Deep Learning
(Introduction)

Sadegh Eskandari
Department of Computer Science, University of Guilan
eskandari@guilan.ac.ir

Today …

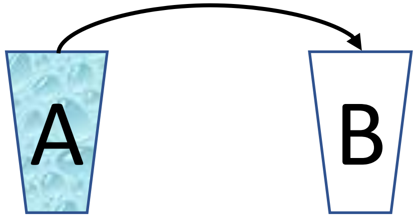
- A high-level review of machine learning (ML)
- A high-level introduction to deep learning (DL)

Machine Learning (ML) …

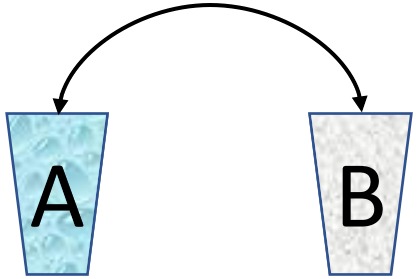
Consider a robot with a single capability: pouring one glass into another

Operator

$B \leftarrow A$



Question: how the robot can swap the contents of two glasses?



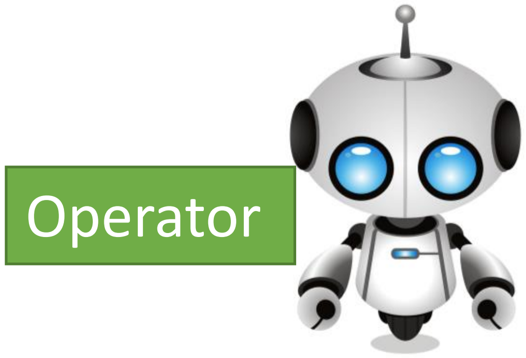
Problem

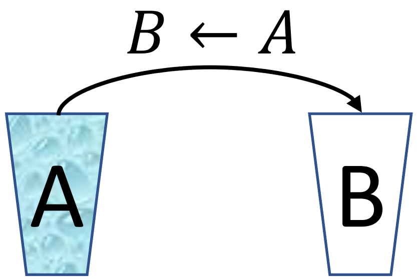
$$C \leftarrow A$$
$$A \leftarrow B$$
$$B \leftarrow C$$

Algorithm

Machine Learning (ML) …

Consider a robot with a single capability: pouring one glass into another





A, B

$\Downarrow$

$\max(A, B)$

Problem

Another Question: How to find the max of two glasses?

The problem is unsolvable by the robot. Why?

- The comparison operation is not defined for the robot
- To solve the problem we should change the operator

Machine Learning (ML) …

Operator (Computer)

Capabilities:

- Input/Output (I/O)
- memory W/R
- Some basic arithmetic and logical operations (+,-,*,/, %, and, or, not, …)

$$A, B$$

$$\Downarrow$$

$$\max(A, B)$$

Problem

Problem: How to find the max of two glasses?

```
read A,B
if A>B:
    max = A
else:
    max = B
print max
```

Algorithm

Machine Learning (ML) …

- A problem is said to be Decidable if we can always construct an algorithm that can solve the problem correctly.
- An example of undecidable problems:

  Can one algorithm specify the output of another algorithm?

- Decidability does not mean simplicity!
  - ❑ Traveling Salesman Problem (TSP): simple to program but hard to execute
  - ❑ Recognizing dogs and cats in an image: simple to do but hard to program

Machine Learning (ML) …

Traveling Salesman Problem (TSP)

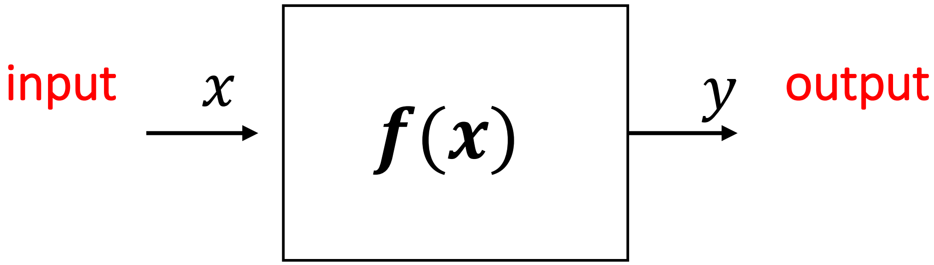
- For a given weighted complete graph with n nodes, find the Hamilton circuit with minimum length.
- An algorithm should compare $(n-1)!$ circuits to find the best one.
- Time required to run this algorithm on a good computer:
  - ❑ $n = 4$ then $time \approx 0.000000007\text{s}$
  - ❑ $n = 99$ then $time \approx 3.1 \times 10^{140}\ years$ ☹

Machine Learning (ML) …

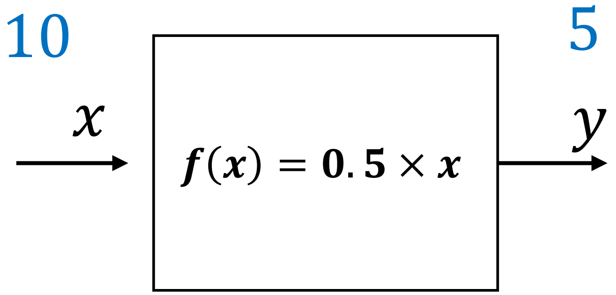
Dogs vs Cats

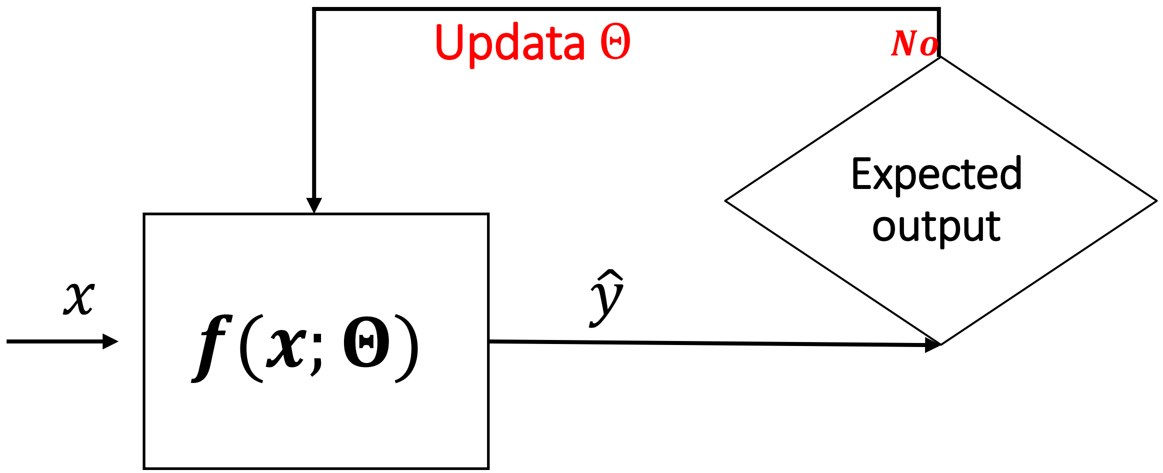
An effective approach: Machine Learning

Machine Learning (ML) …



A typical algorithm or function

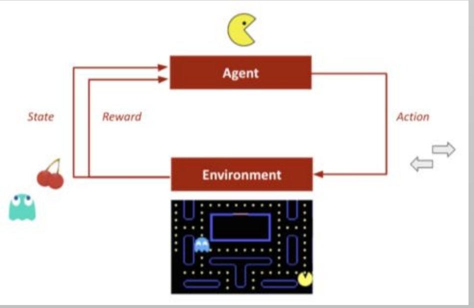




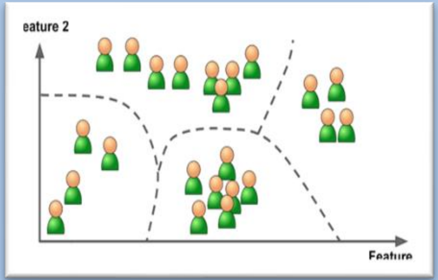
A Parametric Algorithm (Model)

Machine Learning (ML) …

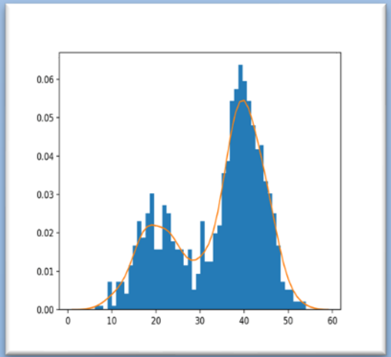
(Reinforcement)



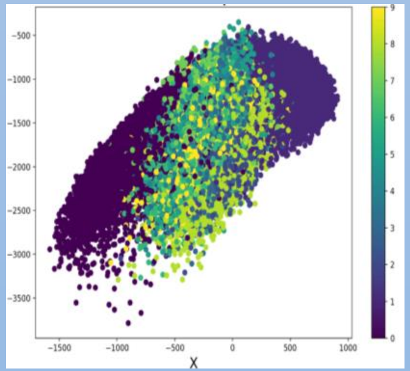
(Unsupervised)



(Clustering)



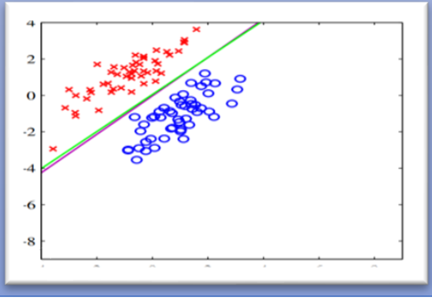
(Density Estimation)



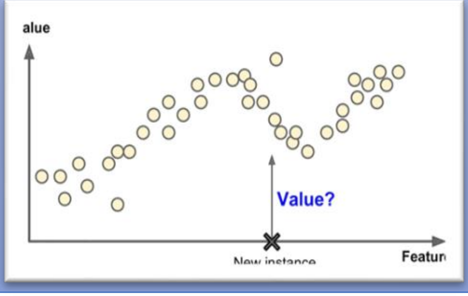
(Visualization)

(Supervised)

(Classification)



(Regression)



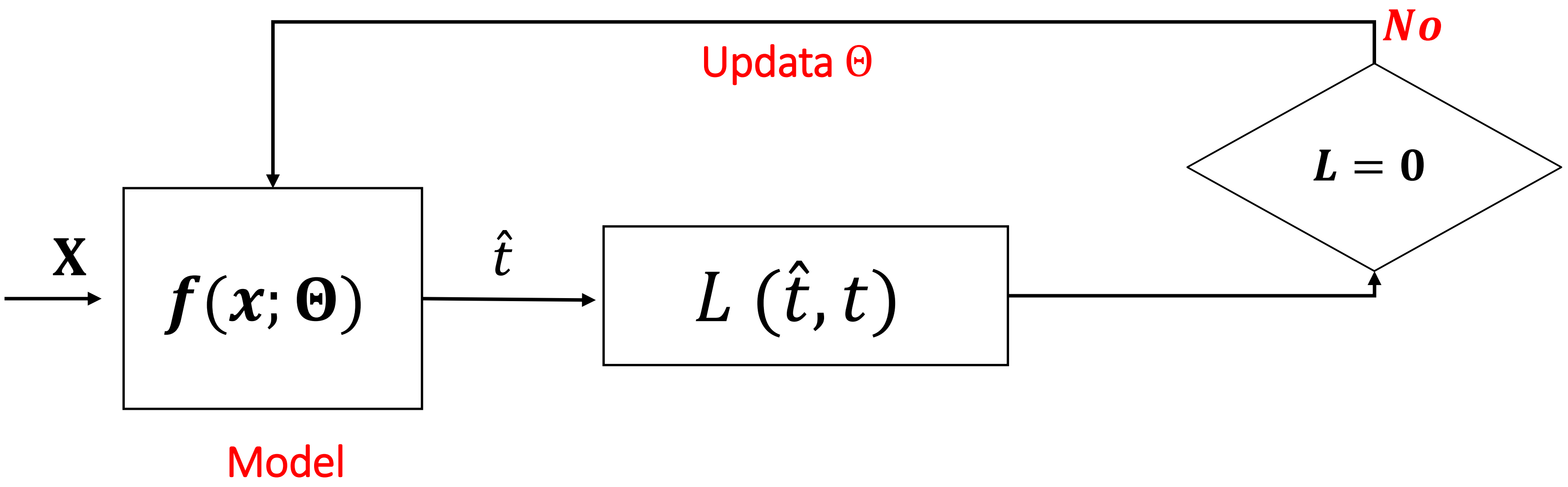
Machine Learning (ML) …

- **Supervised Learning**: Suppose that we are given a training set comprising N observations of random variable x (training set) :

$$\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)^{\boldsymbol{T}}$$

- Moreover, for each observation $\boldsymbol{x}_i$ we are given a target value t_i (training target):

$$\mathbf{t} = (t_1, t_2, \dots, t_N)^{\boldsymbol{T}}$$



Machine Learning (ML) …

- $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ is generated uniformaly in $[0,1]$.
- $\boldsymbol{t} = \{t_i |\ t_i = \sin(2\pi x) + \mathcal{N}(0,0.3)\,, i = 1, 2, \ldots, N\}$

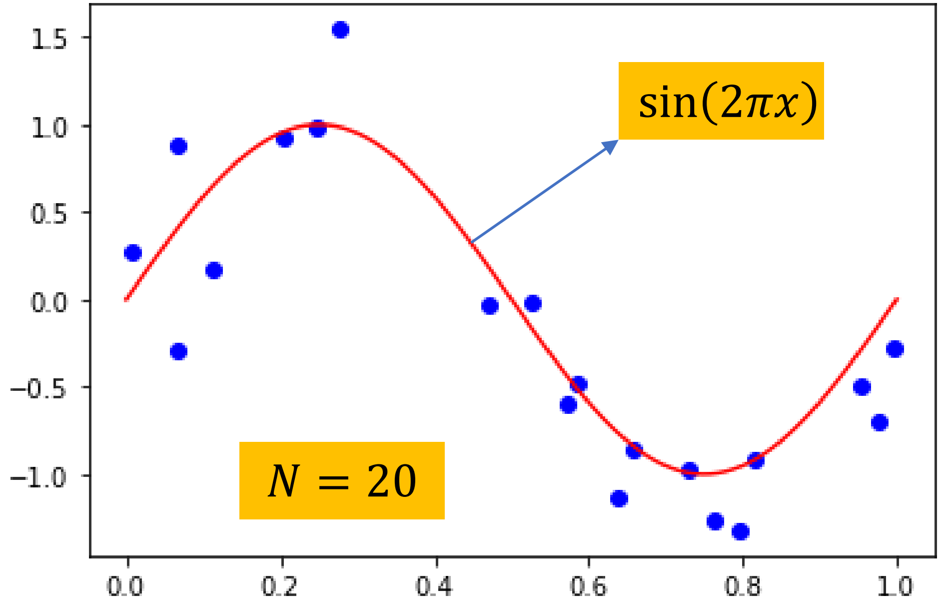
- The generating function in not known and the aim is to estimate it such that:
  - ❑ The estimated function should describe the training data
  - ❑ The estimated function should generalize to new data

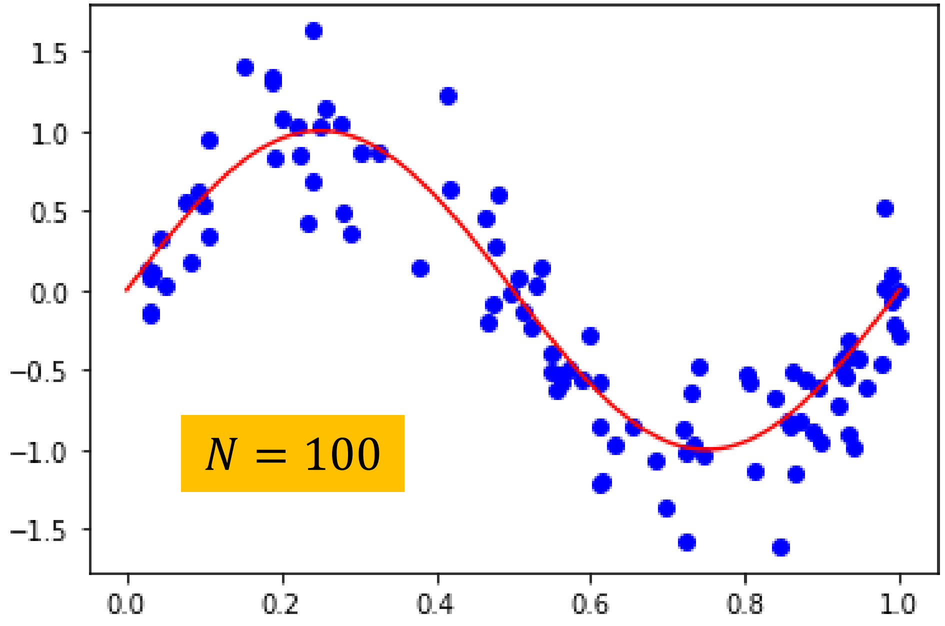
- In particular, we shall fit the data using a polynomial function of the form

$$y(x; \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M$$

  - ❑ M: the order of polynomial
  - ❑ $w \equiv [w_0, w_1, \ldots, w_M]$: The model parameters (unknown in advance)

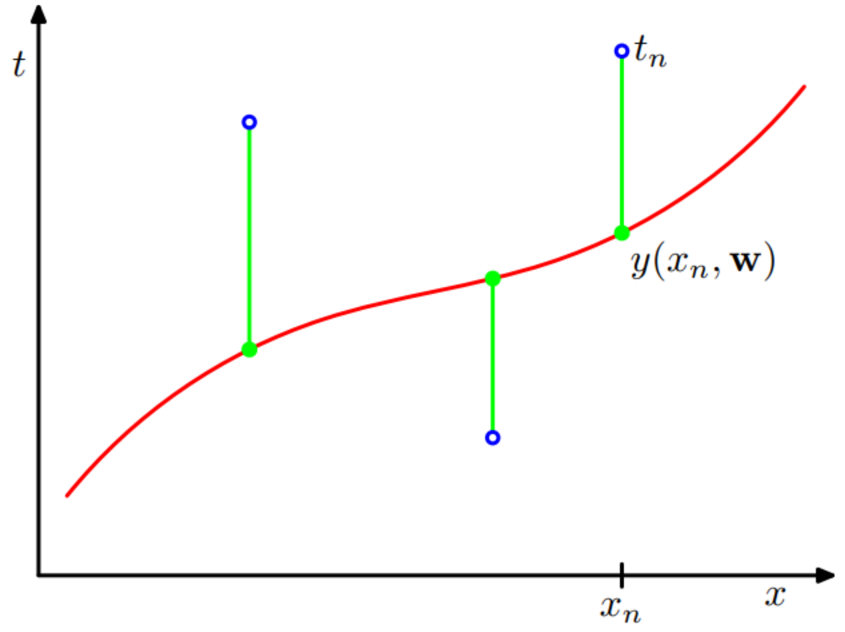
- $y(x, \boldsymbol{w})$ is a linear function of the coefficients $\boldsymbol{w}$. Such functions are called linear models.





Machine Learning (ML) …

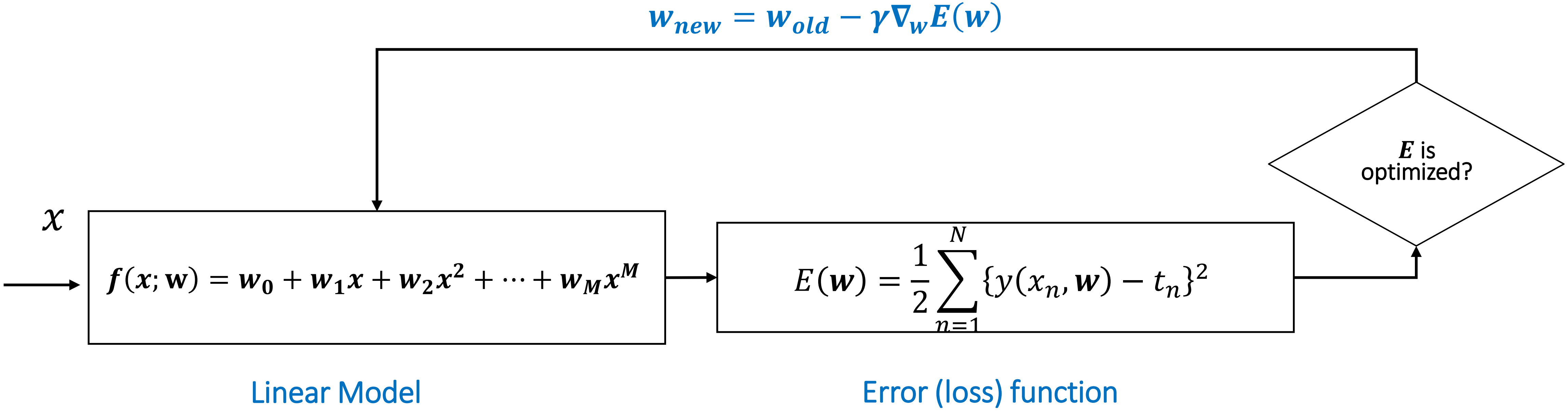
- An error function (loss function) is required to measure the misfit between the function $y(x, \boldsymbol{w})$, for any given $\boldsymbol{w}$, and the training data points.



$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2$$

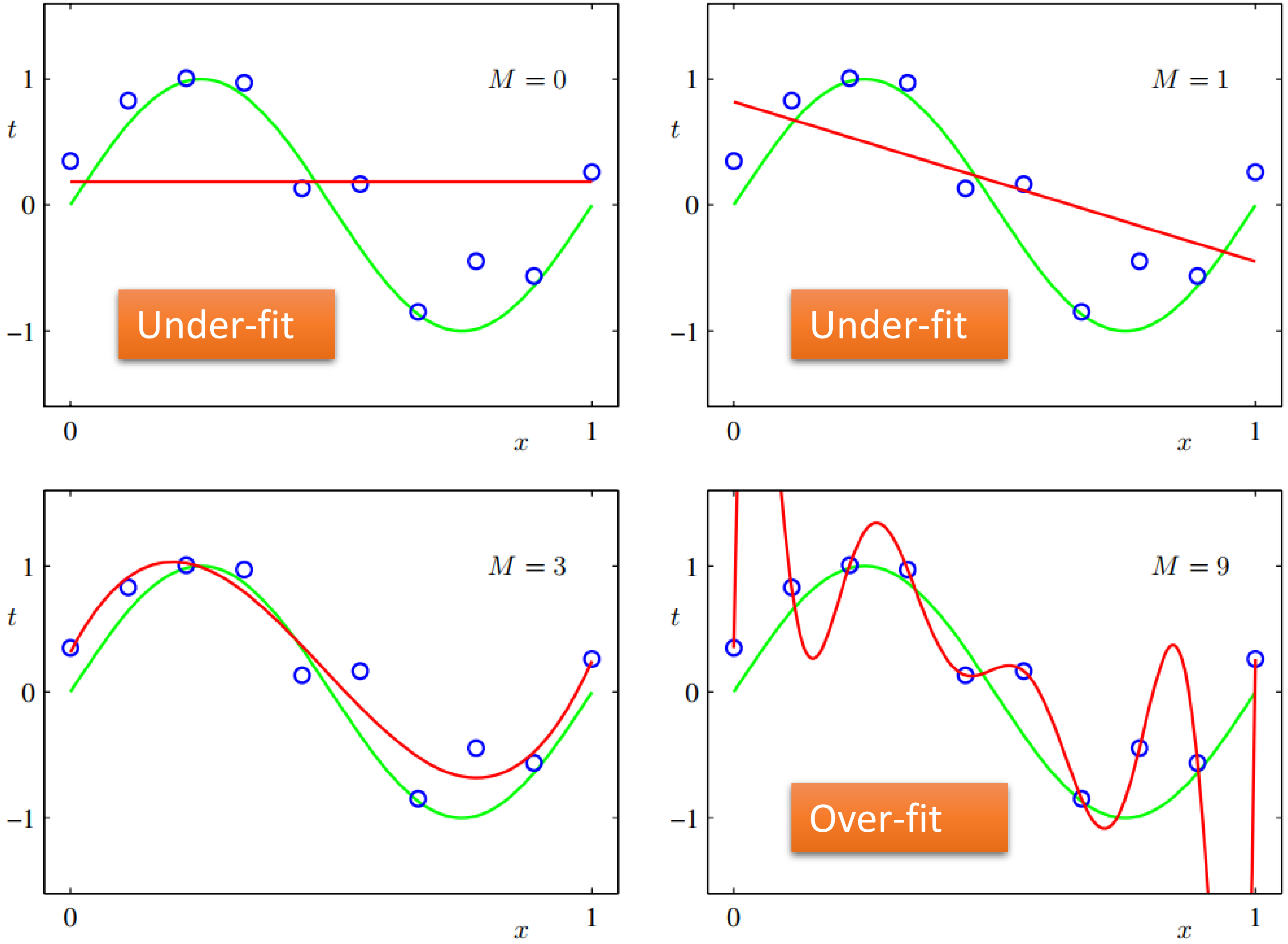
- $E(\boldsymbol{w})$ is a quadratic function of $\boldsymbol{w}$,
- Therefore $\frac{\partial E}{\partial \boldsymbol{w}}$ is linear in the elements of $\boldsymbol{w}$, and so the minimization of the error function has a unique solution, which can be found in closed form.

Machine Learning (ML) …

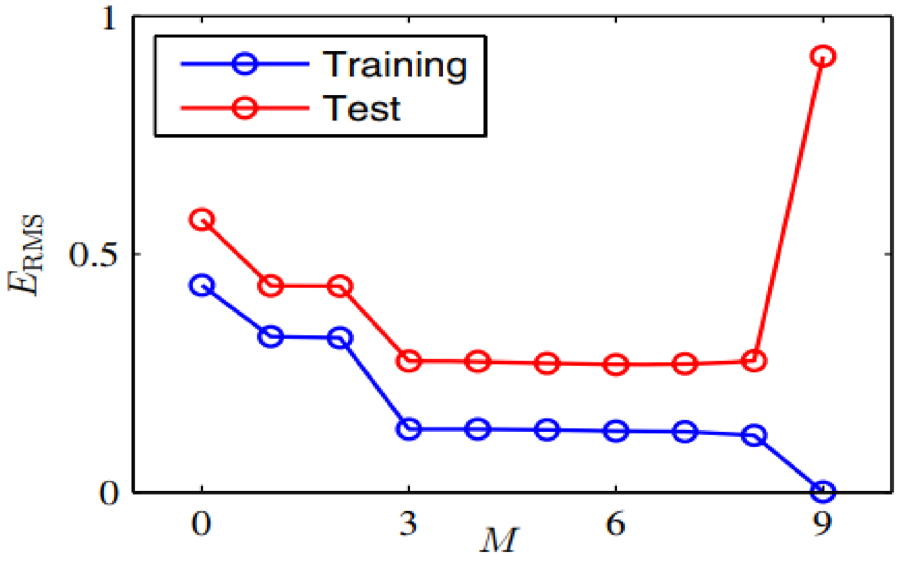


Machine Learning (ML) …

Model Selection (Model Comparison)



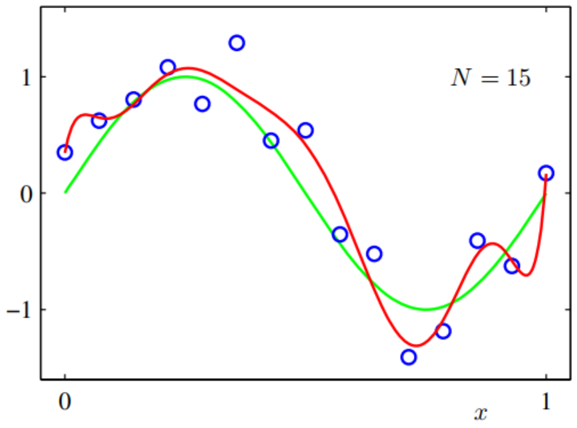
	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

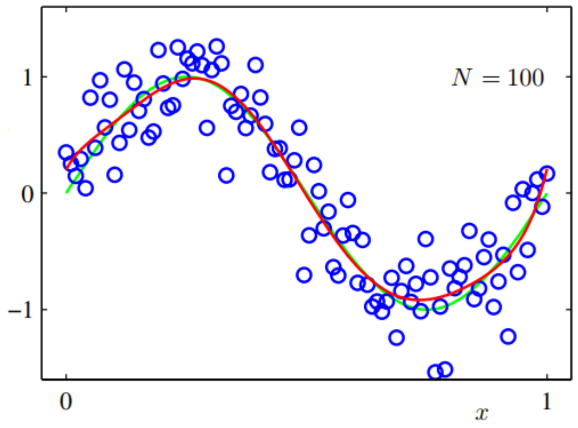


Machine Learning (ML) …

Model Selection (Model Comparison)

- For a given model complexity, the over-fitting problem become less severe as the size of the data set increases.

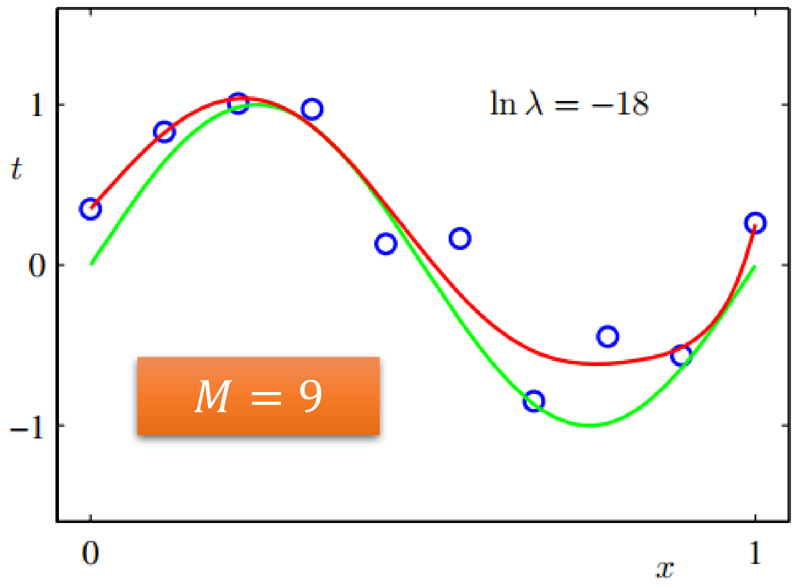




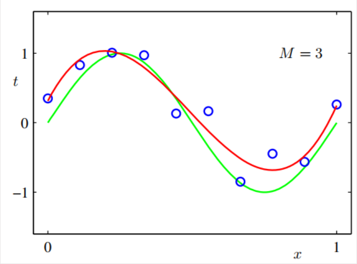
- One technique that to control the over-fitting phenomenon regularization, which involves adding a penalty term to the error function.

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

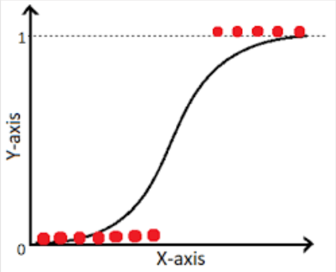
Where $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^T\boldsymbol{w} = w_0^2 + w_1^2 + \cdots + w_M^2$



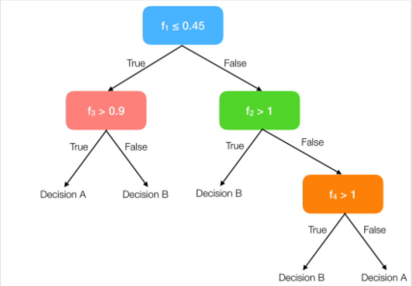
Machine Learning (ML) …



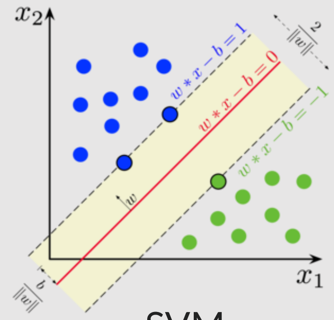
Linear Regression



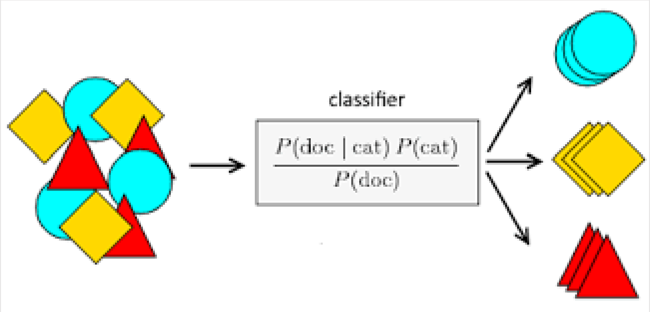
Logistic Regression



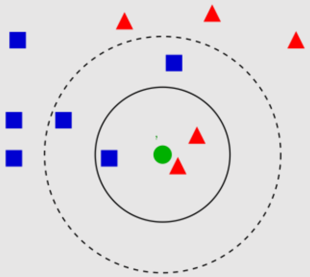
Decision Tree



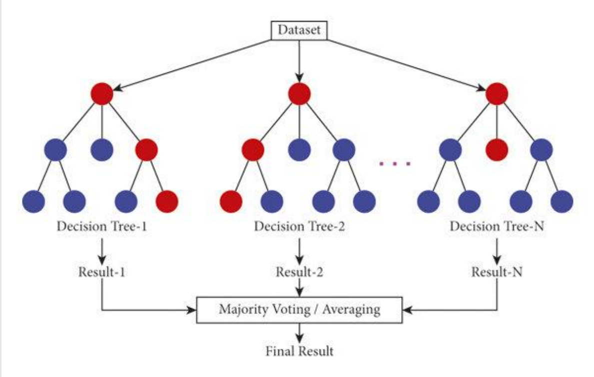
SVM



Naïve Bayes Classifier



KNN



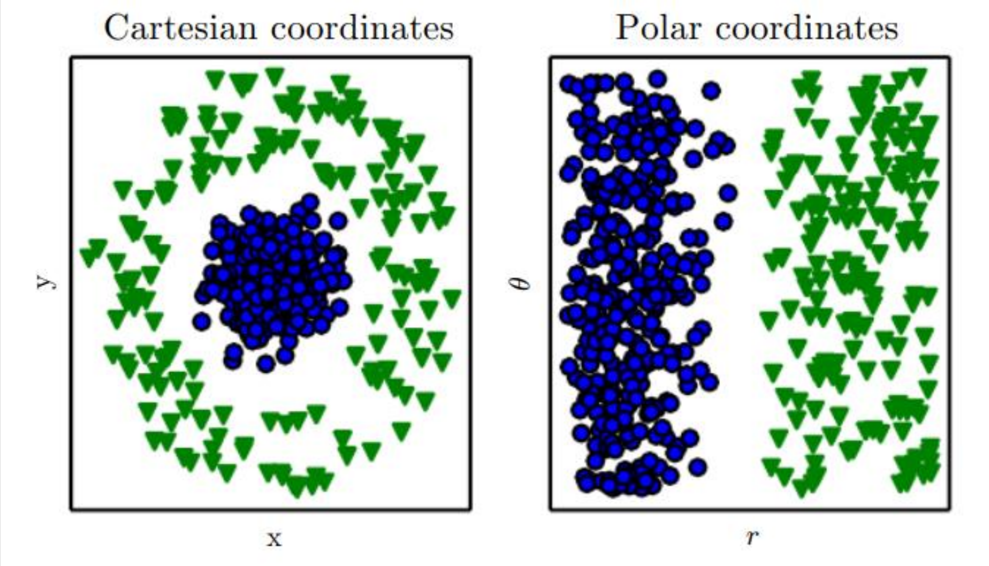
Random Forest

Classic Machine Learning Algorithms (Supervised)

Deep Learning (DL) …

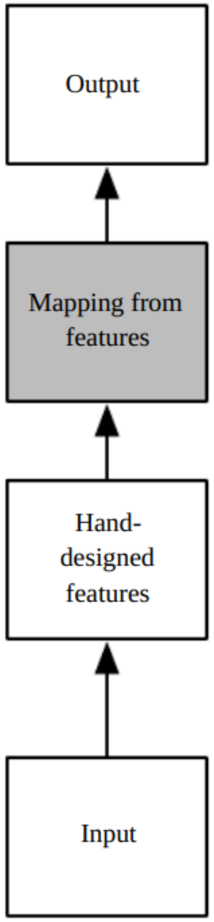
What's the problem with classic machine learning approaches?

- Their input is a set of hand-designed features.
- The performance of these simple machine learning algorithms depends heavily on the representation of the data they are given.
- Therefore, designing a right set of features is the most important in these approaches.



Example: Different feature representation designed to separate two categories of data using a linear classifier

Solution: Representation Learning

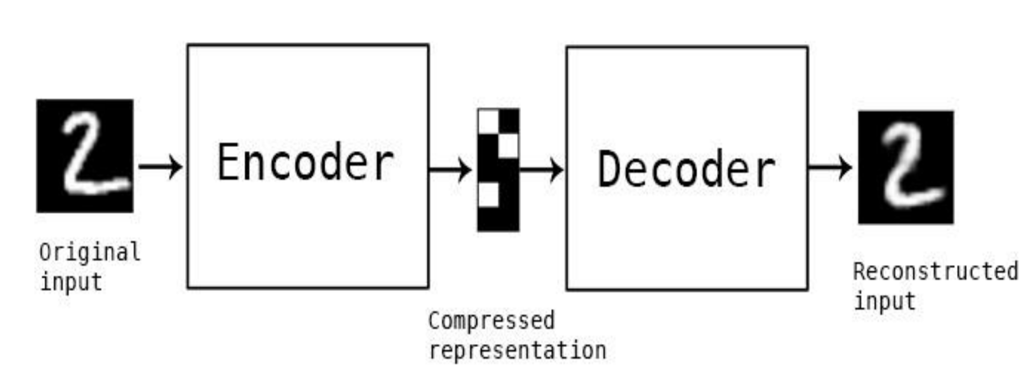


Classic machine learning

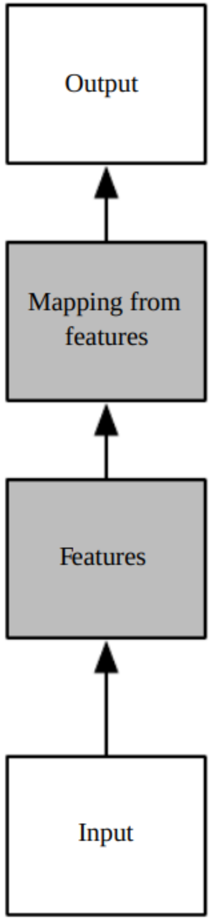
Deep Learning (DL) …

Representation Learning

- Representation learning is to use machine learning to discover not only the mapping from representation to output but also the representation itself.
- Example: Autoencoders



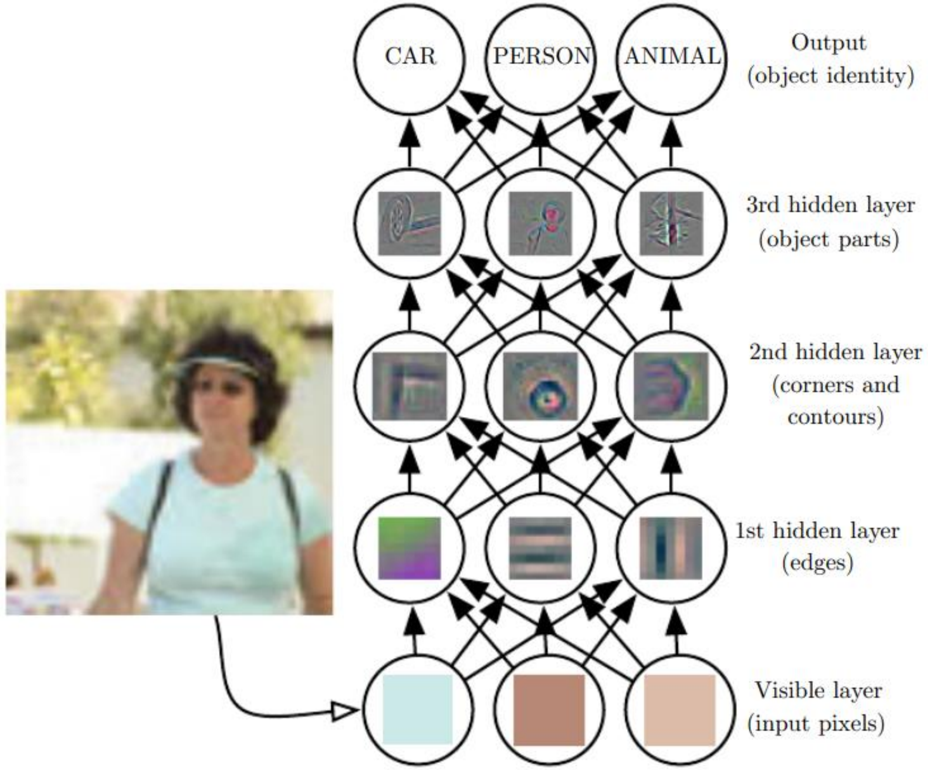
- An autoencoder is the combination of an **encoder** function and a **decoder** function
- The encoder converts the input data into a different representation,
- The decoder converts the new representation back into the original format.
- Autoencoders are trained to preserve as much information as possible when an input is run through the encoder and then the decoder

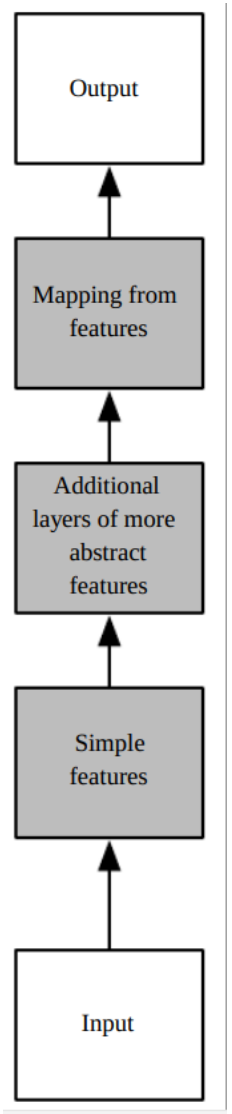


Representation Learning

Deep Learning (DL) …

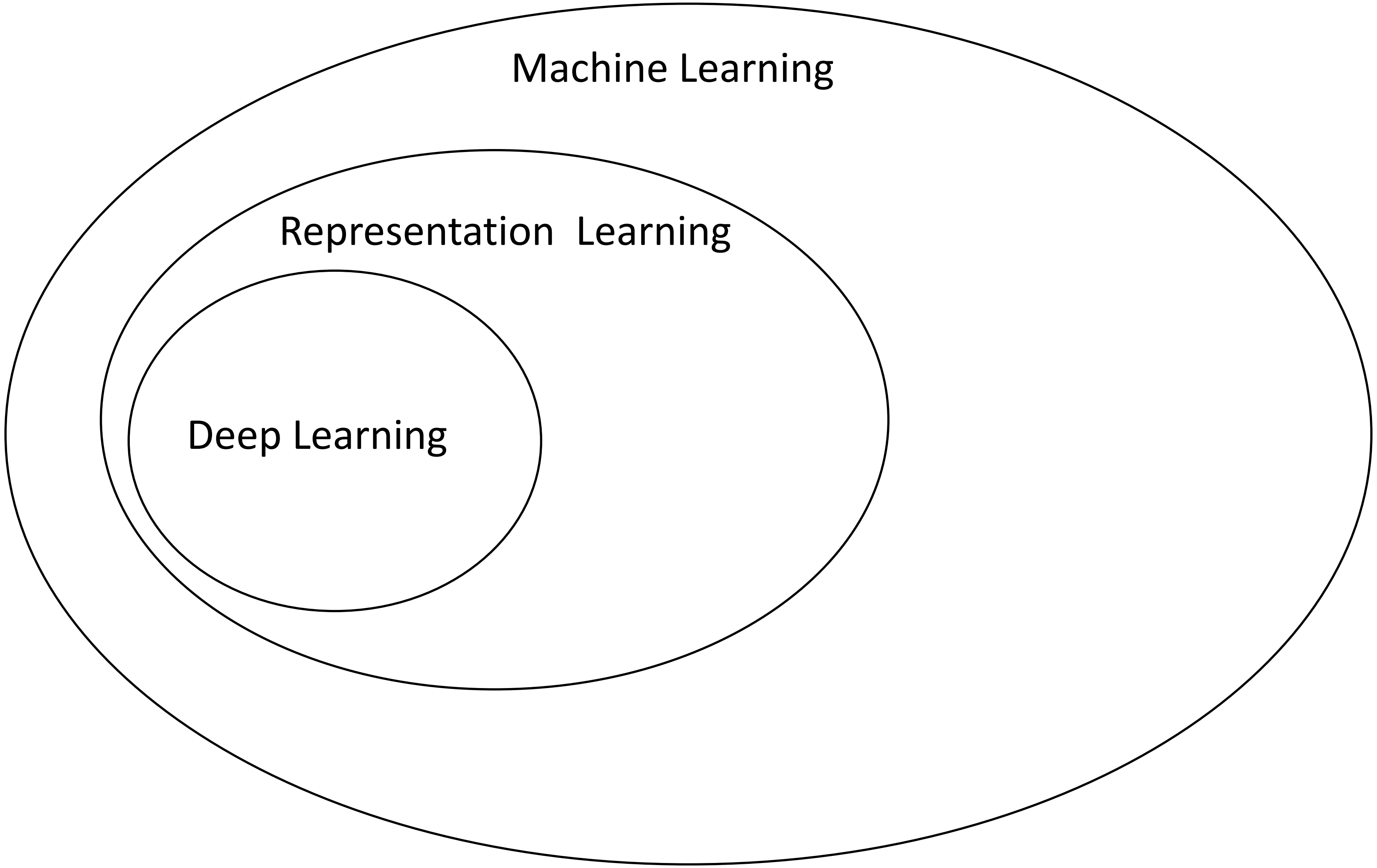
- **Deep learning** is a class of representation learning that uses different levels of representations
- Higher-level representations are expressed in terms of other, lower-level and simpler representations
- Example: Multi-layer perceptron (MLP)





Deep
Learning

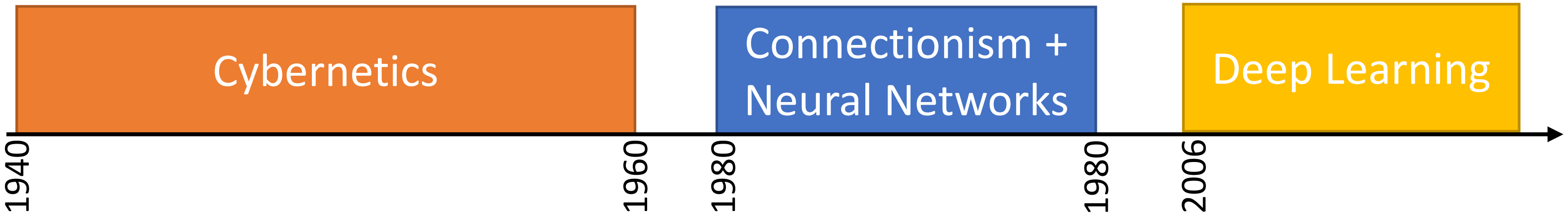
Deep Learning (DL) …



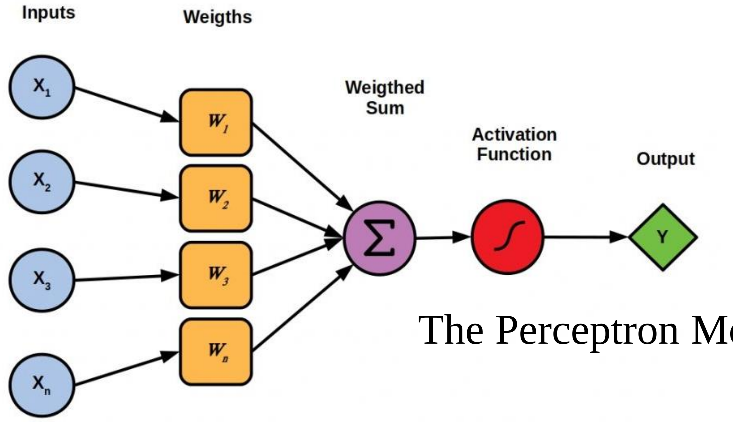
Deep Learning (DL) …

History

- Deep learning has had a long and rich history, but has gone by many names



- Cybernetics (the first wave)
  - Development of theories of biological learning (McCulloch and Pitts, 1943; Hebb, 1949)
  - Perceptron (Rosenblatt, 1958)
  - Universal approximation theorem
    - Any continuous function $f : [0,1]^n \rightarrow [0,1]$ can be approximated arbitrarily well by a neural network with at least 1 hidden layer with a finite number of weights

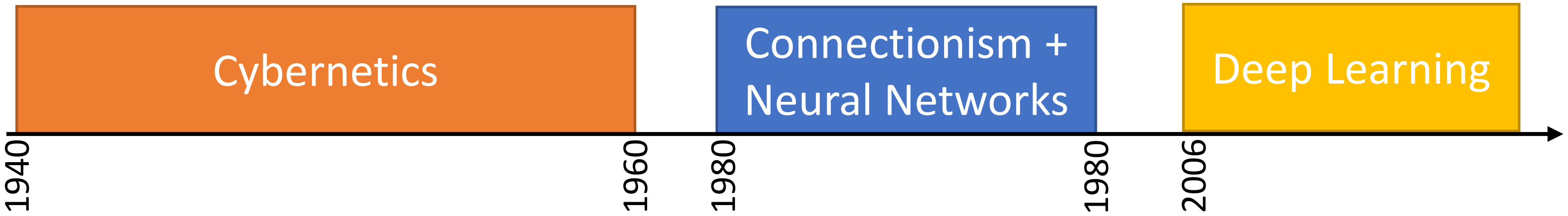


The Perceptron Model

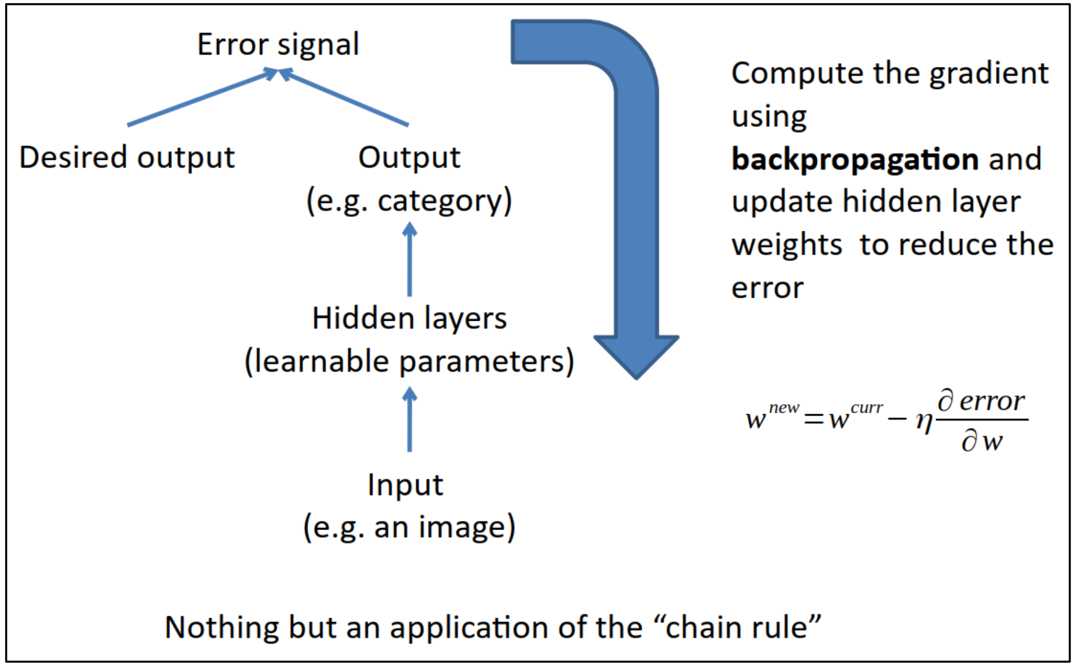
Deep Learning (DL) …

History

- Deep learning has had a long and rich history, but has gone by many names



- Connectionism (the second wave)
  - backpropagation (Rumelhart et al., 1986a)

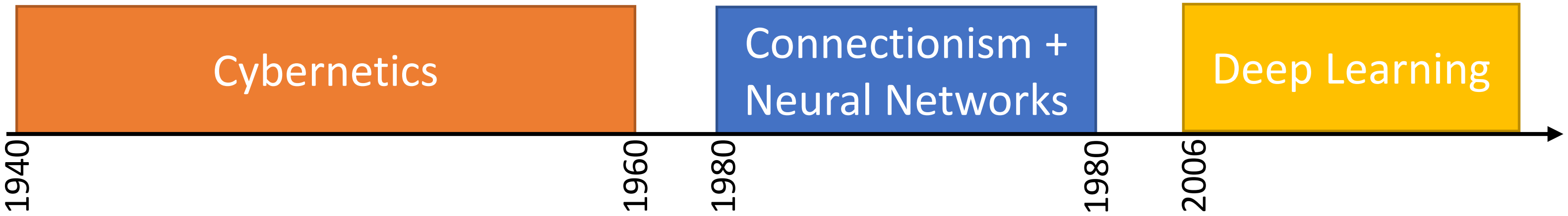


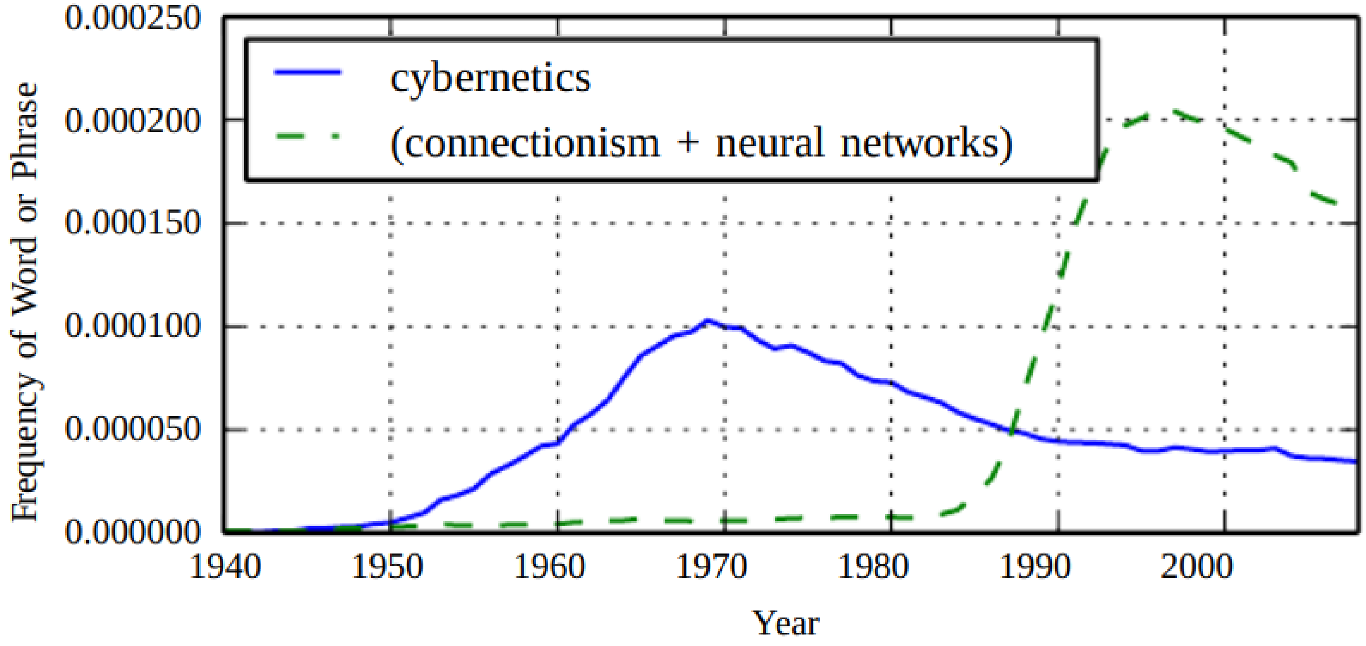
Source: Deep learning course, Emre Akbas, https://user.ceng.metu.edu.tr/~emre/Fall2021-DeepLearning.html

Deep Learning (DL) …

History

- Deep learning has had a long and rich history, but has gone by many names

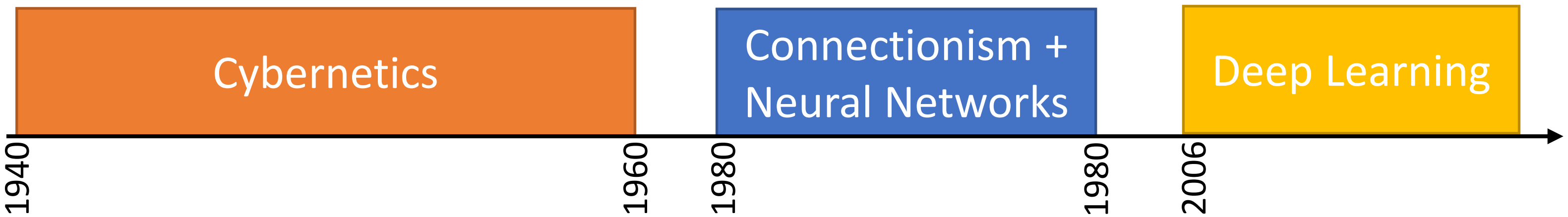




Deep Learning (DL) …

History

- Deep learning has had a long and rich history, but has gone by many names

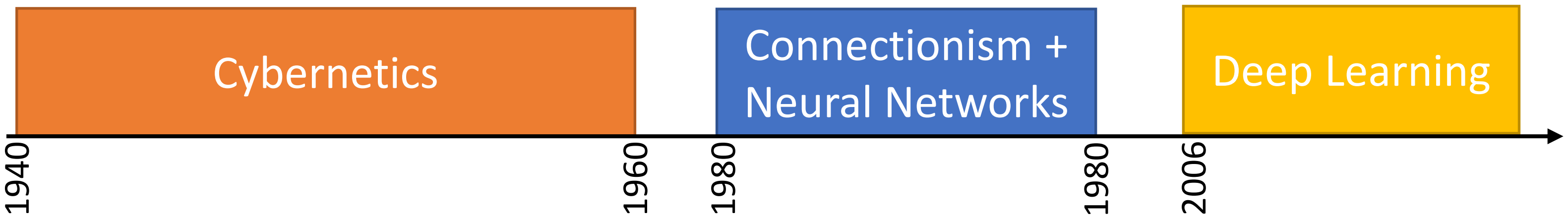


- Why did researchers abandon backpropagation?
  - It was not able to be used in complex networks with multiple hidden layers!
- Today, the researchers have found the **real** reasons:
  - Datasets were too small.
  - Computers were too slow.
  - The weight initialization was wrong
  - The activation functions were ineffective

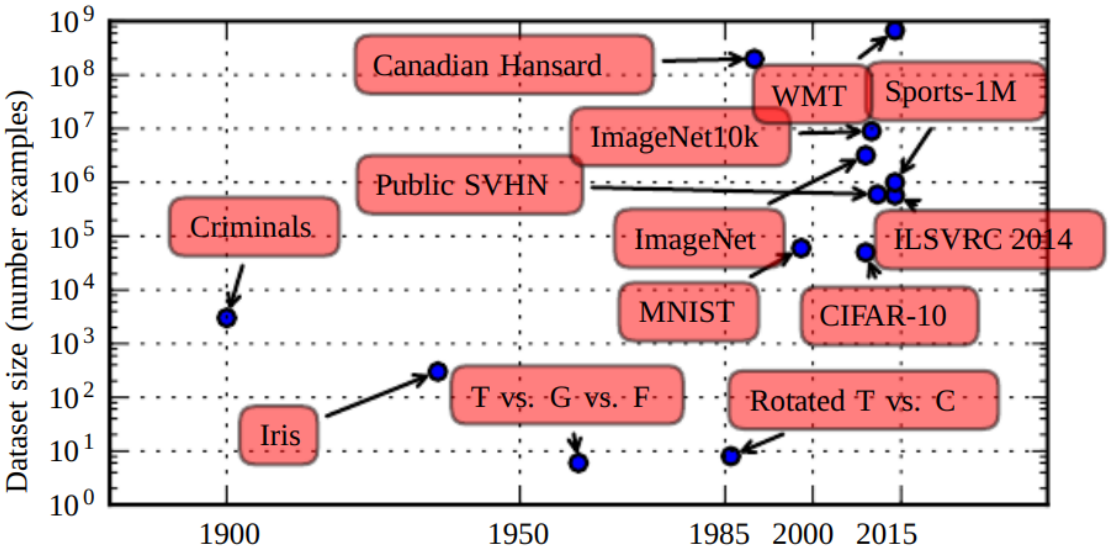
Deep Learning (DL) …

History

- Deep learning has had a long and rich history, but has gone by many names



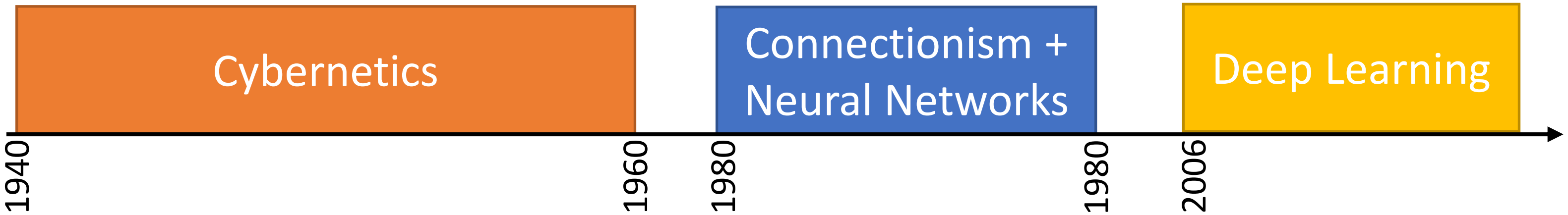
- Deep learning (the third wave)



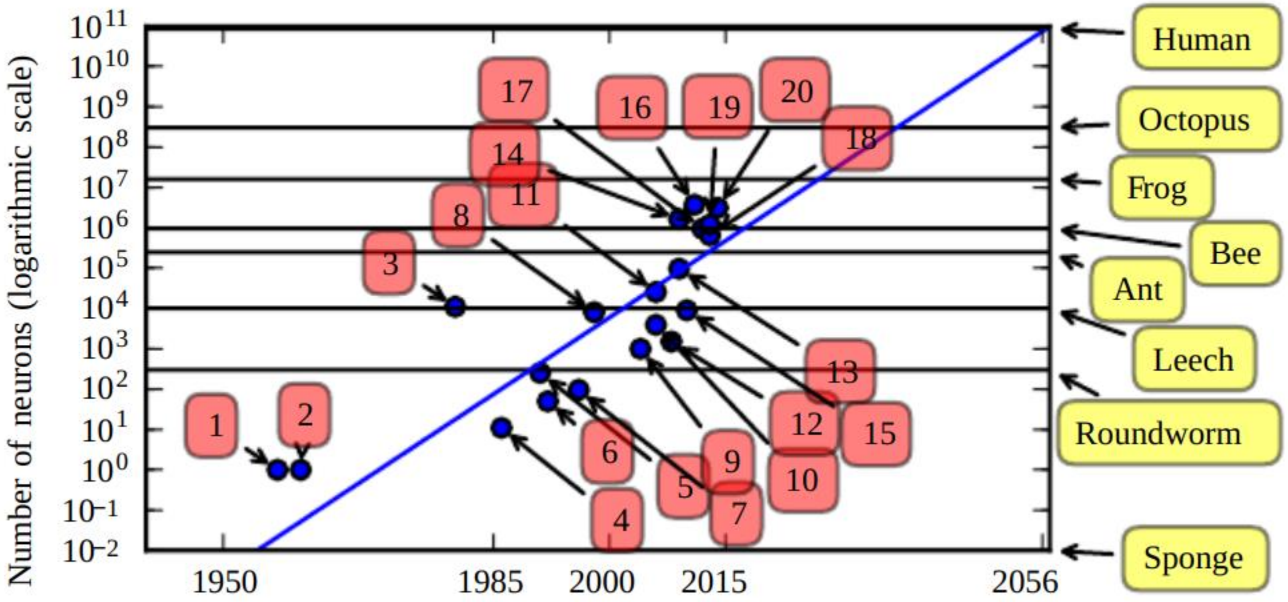
Deep Learning (DL) …

History

- Deep learning has had a long and rich history, but has gone by many names



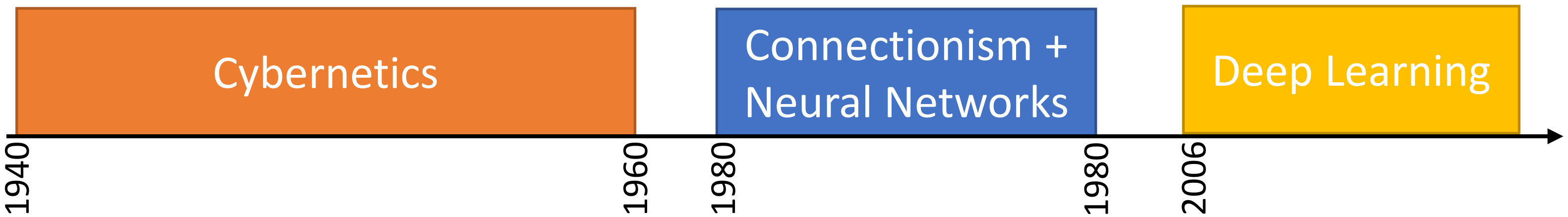
- Deep learning (the third wave)
  - Since the introduction of hidden units, artificial neural networks have doubled in size roughly every 2.4 years.



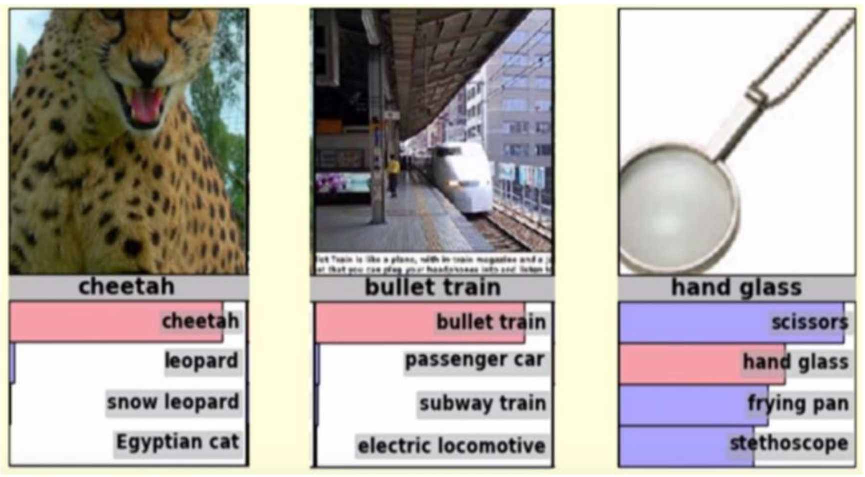
Deep Learning (DL) …

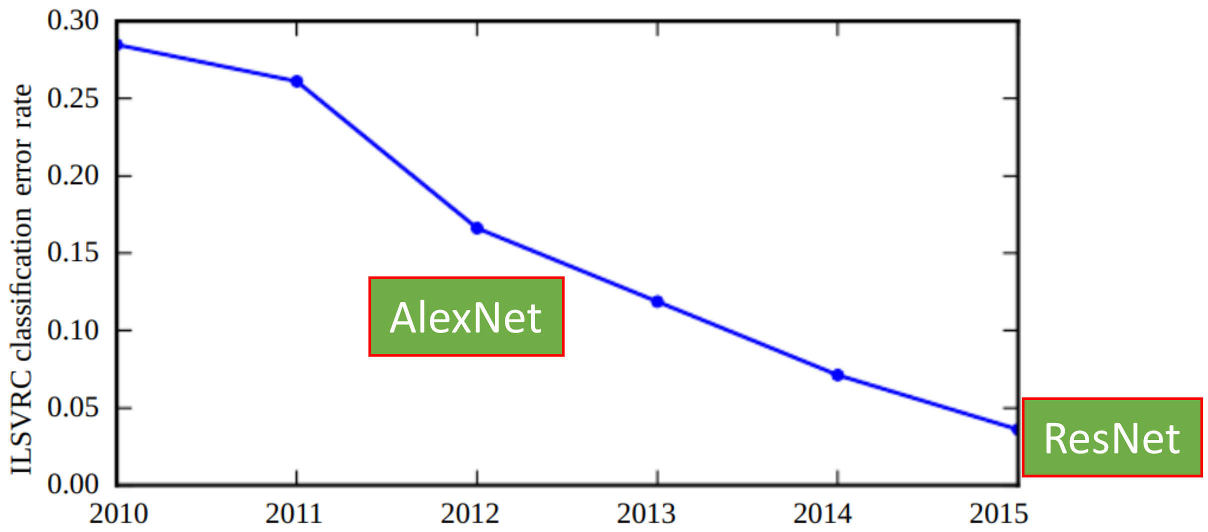
History

- Deep learning has had a long and rich history, but has gone by many names



- Deep learning (the third wave)





Deep Learning (DL) …

Deep learning is everywhere!!