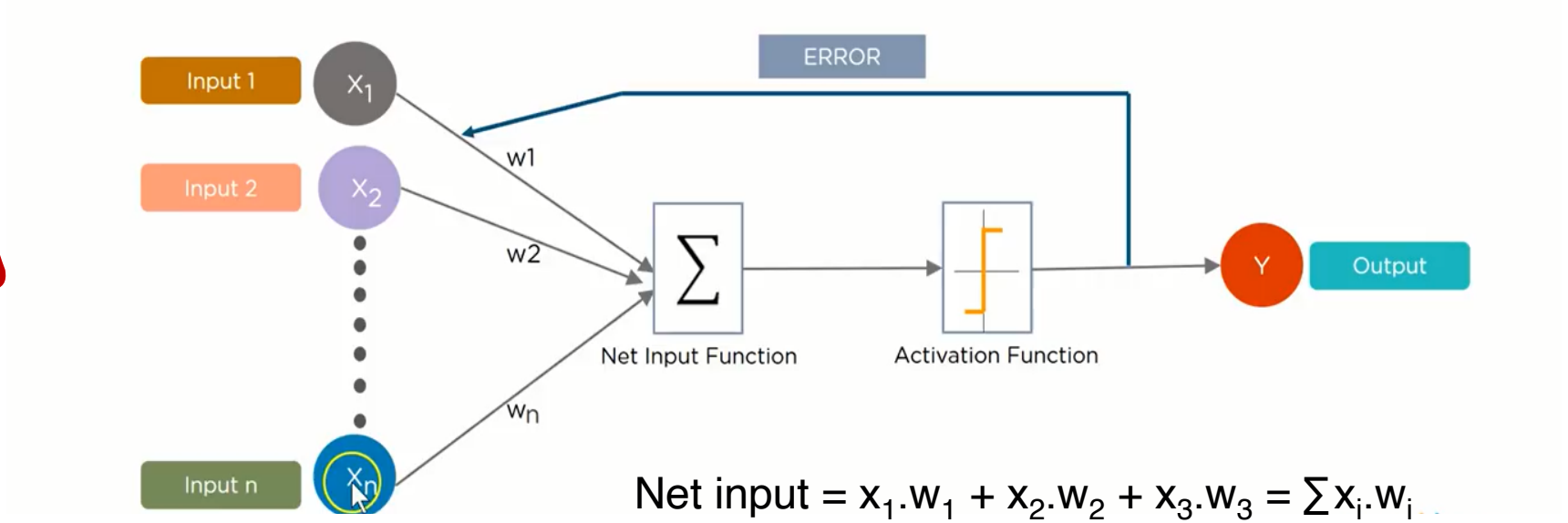
Telkom University

Research Center Digital Business Ecosystem

Artificial Intelligence
Machine Learning
Deep Learning



Component Of Artificial Neural Networks



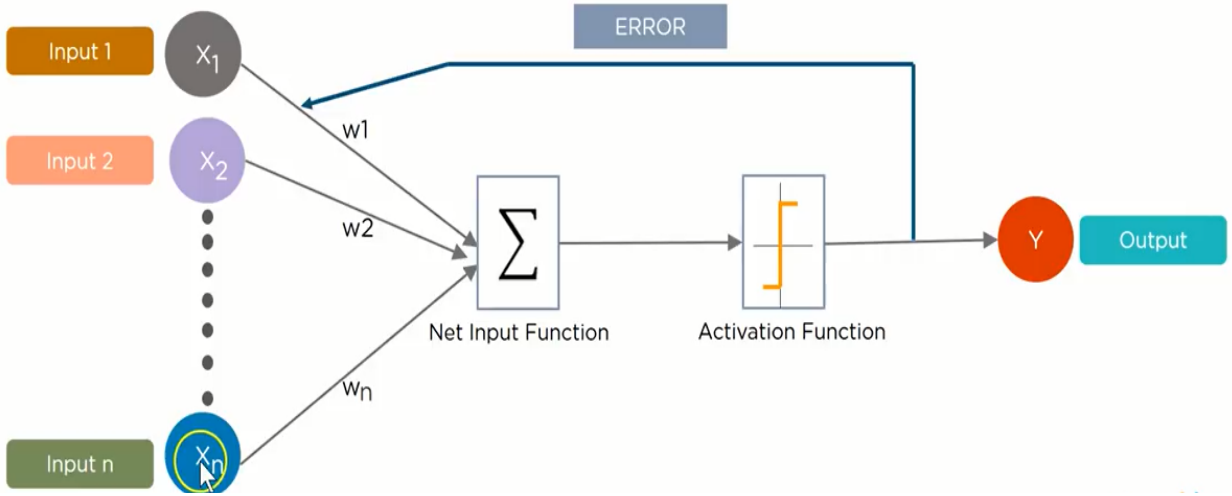
- Input
- Weight
- Bias
- Activation Function

Net input = $x_1.w_1 + x_2.w_2 + x_3.w_3 = \sum x_i.w_i$

Or using simple matrix multiplication

[w1 w2 w3] [x1 x2 x3] $= x_1.w_1 + x_2.w_2 + x_3.w_3$

How It Work Perceptron



- Perceptron is the basic part of a neural network, which represents a single neuron

- A neuron is a computational unit that calculates a piece of information based on weighted input parameters

- Inputs accepted by the neuron are separately weighted.

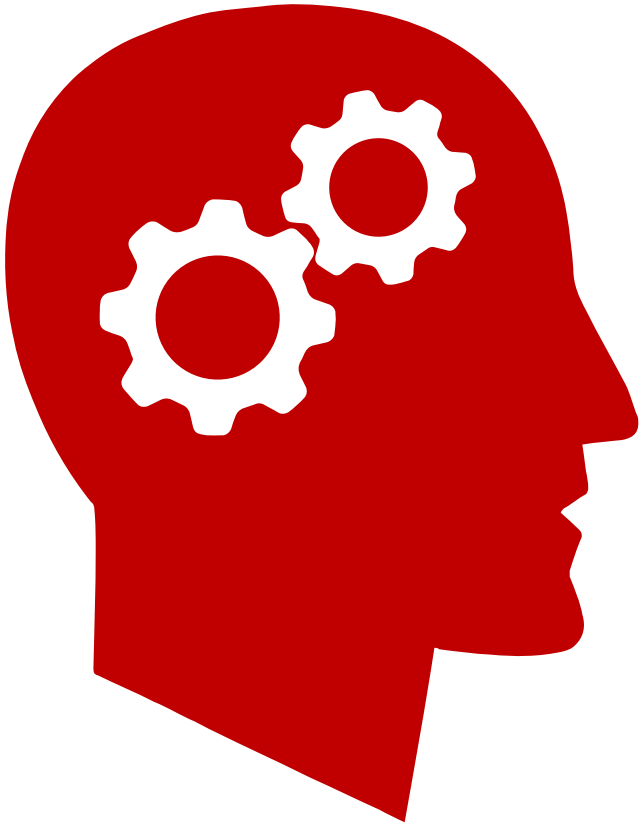
- Weights are real numbers expressing the importance of the respective inputs to the output and can be adjusted during learning phase

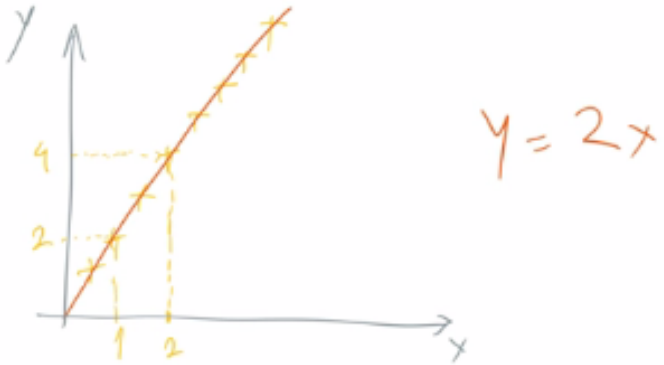
$$y = (x_0 w_0) + (x_1 w_1) + (x_2 w_2)$$

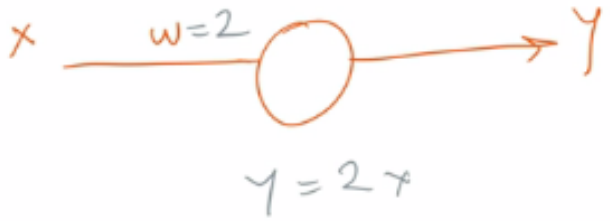
$$= \sum_i x_i \; w_i$$

$$= [w_0 \; w_1 \; w_2] \; \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$$

How It Work
Bias







y = 2x

- A bias value allows you to shift the activation function to the left or right, which may be critical for successful learning.

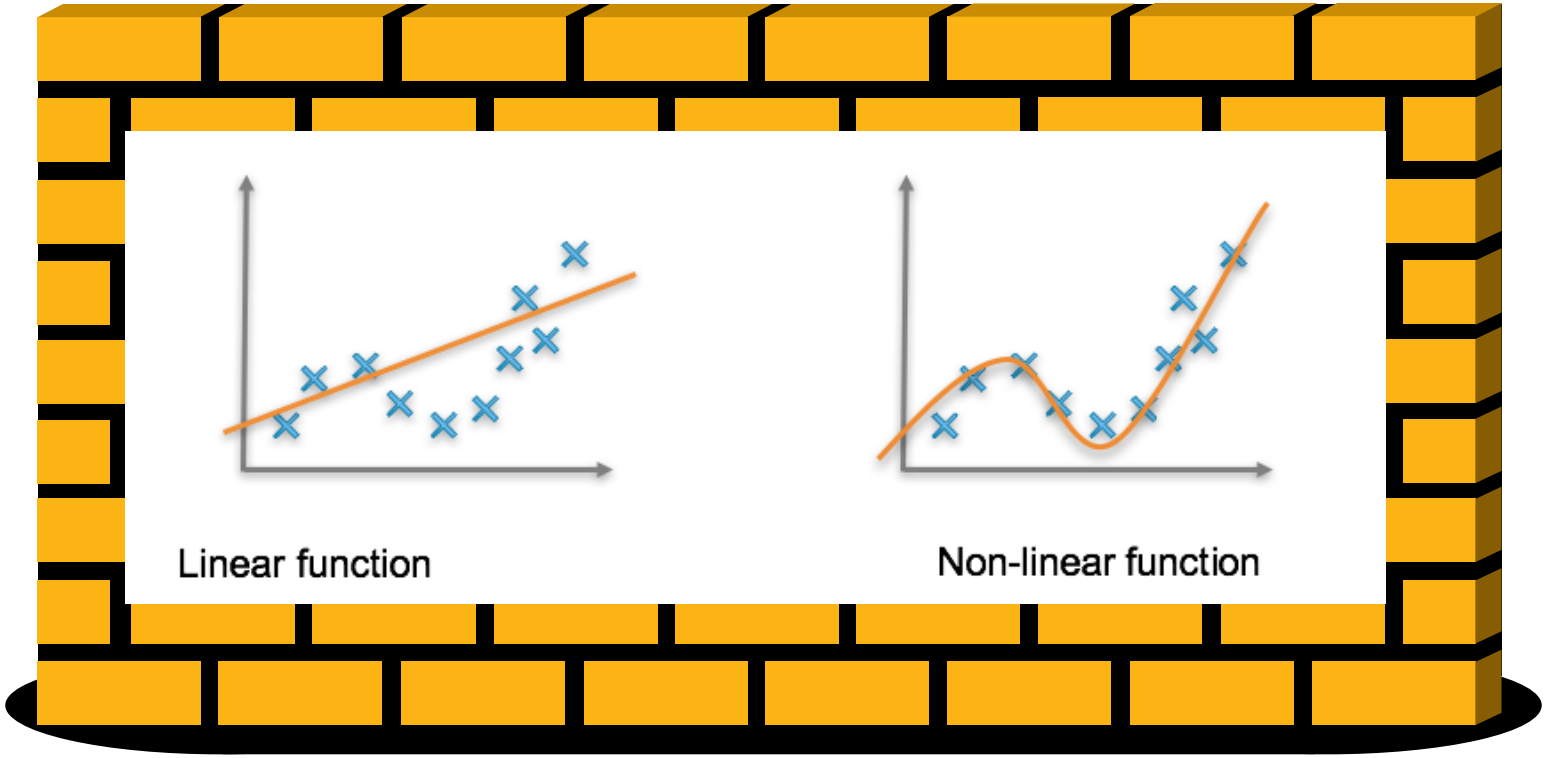
- Changes in weight change the steepness of the curve, while bias shifts the entire curve so that it is more suitable.

- Bias only affects the output value, it does not interact with the actual input data

- Bias value can be adjusted during learning phase

$$y = \sum_i x_i w_i + b$$

Activation Functions

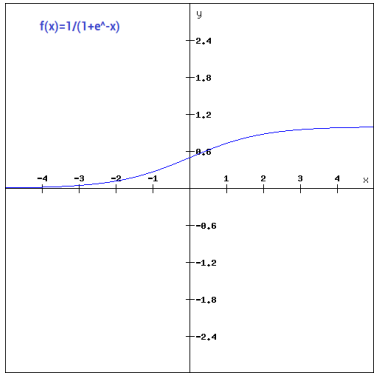


Decides whether a neuron should be activated or not by calculating weighted sum and adding bias with it.

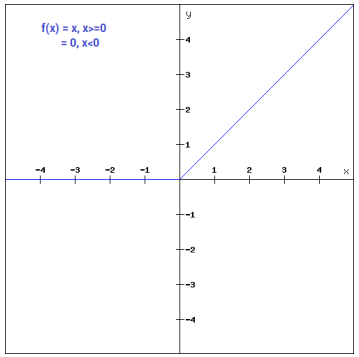
Introduce non-linearity into the output of a neuron → making it capable to learn and perform more complex tasks

If we do not use activation function, the output signal produced is only a linear function

Activation Functions

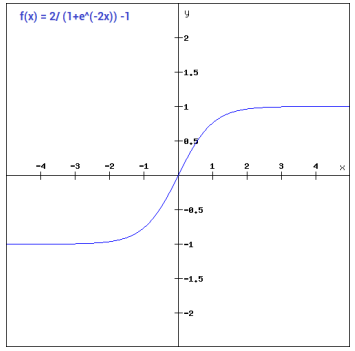


Sigmoid function



ReLu function

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



tanh function

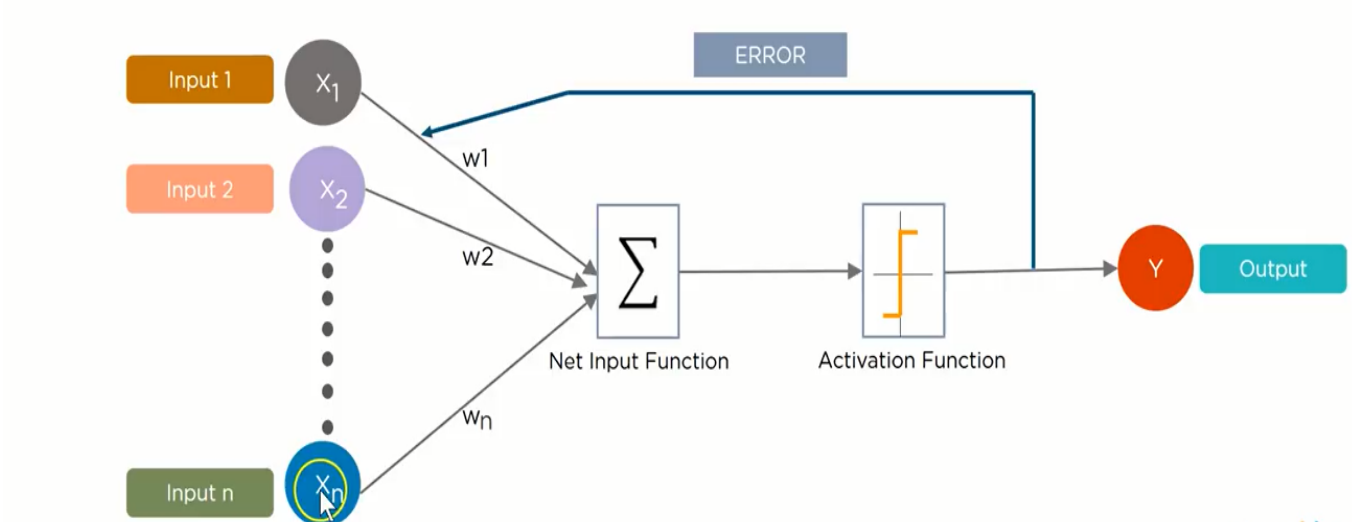
Commonly used activation functions:

- Sigmoid function :
- Tanh function :
- ReLU (Rectified Linear Unit) : A(x) = max(0,x)
- Softmax function : converts an array of values into an array of probabilities (0 - 1)

Learning Rule
2 Steps of Learning

❑ Forward propagation

- weights and bias are initialised randomly
- calculate forward (from input layer to output layer) to get the output
- compare it with the real value to get the rror → using lost function or cost unction



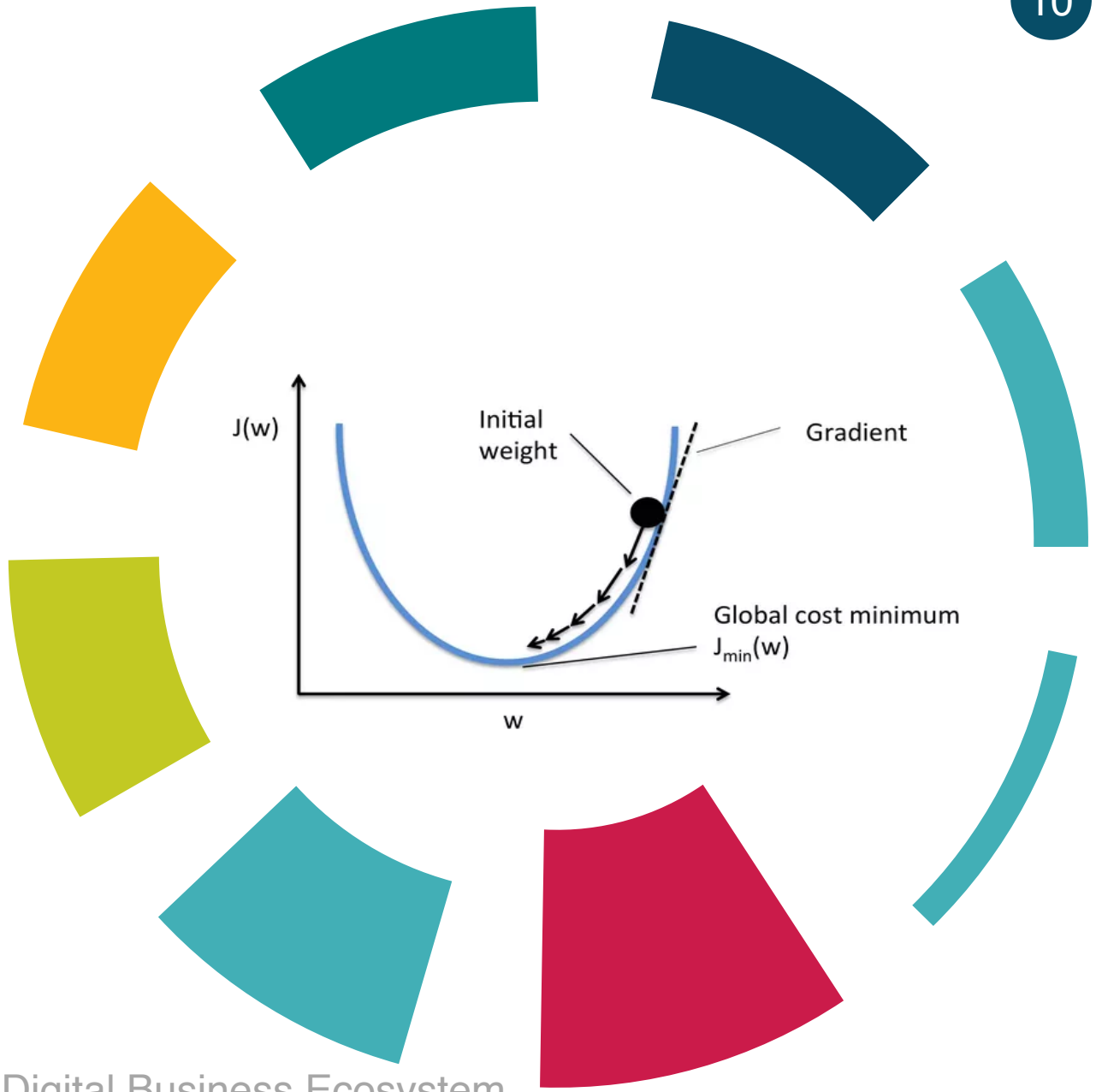
❑ Backward propagation

propagate backwards and do:

- finding the derivative of the lost function with respect to each weight
- update the weight by subtracting this value from the weight value.

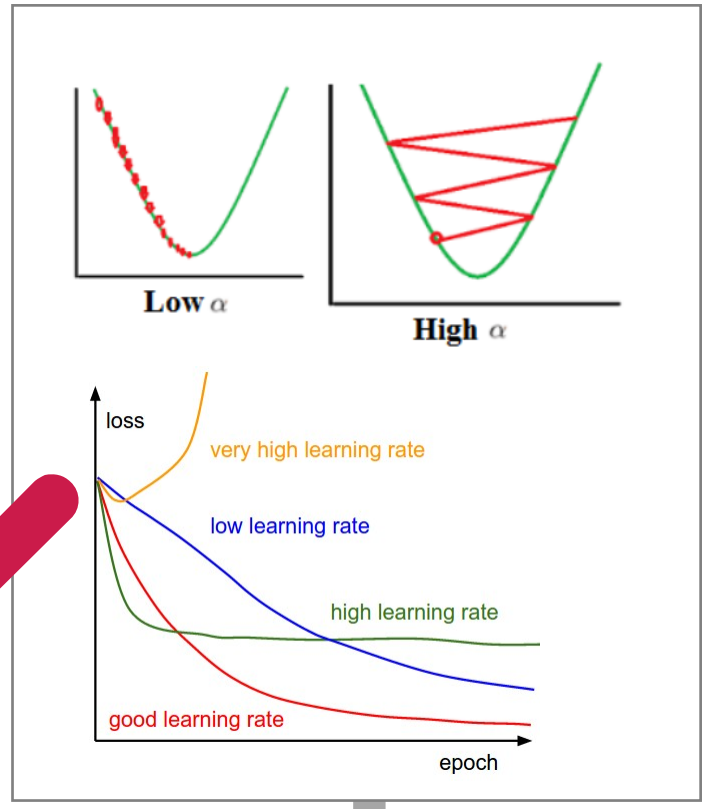
Loss Function and Gradient Descent

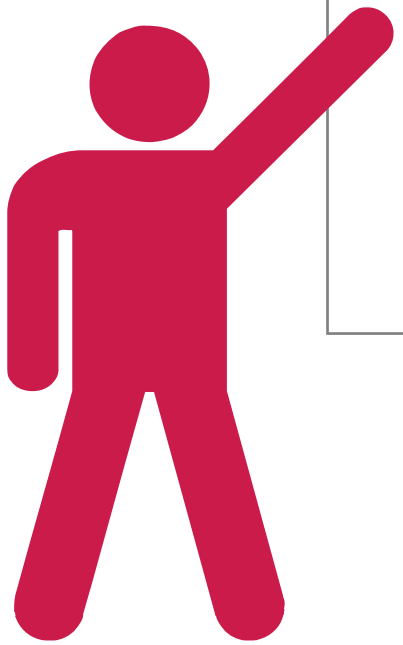
1. Loss function : a method of evaluating how well your algorithm models your dataset.
2. Learning = minimizing loss function
3. How = change the weights* by calculating the gradient (i.e. the partial derivative of loss function with respect to weights)
4. Since we want the value to be minimum → gradient descent. This is the simplest and most popular optimization method for deep neural network
5. Other methods : Newton's method, Conjugate Gradient, Quasi-Newton, Levenberg-Marquardt algorithm.



Learning Rate

the effects of different learning rates





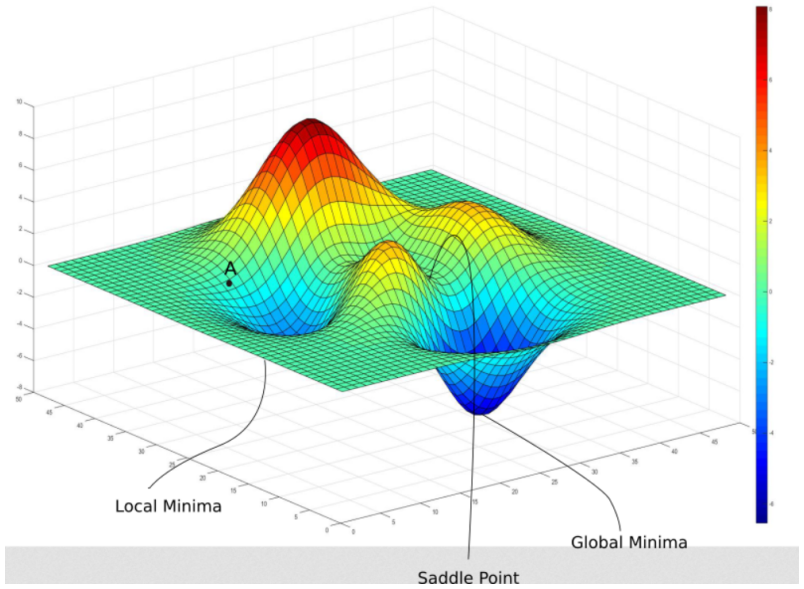
- The gradient told us the direction to change the weight.
- How much we must change is determined by another hyperparameter called learning rate or α
- Picking the value of α is crucial
  - too small → training process too slow
  - too big → diverging away from the minimum / overshoot
- We can pick α manually or using optimization technique that utilize adaptive learning rate (e.g. Adagrad, RMSProp, Adam)
- For example:

$$w' = w - \alpha \, (\partial E / \partial w) \quad \rightarrow \alpha \text{ is learning rate}$$

Gradient Descent Implementation

When we should update the weight? → there are 3 types of gradient descent implementation

Stochastic Gradient Descent : randomized the data, and update the weights for each training instance.



Challenges in Gradient Descent

Batch Gradient Descent : process all training dataset in a single batch, calculate the error and update the weight accordingly.

Minibatch Gradient Descent : splits the training dataset into small batches, calculate error and update weights for each small batches.

1. The complicated nature of deep neural network makes it prone to overfitting.
2. A model is overfit if it performs well on training data but not generalize well on new unseen data
3. Regularization is any modification we make to the learning algorithm that is intended to reduce the generalization error, but not its training error (Ian Goodfellow)
4. The purpose is to control model complexity and reduce overfitting

Overfitting and Regularization

Most Common Regularization Technique

1

Weight penalty L2 & L1

Loss = Loss + Regularization

L2 → Loss = Loss + α ∑w2

L1 → Loss = Loss + α ∑w

2

Dropout

At each training iteration a dropout layer randomly removes some nodes in the network along with all of their incoming and outgoing connections.

3

Early stopping

Interrupting the training procedure once model's performance on a validation set gets worse

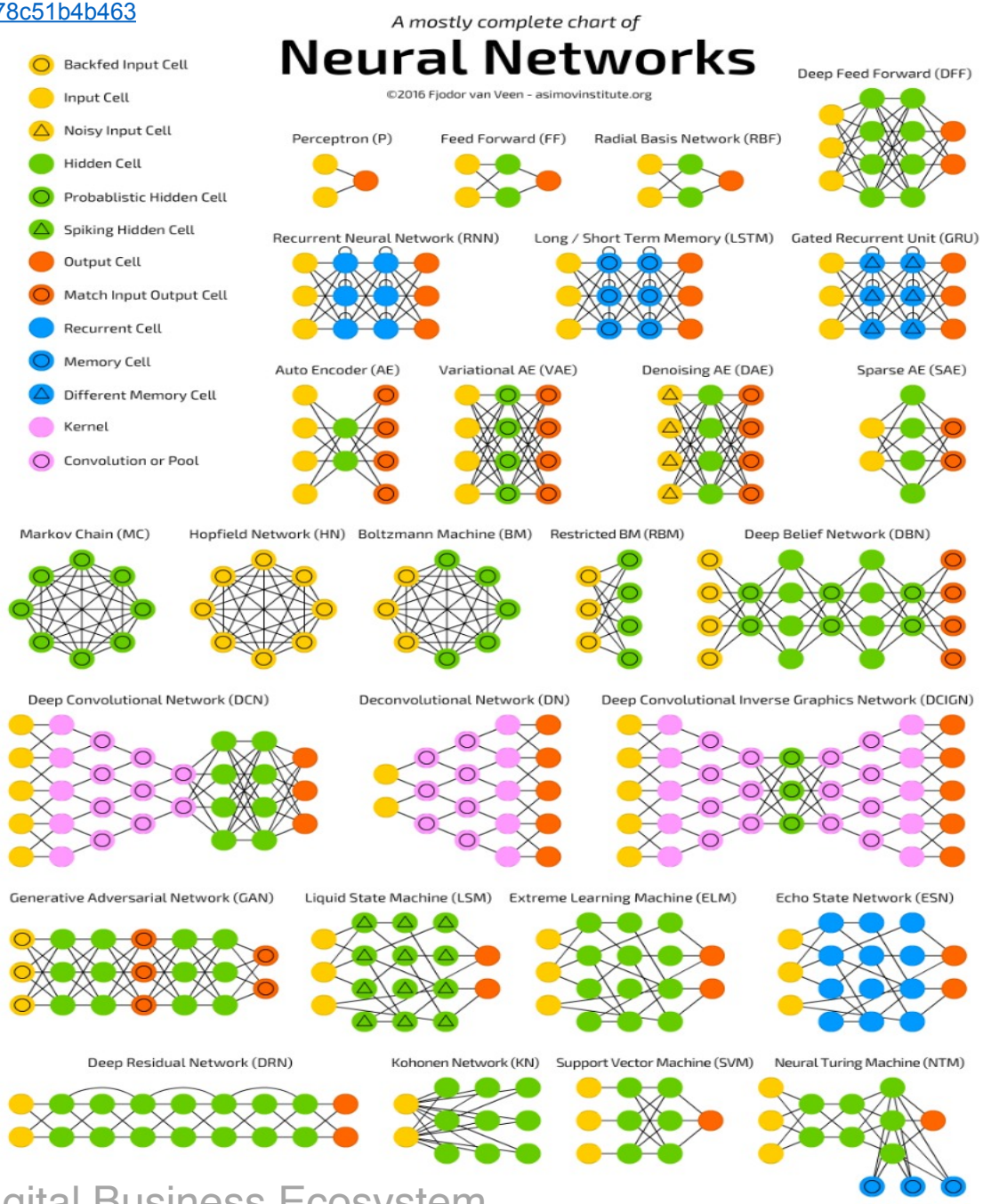
4

Dataset augmentation

Use bigger dataset by using synthetic data

https://becominghuman.ai/cheat-sheets-for-ai-neural-networks-machine-learning-deep-learning-big-data-678c51b4b463

Neural Network Chart

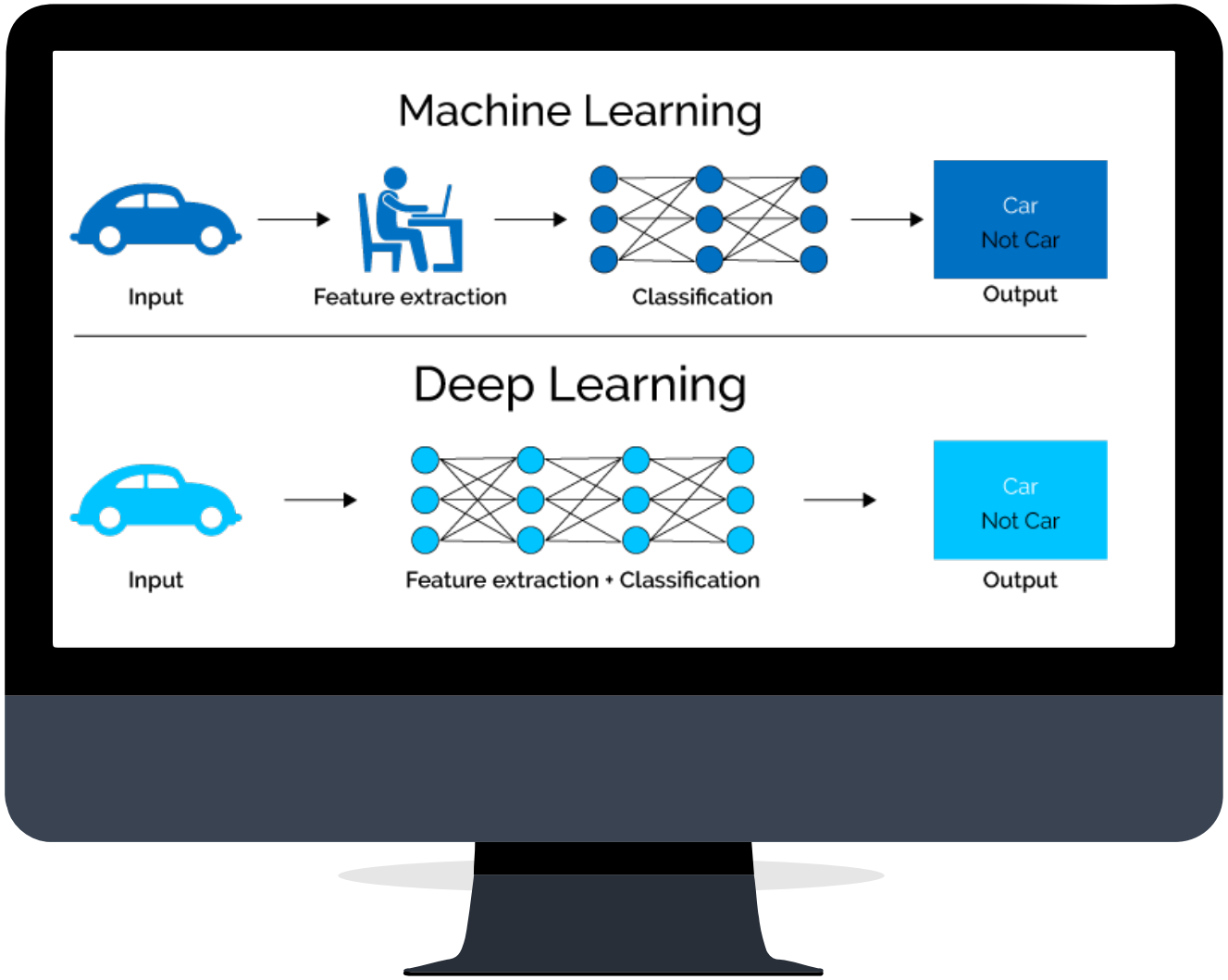


Deep Learning vs Classical Machine Learning

1. In classical machine learning, most of the features used require identification of domain experts

2. Deep networks scale much better with more data than classical ML algorithms

3. Deep learning techniques can be adapted to different domains and applications far more easily than classical ML



Why Now?

1. Exponential data growth (and the ability to Process Structured & Unstructured data)
2. Faster & open distributed systems (Hadoop, Spark, TensorFlow, …)
3. Faster machines and multicore CPU/GPUs
4. New and better models, algorithms, ideas:
   - Better, more flexible learning of intermediate representations
   - Effective end-to-end joint system learning
   - Effective learning methods for using contexts and transferring between tasks

"The analogy to deep learning is that the rocket engine is the deep learning models and the fuel is the huge amounts of data we can feed to these algorithms." - Andrew Ng

What is Deep Learning ?

1. Machine learning algorithms based on learning multiple levels (i.e deep) of representation/ abstraction (1)
2. Learning algorithms derive meaning out of data by using a hierarchy of multiple layers of units (neurons)
3. Each neuron/node computes a weighted sum of its inputs and the weighted sum is passed through a nonlinear function, each layer transforms input data in more and more abstract representations
4. Learning = find optimal weights from data

Why Deep Learning

1. Manually designed features are often over-specified, incomplete and difficult to design and validate. Learned Features are easy to adapt, fast to learn

2. Deep learning provides a very flexible, (almost?) universal, learnable framework for representing world, visual and linguistic information.

3. Insufficiently deep architectures can be exponentially inefficient

4. Distributed representations are necessary to achieve non-local generalization

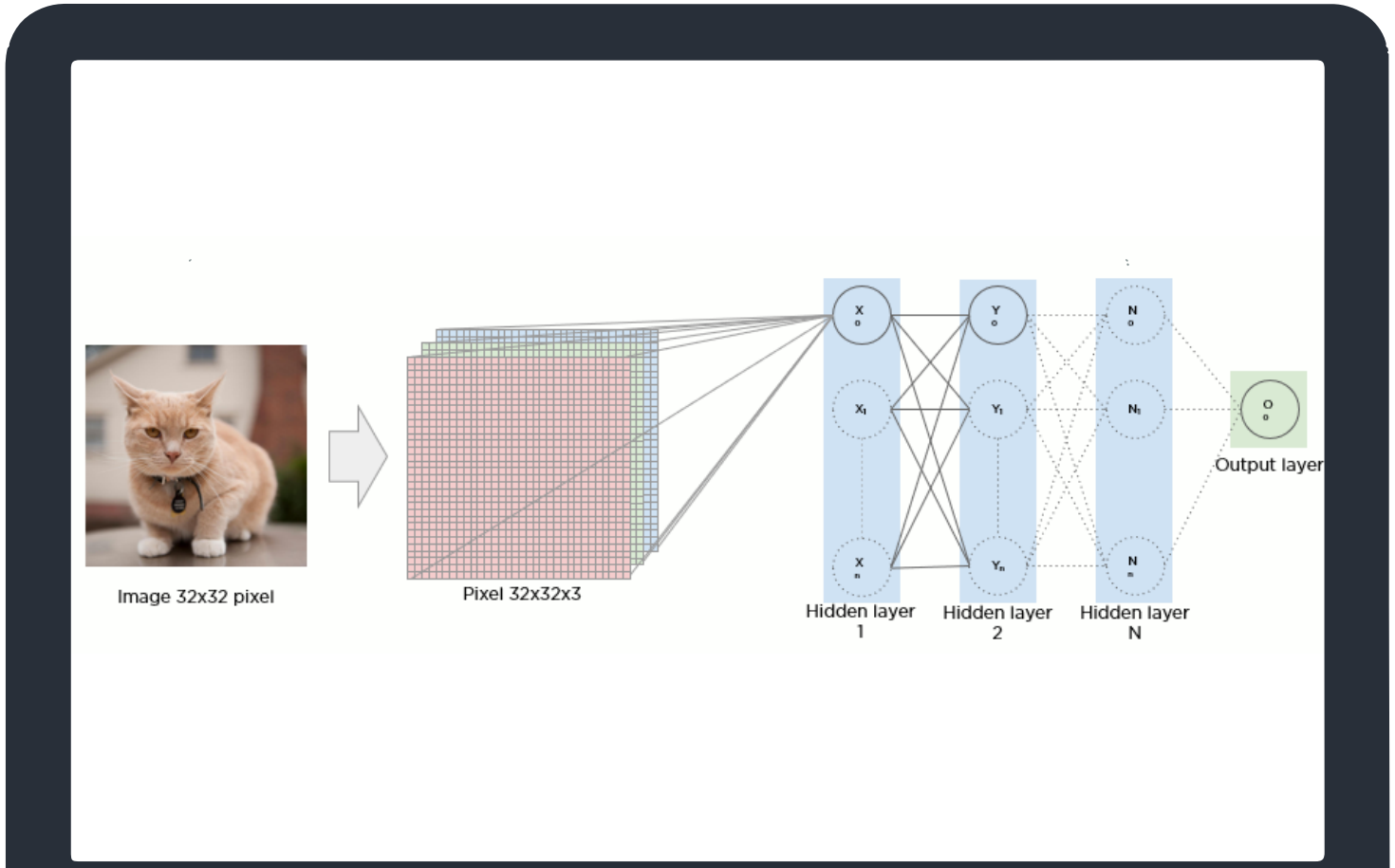


Image Processing in ANN

- Regular Neural Nets don't scale well to full images

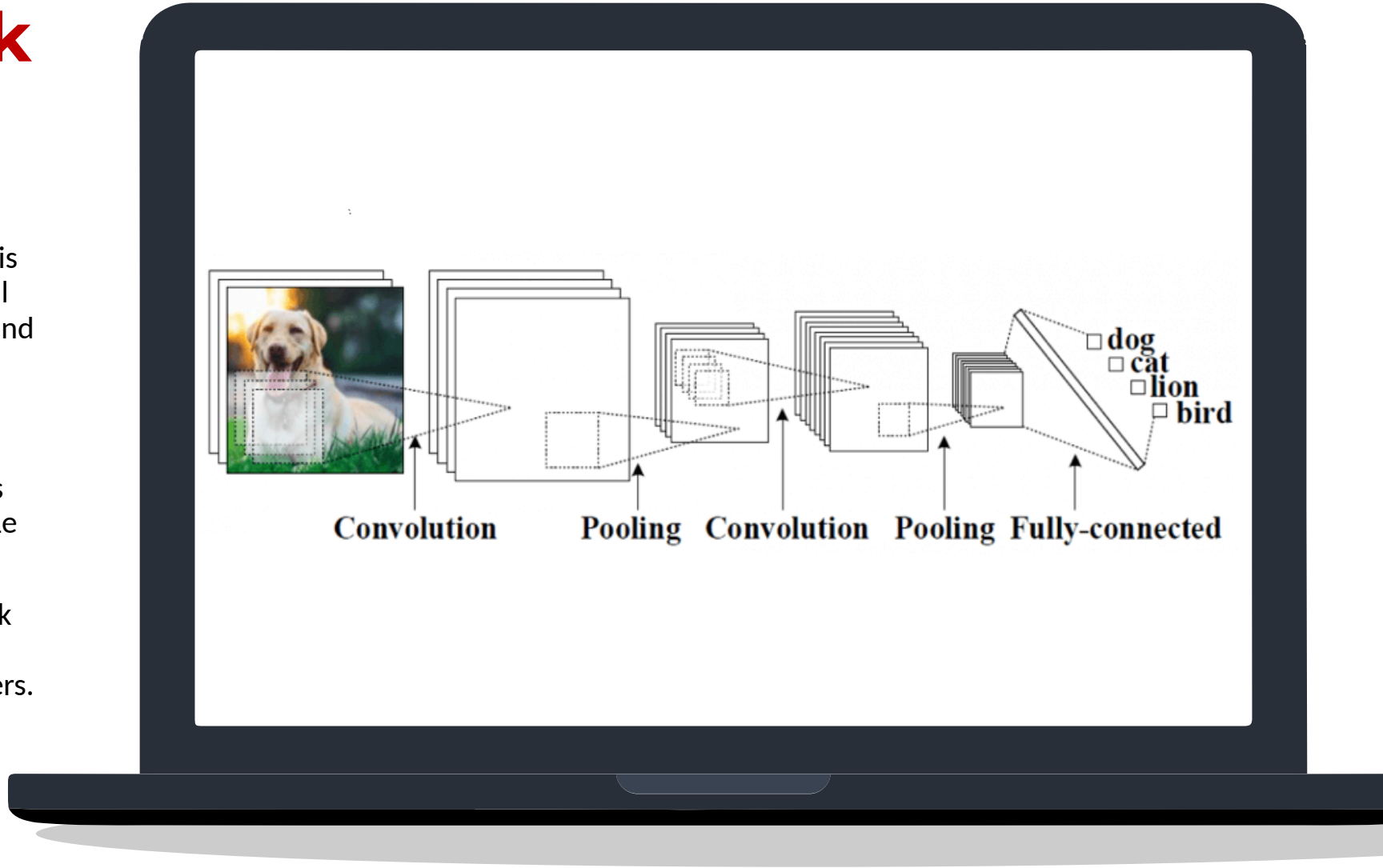
  E.g : CIFAR-10 images size 32x32x3 (32x32, 3 color channels - RGB) → 32*32*3 = 3072 weights in a single neuron in a first hidden layer

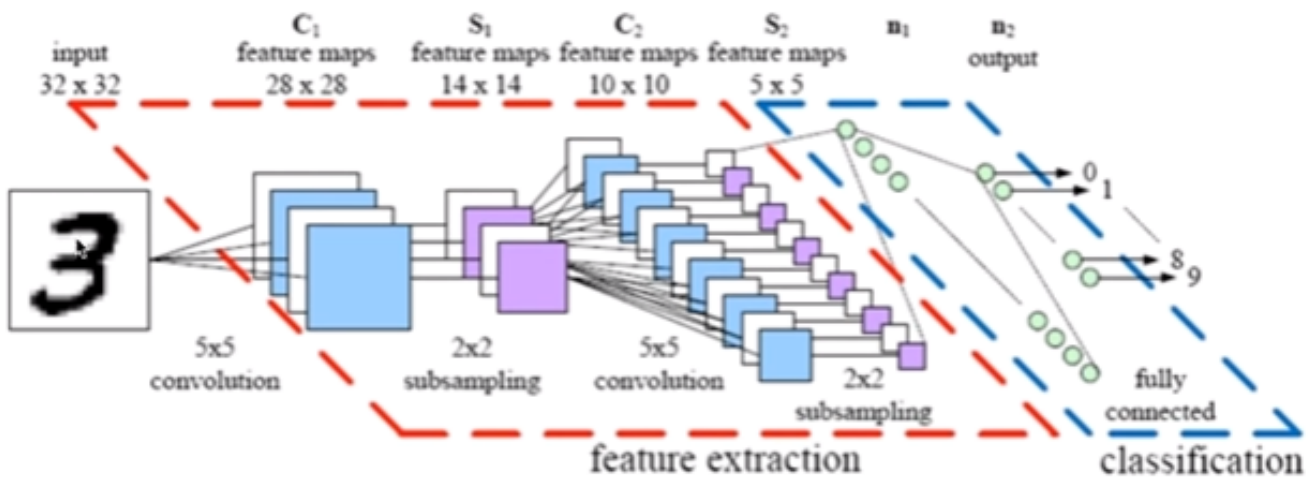
  For 200x200x3 image → 200*200*3 = 120,000 weights

- Parameters would add up quickly → quickly lead to overfitting.
- In order to reduce the number of parameters, it is used by using the convolution method

Convolutional Neural Network Overview

- A Convolutional Neural Network (CNN) is comprised of one or more convolutional layers (often with a subsampling step) and then followed by one or more fully connected layers as in a standard multilayer neural network.
- in 1998, Convolutional Neural Networks were introduced in a paper by Bengio, Le Cun, Bottou and Haffner.
- Their first Convolutional Neural Network was called LeNet-5 and was able to classify digits from hand-written numbers.





CNN Layers

- The common types of layers in CNN architecture : Convolution - Pooling - Fully Connected
  - Convolution layer is the core that does most of the computation
  - Pooling layer performs dimensionality reduction and controls overfitting. Commonly puts between the convolution layers. The Convolution and Pooling layer acts as feature extraction part
  - Fully Connected layer generally ends the CNN. It acts as classification part

Convolution Layer

- The first layer of CNN is always convolution layer
- Convolution layer consist of a set of learnable filters (also called kernel, or weight)
- Filter is a small array of number, and covers all the input depth.

For the CIFAR-10 data, we may have a 5x5x3 filter (5x5 matrix for 3 color channels)

Each neuron will have weights to a [5x5x3] region in the input volume → 5*5*3 = 75 weights → this is what we will learn

- The filter will convolve around the image, performing dot product operation between the filter and the part of the image it convolves with

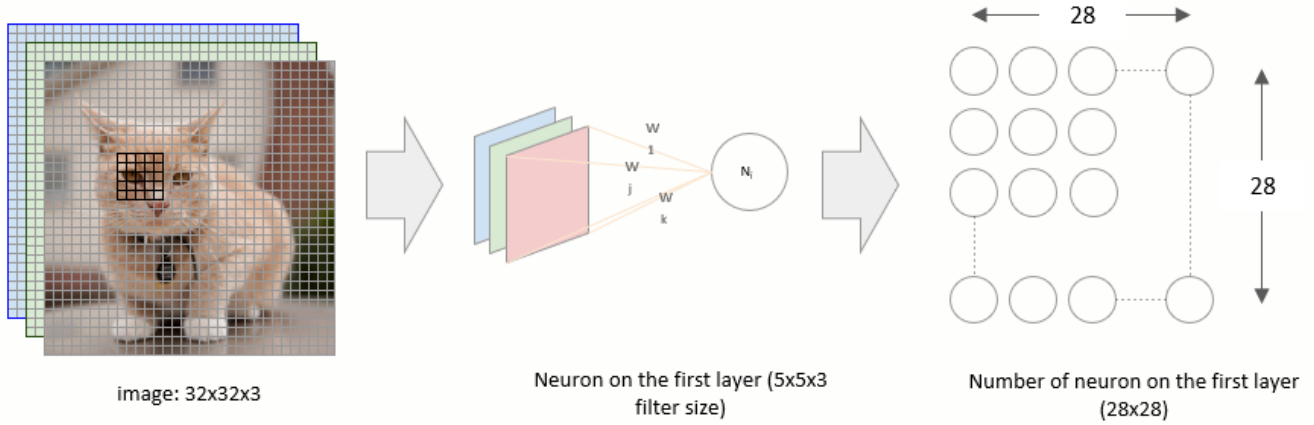
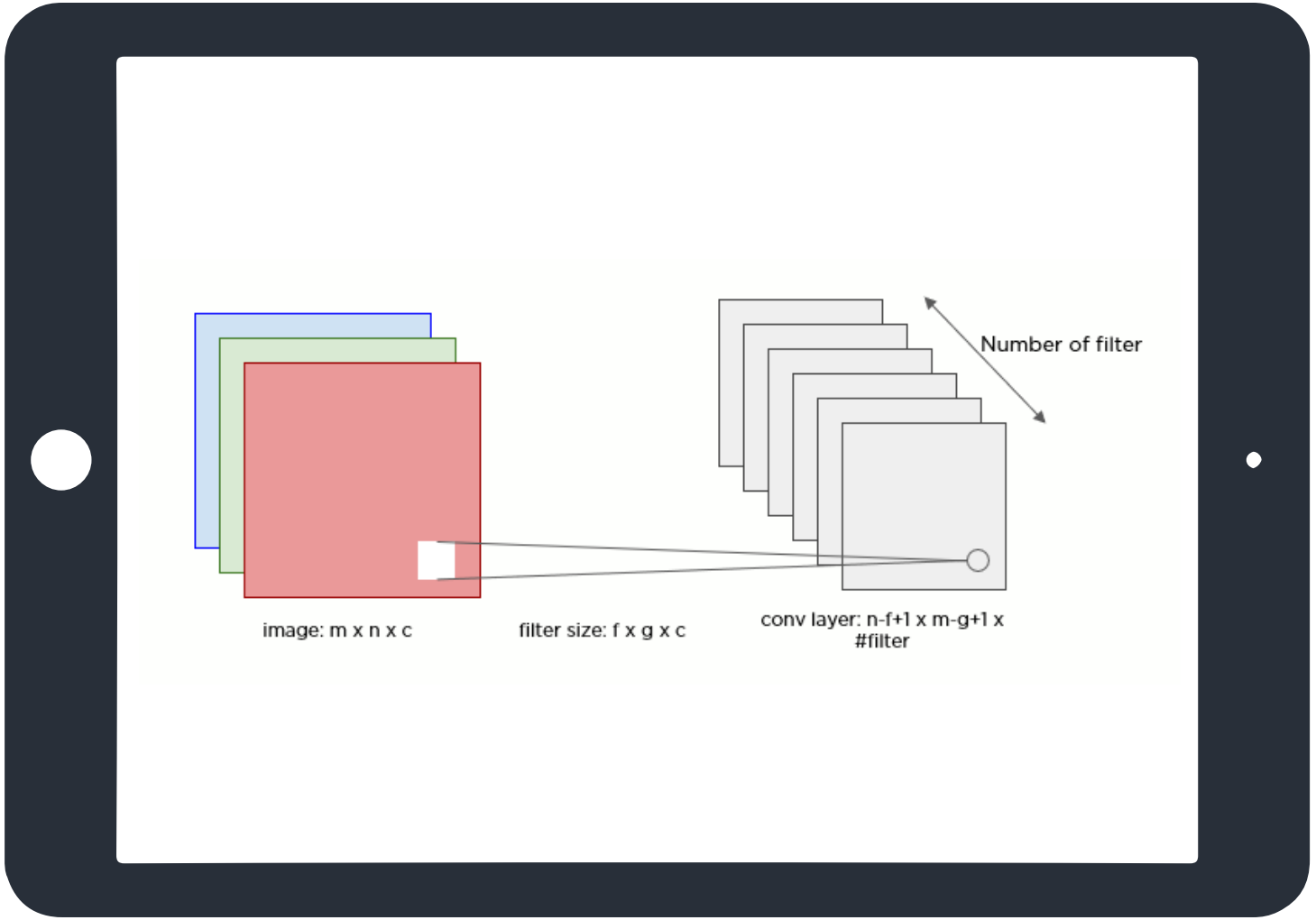


image: 32x32x3

Neuron on the first layer (5x5x3 filter size)

Number of neuron on the first layer (28x28)

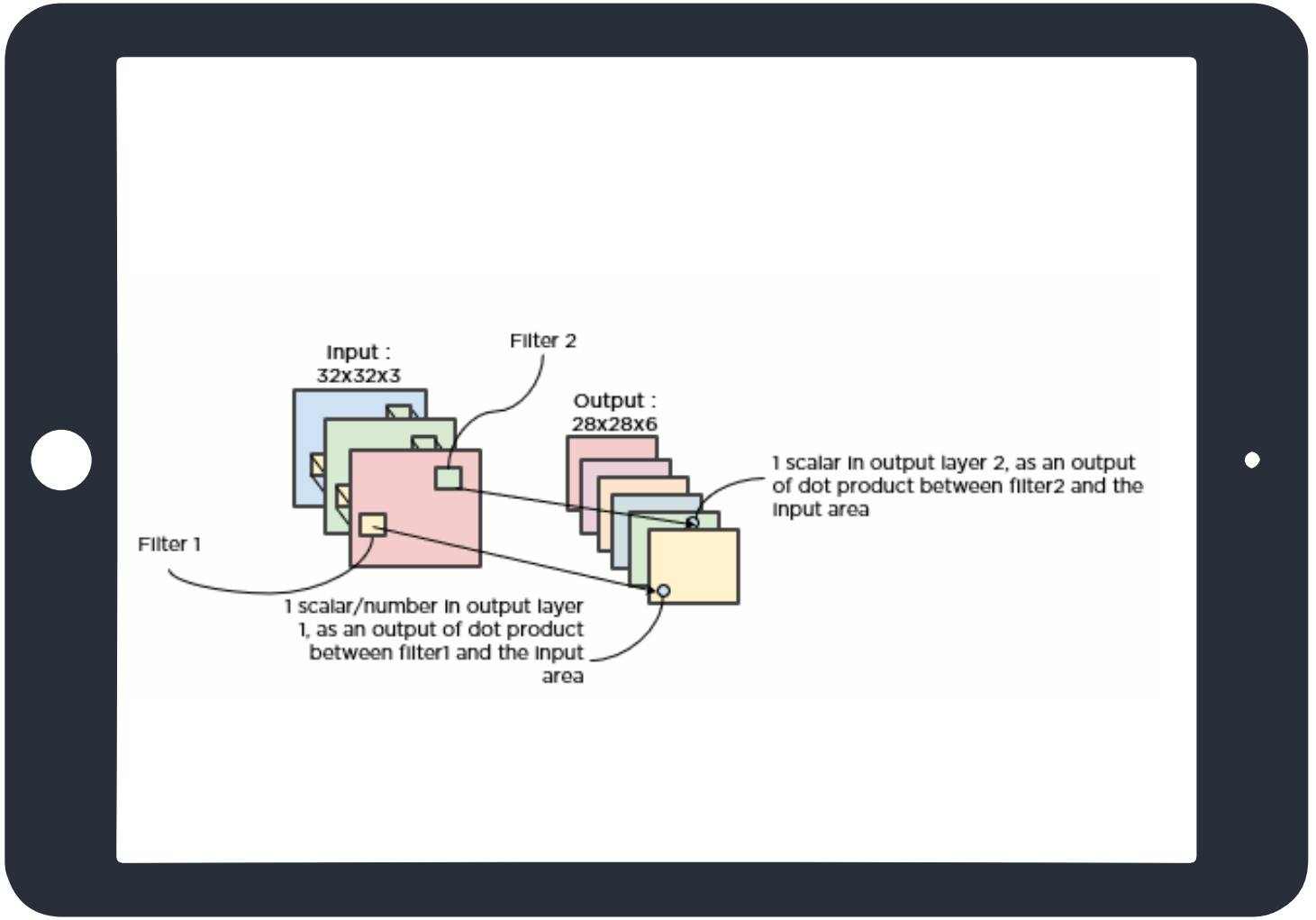


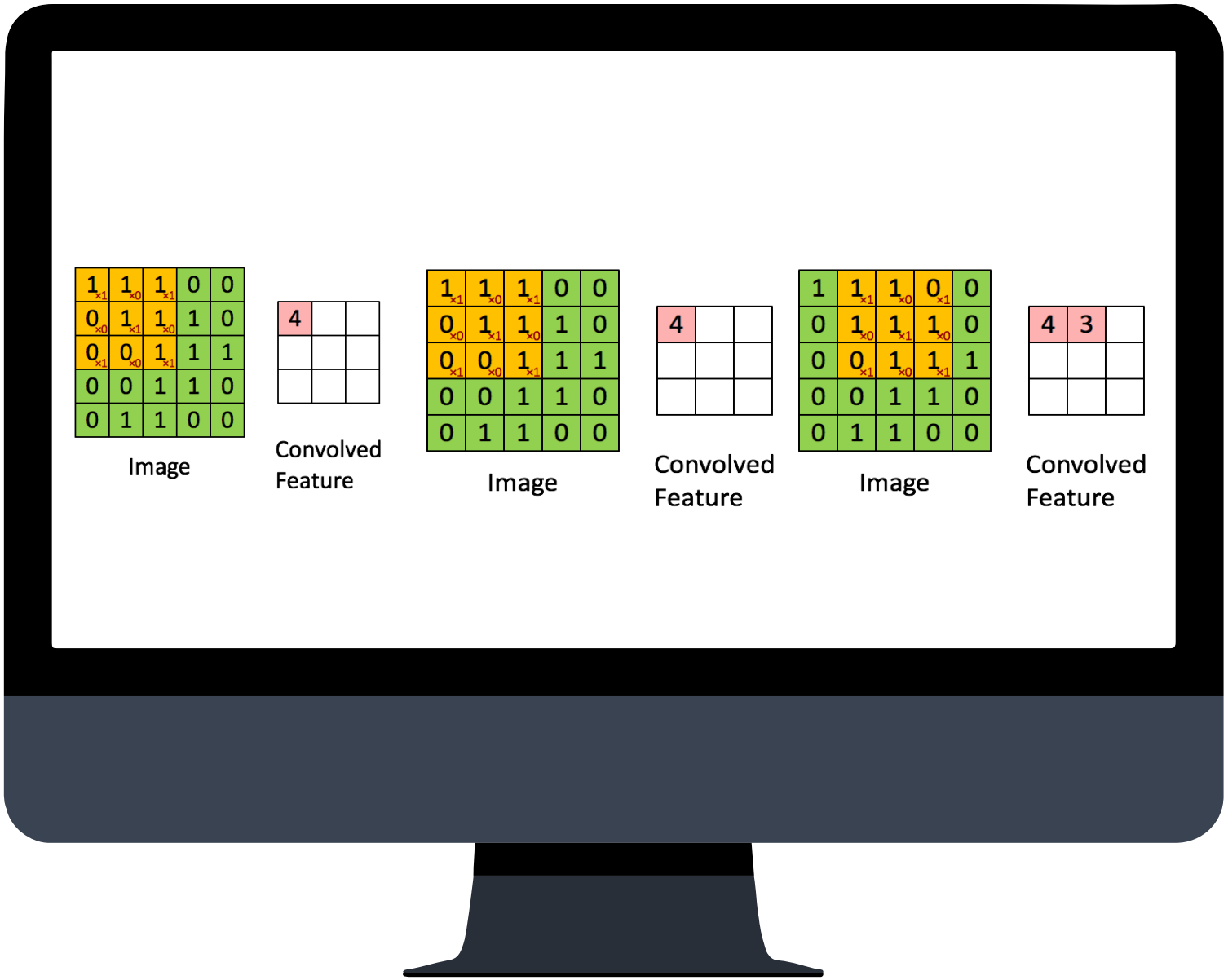
Convolution Layer Hyperparameter

- There are 3 hyperparameters to set in the conv layer: number of filter, stride, and padding
  - Filter size
  - Number of filter
  - Stride
  - Padding

Number of Filters

- The number of filters called depth column (or fibre) → note that it's different from input depth.
- We may have multiple filter for a conv layer, each layer learns different features (e.g. straight edges, blobs of color, curves, etc.)
- The 3rd dimension of output layer of a conv layer = the number of filters
- If we use 6 filters of 5x5x3 for our example, the output will be 28x28x6 (6 layers of 28x28)



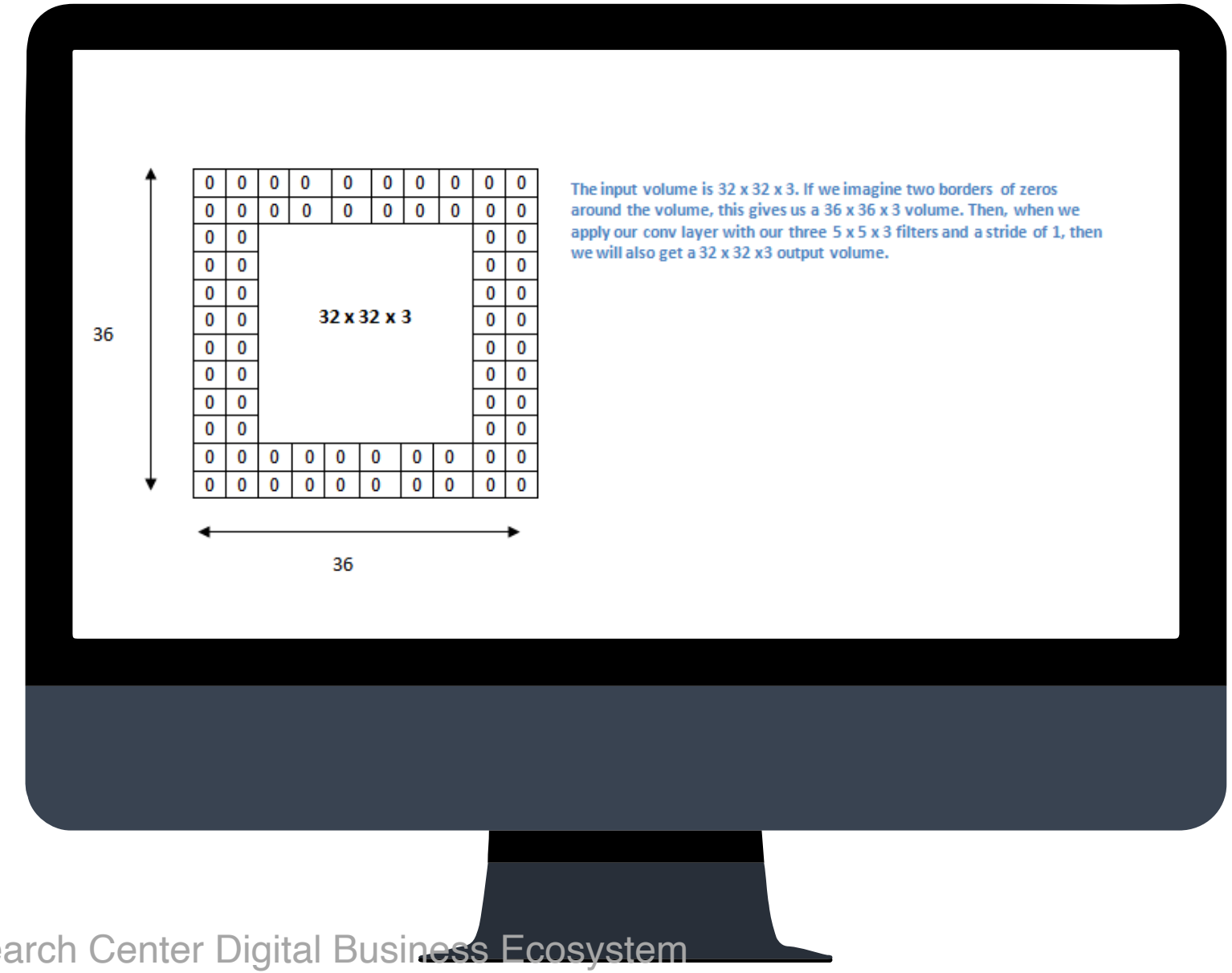


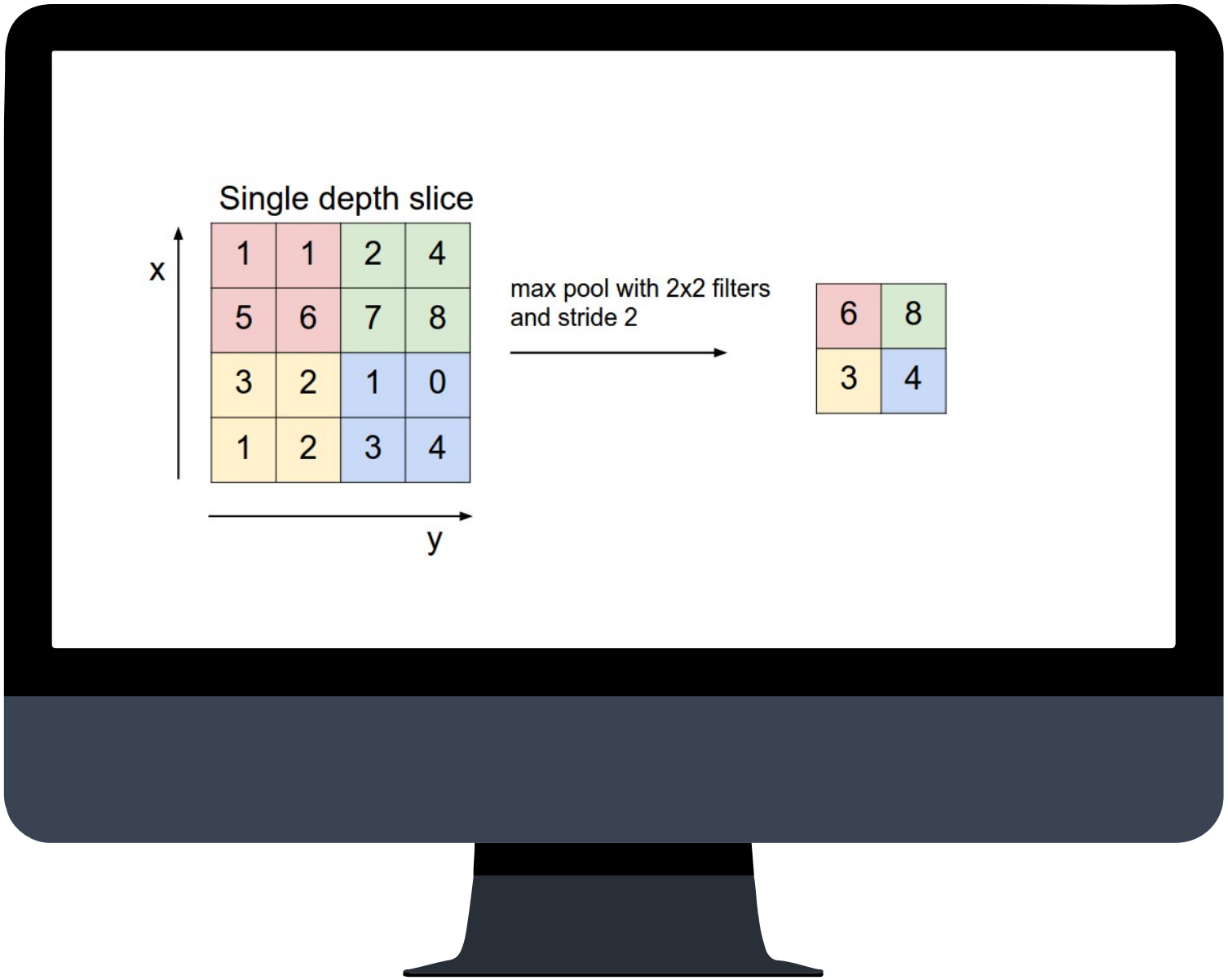
Stride

- Stride = how much we shift the filter at a time. The bigger the stride, the smaller the output.
- Formula to calculate output width and height = (W−F)/S+1, where W = input size (width or height), F = filter size (width or height), and S = stride
- Stride is normally set in a way so that the output volume is an integer and not a fraction.
- For stride = 3, the output will be (32-5)/3+1 = 10x10
- Example for 7x7 input with 3x3 filter :

Padding

- Used to preserve the size of the output.
- The formula : (W − F + 2P) / S + 1, where P = padding
- If we want to preserve the output spatial size to 32x32 with our 5x5 filter, we can use padding=2 and stride=1, so output = (32 - 5 + 4) + 1 = 32



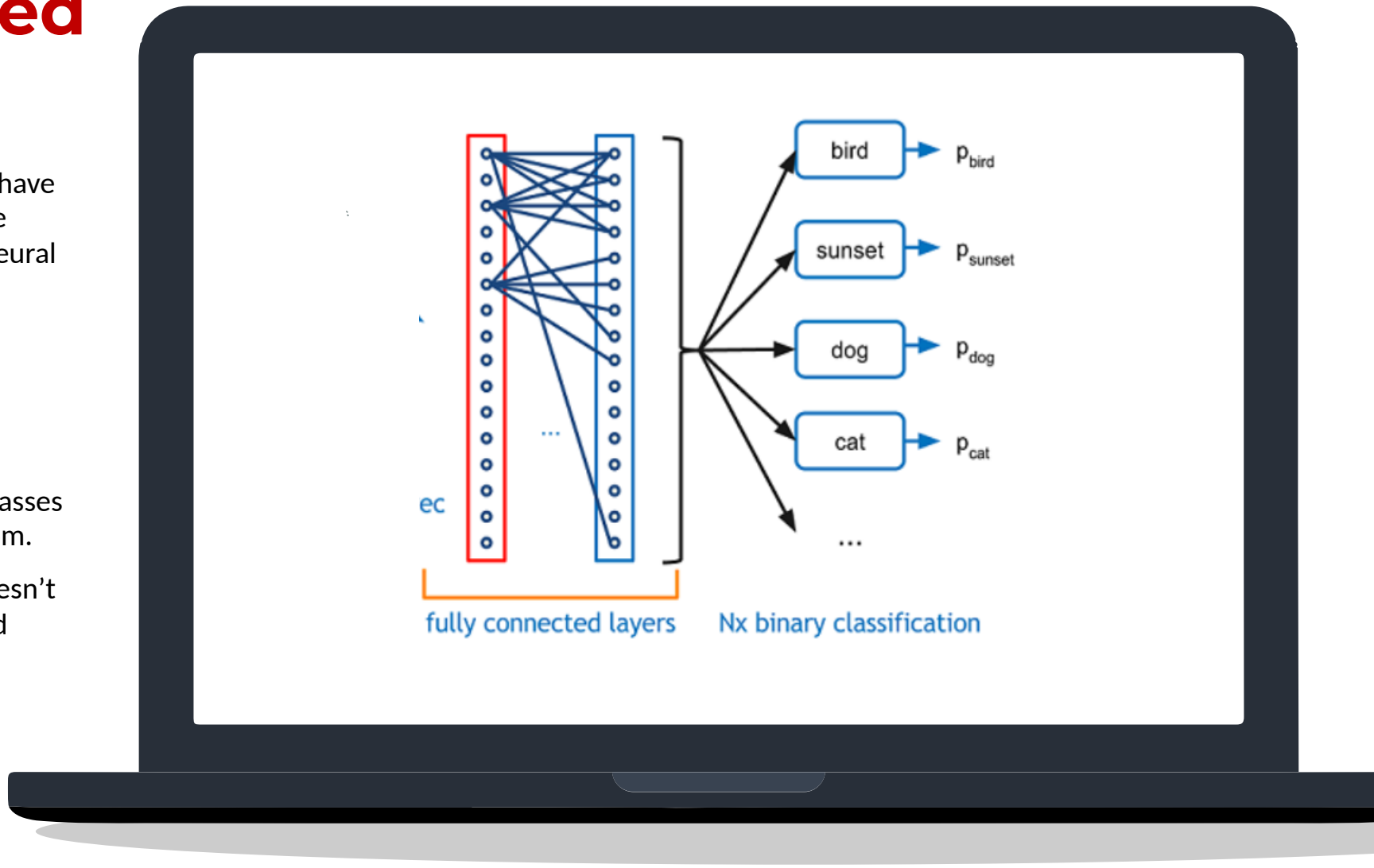


Pooling Layer

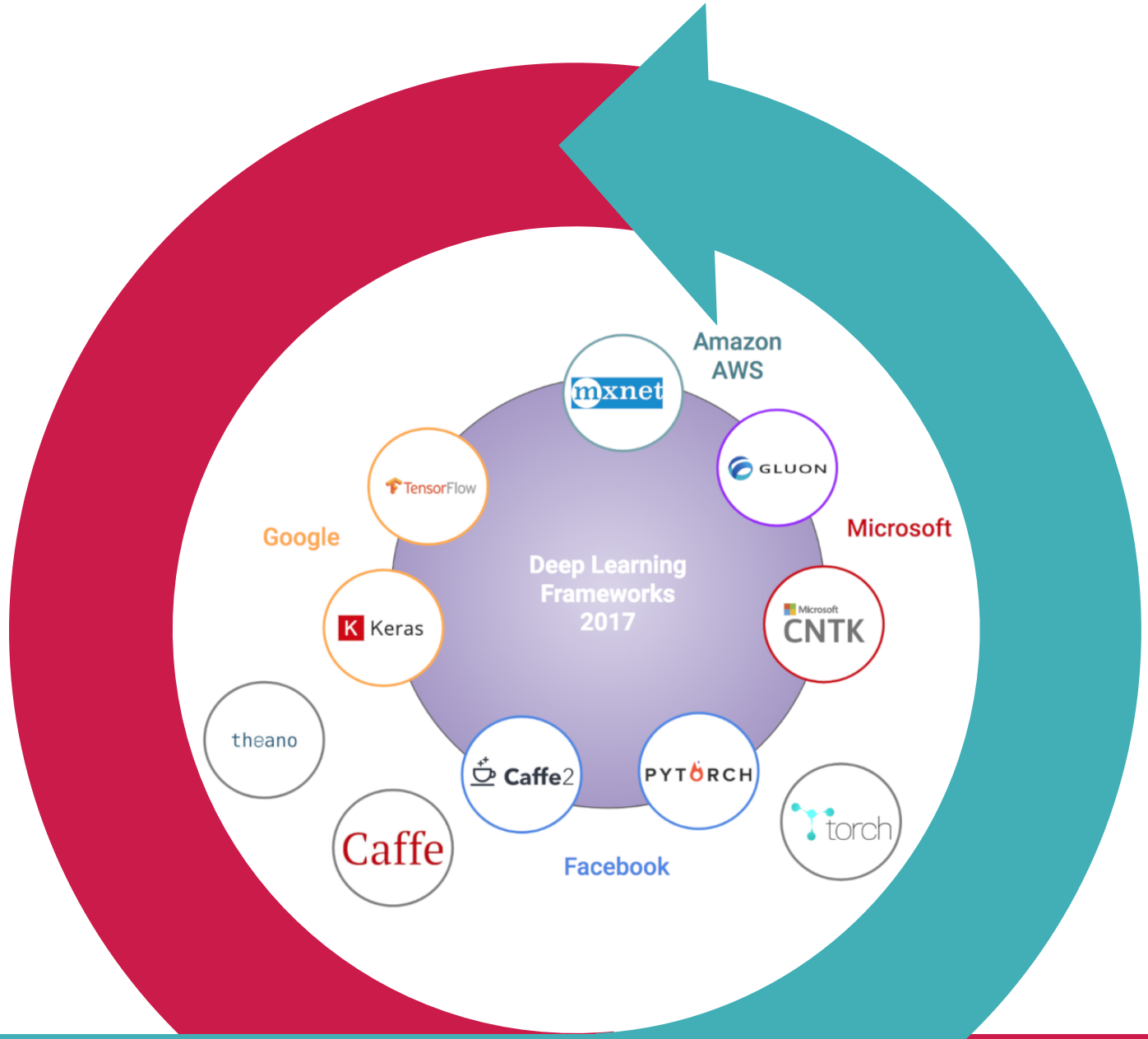
- Also called down sampling layer.
- Reduce the spatial size to reduce the amount of parameters and computation, and to control overfitting.
- It operates independently on every layer/ depth slice of the input → output depth = input depth
- The most common is : 2x2 with stride = 2. Other common setting is F=3x3 ,S=2 (overlapping pooling).
- Most common operation = max pooling. Other less common operation = average pooling

Fully Connected Layer

- Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks.
- Usually put at the end of the CNN architecture to perform the classification.
- FF layer outputs an N dimensional vector where N is the number of classes that the program has to choose from.
- Many newer CNN architectures doesn't use Pooling and/or Fully Connected Layer



Deep Learning Library



Use Case
Self-Driving Cars

01 Uber first announced its intentions to amass a fleet of automatic cars in February 2015

02 The cars themselves were packed with around 20 cameras, seven lasers, a GPS, radar and lidar, a technology that measures the distance reached by outgoing lasers so cars can “see” and interpret the action around them.

03 In March 2018, an Uber self-driving car in Tempe, Arizona struck a pedestrian who was walking outside of a crosswalk at night

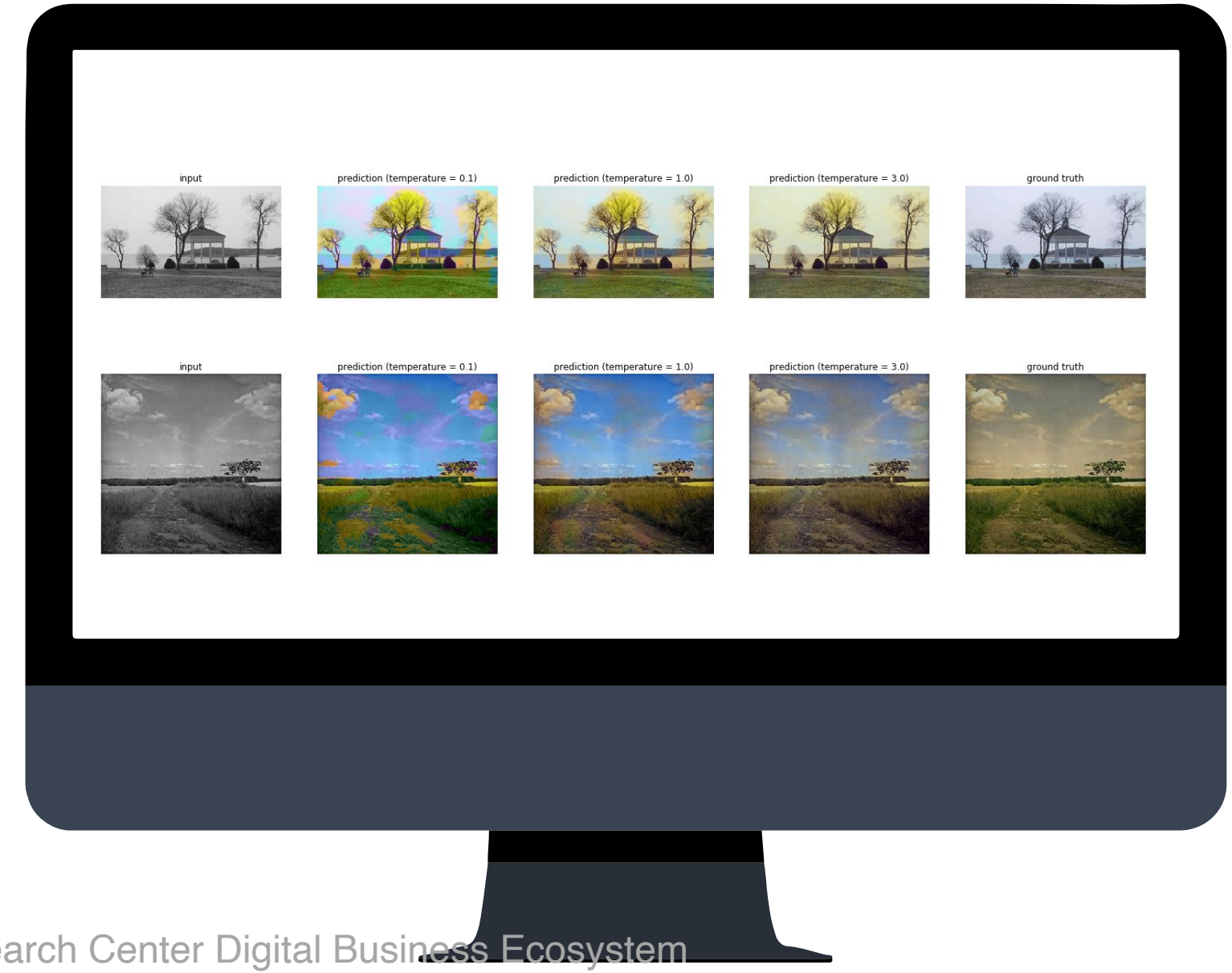
04 Waymo—formerly the Google self-driving car project- paid driverless taxi service could launch in December 2018

Use Case
Machine Translation

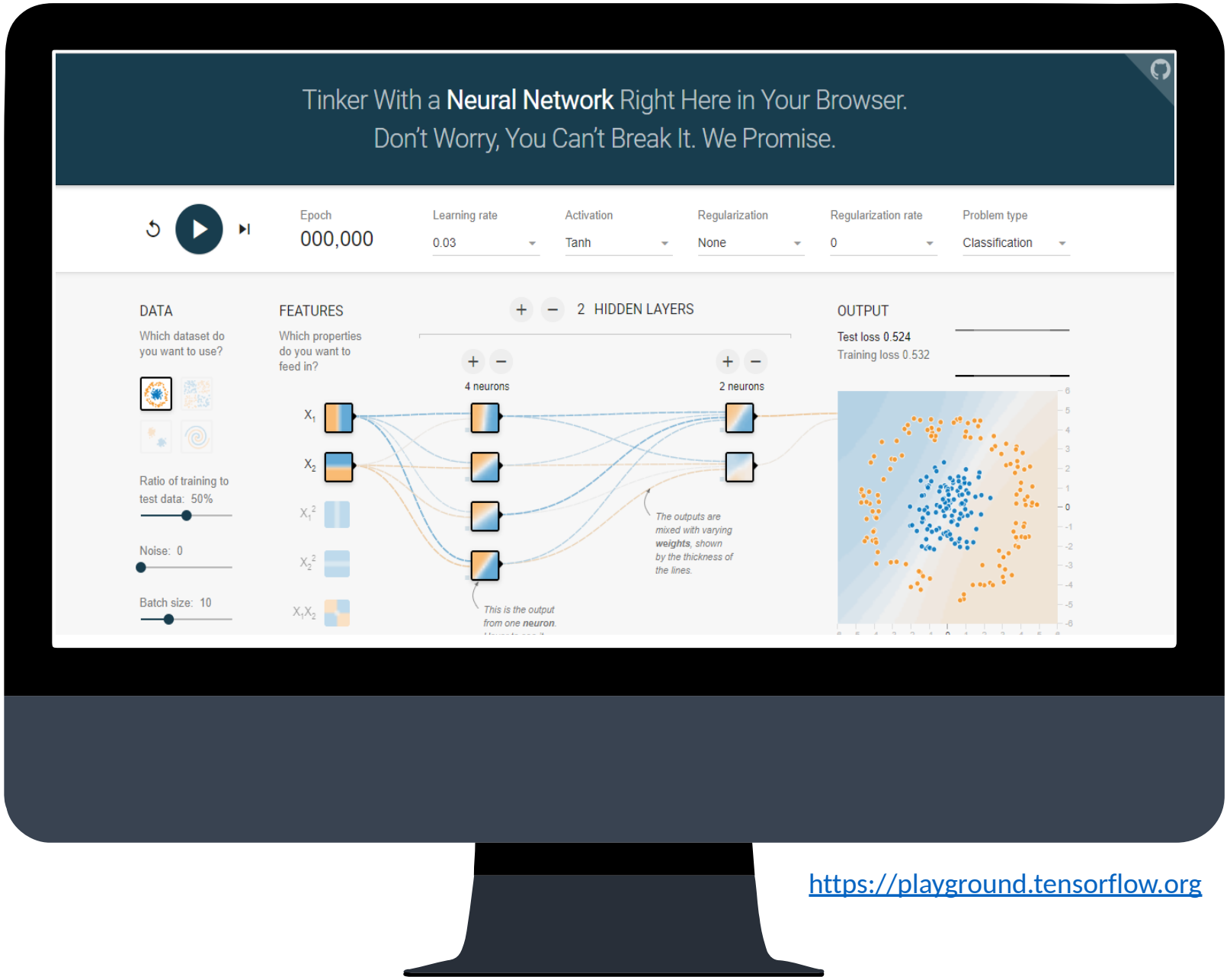
- DeepL Translator is a translation service launched in August 2017 by DeepL GmbH
- Promising to deliver 3x better results compared to Microsoft Translator and Google Translate

Usa Case Colorization Images

- Paper : ColorUNet: A Convolutional Classification Approach to Colorization
- A final project of Computer Vision courses at Stanford University
- Using Convolution Neural Network method classification to colorize grayscale images.



Tensorflow ANN



https://playground.tensorflow.org