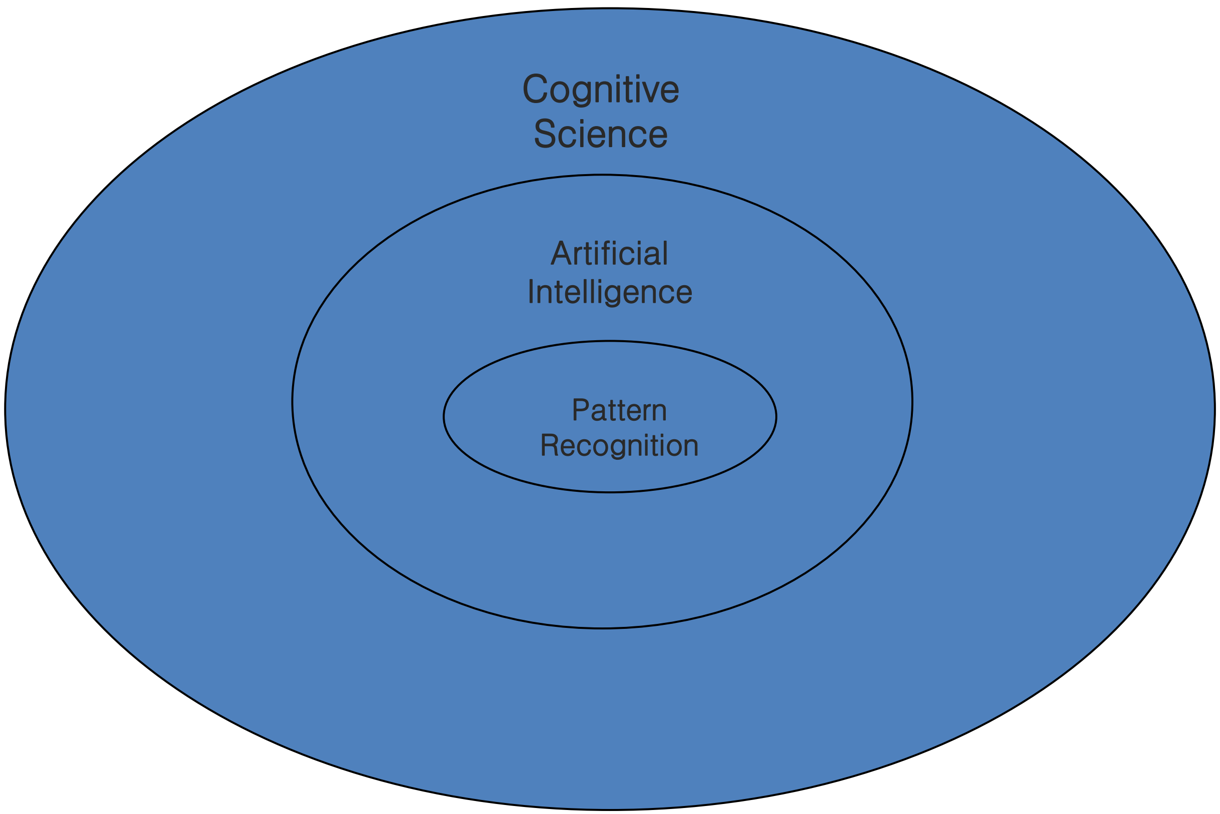
Machine Learning Intro

Prof. Cheolsoo Park

02_Definition of Pattern Recognition

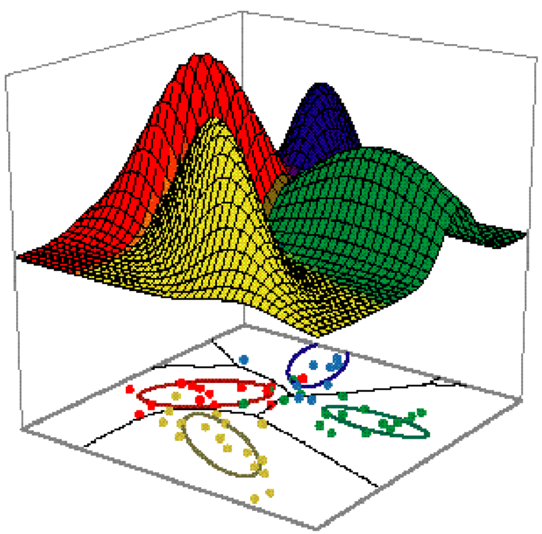
- What is pattern recognition?
  - A study about how a computer or device can recognize an object or phenomenon

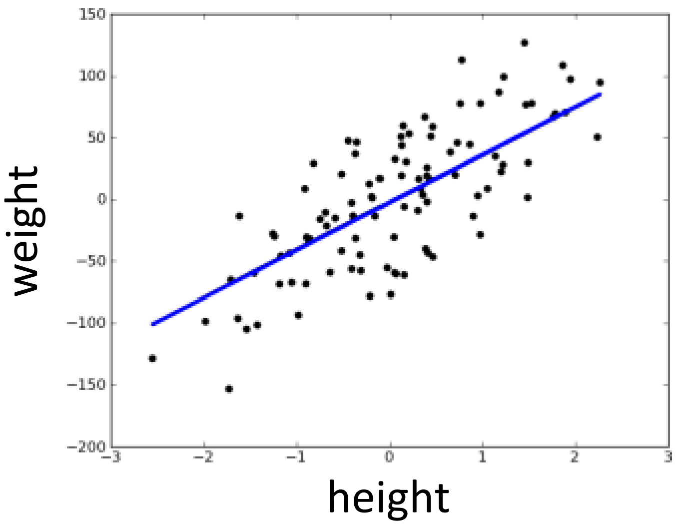


Machine Learning

- ML allows computers to find hidden insights without being explicitly programmed where to look (e.g. pattern recognition)
- A method of automating analytical model building

classification	regression
Y: Classes, Discrete Output	Y: Real Number, Continuous Output



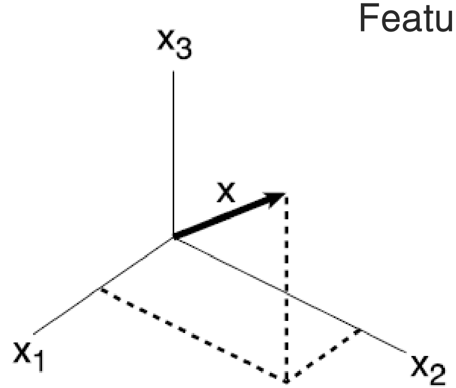


03_Feature and Pattern

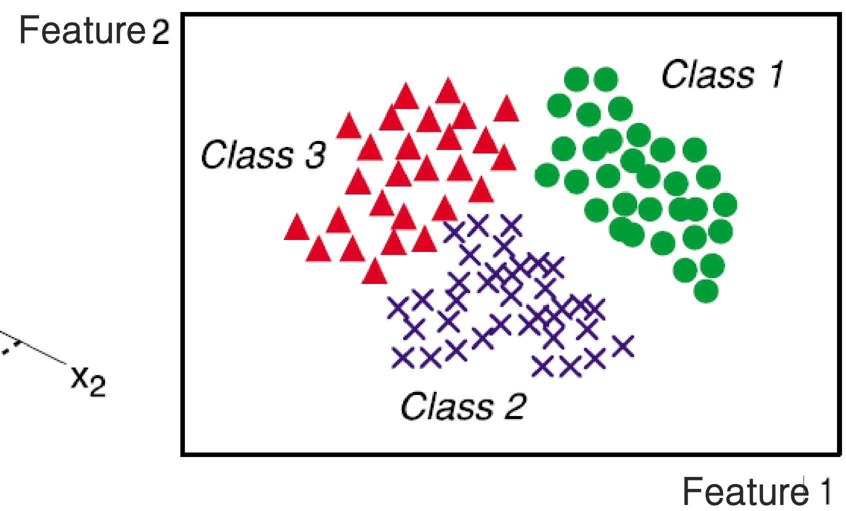
- What is feature?
  - A feature of something is an interesting or important part or characteristic of it.
- What is pattern?
  - Set of traits and features

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

Feature vector

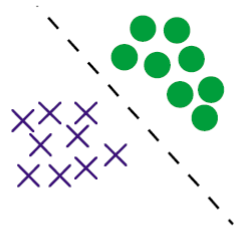


Feature space



03_Feature and Pattern

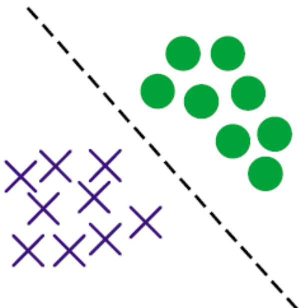
- Good and bad features



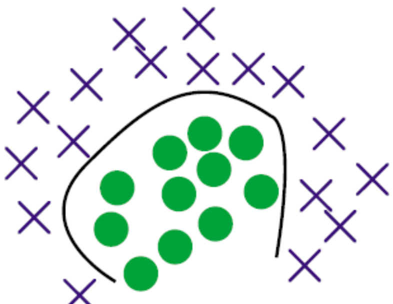
good

bad

- Types of patterns



Linear Classifier



Nonlinear Classifier

Process of ML

Dataset /Cleaning	Feature Manipulation	Classification Regression	Evaluation

feature

instance

? ? ? ? ?

Data matrix
(row: instance)
(col: feature)

Fill missing values

Feature Selection

PCA

normalization

standardization

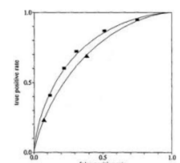
SVM

Decision Tree

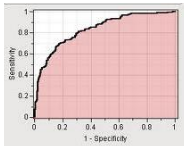
Neural Networks

Accuracy + Cross-validation

ROC Curve



AUC



Process of ML in Material Science

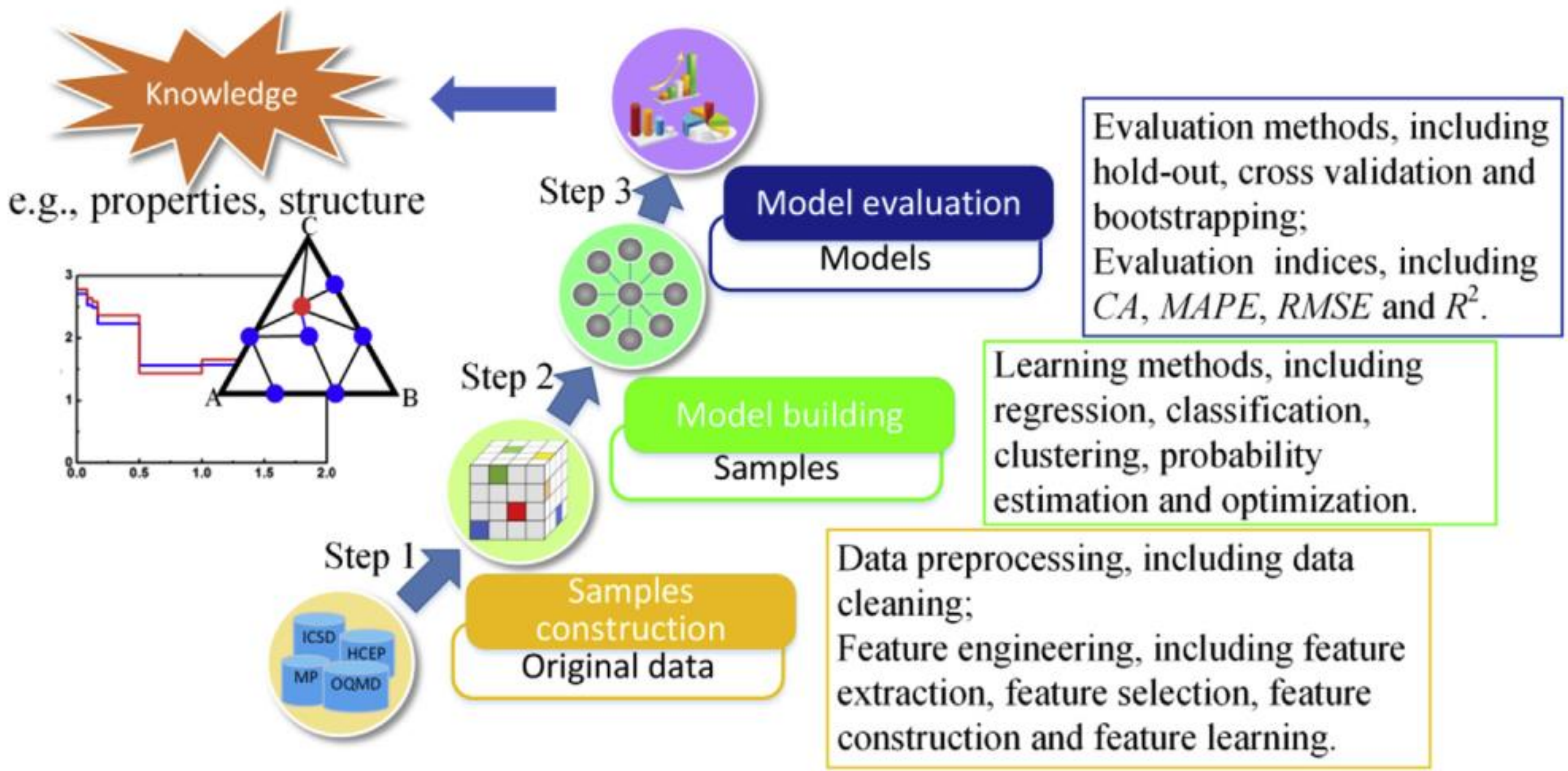
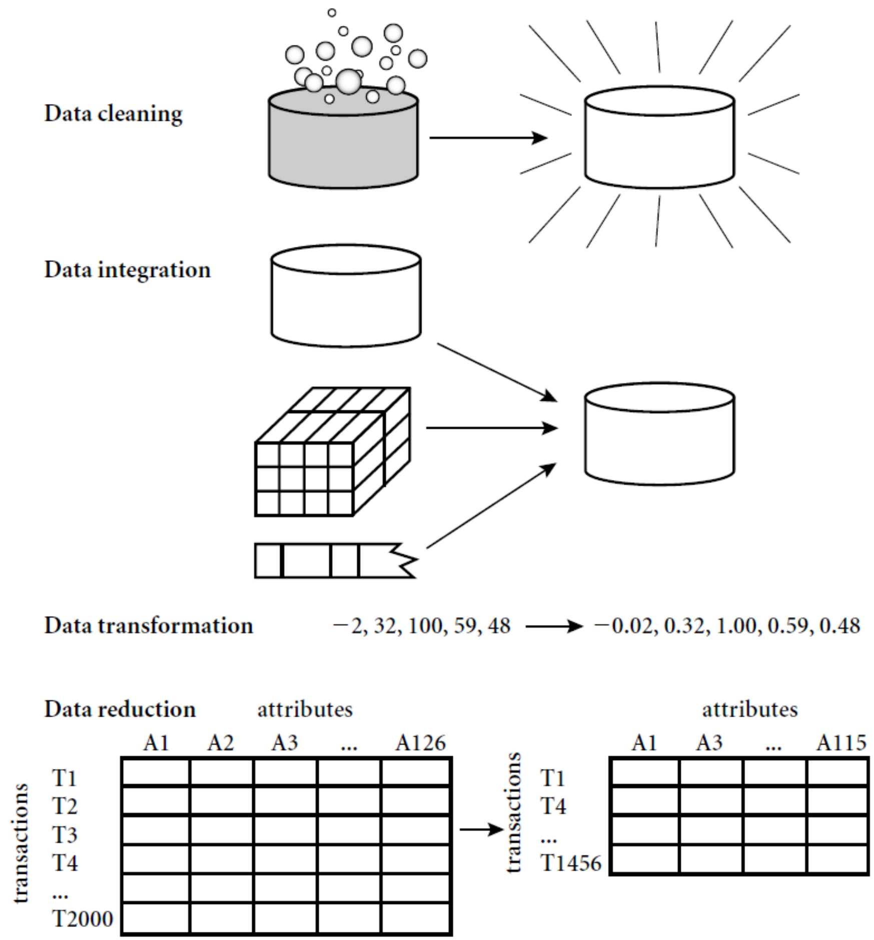


Fig. 2. The general process of machine learning in materials science.

Y. Liu et al. (2017 J. Materiomics)

Data Preprocessing

- Improve the data quality before the prediction model
- Data cleaning
  - incomplete, noisy and inconsistent
  - Fill in missing values
  - Identify or remove outliers
  - Noise filtering
- Data reduction
  - Remove factors, not relevant to the decision attributes
  - Dimensionality reduction
- Data transformation
  - Normalization



Data Cleaning

- **Data in the Real World Is Dirty**: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = " " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary* = "−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age* = "42", *Birthday* = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- The curse of dimensionality

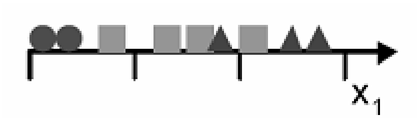
  - A term coined by Bellman in 1961

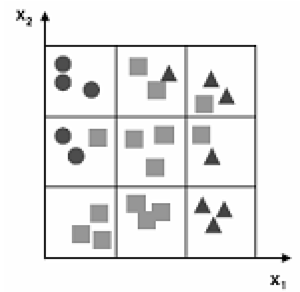
  - Refers to the problems associated with multivariate data analysis as the dimensionality increases

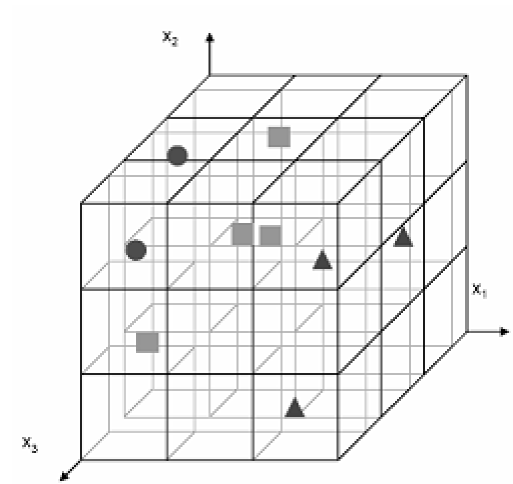
- When the dimensionality increases, the volume of the space increases so fast that the available data become sparse.

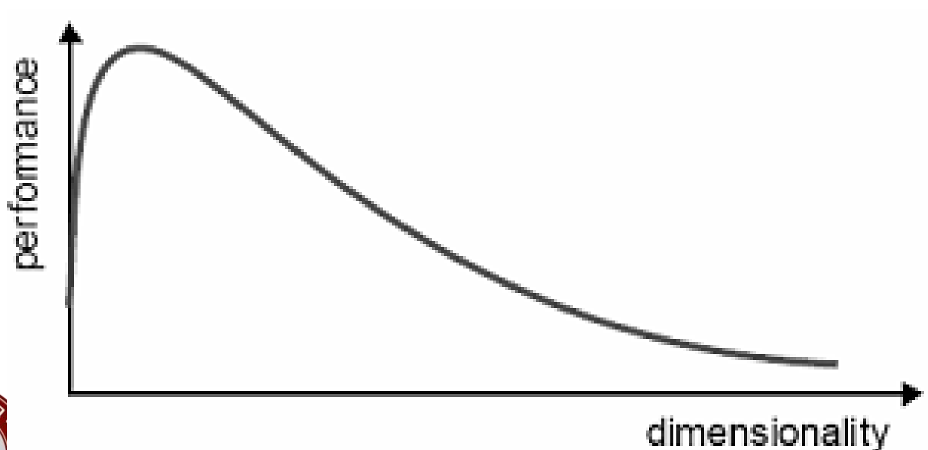
  - This sparsity is problematic for any method that requires statistical significance

Curse of Dimensionality









Data Transformation

- A function that **maps the entire set of values of a given attribute to a new set of replacement values**
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization (standardization)
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

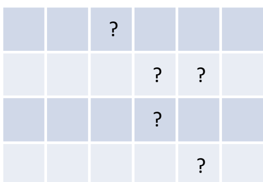
Process of ML

Dataset /Cleaning → Feature Manipulation → Classification Regression → Evaluation

Dataset /Cleaning

feature

instance



Data matrix
(row: instance)
(col: feature)

Fill missing values

Feature Manipulation

Feature Selection

PCA

normalization

standardization

Classification Regression

SVM

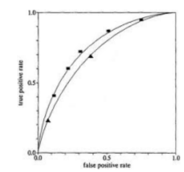
Decision Tree

Neural Networks

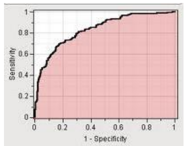
Evaluation

Accuracy + Cross-validation

ROC Curve



AUC



Model Building

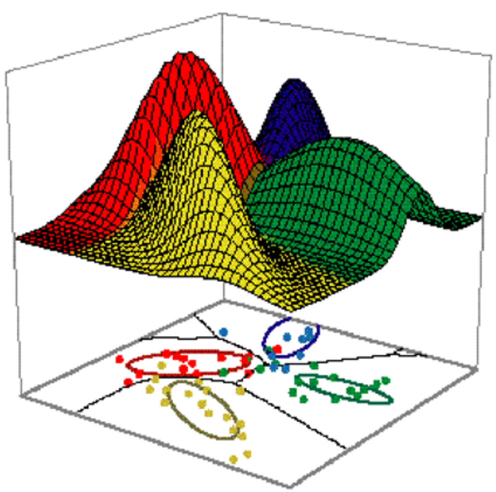
- Types of problems
  - classification
    - Derive specific numbers for the classes
  - regression
    - Generalization of classification

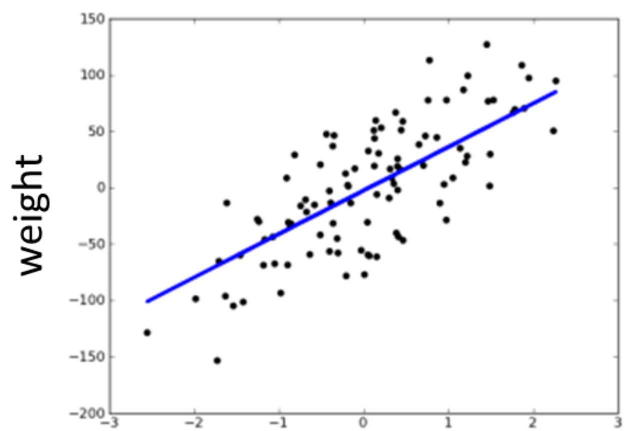
classification

Y: Classes,
Discrete Output

regression

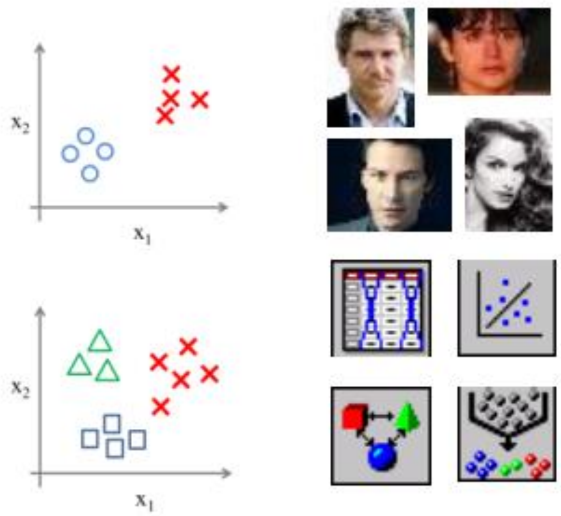
Y: Real Number,
Continuous Output



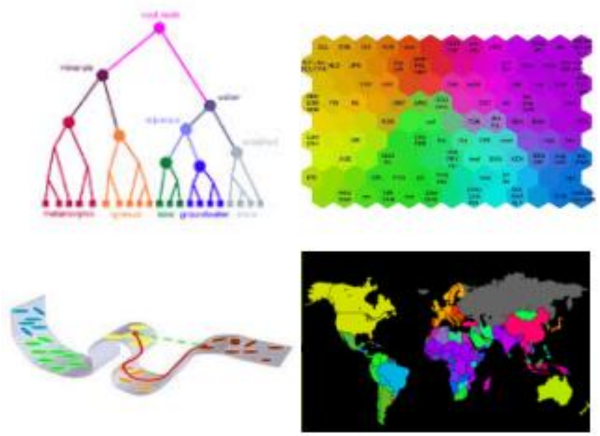


Model Building

- Supervised learning (ex: classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set



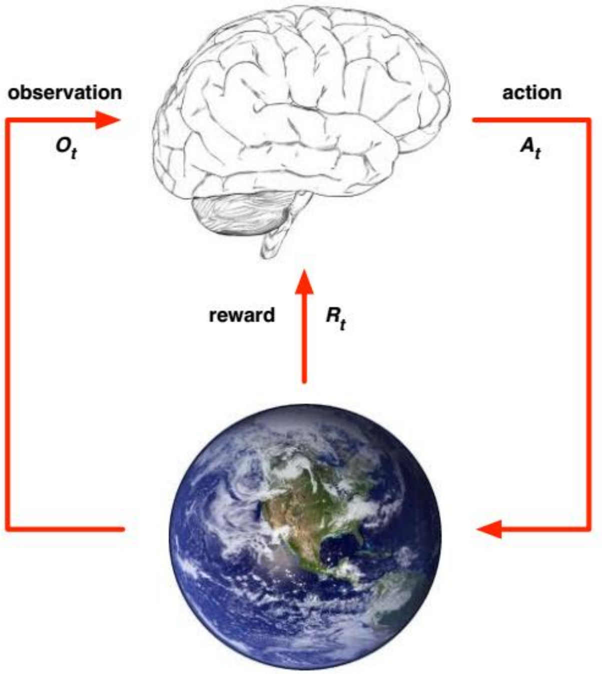
- Unsupervised learning (ex: clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data



Model Building

- Reinforcement Learning
  - No supervisor, only a reward signal
  - Feedback is delayed, not instantaneous
  - Sequential, non independent-and-identical distributed data
  - Examples of Reinforcement Learning

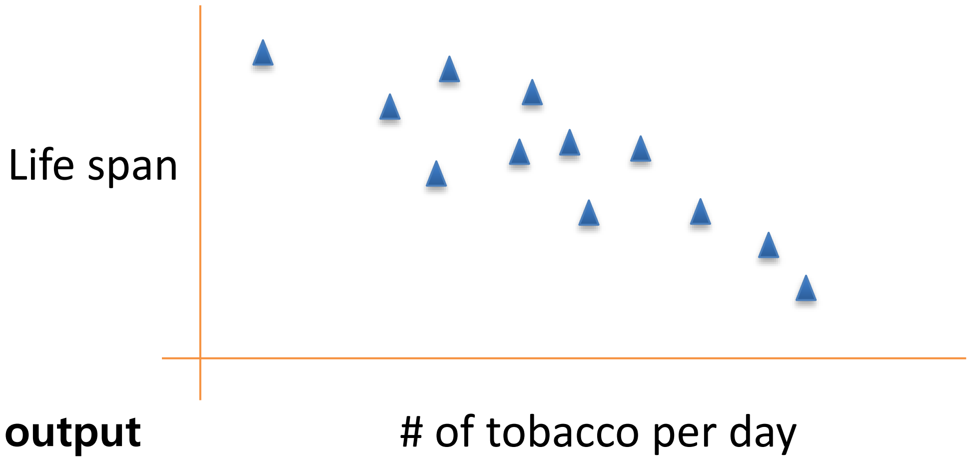
https://youtu.be/2pWv7GOvuf0?list=PL7-jPKtc4r78-wCZcQn5IqyuWhBZ8fOxT&t=935



- At each step t the agent:
  - Executes action A_t
  - Receives observation O_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation O_{t+1}
  - Emits scalar reward R_{t+1}
- t increments at env. step

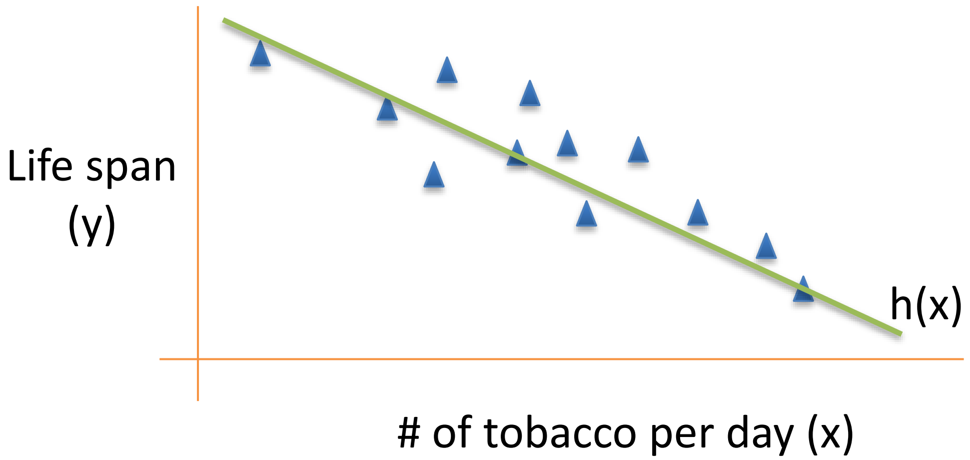
One Variable Linear Regression

- **Smoking vs lifespan**



- **Supervised Learning**
  - **Classification : discrete output**
  - **Regression : continuous value output**

One Variable Linear Regression



- Linear regression with one variable

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Cost function of Linear Regression

- Parameter estimation in linear regression problem
- Hypothesis : $h_\theta(x) = \ \theta_0 + \theta_1 x$

  , where $\theta_i \ (i = 0 \ and \ 1)$ are parameters
- Let’s find the parameters, θ_i
- Cost function

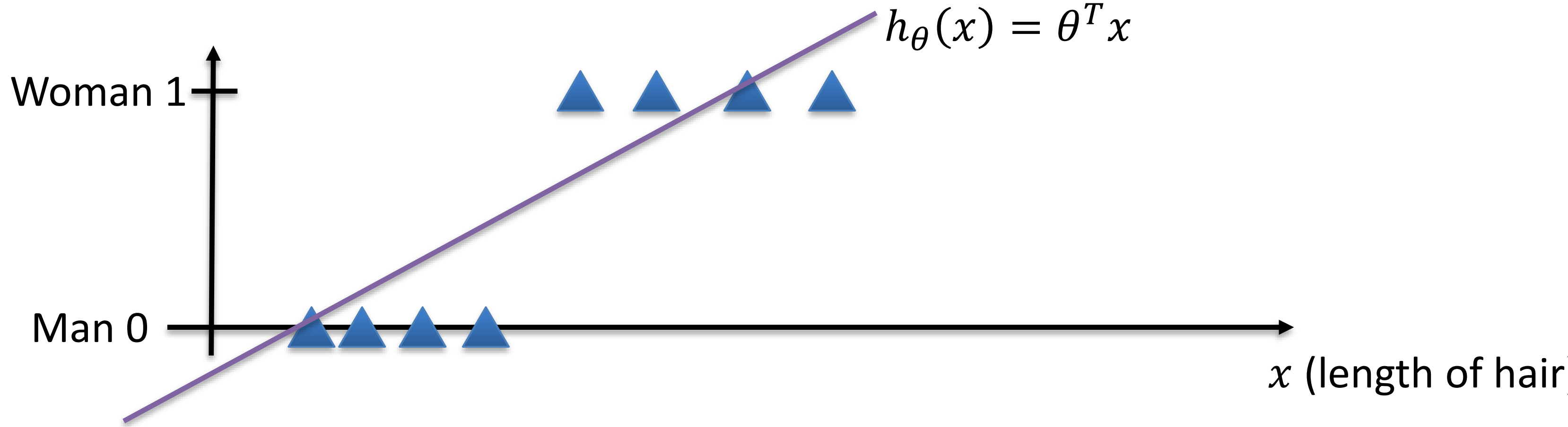
$$J(\theta_0, \theta_1) = \frac{1}{2n} \sum_{k=1}^{n} \left(h_\theta\left(x^{(k)}\right) - y^{(k)}\right)^2$$

- To estimate $\theta_0 \ and \theta_1$, *minimize J(*θ_0*,* θ_1*)*

$$\underset{\theta_0, \theta_1}{minimize} \ J(\theta_0, \theta_1)$$

Classification

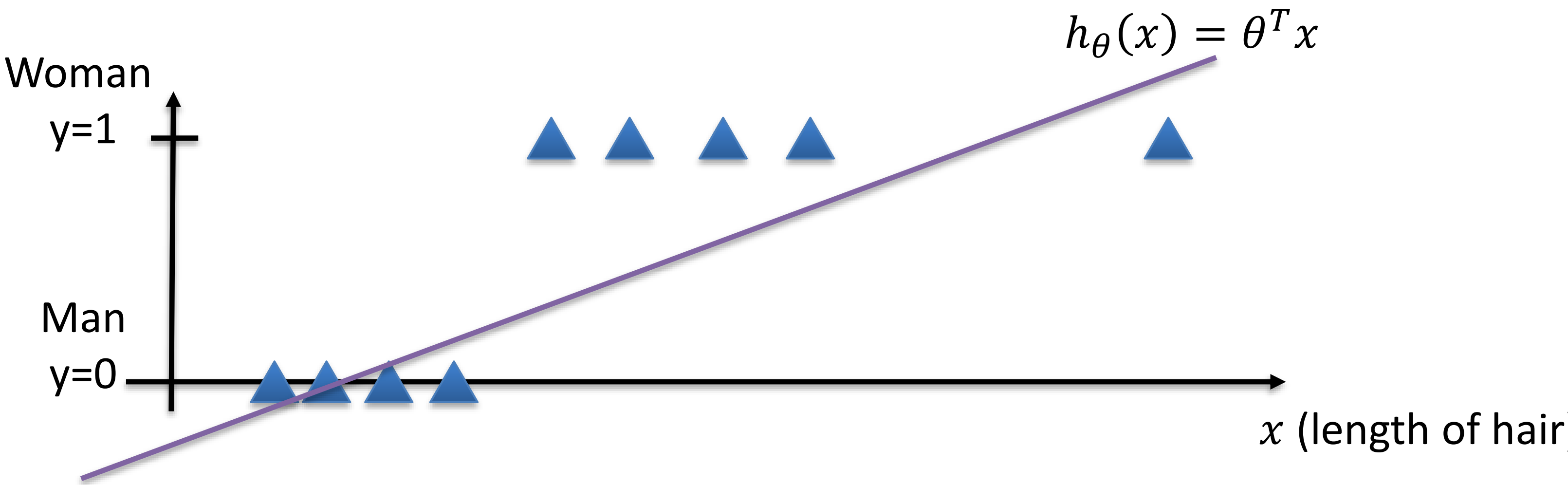
- Classification problem using regression



- Classification using threshold $h_\theta(x)$ = 0.5
  - If $h_\theta(x) \geq 0.5$, then classify the sample into Woman (y=1)
  - If $h_\theta(x) < \ 0.5$, then classify the sample into Man (y=0)

Classification

- Outlier



- Linear regression output $h_\theta(x)$ can be >1 or <0 even though y=0 or 1

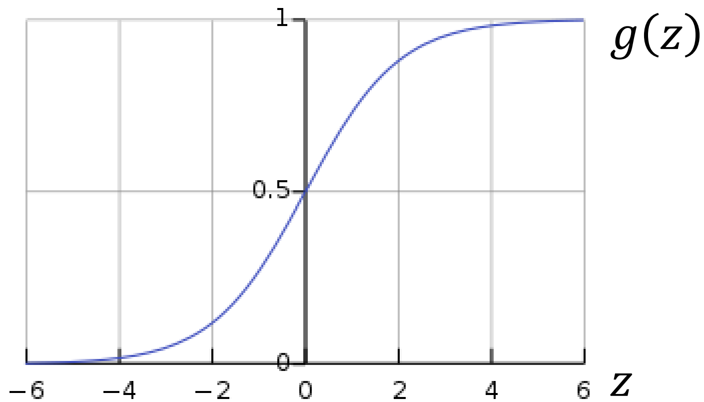
  ➔ Logistic Regression producing $0 \le h_\theta(x) \le 1$

Logistic Regression

- Goal : $0 \le h_\theta(x) \le 1$
- Linear regression $h_\theta(\mathbf{x}) = \boldsymbol{\theta}^T\mathbf{x} \rightarrow \mathrm{g}\left(\boldsymbol{\theta}^T\mathbf{x}\right)$,

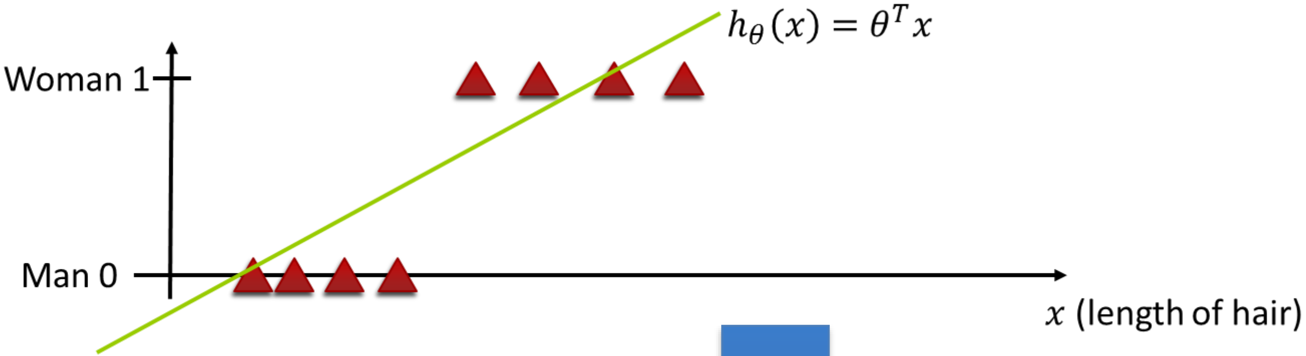
  , $where\ g(z) = \frac{1}{1+e^{-z}}$ (Sigmoid function or Logistic function)
- Logistic Regression Model

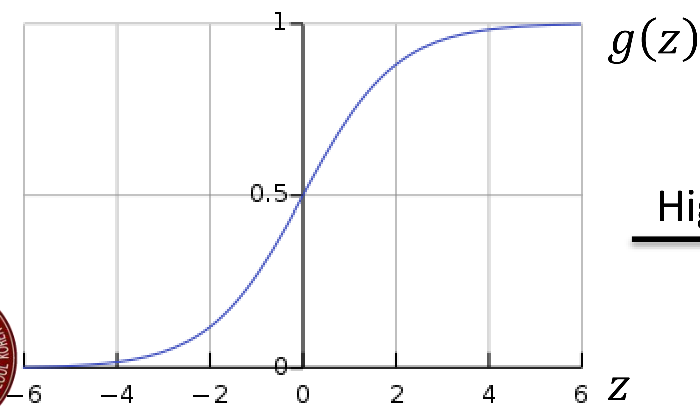
$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$



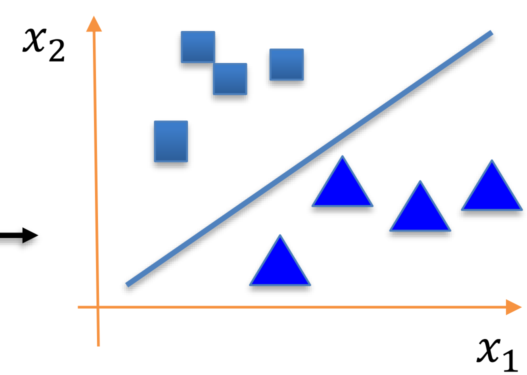
Logistic Regression

- Transform the linear regression problem to binary 0/1 classification problem





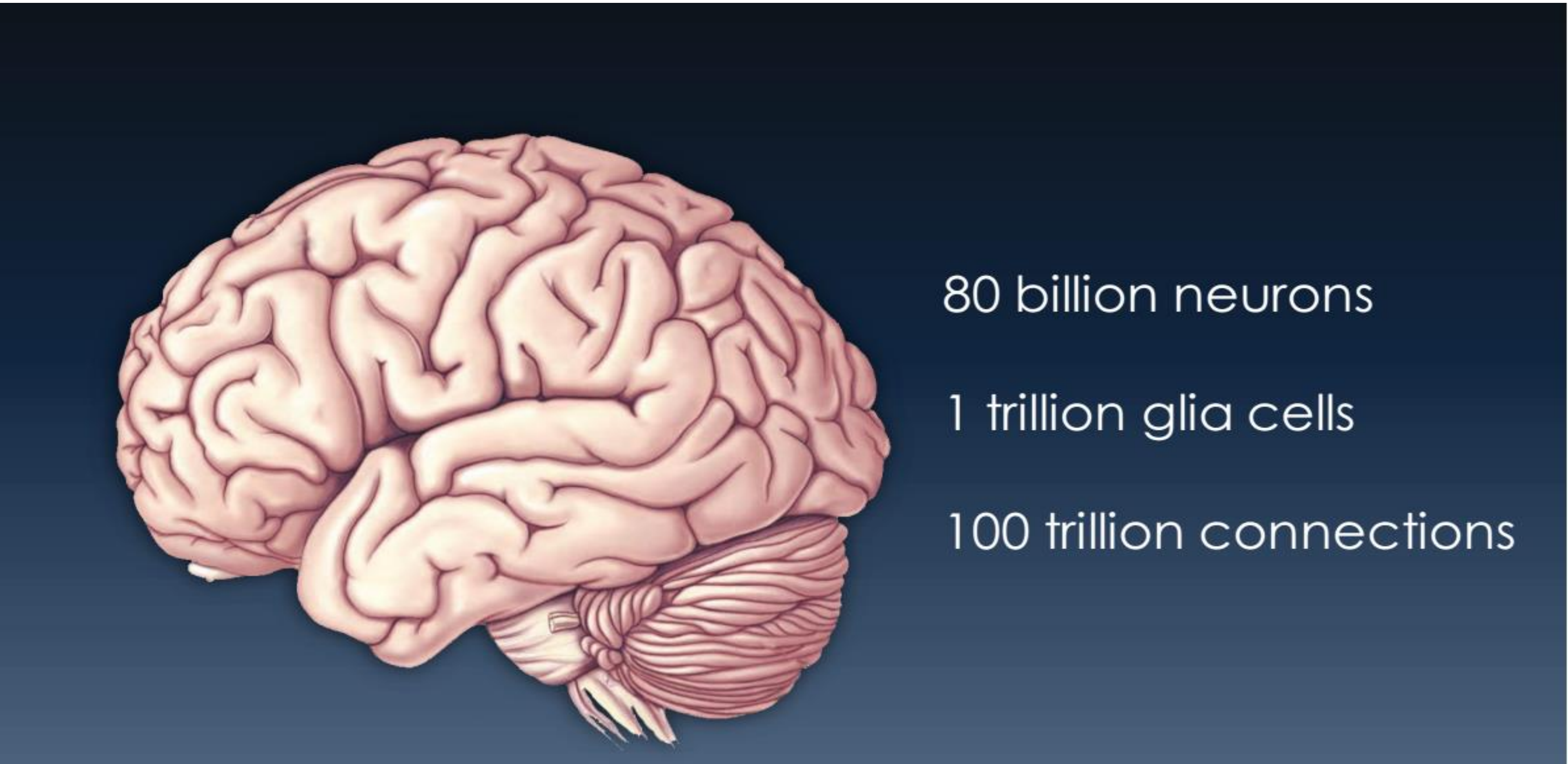
Higher dimension expression



Artificial Neural Network

- Started by psychologists and neurobiologists to develop and test computational analogues of neurons
- A neural network: A set of connected input/output units where each connection has a weight associated with it
- During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples
- Also referred to as connectionist learning due to the connections between units

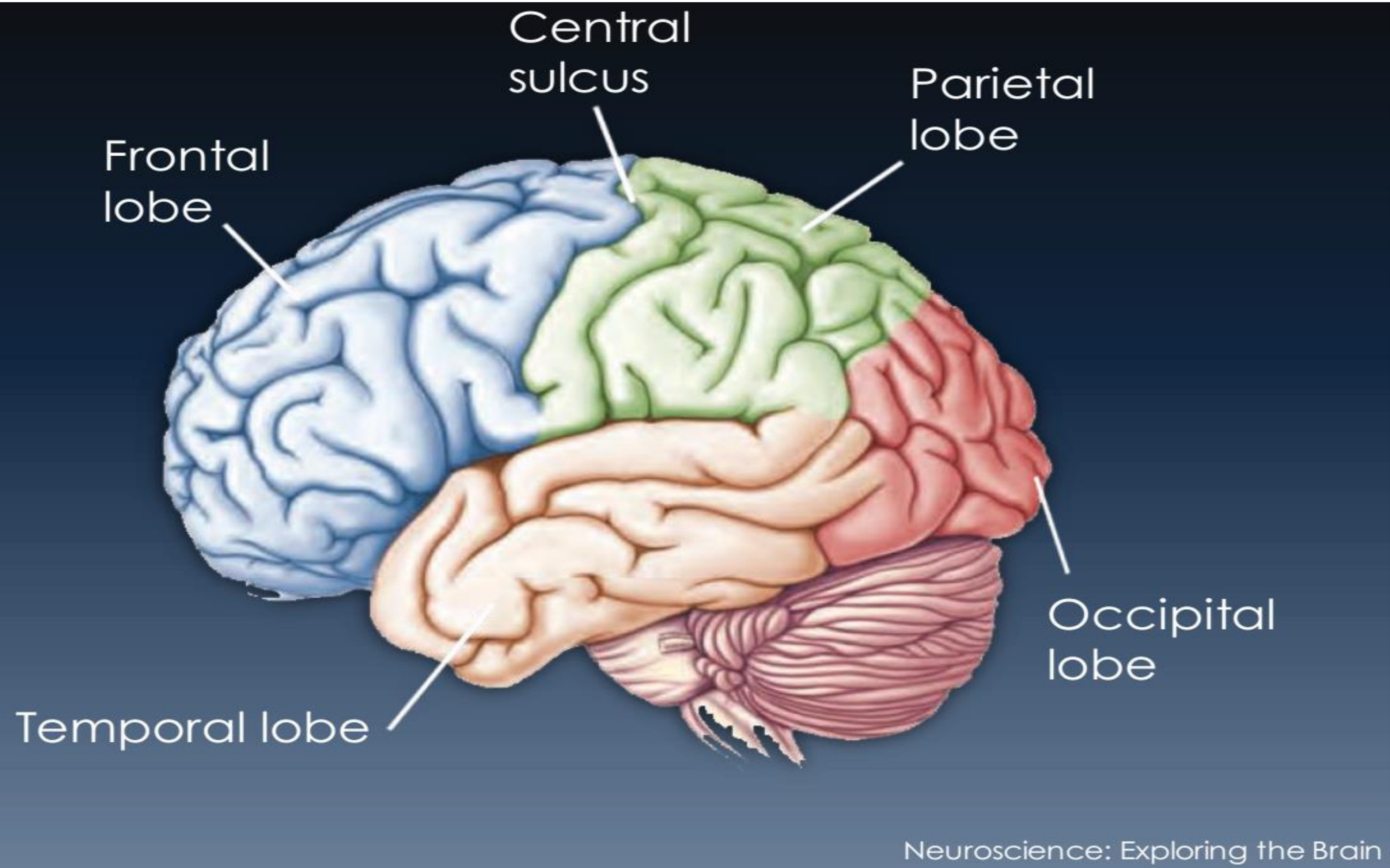
Brain and Neuron



- Much more efficient than computer in terms of learning
- Think about object recognition
- That’s why people would like to mimic the brain to develop software (ANN) and hardware (Neuromorphic)

Neuroscience: Exploring the Brain

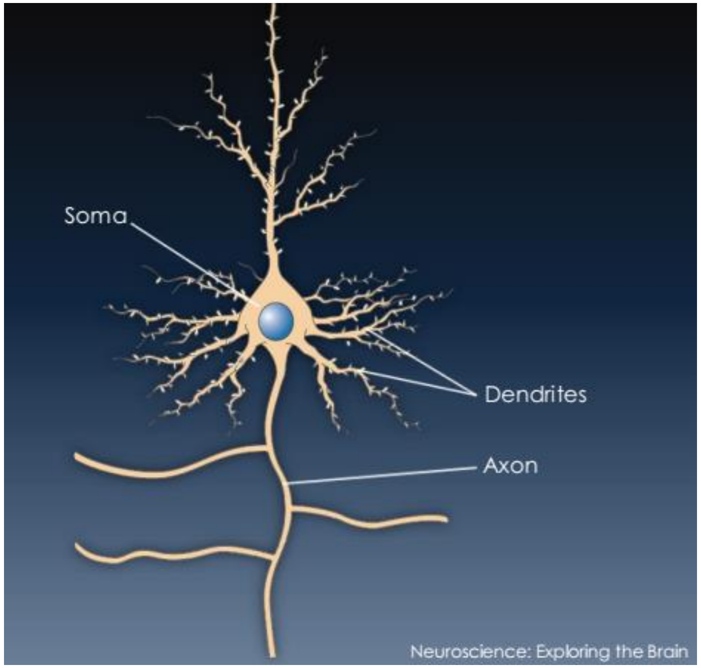
Brain and Neuron



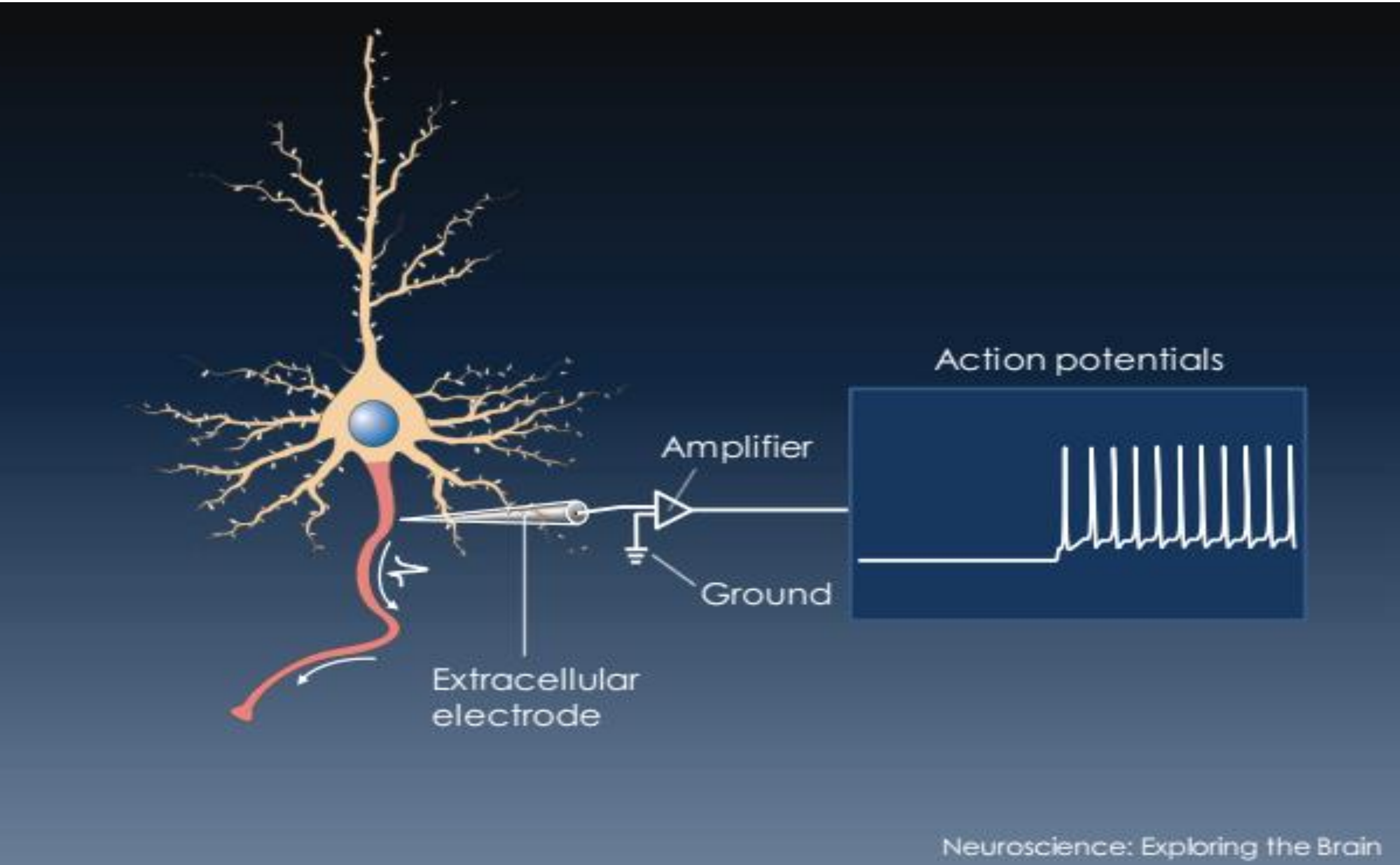
Brain and Neuron

- Neuron
  - Brain consists of neurons
  - Plasticity of neuron
  - Restructuring by learning

- Components of neuron
  - Dendrite : input channel of neural excitation from the next neurons
  - Axon : transferring neural excitation
  - Cell body or soma : accumulating the input signal and transferring the signal to its axon



Brain and Neuron



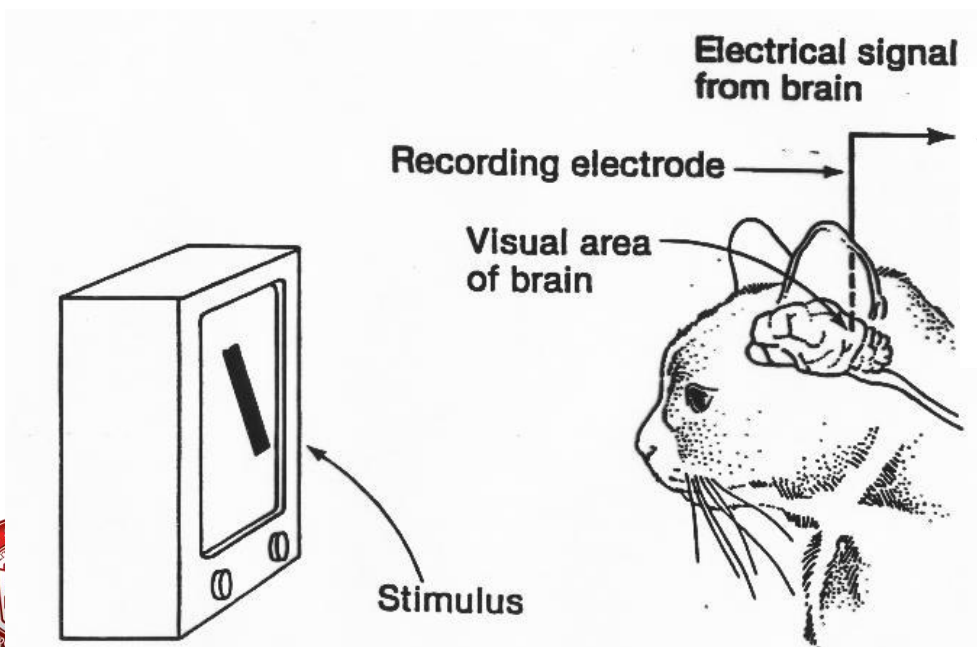
Hubel and Wiesel experiment

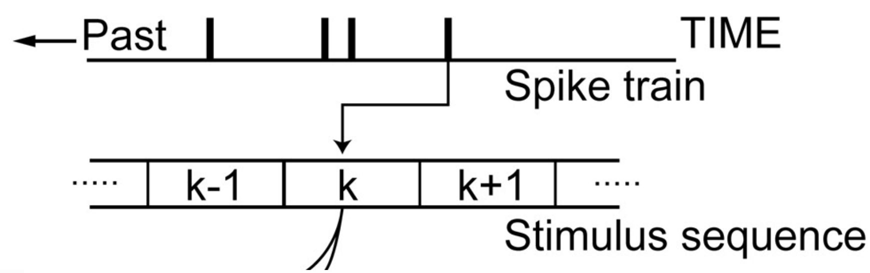
- Found the principles of the visual perception
- Nobel Prize in Physiology and Medicine in 1981
- https://youtu.be/IOHayh06LJ4
- https://youtu.be/Cw5PKV9Rj3o

Hubel and Wiesel experiment

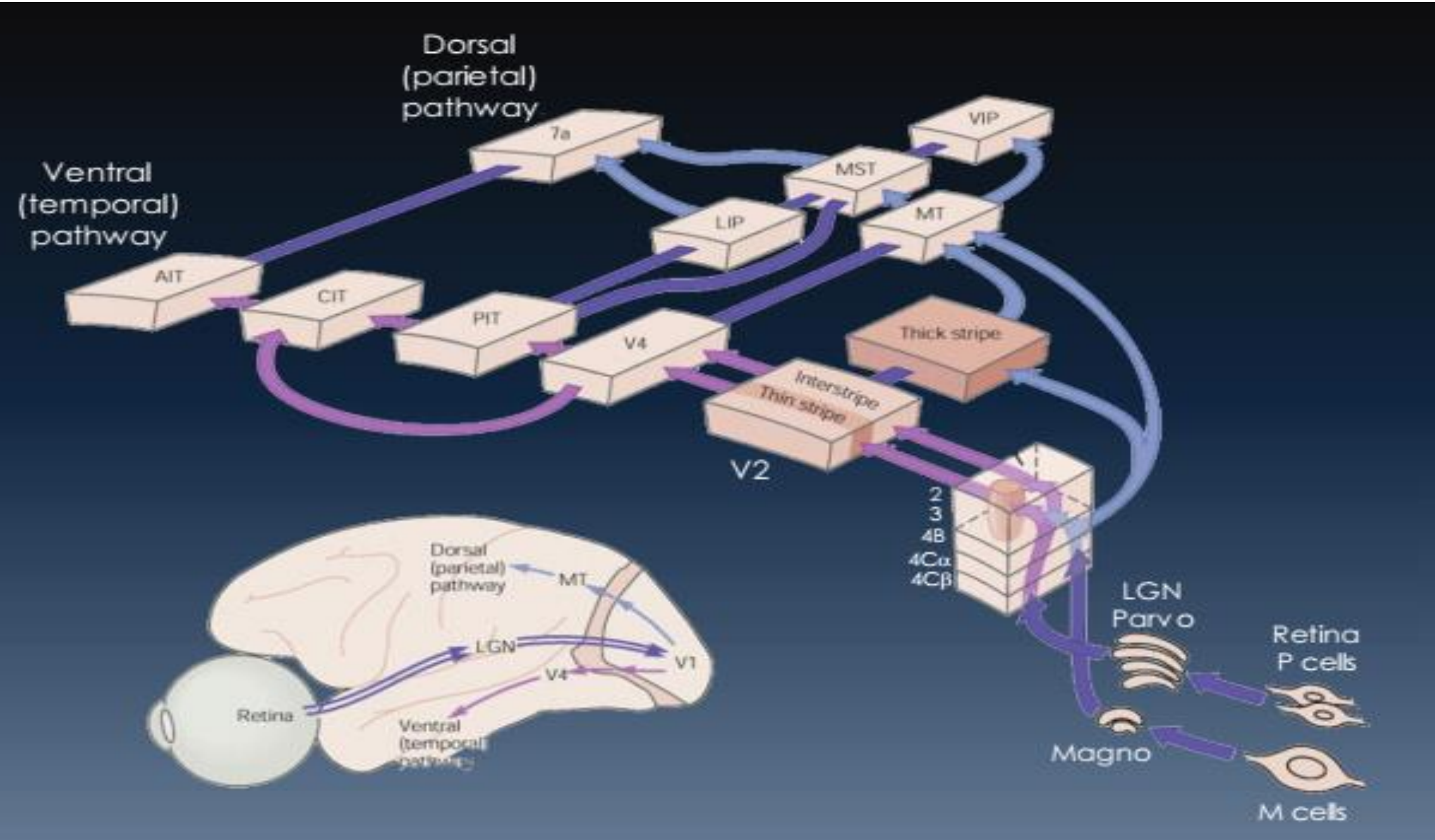
- The neurons fired only when the line was in a particular place on the retina
- The activity of these neurons changed depending on the orientation of the line
- Sometimes the neurons fired only when the line was moving in a particular direction.

“There has been a myth that the brain cannot understand itself. …
The brain can be studied just as the kidney can.”



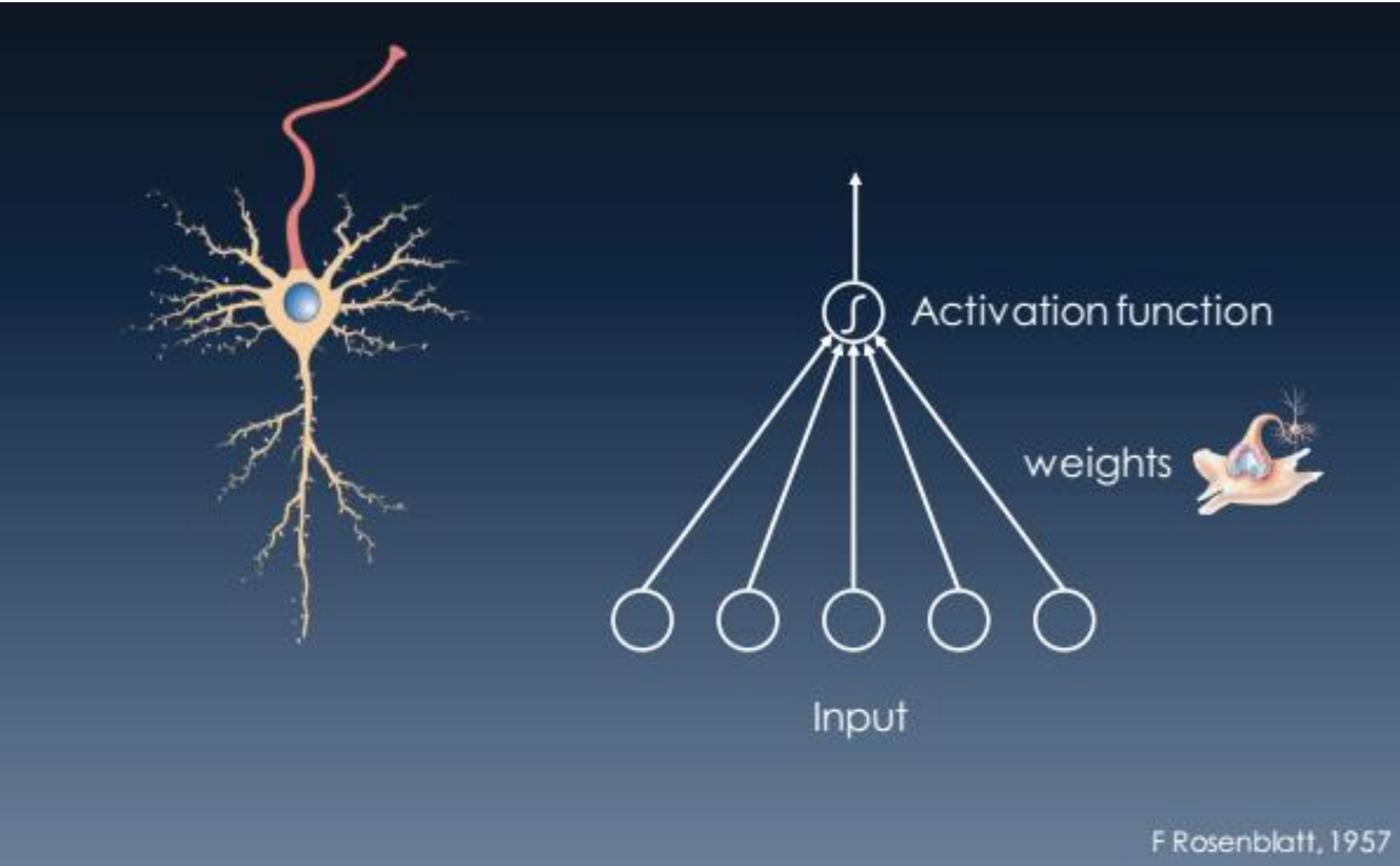


Visual System



Principles of Neural Science

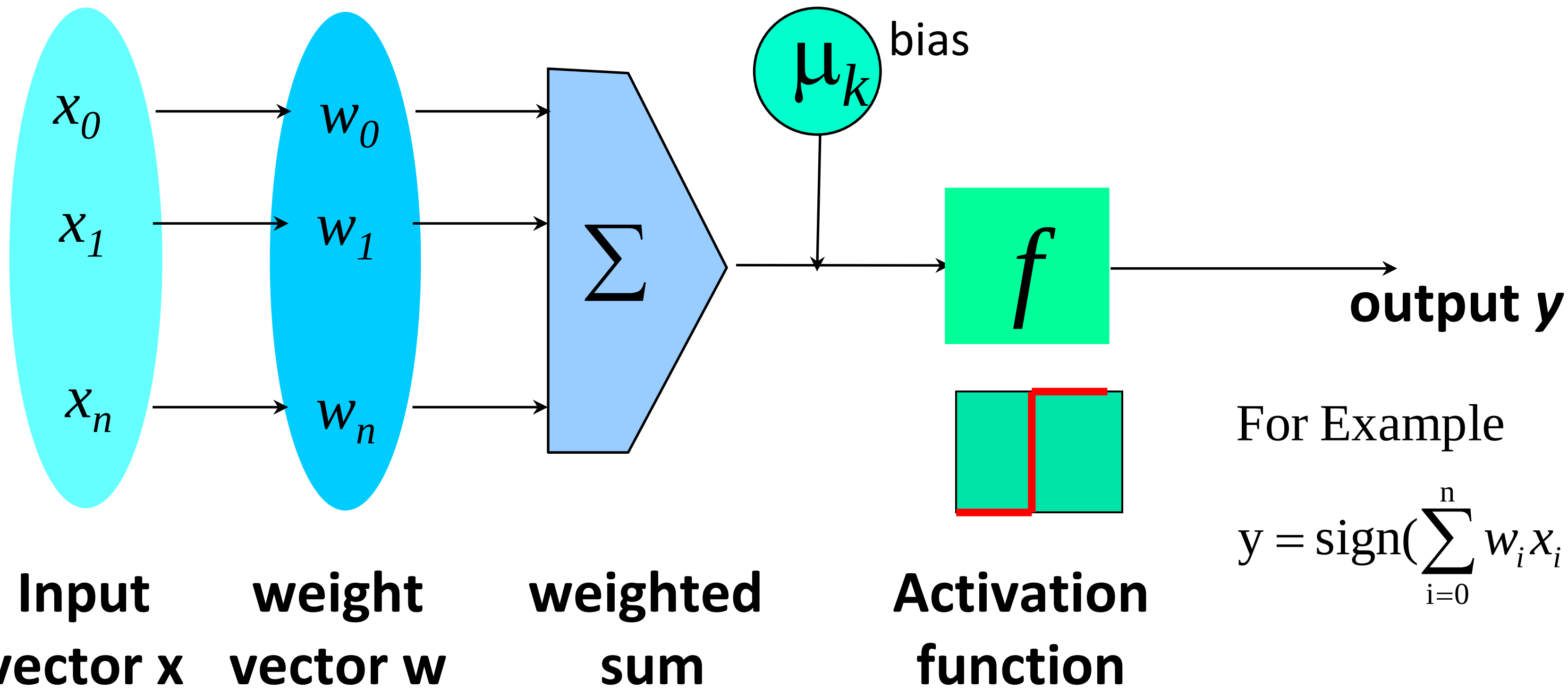
Artificial Neural Network



Artificial Neural Network

- McCulloch-Pitts Model
  - McCulloch : A psychiatrist and neuroanatomist
  - Pitts : A mathematician
  - A logical calculus of ideas immanent in nervous activity, 1943
  - 3 years before the first computer ANIAC (1946)
  - Demonstrated the logic gates like 'AND', 'OR', 'NOT', and ect. using artificial neural network

Artificial Neural Network

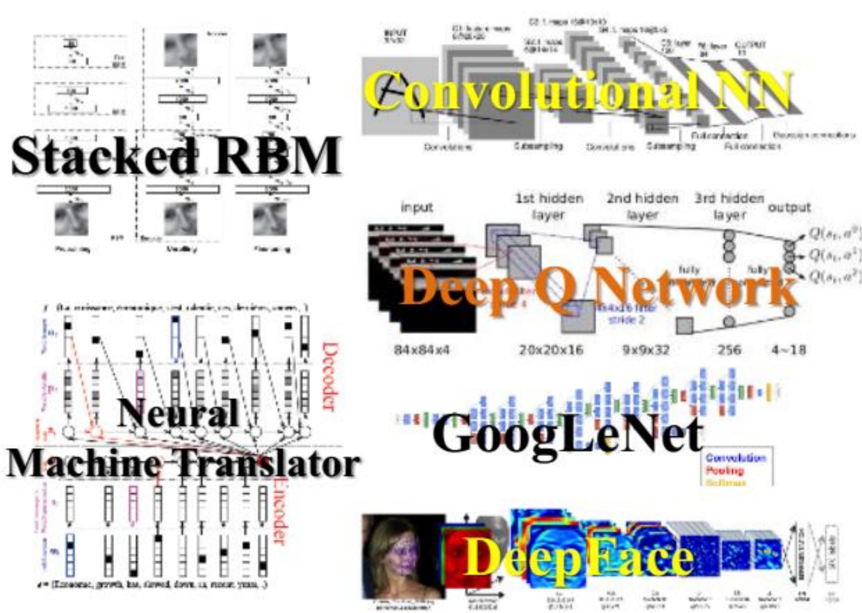


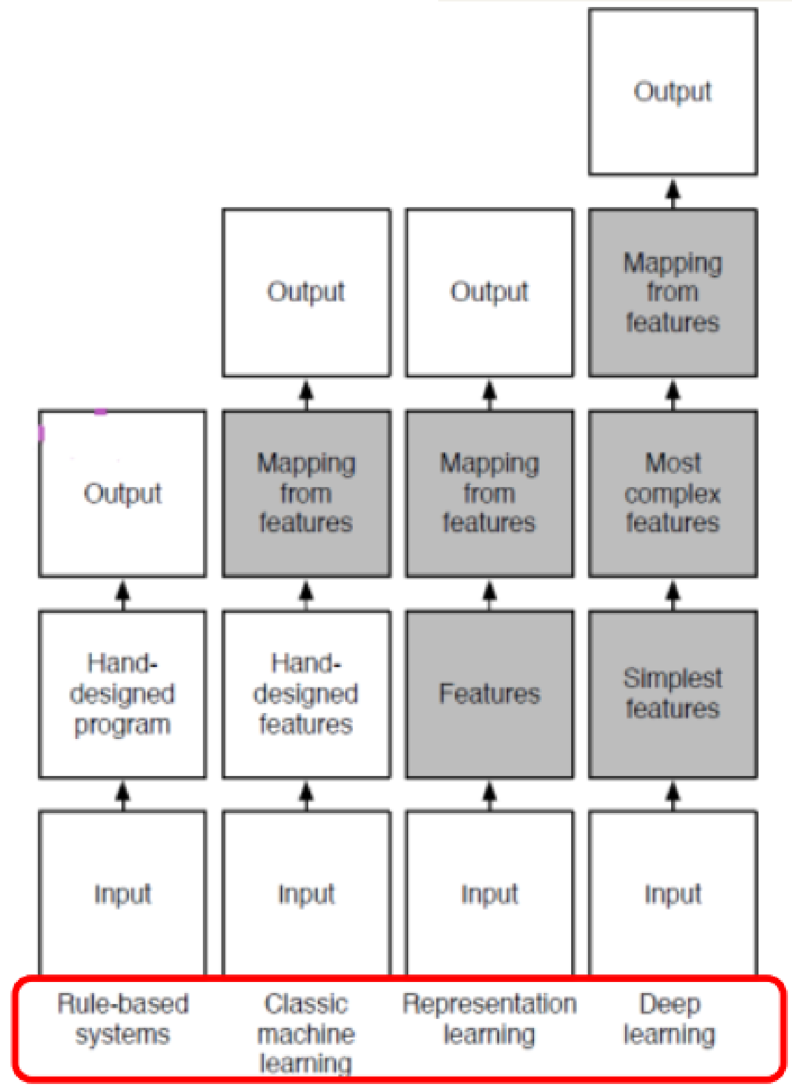
For Example

$$y = \text{sign}\left(\sum_{i=0}^{n} w_i x_i - \mu_k\right)$$

- An *n*-dimensional input vector **x** is mapped into variable y by means of the scalar product and a nonlinear function mapping
- The inputs to unit are outputs from the previous layer. They are multiplied by their corresponding weights to form a weighted sum, which is added to the bias associated with unit. Then a nonlinear activation function is applied to it.

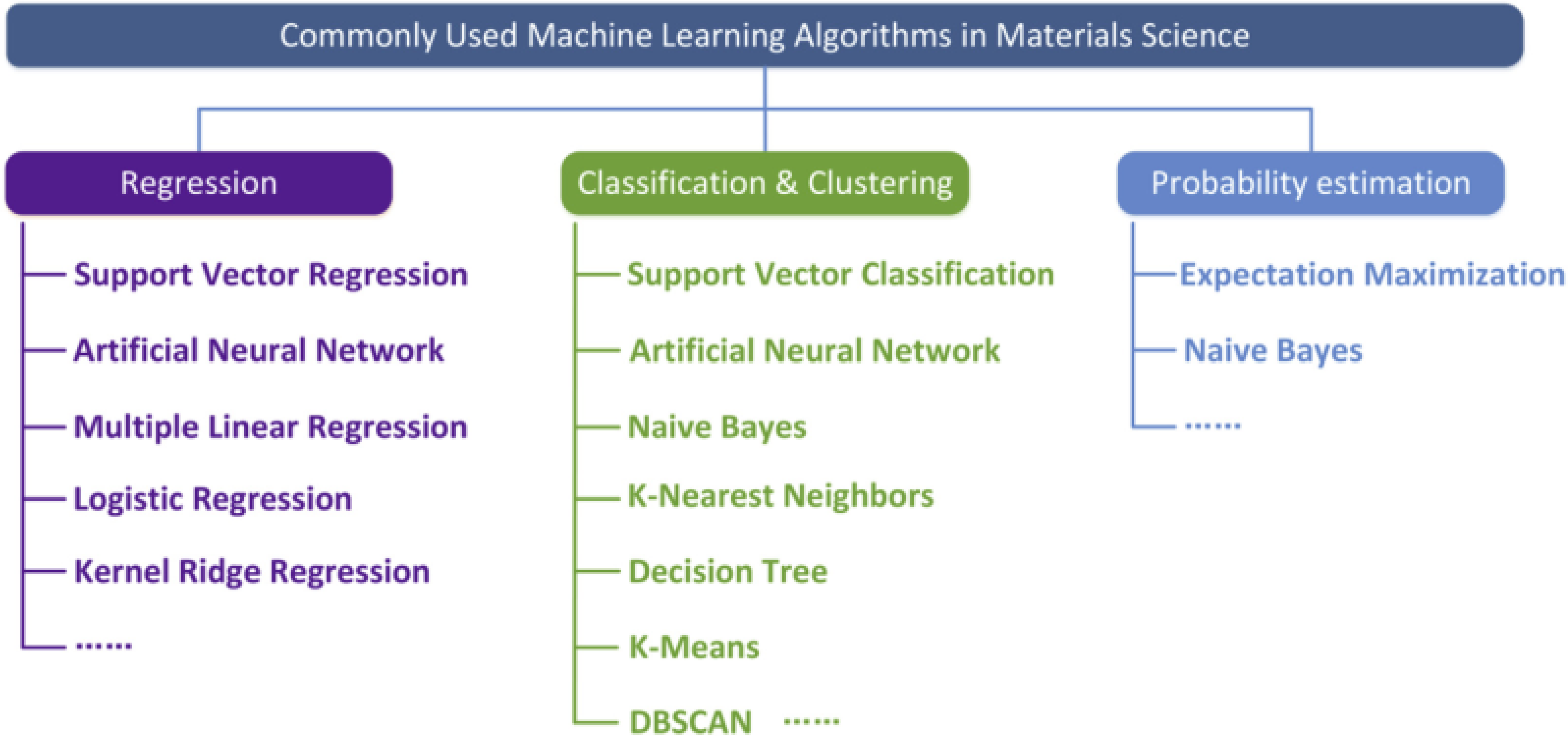
Deep Learning





White box: human controlled
Gray box: machine controlled

Commonly Used ML



Y. Liu et al. (2017 J. Materiomics)

Process of ML

Dataset /Cleaning → **Feature Manipulation** → **Classification Regression** → **Evaluation**

feature

instance

Data matrix
(row: instance)
(col: feature)

Fill missing values

Feature Selection

PCA

normalization

standardization

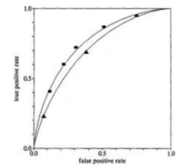
SVM

Decision Tree

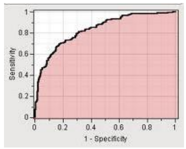
Neural Networks

Accuracy + Cross-validation

ROC Curve



AUC



Model Evaluation

- Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Recall Rate $= \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$

Precision $= \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$

True Positive Rate (TPR) $= \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$

False Positive Rate (FPR) $= \dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$

True Negative Rate (TPR) $= \dfrac{\boldsymbol{TN}}{\boldsymbol{FP}+\boldsymbol{TN}}$

Model Evaluation

- Sensitivity = Recall Rate = True Positive Rate

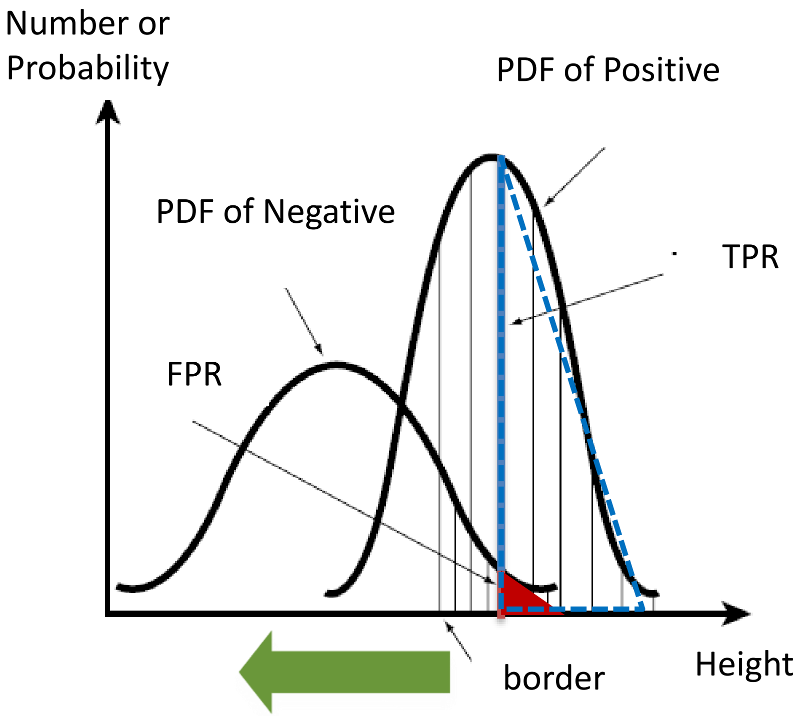
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

- Specificity = True Negative Rate

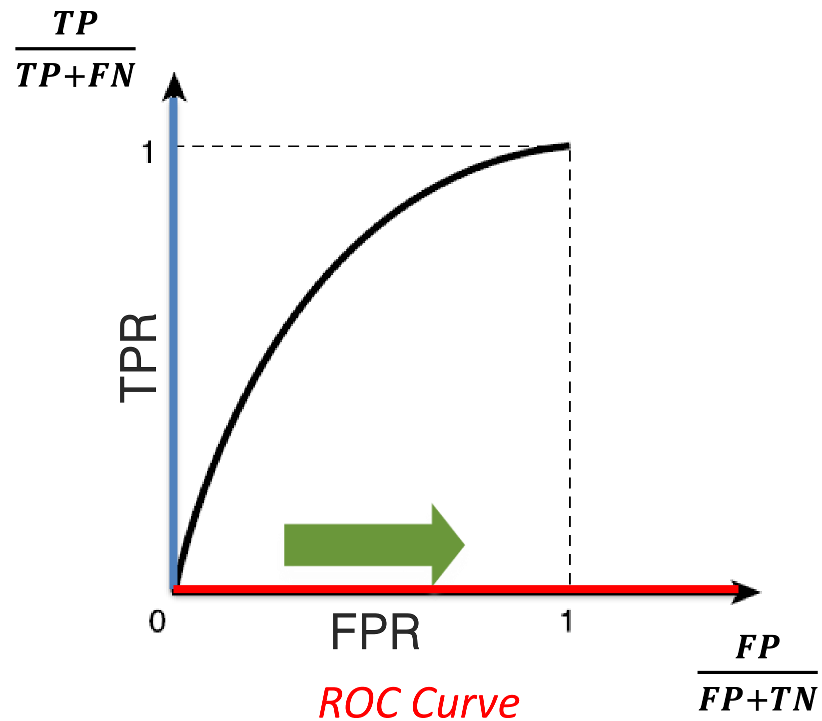
$$\frac{TN}{FP+TN}$$

Model Evaluation

- **Receiver Operating Charateristic (ROC) Curve**



PDF of Positive and Negative



ROC Curve

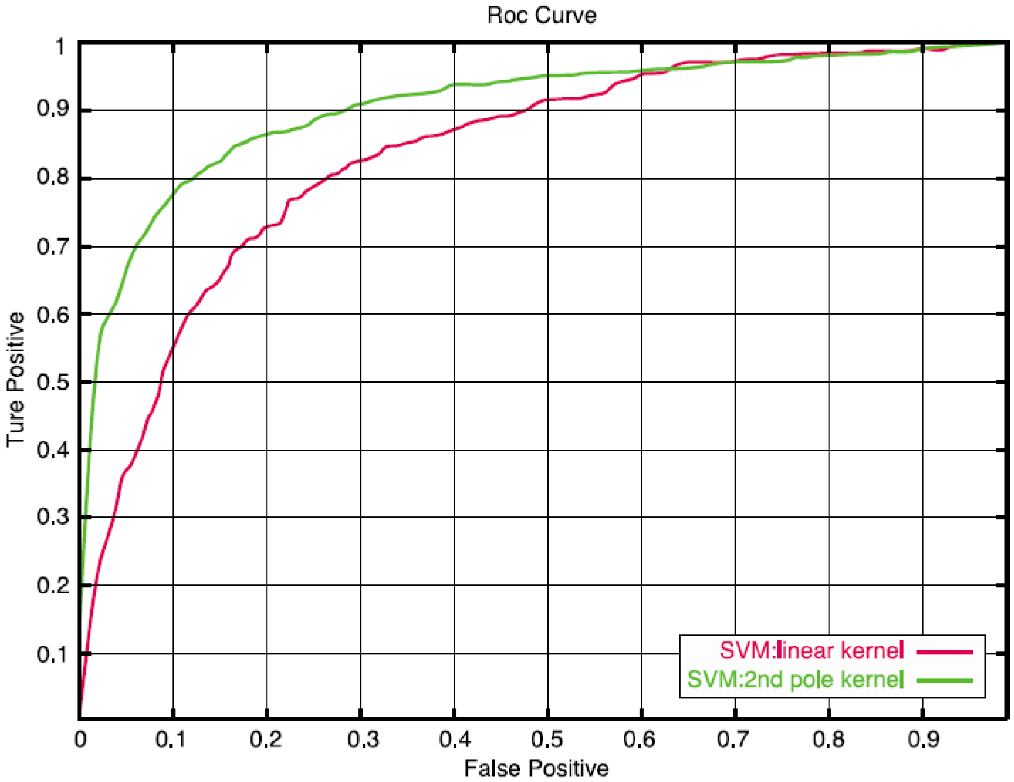
	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

* Example of recruiting Basketball players among good and bad players

Model Evaluation

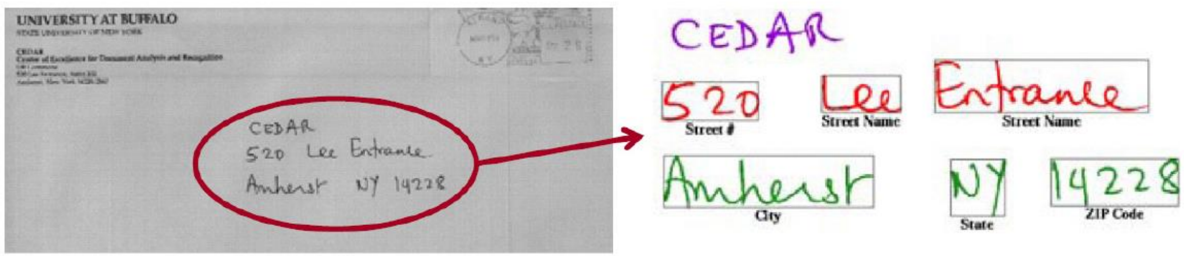
– AUROC

- Area under the ROC curve

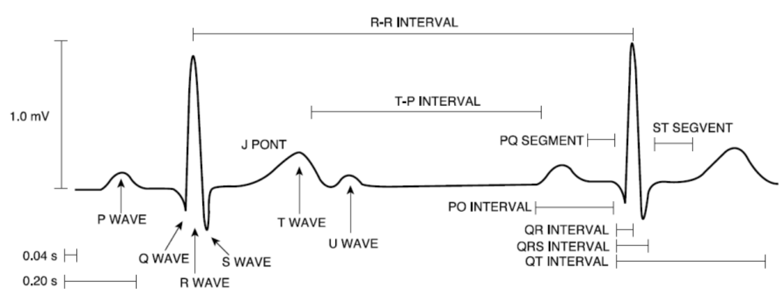


08_Applications

- Applications of pattern recognition
  - Recognition of letters
    - Automatic classification of letters, Producing code and text from an image scanning, ATM machine, recognition of car registration number



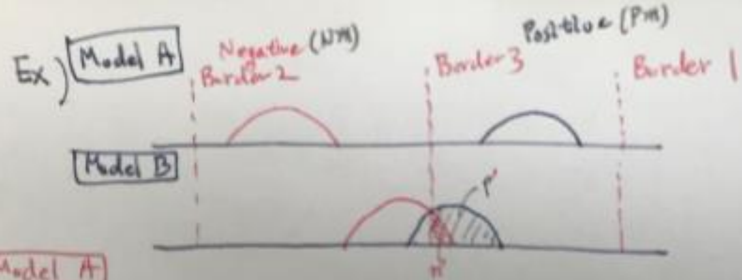
  - Pattern recognition of Bio-physiological signal and Behavior
    - Pattern recognition of speech, finger print, pupil, face, DNA mapping, gait, etc.
    - Voice/finger print/iris/face/walking pattern recognition, DNA mapping, etc
  - Diagnosis
    - Car fault, Clinics, Signal processing of electroencephalogram and electrocardiogram,



Reseach Scopes

Research Scopes	Applications
Adaptive Signal Processing	Image Processing
Machine Learning	Computer Vision
Artificial Neural Network	Audio/Video Recognition
Robotics/Vision	Automatic Target Recognition
Cognitive Science	Optical Letter Recognition
Statistics	Seismic Analysis
Nonlinear Optimization	Chat bot
Data Analytics	Biometric Authentication
Fuzzy/Genetic System	Fintech
Inference/Decision Theory	Medical Diagnosis
Structural Modeling	Etc…
Computational Neuroscience	
Etc…	

Example



Model A

Border 1 $TPR = \frac{0}{P} = 0 \quad FPR = \frac{0}{N} = 0$

~~Model B~~

Border 2 $TPR = \frac{P}{P} = 1 \quad FPR = \frac{N}{N} = 1$

~~Model B~~

Border 3 $TPR = \frac{P}{P} = 1 \quad FPR = \frac{0}{N} = 0$

Model B

Border 1 $TPR = \frac{0}{P} = 0 \quad FPR = \frac{0}{N} = 0$

Border 2 $TPR = \frac{P}{P} = 1 \quad FPR = \frac{N}{N} = 1$

Border 3 $TPR = \frac{p'}{P} \quad FPR = \frac{n'}{N}$

Model C

Border 1 $TPR = \frac{0}{P} = 0, \ FPR = \frac{0}{N} = 0$

Border 2 $TPR = \frac{P}{P} = 1, \ FPR = \frac{N}{N} = 1$

Border 3 $TPR = \frac{p'}{P} = FPR = \frac{n'}{N}$