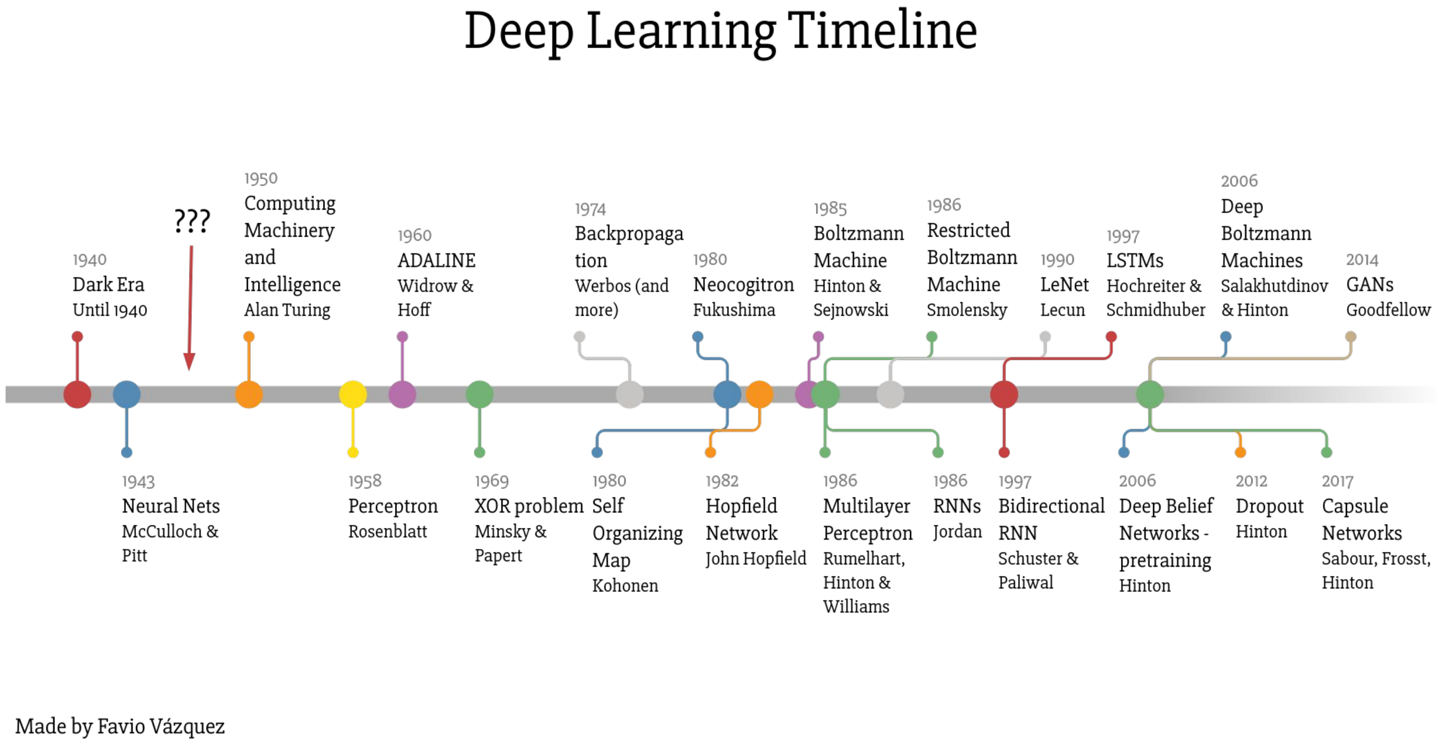
Neural Networks and Deep Learning

DL/NN is not New

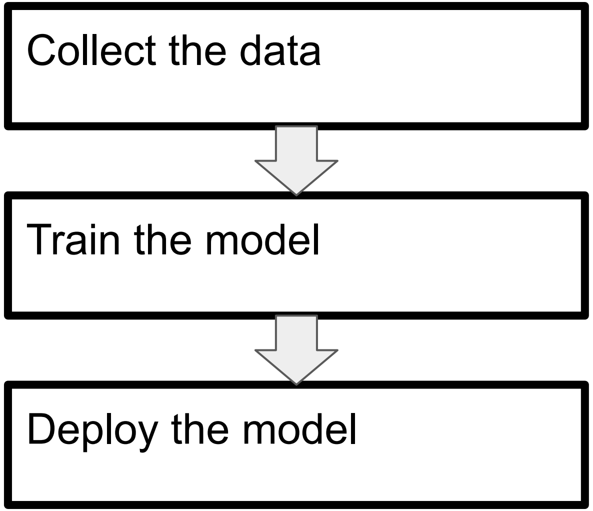


Why DL is Powerful Now?

- Feature engineering require high-level expert knowledge, which are easily over-specified and incomplete.
- Large amounts of training data
- Modern multi-core CPUs/GPUs/TPUs
- Better deep learning ‘tricks’ such as regularization, optimization, transfer learning etc.

Deep Learning Myth: Three Steps

- To deploy deep learning (or other machine learning) systems



The Truth is

1. Select a metric for optimization 💪
2. Collect data 🙇‍♂️
3. Train model 👩‍💻
4. Realize many labels are wrong 😱
5. Relabel data ✍️
6. Train model 👩‍💻
7. Model performs poorly on one class 🤦‍♀️
8. Collect more data for that class 🙇‍♂️
9. Train model 👩‍💻
10. Model performs poorly on most recent data 🤦‍♀️
11. Collect more recent data 🙇‍♂️
12. Train model 👩‍💻
13. Deploy model 👩‍🔧
14. Dream 🤑
15. Get a call at 3am about complaints that model is biased 👦 👱
16. Revert to the older version
17. Collect more data, do more training and testing👨‍💻
18. Deploy model 👩‍🔧
19. Pray 🙆‍♀️
20. Model performs well but revenue decreasing 🤷‍♀️
21. Cry 😭
22. Choose a different metric 💪
23. Start over 👨‍🦲

True three steps in Deep Learning

To approximate the true function, define a function **space**

Neural network structures

Learning Representation

Need a **measure** to evaluate the quality of each potential function in the previous space

Evaluate the function

Objective Function

Search the function space to find the best function based on the measure.

Pick the best function

Optimization

y=ax+b

function space

Find model parameters:

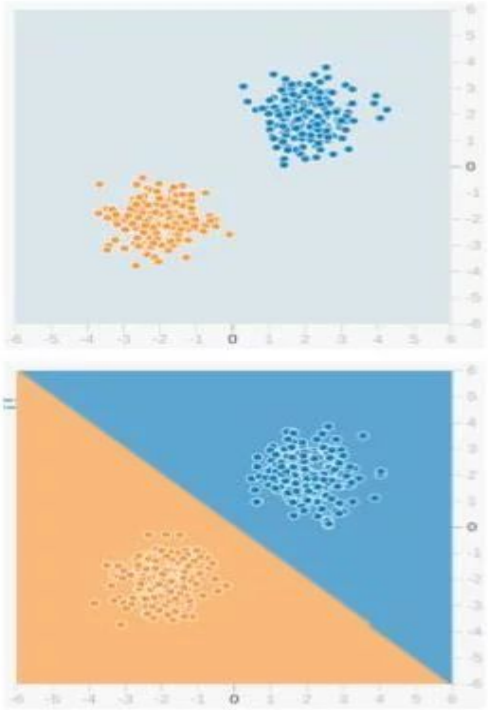
a=2 and b=1

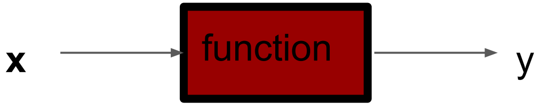
y=2x+1

a function

Neural Networks

A “Simple” Classification Problem





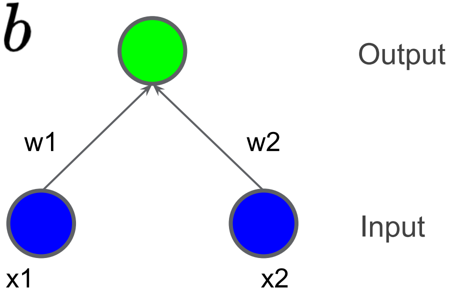
$$y = \mathbf{w}\mathbf{x} + b$$

A Linear Model

- Linear Regression if output is continuous
- Logistic Regression if output is discrete

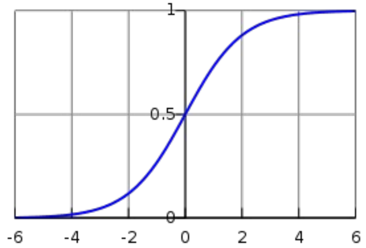
Linear Regression

$$y = w_1 x_1 + w_2 x_2 + b$$



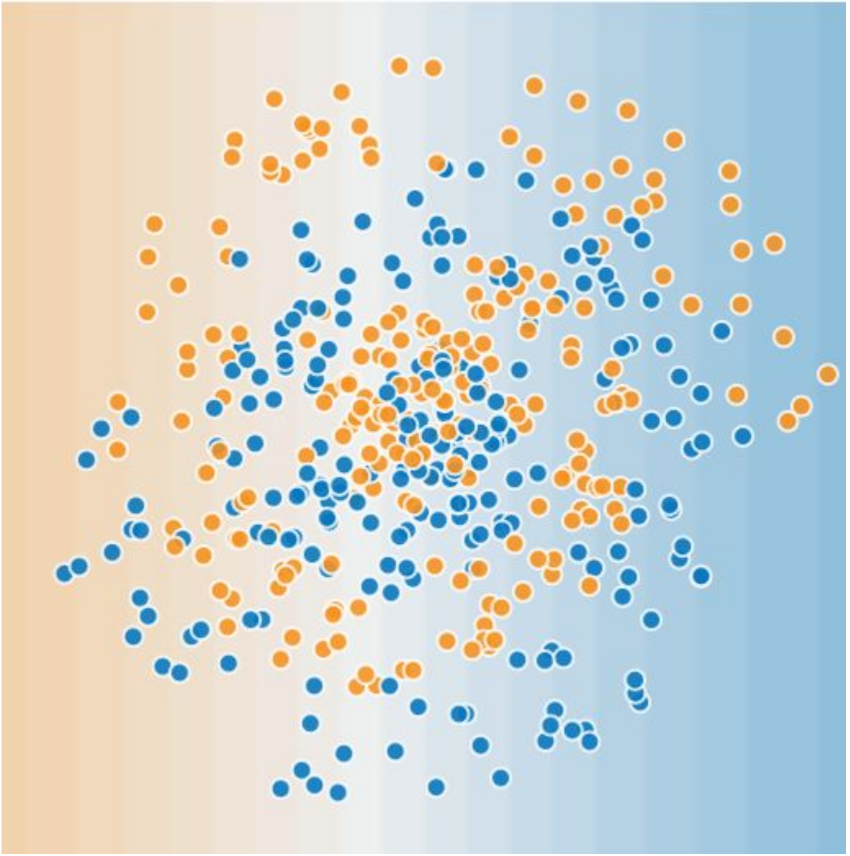
Logistic Regression

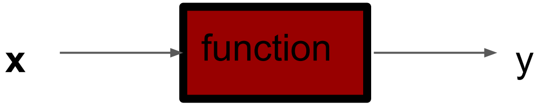
$$y = \sigma(\mathbf{wx} + b)$$



How about this classification problem?

Linear model can not solve the problem





We need **non-linear functions**

Add Complexity

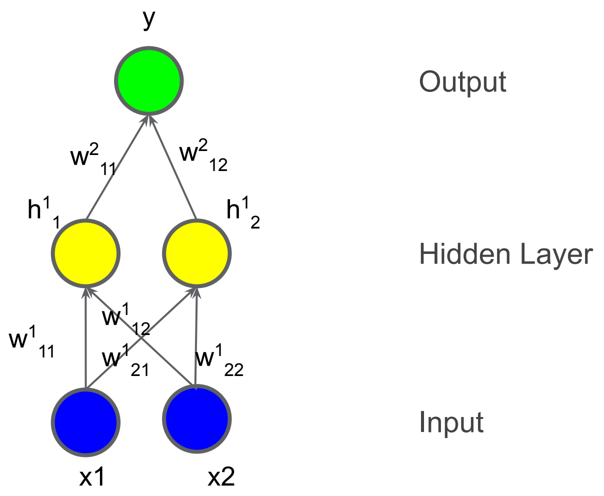
For Simplicity, the bias term is ignored here.

$$h^1_1 = w^1_{11} x1 + w^1_{12} x2$$

$$h^1_2 = w^1_{21} x1 + w^1_{22} x2$$

$$y = w^2_{11} h^1_1 + w^2_{12} h^1_2 \quad = \quad y = \mathbf{wx}$$

$$y = (w^2_{11} w^1_{11} + w^2_{21} w^1_{12}) x1 + (w^2_{12} w^1_{12} + w^2_{12} w^1_{22}) x2$$



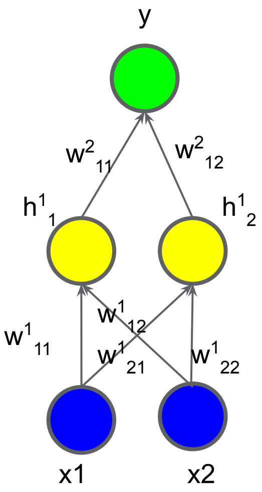
Add Complexity

Associative Law

$$h_1^1 = w_{11}^1 x1 + w_{12}^1 x2$$

$$h_2^1 = w_{21}^1 x1 + w_{22}^1 x2$$

$$y = w_{11}^2 h_1^1 + w_{12}^2 h_2^1$$



Output

Hidden Layer

Input

Matrix Format

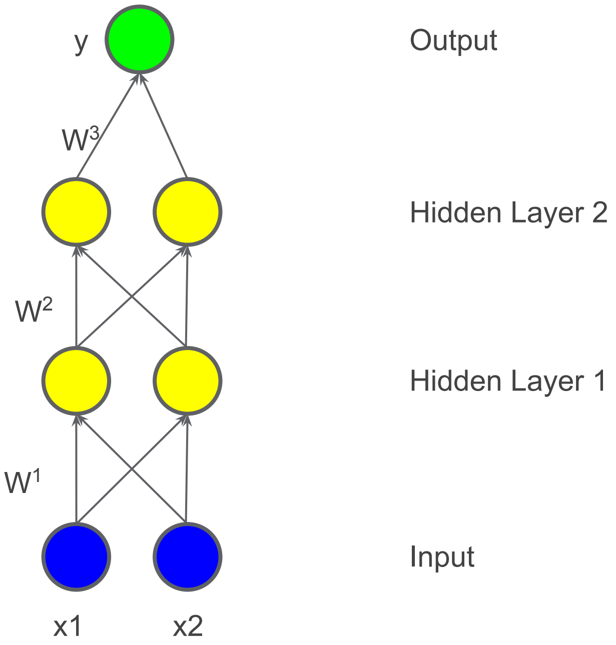
$$\begin{bmatrix} h_1^1 \\ h_2^1 \end{bmatrix} = \begin{bmatrix} w_{11}^1 & w_{12}^1 \\ w_{21}^1 & w_{22}^1 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \end{bmatrix}$$

$$y = \begin{bmatrix} w_{11}^2 & w_{12}^2 \end{bmatrix} \begin{bmatrix} h_1^1 \\ h_2^1 \end{bmatrix}$$

$$y = W^2 W^1 \begin{bmatrix} x1 \\ x2 \end{bmatrix} = (W^2 W^1) \begin{bmatrix} x1 \\ x2 \end{bmatrix} = W \begin{bmatrix} x1 \\ x2 \end{bmatrix}$$

How about now?

$$y = \mathbf{W}^3\mathbf{W}^2\mathbf{W}^1 \begin{bmatrix} x1 \\ x2 \end{bmatrix} = (\mathbf{W}^3\mathbf{W}^2\mathbf{W}^1) \begin{bmatrix} x1 \\ x2 \end{bmatrix}$$



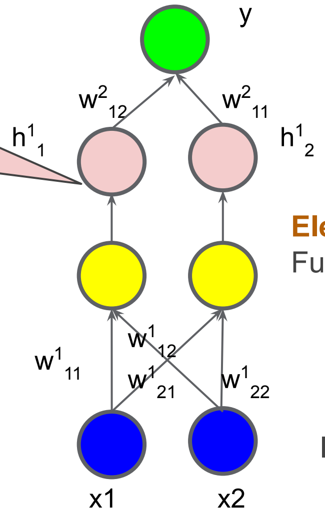
Make it non-linear

$$h_1^1 = \sigma(w_{11}^1 x1 + w_{12}^1 x2)$$

$$h_2^1 = \sigma(w_{21}^1 x1 + w_{22}^1 x2)$$

$$y = w_{11}^2 h_1^1 + w_{12}^2 h_2^1 \neq y = \mathbf{wx}$$

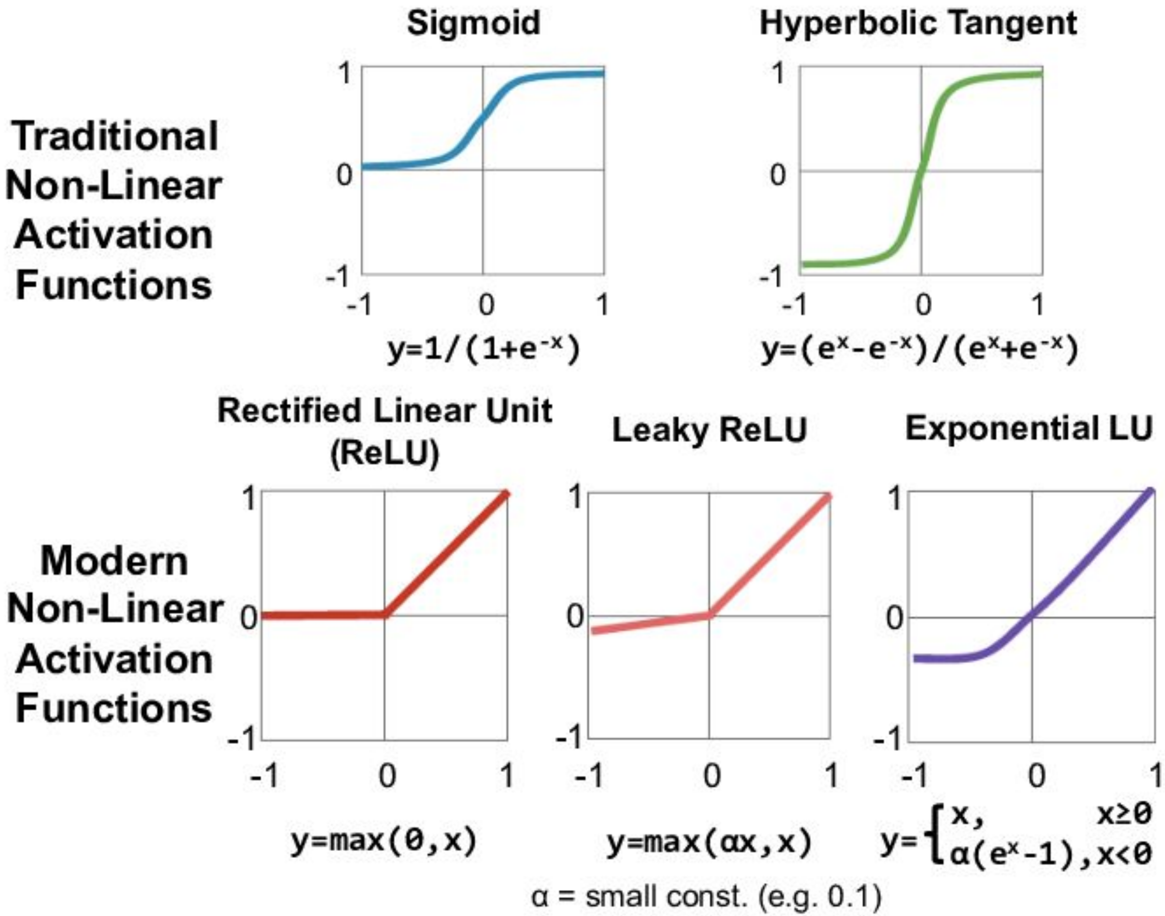
We Usually Don't Draw Non-Linear Transforms



Element-wise Non-linear Activation Function \sigma

Input

Non-linear Activation Functions



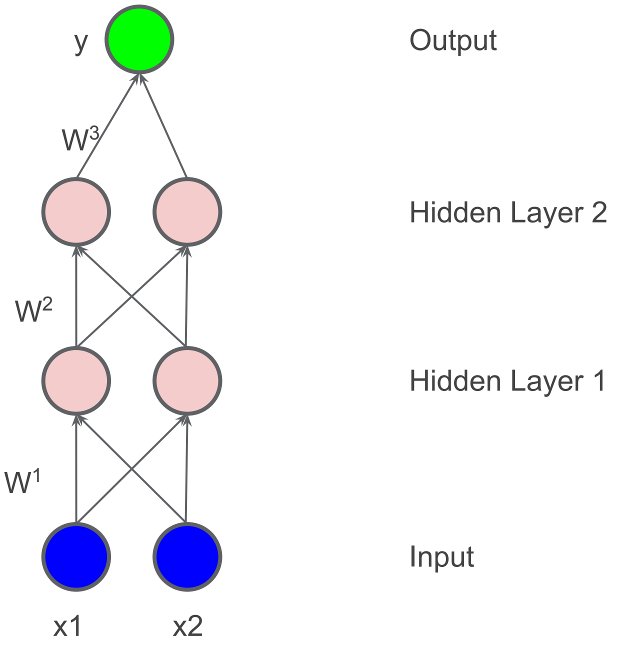
Add Non-linear Activation Function

$$y = \mathbf{W}^3 \sigma(\mathbf{W}^2 \sigma(\mathbf{W}^1 \mathbf{x}))$$

$$y = \mathbf{W}^3 \mathbf{h}^2$$

$$\mathbf{h}^2 = \sigma(\mathbf{W}^2 \mathbf{h}^1)$$

$$\mathbf{h}^1 = \sigma(\mathbf{W}^1 \mathbf{x})$$



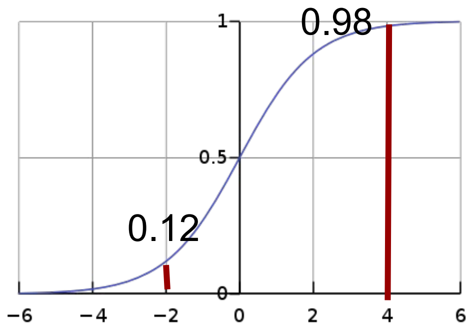
Why Non-linear Activation

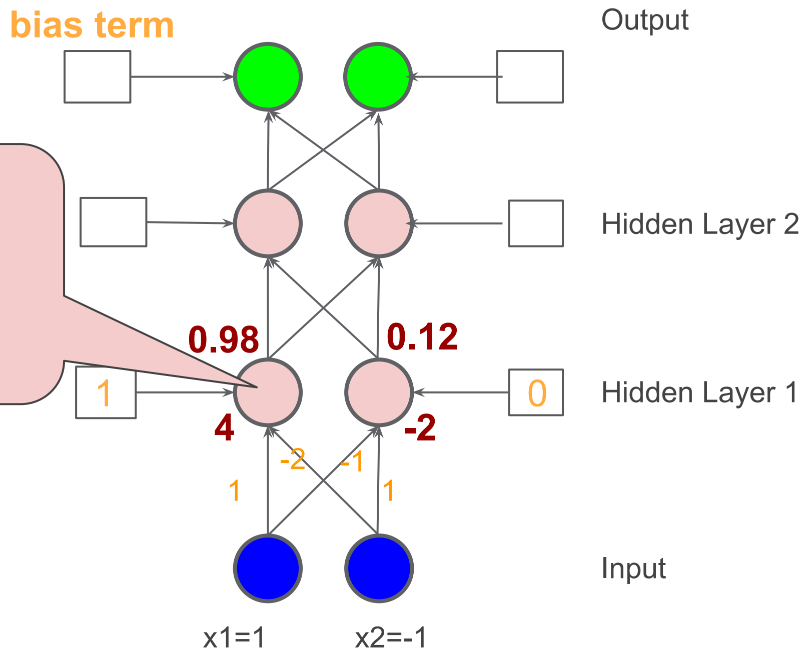
- The non-linearities activation function increases the capacity of model
- Without non-linearities, deep neural networks is meaningless: each extra layer is just one linear transform.
- How to select activation functions?
  You can select an activation function which will approximate the distribution faster leading to faster training process.

Forward Computation

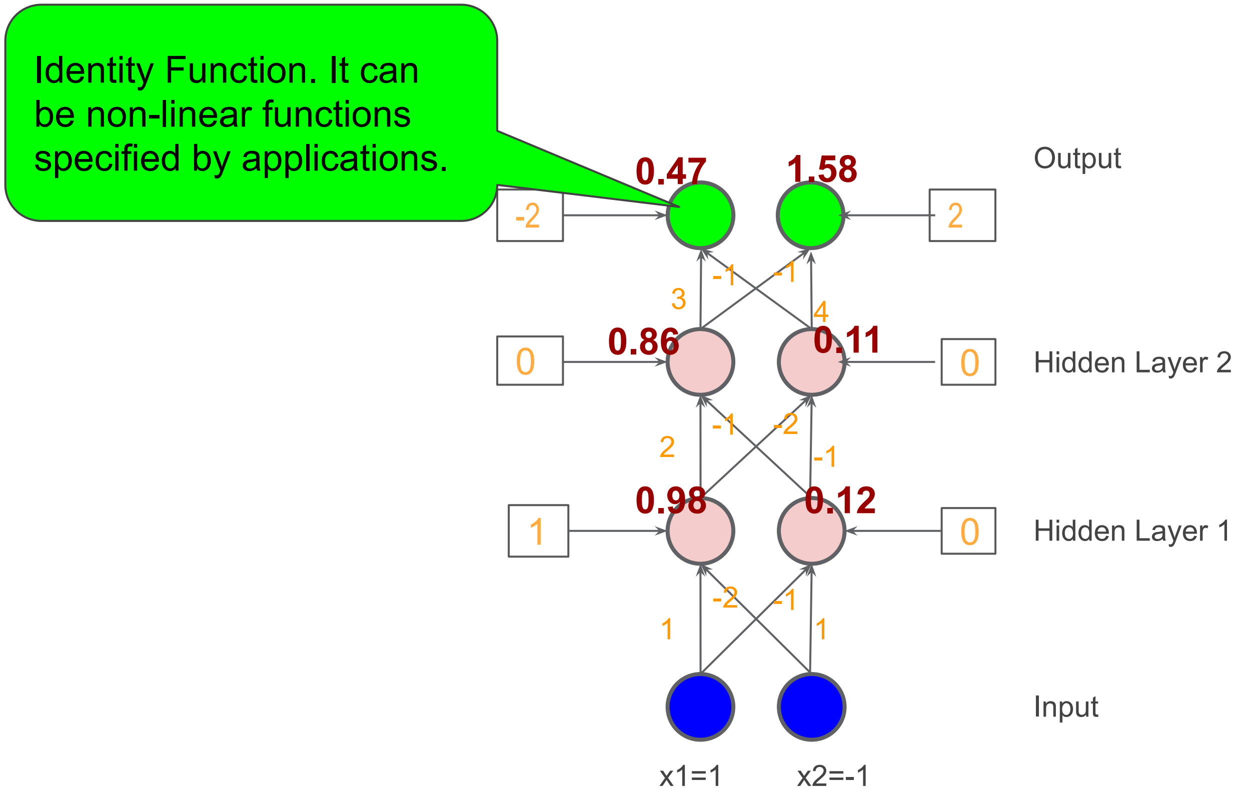
Sigmoid function:

$$h = \sigma(z) = \frac{1}{1+e^{-z}}$$





Forward Computation



Forward Computation

1. Neural Network acts as a function that transforms the input vector into the output vector (target)

$$\begin{bmatrix} 0.47 \\ 1.58 \end{bmatrix} = f_\theta(\begin{bmatrix} 1 \\ -1 \end{bmatrix})$$

$$\begin{bmatrix} 0.04 \\ 1.7 \end{bmatrix} = f_\theta(\begin{bmatrix} 0 \\ 0 \end{bmatrix})$$

$$\mathbf{W}^1 = \begin{bmatrix} 1 & -2 \\ -1 & 1 \end{bmatrix} \quad \mathbf{b}^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{W}^2 = \begin{bmatrix} 2 & -1 \\ -2 & -1 \end{bmatrix} \quad \mathbf{b}^2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

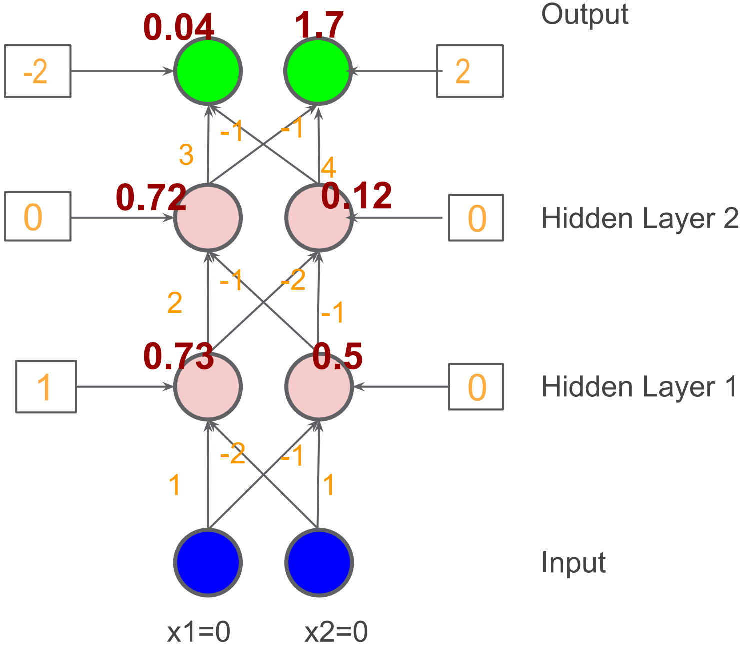
$$\mathbf{W}^3 = \begin{bmatrix} 3 & -1 \\ -1 & 4 \end{bmatrix} \quad \mathbf{b}^3 = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

one param. set

2. It is actually a function space parameterized by weights matrices and bias vectors.

$$\theta = \{\mathbf{W}^1, \mathbf{b}^1, \mathbf{W}^2, \mathbf{b}^2, \mathbf{W}^3, \mathbf{b}^3\}$$

Parameter space



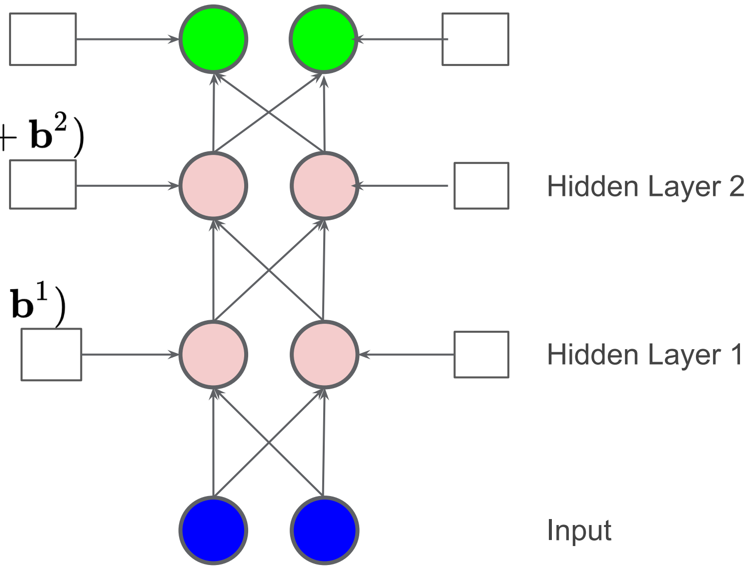
Forward Computation

$$y = \mathbf{W}^3\mathbf{h}^2 + \mathbf{b}^3$$

$$\mathbf{h}^2 = \sigma(\mathbf{W}^2\mathbf{h}^1 + \mathbf{b}^2)$$

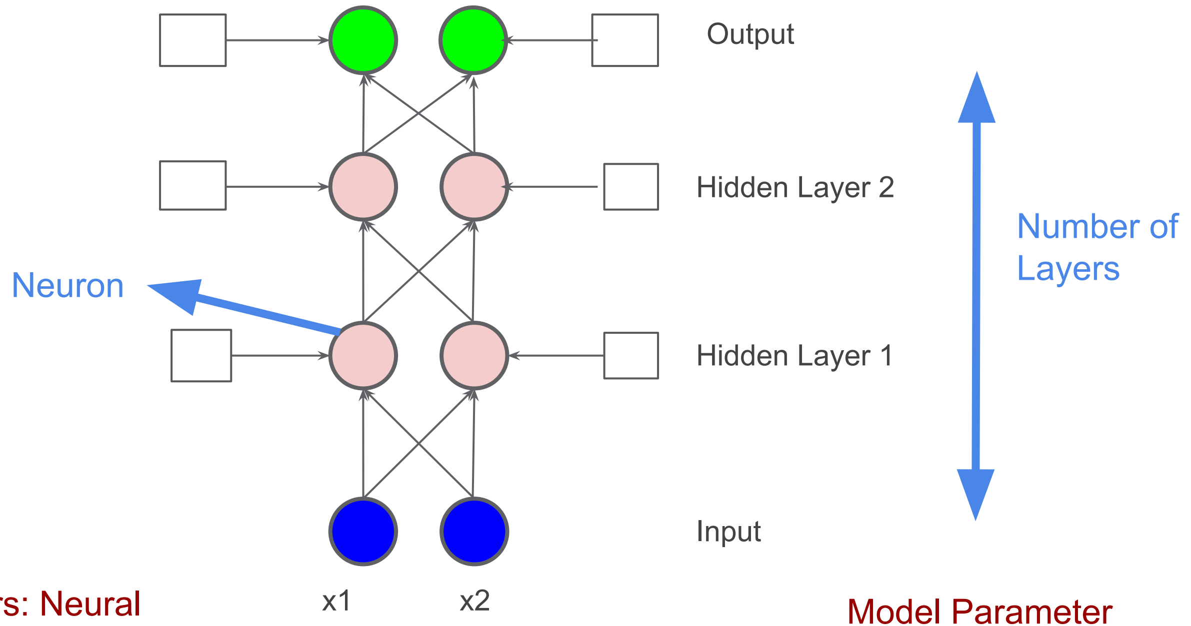
$$\mathbf{h}^1 = \sigma(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1)$$

$$y = \mathbf{W}^3\sigma(\mathbf{W}^2\sigma(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) + \mathbf{b}^2) + \mathbf{b}^3$$



1. Neural Network is a model that **recursively** applies the matrix multiplication and non-linear activation function.
2. Parallel computing techniques can be used to speed up matrix operation.

Neural network: Function Set



Hyperparameters: Neural Network Structure

Model Parameter

1.Number of Layers
2.Number of neurons in each layer
3.Non-linear Activation function in each layer

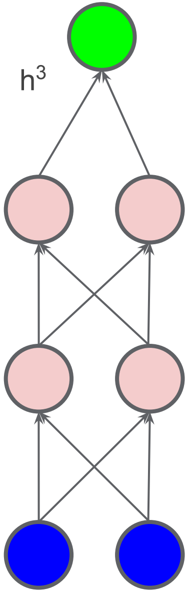
→ **A function space containing various functions**

→ **A specific function mapping from input data to targets.**

Output Layer

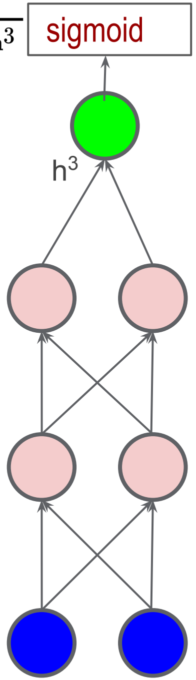
Regression

$$\tilde{y} = \mathbf{h}^3$$



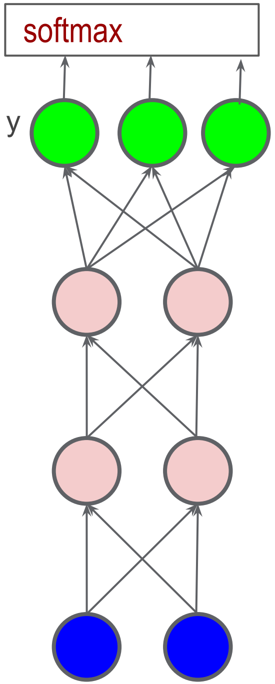
Binary Classification

$$\tilde{y} = \frac{1}{1+e^{-\mathbf{h}^3}}$$



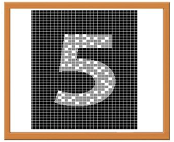
Multi-label Classification

$$\tilde{\mathbf{y}} = [\frac{e^{h_1^3}}{e^{h_1^3}+e^{h_2^3}+e^{h_3^3}}, \frac{e^{h_2^3}}{e^{h_1^3}+e^{h_2^3}+e^{h_3^3}}, \frac{e^{h_3^3}}{e^{h_1^3}+e^{h_2^3}+e^{h_3^3}}]$$



$$\tilde{y}_i = \frac{e^{h_i^3}}{\sum_{j=1}^{K} e^{h_j^3}}$$

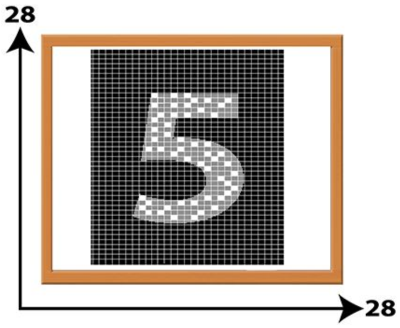
Example: MINST Dataset



Machine Learning Model

Output: 5

Input Vector



28*28=784

Pixel
Pixel
Pixel
Pixel
Pixel
⋮
Pixel

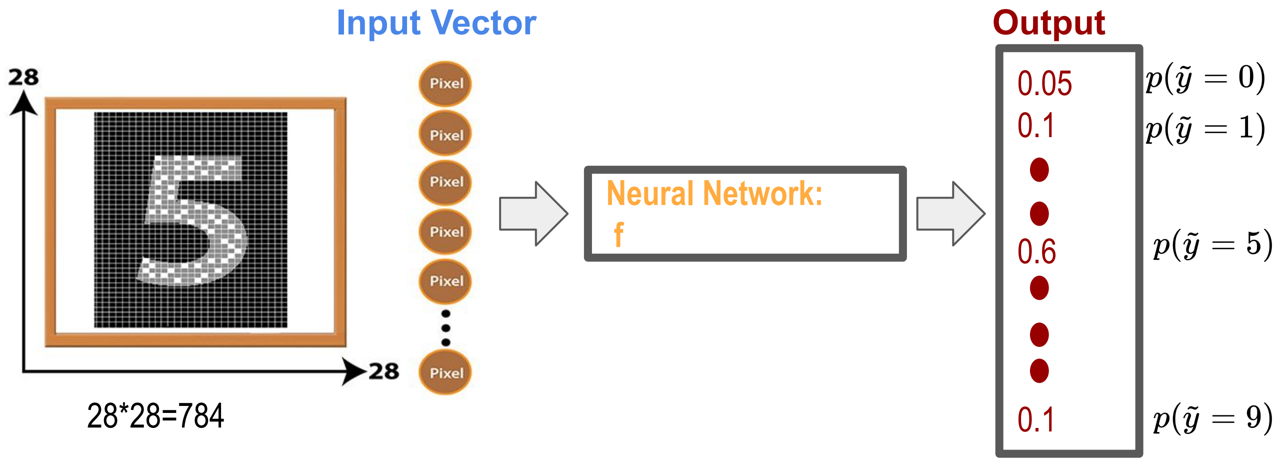
Neural Network: f

Output

0.05	$p(\tilde{y} = 0)$
0.1	$p(\tilde{y} = 1)$
●	
●	
0.6	$p(\tilde{y} = 5)$
●	
●	
●	
0.1	$p(\tilde{y} = 9)$

The probability distribution over the ten digits

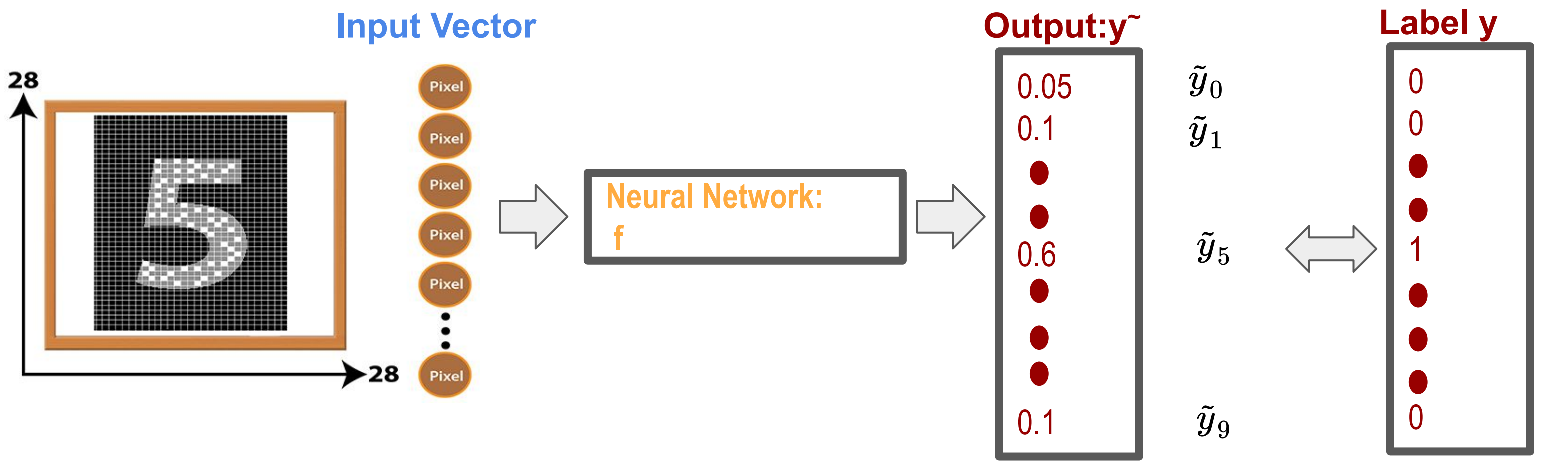
Example: MINST Dataset



1. In this task, the neural network is a function mapping from the input 784-dim vector to the output 10-dim vector.
2. The neural network structure should be decided to make sure the best function exists in the function set.

Evaluation of Functions

Cross-Entropy Loss



Given a set of parameters and one training sample,

$$loss(\tilde{\mathbf{y}}, \mathbf{y}) = -\sum_{i=0}^{9} y_i ln(\tilde{y}_i)$$

Total Loss

1. Training dataset contains N training samples
2. The total loss is:

$$J = \sum_{n=1}^{N} loss(\tilde{\mathbf{y}_n}, \mathbf{y}_n)$$

3. Find a function in the function set that minimizes the total loss J
4. Find the network parameters θ that minimizes the total loss J

In E, training data are fixed and model parameters are unknown.

$$argmin_{\theta} J$$

Optimization

Gradient Descent

Gradient for the total loss function over parameters, which computed by Backpropagation algorithm

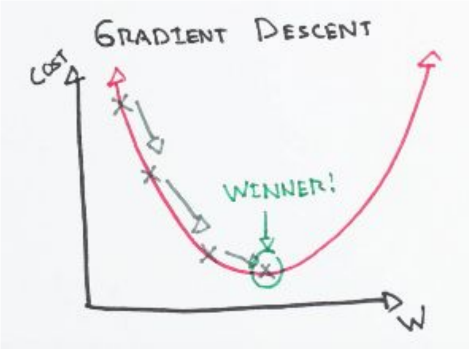
$$\theta_{t+1} = \theta_t - \alpha \boxed{\nabla J(\theta_t)}$$

New Parameters Guess

Current Parameters Guess

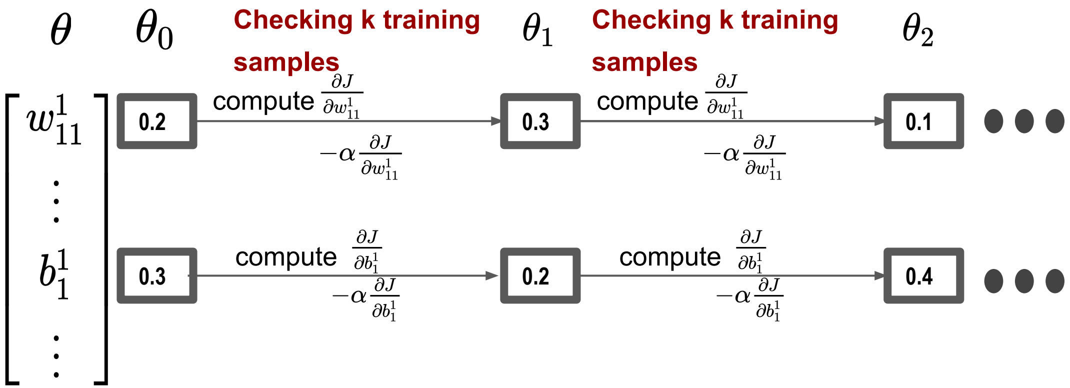
Learning Rate

Like hiking down a mountain



Credit:https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html

Gradient Descent



Backpropagation is used to compute gradients in an efficient way. $$\frac{\partial J}{\partial \theta}$$

Batch size: **k**

A dataset is [1,2,3,4,5,6] and the batch size is 2, one batch shuffle could be:
batch0=[2,1], batch1=[3,6], batch2=[4,5]

Backpropagation

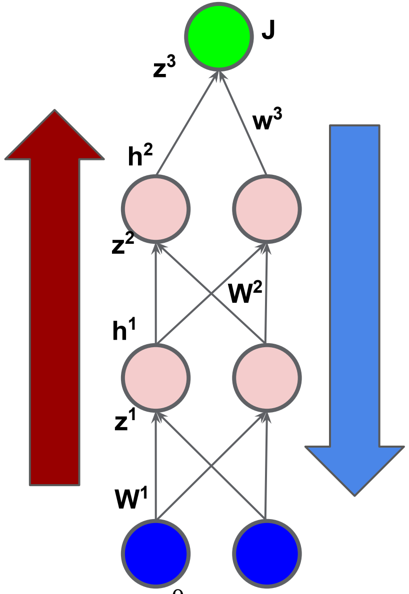
Definition (from wiki):

By computing the gradient of the loss function with respect to each weight by the **chain rule**, computing the gradient one layer at a time, iterating backward from the last layer to avoid redundant calculations of intermediate terms in the chain rule

Predictions flows forward

$$h^l = \sigma(\mathbf{z}^l)$$

$$\mathbf{z}^l = \mathbf{W}^l h^{l-1} + \mathbf{b}^l$$



$$\mathbf{x} = \mathbf{h}^0$$

Error flows backward

$$\frac{\partial J}{\partial \mathbf{w}^3} = \frac{\partial J}{\partial \mathbf{z}^3}\frac{\partial \mathbf{z}^3}{\partial \mathbf{w}^3}$$

$$\frac{\partial J}{\partial \mathbf{W}^2} = \underbrace{\frac{\partial J}{\partial \mathbf{z}^3}}_{\text{From } \mathbf{w}^3}\frac{\partial \mathbf{z}^3}{\partial \mathbf{h}^2}\frac{\partial \mathbf{h}^2}{\partial \mathbf{z}^2}\frac{\partial \mathbf{z}^2}{\partial \mathbf{W}^2}$$

$$\frac{\partial J}{\partial \mathbf{W}^1} = \underbrace{\frac{\partial J}{\partial \mathbf{z}^3}}_{\text{From } \mathbf{w}^3}\underbrace{\frac{\partial \mathbf{z}^3}{\partial \mathbf{h}^2}\frac{\partial \mathbf{h}^2}{\partial \mathbf{z}^2}}_{\text{From } \mathbf{W}^2}\frac{\partial \mathbf{z}^2}{\partial \mathbf{h}^1}\frac{\partial \mathbf{h}^1}{\partial \mathbf{z}^1}\frac{\partial \mathbf{z}^1}{\partial \mathbf{W}^1}$$

More details:
https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/

Batch Size

Three approaches to select batch sizes:

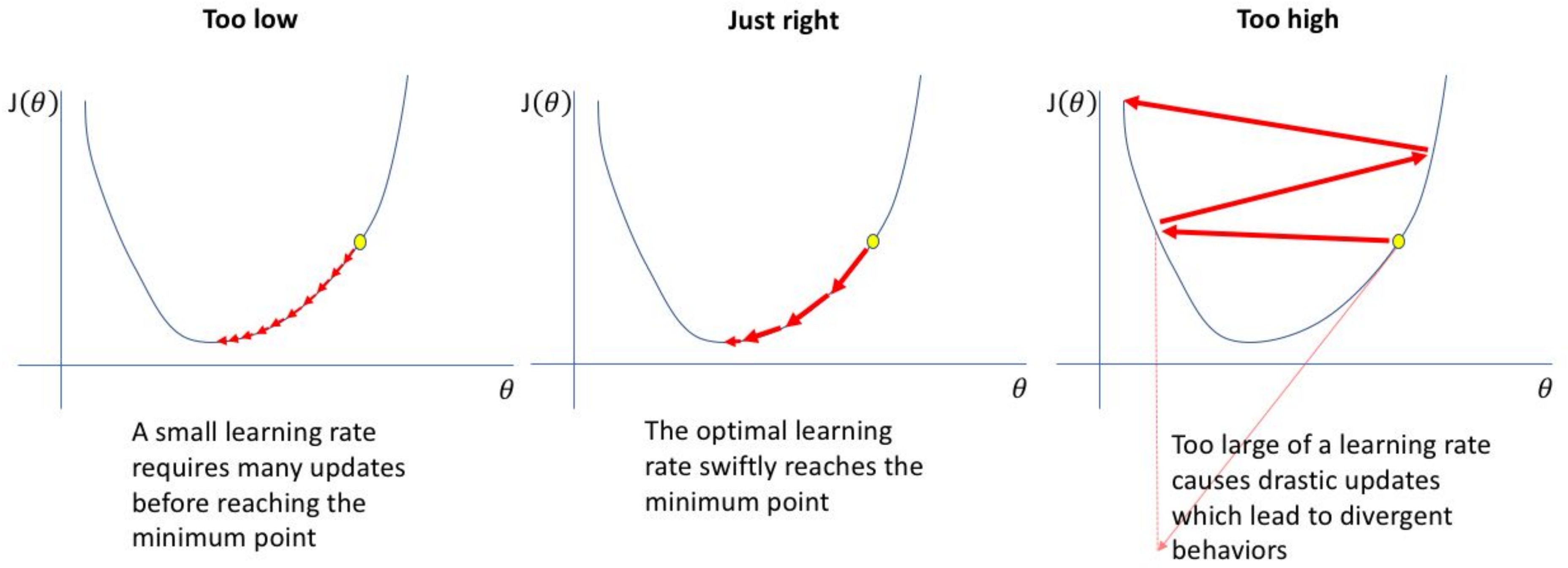
1. Batch Gradient Descent — ***batch size*** = *Number of training data*
2. Mini-batch Gradient Descent — 1<***batch size***< *number of training data*
3. Stochastic Gradient Descent — ***batch size*** = *1*

Training Process

1. Initialize neural network randomly
2. For _ in range(number of epoch):
    - Shuffle all the training dataset into a list of batches
    - For _ in range(number of batches)
        - Get output with the input data in the batch
        - Compare outputs with ground truth in training data
        - Compute loss function with the batch data
        - Update weights with backpropagation and gradient descent algorithm

Iteratively perform

How to find learning rate?



https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/

Except SGD

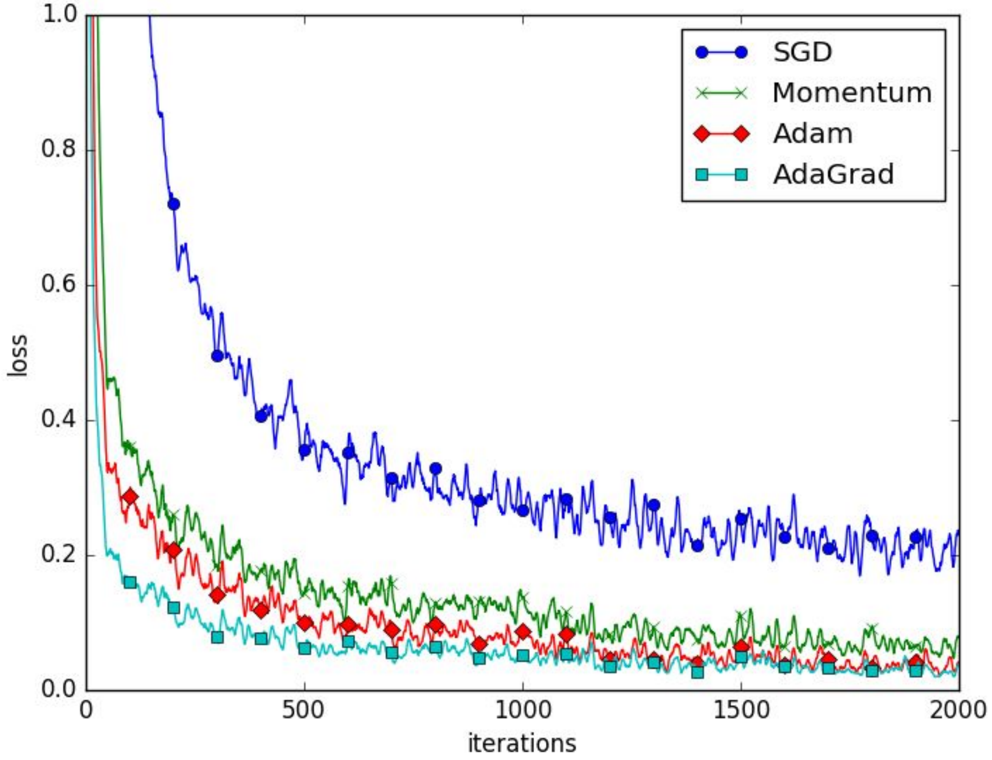
SGD

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha \bigtriangledown f(\mathbf{x}_n)$$

Different Variants

Momentum, Adam, AdaGrad, RMSProp

Auto-tune learning rates



Neural Network Visualization

Playground

Deep Learning/Deep Neural Networks

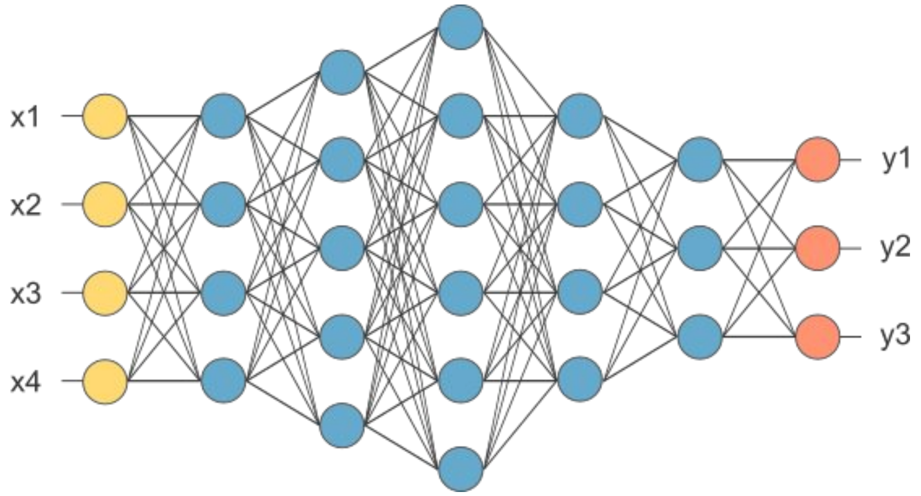
Neural Network

1. From Wiki:

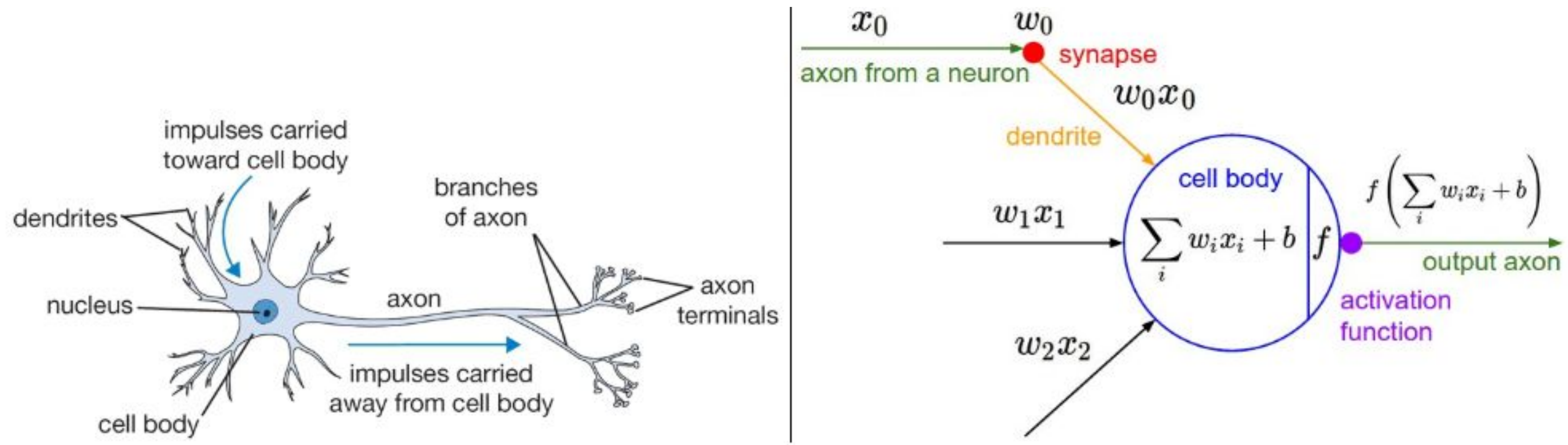
- NN is based on a collection of connected units of nodes called artificial **neurons** which loosely model the neurons in a biological brain.

2. From another way:

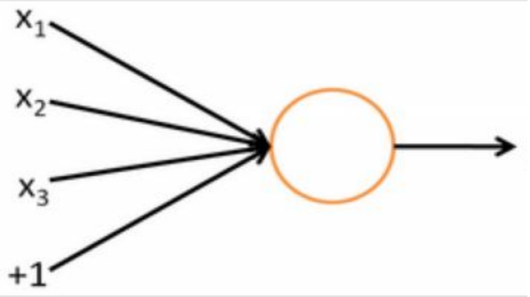
- NN is running several ‘logistic regression’ at the same time (expanding at width and depth dimensions).



Neural Computation



A cartoon drawing of a biological neuron (left) and its mathematical model (right).

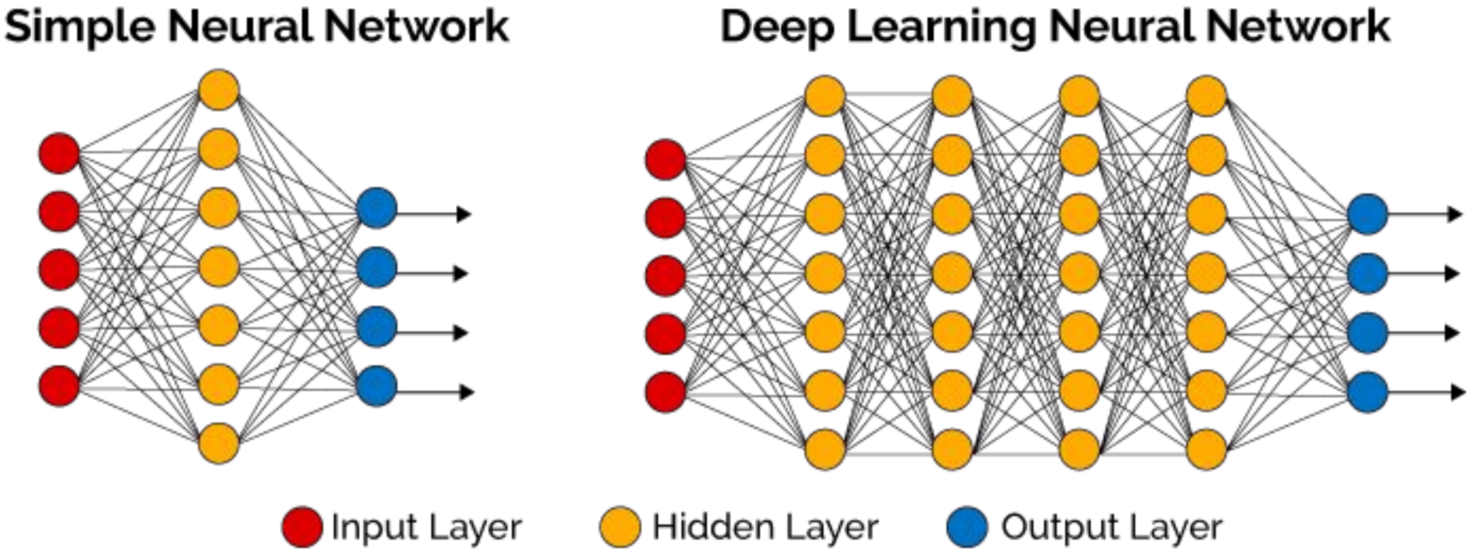


The fact that a neuron is essentially a logistic regression unit:

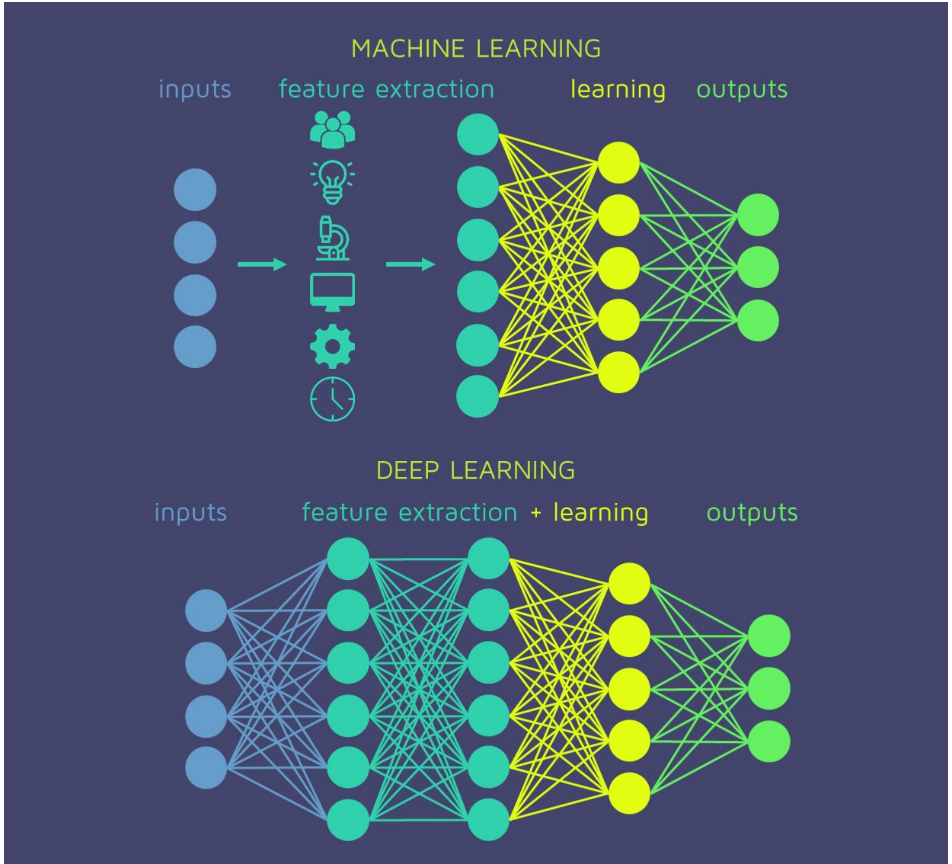
1 performs a dot product with the input and its weights

2 adds the bias and apply the non-linearity

Shallow vs Deep

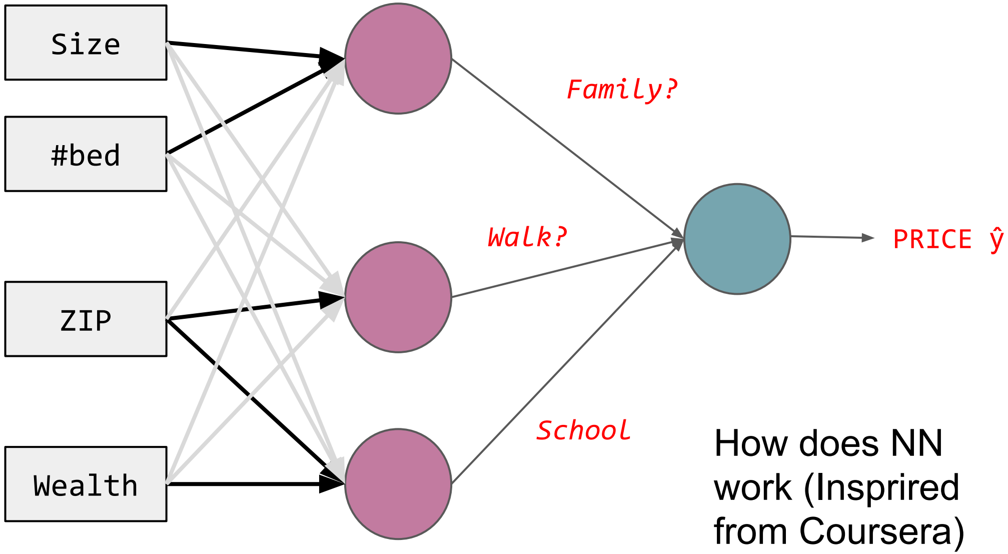


End-to-End Learning



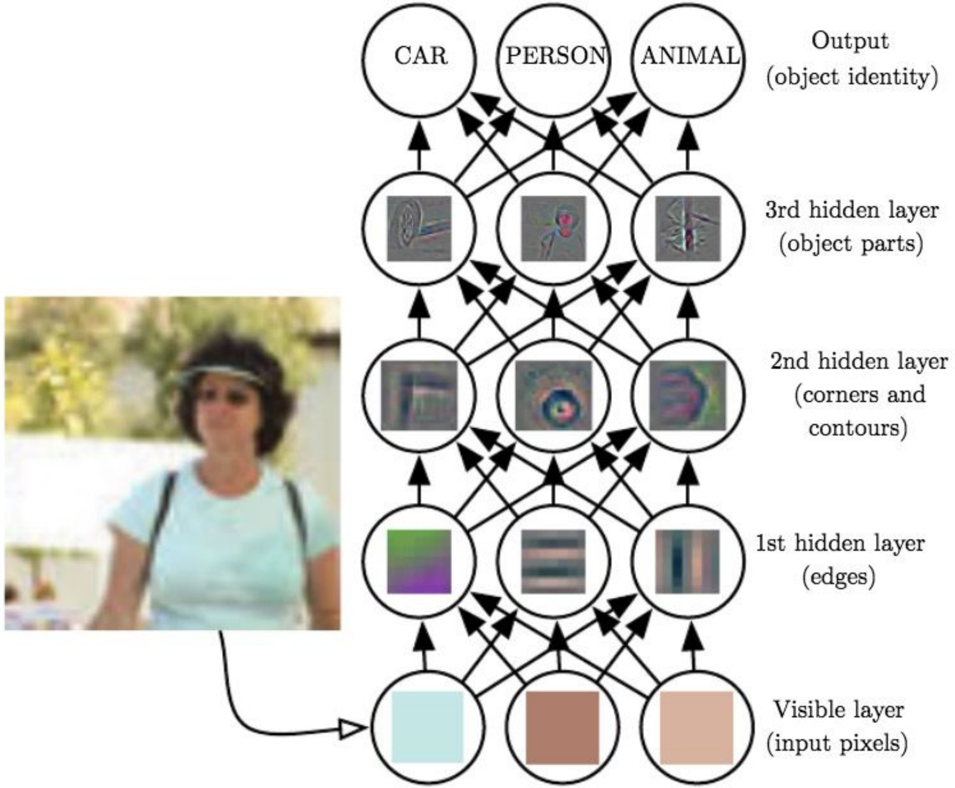
From Aporras

Representation Learning in DL



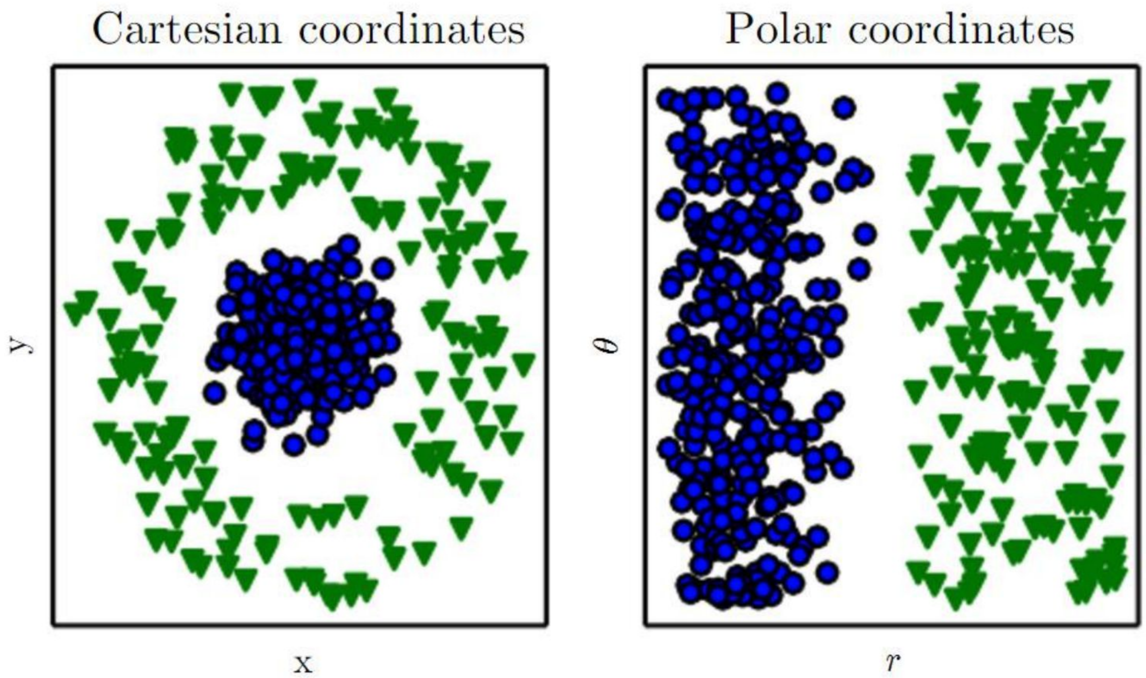
How does NN work (Insprired from Coursera)

Representation Learning in DL

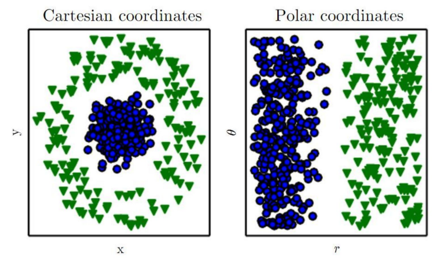


From Deep Learning (Goodfellow)

Representation Matters

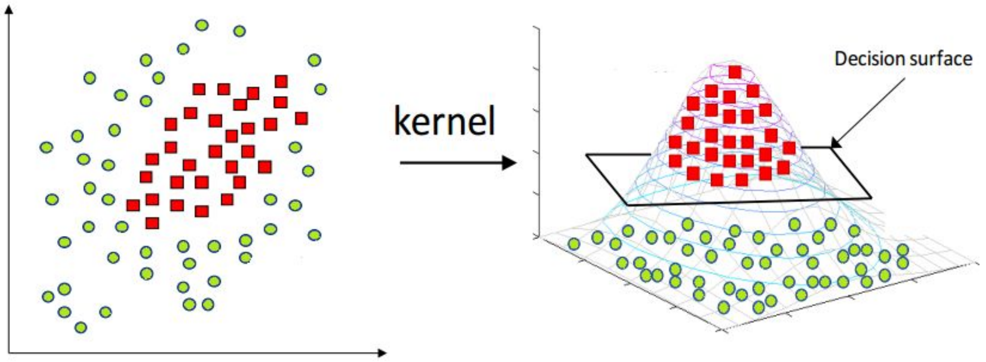


Task: Draw a line to separate the **green triangles** and **blue circles**.



We want to project the data into the new feature/vector space that data is linearly separated

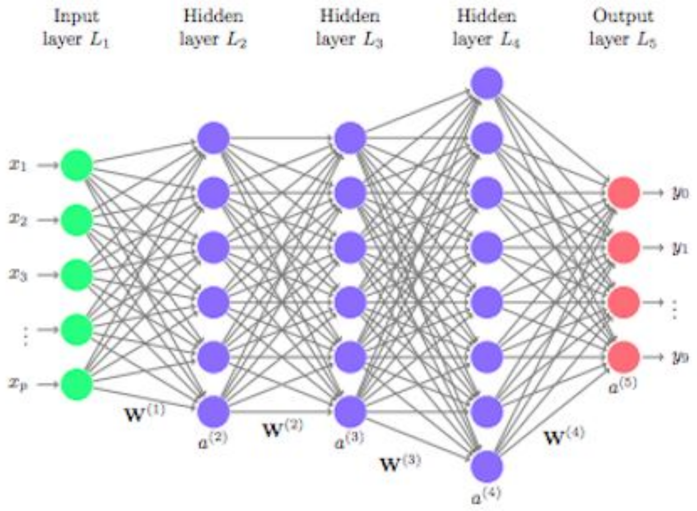
Kernel Tricks in SVM



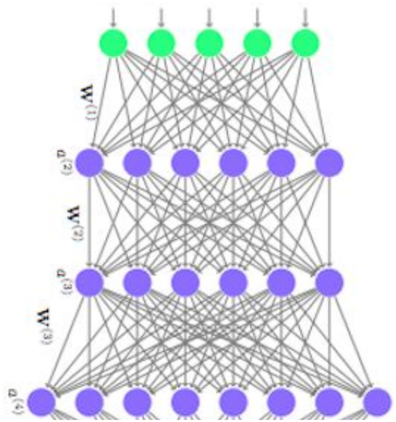
Low-dim, Original Space

High-dim, **Linearly Separated** Space

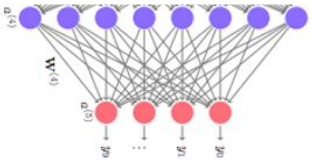
“Trick” in Deep Learning



Low-dim, Original Space

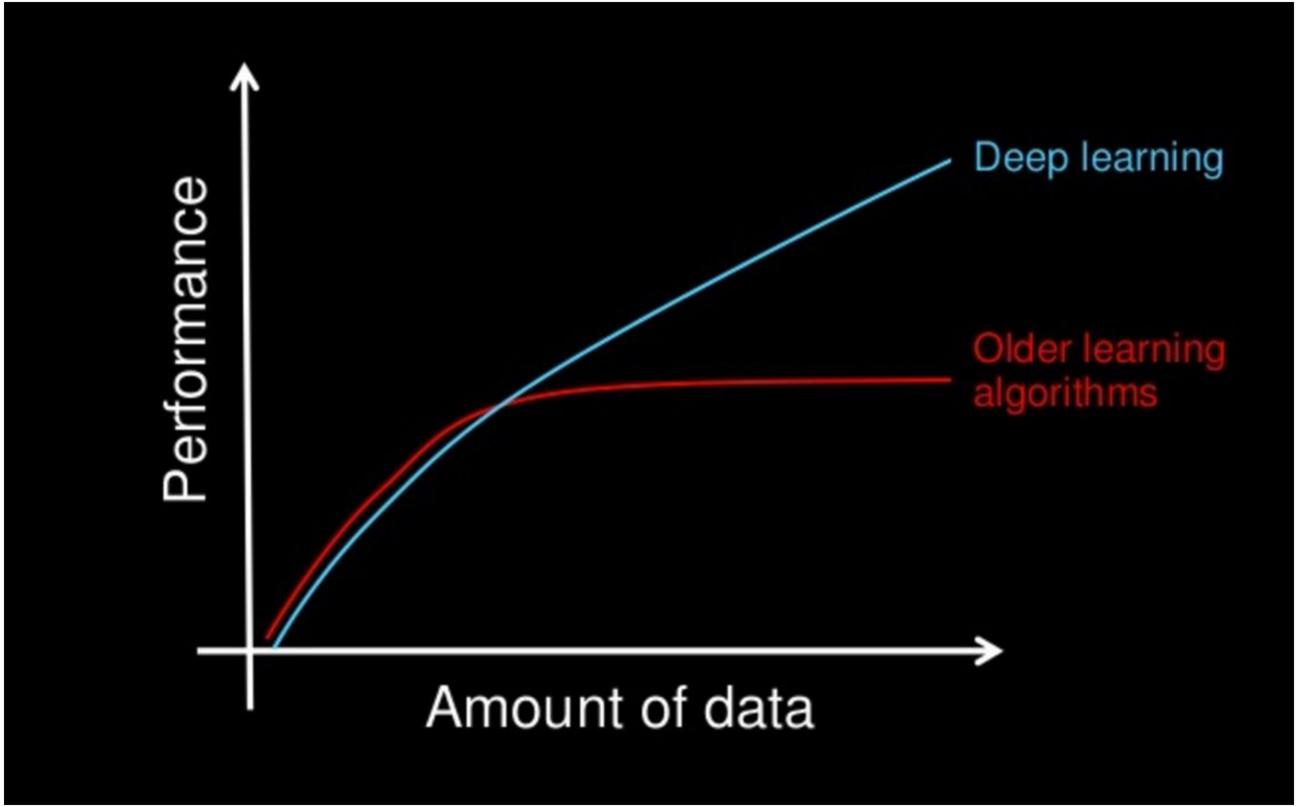


High-dim, **Linearly Separated** Space



Softmax Classifier
(Linear Model)

Why Deep Learning

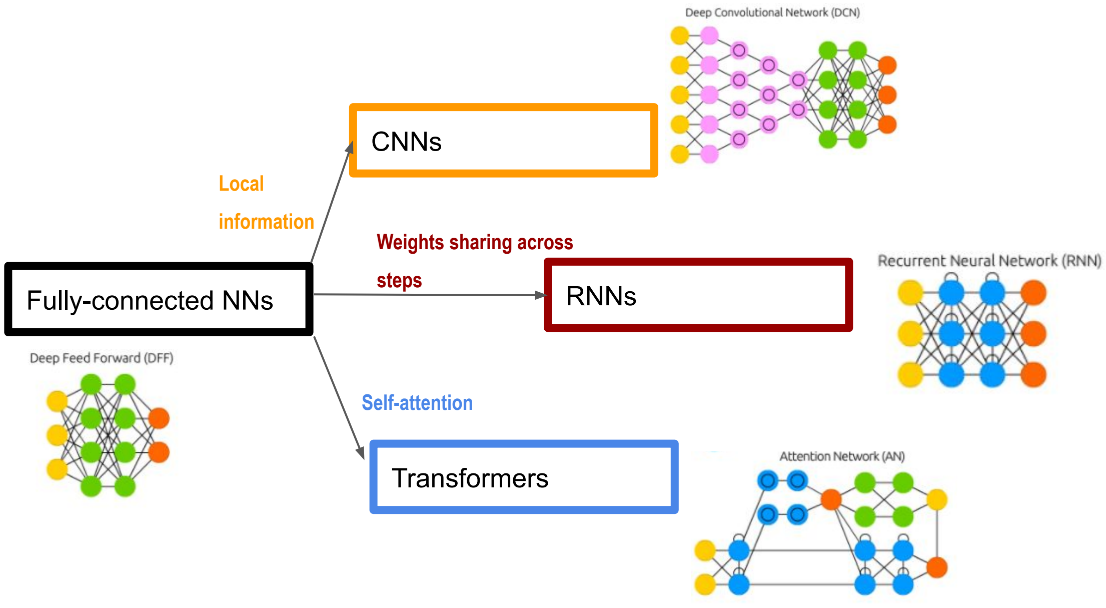


From Andrew Ng

Deep Learning

- Deep learning is a subfield of machine learning
- Most machine learning methods work well because of high-quality feature engineering/representation learning.
- Deep learning is an **end-to-end** structure, which supports automatic representation learning
- Different network structures: CNN, RNN, LSTM, GRU, Attention model, etc

Deep Learning Structures



https://www.asimovinstitute.org/author/fjodorvanveen/

Deep learning with Bayesian Principles

Bayesian learning	Deep learning
Bayesian models	Deep models
(GPs, BayesNets, PGMs,)	(MLP, CNN, RNN etc.)
Bayesian inference	Stochastic training
(Bayes rule)	(SGD, RMSprop, Adam)

	Bayes	DL
Can handle large data and complex models?	✗	✓
Scalable training?	✗	✓
Can estimate uncertainty?	✓	✗
Can perform sequential / active /online / incremental learning?	✓	✗

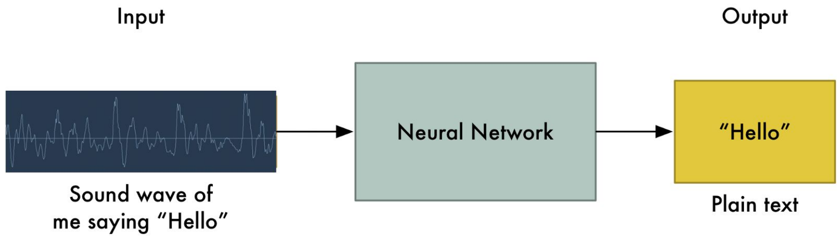
Use a probability distribution over the network weights and output an average prediction of all sets of weights in that distribution

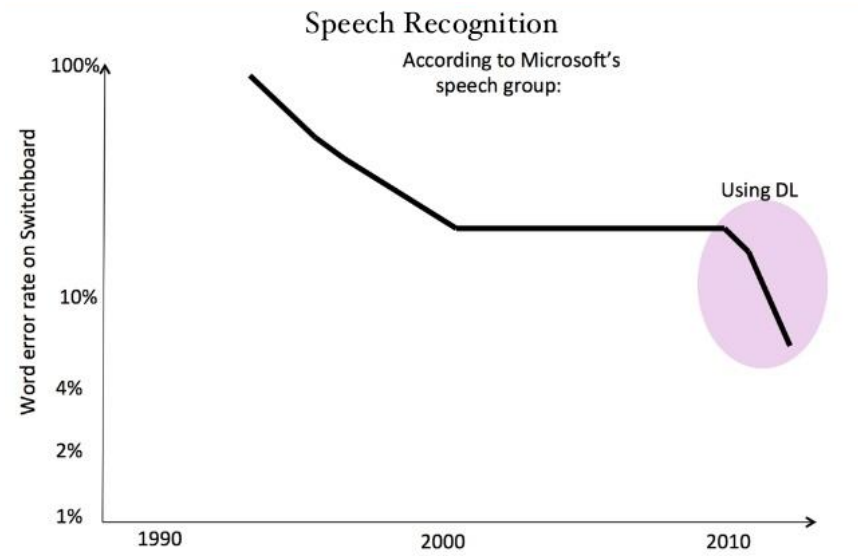
https://slideslive.com/38923183/deep-learning-with-bayesian-principles

Applications of DL

Deep Learning for Speech

The first real-world tasks addressed by deep learning is speech recognition

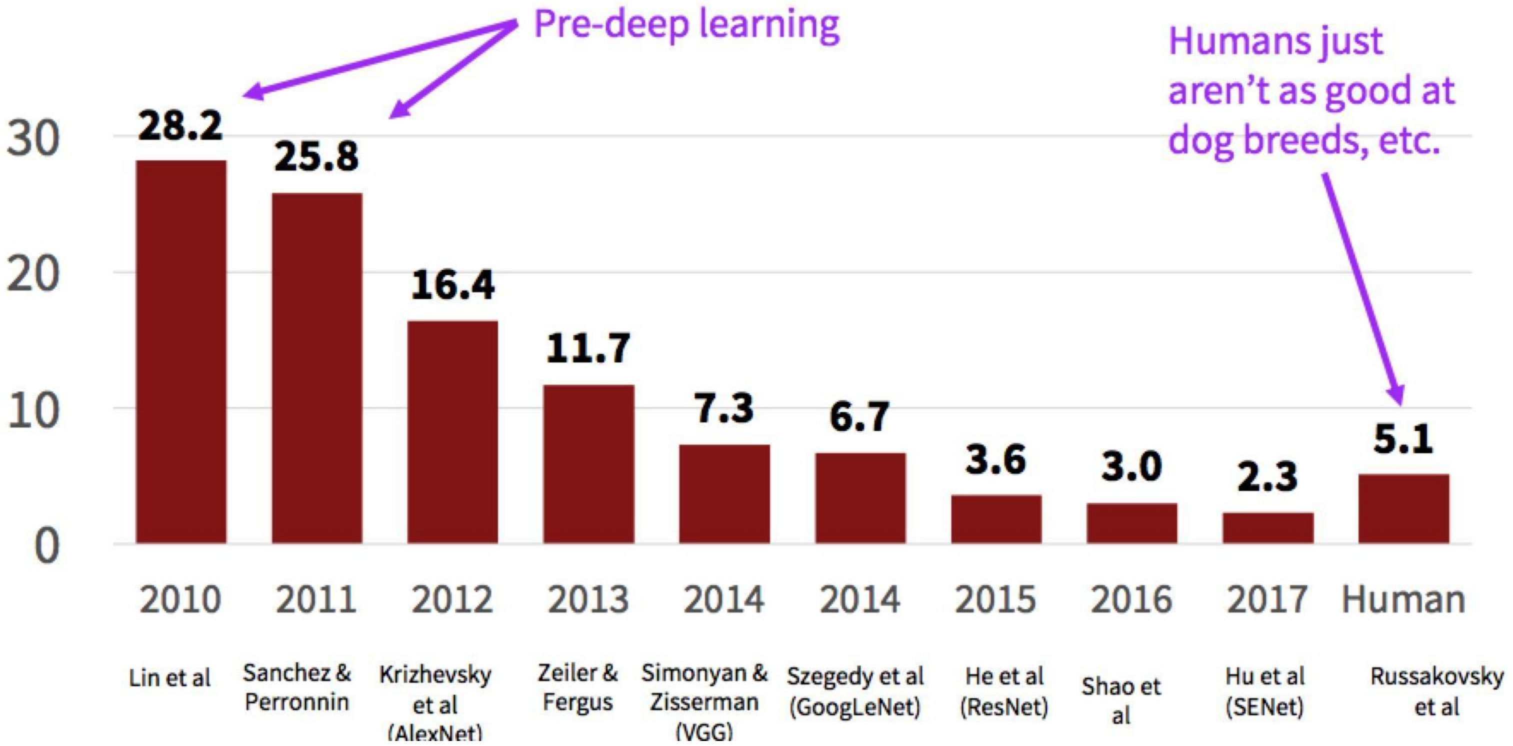




Deep Learning for Computer Vision

- Computer vision may be the most well-known breakthrough of DL.
- ImageNet Classification with Deep Convolutional Neural Networks.

ImageNet Scoreboard



Deep Learning For Arts

Style transfer based on Deep Learning: use one image to stylize another.

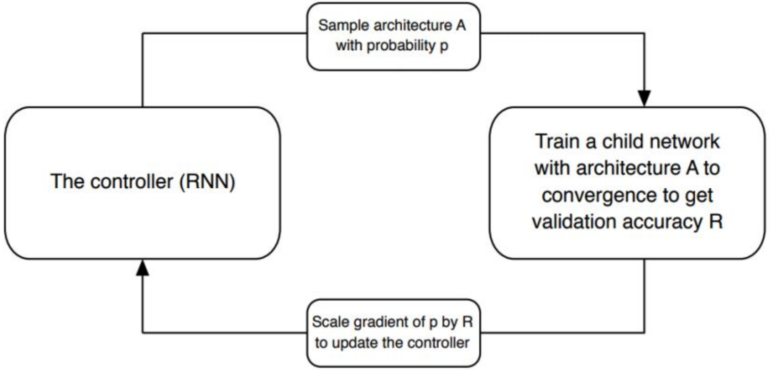
The now iconic examples from Figure 2 of Gatys et al (2015).

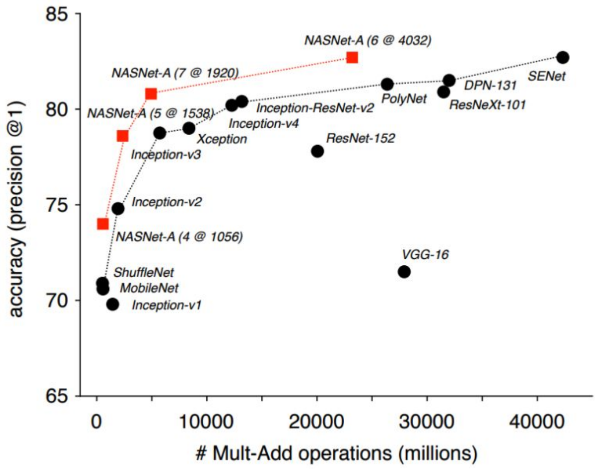
Deep Learning For Data Generation

Given training data, generate new data samples from same distribution

Examples of Photorealistic GAN-Generated Faces.

AutoML and Neural Architecture Search





Source:Lex Fridman