



# Regression Techniques



# Table of Contents

- What is Regression Analysis?
- Why do we use Regression Analysis?
- What are the types of Regressions?
  - Linear Regression
  - Logistic Regression
  - Polynomial Regression
  - Stepwise Regression
  - Ridge Regression
  - Lasso Regression
  - ElasticNet Regression
- How to select the right Regression Model?

# What is Regression Analysis?

- Regression analysis is a form of **predictive modelling technique** which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the **causal effect relationship** between the variables. For **example**, relationship between rash driving and number of road accidents by a driver is best studied through regression.



- Regression analysis is an important tool for **modelling** and **analyzing data**. Here, we **fit a curve / line to the data points**, in such a manner that the differences between the **distances of data points from the curve or line is minimized**.

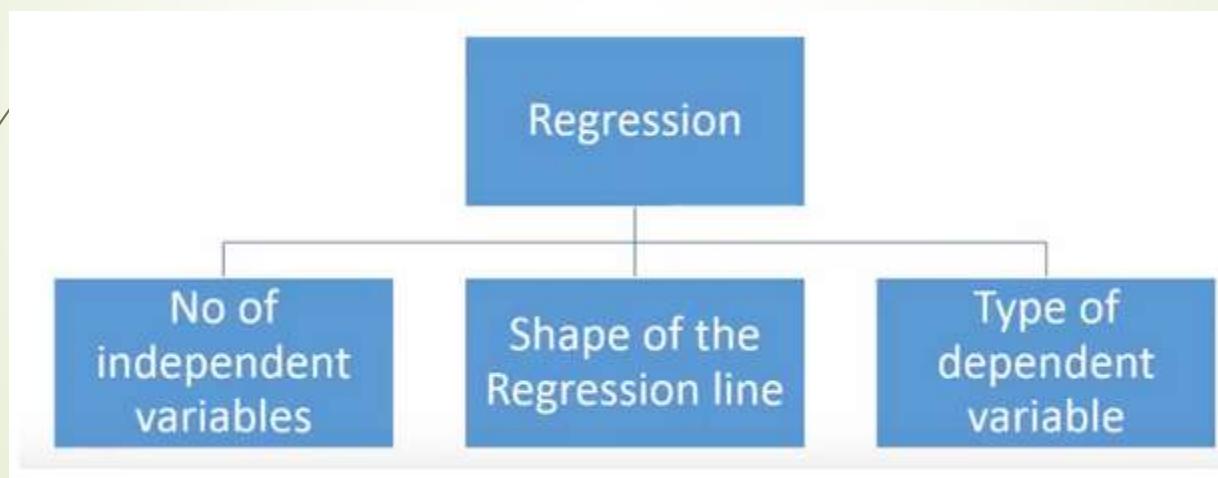


# Why do we use Regression Analysis?

- As mentioned above, regression analysis estimates the **relationship between two or more variables**. Let's understand this with an easy **example**:
- Let's say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict **future sales of the company based on current & past information**.
- There are multiple benefits of using regression analysis. They are as follows:
- It indicates the **significant relationships** between dependent variable and independent variable.
- It indicates the **strength of impact** of multiple independent variables on a dependent variable.
- Regression analysis also allows us to compare the effects of variables measured on different scales, such as the **effect of price changes** and the **number of promotional activities**. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

# How many types of regression techniques do we have?

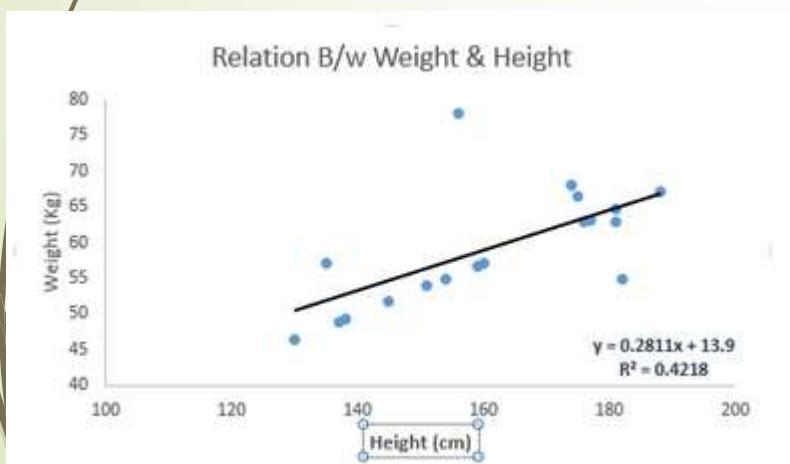
- There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by **three metrics** (number of independent variables, type of dependent variables and shape of regression line). We'll discuss them in detail in the following sections.



- For the **creative ones**, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven't used before.

# 1. Linear Regression

- It is one of the **most widely** known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the **dependent variable is continuous, independent variable(s) can be continuous or discrete**, and nature of regression line is linear.
- Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as **regression line**).
- It is represented by an equation  $Y=a+b*X + e$ , where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).



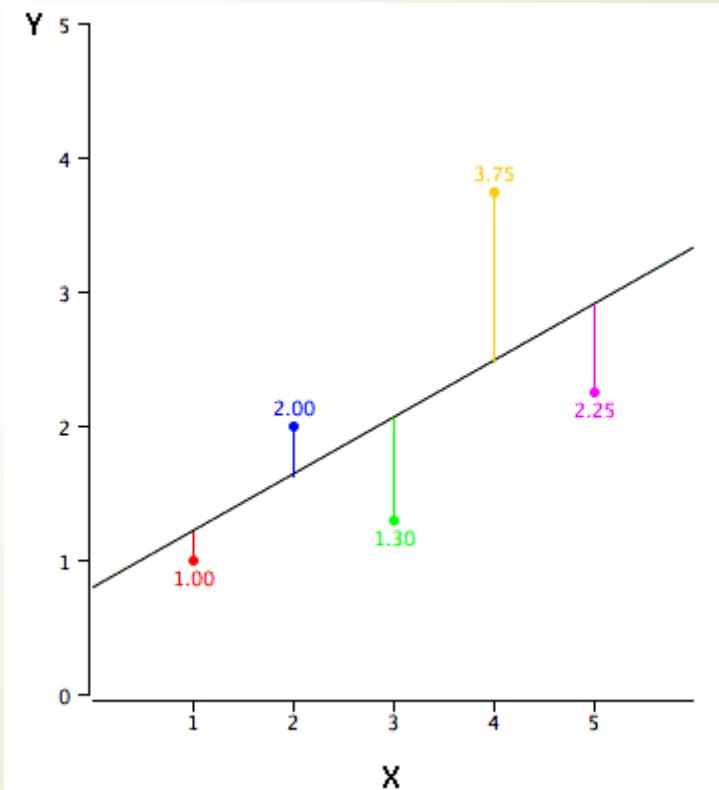
The difference between **simple linear regression** and **multiple linear regression** is that, multiple linear regression has ( $>1$ ) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is “**How do we obtain best fit line?**”.

# 1. Linear Regression

- How to obtain best fit line (Value of a and b)?
- This task can be easily accomplished by **Least Square Method**. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the **sum of the squares of the vertical deviations** from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values

$$\min_w \|Xw - y\|_2^2$$

We can evaluate the model performance using the metric **R-square**



## 2. Logistic Regression

- Logistic regression is used to **find the probability of event=Success and event=Failure**. We should use logistic regression when the dependent variable is **binary** (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can be represented by following equation.

$\text{odds} = p / (1-p) = \text{probability of event occurrence} / \text{probability of not event occurrence}$

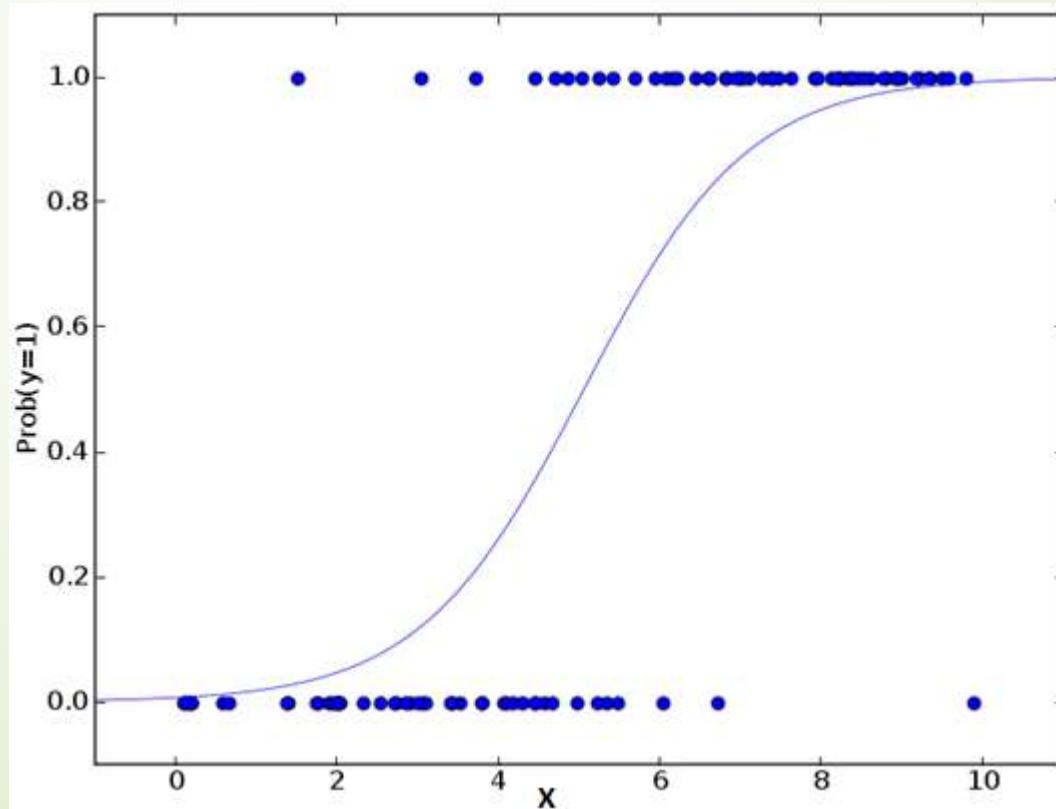
$$\ln(\text{odds}) = \ln(p/(1-p))$$

$$\text{logit}(p) = \ln(p/(1-p)) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

- Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

## 2. Logistic Regression

- Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is **logit** function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).

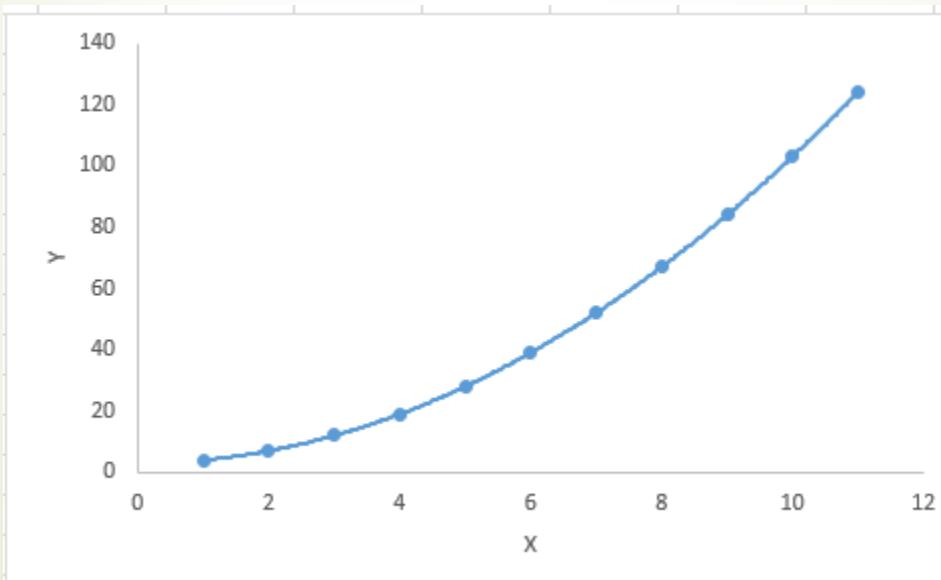


# 3. Polynomial Regression

- A regression equation is a polynomial regression equation if the **power of independent variable is more than 1**. The equation below represents a polynomial equation:

$$y=a+b*x^2$$

- In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



# 4. Stepwise Regression

- This form of regression is used when we deal with **multiple independent variables**. In this technique, the selection of independent variables is done with the help of an automatic process, which **involves no human intervention**.
- This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by **adding/dropping co-variates one at a time based on a specified criterion**. Some of the most commonly used Stepwise regression methods are listed below:
- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
- **Forward selection** starts with most significant predictor in the model and adds variable for each step.
- **Backward elimination** starts with all predictors in the model and removes the least significant variable for each step.
- **The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables.** It is one of the method to handle higher dimensionality of data set

# 5. Ridge Regression

- Ridge Regression is a technique used when the **data suffers from multicollinearity (independent variables are highly correlated)**. In multicollinearity, even though the **least squares estimates** (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. **By adding a degree of bias to the regression estimates**, ridge regression reduces the standard errors.
- Above, we saw the equation for linear regression. Remember? It can be represented as:
- $y=a+ b*x$
- This equation also has an error term. The complete equation becomes:

$y=a+b*x+e$  (error term), [error term is the value needed to correct for a prediction error between the observed and predicted value]

$\Rightarrow y=a+y= a+ b_1x_1+ b_2x_2+....+e$ , for multiple independent variables.

In a linear equation, prediction errors can be decomposed into two sub components. First is due to the **biased** and second is due to the **variance**. Prediction error can occur due to any one of these two or both components. Here, we'll discuss about the error caused due to variance.

# 5. Ridge Regression

- Ridge regression solves the **multicollinearity problem** through shrinkage parameter  $\lambda$  (lambda). Look at the equation below.

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

- In this equation, we have two components. First one is **least square term** and other one is **lambda of the summation** of  $\beta^2$  (beta- square) where  $\beta$  is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

# 6. Lasso Regression

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

- Similar to Ridge Regression, **Lasso (Least Absolute Shrinkage and Selection Operator)** also penalizes the absolute size of the regression coefficients. In addition, it is **capable of reducing the variability and improving the accuracy of linear regression models**. Look at the equation below: Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. **Larger the penalty applied, further the estimates get shrunk towards absolute zero.** This results to variable selection out of given n variables.

# 7. ElasticNet Regression

- ElasticNet is **hybrid of Lasso and Ridge Regression techniques**. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are **correlated**. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} (\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$

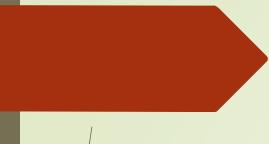
- A practical advantage of **trading-off between Lasso and Ridge** is that, it allows Elastic-Net to inherit some of Ridge's stability under rotation.

## Important Points:

- It encourages group effect in case of **highly correlated variables**
- There are no limitations on the **number of selected variables**
- It can suffer with **double shrinkage**

# How to select the right regression model?

- Life is usually simple, when you know only one or two techniques. One of the training institutes I know of tells their students – if the outcome is continuous – **apply linear regression**. If it is **binary – use logistic regression!** However, higher the number of options available at our disposal, more difficult it becomes to choose the right one. A similar case happens with regression models.
- Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:
  1. **Data exploration** is an inevitable part of building predictive model. It should be your **first step before selecting the right model** like identify the relationship and impact of variables



# How to select the right regression model?

2. To compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the Mallow's Cp criterion. This **essentially checks for possible bias in your model**, by comparing the model with all possible submodels (or a careful selection of them).
3. **Cross-validation** is the best way to evaluate models used for prediction. Here you divide your data set into two group (**train and validate**). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.
4. If your **data set has multiple confounding variables**, you should not choose automatic model selection method because you do not want to put these in a model at the same time.
5. It'll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a **highly statistically significant model**.
6. **Regression regularization methods**(Lasso, Ridge and ElasticNet) works well in case of **high dimensionality** and **multicollinearity** among the variables in the data set.

## Solved Example Problems for Regression Analysis

### Example 9.9

Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

*Solution:*

X	Y	$X^2$	$Y^2$	$XY$
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98

$\sum X = 28 \sum Y = 77 \sum X^2 = 140 \sum Y^2 = 875 \sum XY = 334$

Table 9.7

$$\bar{X} = \frac{\sum X}{N} = \frac{28}{7} = 4 ,$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{77}{7} = 11$$

(i) Regression coefficient of X on Y

$$b_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{N \sum Y^2 - (\sum Y)^2}$$

$$= \frac{7(334) - (28)(77)}{7(875) - (77)^2}$$

$$= \frac{2338 - 2156}{6125 - 5929}$$

$$= \frac{182}{196}$$

$$b_{xy} = 0.929$$

**(ii) Regression equation of  $X$  on  $Y$**

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 4 = 0.929(Y - 11)$$

$$X - 4 = 0.929Y - 10.219$$

$\therefore$  The regression equation  $X$  on  $Y$  is  $X = 0.929Y - 6.219$

**(iii) Regression coefficient of  $Y$  on  $X$**

$$b_{yx} = \frac{N\sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2}$$

$$= \frac{7(334) - (28)(77)}{7(140) - (28)^2}$$

$$= \frac{2338 - 2156}{980 - 784}$$

$$= \frac{182}{196}$$

$$\therefore b_{yx} = 0.929$$

**(iv) Regression equation of  $Y$  on  $X$**

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 11 = 0.929(X - 4)$$

$$Y = 0.929X - 3.716 + 11$$

$$= 0.929X + 7.284$$

The regression equation of  $Y$  on  $X$  is  $Y = 0.929X + 7.284$

# Solved Example Problems for Regression Analysis

## Example 9.9

Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

*Solution:*

X	Y	$X^2$	$Y^2$	$XY$
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98

$\sum X = 28$   $\sum Y = 77$   $\sum X^2 = 140$   $\sum Y^2 = 875$   $\sum XY = 334$

Table 9.7

$$\bar{X} = \frac{\sum X}{N} = \frac{28}{7} = 4,$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{77}{7} = 11$$

### **Regression coefficient of $X$ on $Y$**

$$b_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{N \sum Y^2 - (\sum Y)^2}$$

$$= \frac{7(334) - (28)(77)}{7(875) - (77)^2}$$

$$= \frac{2338 - 2156}{6125 - 5929}$$

$$= \frac{182}{196}$$

$$b_{xy} = 0.929$$

#### **(i) Regression equation of $X$ on $Y$**

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 4 = 0.929(Y - 11)$$

$$X - 4 = 0.929Y - 10.219$$

$\therefore$  The regression equation  $X$  on  $Y$  is  $X = 0.929Y - 6.219$

#### **(ii) Regression coefficient of $Y$ on $X$**

$$\begin{aligned}
 b_{yx} &= \frac{N\sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \\
 &= \frac{7(334) - (28)(77)}{7(140) - (28)^2} \\
 &= \frac{2338 - 2156}{980 - 784} \\
 &= \frac{182}{196} \\
 \therefore b_{yx} &= 0.929
 \end{aligned}$$

### (iii) Regression equation of $Y$ on $X$

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 11 = 0.929(X - 4)$$

$$Y = 0.929X - 3.716 + 11$$

$$= 0.929X + 7.284$$

The regression equation of  $Y$  on  $X$  is  $Y = 0.929X + 7.284$

### Example 9.10

Calculate the two regression equations of  $X$  on  $Y$  and  $Y$  on  $X$  from the data given below, taking deviations from actual means of  $X$  and  $Y$ .

<b>Price(Rs.)</b>	10	12	13	12	16	15
<b>Amount demanded</b>	40	38	43	45	37	43

Estimate the likely demand when the price is Rs.20.

**Solution:**

Calculation of Regression equation

X	$x = (X - 13)$	$x^2$	Y	$y = (Y - 41)$	$y^2$	$xy$
10	-3	9	40	-1	1	3
12	-1	1	38	-3	9	3
13	0	0	43	2	4	0
12	-1	1	45	4	16	-4
16	3	9	37	-4	16	-12
15	2	4	43	2	4	4
$\sum X = 78$	$\sum x = 0$	$\sum x^2 = 24$	$\sum Y = 246$	$\sum y = 0$	$\sum y^2 = 50$	$\sum xy = -6$

Table 9.8

### (i) Regression equation of X on Y

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

$$\bar{X} = \frac{78}{6} = 13, \quad \bar{Y} = \frac{246}{6} = 41$$

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-6}{50} = -0.12$$

$$X - 13 = -0.12 (Y - 41)$$

$$X - 13 = -0.12Y + 4.92$$

$$X = -0.12Y + 17.92$$

## (ii) Regression Equation of Y on X

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = -\frac{6}{24} = -0.25$$

$$Y - 41 = -0.25 (X - 13)$$

$$Y - 41 = -0.25 X + 3.25$$

$$Y = -0.25 X + 44.25$$

When  $X$  is 20,  $Y$  will be

$$= -0.25 (20) + 44.25$$

$$= -5 + 44.25$$

= 39.25 (when the price is Rs. 20, the likely demand is 39.25)

## Example 9.12

Find the means of  $X$  and  $Y$  variables and the coefficient of correlation between them from the following two regression equations:

$$2Y - X - 50 = 0$$

$$3Y - 2X - 10 = 0.$$

**Solution:**

We are given

$$2Y - X - 50 = 0 \dots (1)$$

$$3Y - 2X - 10 = 0 \dots (2)$$

Solving equation (1) and (2)

We get  $Y = 90$

Putting the value of  $Y$  in equation (1)

We get  $X = 130$

Hence  $\bar{X} = 130$  and  $\bar{Y} = 90$

**Calculating correlation coefficient**

Let us assume equation (1) be the regression equation of  $Y$  on  $X$

$$2Y = X + 50$$

$$2Y = X+50$$

$$Y = \frac{1}{2}X + 25 \text{ therefore } b_{yx} = \frac{1}{2}$$

Clearly equation ( 2) would be treated as regression equation of  $X$  on  $Y$

$$3Y - 2X - 10 = 0$$

$$2X = 3Y - 10$$

$$X = \frac{3}{2}Y - 5 \text{ therefore } b_{xy} = \frac{3}{2}$$

The Correlation coefficient  $r = \pm \sqrt{b_{xy} \times b_{yx}}$

$$r = \sqrt{\frac{1}{2} \times \frac{3}{2}} = 0.866$$

### NOTE

It may be noted that in the above problem one of the regression coefficient is greater than 1 and the other is less than 1. Therefore our assumption on given equations are correct.

### Example 9.14

The following table shows the sales and advertisement expenditure of a firm

	Sales	Advertisement expenditure ( Rs. Crores)
Mean	40	6
SD	10	1.5

Coefficient of correlation  $r = 0.9$ . Estimate the likely sales for a proposed advertisement expenditure of Rs. 10 crores.

**Solution:**

Given  $\bar{X} = 40$ ,  $\bar{Y} = 6$ ,  $\sigma_x = 10$ ,  $\sigma_y = 1.5$  and  $r = 0.9$

Equation of line of regression  $x$  on  $y$  is

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

$$X - 40 = (0.9) \frac{10}{1.5} (Y - 6)$$

$$X - 40 = 6Y - 36$$

$$X = 6Y + 4$$

When advertisement expenditure is 10 crores i.e.,  $Y=10$  then sales  $X=6(10)+4=64$  which implies sales is 64.

### **Example 9.15**

There are two series of index numbers  $P$  for price index and  $S$  for stock of the commodity. The mean and standard deviation of  $P$  are 100 and 8 and of  $S$  are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these data obtain the regression lines of  $P$  on  $S$  and  $S$  on  $P$ .

**Solution:**

Let us consider  $X$  for price  $P$  and  $Y$  for stock  $S$ . Then the mean and  $SD$  for  $P$  is considered as  $X\text{-Bar} = 100$  and  $\sigma_x = 8$ , respectively and the mean and  $SD$  of  $S$  is considered as  $Y\text{-Bar} = 103$  and  $\sigma_y = 4$ . The correlation coefficient between the series is  $r(X,Y) = 0.4$

Let the regression line  $X$  on  $Y$  be

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

$$X - 100 = (0.4) \frac{8}{4} (Y - 103)$$

$$X - 100 = 0.8(Y - 103)$$

$$X - 0.8Y - 17.6 = 0 \quad \text{or} \quad X = 0.8Y + 17.6$$

The regression line  $Y$  on  $X$  be  $Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$

$$Y - 103 = (0.4) \frac{4}{8} (X - 100)$$

$$Y - 103 = 0.2 (X - 100)$$

$$Y - 103 = 0.2 X - 20$$

$$Y = 0.2 X + 83 \quad \text{or} \quad 0.2 X - Y + 83 = 0$$

### Example 9.16

For 5 pairs of observations the following results are obtained  $\sum X = 15$ ,  $\sum Y = 25$ ,  $\sum X^2 = 55$ ,  $\sum Y^2 = 135$ ,  $\sum XY = 83$ . Find the equation of the lines of regression and estimate the value of  $X$  on the first line when  $Y = 12$  and value of  $Y$  on the second line if  $X = 8$ .

**Solution:**

Here  $N=5$ ,  $\bar{X} = \frac{\Sigma X}{N} = \frac{15}{5} = 3$ ,  $\bar{Y} = \frac{\Sigma Y}{N} = \frac{25}{5} = 5$

and the regression coefficient

$$b_{xy} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2} = \frac{5(83) - (15)(25)}{5(135) - (25)^2} = 0.8$$

The regression line of  $X$  on  $Y$  is

$$X - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$X - 3 = 0.8(Y - 5)$$

$$X = 0.8 Y - 1$$

When  $Y=12$ , the value of  $X$  is estimated as

$$X = 0.8(12) - 1 = 8.6$$

The regression coefficient

$$\begin{aligned} b_{yx} &= \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} \\ &= \frac{5(83) - (15)(25)}{5(55) - (15)^2} = 0.8 \end{aligned}$$

Thus  $b_{yx} = 0.8$  then the regression line  $Y$  on  $X$  is

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$Y - 5 = 0.8(X - 3)$$

$$Y = 0.8X + 2.6$$

When  $X=8$  the value of  $Y$  is estimated as

$$Y = 0.8(8) + 2.6$$

$$Y = 9$$

$$Y-5 = 0.8(X-3)$$

$$= 0.8X+2.6$$

When  $X=8$  the value of  $Y$  is estimated as

$$= 0.8(8)+2.6$$

$$= 9$$

### Example 9.19

For the given lines of regression  $3X-2Y=5$  and  $X-4Y=7$ . Find

(i) Regression coefficients

(ii) Coefficient of correlation

**Solution:**

(i) First convert the given equations  $Y$  on  $X$  and  $X$  on  $Y$  in standard form and find their regression coefficients respectively.

Given regression lines are

$$3X-2Y = 5 \dots (1)$$

$$X-4Y = 7 \dots (2)$$

Let the line of regression of  $X$  on  $Y$  is

$$3X-2Y = 5$$

$$3X = 2Y+5$$

$$3X - 2Y = 5$$

$$3X = 2Y + 5$$

$$X = \frac{1}{3}(2Y + 5)$$

$$X = \frac{1}{3}(2Y + 5)$$

$$X = \frac{2}{3}Y + \frac{5}{3}$$

∴ Regression coefficient of  $X$  on  $Y$  is

$$b_{xy} = \frac{2}{3} (<1)$$

Let the line of regression of  $Y$  on  $X$  is

$$X - 4Y = 7$$

$$-4Y = -X + 7$$

$$4Y = X - 7$$

$$Y = \frac{1}{4}(X - 7)$$

$$Y = \frac{1}{4}X - \frac{7}{4}$$

∴ Regression coefficient of  $Y$  on  $X$  is

$$b_{yx} = \frac{1}{4} (<1)$$

## Coefficient of correlation

Since the two regression coefficients are positive then the correlation coefficient is also positive and it is given by

$$r = \sqrt{b_{yx} \cdot b_{xy}}$$

$$= \sqrt{\frac{2}{3} \cdot \frac{1}{4}}$$

$$= \sqrt{\frac{1}{6}}$$

$$= 0.4082$$

$$r = 0.4082$$

### Exercise

1. From the data given below

Marks in Economics:	25	28	35	32	31	36	29	38	34	32
Marks in Statistics:	43	46	49	41	36	32	31	30	33	39

Find (a) The two regression equations, (b) The coefficient of correlation between marks in Economics and statistics, (c) The mostly likely marks in Statistics when the marks in Economics is 30.

2. The heights ( in cm.) of a group of fathers and sons are given below

Heights of fathers:	158	166	163	165	167	170	167	172	177	181
Heights of Sons :	163	158	167	170	160	180	170	175	172	175

Find the lines of regression and estimate the height of son when the height of the father is 164 cm.

3. The following data give the height in inches ( $X$ ) and the weight in lb. ( $Y$ ) of a random sample of 10 students from a large group of students of age 17 years:

X	61	68	68	64	65	70	63	62	64	67
Y	112	123	130	115	110	125	100	113	116	125

Estimate weight of the student of a height 69 inches.

4. Obtain the two regression lines from the following data  $N=20$ ,  $\sum X=80$ ,  $\sum Y=40$ ,  $\sum X^2=1680$ ,  $\sum Y^2=320$  and  $\sum XY=480$

5. Given the following data, what will be the possible yield when the rainfall is 29 $\text{₹}$

	Rainfall	Production
Mean	25 $\text{``}$	40 units per acre
Standard Deviation	3 $\text{``}$	6 units per acre

Coefficient of correlation between rainfall and production is 0.8

6. The following data relate to advertisement expenditure(in lakh of rupees) and their corresponding sales( in crores of rupees)

Advertisement expenditure	40	50	38	60	65	50	35
Sales	38	60	55	70	60	48	30

Estimate the sales corresponding to advertising expenditure of Rs. 30 lakh.

7. You are given the following data:

	X	Y
Arithmetic Mean	36	85
Standard Deviation	11	8

If the Correlation coefficient between  $X$  and  $Y$  is 0.66, then find (i) the two regression coefficients, (ii) the most likely value of  $Y$  when  $X=10$

8. Find the equation of the regression line of  $Y$  on  $X$ , if the observations ( $X_i, Y_i$ ) are the following (1,4) (2,8) (3,2) (4,12) (5, 10) (6, 14) (7, 16) (8, 6) (9, 18)

9. A survey was conducted to study the relationship between expenditure on accommodation ( $X$ ) and expenditure on Food and Entertainment ( $Y$ ) and the following results were obtained:

	Mean	SD
Expenditure on Accommodation	Rs. 178	63.15
Expenditure on Food and Entertainment	Rs 47.8	22.98
Coefficient of Correlation	0.43	

Write down the regression equation and estimate the expenditure on Food and Entertainment, if the expenditure on accommodation is Rs. 200.

10. For 5 observations of pairs of ( $X, Y$ ) of variables  $X$  and  $Y$  the following results are obtained.  $\sum X=15$ ,  $\sum Y=25$ ,  $\sum X_2=55$ ,  $\sum Y_2=135$ ,  $\sum XY=83$ . Find the equation of the lines of regression and estimate the values of  $X$  and  $Y$  if  $Y=8$ ;  $X=12$ .

11. The two regression lines were found to be  $4X-5Y+33=0$  and  $20X-9Y-107=0$ . Find the mean values and coefficient of correlation between  $X$  and  $Y$ .
12. The equations of two lines of regression obtained in a correlation analysis are the following  $2X=8-3Y$  and  $2Y=5-X$ . Obtain the value of the regression coefficients and correlation coefficient.

Multivariate analysis is different from Univariate

- Examines the interrelatedness between and within sets of variables

Multivariate analysis exposes the structure and meaning of the data from the application and interpretation of various statistical methods, beyond the ones in a basic statistics course

Multivariate Analysis focuses on the many variables within a dataset.

Multivariate statistical analysis focuses on establishing relationship among a set of variables, or grouping entities into distinct subgroups

## Difference between Univariate(UV) and Multivariate (MV)

---

Univariate analysis focuses on a single dependent variable, although some consider any dataset with multiple variables multivariate. We prefer the term multivariate for dealing with multiple variables, especially when considering multiple dependent variables.

Data:  $a_1, a_2, a_3, \dots, a_n$

Multivariate deals with a larger data set consisting of many variables. Generally looking like a table format

Data:	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
	..	..	..	..	..	..
	..	..	..	..	..	..
	..	..	..	..	..	..
	..	..	..	..	..	..

## Goals of Multivariate Analysis

---

**Data reduction:** identify one or a few new variables that capture most of the variability in a large data set.

**Sorting and grouping:** develop rules for splitting a large data set into relatively homogeneous groups.

**Dependence:** characterize relationships among variables. **Prediction:** use dependence to predict.

**Hypothesis testing:** multiple responses in a designed experiment

# When should we use MV techniques

---

## **Complexity**

The subject/data studied may be more complex than what univariate methods can offer in terms of analysis

## **Reality**

In some cases it would be inappropriate to conduct univariate analysis as the data/research demand a multivariate analysis

## **Experiments vs Empirical Data**

Experimental data generally manipulates a single variable (drug vs placebo) making univariate analysis easier to imply some causality, empirical data however (crime or census) may require MV for accurate correlation analysis.

## **History**

MV calculations were much harder due to lack of availability of statistical computing. Advances in computing and software make it much easier to perform.

## MV Analysis - Caution

---

### Ambiguity

MV analysis may result in a less clear understanding of the data, e.g. group differences on a linear combination of DVs (Manova). It's sometimes easier to explain group differences in a univariate context.

### Complexity

Just because a technique is popular doesn't mean you need to use it, e.g. using a tank to destroy a wall.

### Assumptions

MV analyses have rules and assumptions just like UV.

## Overview of MV

---

### Advantages

Richer realistic design

Looks at phenomena in an overarching way (provides multiple levels of analysis)

Each method differs in amount or type of IV and DV\*

Can help control for Type I Error

### Disadvantages

Larger Ns are often required

More difficult to interpret

Less known about the robustness of assumptions

## Foundations

---

**Terminology:** We measure several variables for each item in a sample  
 $p$  variables,  $n$  items.

**Notation:**  $x_{jk}$  = measurement of  $k^{\text{th}}$  variable on  $j^{\text{th}}$  item.

**Data array:** Columns are the variables, and observations are the rows.

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
$\text{Obs}_1$	..	..	..	..	..	..
$\text{Obs}_2$	..	..	..	..	..	..
...	..	..	..	..	..	..
$\text{Obs}_n$	..	..	..	..	..	..

-Used to analyze your data when you have made multiple measurements on items or subjects

**Factor Analysis: Academic record, Appearance, Communication, Company Fit, Experience, Job Fit, Letter, Likeability, Organization, Potential, Resume, Self-Confidence**

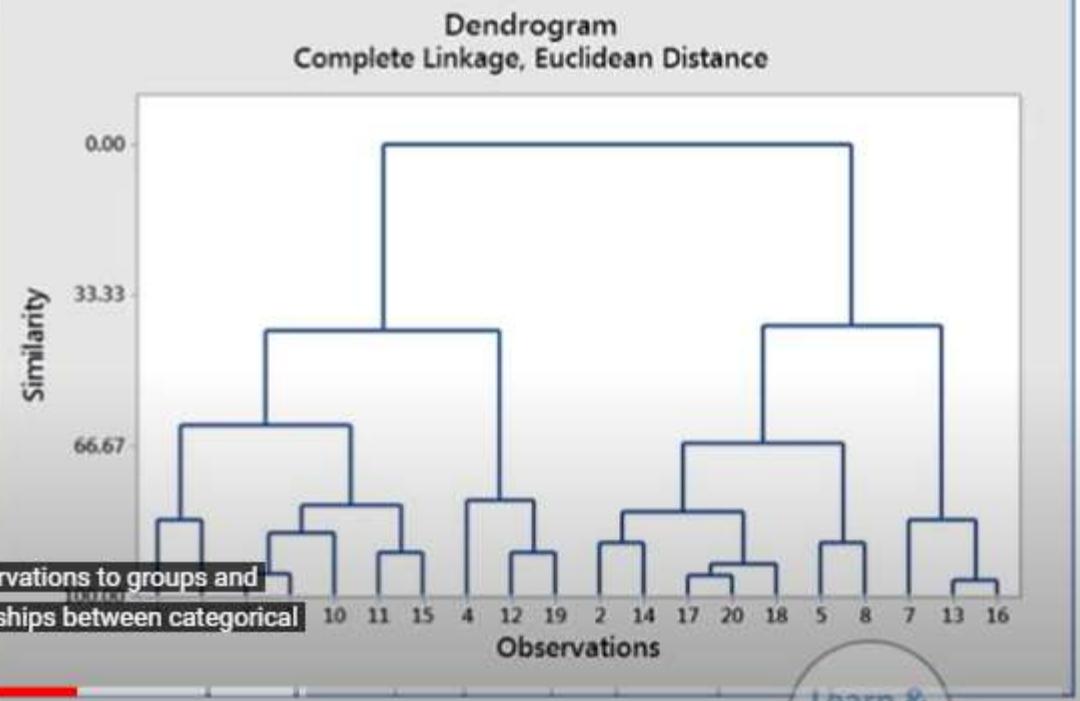
Maximum Likelihood Factor Analysis of the Correlation Matrix

**Unrotated Factor Loadings and Communalities**

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Academic record	0.380	0.455	0.340	0.259	0.534
Appearance	0.359	0.530	-0.040	0.523	0.685
Communication	0.465	0.660	-0.377	-0.023	0.795
Company Fit	0.523	0.677	0.266	-0.253	0.866
Experience	0.508	0.194	0.450	0.232	0.553
Job Fit	0.532	0.632	0.415	-0.201	0.895
Letter	0.992	-0.094	-0.012	-0.007	0.994
Likeability	0.412	0.529	0.032	0.377	0.593
Organization	0.406	0.761	-0.424	-0.055	0.926
Potential	0.446	0.548	0.431	0.172	0.714
Resume	0.850	0.040	0.096	0.283	0.814
Self-Confidence	0.293	0.575	0.083	0.506	0.679
Variance	3.6320	3.3193	1.0883	1.0095	9.0491
% Var	0.303	0.277	0.091	0.084	0.754

• Assign observations to groups and

• Explore relationships between categorical



## Multivariate Tools:



### 1) Principal Component Analysis:

-Used to form a smaller number of uncorrelated variables from a large set of data

**Principal Component Analysis: Income, Education, Age, Residence, Employ, Savings, Debt, Credit cards**

#### Eigenanalysis of the Correlation Matrix

Eigenvalue	3.5476	2.1320	1.0447	0.5315	0.4112	0.1665	0.1254	0.0411
Proportion	0.443	0.266	0.131	0.066	0.051	0.021	0.016	0.005
Cumulative	0.443	0.710	0.841	0.907	0.958	0.979	0.995	1.000

#### Eigenvectors

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Income	0.314	0.145	-0.676	-0.347	-0.241	0.494	0.018	-0.030
Education	0.237	0.444	-0.401	0.240	0.622	-0.357	0.103	0.057
Age	0.484	-0.135	-0.004	-0.212	-0.175	-0.487	-0.657	-0.052
Residence	0.466	-0.277	0.091	0.116	-0.035	-0.085	0.487	-0.662
Employ	0.459	-0.304	0.122	-0.017	-0.014	-0.023	0.368	0.739
Savings	0.404	0.219	0.366	0.436	0.143	0.568	-0.348	-0.017
Debt	-0.067	-0.585	-0.078	-0.2	Principal components have wide applicability	5		
Credit cards	-0.123	-0.452	-0.468	0.703	in social sciences and market research.	058		

-To explain the maximum amount of variance with the fewest number of Principal Components



## 2) Factor Analysis:

-Used to determine the underlying factors responsible for correlations in the data

**Factor Analysis: Academic record, Appearance, Communication, Company Fit, Experience, Job Fit, Letter, Likeability, Organization, Potential, Resume, Self-Confidence**

Maximum Likelihood Factor Analysis of the Correlation Matrix

### Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Academic record	0.380	0.455	0.340	0.259	0.534
Appearance	0.359	0.530	-0.040	0.523	0.685
Communication	0.465	0.660	-0.377	-0.023	0.795
Company Fit	0.523	0.677	0.266	-0.253	0.866
Experience	0.508	0.194	0.450	0.232	0.553
Job Fit	0.532	0.632	0.415	-0.201	0.895
Letter	0.992	-0.094	-0.012	-0.007	0.994
Likeability	0.412	0.529	0.032	0.377	0.593
Organization	0.406	0.761	-0.424	-0.055	0.926
Potential	0.446	0.548	0.431		
Resume	0.850	0.040	0.096		
Self-Confidence	0.293	0.575	0.083	0.506	0.679

-Summarizes data into a few dimensions by compressing a large number of variables into a smaller set of hidden factors



### 3) Item Analysis:



-To assess how well the multiple items in a survey measure the same characteristic

#### Item Analysis of Item 1, Item 2, Item 3

##### Correlation Matrix

	Item 1	Item 2
Item 2	0.903	
Item 3	0.867	0.864

##### Cell Contents

Pearson correlation

#### Item and Total Statistics

Variable	Total Count	Total Mean	Total StDev
Item 1	50	3.1600	1.2675
Item 2	50	2.8400	1.3607
Item 3	50	2.9400	1.3463
Total	50	8.9400	3.8087

#### Cronbach's Alpha

##### Alpha

0.9550

#### Omitted Item Statistics

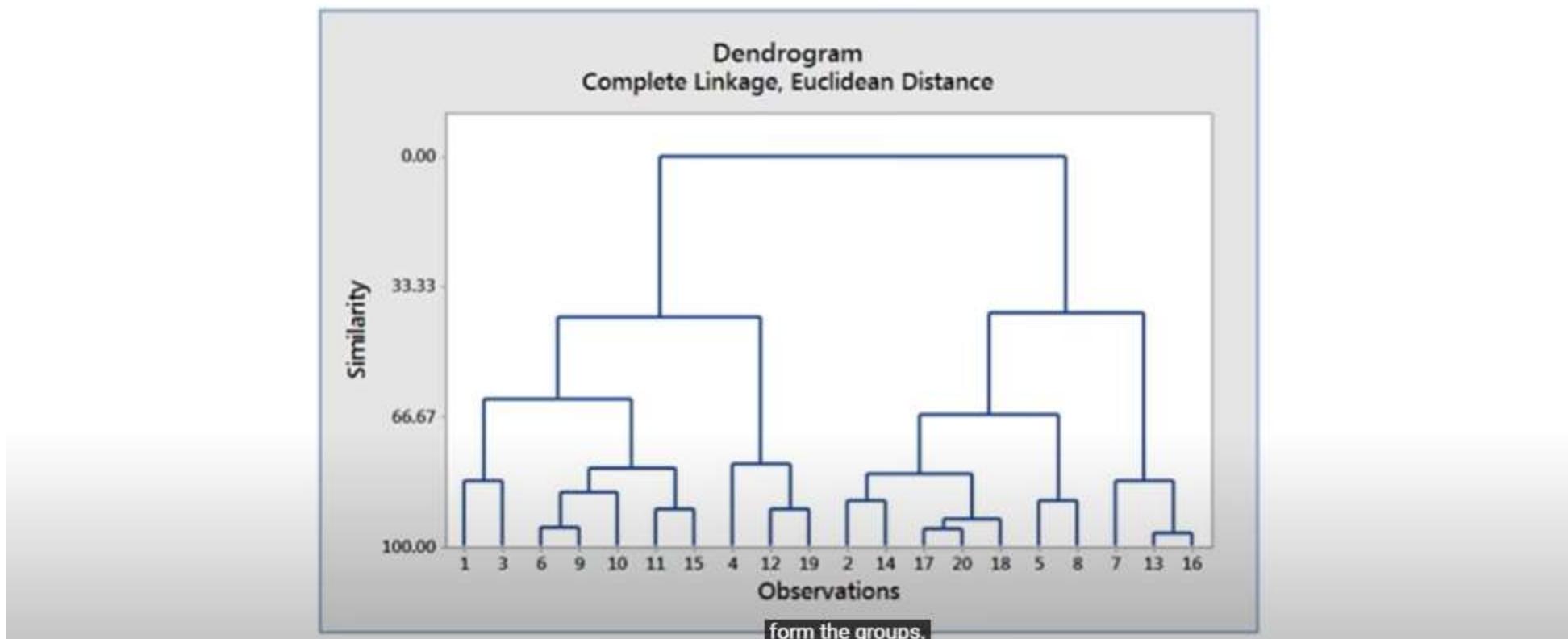
Omitted Variable	Adj. Total	Adj. Total	Item-Adj. Total	Item-Adj. Total Corr	Squared Multiple Corr
Mean	StDev	Mean	StDev	Mean	StDev

• Determine whether omitting items improve internal consistency.

- Assess the strength and direction of the relationship between pairs of items
- Evaluate the overall internal consistency of the test or survey
- Determine whether omitting items improve internal consistency

#### 4) Cluster Observations:

-To join observations that share common characteristics into groups



5. Cluster Variables

6. Cluster K-Means

7. Descriminant Analysis

8. Simple Correspondance Analysis

9. Multiple Correspondance Analysis

## Dimension Reduction-

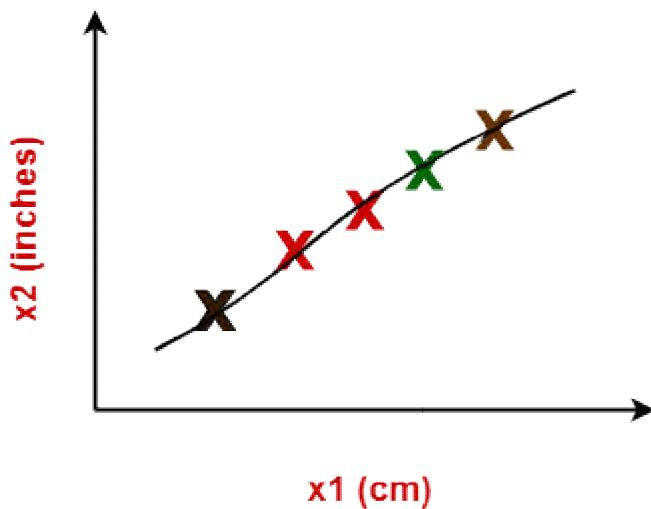
In pattern recognition, Dimension Reduction is defined as-

- It is a process of converting a data set having vast dimensions into a data set with lesser dimensions.
- It ensures that the converted data set conveys similar information concisely.

## Example-

Consider the following example-

- The following graph shows two dimensions  $x_1$  and  $x_2$ .
- $x_1$  represents the measurement of several objects in cm.
- $x_2$  represents the measurement of several objects in inches.



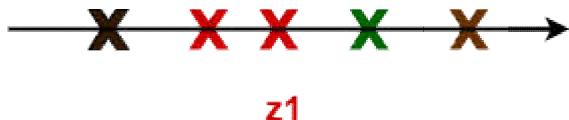
In machine learning,

- Using both these dimensions convey similar information.

- Also, they introduce a lot of noise in the system.
- So, it is better to use just one dimension.

Using dimension reduction techniques-

- We convert the dimensions of data from 2 dimensions ( $x_1$  and  $x_2$ ) to 1 dimension ( $z_1$ ).
- It makes the data relatively easier to explain.



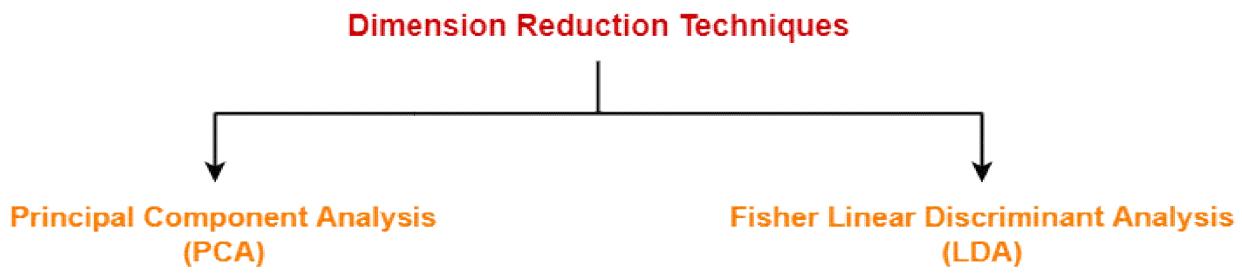
## **Benefits-**

Dimension reduction offers several benefits such as-

- It compresses the data and thus reduces the storage space requirements.
- It reduces the time required for computation since less dimensions require less computation.
- It eliminates the redundant features.
- It improves the model performance.

## **Dimension Reduction Techniques-**

The two popular and well-known dimension reduction techniques are-



1. Principal Component Analysis (PCA)
2. Fisher Linear Discriminant Analysis (LDA)

In this article, we will discuss about Principal Component Analysis.

## Principal Component Analysis-

- Principal Component Analysis is a well-known dimension reduction technique.
- It transforms the variables into a new set of variables called as principal components.
- These principal components are linear combination of original variables and are orthogonal.
- The first principal component accounts for most of the possible variation of original data.
- The second principal component does its best to capture the variance in the data.
- There can be only two principal components for a two-dimensional data set.

## PCA Algorithm-

The steps involved in PCA Algorithm are as follows-

**Step-01:** Get data.

**Step-02:** Compute the mean vector ( $\mu$ ).

**Step-03:** Subtract mean from the given data.

**Step-04:** Calculate the covariance matrix.

**Step-05:** Calculate the eigen vectors and eigen values of the covariance matrix.

**Step-06:** Choosing components and forming a feature vector.

**Step-07:** Deriving the new data set.

## **PRACTICE PROBLEMS BASED ON PRINCIPAL COMPONENT ANALYSIS-**

### **Problem-01:**

Given data = { 2, 3, 4, 5, 6, 7 ; 1, 5, 3, 6, 7, 8 }.

Compute the principal component using PCA Algorithm.

**OR**

Consider the two dimensional patterns (2, 1), (3, 5), (4, 3), (5, 6), (6, 7), (7, 8).

Compute the principal component using PCA Algorithm.

**OR**

Compute the principal component of following data-

CLASS 1

$$X = 2, 3, 4$$

$$Y = 1, 5, 3$$

CLASS 2

$$X = 5, 6, 7$$

$$Y = 6, 7, 8$$

## Solution-

We use the above discussed PCA Algorithm-

### Step-01:

Get data.

The given feature vectors are-

- $x_1 = (2, 1)$
- $x_2 = (3, 5)$
- $x_3 = (4, 3)$
- $x_4 = (5, 6)$
- $x_5 = (6, 7)$
- $x_6 = (7, 8)$

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} \begin{bmatrix} 6 \\ 7 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

### Step-02:

Calculate the mean vector ( $\mu$ ).

Mean vector ( $\mu$ )

$$\begin{aligned} &= ((2 + 3 + 4 + 5 + 6 + 7) / 6, (1 + 5 + 3 + 6 + 7 + 8) / 6) \\ &= (4.5, 5) \end{aligned}$$

Thus,

Mean vector ( $\mu$ ) = 
$$\begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$$

### Step-03:

Subtract mean vector ( $\mu$ ) from the given feature vectors.

- $x_1 - \mu = (2 - 4.5, 1 - 5) = (-2.5, -4)$
- $x_2 - \mu = (3 - 4.5, 5 - 5) = (-1.5, 0)$
- $x_3 - \mu = (4 - 4.5, 3 - 5) = (-0.5, -2)$
- $x_4 - \mu = (5 - 4.5, 6 - 5) = (0.5, 1)$
- $x_5 - \mu = (6 - 4.5, 7 - 5) = (1.5, 2)$
- $x_6 - \mu = (7 - 4.5, 8 - 5) = (2.5, 3)$

Feature vectors ( $x_i$ ) after subtracting mean vector ( $\mu$ ) are-

$$\begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$$

#### Step-04:

Calculate the covariance matrix.

Covariance matrix is given by-

$$\text{Covariance Matrix} = \frac{\sum (x_i - \mu)(x_i - \mu)^t}{n}$$

Now,

$$m_1 = (x_1 - \mu)(x_1 - \mu)^t = \begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -2.5 & -4 \end{bmatrix} = \begin{bmatrix} 6.25 & 10 \\ 10 & 16 \end{bmatrix}$$

$$m_2 = (x_2 - \mu)(x_2 - \mu)^t = \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -1.5 & 0 \end{bmatrix} = \begin{bmatrix} 2.25 & 0 \\ 0 & 0 \end{bmatrix}$$

$$m_3 = (x_3 - \mu)(x_3 - \mu)^t = \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} -0.5 & -2 \end{bmatrix} = \begin{bmatrix} 0.25 & 1 \\ 1 & 4 \end{bmatrix}$$

$$m_4 = (x_4 - \mu)(x_4 - \mu)^t = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 0.25 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$m_5 = (x_5 - \mu)(x_5 - \mu)^t = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 1.5 & 2 \end{bmatrix} = \begin{bmatrix} 2.25 & 3 \\ 3 & 4 \end{bmatrix}$$

$$m_6 = (x_6 - \mu)(x_6 - \mu)^t = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix} \begin{bmatrix} 2.5 & 3 \end{bmatrix} = \begin{bmatrix} 6.25 & 7.5 \\ 7.5 & 9 \end{bmatrix}$$

Now,

Covariance matrix

$$= (m_1 + m_2 + m_3 + m_4 + m_5 + m_6) / 6$$

On adding the above matrices and dividing by 6, we get-

$$\text{Covariance Matrix} = \frac{1}{6} \begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix}$$

$$\text{Covariance Matrix} = \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix}$$

### Step-05:

Calculate the eigen values and eigen vectors of the covariance matrix.

$\lambda$  is an eigen value for a matrix M if it is a solution of the characteristic equation  $|M - \lambda I| = 0$ .

So, we have-

$$\begin{vmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{vmatrix} - \begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix} = 0$$

$$\begin{vmatrix} 2.92 - \lambda & 3.67 \\ 3.67 & 5.67 - \lambda \end{vmatrix} = 0$$

From here,

$$(2.92 - \lambda)(5.67 - \lambda) - (3.67 \times 3.67) = 0$$

$$16.56 - 2.92\lambda - 5.67\lambda + \lambda^2 - 13.47 = 0$$

$$\lambda^2 - 8.59\lambda + 3.09 = 0$$

Solving this quadratic equation, we get  $\lambda = 8.22, 0.38$

Thus, two eigen values are  $\lambda_1 = 8.22$  and  $\lambda_2 = 0.38$ .

Clearly, the second eigen value is very small compared to the first eigen value.

So, the second eigen vector can be left out.

Eigen vector corresponding to the greatest eigen value is the principal component for the given data set.

So, we find the eigen vector corresponding to eigen value  $\lambda_1$ .

We use the following equation to find the eigen vector-

$$MX = \lambda X$$

where-

- M = Covariance Matrix
- X = Eigen vector
- $\lambda$  = Eigen value

Substituting the values in the above equation, we get-

$$\begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 8.22 \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Solving these, we get-

$$2.92X_1 + 3.67X_2 = 8.22X_1$$

$$3.67X_1 + 5.67X_2 = 8.22X_2$$

On simplification, we get-

$$5.3X_1 = 3.67X_2 \dots\dots\dots(1)$$

$$3.67X_1 = 2.55X_2 \dots\dots\dots(2)$$

From (1) and (2),  $X_1 = 0.69X_2$

From (2), the eigen vector is-

Eigen Vector :

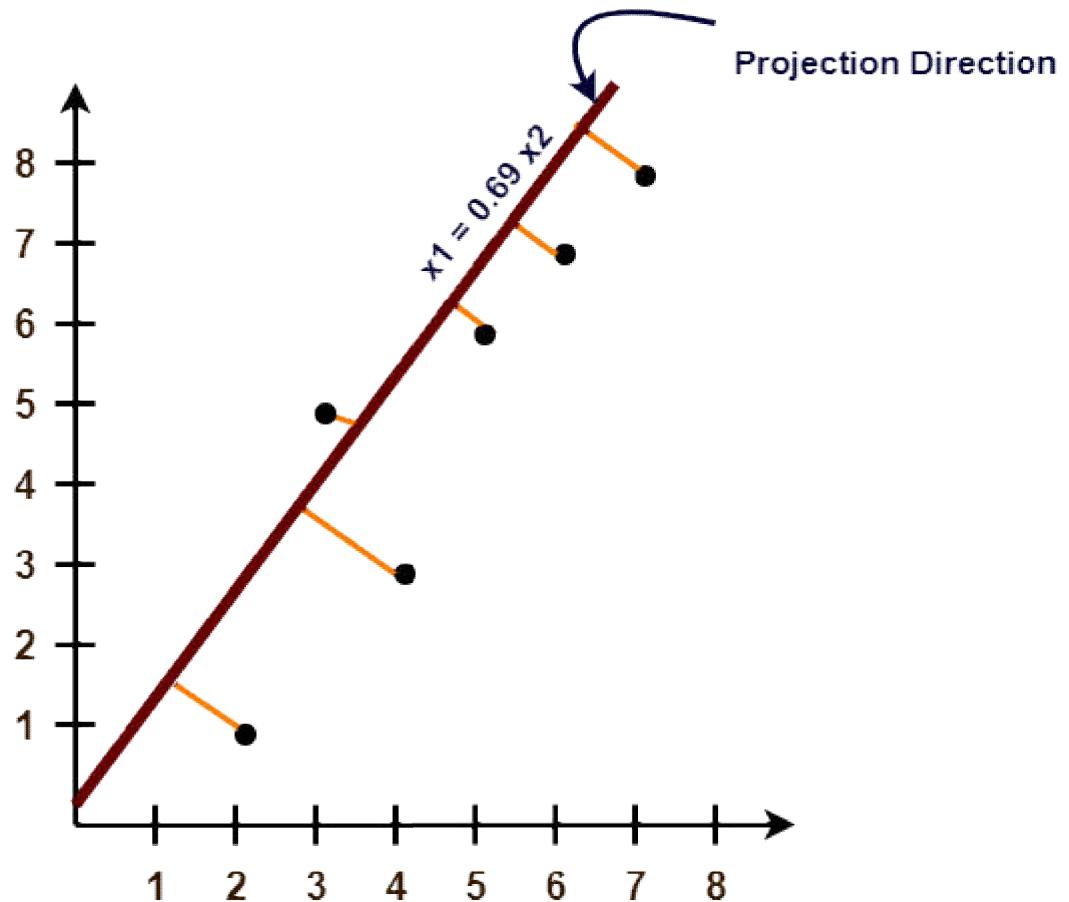
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$$

Thus, principal component for the given data set is-

Principal Component :

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$$

Lastly, we project the data points onto the new subspace as-



# Bayesian Statistics

## Overview

- The drawbacks of frequentist statistics lead to the need for Bayesian Statistics
- Discover Bayesian Statistics and Bayesian Inference
- There are various methods to test the significance of the model like p-value, confidence interval, etc

## Introduction

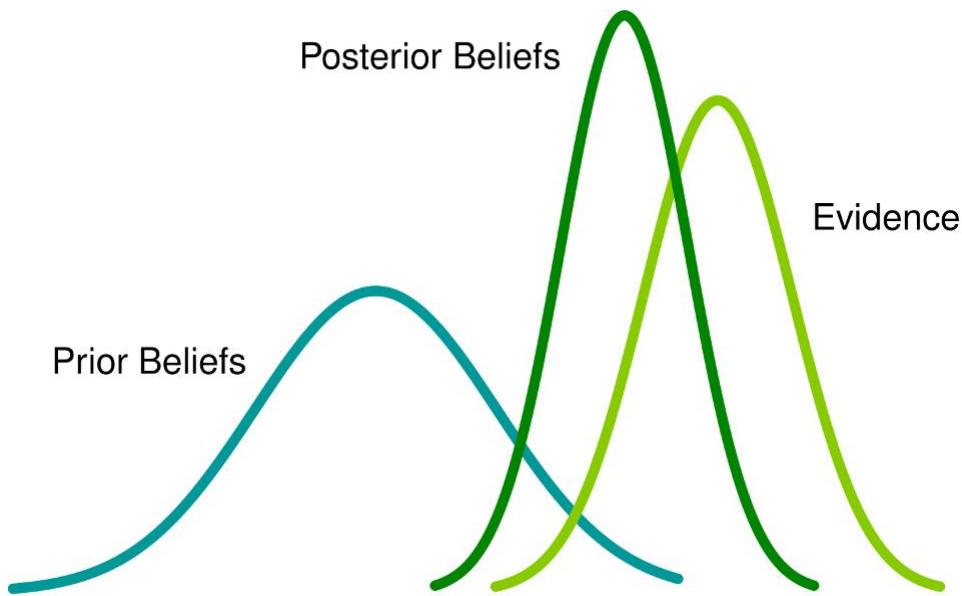
Bayesian Statistics continues to remain incomprehensible in the ignited minds of many analysts. Being amazed by the incredible power of [machine learning](#), a lot of us have become unfaithful to statistics. Our focus has narrowed down to exploring machine learning. Isn't it true?

We fail to understand that machine learning is not the only way to solve real world problems. In several situations, it does not help us solve business problems, even though there is data involved in these problems. To say the least, [knowledge of statistics](#) will allow you to work on complex analytical problems, irrespective of the size of data.

In 1770s, Thomas Bayes introduced ‘Bayes Theorem’. Even after centuries later, the importance of ‘Bayesian Statistics’ hasn’t faded away. In fact, today this topic is being taught in great depths in some of the world’s leading universities.

With this idea, I’ve created this beginner’s guide on Bayesian Statistics. I’ve tried to explain the concepts in a simplistic manner with examples. Prior knowledge of [basic probability & statistics](#) is desirable. You should check out [this course](#) to get a comprehensive low down on statistics and probability.

By the end of this article, you will have a concrete understanding of Bayesian Statistics and its associated concepts.



# Table of Contents

1. Frequentist Statistics
2. The Inherent Flaws in Frequentist Statistics
3. Bayesian Statistics
  - o Conditional Probability
  - o Bayes Theorem
4. Bayesian Inference
  - o Bernoulli likelihood function
  - o Prior Belief Distribution
  - o Posterior belief Distribution
5. Test for Significance – Frequentist vs Bayesian
  - o p-value
  - o Confidence Intervals
  - o Bayes Factor
  - o High Density Interval (HDI)

Before we actually delve in Bayesian Statistics, let us spend a few minutes understanding *Frequentist Statistics*, the more popular version of statistics most of us come across and the inherent problems in that.

# 1. Frequentist Statistics

The debate between *frequentist* and *bayesian* have haunted beginners for centuries. Therefore, it is important to understand the difference between the two and how does there exists a thin line of demarcation!

It is the most widely used inferential technique in the statistical world. Infact, generally it is the first school of thought that a person entering into the statistics world comes across.

Frequentist Statistics tests whether an event (hypothesis) occurs or not. It calculates the probability of an event in the long run of the experiment (i.e the experiment is repeated under the same conditions to obtain the outcome).

Here, the sampling distributions of fixed size are taken. Then, the experiment is theoretically repeated infinite number of times but practically done with a stopping intention. For example, I perform an experiment with a stopping intention in mind that I will stop the experiment when it is repeated 1000 times or I see minimum 300 heads in a coin toss.

Let's go deeper now.

Now, we'll understand *frequentist statistics* using an example of coin toss. The objective is to estimate the fairness of the coin. Below is a table representing the frequency of heads:

no. of tosses	no. of heads	difference
10	4	-1
50	25	0
100	44	-6
500	255	5
1000	502	2
5000	2533	33
10000	5067	67

We know that probability of getting a head on tossing a fair coin is 0.5. **No. of heads** represents the actual number of heads obtained. **Difference** is the difference between  $0.5 * (\text{No. of tosses}) - \text{no. of heads}$ .

An important thing is to note that, though the difference between the actual number of heads and expected number of heads( 50% of number of tosses) increases as the number of tosses are increased, the proportion of number of heads to total number of tosses approaches 0.5 (for a fair coin).

This experiment presents us with a very common flaw found in frequentist approach i.e. *Dependence of the result of an experiment on the number of times the experiment is repeated.*

To know more about frequentist statistical methods, you can head to this [excellent course](#) on inferential statistics.

## 2. The Inherent Flaws in Frequentist Statistics

Till here, we've seen just one flaw in *frequentist statistics*. Well, it's just the beginning.

20th century saw a massive upsurge in the *frequentist statistics* being applied to numerical models to check whether one sample is different from the other, a parameter is important enough to be kept in the model and various other manifestations of hypothesis testing. But *frequentist statistics* suffered some great flaws in its design and interpretation which posed a serious concern in all real life problems. For example:

1. *p-values* measured against a sample (fixed size) statistic with some stopping intention changes with change in intention and sample size. i.e If two persons work on the same data and have different stopping intention, they may get two different *p- values* for the same data, which is undesirable.

For example: Person A may choose to stop tossing a coin when the total count reaches 100 while B stops at 1000. For different sample sizes, we get different t-scores and different p-values. Similarly, intention to stop may change from fixed

number of flips to total duration of flipping. In this case too, we are bound to get different *p-values*.

2- Confidence Interval (C.I) like *p-value* depends heavily on the sample size. This makes the stopping potential absolutely absurd since no matter how many persons perform the tests on the same data, the results should be consistent.

3- Confidence Intervals (C.I) are not probability distributions therefore they do not provide the most probable value for a parameter and the most probable values.

These three reasons are enough to get you going into thinking about the drawbacks of the *frequentist approach* and why is there a need for *bayesian approach*. Let's find it out.

From here, we'll first understand the basics of Bayesian Statistics.

### 3. Bayesian Statistics

"Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data."

You got that? Let me explain it with an example:

Suppose, out of all the 4 championship races (F1) between [Niki Lauda](#) and [James hunt](#), Niki won 3 times while James managed only 1.

So, if you were to bet on the winner of next race, who would he be ?

I bet you would say Niki Lauda.

Here's the twist. What if you are told that it rained once when James won and once when Niki won and it is definite that it will rain on the next date. So, who would you bet your money on now ?

By intuition, it is easy to see that chances of winning for James have increased drastically. But the question is: how much ?

To understand the problem at hand, we need to become familiar with some concepts, first of which is conditional probability (explained below).

In addition, there are certain pre-requisites:

Pre-Requisites:

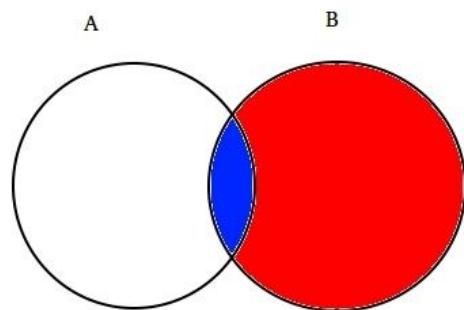
1. Linear Algebra : To refresh your basics, you can check out [Khan's Academy Algebra](#).
2. Probability and Basic Statistics : To refresh your basics, you can check out [another course](#) by Khan Academy.

### 3.1 Conditional Probability

It is defined as the: Probability of an event A given B equals the probability of B and A happening together divided by the probability of B."

For example: Assume two partially intersecting sets A and B as shown below.

Set A represents one set of events and Set B represents another. We wish to calculate the probability of A given B has already happened. Lets represent the happening of event B by shading it with red.



Now since B has happened, the part which now matters for A is the part shaded in blue which is interestingly  $A \cap B$ . So, the probability of A given B turns out to be:

$$\frac{\text{BlueArea}}{\text{RedArea} + \text{BlueArea}}$$

Therefore, we can write the formula for event B given A has already occurred by:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

or

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Now, the second equation can be rewritten as :

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

This is known as Conditional Probability.

Let's try to answer a betting problem with this technique.

Suppose, B be the event of *winning of James Hunt*. A be the event of *raining*.

Therefore,

1.  $P(A) = 1/2$ , since it rained twice out of four days.
2.  $P(B)$  is  $1/4$ , since James won only one race out of four.
3.  $P(A|B)=1$ , since it rained every time when James won.

Substituting the values in the conditional probability formula, we get the probability to be around 50%, which is almost the double of 25% when rain was not taken into account (Solve it at your end).

This further strengthened our belief of James winning in the light of new evidence i.e rain. You must be wondering that this formula bears close resemblance to something you might have heard a lot about. Think!

Probably, you guessed it right. It looks like Bayes Theorem.

Bayes theorem is built on top of conditional probability and lies in the heart of Bayesian Inference. Let's understand it in detail now.

### 3.2 Bayes Theorem

Bayes Theorem comes into effect when multiple events  $A_i$  form an exhaustive set with another event B. This could be understood with the help of the below diagram.

<b>A1</b>	<b>B</b>	
<b>A2</b>		
<b>A3</b>		

Now, B can be written as

$$B = \sum_{i=1}^n B \cap A_i$$

So, probability of B can be written as,

$$P(B) = \sum_{i=1}^n P(B \cap A_i)$$

But

$$P(B \cap A_i) = P(B|A_i) \times P(A_i)$$

So, replacing P(B) in the equation of conditional probability we get

$$P(A_i|B) = (P(B|A_i) \times P(A_i)) / (\sum_{i=1}^n (P(B|A_i) \times P(A_i)))$$

This is the equation of Bayes Theorem.

## 4. Bayesian Inference

There is no point in diving into the theoretical aspect of it. So, we'll learn how it works! Let's take an example of coin tossing to understand the idea behind *bayesian inference*.

An important part of *bayesian inference* is the establishment of *parameters* and *models*.

Models are the mathematical formulation of the observed events. Parameters are the factors in the models affecting the observed data. For example, in tossing a coin, fairness of coin may be defined as the parameter of coin denoted by  $\theta$ . The outcome of the events may be denoted by  $D$ .

Answer this now. What is the probability of 4 heads out of 9 tosses(D) given the fairness of coin ( $\theta$ ). i.e  $P(D|\theta)$

Wait, did I ask the right question? No.

We should be more interested in knowing : Given an outcome (D) what is the probability of coin being fair ( $\theta=0.5$ )

Lets represent it using Bayes Theorem:

$$P(\theta|D) = (P(D|\theta) \times P(\theta)) / P(D)$$

Here,  $P(\theta)$  is the *prior* i.e the strength of our belief in the fairness of coin before the toss. It is perfectly okay to believe that coin can have any degree of fairness between 0 and 1.

$P(D|\theta)$  is the likelihood of observing our result given our distribution for  $\theta$ . If we knew that coin was fair, this gives the probability of observing the number of heads in a particular number of flips.

$P(D)$  is the evidence. This is the probability of data as determined by summing (or integrating) across all possible values of  $\theta$ , weighted by how strongly we believe in those particular values of  $\theta$ .

*If we had multiple views of what the fairness of the coin is (but didn't know for sure), then this tells us the probability of seeing a certain sequence of flips for all possibilities of our belief in the coin's fairness.*

$P(\theta | D)$  is the posterior belief of our parameters after observing the evidence i.e the number of heads .

From here, we'll dive deeper into mathematical implications of this concept. Don't worry. Once you understand them, getting to its *mathematics* is pretty easy.

To define our model correctly , we need two mathematical models before hand. One to represent the *likelihood function*  $P(D|\theta)$  and the other for representing the distribution of *prior beliefs* . The product of these two gives the *posterior belief*  $P(\theta|D)$  distribution.

Since prior and posterior are both beliefs about the distribution of fairness of coin, intuition tells us that both should have the same mathematical form. Keep this in mind. We will come back to it again.

So, there are several functions which support the existence of bayes theorem. Knowing them is important, hence I have explained them in detail.

## 4.1. Bernoulli likelihood function

Lets recap what we learned about the likelihood function. So, we learned that:

*It is the probability of observing a particular number of heads in a particular number of flips for a given fairness of coin. This means our probability of observing heads/tails depends upon the fairness of coin ( $\theta$ ).*

$$P(y=1 | \theta) = \theta^y \quad [\text{If coin is fair } \theta=0.5, \text{ probability of observing heads (y=1) is 0.5}]$$

$$P(y=0 | \theta) = (1 - \theta)^{1-y} \quad [\text{If coin is fair } \theta=0.5, \text{ probability of observing tails(y=0) is 0.5}]$$

It is worth noticing that representing 1 as heads and 0 as tails is just a mathematical notation to formulate a model. We can combine the above mathematical definitions into a single definition to represent the probability of both the outcomes.

$$P(y|\theta) = \theta^y \cdot (1 - \theta)^{1-y}$$

This is called the Bernoulli Likelihood Function and the task of coin flipping is called Bernoulli's trials.

$$y=\{0,1\}, \theta=(0,1)$$

And, when we want to see a series of heads or flips, its probability is given by:

$$P(y_1, y_2, \dots, y_n | \theta) = \prod_1^n P(y_i | \theta)$$

$$P(y_1, y_2, \dots, y_n | \theta) = \prod_1^n \theta^{y_i} \cdot (1 - \theta)^{1-y_i}$$

Furthermore, if we are interested in the probability of number of heads  $z$  turning up in  $N$  number of flips then the probability is given by:

$$P(z, N | \theta) = \theta^z \cdot (1 - \theta)^{N-z}$$

## 4.2. Prior Belief Distribution

This distribution is used to represent our strengths on beliefs about the parameters based on the previous experience.

But, what if one has no previous experience?

Don't worry. Mathematicians have devised methods to mitigate this problem too. It is known as [uninformative priors](#). I would like to inform you beforehand that it is just a misnomer. Every uninformative prior always provides some information even the constant distribution prior.

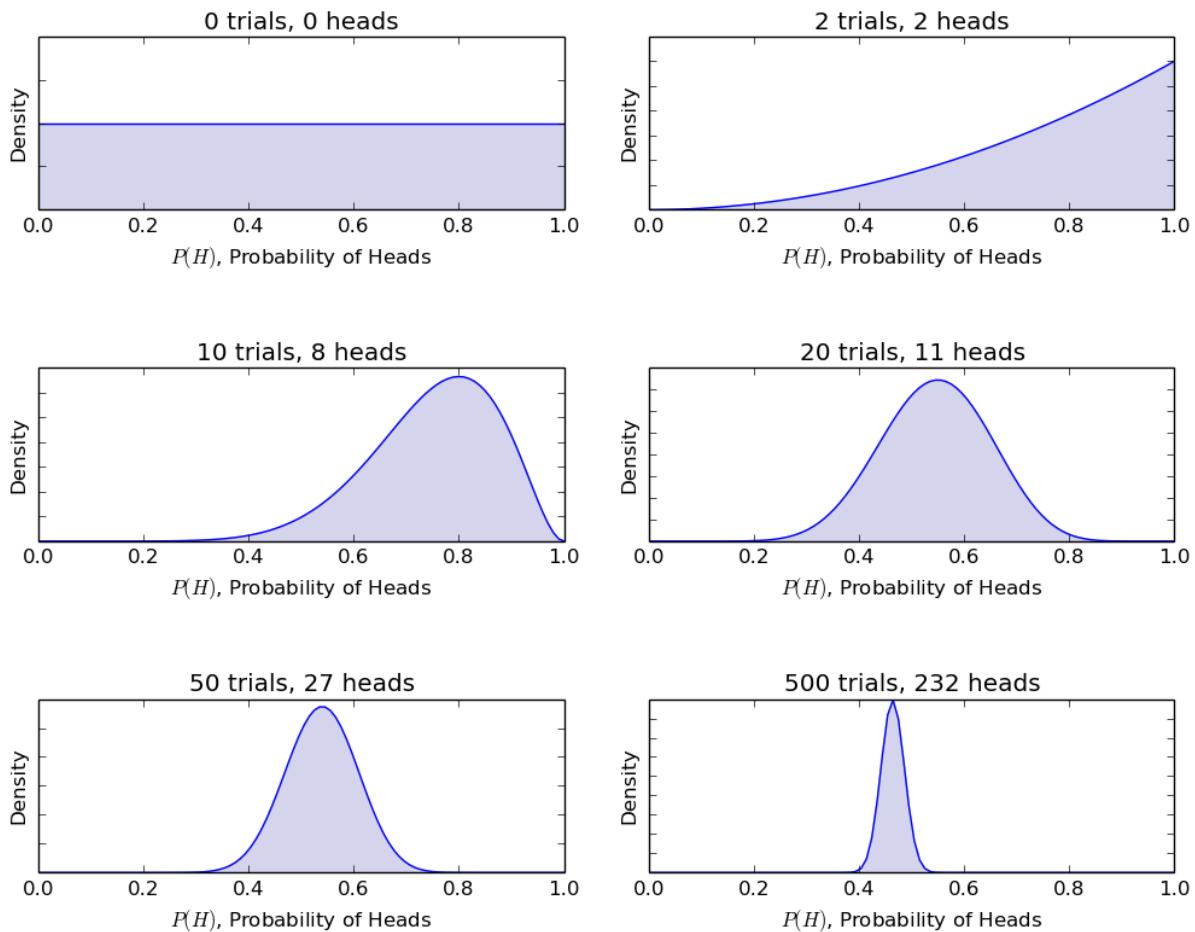
Well, the mathematical function used to represent the prior beliefs is known as *beta distribution*. It has some very nice mathematical properties which enable us to model our beliefs about a binomial distribution.

Probability density function of beta distribution is of the form :

$$x^{\alpha-1} \cdot (1-x)^{\beta-1} / B(\alpha, \beta)$$

where, our focus stays on numerator. The denominator is there just to ensure that the total probability density function upon integration evaluates to 1.

$\alpha$  and  $\beta$  are called the shape deciding parameters of the density function. Here  $\alpha$  is analogous to number of heads in the trials and  $\beta$  corresponds to the number of tails. The diagrams below will help you visualize the beta distributions for different values of  $\alpha$  and  $\beta$ .



You too can draw the beta distribution for yourself using the following code in R:

```
> library(stats)

> par(mfrow=c(3,2))

> x=seq(0,1,by=0.1)

> alpha=c(0,2,10,20,50,500)
```

```

> beta=c(0,2,8,11,27,232)

> for(i in 1:length(alpha)) {

  y<-dbeta(x,shape1=alpha[i],shape2=beta[i])

  plot(x,y,type="l")

}

```

Note:  $\alpha$  and  $\beta$  are intuitive to understand since they can be calculated by knowing the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of the distribution. In fact, they are related as :

$$\mu = \frac{\alpha}{\alpha + \beta}$$

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}}$$

If mean and standard deviation of a distribution are known , then there shape parameters can be easily calculated.

Inference drawn from graphs above:

1. When there was no toss we believed that every fairness of coin is possible as depicted by the flat line.
2. When there were more number of heads than the tails, the graph showed a peak shifted towards the right side, indicating higher probability of heads and that coin is not fair.
3. As more tosses are done, and heads continue to come in larger proportion the peak narrows increasing our confidence in the fairness of the coin value.

### 4.3. Posterior Belief Distribution

The reason that we chose prior belief is to obtain a beta distribution. This is because when we multiply it with a likelihood function, posterior distribution yields a form similar to the prior distribution which is much easier to relate to and understand. If this much information whets your appetite, I'm sure you are ready to walk an extra mile.

Let's calculate posterior belief using bayes theorem.

Calculating posterior belief using Bayes Theorem

$$P(\theta|z, N) = P(z, N|\theta)P(\theta)/P(z, N)$$

$$= \theta^z (1 - \theta)^{N-z} \cdot \theta^{\alpha-1} (1 - \theta)^{\beta-1} / [B(\alpha, \beta) P(z, N)]$$

$$= \theta^{z+\alpha-1} (1 - \theta)^{N-z+\beta-1} / [B(z + \alpha, N - z + \beta)]$$

Now, our posterior belief becomes,

$$P(\theta | z + \alpha, N - z + \beta)$$

This is interesting. Just knowing the mean and standard distribution of our belief about the parameter  $\theta$  and by observing the number of heads in  $N$  flips, we can update our belief about the model parameter( $\theta$ ).

Lets understand this with the help of a simple example:

Suppose, you think that a coin is biased. It has a mean ( $\mu$ ) bias of around 0.6 with standard deviation of 0.1.

Then ,

$$\alpha = 13.8, \beta = 9.2$$

i.e our distribution will be biased on the right side. Suppose, you observed 80 heads ( $z=80$ ) in 100 flips( $N=100$ ). Let's see how our prior and posterior beliefs are going to look:

```
prior = P(θ | α, β) = P(θ | 13.8, 9.2)
```

```
Posterior = P(θ | z+α, N-z+β) = P(θ | 93.8, 29.2)
```

Lets visualize both the beliefs on a graph:

The R code for the above graph is as:

```
> library(stats)

> x=seq(0,1,by=0.1)

> alpha=c(13.8,93.8)

> beta=c(9.2,29.2)

> for(i in 1:length(alpha)) {

  y<-dbeta(x, shape1=alpha[i], shape2=beta[i])
```

```
plot(x,y,type="l",xlab = "theta",ylab = "density")  
}
```

As more and more flips are made and new data is observed, our beliefs get updated. This is the real power of Bayesian Inference.

Conditional probability of evidence given hypothesis

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Prior probability of A before evidence observed

Posterior probability given the evidence

$$\therefore P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$P(A \cap B)$  = Marginal probability (Given data probability)

$$= P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Condition of probability

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$\Rightarrow P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(\text{King}|\text{face})$$

$$= \frac{P(\text{Face}|\text{King}) \cdot P(\text{King})}{P(\text{face})}$$

$$= \frac{1 \cdot \frac{4}{52}}{\frac{12}{52}}$$

$$= \frac{1}{13} / \frac{1}{13} = \frac{1}{3}$$

King, Jack, Queen,

Bayesian inference is a method of statistical inference in which Bayes theorem is used to update the probability for a hypothesis as more evidence or information become available.

### Tips and Tricks (cards)

#### terminology

- \* There are 13 cards of each suit, consisting of 1 ace, 3 face cards, and 9 number cards
- \* There are 4 aces, 12 face cards, and 36 number cards in a 52 card deck.
- \* Probability of drawing any card will always lie between 0 & 1
- \* The number of spades, hearts, diamonds, and clubs is same in every pack of 52 cards

## Conditional probability

It is probability of occurrence of a certain event say A, based on the occurrence of some other event say B.

This can be written as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

By Bayes theorem

Bayes theorem derived from the conditional probability of events. This theorem includes two conditional probabilities.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \rightarrow ①$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \rightarrow ②$$

from (1) & (2)

$$P(A|B) \cdot P(B) = P(A \cap B)$$

$$P(B|A) \cdot P(A) = P(A \cap B)$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad -(3)$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

### Problem

Suppose, B be the event of winning of James Hunt. A be the event of raining. Therefore

1.  $P(A) = 1/2$ , since it rained twice out of four days.

2.  $P(B) = 1/4$ , since James won only one race out of four.

3.  $P(A|B) = 1$ , since it rained every time when James won.

$$P(B|A) = \frac{P(A|B) P(B)}{P(A)}$$

$$= \frac{1 \cdot \frac{1}{2}}{\frac{1}{2}} = \frac{1}{2} = 0.5$$

Substituting the values in the conditional probability formula, we get the probability to be around 50% which is almost the double of 25% when rain was not taken into account (solve it on your own).

Bayes theorem is built on top of conditional probability and lies in the heart of Bayesian Inference.

# Bayesian Belief Network in artificial intelligence

Bayesian belief network is key computer technology for dealing with probabilistic events and to solve a problem which has uncertainty. We can define a Bayesian network as:

"A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph."

It is also called a **Bayes network, belief network, decision network, or Bayesian model.**

Bayesian networks are probabilistic, because these networks are built from a **probability distribution**, and also use probability theory for prediction and anomaly detection.

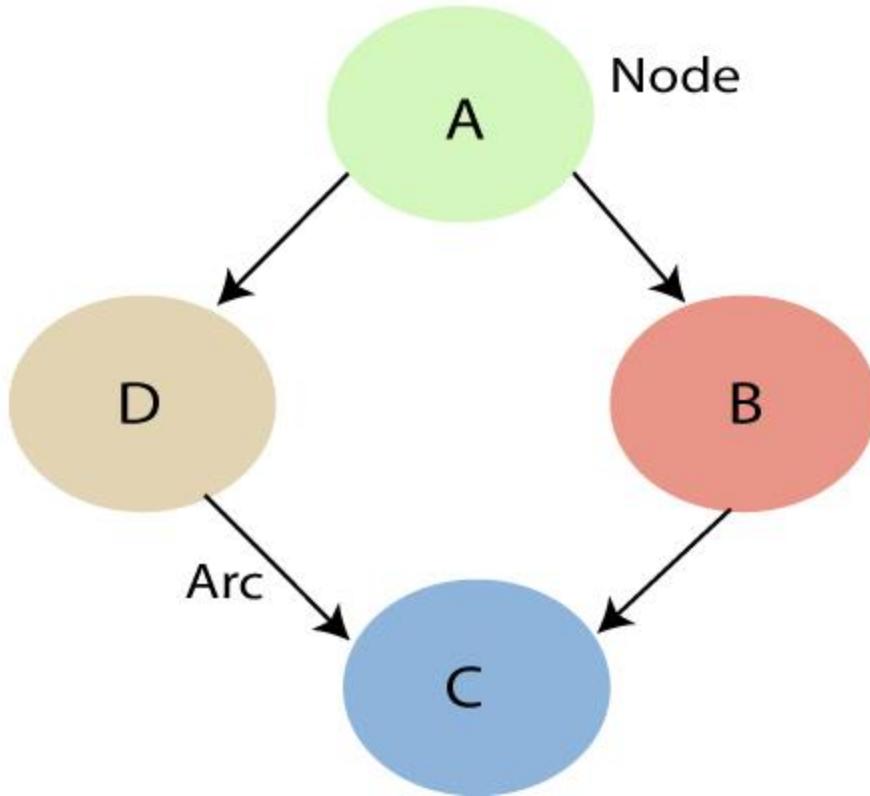
Real world applications are probabilistic in nature, and to represent the relationship between multiple events, we need a Bayesian network. It can also be used in various tasks including **prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction, and decision making under uncertainty.**

Bayesian Network can be used for building models from data and experts opinions, and it consists of two parts:

- **Directed Acyclic Graph**
- **Table of conditional probabilities.**

The generalized form of Bayesian network that represents and solve decision problems under uncertain knowledge is known as an **Influence diagram.**

**A Bayesian network graph is made up of nodes and Arcs (directed links), where:**



- Each **node** corresponds to the random variables, and a variable can be **continuous or discrete**.
- **Arc or directed arrows** represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph.

These links represent that one node directly influence the other node, and if there is no directed link that means that nodes are independent with each other

- In the above diagram, A, B, C, and D are random variables represented by the nodes of the network graph.
- If we are considering node B, which is connected with node A by a directed arrow, then node A is called the parent of Node B.
- Node C is independent of node A.

Note: The Bayesian network graph does not contain any cyclic graph. Hence, it is known as a **directed acyclic graph or DAG**.

The Bayesian network has mainly two components:

- **Causal Component**
- **Actual numbers**

Each node in the Bayesian network has condition probability distribution  $P(X_i | \text{Parent}(X_i))$ , which determines the effect of the parent on that node.

Bayesian network is based on Joint probability distribution and conditional probability. So let's first understand the joint probability distribution:

## Joint probability distribution:

If we have variables  $x_1, x_2, x_3, \dots, x_n$ , then the probabilities of a different combination of  $x_1, x_2, x_3, \dots, x_n$ , are known as Joint probability distribution.

$P[x_1, x_2, x_3, \dots, x_n]$ , it can be written as the following way in terms of the joint probability distribution.

$$\begin{aligned} &= P[x_1 | x_2, x_3, \dots, x_n] P[x_2 | x_3, \dots, x_n] \\ &= P[x_1 | x_2, x_3, \dots, x_n] P[x_2 | x_3, \dots, x_n] \dots P[x_{n-1} | x_n] P[x_n]. \end{aligned}$$

In general for each variable  $X_i$ , we can write the equation as:

$$P(X_i | X_{i-1}, \dots, X_1) = P(X_i | \text{Parents}(X_i))$$

## Explanation of Bayesian network:

Let's understand the Bayesian network through an example by creating a directed acyclic graph:

**Example:** Harry installed a new burglar alarm at his home to detect burglary. The alarm reliably responds at detecting a burglary but also responds for minor earthquakes. Harry has two neighbors David and Sophia, who have taken a responsibility to inform Harry at work when they hear the alarm. David always calls Harry when he hears the alarm, but sometimes he got confused with the phone ringing and calls at that time too. On the other hand, Sophia likes to listen to high music, so sometimes she misses to hear the alarm. Here we would like to compute the probability of Burglary Alarm.

**Problem:**

**Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and David and Sophia both called the Harry.**

**Solution:**

- The Bayesian network for the above problem is given below. The network structure is showing that burglary and earthquake is the parent node of the alarm and directly affecting the probability of alarm's going off, but David and Sophia's calls depend on alarm probability.
- The network is representing that our assumptions do not directly perceive the burglary and also do not notice the minor earthquake, and they also not confer before calling.
- The conditional distributions for each node are given as conditional probabilities table or CPT.
- Each row in the CPT must be sum to 1 because all the entries in the table represent an exhaustive set of cases for the variable.
- In CPT, a boolean variable with  $k$  boolean parents contains  $2^k$  probabilities. Hence, if there are two parents, then CPT will contain 4 probability values

**List of all events occurring in this network:**

- **Burglary (B)**
- **Earthquake(E)**

- Alarm(A)
- David Calls(D)
- Sophia calls(S)

We can write the events of problem statement in the form of probability:  $P[D, S, A, B, E]$ , can rewrite the above probability statement using joint probability distribution:

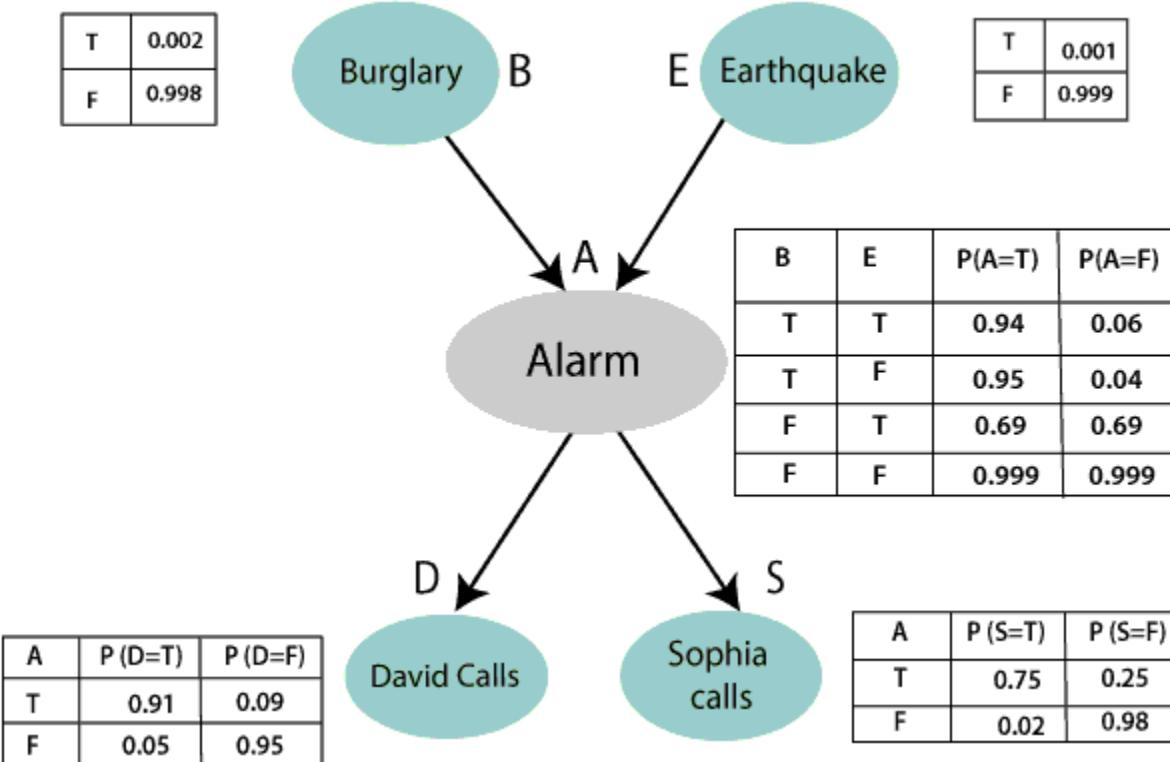
$$P[D, S, A, B, E] = P[D | S, A, B, E] \cdot P[S, A, B, E]$$

$$= P[D | S, A, B, E] \cdot P[S | A, B, E] \cdot P[A, B, E]$$

$$= P[D | A] \cdot P[S | A, B, E] \cdot P[A, B, E]$$

$$= P[D | A] \cdot P[S | A] \cdot P[A | B, E] \cdot P[B, E]$$

$$= P[D | A] \cdot P[S | A] \cdot P[A | B, E] \cdot P[B | E] \cdot P[E]$$



Let's take the observed probability for the Burglary and earthquake component:

$P(B= \text{True}) = 0.002$ , which is the probability of burglary.

$P(B= \text{False}) = 0.998$ , which is the probability of no burglary.

$P(E= \text{True}) = 0.001$ , which is the probability of a minor earthquake

$P(E= \text{False}) = 0.999$ , Which is the probability that an earthquake not occurred.

We can provide the conditional probabilities as per the below tables:

#### **Conditional probability table for Alarm A:**

The Conditional probability of Alarm A depends on Burglar and earthquake:

B	E	$P(A= \text{True})$	$P(A= \text{False})$
True	True	0.94	0.06
True	False	0.95	0.04
False	True	0.31	0.69
False	False	0.001	0.999

#### **Conditional probability table for David Calls:**

The Conditional probability of David that he will call depends on the probability of Alarm.

A	$P(D= \text{True})$	$P(D= \text{False})$
True	0.91	0.09
False	0.05	0.95

#### **Conditional probability table for Sophia Calls:**

The Conditional probability of Sophia that she calls is depending on its Parent Node "Alarm."

A	P(S= True)	P(S= False)
True	0.75	0.25
False	0.02	0.98

From the formula of joint distribution, we can write the problem statement in the form of probability distribution:

$$\begin{aligned}
 P(S, D, A, \neg B, \neg E) &= P(S|A) * P(D|A) * P(A|\neg B \wedge \neg E) * P(\neg B) * P(\neg E) \\
 &= 0.75 * 0.91 * 0.001 * 0.998 * 0.999 \\
 &= 0.00068045.
 \end{aligned}$$

**Hence, a Bayesian network can answer any query about the domain by using Joint distribution.**

#### **The semantics of Bayesian Network:**

There are two ways to understand the semantics of the Bayesian network, which is given below:

##### **1. To understand the network as the representation of the Joint probability distribution.**

It is helpful to understand how to construct the network.

##### **2. To understand the network as an encoding of a collection of conditional independence statements.**

It is helpful in designing inference procedure.

# Bayes' Theorem

The concept of *conditional probability* is introduced in *Elementary Statistics*. We noted that the conditional probability of an event is a probability obtained with the additional information that some other event has already occurred. We used  $P(B|A)$  to denote the conditional probability of event  $B$  occurring, given that event  $A$  has already occurred. The following formula was provided for finding  $P(B|A)$ :

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

In addition to the above formal rule, the textbook also included this "intuitive approach for finding a conditional probability":

The conditional probability of  $B$  given  $A$  can be found by assuming that event  $A$  has occurred and, working under that assumption, calculating the probability that event  $B$  will occur.

In this section we extend the discussion of conditional probability to include applications of *Bayes' theorem* (or *Bayes' rule*), which we use for revising a probability value based on additional information that is later obtained. One key to understanding the essence of Bayes' theorem is to recognize that we are dealing with *sequential events*, whereby new additional information is obtained for a subsequent event, and that new information is used to revise the probability of the initial event. In this context, the terms *prior probability* and *posterior probability* are commonly used.

## Definitions

**A prior probability** is an initial probability value originally obtained before any additional information is obtained.

**A posterior probability** is a probability value that has been revised by using additional information that is later obtained.

## Example 1

The Gallup organization randomly selects an adult American for a survey about credit card usage. Use subjective probabilities to estimate the following.

- What is the probability that the selected subject is a male?
- After selecting a subject, it is later learned that this person was smoking a cigar during the interview. What is the probability that the selected subject is a male?
- Which of the preceding two results is a prior probability? Which is a posterior probability?

## Solution

- a. Roughly half of all Americans are males, so we estimate the probability of selecting a male subject to be 0.5. Denoting a male by M, we can express this probability as follows:  $P(M) = 0.5$ .
- b. Although some women smoke cigars, the vast majority of cigar smokers are males. A reasonable guess is that 85% of cigar smokers are males. Based on this additional subsequent information that the survey respondent was smoking a cigar, we estimate the probability of this person being a male as 0.85. Denoting a male by M and denoting a cigar smoker by C, we can express this result as follows:  $P(M | C) = 0.85$ .
- c. In part (a), the value of 0.5 is the initial probability, so we refer to it as the prior probability. Because the probability of 0.85 in part (b) is a revised probability based on the additional information that the survey subject was smoking a cigar, this value of 0.85 is referred to a posterior probability.

The Reverend Thomas Bayes [1701 (approximately) – 1761] was an English minister and mathematician. Although none of his work was published during his lifetime, later (posterior?) publications included the following theorem (or rule) that he developed for determining probabilities of events by incorporating information about subsequent events.

## Bayes' Theorem

The probability of event A, given that event B has subsequently occurred, is

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{[P(A) \cdot P(B|A)] + [P(\bar{A}) \cdot P(B|\bar{A})]}$$

That's a formidable expression, but we will simplify its calculation. See the following example, which illustrates use of the above expression, but also see the alternative method based on a more intuitive application of Bayes' theorem.

## Example 2

In Orange County, 51% of the adults are males. (It doesn't take too much advanced mathematics to deduce that the other 49% are females.) One adult is randomly selected for a survey involving credit card usage.

- a. Find the prior probability that the selected person is a male.
- b. It is later learned that the selected survey subject was smoking a cigar. Also, 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars (based on data from the Substance Abuse and Mental Health Services Administration). Use this additional information to find the probability that the selected subject is a male.

## Solution

Let's use the following notation:

$$\begin{array}{ll} M = \text{male} & \bar{M} = \text{female (or not male)} \\ C = \text{cigar smoker} & \bar{C} = \text{not a cigar smoker.} \end{array}$$

- a. Before using the information given in part b, we know only that 51% of the adults in Orange County are males, so the probability of randomly selecting an adult and getting a male is given by  $P(M) = 0.51$ .
- b. Based on the additional given information, we have the following:

$$P(M) = 0.51 \quad \text{because 51\% of the adults are males}$$

$$P(\bar{M}) = 0.49 \quad \text{because 49\% of the adults are females (not males)}$$

$$P(C|M) = 0.095 \quad \text{because 9.5\% of the males smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a male, is 0.095.)}$$

$$P(C|\bar{M}) = 0.017. \quad \text{because 1.7\% of the females smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a female, is 0.017.)}$$

Let's now apply Bayes' theorem by using the preceding formula with M in place of A, and C in place of B. We get the following result:

$$\begin{aligned} P(M | C) &= \frac{P(M)! P(C|M)}{[P(M)! P(C|M)] + [P(\bar{M})! P(C|\bar{M})]} \\ &= \frac{0.51! 0.095}{[0.51! 0.095] + [0.49! 0.017]} \\ &= 0.85329341 \\ &= 0.853 \text{ (rounded)} \end{aligned}$$

Before we knew that the survey subject smoked a cigar, there is a 0.51 probability that the survey subject is male (because 51% of the adults in Orange County are males). However, after learning that the subject smoked a cigar, we revised the probability to 0.853. There is a 0.853 probability that the cigar-smoking respondent is a male. This makes sense, because the likelihood of a male increases dramatically with the additional information that the subject smokes cigars (because so many more males smoke cigars than females).

## Intuitive Bayes Theorem

The preceding solution illustrates the application of Bayes' theorem with its calculation using the formula. Unfortunately, that calculation is complicated enough to create an abundance of opportunities for errors and/or incorrect substitution of the involved probability values. Fortunately, here is another approach that is much more intuitive and easier:

**Assume some convenient value for the total of all items involved, then construct a table of rows and columns with the individual cell frequencies based on the known probabilities.**

For the preceding example, simply assume some value for the adult population of Orange County, such as 100,000, then use the given information to construct a table, such as the one shown below.

*Finding the number of males who smoke cigars:* If 51% of the 100,000 adults are males, then there are 51,000 males. If 9.5% of the males smoke cigars, then the number of cigar-smoking males is 9.5% of 51,000, or  $0.095 \times 51,000 = 4845$ . See the entry of 4845 in the table. The other males who do *not* smoke cigars must be  $51,000 - 4845 = 46,155$ . See the value of 46,155 in the table.

*Finding the number of females who smoke cigars:* Using similar reasoning, 49% of the 100,000 adults are females, so the number of females is 49,000. Given that 1.7% of the females smoke cigars, the number of cigar-smoking females is  $0.017 \times 49,000 = 833$ . The number of females who do *not* smoke cigars is  $49,000 - 833 = 48,167$ . See the entries of 833 and 48,167 in the table.

	C (Cigar Smoker)	$\bar{C}$ (Not a Cigar Smoker)	Total
M (male)	4845	46,155	<b>51,000</b>
$\bar{M}$ (female)	833	48,167	<b>49,000</b>
<b>Total</b>	<b>5678</b>	<b>94,322</b>	<b>100,000</b>

The above table involves relatively simple arithmetic. Simply partition the assumed population into the different cell categories by finding suitable percentages.

Now we can easily address the key question as follows: To find the probability of getting a male subject, given that the subject smokes cigars, simply use the same conditional probability described in the textbook. To find the probability of getting a male given that the subject smokes, restrict the table to the column of cigar smokers, then find the probability of getting a male in that column. Among the 5678 cigar smokers, there are 4845 males, so the probability we seek is  $4845/5678 = 0.85329341$ . That is,  $P(M | C) = 4845/5678 = 0.85329341 = 0.853$  (rounded).

## Bayes' Theorem Generalized

The preceding formula for Bayes' theorem and the preceding example use exactly two categories for event  $A$  (male and female), but the formula can be extended to include more than two categories. The following example illustrates this extension and it also illustrates a practical application of Bayes' theorem to quality control in industry. When dealing with more than the two events of  $A$  and  $\bar{A}$ , we must be sure that the multiple events satisfy two important conditions:

1. The events must be *disjoint* (with no overlapping).
2. The events must be *exhaustive*, which means that they combine to include all possibilities.

## Example 3

An aircraft emergency locator transmitter (ELT) is a device designed to transmit a signal in the case of a crash. The Altigauge Manufacturing Company makes 80% of the ELTs, the Bryant Company makes 15% of them, and the Chartair Company makes the other 5%. The ELTs made by Altigauge have a 4% rate of defects, the Bryant ELTs have a 6% rate of defects, and the Chartair ELTs have a 9% rate of defects (which helps to explain why Chartair has the lowest market share).

- a. If an ELT is randomly selected from the general population of all ELTs, find the probability that it was made by the Altigauge Manufacturing Company.
- b. If a randomly selected ELT is then tested and is found to be defective, find the probability that it was made by the Altigauge Manufacturing Company.

## Solution

We use the following notation:

$A$  = ELT manufactured by Altigauge

$B$  = ELT manufactured by Bryant

$C$  = ELT manufactured by Chartair

$D$  = ELT is defective

$\bar{D}$  = ELT is not defective (or it is good)

- a. If an ELT is randomly selected from the general population of all ELTs, the probability that it was made by Altigauge is 0.8 (because Altigauge manufactures 80% of them).
- b. If we now have the additional information that the ELT was tested and was found to be defective, we want to revise the probability from part (a) so that the new information can be used. We want to find the value of  $P(A|D)$ , which is the probability that the ELT was made by the Altigauge company given that it is defective. Based on the given information, we know these probabilities:

$P(A) = 0.80$  because Altigauge makes 80% of the ELTs  
 $P(B) = 0.15$  because Bryant makes 15% of the ELTs  
 $P(C) = 0.05$  because Chartair makes 5% of the ELTs

$P(D|A) = 0.04$  because 4% of the Altigauge ELTs are defective  
 $P(D|B) = 0.06$  because 6% of the Bryant ELTs are defective  
 $P(D|C) = 0.09$  because 9% of the Chartair ELTs are defective

Here is Bayes' theorem extended to include three events corresponding to the selection of ELTs from the three manufacturers (A, B, C):

$$\begin{aligned} P(A|D) &= \frac{P(A) \cdot P(D|A)}{P(A) \cdot P(D|A) + P(B) \cdot P(D|B) + P(C) \cdot P(D|C)} \\ &= \frac{0.80 \cdot 0.04}{[0.80 \cdot 0.04] + [0.15 \cdot 0.06] + [0.05 \cdot 0.09]} \\ &= 0.703 \text{ (rounded)} \end{aligned}$$

*Intuitive Baye's Theorem:* Now let's find  $P(A|D)$  by using a table. Let's arbitrarily assume that 10,000 ELTs were manufactured. (The solution doesn't depend on the number selected, but it's helpful to select a number large enough so that the cells in the table are all whole numbers.) Because 80% of the ELTs are made by Altigauge, we have 8000 ELTs made by Altigauge, and 4% of them (or 320) are defective. Also, if 320 of the Altigauge ELTs are defective, the other 7680 are not defective. See the values of 320 and 7680 in the table below. The other values are found using the same reasoning.

	D (defective)	$\bar{D}$ (not defective)	Total
A (Altigauge)	320	7680	<b>8,000</b>
B (Bryant)	90	1410	<b>1,500</b>
C (Chartair)	45	455	<b>500</b>
<b>Total</b>	<b>455</b>	<b>9545</b>	<b>10,000</b>

We want to find the probability that an ELT was made by Altigauge, given that it is known to be defective. Because we know the condition that the ELT is defective, we can refer to the first column of values where we see that among the 455 total defective ELTs, 320 were made by Altigauge, so that the probability is  $320/455 = 0.703$  (rounded). This is the same result obtained with the formula from Bayes' theorem.

The preceding example involve an extension of Bayes' theorem to three events denoted by A, B, C. Based on the format of the formula used in the solution, it is easy to extend Bayes' theorem so that it can be used with four or more events. (See Exercises 11 and 12.)

## Exercises

**Pregnancy Test Results.** In Exercises 1 and 2, refer to the results summarized in the table below.

	Positive Test Result (Pregnancy is indicated)	Negative Test Result (Pregnancy is not indicated)
Subject is Pregnant	80	5
Subject is Not Pregnant	3	11

1. a. If one of the 99 test subjects is randomly selected, what is the probability of getting a subject who is pregnant?  
b. A test subject is randomly selected and is given a pregnancy test. What is the probability of getting a subject who is pregnant, given that the test result is positive?
2. a. One of the 99 test subjects is randomly selected. What is the probability of getting a subject who is not pregnant?  
b. A test subject is randomly selected and is given a pregnancy test. What is the probability of getting a subject who is not pregnant, given that the test result is negative?
3. **Survey Results** In Orange County, 51% of the adults are males. One adult is randomly selected for a survey involving credit card usage. (See Example 2 in this section.)  
a. Find the prior probability that the selected person is a female.  
b. It is later learned that the selected survey subject was smoking a cigar. Also, 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars (based on data from the Substance Abuse and Mental Health Services Administration). Use this additional information to find the probability that the selected subject is a female.
4. **Emergency Locator Transmitters** An aircraft emergency locator transmitter (ELT) is a device designed to transmit a signal in the case of a crash. The Altigauge Manufacturing Company makes 80% of the ELTs, the Bryant Company makes 15% of them, and the Chartair Company makes the other 5%. The ELTs made by Altigauge have a 4% rate of defects, the Bryant ELTs have a 6% rate of defects, and the Chartair ELTs have a 9% rate of defects. (These are the same results from Example 3 in this section.)  
a. Find the probability of randomly selecting an ELT and getting one manufactured by the Bryant Company.  
b. If an ELT is randomly selected and tested, find the probability that it was manufactured by the Bryant Company if the test indicates that the ELT is defective.

5. ***Emergency Locator Transmitters*** Use the same ELT data from Exercise 4.
  - a. Find the probability of randomly selecting an ELT and getting one manufactured by the Chartair Company.
  - b. An ELT is randomly selected and tested. If the test indicates that the ELT is defective, find the probability that it was manufactured by the Chartair Company.
6. ***Emergency Locator Transmitters*** Use the same ELT data from Exercise 4. An ELT is randomly selected and tested. If the test indicates that the ELT is *not* defective, find the probability that it is from the Altigauge Company.
7. ***Pleas and Sentences*** In a study of pleas and prison sentences, it is found that 45% of the subjects studied were sent to prison. Among those sent to prison, 40% chose to plead guilty. Among those not sent to prison, 55% chose to plead guilty.
  - a. If one of the study subjects is randomly selected, find the probability of getting someone who was not sent to prison.
  - b. If a study subject is randomly selected and it is then found that the subject entered a guilty plea, find the probability that this person was not sent to prison.
8. ***Pleas and Sentences*** Use the same data given in Exercise 7.
  - a. If one of the study subjects is randomly selected, find the probability of getting someone who was sent to prison.
  - b. If a study subject is randomly selected and it is then found that the subject entered a guilty plea, find the probability that this person was sent to prison.
9. ***HIV*** The New York State Health Department reports a 10% rate of the HIV virus for the “at-risk” population. Under certain conditions, a preliminary screening test for the HIV virus is correct 95% of the time. (Subjects are not told that they are HIV infected until additional tests verify the results.) If someone is randomly selected from the at-risk population, what is the probability that they have the HIV virus if it is known that they have tested positive in the initial screening?
10. ***HIV*** Use the same data from Exercise 9. If someone is randomly selected from the at-risk population, what is the probability that they have the HIV virus if it is known that they have tested negative in the initial screening?
11. ***Extending Bayes' Theorem*** Example 3 in this section included an extension of Bayes' theorem to include three events, denoted by A, B, C. Write an expression that extends Bayes' theorem so that it can be used to find  $P(A|Z)$ , assuming that the initial event can occur in one of four ways: A, B, C, D.
12. ***Extensions of Bayes' Theorem*** In Example 2, we used only the initial events of A and  $\bar{A}$ . In Example 3, we used initial events of A, B, and C. If events B and C in Example 3 are combined and denoted as  $\bar{A}$ , we can find  $P(A|D)$  using the simpler format of Bayes' theorem given in Example 2. How would the resulting value of  $P(A|D)$  in Example 3 be affected by using this simplified approach?

## Answers to Odd–Numbered Exercises

1. a.  $85/99$  or  $0.859$

b.  $80/83$  or  $0.964$

3. a.  $0.49$

b.  $0.147$

5. a.  $0.05$

b.  $0.0989$

7. a.  $0.55$

b.  $0.627$

9.  $0.679$

11. 
$$P(A|Z) = \frac{P(A)! P(Z|A)}{P(C)! P(Z|C) + P(D)! P(Z|D)} [P(A)! P(Z|A)] + [P(B)! P(Z|B)] + [P(C)! P(Z|C)] + [P(D)! P(Z|D)]$$

**PRACTICE QUESTIONS ON BAYES'S FORMULA AND ON  
PROBABILITY  
(NOT TO BE HANDED IN )**

1. REMARKS

If you find any errors in this document, please alert me.

**Remark 1.** First, I'll make a remark about question 40 from section 12.4 in the book. Let A= event that first card is a spade and B=event that second card is a spade. As part of this question, you computed (presumably using the total law of probability) that

$$P(B) = P(A)P(B | A) + P(A^c)P(B | A^c) = \frac{13}{52} \times \frac{12}{51} + \frac{39}{52} \times \frac{13}{51} = \frac{1}{4}.$$

Note that in this case, of course, you already knew actually that

$$P(B) = \frac{13}{52} = \frac{1}{4},$$

since there are 13 spades in 52 cards, therefore the *unconditional* probability of B is  $\frac{13}{52}$ .

The law of total of total probability gives you a method for computing the unconditional (or total) probability of an event B if you know its conditional probabilities with respect to some other event A and the probability of A. In this case, we knew directly what  $P(B)$  is (because we had enough information- we know how many cards there are and how many spades) and you can see how it agrees with what the total law of probability gives you.

However, in most of the other examples, such as the one with the test for a virus we did in class, it's not possible to compute the probability of B (in that case, that the test is positive) directly because you don't have enough information (we don't know how many tests come out positive and how many tests are being administered, i.e., we don't know the percentage of tests that come out positive). What we know are the *conditional probabilities* of the test coming out positive with the conditions that the person taking it was infected or not. And we know the probability of this condition happening, i.e., we know the probability that someone is infected. So the information you have here consists of precisely the pieces that you need in order to use the total law of probability to compute the probability that a test comes out positive, and there's no other way to know this probability.

**Remark 2.** For all the following questions, the easiest way to think about them is to draw the tree diagram. Please do so when you try to do them, or when you read the solutions – draw the diagram to try to follow what's happening.

2. SOLUTIONS

**Exercise 1.** A doctor is called to see a sick child. The doctor has prior information that 90% of sick children in that neighborhood have the flu, while the other 10% are sick with

measles. Let  $F$  stand for an event of a child being sick with flu and  $M$  stand for an event of a child being sick with measles. Assume for simplicity that  $F \cup M = \Omega$ , i.e., that there no other maladies in that neighborhood.

A well-known symptom of measles is a rash (the event of having which we denote  $R$ ). Assume that the probability of having a rash if one has measles is  $P(R | M) = 0.95$ . However, occasionally children with flu also develop rash, and the probability of having a rash if one has flu is  $P(R | F) = 0.08$ .

Upon examining the child, the doctor finds a rash. What is the probability that the child has measles?

**Solution.**

We use Bayes's formula.

$$\begin{aligned} P(M | R) &= \frac{P(R | M)P(M)}{(P(R | M)P(M) + P(R | F)P(F))} \\ &= \frac{0.95 \times 0.10}{(0.95 \times 0.10 + 0.08 \times 0.90)} \simeq 0.57. \end{aligned}$$

Which is nowhere close to 95% of  $P(R|M)$ .

**Exercise 2.** In a study, physicians were asked what the odds of breast cancer would be in a woman who was initially thought to have a 1% risk of cancer but who ended up with a positive mammogram result (a mammogram accurately classifies about 80% of cancerous tumors and 90% of benign tumors.)

95 out of a hundred physicians estimated the probability of cancer to be about 75%. Do you agree?

**Solution.**

Introduce the events:

$+$  = mammogram result is positive,

$B$  = tumor is benign,

$M$  = tumor is malignant.

Note that  $B^c = M$ . We are given  $P(M) = .01$ , so  $P(B) = 1 - P(M) = .99$ . We are also given the conditional probabilities  $P(+ | M) = .80$  and  $P(- | B) = .90$ , where the event  $-$  is the complement of  $+$ , thus  $P(+ | B) = .10$

Bayes' formula in this case is

$$\begin{aligned} P(M | +) &= \frac{P(+ | M)P(M)}{(P(+ | M)P(M) + P(+ | B)P(B))} \\ &= \frac{0.80 \times 0.01}{(0.80 \times 0.01 + 0.10 \times 0.99)} \\ &\simeq 0.075 \end{aligned}$$

So the chance would be 7.5%. A far cry from a common estimate of 75

**Exercise 3.** Suppose we have 3 cards identical in form except that both sides of the first card are colored red, both sides of the second card are colored black, and one side of the third card is colored red and the other side is colored black.

The 3 cards are mixed up in a hat, and 1 card is randomly selected and put down on the ground. If the upper side of the chosen card is colored red, what is the probability that the other side is colored black?

**Solution.**

Let RR, BB, and RB denote, respectively, the events that the chosen card is the red-red, the black-black, or the red-black card. Letting R be the event that the upturned side of the chosen card is red, we have that the desired probability is obtained by

$$\begin{aligned} P(RB | R) &= \frac{P(RB \cap R)}{P(R)} \\ &= \frac{P(R | RB)P(RB)}{P(R | RR)P(RR) + P(R | RB)P(RB) + P(R | BB)P(BB)} \\ &= \frac{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)}{\left(1\right)\left(\frac{1}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + 0\left(\frac{1}{3}\right)} = \frac{1}{3} \end{aligned}$$

This question was actually just like the Monty Hall problem!

**Exercise 4.** It is estimated that 50% of emails are spam emails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%.

Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?

**Solution.**

Define events

$A$  = event that an email is detected as spam,

$B$  = event that an email is spam,

$B^c$  = event that an email is not spam.

We know  $P(B) = P(B^c) = .5$ ,  $P(A | B) = 0.99$ ,  $P(A | B^c) = 0.05$ .

Hence by the Bayes's formula we have

$$\begin{aligned} P(B^c | A) &= \frac{P(A | B^c)P(B^c)}{P(A | B)P(B) + P(A | B^c)P(B^c)} \\ &= \frac{0.05 \times 0.5}{0.05 \times 0.5 + 0.99 \times 0.5} \\ &= \frac{5}{104}. \end{aligned}$$

## **What is an SVM?**

Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. All of these are common tasks in machine learning.

You can use them to detect cancerous cells based on millions of images or you can use them to predict future driving routes with a well-fitted regression model.

There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC).

The main thing to keep in mind here is that these are just math equations tuned to give you the most accurate answer possible as quickly as possible.

SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

## **How an SVM works**

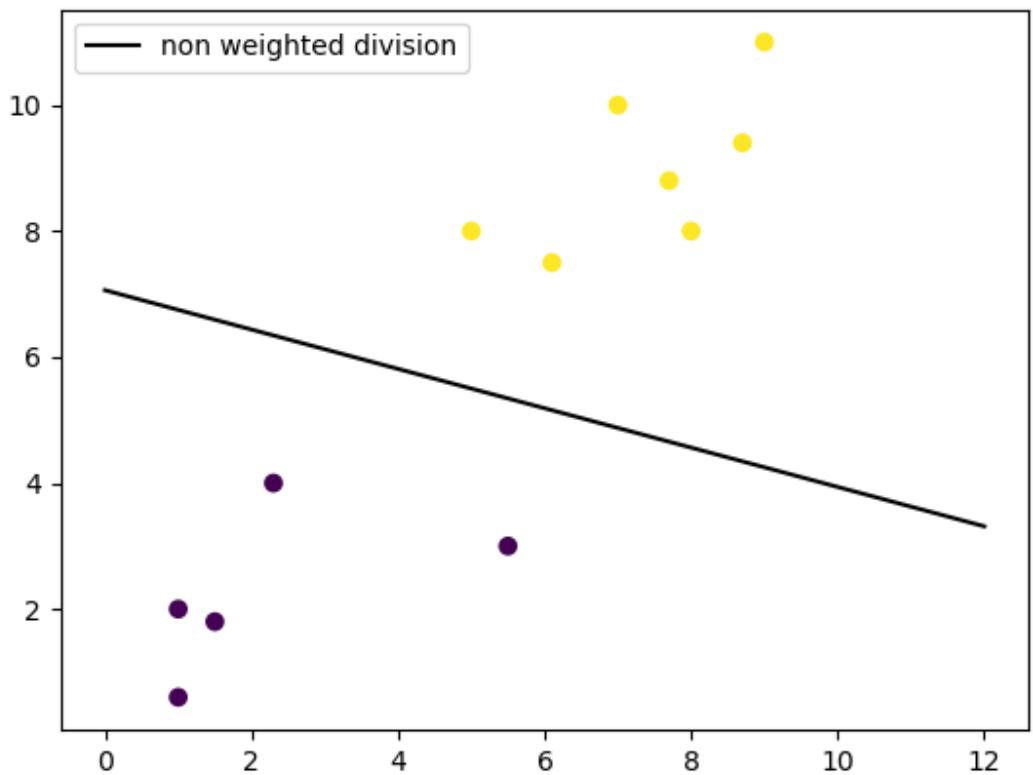
A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on

the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from.

What makes the linear SVM algorithm better than some of the other algorithms, like k-nearest neighbors, is that it chooses the best line to classify your data points. It chooses the line that separates the data and is the furthest away from the closest data points as possible.

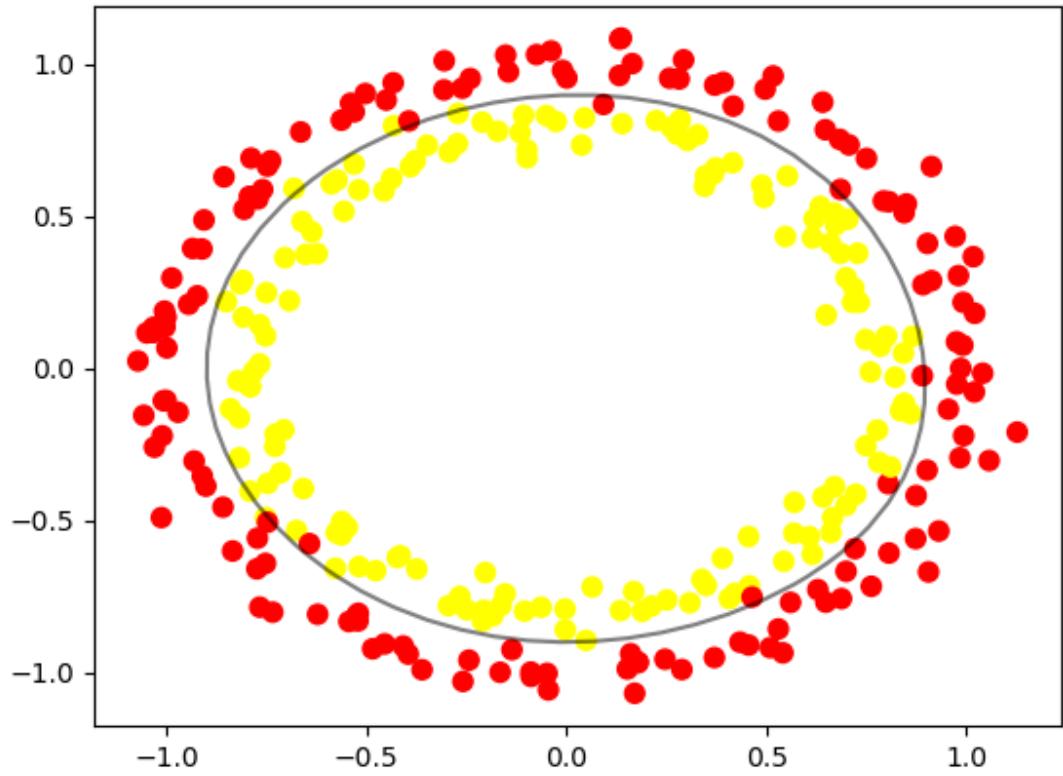
A 2-D example helps to make sense of all the machine learning jargon. Basically you have some data points on a grid. You're trying to separate these data points by the category they should fit in, but you don't want to have any data in the wrong category. That means you're trying to find the line between the two closest points that keeps the other data points separated.

So the two closest data points give you the support vectors you'll use to find that line. That line is called the decision boundary.



### linear SVM

The decision boundary doesn't have to be a line. It's also referred to as a hyperplane because you can find the decision boundary with any number of features, not just two.



non-linear SVM using RBF kernel

### Types of SVMs

There are two different types of SVMs, each used for different things:

- Simple SVM: Typically used for linear regression and classification problems.
- Kernel SVM: Has more flexibility for non-linear data because you can add more features to fit a hyperplane instead of a two-dimensional space.

## Why SVMs are used in machine learning

SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. This is one of the reasons we use SVMs in machine learning. It can handle both classification and regression on linear and non-linear data.

Another reason we use SVMs is because they can find complex relationships between your data without you needing to do a lot of transformations on your own. It's a great option when you are working with smaller datasets that have tens to hundreds of thousands of features. They typically find more accurate results when compared to other algorithms because of their ability to handle small, complex datasets.

Here are some of the pros and cons for using SVMs.

### Pros

- Effective on datasets with multiple features, like financial or medical data.
- Effective in cases where number of features is greater than the number of data points.
- Uses a subset of training points in the decision function called support vectors which makes it memory efficient.

- Different kernel functions can be specified for the decision function. You can use common kernels, but it's also possible to specify custom kernels.

### Cons

- If the number of features is a lot bigger than the number of data points, avoiding over-fitting when choosing kernel functions and regularization term is crucial.
- SVMs don't directly provide probability estimates. Those are calculated using an expensive five-fold cross-validation.
- Works best on small sample sets because of its high training time.

Since SVMs can use any number of kernels, it's important that you know about a few of them.

## Kernel functions

### Linear

These are commonly recommended for text classification because most of these types of classification problems are linearly separable.

The linear kernel works really well when there are a lot of features, and text classification problems have a lot of features. Linear kernel functions are faster than most of the others and you have fewer parameters to optimize.

Here's the function that defines the linear kernel:

$$f(X) = w^T * X + b$$

In this equation, **w** is the weight vector that you want to minimize, **X** is the data that you're trying to classify, and **b** is the linear coefficient estimated from the training data. This equation defines the decision boundary that the SVM returns.

### Polynomial

The polynomial kernel isn't used in practice very often because it isn't as computationally efficient as other kernels and its predictions aren't as accurate.

Here's the function for a polynomial kernel:

$$f(X_1, X_2) = (a + X_1^T * X_2) ^ b$$

This is one of the more simple polynomial kernel equations you can use.  $f(\mathbf{X}_1, \mathbf{X}_2)$  represents the polynomial decision boundary that will separate your data.  $\mathbf{X}_1$  and  $\mathbf{X}_2$  represent your data.

### Gaussian Radial Basis Function (RBF)

One of the most powerful and commonly used kernels in SVMs. Usually the choice for non-linear data.

Here's the equation for an RBF kernel:

$$f(\mathbf{X}_1, \mathbf{X}_2) = \exp(-\gamma * \|\mathbf{X}_1 - \mathbf{X}_2\|^2)$$

In this equation, **gamma** specifies how much a single training point has on the other data points around it.  $\|\mathbf{X}_1 - \mathbf{X}_2\|$  is the dot product between your features.

### Sigmoid

More useful in neural networks than in support vector machines, but there are occasional specific use cases.

Here's the function for a sigmoid kernel:

$$f(\mathbf{x}, \mathbf{y}) = \tanh(\alpha * \mathbf{x}^T * \mathbf{y} + C)$$

In this function, **alpha** is a weight vector and **C** is an offset value to account for some mis-classification of data that can happen.

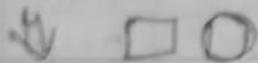
### Others

There are plenty of other kernels you can use for your project. This might be a decision to make when you need to meet certain error constraints, you want to try and speed up the training time, or you want to super tune parameters.

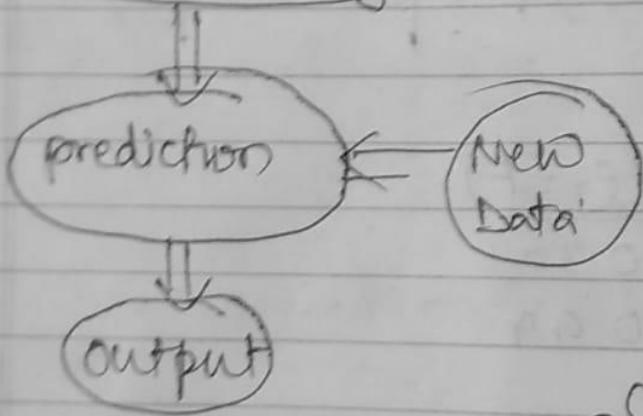
# Support vector Machine

- Used for classification and regression analysis

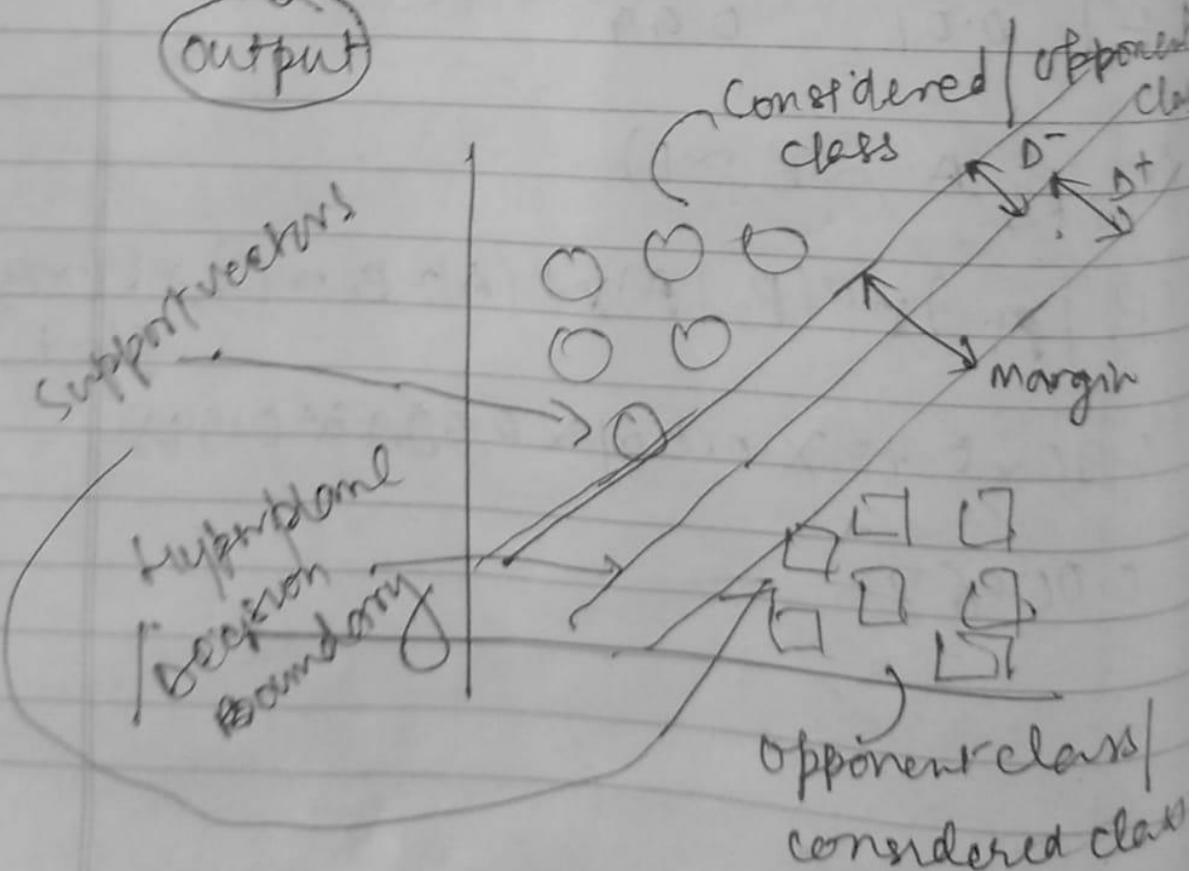
- Label data



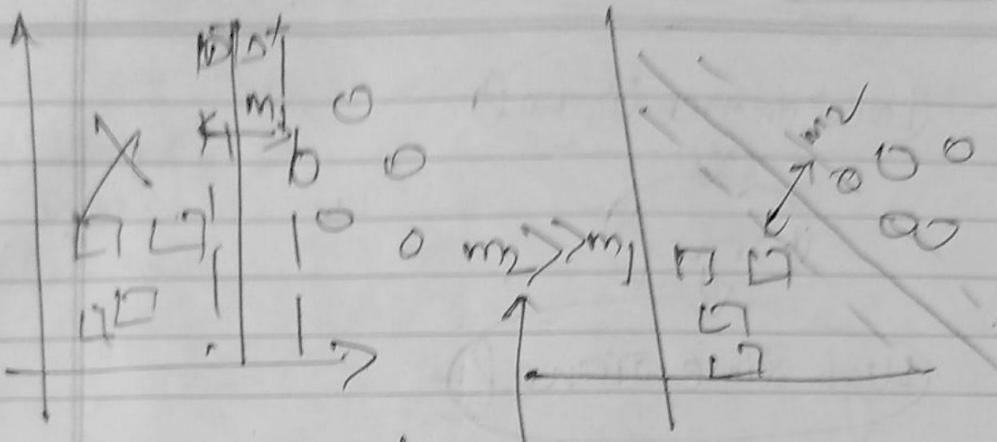
- Training purpose  
(Model Training)



□ ○ ?



## Linear separable data

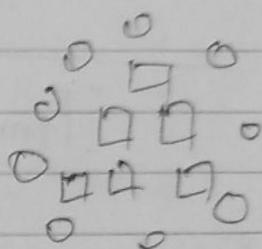


hyperplane angle

select  
the max size  
margin for  
future prediction/ classification

max margin hyperplane

should get selected bcz Non-linear separable  
reduce error & more wonder future  
prediction.



## Synopsis of Time Series Analysis

- A Time-Series represents a series of time-based orders. It would be Years, Months, Weeks, Days, Hours, Minutes, and Seconds
- A time series is an observation from the sequence of discrete-time of successive intervals.
- A time series is a running chart.
- The time variable/feature is the independent variable and supports the target variable to predict the results.
- Time Series Analysis (TSA) is used in different fields for time-based predictions – like Weather Forecasting, Financial, Signal processing, Engineering domain – Control Systems, Communications Systems.
- Since TSA involves producing the set of information in a particular sequence, it makes a distinct from spatial and other analyses.
- Using AR, MA, ARMA, and ARIMA models, we could predict the future.

## Introduction to Time Series Analysis

Time Series Analysis is the way of studying the characteristics of the response variable with respect to time, as the independent variable. To estimate the target variable in the name of predicting or forecasting, use the time variable as the point of reference. In this article we will discuss in detail TSA Objectives, Assumptions, Components (stationary, and Non- stationary). Along with the TSA algorithm and specific use cases in Python.

### Table of Contents

1. What is Time Series Analysis (TSA) and its assumption
2. How to analyze Time Series (TSA)?
3. Time Series Analysis Significance and its types.
4. Components of Time Series
5. What are the limitations of time series?
6. Detailed Study of Time Series Data types.
7. Discussion on stationary and Non- stationary components
8. Conversion of Non- stationary into stationary

9. Why is Time Series Analysis used in Data Science and Machine Learning?
10. Time Series Analysis in Data Science and Machine Learning
11. Implementation of Auto-Regressive model
12. Implementation of Moving Average (WEIGHTS – SIMPLE MOVING AVERAGE)
13. Understanding ARMA and ARIMA
14. Implementation steps for ARIMA
15. Time Series Analysis – Process flow (Re-gap)

## What is Time Series Analysis

Definition: If you see, there are many more definitions for TSA. But make it simple.

A *time series* is nothing but a sequence of various data points that occurred in a successive order for a given period of time

Objectives:

- To understand how time series works, what factors are affecting a certain variable(s) at different points of time.
- Time series analysis will provide the consequences and insights of features of the given dataset that changes over time.
- Supporting to derive the predicting the future values of the time series variable.
- Assumptions: There is one and the only assumption that is “stationary”, which means that the origin of time, does not affect the properties of the process under the statistical factor.

## How to analyze Time Series?

Quick steps here for your reference, anyway. Will see this in detail in this article later.

- Collecting the data and cleaning it
- Preparing Visualization with respect to time vs key feature
- Observing the stationarity of the series
- Developing charts to understand its nature.
- Model building – AR, MA, ARMA and ARIMA
- Extracting insights from prediction

# Significance of Time Series and its types

TSA is the backbone for prediction and forecasting analysis, specific to the time-based problem statements.

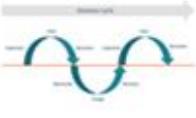
- Analyzing the historical dataset and its patterns
- Understanding and matching the current situation with patterns derived from the previous stage.
- Understanding the factor or factors influencing certain variable(s) in different periods.

With help of “Time Series” we can prepare numerous time-based analyses and results.

- Forecasting
- Segmentation
- Classification
- Descriptive analysis`
- Intervention analysis

## Components of Time Series Analysis

- Trend
- Seasonality
- Cyclical
- Irregularity
- Trend: In which there is no fixed interval and any divergence within the given dataset is a continuous timeline. The trend would be Negative or Positive or Null Trend
- Seasonality: In which regular or fixed interval shifts within the dataset in a continuous timeline. Would be bell curve or saw tooth
- Cyclical: In which there is no fixed interval, uncertainty in movement and its pattern
- Irregularity: Unexpected situations/events/scenarios and spikes in a short time span.

	Trend	Seasonality	Cyclical	Irregularity
Time	Fixed Time Interval	Fixed Time Interval	Not Fixed Time Interval	Not Fixed Time Interval
Duration	Long and Short Term	Short Term	Long and Short Term	Regular/Irregular
Visualization				
Nature - I	Gradual	Swings between Up or Down	Repeating Up and Down	Errored or High Fluctuation
Nature - II	Upward/Down Trend	Pattern repeatable	No fixed period	Short and Not repeatable
Prediction Capability	Predictable	Predictable	Challenging	Challenging
Market Model				Highly random/Unforeseen Events – along with white noise.

## What are the limitations of Time Series Analysis?

Time series has the below-mentioned limitations, we have to take care of those during our analysis,

- Similar to other models, the missing values are not supported by TSA
- The data points must be linear in their relationship.
- Data transformations are mandatory, so a little expensive.
- Models mostly work on Uni-variate data.

## Data Types of Time Series

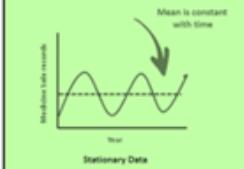
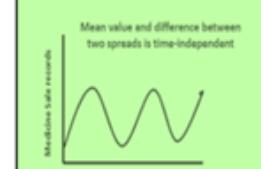
Let's discuss the time series' data types and their influence. While discussing TS data-types, there are two major types.

- Stationary
- Non- Stationary

6.1 Stationary: A dataset should follow the below thumb rules, without having Trend, Seasonality, Cyclical, and Irregularity component of time series

- The MEAN value of them should be completely constant in the data during the analysis
- The VARIANCE should be constant with respect to the time-frame
- The COVARIANCE measures the relationship between two variables.

6.2 Non- Stationary: This is just the opposite of Stationary.

	MEAN	Variance	Covariance
Stationary			
Non-Stationary			

## Methods to check Stationarity

During the TSA model preparation workflow, we must access if the given dataset is Stationary or NOT. Using Statistical and Plots test.

7.1 Statistical Test: There are two tests available to test if the dataset is Stationary or NOT.

- Augmented Dickey-Fuller (ADF) Test
  - Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test
- 7.1.1 Augmented Dickey-Fuller (ADF) Test or Unit Root Test: The ADF test is the most popular statistical test and with the following assumptions.
- Null Hypothesis ( $H_0$ ): Series is non-stationary
  - Alternate Hypothesis ( $H_A$ ): Series is stationary
    - p-value  $> 0.05$  Fail to reject ( $H_0$ )
    - p-value  $\leq 0.05$  Accept ( $H_1$ )

7.1.2 Kwiatkowski–Phillips–Schmidt–Shin (KPSS): these tests are used for testing a NULL Hypothesis ( $H_0$ ), that will perceive the time-series, as stationary around a deterministic trend against the alternative of a unit root. Since TSA looking for Stationary Data for its further analysis, we have to make sure that the dataset should be stationary.

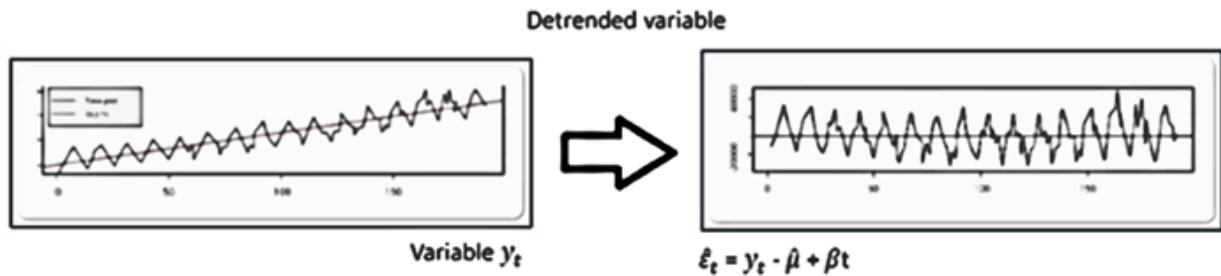
## Converting Non- stationary into stationary

Let's discuss quickly how to convert Non- stationary into stationary for effective time series modeling. There are two major methods available for this conversion.

- Detrending

- Differencing
- Transformation

8.1 Detrending: It involves removing the trend effects from the given dataset and showing only the differences in values from the trend. it always allows the cyclical patterns to be identified.



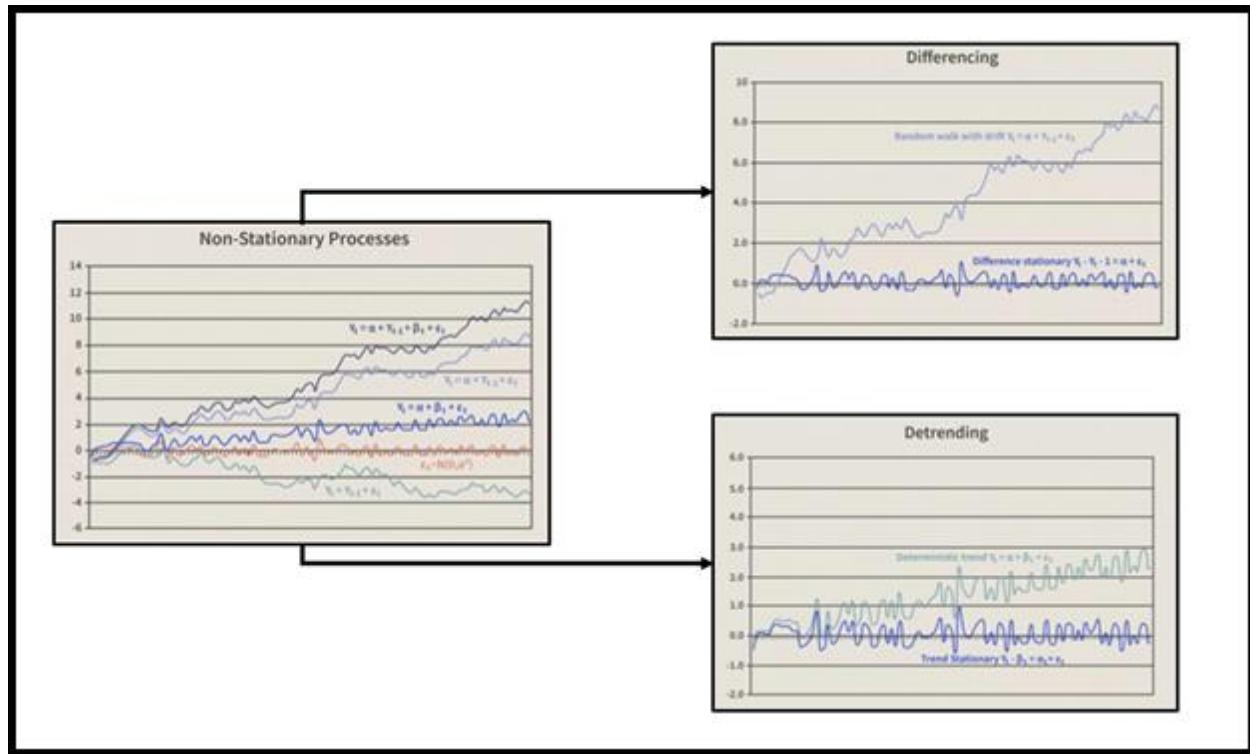
Designed by Author (Shanthababu)

8.2 Differencing: This is a simple transformation of the series into a new time series, which we use to remove the series dependence on time and stabilize the mean of the time series, so trend and seasonality are reduced during this transformation.

$$Y_t = Y_t - Y_{t-1}$$

$Y_t$ =Value with time

Detrending and Differencing extractions



8.3 Transformation: This includes three different methods they are Power Transform, Square Root, and Log Transfer., most commonly used one is Log Transfer.

## Time series Analysis

\* Future prediction

eg: stock price

\* One variable study

eg:- demand of the product

\* Sequence of data maintained

\* Analysis done yearly, quarterly, monthly

eg GDP, closing price, temperature

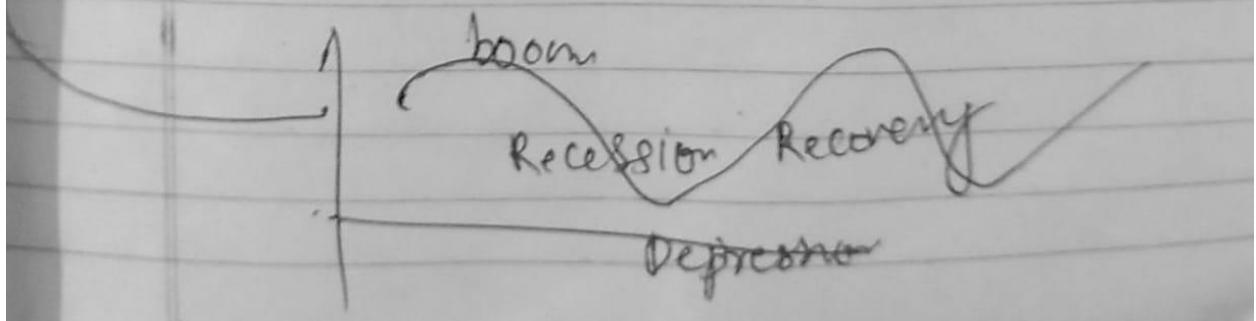
- Understand, interpret and access chronological changes in the value of a variable in the past, so that reliable prediction can be made about future values.
- ~~Examine~~ ~~Examine~~ ~~traces~~ ?

- Components of time series

- i) T - secular trend: long term
- ii) S - seasonal variations <sup>time lag</sup>
- iii) C - cyclical variations → Business cycle
- iv) I - irregular variations <sup>random</sup>  
Random → earthquakes ~~Random~~  
                  floods ~~floods~~

$$X_t = T \times S \times C \times I \quad \text{multiplicative model}$$

$$X_t = T + S + C + I \quad \text{Addition Model}$$



## 3 methods

1. Graphical method (free hand)
2. moving Average = average of every period
3. Least Square method.

### 1. Graphical

Year Y

1999 19.3

2000 20.9

2001 17.8

2002 16.1

2003 17.6

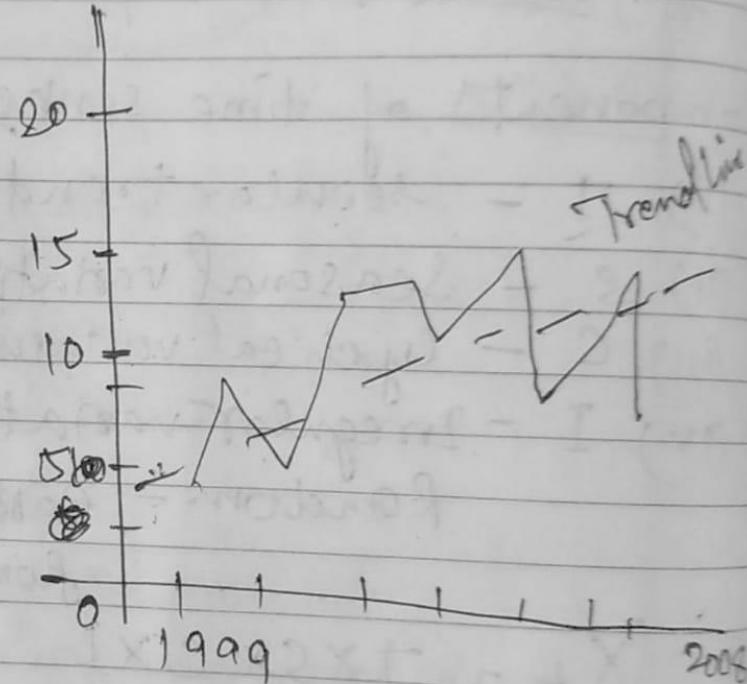
2004 17.8

2005 18.3

2006 17.3

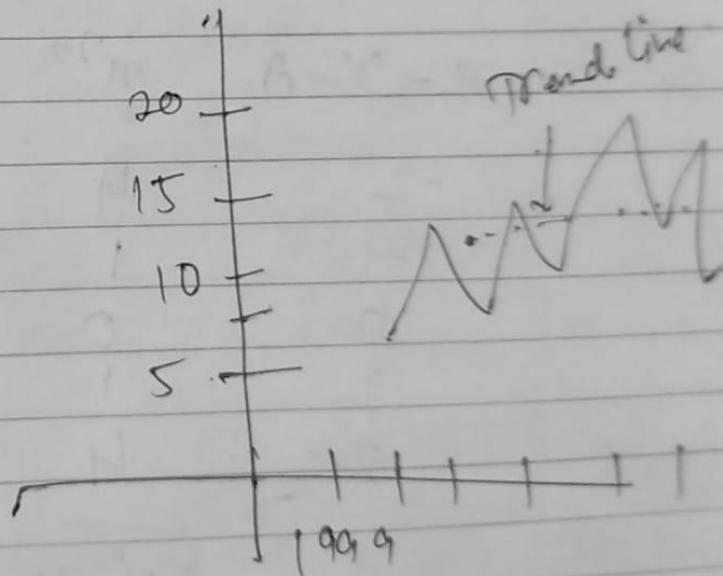
2007 21.4

2008 19.3



## ② moving Average

Year	Y	5 year moving total	5 Average
1999	19.3		
2000	20.9		
2001	17.8	91.7	$91.7 \div 5 = 18.34$
2002	16.1	90.2	$90.2 \div 5 = 18.04$
2003	17.6	87.6	$87.6 \div 5 = 17.52$
2004	17.8	87.1	$87.1 \div 5 = 17.42$
2005	18.3	92.4	$92.4 \div 5 = 18.48$
2006	17.3	94.1	$94.1 \div 5 = 18.82$
2007	21.4		
2008	19.3		





# least square method

deviations  
from  
mean

BEST

X      Y  
Year   Sales

2015	30
2016	50
2017	75
2018	80
2019	40

$$\sum Y = 275$$

trend equation -  $y = a + bx$

$$Y = a + b \times x$$

$$a = \frac{\sum Y}{N} \quad b = \frac{\sum x \cdot Y}{\sum x^2}$$

$$= \frac{275}{5} \quad = 55$$

$$b = \frac{50}{10}$$

$$= 5$$

$$x = X - A \quad x^2 \quad x \cdot Y \quad \text{Trend Eqn}$$

$$y = a + bx$$

-2	4	-60	$55 + 5 \times -2 = 45$
-1	1	-50	$55 + 5 \times -1 = 50$
0	0	0	$55 + 5 \times 0 = 55$
1	1	80	$55 + 5 \times 1 = 60$
2	4	88	$55 + 5 \times 2 = 65$

$$\sum x^2 = 10 \quad \sum x \cdot Y = 50$$

future  
Sale

2025 - Trend value

# **SOFT COMPUTING**

By

K.Sai Saranya,Assistant Professor,Department of CSE

# **UNIT-1**

## **INTRODUCTION TO ARTIFICIAL NEURAL NETWORKS (ANN)**

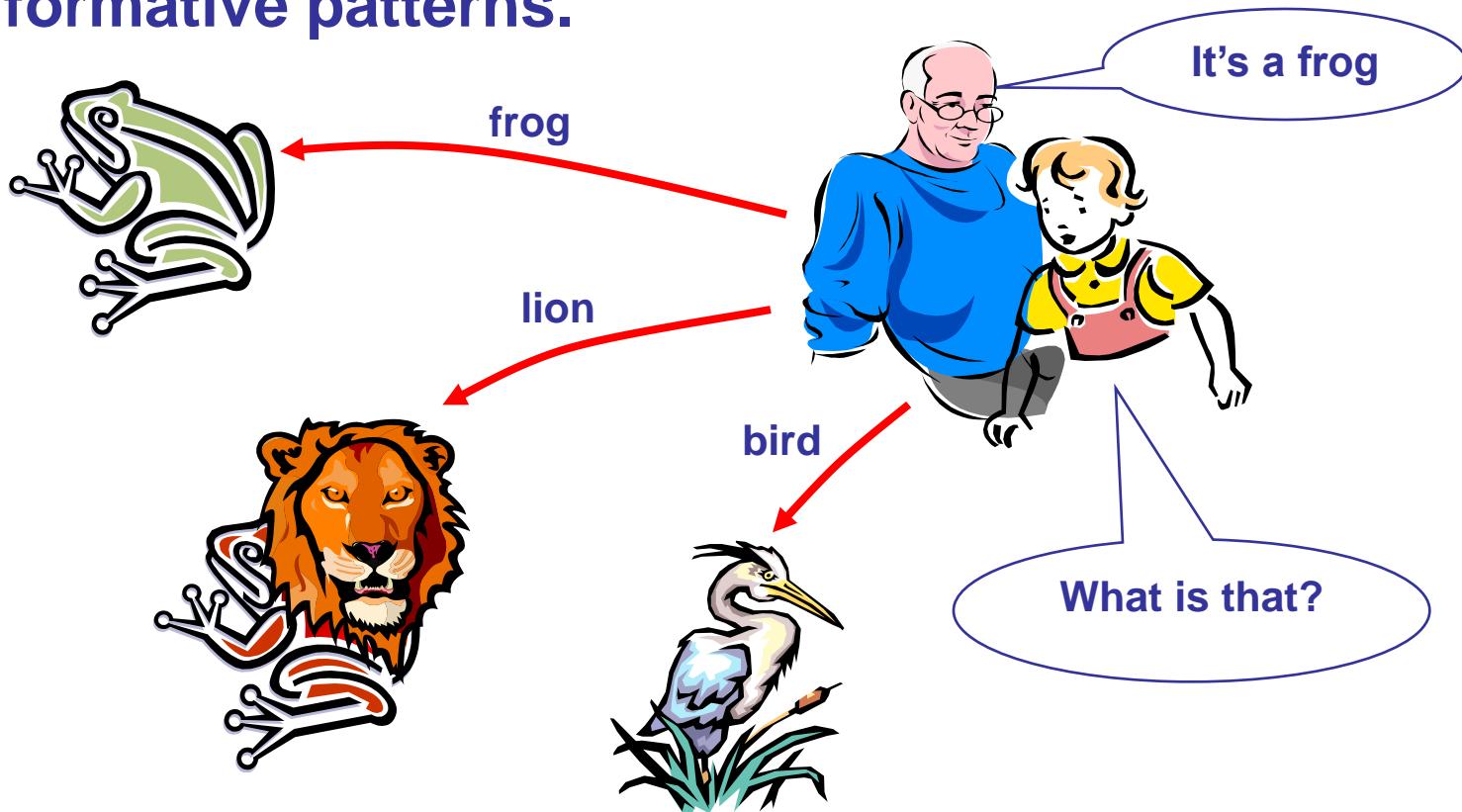
# *Outline*

---

- **Definition, why and how are neural networks being used in solving problems**
- **Human biological neuron**
- **Artificial Neuron**
- **Applications of ANN**
- **Comparison of ANN vs conventional AI methods**

# *The idea of ANNs..?*

- NNs learn relationship between cause and effect or organize large volumes of data into orderly and informative patterns.



# *Neural networks to the rescue...*

- **Neural network:** *information processing paradigm inspired by biological nervous systems, such as our brain*
- Structure: large number of highly interconnected processing elements (*neurons*) working together
- Like people, they learn *from experience* (by example)

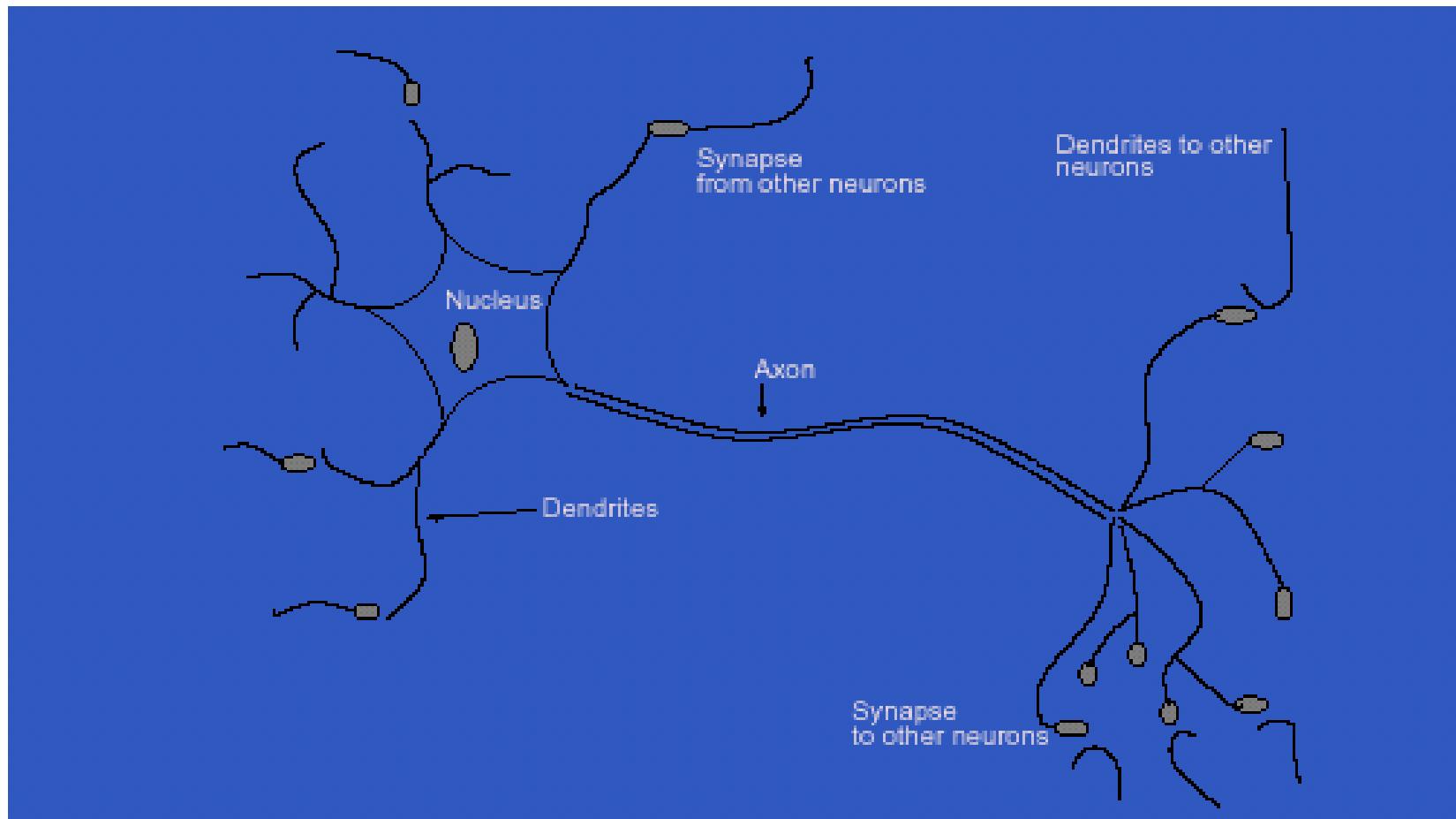
# *Definition of ANN*

“Data processing system consisting of a large number of simple, highly interconnected processing elements (artificial neurons) in an architecture inspired by the structure of the cerebral cortex of the brain”

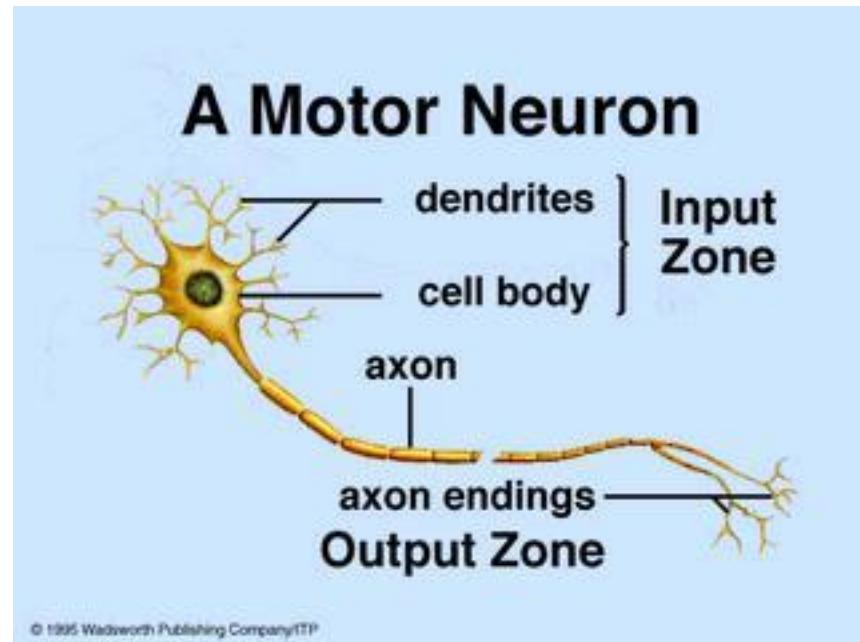
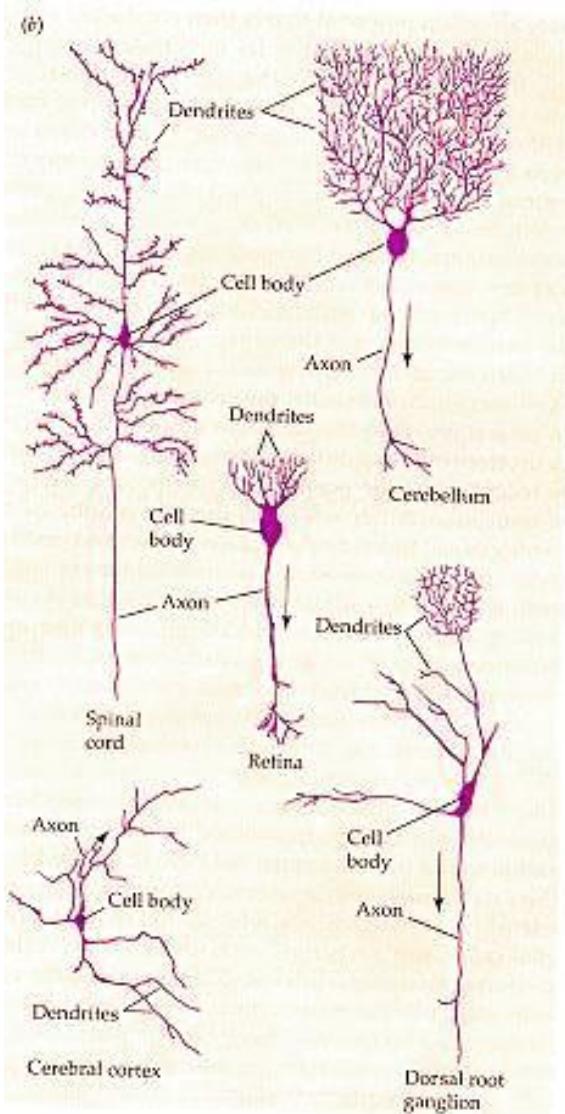
(Tsoukalas & Uhrig, 1997).

# *Inspiration from Neurobiology*

Human Biological Neuron



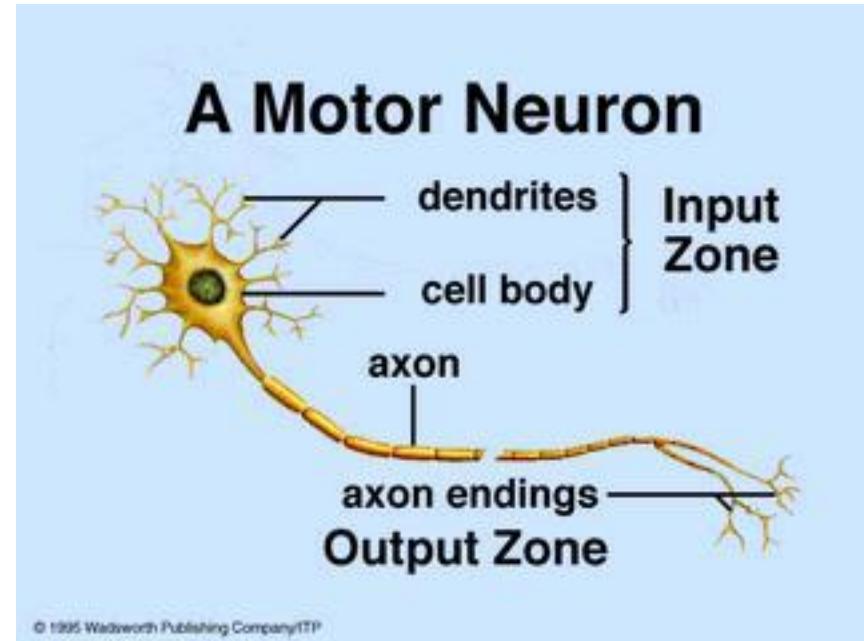
# *Biological Neural Networks*



**Biological neuron**

# *Biological Neural Networks*

- A biological neuron has three types of main components; dendrites, soma (or cell body) and axon.
- Dendrites receives signals from other neurons.
- The soma, sums the incoming signals. When sufficient input is received, the cell fires; that is it transmit a signal over its axon to other cells.



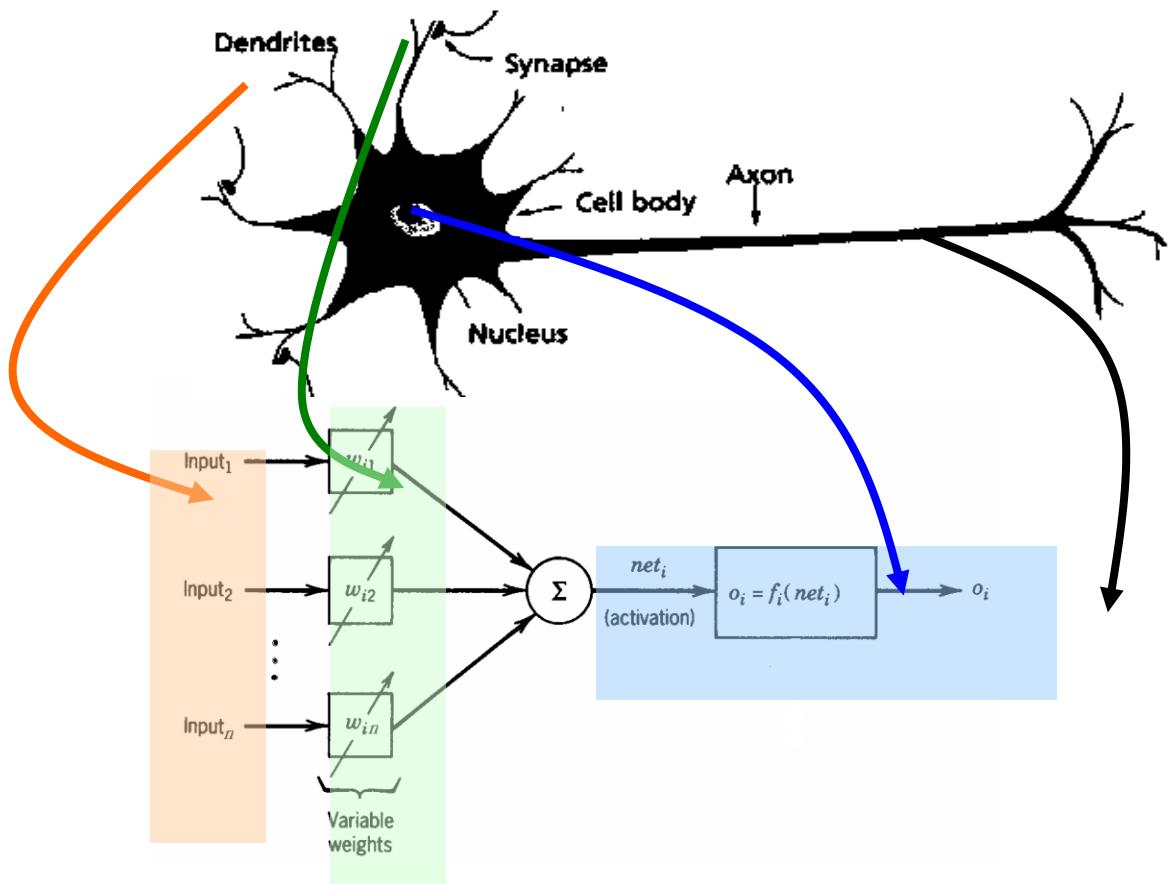
# *Artificial Neurons*

- ANN is an information processing system that has certain performance characteristics in common with biological nets.
- Several key features of the processing elements of ANN are suggested by the properties of biological neurons:
  1. The processing element receives many signals.
  2. Signals may be modified by a weight at the receiving synapse.
  3. The processing element sums the weighted inputs.
  4. Under appropriate circumstances (sufficient input), the neuron transmits a single output.
  5. The output from a particular neuron may go to many other neurons.

# Artificial Neurons

- From experience:  
examples / training  
data
- Strength of connection  
between the neurons  
is stored as a weight-  
value for the specific  
connection.
- Learning the solution  
to a problem =  
changing the  
connection weights

A physical neuron

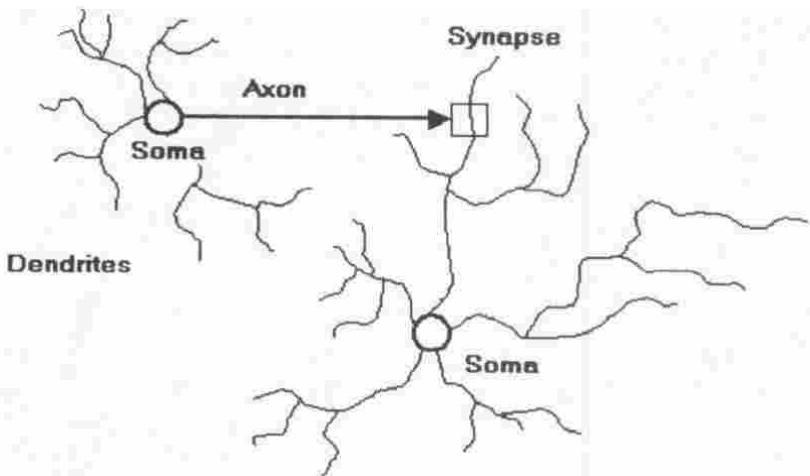


An artificial neuron

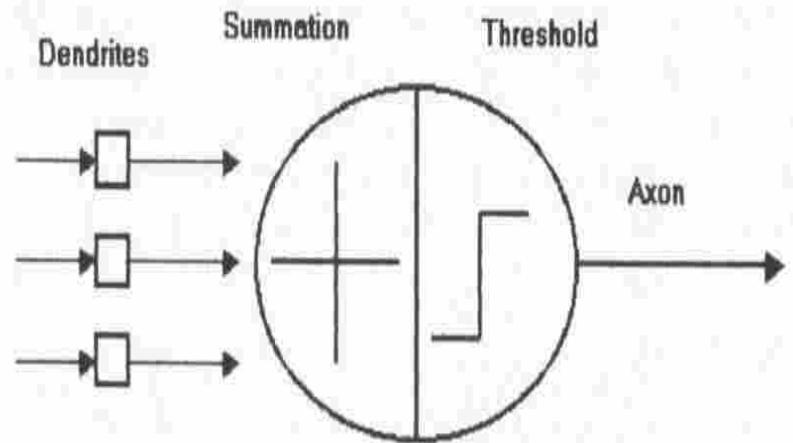
# *Artificial Neurons*

- ANNs have been developed as generalizations of mathematical models of neural biology, based on the assumptions that:
  1. Information processing occurs at many simple elements called neurons.
  2. Signals are passed between neurons over connection links.
  3. Each connection link has an associated weight, which, in typical neural net, multiplies the signal transmitted.
  4. Each neuron applies an activation function to its net input to determine its output signal.

## Artificial Neuron

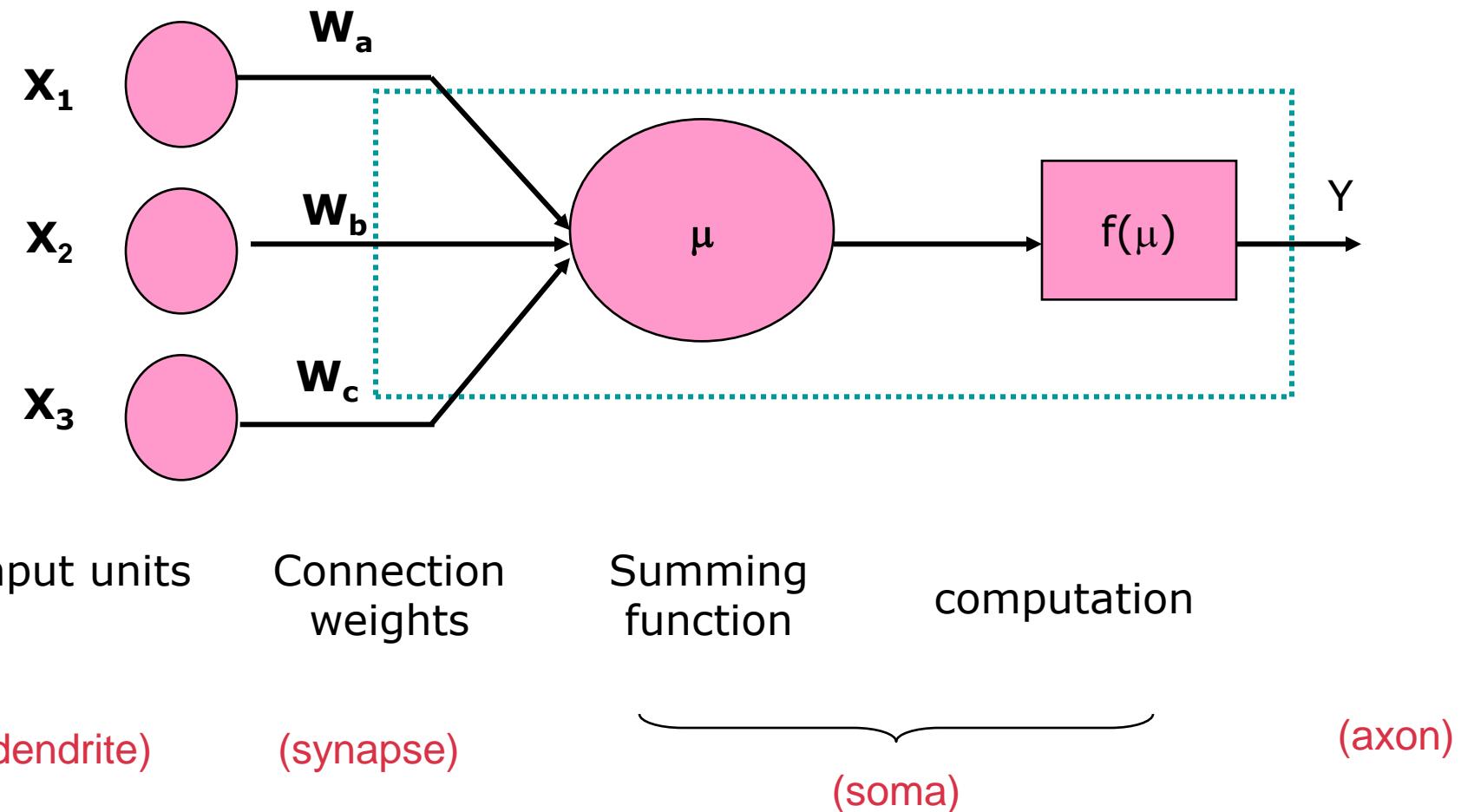


Four basic components of a human biological neuron



The components of a basic artificial neuron

# *Model Of A Neuron*

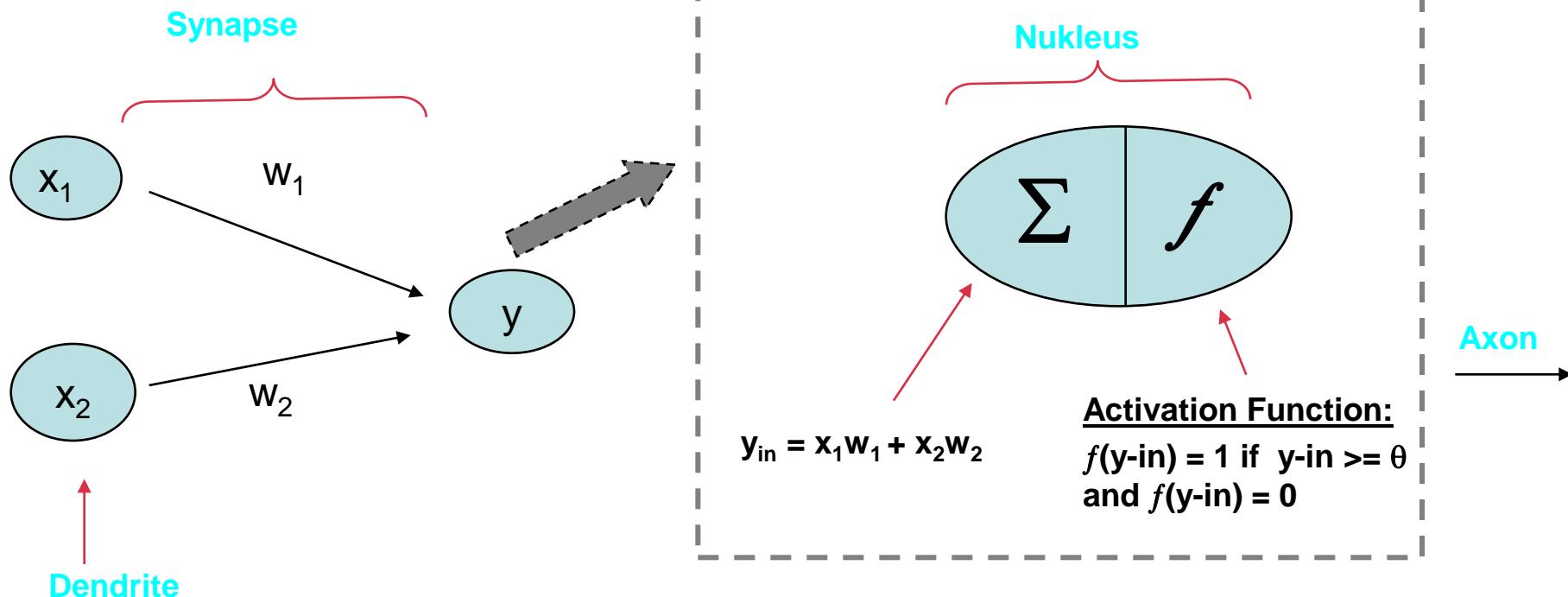


- A neural net consists of a large number of simple processing elements called neurons, units, cells or nodes.
- Each neuron is connected to other neurons by means of directed communication links, each with associated weight.
- The weight represent information being used by the net to solve a problem.

- Each neuron has an internal state, called its activation or activity level, which is a function of the inputs it has received. Typically, a neuron sends its activation as a signal to several other neurons.
- It is important to note that a neuron can send only one signal at a time, although that signal is broadcast to several other neurons.

- Neural networks are configured for a specific application, such as pattern recognition or data classification, through a **learning process**
  - In a biological system, learning involves adjustments to the synaptic connections between neurons
- same for artificial neural networks (ANNs)

# Artificial Neural Network



-A neuron receives input, determines the strength or the weight of the input, calculates the total weighted input, and compares the total weighted with a value (threshold)

-The value is in the range of 0 and 1

- If the total weighted input greater than or equal the threshold value, the neuron will produce the output, and if the total weighted input less than the threshold value, no output will be produced

# *History*

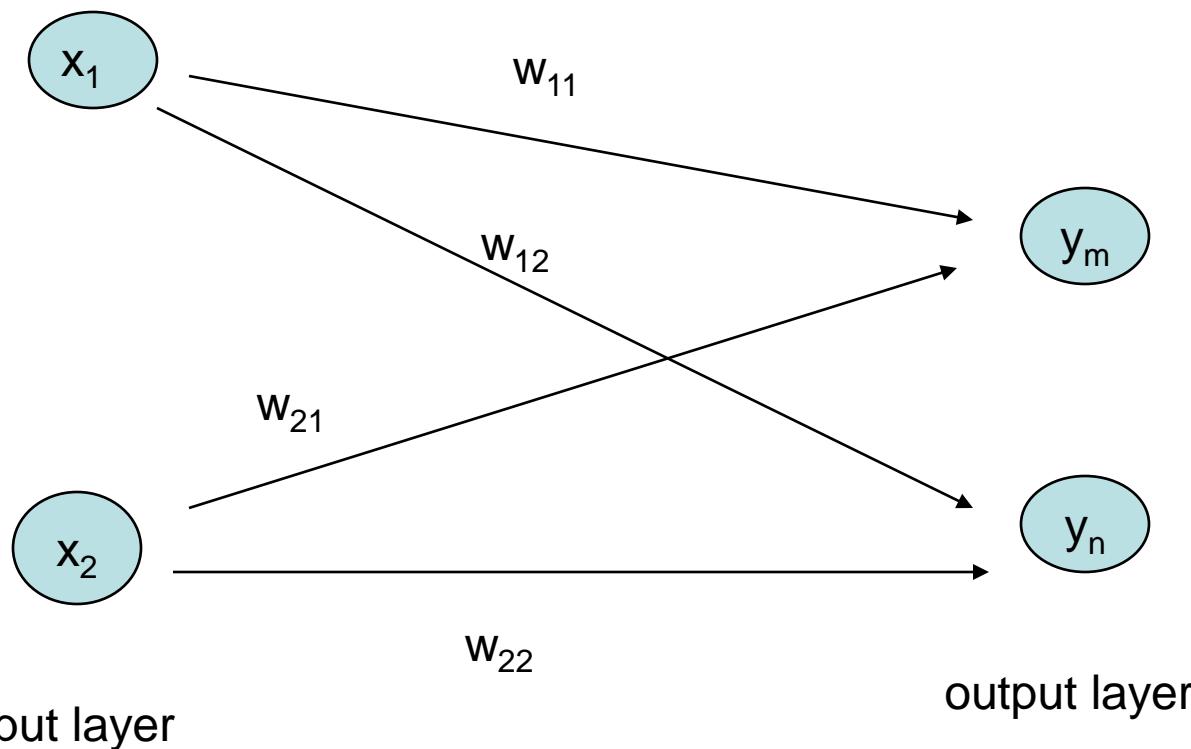
- 1943 McCulloch-Pitts neurons
- 1949 Hebb's law
- 1958 Perceptron (Rosenblatt)
- 1960 Adaline, better learning rule (Widrow, Huff)
- 1969 Limitations (Minsky, Papert)
- 1972 Kohonen nets, associative memory

- 1977 Brain State in a Box (Anderson)
- 1982 Hopfield net, constraint satisfaction
- 1985 ART (Carpenter, Grossfield)
- 1986 Backpropagation (Rumelhart, Hinton, McClelland)
- 1988 Neocognitron, character recognition (Fukushima)

# *Characterization*

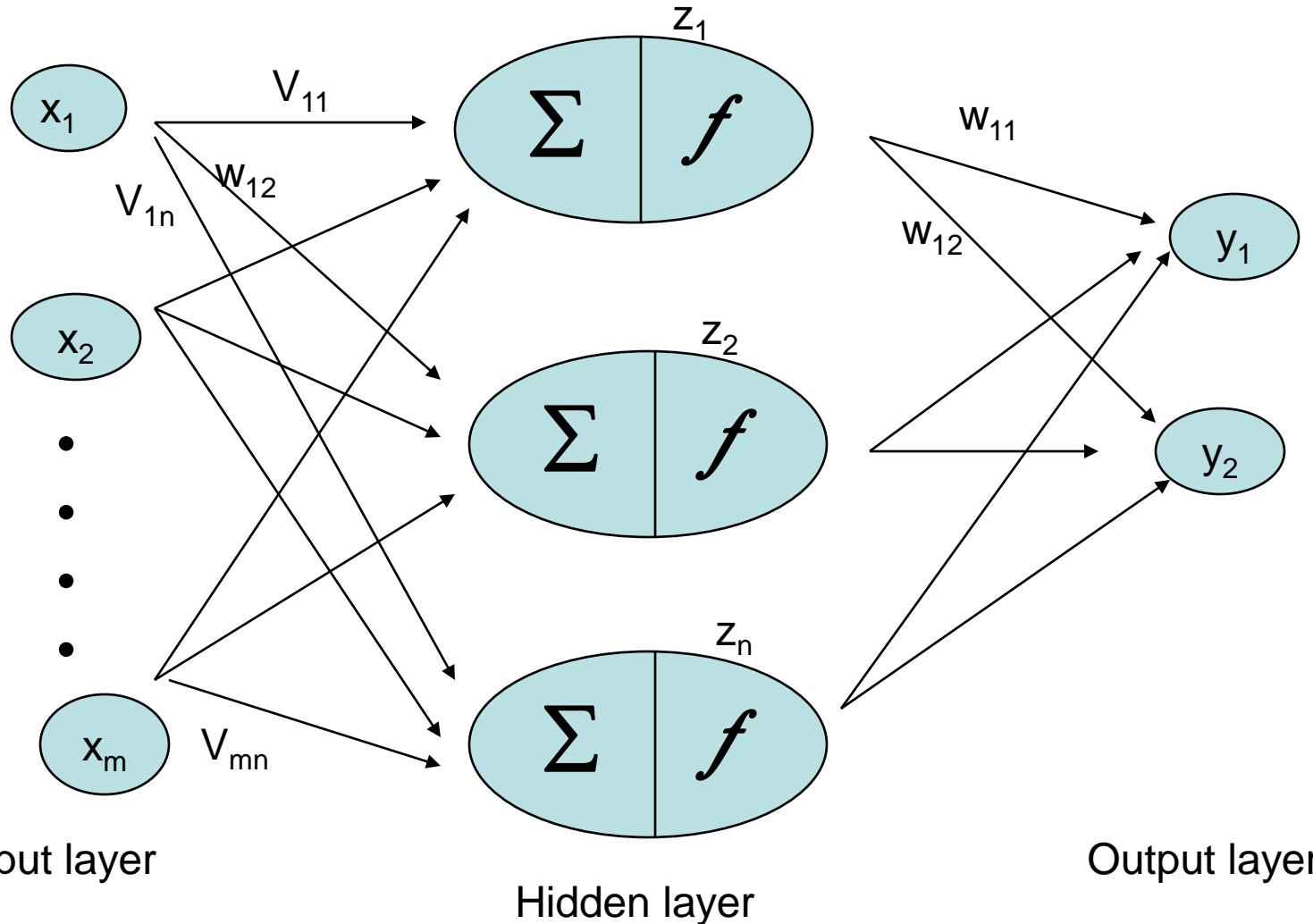
- **Architecture**
  - a pattern of connections between neurons
    - Single Layer Feedforward
    - Multilayer Feedforward
    - Recurrent
- **Strategy / Learning Algorithm**
  - a method of determining the connection weights
    - Supervised
    - Unsupervised
    - Reinforcement
- **Activation Function**
  - Function to compute output signal from input signal

# *Single Layer Feedforward NN*



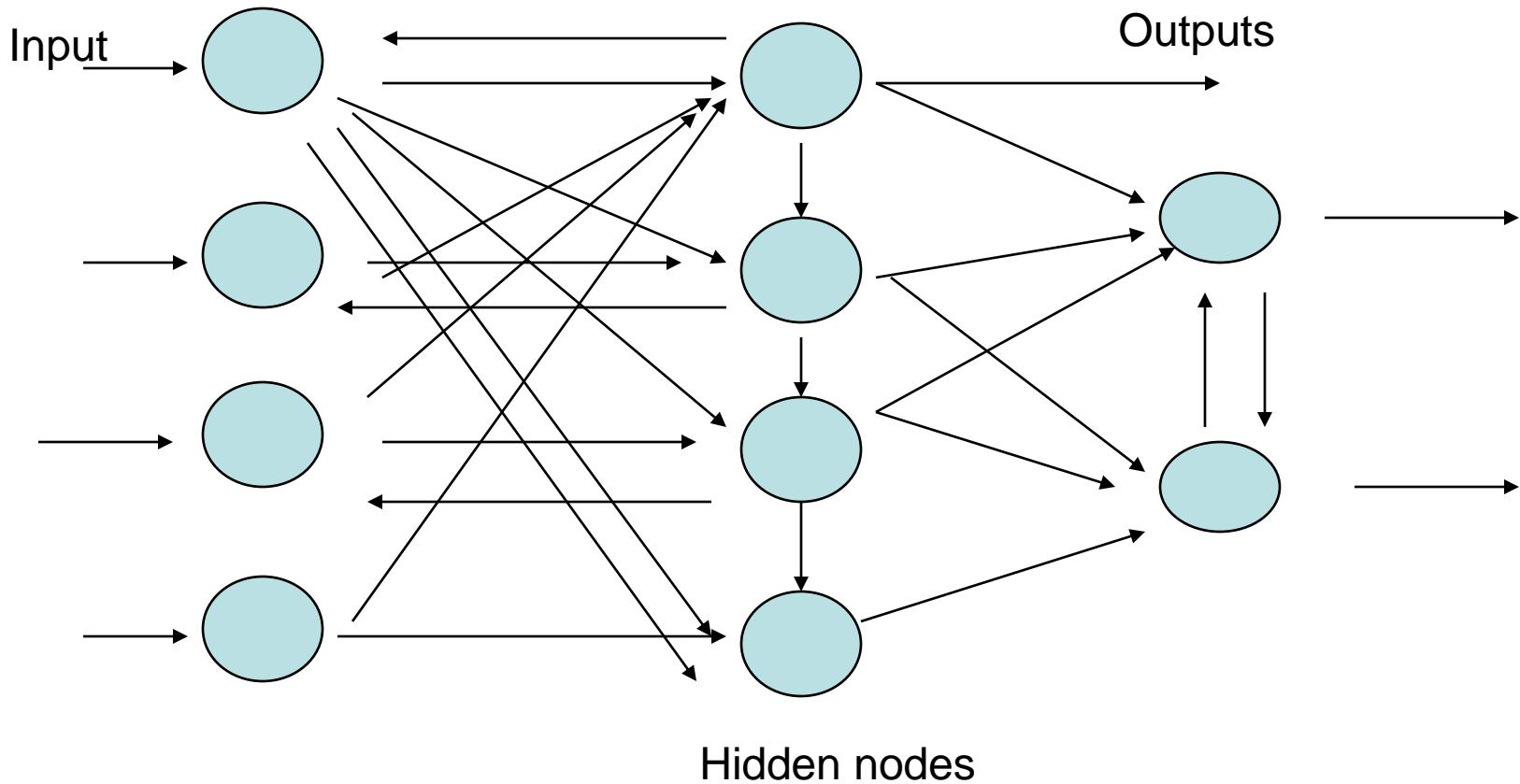
Contoh: **ADALINE, AM, Hopfield, LVQ, Perceptron, SOFM**

# Multilayer Neural Network



Contoh: **CCN, GRNN, MADALINE, MLFF with BP, Neocognitron, RBF, RCE**

# *Recurrent NN*



Contoh: **ART, BAM, BSB, Boltzman Machine, Cauchy Machine, Hopfield, RNN**

# *Strategy / Learning Algorithm*

## Supervised Learning

- Learning is performed by presenting pattern with target
- During learning, produced output is compared with the desired output
  - The difference between both output is used to modify learning weights according to the learning algorithm
- Recognizing hand-written digits, pattern recognition and etc.
- Neural Network models: **perceptron, feed-forward, radial basis function, support vector machine.**

## Unsupervised Learning

- Targets are not provided
- Appropriate for clustering task
  - Find similar groups of documents in the web, content addressable memory, clustering.
- Neural Network models: **Kohonen, self organizing maps, Hopfield networks.**

## Reinforcement Learning

- Target is provided, but the desired output is absent.
- The net is only provided with guidance to determine the produced output is correct or vice versa.
- Weights are modified in the units that have errors

# *Activation Functions*

- Identity

$$f(x) = x$$

- Binary step

$$f(x) = 1 \text{ if } x \geq \theta$$

$$f(x) = 0 \text{ otherwise}$$

- Binary sigmoid

$$f(x) = 1 / (1 + e^{-\sigma x})$$

- Bipolar sigmoid

$$f(x) = -1 + 2 / (1 + e^{-\sigma x})$$

- Hyperbolic tangent

$$f(x) = (e^x - e^{-x}) / (e^x + e^{-x})$$

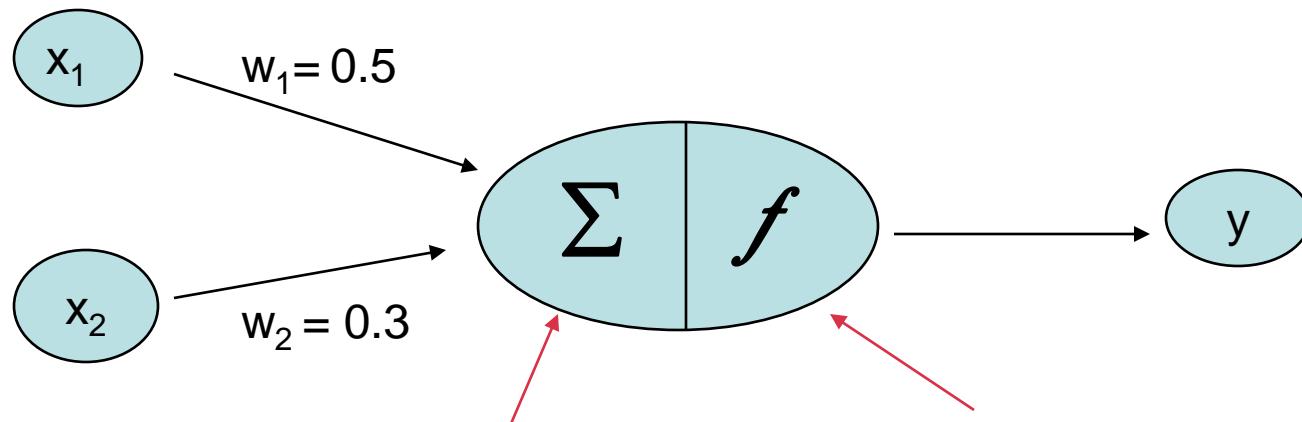
# *Exercise*

- 2 input AND

1	1	1
1	0	0
0	1	0
0	0	0

- 2 input OR

1	1	1
1	0	1
0	1	1
0	0	0



$$y_{in} = x_1w_1 + x_2w_2$$

**Activation Function:**  
**Binary Step Function**  
 $\theta = 0.5,$

$f(y-in) = 1 \text{ if } y-in \geq \theta$   
 dan  $f(y-in) = 0$

## *Where can neural network systems help...*

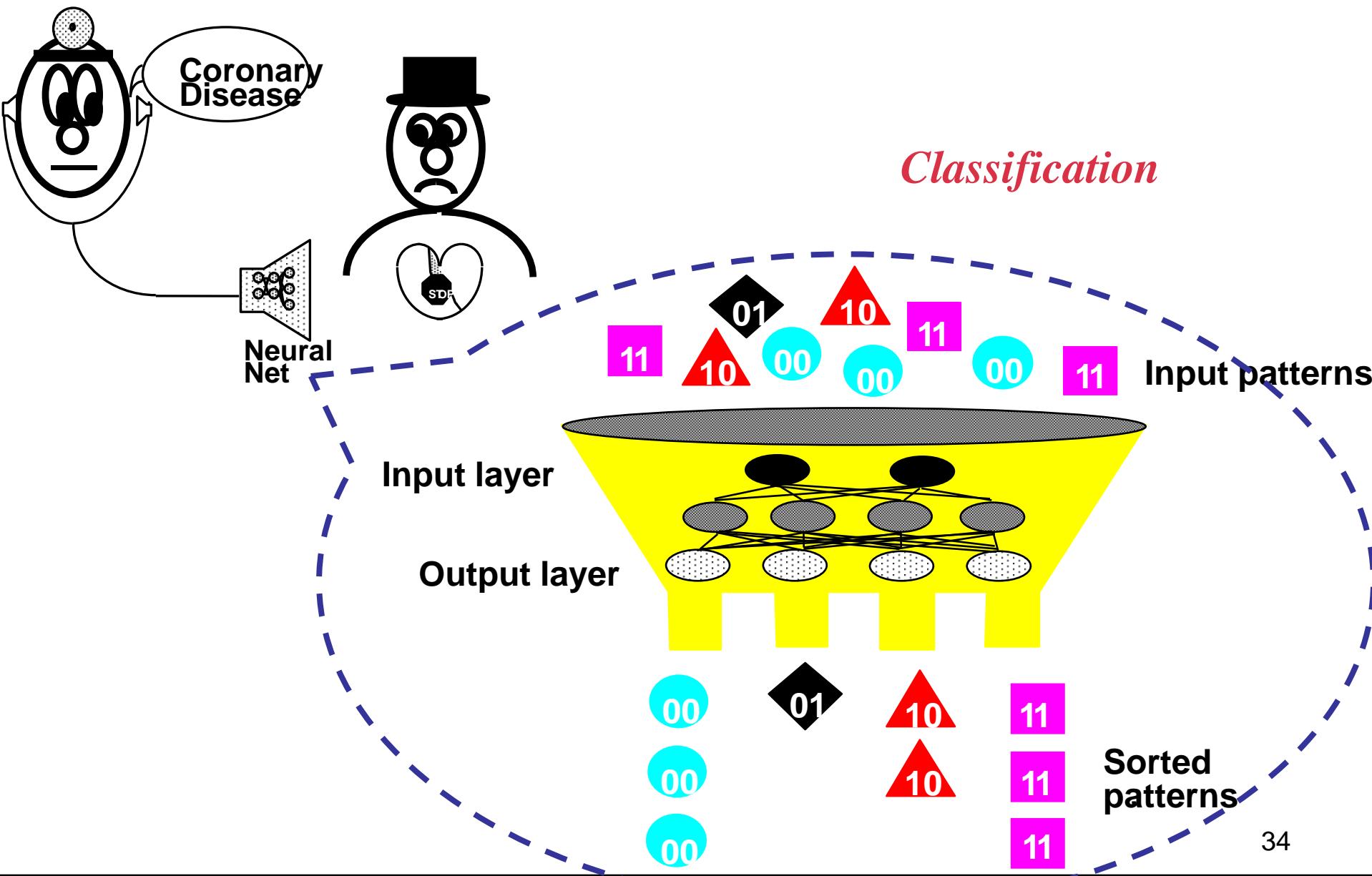
- when we can't formulate an algorithmic solution.
- when we **can** get lots of examples of the behavior we require.  
  
**'learning from experience'**
- when we need to pick out the structure from existing data.

# *Who is interested?...*

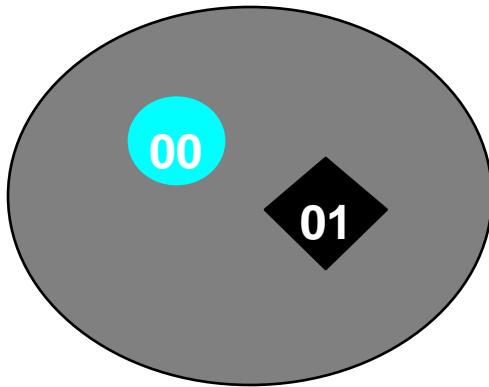
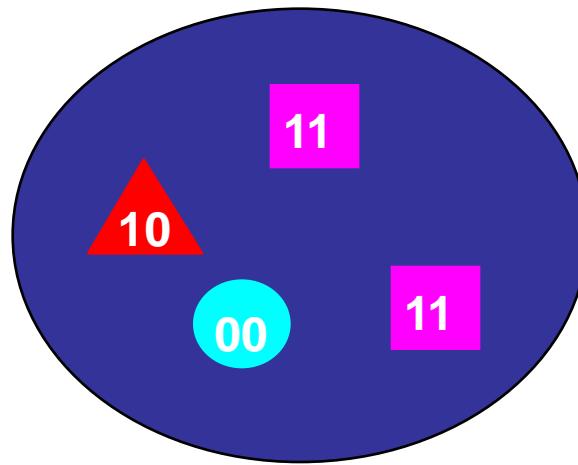
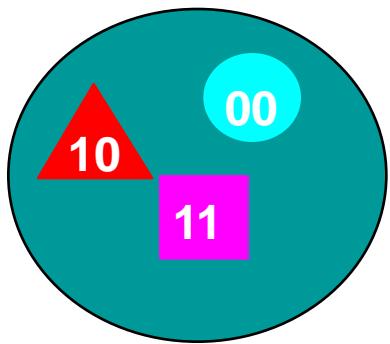
- Electrical Engineers – signal processing, control theory
- Computer Engineers – robotics
- Computer Scientists – artificial intelligence, pattern recognition
- Mathematicians – modelling tool when explicit relationships are unknown

# *Problem Domains*

- Storing and recalling patterns
- Classifying patterns
- Mapping inputs onto outputs
- Grouping similar patterns
- Finding solutions to constrained optimization problems



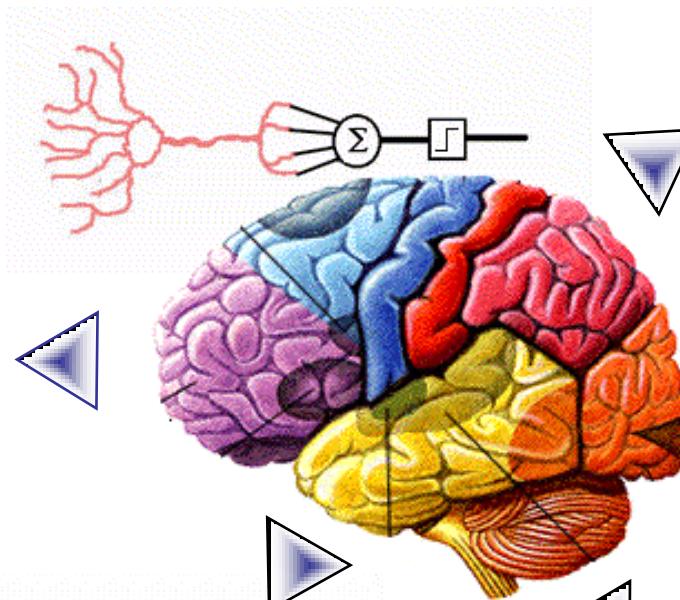
## *Clustering*



# *ANN Applications*



Chemistry



Education



Medical Applications



Information  
Searching & retrieval

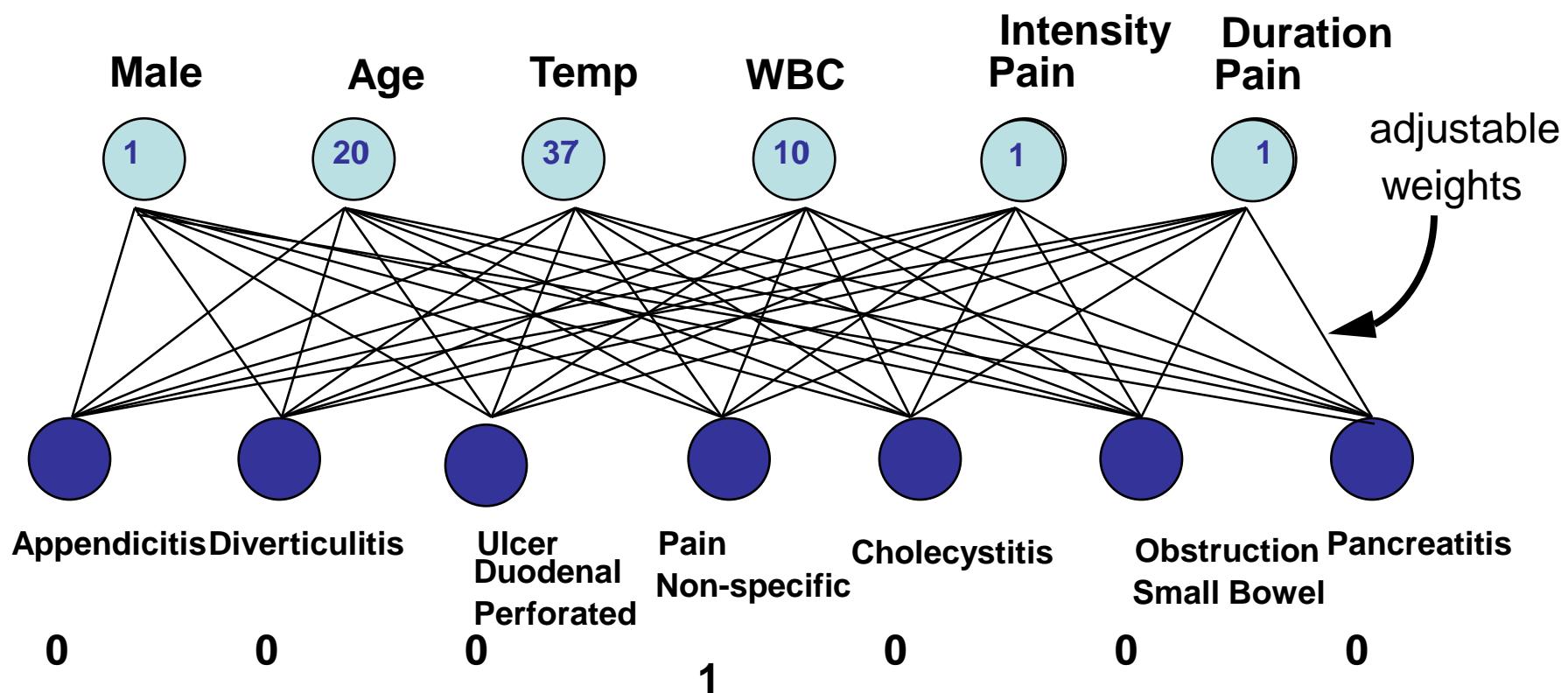


Business & Management

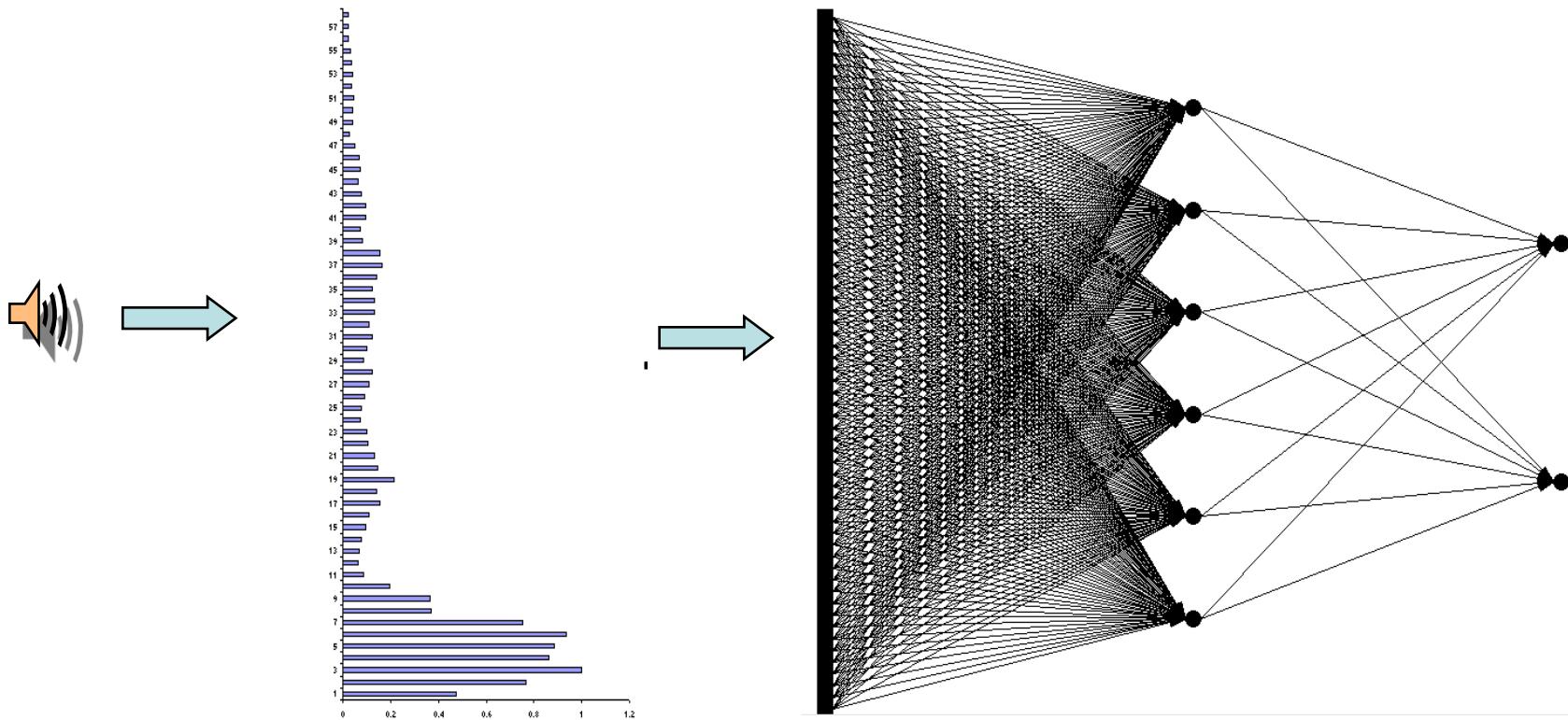
## *Applications of ANNs*

- Signal processing
- Pattern recognition, e.g. handwritten characters or face identification.
- Diagnosis or mapping symptoms to a medical case.
- Speech recognition
- Human Emotion Detection
- Educational Loan Forecasting

## Abdominal Pain Prediction



## Voice Recognition



# Educational Loan Forecasting System



# *Advantages Of NN*

## NON-LINEARITY

It can model non-linear systems

## INPUT-OUTPUT MAPPING

It can derive a relationship between a set of input & output responses

## ADAPTIVITY

The ability to learn allows the network to adapt to changes in the surrounding environment

## EVIDENTIAL RESPONSE

It can provide a confidence level to a given solution

# *Advantages Of NN*

## CONTEXTUAL INFORMATION

Knowledge is presented by the structure of the network. Every neuron in the network is potentially affected by the global activity of all other neurons in the network. Consequently, contextual information is dealt with naturally in the network.

## FAULT TOLERANCE

Distributed nature of the NN gives it fault tolerant capabilities

## NEUROBIOLOGY ANALOGY

Models the architecture of the brain

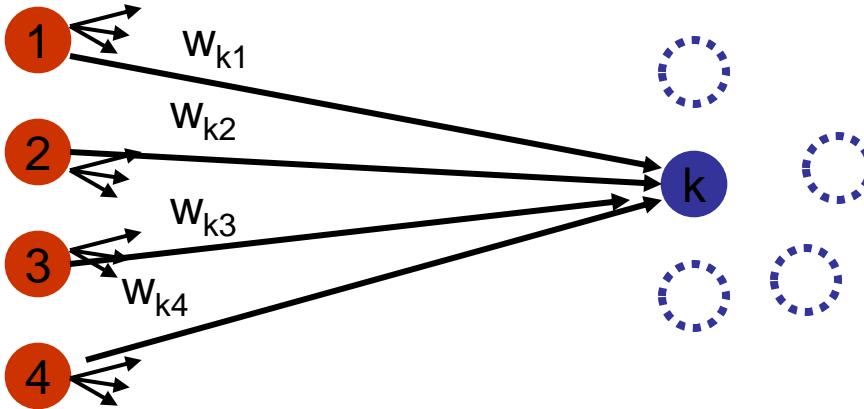
# *Comparison of ANN with conventional AI methods*

CHARACTERISTICS	TRADITIONAL COMPUTING (including Expert Systems)	ARTIFICIAL NEURAL NETWORKS
Processing style	Sequential	Parallel
Functions	Logically (left brained) via Rules Concepts Calculations	Gestault (right brained) via Images Pictures Controls
Learning Method	by rules (didactically)	by example (Socratically)
Applications	Accounting, word processing, math, inventory, digital communications	Sensor processing, speech recognition, pattern recognition, text recognition

# **UNIT-2**

**ASSOCIATIVE MEMORY AND  
UNSUPERVISED LEARNING  
NETWORKS**

# Weight Vectors in Clustering Networks



- Node  $k$  represents a particular class of input vectors, and the weights into  $k$  encode a prototype/centroid of that class.
- So if  $\text{prototype}(\text{class}(k)) = [i_{k1}, i_{k2}, i_{k3}, i_{k4}]$ , then:  
 $w_{km} = f_e(i_{km})$  for  $m = 1..4$ , where  $f_e$  is the encoding function.
- In some cases, the encoding function involves normalization.  
Hence  $w_{km} = f_e(i_{k1} \dots i_{k4})$ .
- The weight vectors are learned during the unsupervised training phase.

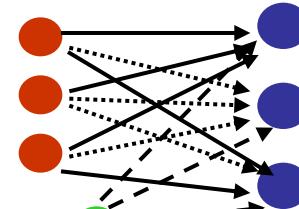
# Unsupervised Learning with Artificial Neural Networks

- The ANN is given a set of patterns,  $P$ , from space,  $S$ , but little/no information about their classification, evaluation, interesting features, etc. It must learn these by itself!
- Tasks
  - Clustering - Group patterns based on similarity (Focus of this lecture)
  - Vector Quantization - Fully divide up  $S$  into a small set of regions (defined by codebook vectors) that also helps cluster  $P$ .
  - Probability Density Approximation - Find small set of points whose distribution matches that of  $P$ .

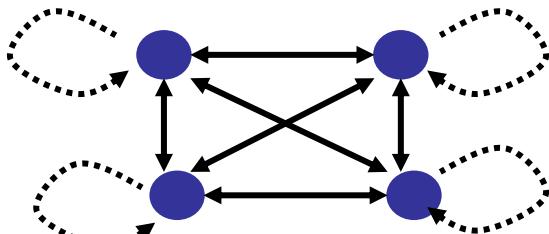
# Network Types for Clustering

- Winner-Take-All Networks

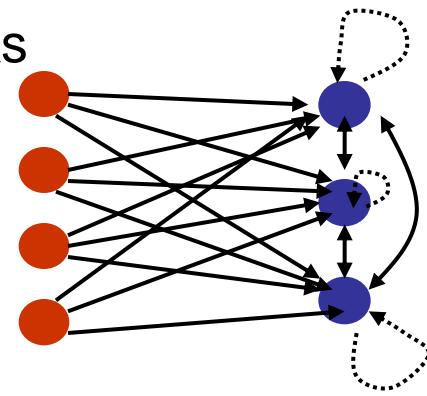
- Hamming Networks



- Maxnet

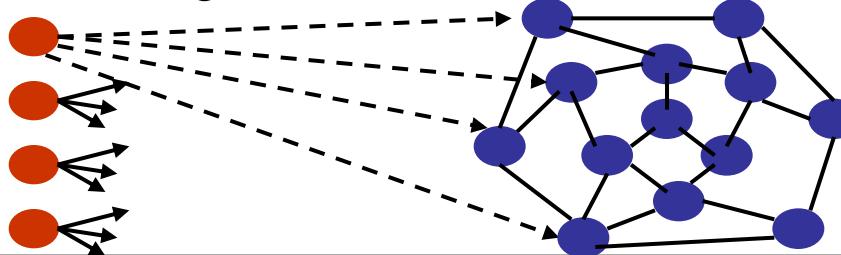


- Simple Competitive Learning Networks



- Topologically Organized Networks

- Winner & its neighbors “take some”



# Hamming Networks

Given: a set of patterns  $m$  patterns,  $P$ , from an  $n$ -dim input space,  $S$ .

Create: a network with  $n$  input nodes and  $m$  simple linear output nodes (one per pattern), where the incoming weights to the output node for pattern  $p$  is based on the  $n$  features of  $p$ .

$i_{pj}$  = jth input bit of the pth pattern.  $i_{pj} = 1$  or  $-1$

Set  $w_{pj} = i_{pj}/2$ .

Also include a threshold input of  $-n/2$  at each output node.

Testing: enter an input pattern,  $I$ , and use the network to determine which member of  $P$  that is closest to  $I$ . Closeness is based on the Hamming distance (# non-matching bits in the two patterns).

Given input  $I$ , the output value of the output node for pattern  $p$  =

$$\sum_{k=1}^n w_{pk} I_k - \frac{n}{2} \equiv \sum_{k=1}^n \frac{i_{pk}}{2} I_k - \frac{n}{2} \equiv \frac{1}{2} \left( \sum_{k=1}^n i_{pk} I_k - n \right)$$

# Hamming Networks (2)

Proof: (that output of output node  $p$  is the negative of the Hamming distance between  $p$  and input vector  $I$ ).

Assume:  $k$  bits match.

Then:  $n - k$  bits do not match, and  $n - k$  is the Hamming distance.

And: the output value of  $p$ 's output node is:

$$\frac{1}{2} \left( \sum_{k=1}^n i_{pk} I_k - n \right) \equiv \frac{1}{2} (k - (n - k) - n) \equiv k - n \equiv -(n - k)$$

$k$  matches, where each match gives  $(1)(1)$  or  $(-1)(-1) = 1$

Neg. Hamming distance

$n - k$  mismatches, where each gives  $(-1)(1)$  or  $(1)(-1) = -1$

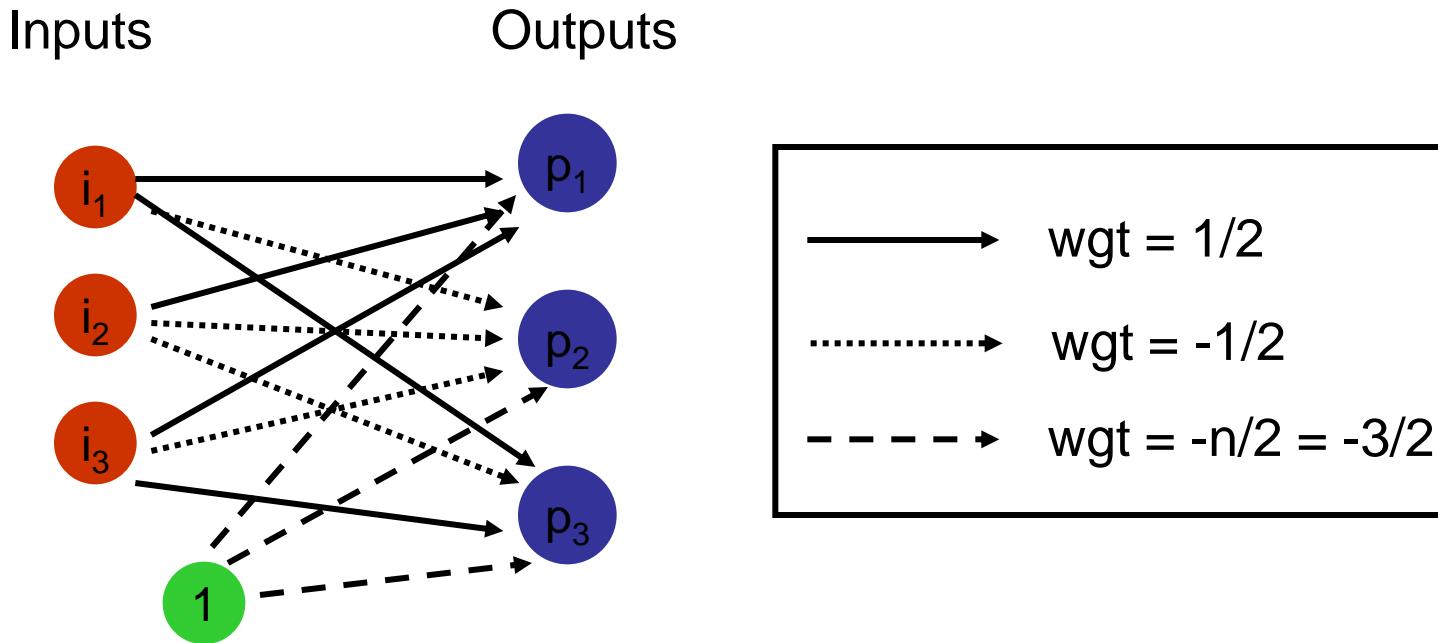
The pattern  $p^*$  with the largest negative Hamming distance to  $I$  is thus the pattern with the smallest Hamming distance to  $I$  (i.e. the nearest to  $I$ ).

Hence,

the output node that represents  $p^*$  will have the highest output value of all

# Hamming Network Example

$P = \{(1 1 1), (-1 -1 -1), (1 -1 1)\} = 3$  patterns of length 3



Given: input pattern  $I = (-1 1 1)$

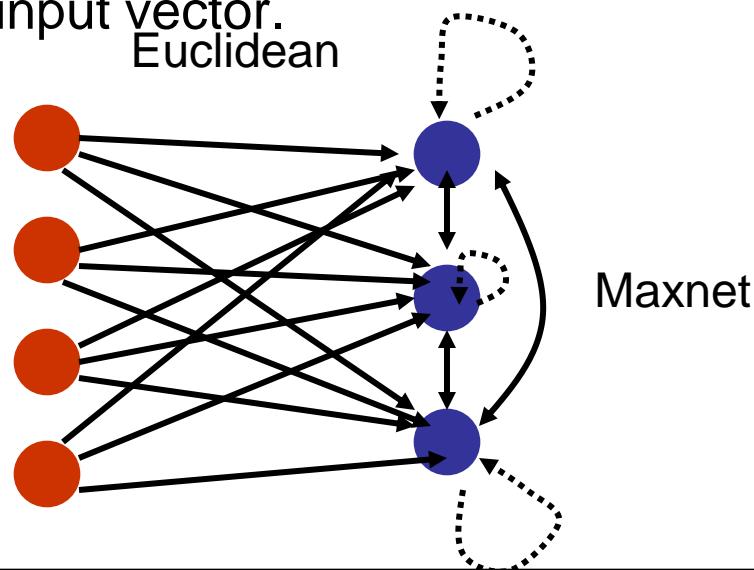
$$\text{Output } (p_1) = -1/2 + 1/2 + 1/2 - 3/2 = -1 \text{ (Winner)}$$

$$\text{Output } (p_2) = 1/2 - 1/2 - 1/2 - 3/2 = -2$$

$$\text{Output } (p_3) = -1/2 - 1/2 + 1/2 - 3/2 = -2$$

# Simple Competitive Learning

- Combination Hamming-like Net + Maxnet with learning of the input-to-output weights.
  - Inputs can be real valued, not just 1, -1.
    - So distance metric is actually Euclidean or Manhattan, not Hamming.
- Each output node represents a centroid for input patterns it wins on.
- Learning: winner node's incoming weights are updated to move closer to the input vector.



# Winning & Learning

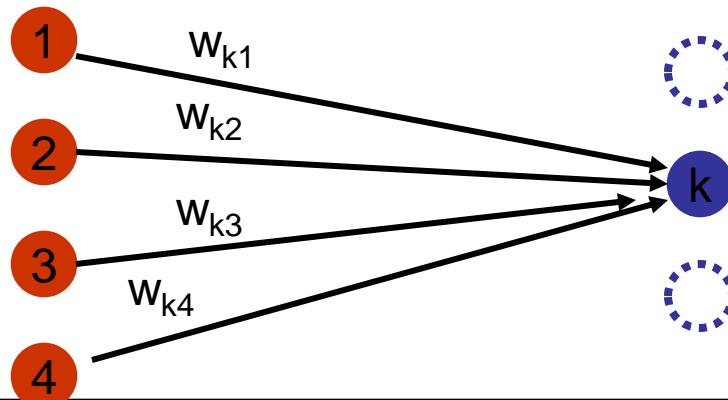
``Winning isn't everything...it's the ONLY thing'' - Vince Lombardi

- Only the incoming weights of the winner node are modified.
- Winner = output node whose incoming weights are the shortest Euclidean distance from the input vector.

$$\sqrt{\sum_{i=1}^n (I_i - w_{ki})^2} \quad = \text{Euclidean distance from input vector } I \text{ to the vector represented by output node } k \text{'s incoming weights}$$

- Update formula: If j is the winning output node:

$$\forall i : w_{ji}(\text{new}) = w_{ji}(\text{old}) + \eta(I_i - w_{ji}(\text{old}))$$



Note: The use of real-valued inputs & Euclidean distance means that the simple product of weights and inputs does not correlate with "closeness" as in binary networks using Hamming distance.

# SCL Examples (1)

6 Cases:

(0 1 1)	(1 1 0.5)
(0.2 0.2 0.2)	(0.5 0.5 0.5)
(0.4 0.6 0.5)	(0 0 0)

Learning Rate: 0.5

Initial Randomly-Generated Weight Vectors:

[ 0.14 0.75 0.71 ]	
[ 0.99 0.51 0.37 ]	Hence, there are 3 classes to be learned
[ 0.73 0.81 0.87 ]	

Training on Input Vectors

Input vector # 1: [ 0.00 1.00 1.00 ]  
Winning weight vector # 1: [ 0.14 0.75 0.71 ] Distance: 0.41  
Updated weight vector: [ 0.07 0.87 0.85 ]

Input vector # 2: [ 1.00 1.00 0.50 ]  
Winning weight vector # 3: [ 0.73 0.81 0.87 ] Distance: 0.50  
Updated weight vector: [ 0.87 0.90 0.69 ]

# SCL Examples (2)

Input vector # 3: [ 0.20 0.20 0.20 ]

Winning weight vector # 2: [ 0.99 0.51 0.37 ] Distance: 0.86

Updated weight vector: [ 0.59 0.36 0.29 ]

Input vector # 4: [ 0.50 0.50 0.50 ]

Winning weight vector # 2: [ 0.59 0.36 0.29 ] Distance: 0.27

Updated weight vector: [ 0.55 0.43 0.39 ]

Input vector # 5: [ 0.40 0.60 0.50 ]

Winning weight vector # 2: [ 0.55 0.43 0.39 ] Distance: 0.25

Updated weight vector: [ 0.47 0.51 0.45 ]

Input vector # 6: [ 0.00 0.00 0.00 ]

Winning weight vector # 2: [ 0.47 0.51 0.45 ] Distance: 0.83

Updated weight vector: [ 0.24 0.26 0.22 ]

Weight Vectors after epoch 1:

[ 0.07 0.87 0.85 ]

[ 0.24 0.26 0.22 ]

[ 0.87 0.90 0.69 ]

# SCL Examples (3)

Clusters after epoch 1:

Weight vector # 1: [ 0.07 0.87 0.85 ]  
Input vector # 1: [ 0.00 1.00 1.00 ]  
Weight vector # 2: [ 0.24 0.26 0.22 ]  
Input vector # 3: [ 0.20 0.20 0.20 ]  
Input vector # 4: [ 0.50 0.50 0.50 ]  
Input vector # 5: [ 0.40 0.60 0.50 ]  
Input vector # 6: [ 0.00 0.00 0.00 ]  
Weight vector # 3: [ 0.87 0.90 0.69 ]  
Input vector # 2: [ 1.00 1.00 0.50 ]

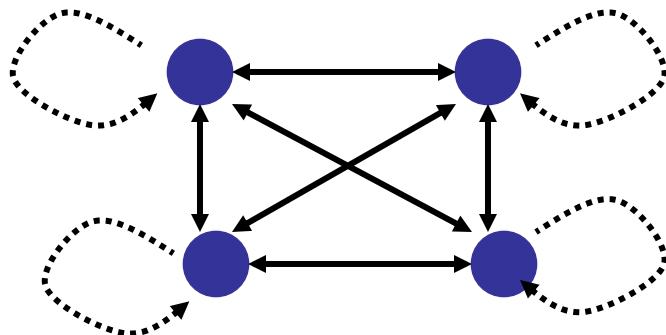
Weight Vectors after epoch 2:

[ 0.03 0.94 0.93 ]  
[ 0.19 0.24 0.21 ]  
[ 0.93 0.95 0.59 ]

Clusters after epoch 2:

unchanged.

# Maxnet



$$\begin{array}{l} \longleftrightarrow \quad wgt = -\varepsilon \\ \cdots \rightarrow \quad wgt = \theta \\ \text{e.g. } \theta = 1 \wedge \varepsilon \leq \frac{1}{n} \end{array}$$

Simple network to find node with largest initial input value.

Topology: clique with self-arcs, where all self-arcs have a small positive (excitatory) weight, and all other arcs have a small negative (inhibitory) weight.

Nodes: have transfer function  $f_T = \max(\sum, 0)$

Algorithm:

Load initial values into the clique

Repeat:

Synchronously update all node values via  $f_T$

Until: all but one node has a value of 0

Winner = the non-zero node

# Maxnet Examples

- Input values: (1, 2, 5, 4, 3) with epsilon = 1/5 and theta = 1

0.000	0.000	3.000	1.800	0.600
0.000	0.000	2.520	1.080	0.000
0.000	0.000	2.304	0.576	0.000
0.000	0.000	2.189	0.115	0.000
0.000	0.000	2.166	0.000	0.000
0.000	0.000	2.166	0.000	0.000

$$= (1)5 - (0.2)(1+2+4+3)$$

Stable attractor

- Input values: (1, 2, 5, 4.5, 4.7) with epsilon = 1/5 and theta = 1

0.000	0.000	2.560	1.960	2.200
0.000	0.000	1.728	1.008	1.296
0.000	0.000	1.267	0.403	0.749
0.000	0.000	1.037	0.000	0.415
0.000	0.000	0.954	0.000	0.207
0.000	0.000	0.912	0.000	0.017
0.000	0.000	0.909	0.000	0.000
0.000	0.000	0.909	0.000	0.000

Stable attractor

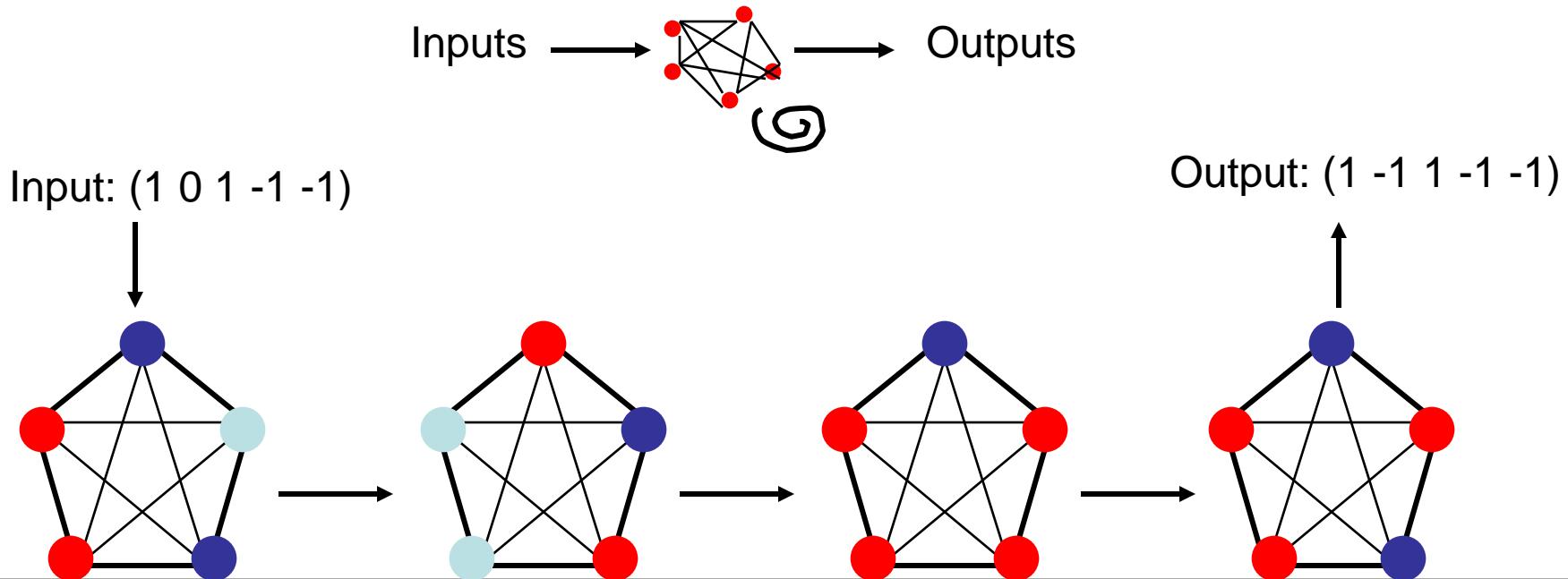
# Associative-Memory Networks

Input: Pattern (often noisy/corrupted)

Output: Corresponding pattern (complete / relatively noise-free)

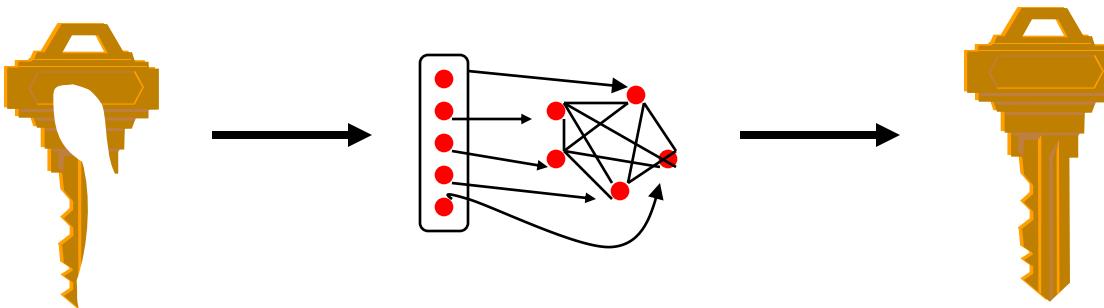
Process

1. Load input pattern onto core group of highly-interconnected neurons.
2. Run core neurons until they reach a steady state.
3. Read output off of the states of the core neurons.



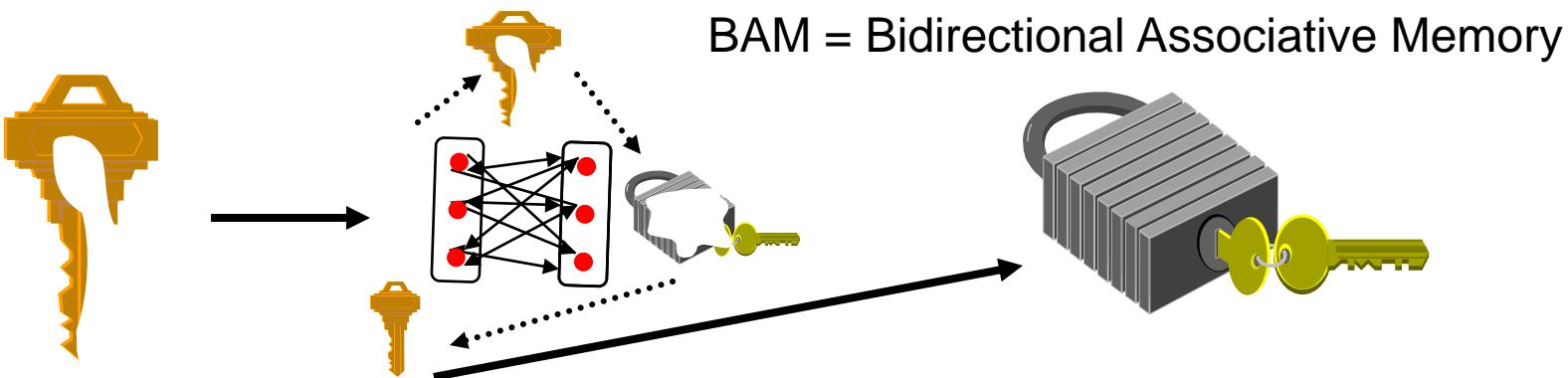
# Associative Network Types

1. Auto-associative:  $X = Y$



\*Recognize noisy versions of a pattern

2. Hetero-associative Bidirectional:  $X < > Y$



\*Iterative correction of input and output

# Hebb's Rule

Connection Weights ~ Correlations

``When one cell repeatedly assists in firing another, the axon of the first cell develops synaptic knobs (or enlarges them if they already exist) in contact with the soma of the second cell." (Hebb, 1949)



In an associative neural net, if we compare two pattern components (e.g. pixels) within many patterns and find that they are frequently in:

- a) the same state, then the arc weight between their NN nodes should be positive
- b) different states, then " " " negative

## Matrix Memory:

The weights must store the average correlations between all pattern components across all patterns. A net presented with a partial pattern can then use the correlations to recreate the entire pattern.

# Quantifying Hebb's Rule

Compare two nodes to calc a weight change that reflects the state correlation:

Auto-Association:  $\Delta w_{jk} \propto i_{pk} i_{pj}$  \* When the two components are the same (different), increase (decrease) the weight

Hetero-Association:  $\Delta w_{jk} \propto i_{pk} o_{pj}$  i = input component  
o = output component

Ideally, the weights will record the average correlations across all patterns:

$$\text{Auto: } w_{jk} \propto \sum_{p=1}^P i_{pk} i_{pj}$$

$$\text{Hetero: } w_{jk} \propto \sum_{p=1}^P i_{pk} o_{pj}$$

Hebbian Principle: If all the input patterns are known prior to retrieval time, then init weights as:

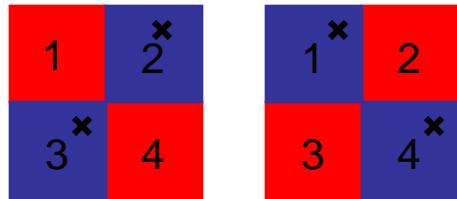
$$\text{Auto: } w_{jk} \equiv \frac{1}{P} \sum_{p=1}^P i_{pk} i_{pj}$$

$$\text{Hetero: } w_{jk} \equiv \frac{1}{P} \sum_{p=1}^P i_{pk} o_{pj}$$

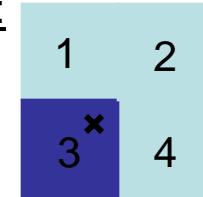
Weights = Average Correlations

# Auto-Associative Memory

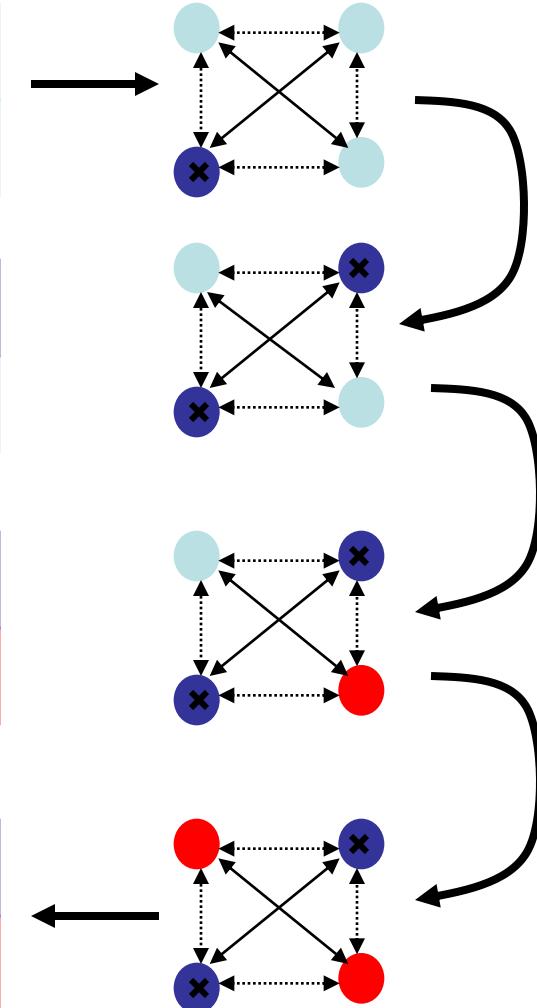
## 1. Auto-Associative Patterns to Remember



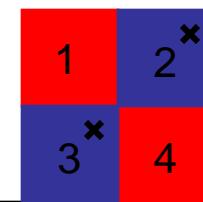
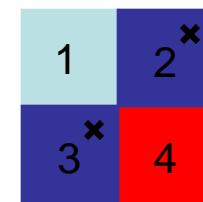
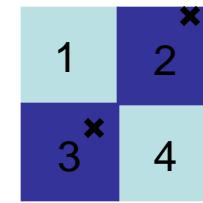
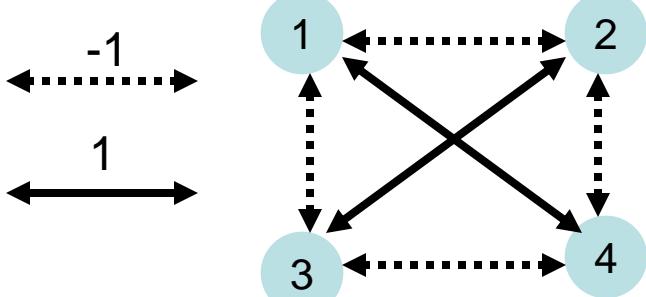
Comp/Node value legend:  
dark (blue) with  $x \Rightarrow +1$   
dark (red) w/o  $x \Rightarrow -1$   
light (green)  $\Rightarrow 0$



## 3. Retrieval



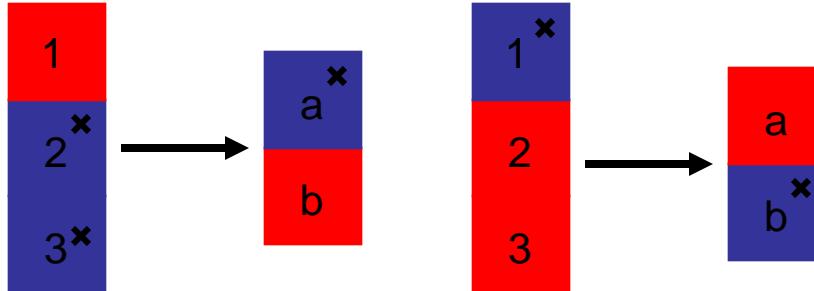
## 2. Distributed Storage of All Patterns:



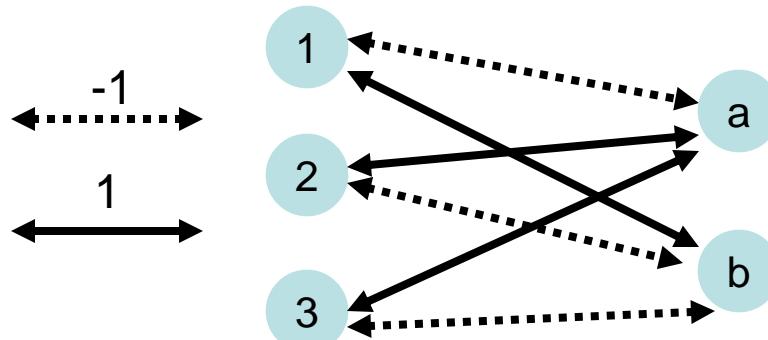
- 1 node per pattern unit
- Fully connected: clique
- Weights = avg correlations across all patterns of the corresponding units

# Hetero-Associative Memory

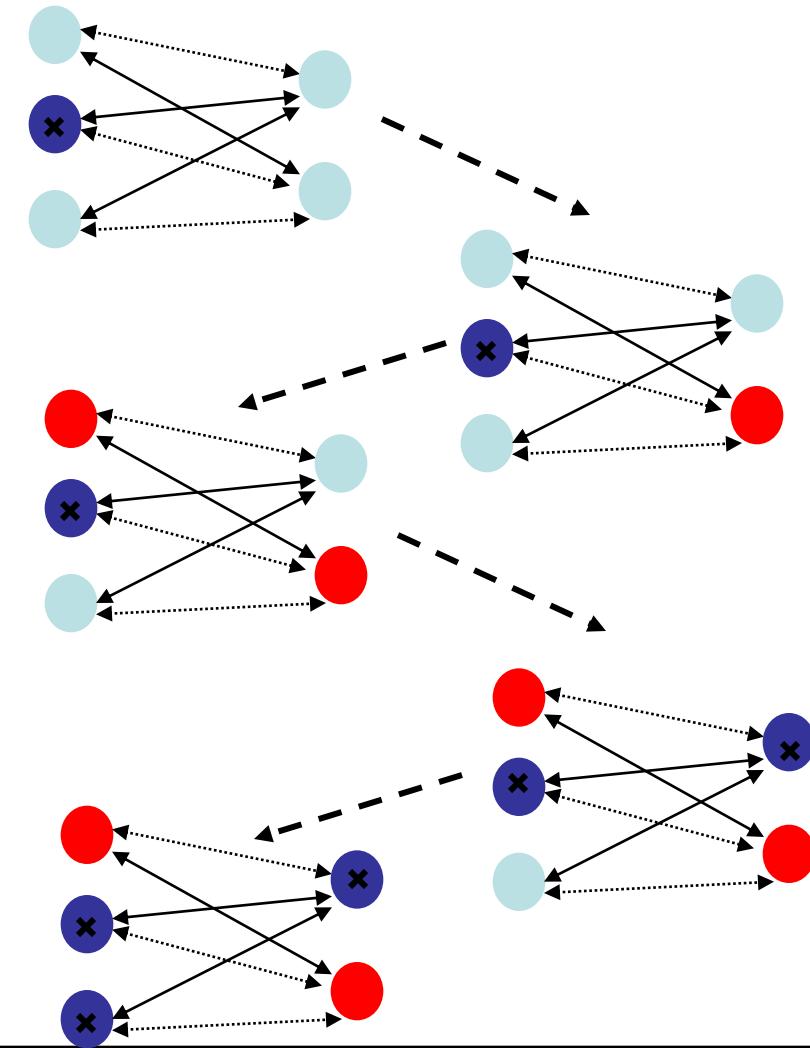
1. Hetero-Associative Patterns (Pairs) to Remember    3. Retrieval



2. Distributed Storage of All Patterns:

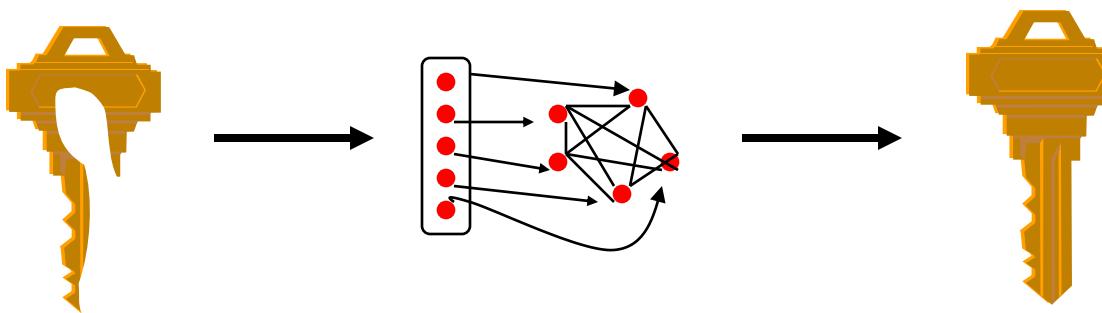


- 1 node per pattern unit for X & Y
- Full inter-layer connection
- Weights = avg correlations across all patterns of the corresponding units

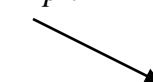


# Hopfield Networks

- Auto-Association Network
  - Fully-connected (clique) with symmetric weights
  - State of node =  $f(\text{inputs})$
  - Weight values based on Hebbian principle
  - Performance: Must iterate a bit to converge on a pattern, but generally much less computation than in back-propagation networks.
- Input      Output (after many iterations)



Discrete node update rule:  $x_{pk}(t+1) = \text{sgn}\left(\sum_{j=1}^n w_{kj}x_{pj}(t) + I_{pk}\right)$



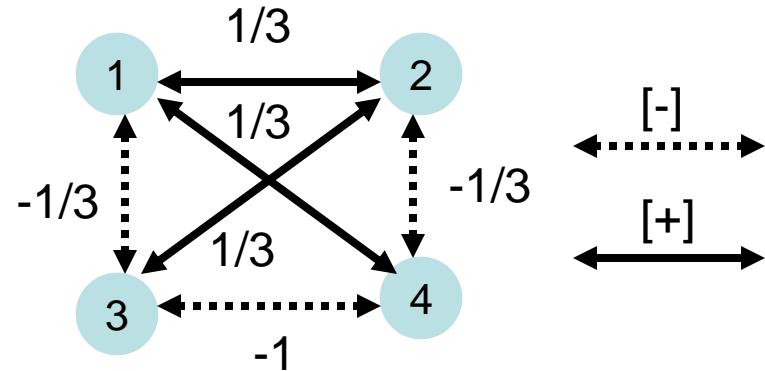
Input value

# Hopfield Network Example

## 1. Patterns to Remember

$p_1$	$p_2$	$p_3$
$\begin{matrix} 1^* & 2^* \\ 3^* & 4 \end{matrix}$	$\begin{matrix} 1^* & 2^* \\ 3 & 4^* \end{matrix}$	$\begin{matrix} 1 & 2 \\ 3^* & 4 \end{matrix}$

## 3. Build Network



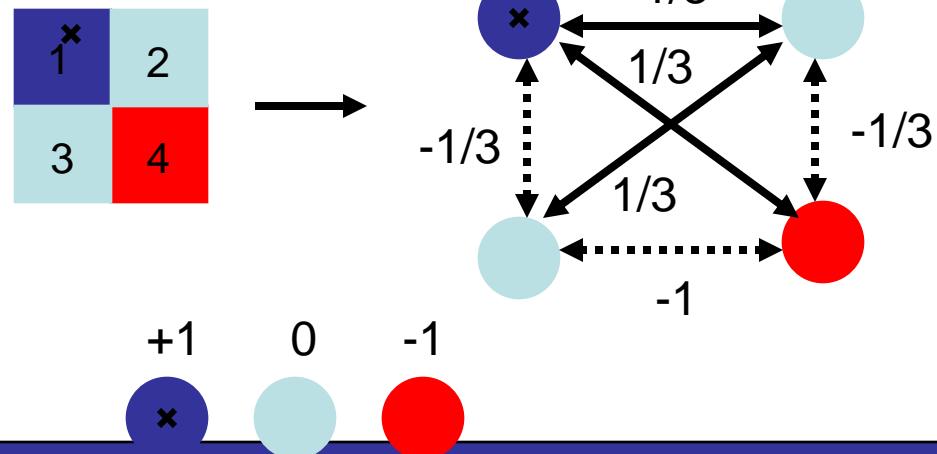
## 2. Hebbian Weight Init:

Avg Correlations across 3 patterns

$p_1$	$p_2$	$p_3$	Avg
-------	-------	-------	-----

$W_{12}$	1	1	-1	1/3
$W_{13}$	1	-1	-1	-1/3
$W_{14}$	-1	1	1	1/3
$W_{23}$	1	-1	1	1/3
$W_{24}$	-1	1	-1	-1/3
$W_{34}$	-1	-1	-1	-1

## 4. Enter Test Pattern



# Storage Capacity of Hopfield Networks

Capacity = Relationship between # patterns that can be stored & retrieved  
without error to the size of the network.

Capacity = # patterns / # nodes or # patterns / # weights

- If we use the following definition of 100% correct retrieval:  
When any of the stored patterns is entered completely (no noise), then that same pattern is returned by the network; i.e. The pattern is a stable attractor.
  - A detailed proof shows that a Hopfield network of  $N$  nodes can achieve 100% correct retrieval on  $P$  patterns if:  $P < N/(4*\ln(N))$
- In general, as more patterns are added to a network, the avg correlations will be less likely to match the correlations in any particular pattern. Hence, the likelihood of retrieval error will increase.  
=> The key to perfect recall is selective ignorance!!

<u>N</u>	<u>Max P</u>
10	1
100	5
1000	36
10000	271
$10^{11}$	$10^9$

# Stochastic Hopfield Networks

Node state is stochastically determined by sum of inputs:

Node fires with probability:

$$P \equiv \frac{1}{1 + e^{-2\beta \sum_k m_k}}$$

For these networks, effective retrieval is obtained when  $P < 0.138N$ , which is an improvement over standard Hopfield nets.

## Boltzmann Machines:

Similar to Hopfield nets but with hidden layers.

State changes occur either:

- Deterministically when  $\Delta E < 0$
- Stochastically with probability =  $\frac{1}{1 + e^{-\Delta E / \tau}}$

Where  $t$  is a decreasing temperature variable and  $\Delta E$

is the expected change in energy if the change is made.

The non-determinism allows the system to "jiggle" out of local

minima

# Unit -3

Fuzzy logic

# Overview

- What is Fuzzy Logic?
  - Where did it begin?
  - Fuzzy Logic vs. Neural Networks
  - Fuzzy Logic in Control Systems
  - Fuzzy Logic in Other Fields
  - Future

# WHAT IS FUZZY LOGIC?

- Definition of fuzzy
  - Fuzzy – “not clear, distinct, or precise; blurred”
- Definition of fuzzy logic
  - A form of knowledge representation suitable for notions that cannot be defined precisely, but which depend upon their contexts.

# TRADITIONAL REPRESENTATION OF LOGIC



Slow

Speed = 0



Fast

Speed = 1

```
bool speed;  
get the speed  
if ( speed == 0) {  
//  speed is slow  
}  
else {  
//  speed is fast  
}
```

# FUZZY LOGIC REPRESENTATION

- For every problem must represent in terms of fuzzy sets.
- What are fuzzy sets?



Slowest

[ 0.0 – 0.25 ]



Slow

[ 0.25 – 0.50 ]



Fast

[ 0.50 – 0.75 ]

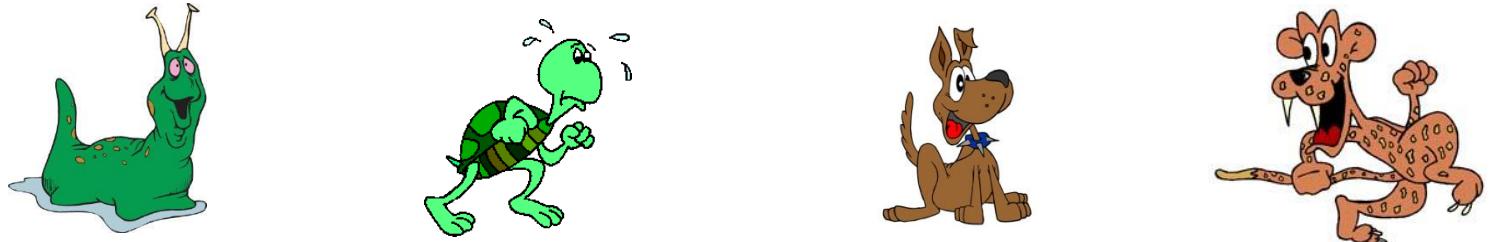


Fastest

[ 0.75 – 1.00 ]

# FUZZY LOGIC REPRESENTATION

## CONT.



Slowest

```
float speed;  
get the speed  
if ((speed >= 0.0)&&(speed < 0.25)) {  
// speed is slowest  
}  
else if ((speed >= 0.25)&&(speed < 0.5))  
{  
// speed is slow  
}  
else if ((speed >= 0.5)&&(speed < 0.75))  
{  
// speed is fast  
}  
else // speed >= 0.75 && speed < 1.0  
{  
// speed is fastest  
}
```

Fast

Fastest

# ORIGINS OF FUZZY LOGIC

- Traces back to Ancient Greece
- Lotfi Asker Zadeh ( 1965 )
  - First to publish ideas of fuzzy logic.
- Professor Toshire Terano ( 1972 )
  - Organized the world's first working group on fuzzy systems.
- F.L. Smidth & Co. ( 1980 )
  - First to market fuzzy expert systems.

# FUZZY LOGIC VS. NEURAL NETWORKS

- How does a Neural Network work?
- Both model the human brain.
  - Fuzzy Logic
  - Neural Networks
- Both used to create behavioral systems.

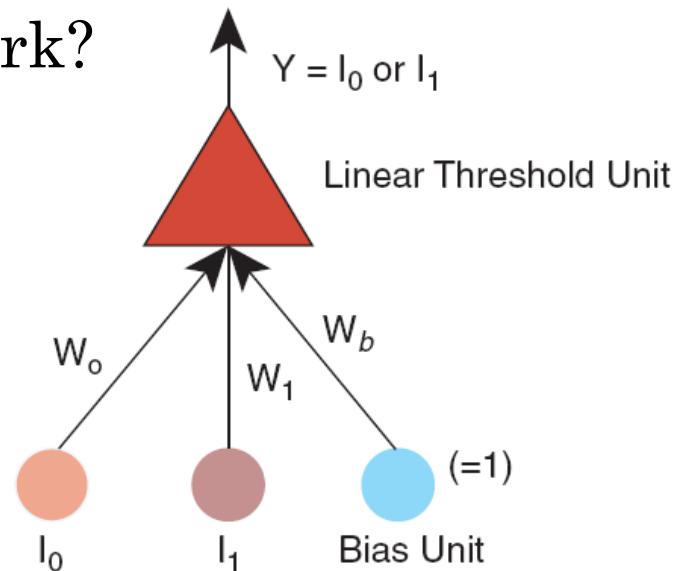


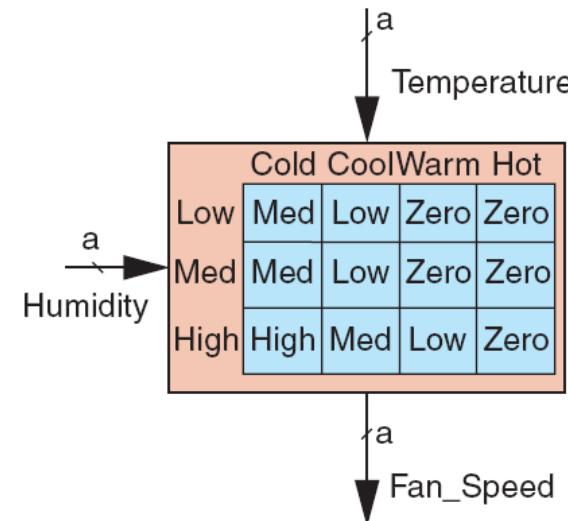
Fig. 2 A simple, single-unit adaptive network

# FUZZY LOGIC IN CONTROL SYSTEMS

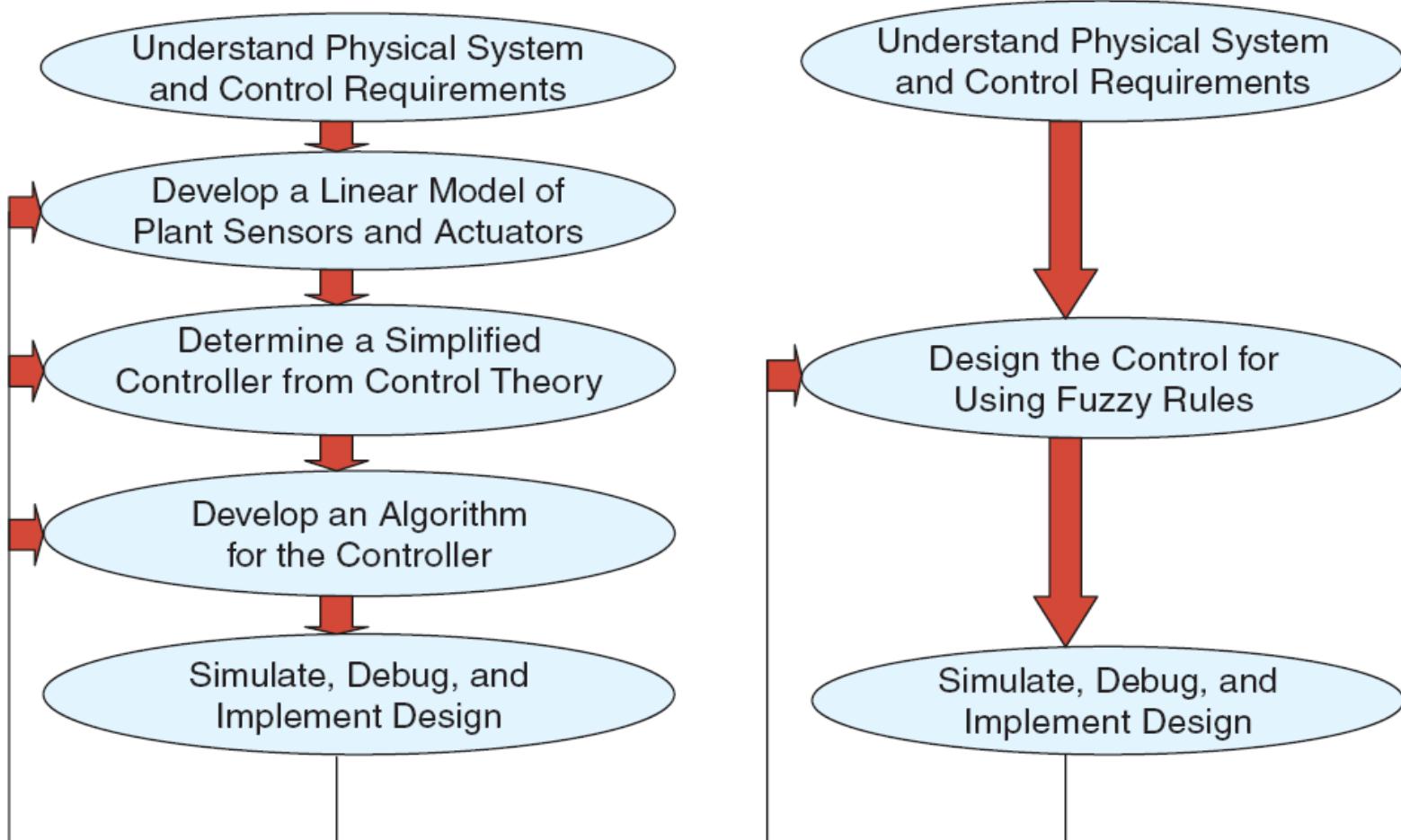
- Fuzzy Logic provides a more efficient and resourceful way to solve Control Systems.
- Some Examples
  - Temperature Controller
  - Anti – Lock Break System ( ABS )

# TEMPERATURE CONTROLLER

- The problem
  - Change the speed of a heater fan, based off the room temperature and humidity.
- A temperature control system has four settings
  - Cold, Cool, Warm, and Hot
- Humidity can be defined by:
  - Low, Medium, and High
- Using this we can define the fuzzy set.



# BENEFITS OF USING FUZZY LOGIC



# FUZZY LOGIC IN OTHER FIELDS

- Business
- Hybrid Modeling
- Expert Systems

# Fuzzy Logic Example

## Automotive Speed Controller

3 inputs:

- speed (5 levels)

- acceleration (3 levels)

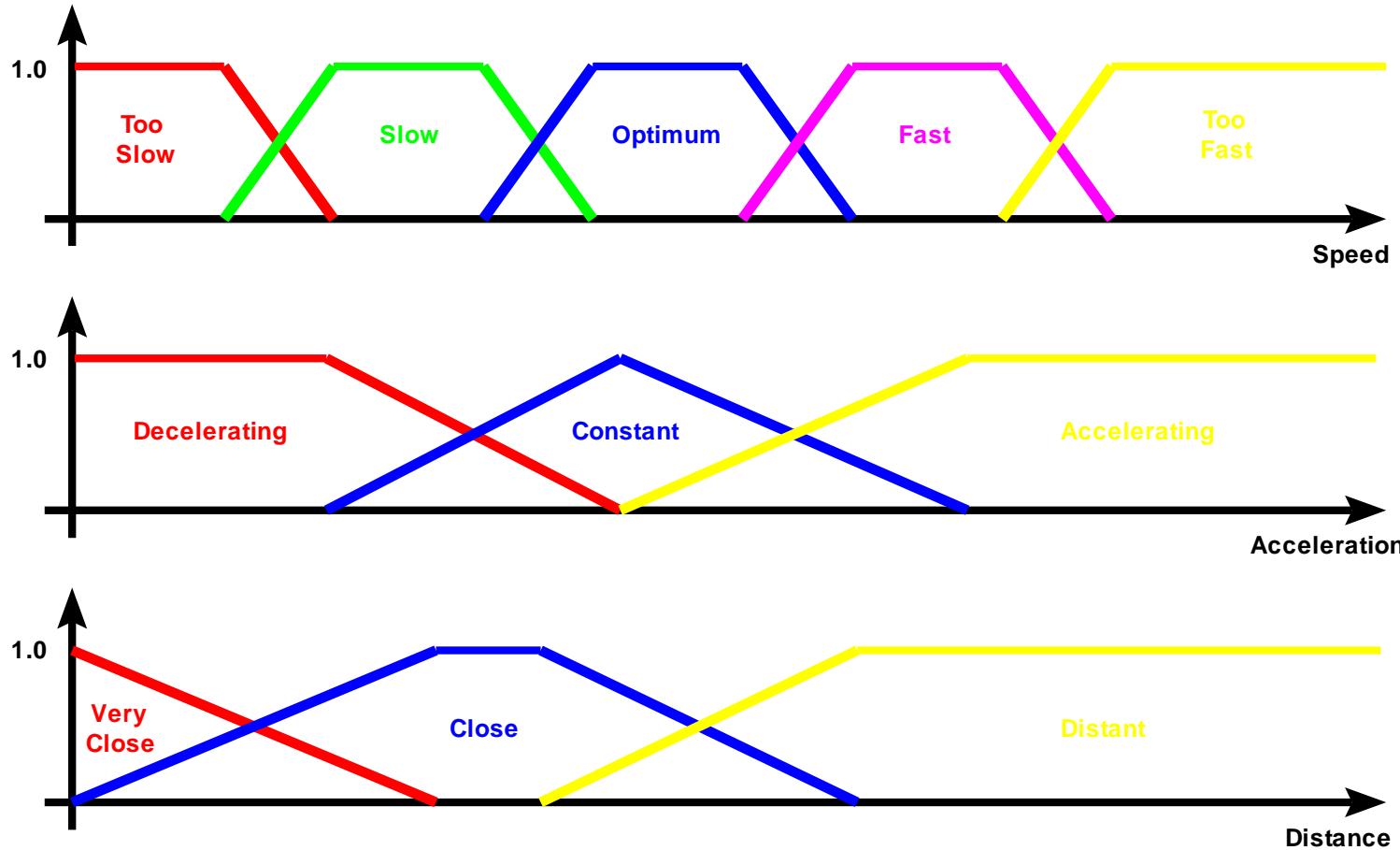
- distance to destination (3 levels)

1 output:

- power (fuel flow to engine)

Set of rules to determine output based on input values

# Fuzzy Logic Example



# Fuzzy Logic Example

## Example Rules

IF speed is TOO SLOW and acceleration is DECELERATING,  
THEN INCREASE POWER GREATLY

IF speed is SLOW and acceleration is DECREASING,  
THEN INCREASE POWER SLIGHTLY

IF distance is CLOSE,  
THEN DECREASE POWER SLIGHTLY

...

# Fuzzy Logic Example

Note there would be a total of 95 different rules for all combinations of inputs of 1, 2, or 3 at a time.

$$( 5 \times 3 \times 3 + 5 \times 3 + 5 \times 3 + 3 \times 3 + 5 + 3 + 3 )$$

In practice, a system won't require all the rules.

System tweaked by adding or changing rules and by adjusting set boundaries.

System performance can be very good but not usually optimized by traditional metrics (minimize RMS error).



# Fuzzy Logic Summary

Doesn't require an understanding of process but any knowledge will help formulate rules.

Complicated systems may require several iterations to find a set of rules resulting in a stable system.

Combining Neural Networks with fuzzy logic reduces time to establish rules by analyzing clusters of data.

Possible applications: Master Production Schedule, Material Requirements Planning, Inventory Capacity Planning

# CONCLUSION

- Fuzzy logic provides an alternative way to represent linguistic and subjective attributes of the real world in computing.
- It is able to be applied to control systems and other applications in order to improve the efficiency and simplicity of the design process.

# UNIT-4

## FUZZY ARITHMETIC

# Definition

- Fuzzy Number
  - Convex and normal fuzzy set defined on  $\mathbb{R}$
  - Equivalently it satisfies
    - Normal fuzzy set on  $\mathbb{R}$
    - Every alpha-cut must be a closed interval
    - Support must be bounded
- Applications of fuzzy number
  - Fuzzy Control, Decision Making, Optimizations

# Arithmetic Operations

- Interval Operations       $A = [a_1, a_3], \quad B = [b_1, b_3]$

Addition

$$[a_1, a_3](+)[b_1, b_3] = [a_1 + b_1, a_3 + b_3]$$

Subtraction

$$[a_1, a_3](-)[b_1, b_3] = [a_1 - b_3, a_3 - b_1]$$

Multiplication

$$\begin{aligned} [a_1, a_3](\bullet)[b_1, b_3] = & [a_1 \bullet b_1 \wedge a_1 \bullet b_3 \wedge a_3 \bullet b_1 \wedge a_3 \bullet b_3, \\ & a_1 \bullet b_1 \vee a_1 \bullet b_3 \vee a_3 \bullet b_1 \vee a_3 \bullet b_3] \end{aligned}$$

Division

$$\begin{aligned} [a_1, a_3](/)[b_1, b_3] = & [a_1 / b_1 \wedge a_1 / b_3 \wedge a_3 / b_1 \wedge a_3 / b_3, \\ & a_1 / b_1 \vee a_1 / b_3 \vee a_3 / b_1 \vee a_3 / b_3] \\ \text{except } & b_1 = b_3 = 0 \end{aligned}$$

# Examples

- Addition

$$[2,5]+[1,3]=[3,8] \quad [0,1]+[-6,5]=[-6,6]$$

- Subtraction

$$[2,5]-[1,3]=[-1,4] \quad [0,1]-[-6,5]=[-5,7]$$

- Multiplication

$$[-1,1]*[-2,-0.5]=[-2,2] \quad [3,4]*[2,2]=[6,8]$$

- Division

$$[-1,1]/[-2,-0.5]=[-2,2] \quad [4,10]*[1,2]=[2,10]$$

# Properties of Interval Operations

Commutative

$$A + B = B + A \quad A \cdot B = B \cdot A$$

Assocoative

$$(A + B) + C = A + (B + C) \quad (A \cdot B) \cdot C = A \cdot (B \cdot C)$$

Identity  $0 = [0,0]$   $1 = [1,1]$

$$A = A + 0 = 0 + A \quad A = A \cdot 1 = 1 \cdot A$$

Subdistributive

$$A \cdot (B + C) \subseteq A \cdot B + A \cdot C$$

Inverse

$$0 \in A - A \quad 1 \in A/A$$

Monotonicity for any operations\*

If  $A \subseteq E$  and  $B \subseteq F$  then  $A * B \subseteq E * F$

# Arithmetic Operation on Fuzzy Numbers

- Interval operations of alpha-level sets

${}^{\alpha}(A * B)={}^{\alpha}A * {}^{\alpha}B$  for any  $\alpha \in (0,1]$ .

When  $* = /$ ,  $0 \notin {}^{\alpha}B$  for all  $\alpha \in (0,1]$ .

$$A * B = \bigcup_{\alpha \in (0,1]} (A * {}^{\alpha}B)$$

- Note: The Result is a fuzzy number.
- Example: See Text pp. 105 and Fig. 4.5

# Example

- $A+B = \{1/5, 0.8/6, 0.5/7\}$

i)  $z < 5$       No such case.

$$\mu_{A(+)}(z) = 0$$

ii)  $z = 5$

$$x + y = 2 + 3$$

$$\mu_A(2) \wedge \mu_B(3) = 1$$

iii)  $z = 6$

$$x + y = 3 + 3 \quad \text{or} \quad x + y = 2 + 4$$

$$\mu_A(3) \wedge \mu_B(3) = 0.5$$

$$\mu_A(2) \wedge \mu_B(4) = 0.8 \quad \mu_{A(+)}(6) = \bigvee_{\substack{6=3+3 \\ 6=2+4}} (0.5, 0.8) = 0.8$$

iv)  $z = 7$

$$\mu_A(3) \wedge \mu_B(4) = \min(0.5, 0.8) = 0.5$$

# Example

- $\text{Max } (A, B) = \{ (3, 1), (4, 0.5) \}$

i)  $z \leq 2$  No such case.

$$\mu_{A(\vee)B}(z) = 0$$

ii)  $z = 3$

$$x \vee y = 2 \vee 3 \quad \text{or} \quad x \vee y = 3 \vee 3$$

$$\mu_A(2) \wedge \mu_B(3) = 1 \wedge 1 = 1 \quad \mu_A(3) \wedge \mu_B(3) = 0.5 \wedge 1 = 0.5$$

$$\mu_{A(\vee)B}(3) = \bigvee_{\substack{3=2 \vee 3 \\ 3=3 \vee 3}} (1, 0.5) = 1$$

iii)  $z = 4$

$$x \vee y = 2 \vee 4 \quad \text{or} \quad x \vee y = 3 \vee 4$$

$$\mu_A(2) \wedge \mu_B(4) = 1 \wedge 0.5 = 0.5 \quad \mu_A(3) \wedge \mu_B(4) = 0.5 \wedge 0.5 = 0.5$$

$$\mu_{A(\vee)B}(4) = \bigvee_{\substack{4=2 \vee 4 \\ 4=3 \vee 4}} (0.5, 0.5) = 0.5$$

iv)  $z \geq 5$  No such case.

$$\mu_{A(\vee)B}(z) = 0$$

# Typical Fuzzy Numbers

- **Triangular Fuzzy Number**

- Fig. 4.5

- **Trapezoidal Fuzzy Numbers: Fig. 4.4**

- Linguistic variable: "Performance"
  - Linguistic values (terms):  
"very small" ... "very large"
  - Semantic Rules:  
Terms maps on trapezoidal fuzzy numbers
  - Syntactic rules (grammar):  
Rules for other terms such as "not small"

# Lattice of Fuzzy Numbers

- Lattice
  - Partially ordered set with ordering relation
  - Meet(g.l.b) and Join(l.u.b) operations
  - Example:  
Real number and “less than or equal to”
- Lattice of fuzzy numbers

$$MIN(A, B)(z) = \sup_{z=\min(x, y)} \min[A(x), B(y)] = MEET(A, B)$$

$$MAX(A, B)(z) = \sup_{z=\max(x, y)} \min[A(x), B(y)] = JOIN(A, B)$$

# Fuzzy Equations

- Addition  $A + X = B$ 
  - $X = B - A$  is not a solution because  $A + (B - A)$  is not  $B$ .
  - Conditions to have a solution

For any  $\alpha \in (0,1]$ , let  ${}^\alpha A = [{}^\alpha a_1, {}^\alpha a_2]$ ,  ${}^\alpha B = [{}^\alpha b_1, {}^\alpha b_2]$  and  ${}^\alpha X = [{}^\alpha x_1, {}^\alpha x_2]$

(i)  ${}^\alpha b_1 - {}^\alpha a_1 \leq {}^\alpha b_2 - {}^\alpha a_2$  for any  $\alpha \in (0,1]$

(ii)  $\alpha \leq \beta$  implies  ${}^\alpha b_1 - {}^\alpha a_1 \leq {}^\beta b_1 - {}^\beta a_1 \leq {}^\beta b_2 - {}^\beta a_2 \leq {}^\alpha b_2 - {}^\alpha a_2$

- Solution

Suppose  ${}^\alpha X = [{}^\alpha b_1 - {}^\alpha a_1, {}^\alpha b_2 - {}^\alpha a_2]$  is a solution of  ${}^\alpha A + {}^\alpha X = {}^\alpha B$  for any  $\alpha \in (0,1]$ . Then

$$X = \bigcup_{\alpha \in (0,1]} {}^\alpha X$$

# Fuzzy Equations

- Multiplication  $A \cdot X = B$

- $X = B/A$  is not a solution.
- Conditions to have a solution

For any  $\alpha \in (0,1]$ , let  ${}^\alpha A = [{}^\alpha a_1, {}^\alpha a_2]$ ,  ${}^\alpha B = [{}^\alpha b_1, {}^\alpha b_2]$  and  ${}^\alpha X = [{}^\alpha x_1, {}^\alpha x_2]$

(i)  ${}^\alpha b_1 / {}^\alpha a_1 \leq {}^\alpha b_2 / {}^\alpha a_2$  for any  $\alpha \in (0,1]$

(ii)  $\alpha \leq \beta$  implies  ${}^\alpha b_1 / {}^\alpha a_1 \leq {}^\beta b_1 / {}^\beta a_1 \leq {}^\beta b_2 / {}^\beta a_2 \leq {}^\alpha b_2 / {}^\alpha a_2$

- Solution

Suppose  ${}^\alpha X = [{}^\alpha b_1 / {}^\alpha a_1, {}^\alpha b_2 / {}^\alpha a_2]$  is a solution of  
 ${}^\alpha A \cdot {}^\alpha X = {}^\alpha B$  for any  $\alpha \in (0,1]$ . Then

$$X = \bigcup_{\alpha \in (0,1]} {}^\alpha X$$

# Fuzzification

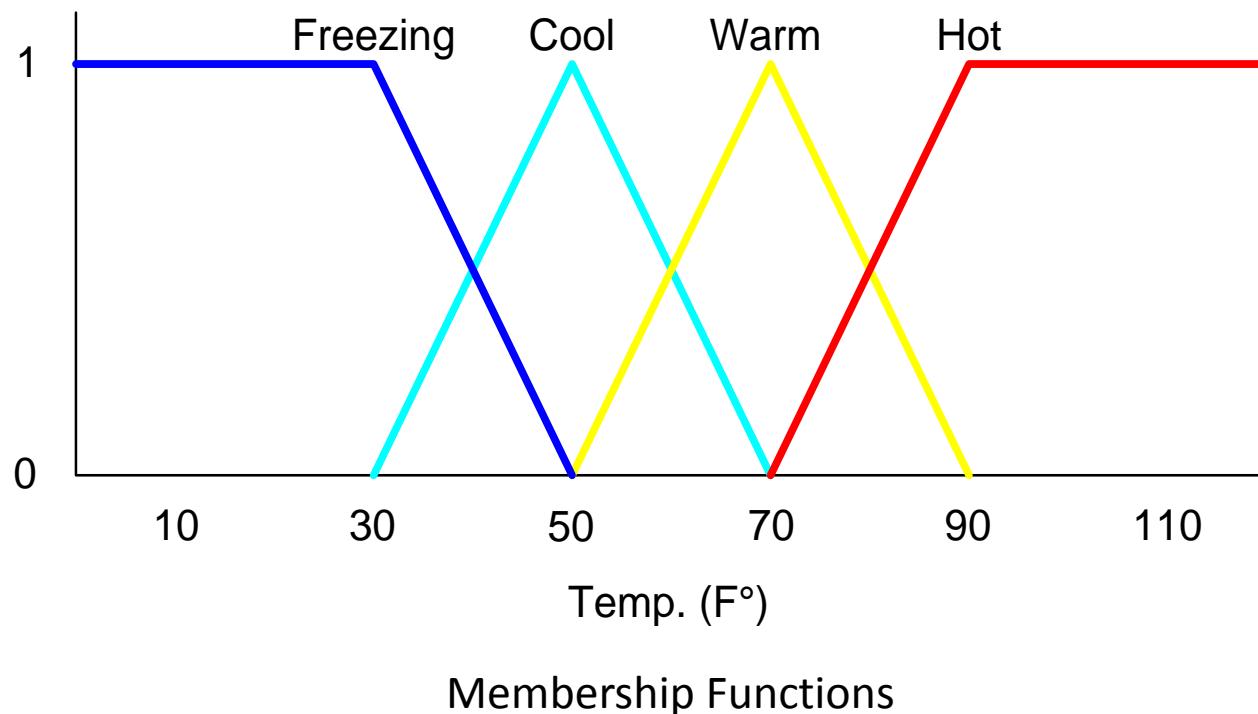
**Fuzzification** is the process of changing a real scalar value into a fuzzy value.

This is achieved with the different types of **fuzzifiers (membership functions)**.

# Fuzzification

Temp: {Freezing, Cool, Warm, Hot}

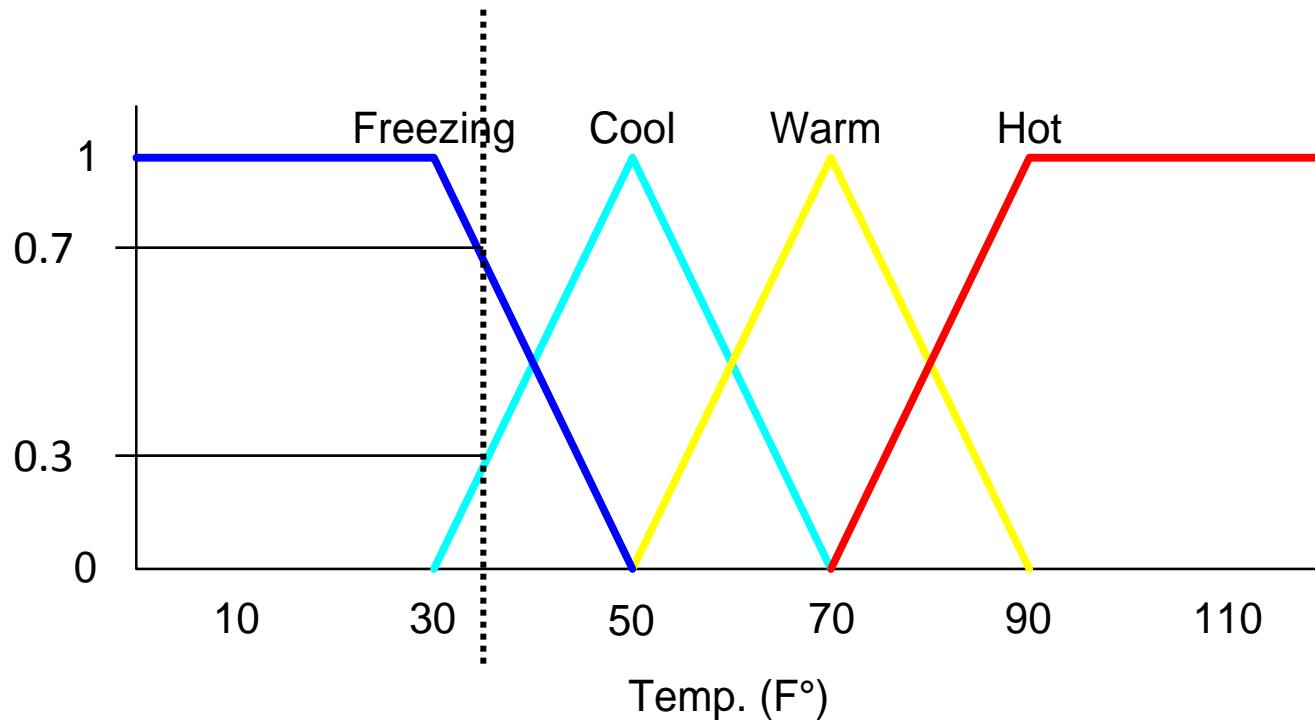
Degree of Truth or "Membership"



# Fuzzification

How cool is 36 F° ?

It is 30% Cool and 70% Freezing



# Fuzzification

## Membership Functions

The MATLAB toolbox includes 11 built-in membership function types. These 11 functions are, in turn, built from several basic functions:

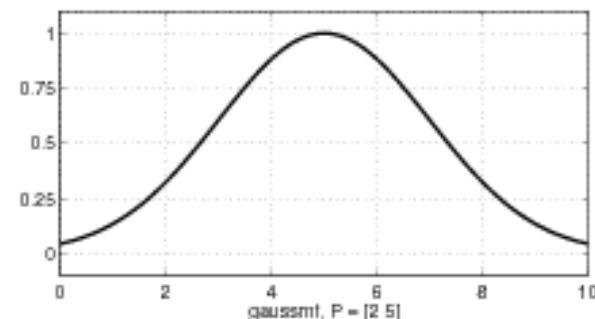
- piecewise linear functions
- the Gaussian distribution function
- the sigmoid curve
- quadratic and cubic polynomial curves

# Fuzzification

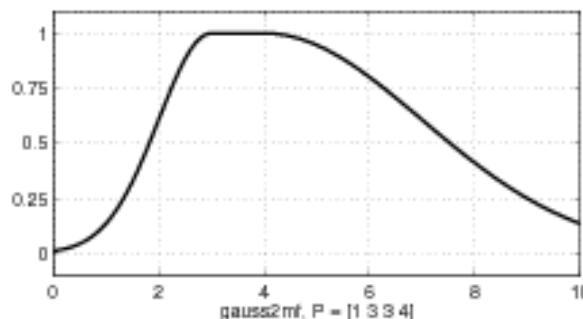
## Membership Functions

Two membership functions are built on the Gaussian distribution curve: a simple Gaussian curve and a two-sided composite of two different Gaussian curves. The two functions are **gaussmf** and **gauss2mf**. The generalized bell membership has the function name **gbellmf**.

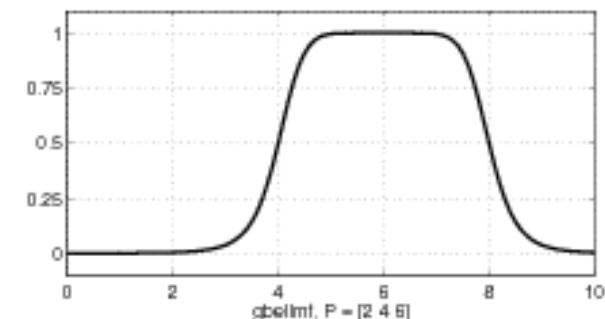
Because of their smoothness and concise notation, Gaussian and bell membership functions are popular methods for specifying fuzzy sets. Both of these curves have the advantage of being smooth and nonzero at all points.



gaussmf



gauss2mf



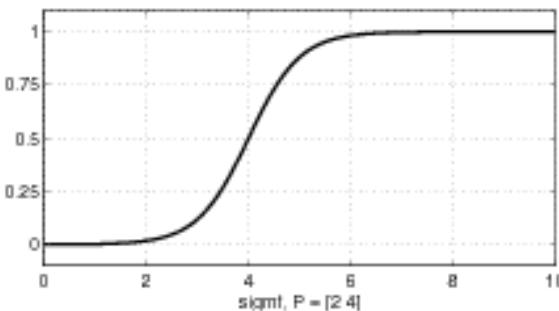
gbellmf

# Fuzzification

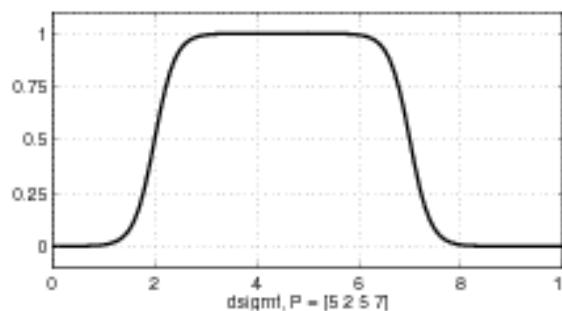
## Membership Functions

Although the Gaussian membership functions and bell membership functions achieve smoothness, they are unable to specify asymmetric membership functions, which are important in certain applications.

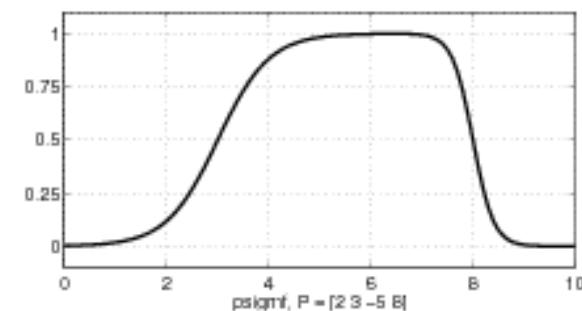
the **sigmoidal membership function** is defined, which is either open left or right. Asymmetric and closed (i.e. not open to the left or right) membership functions can be synthesized using two sigmoidal functions, so in addition to the basic **sigmf**, you also have the difference between two sigmoidal functions, **dsigmf**, and the product of two sigmoidal functions **psigmf**.



**sigmf**



**dsigmf**



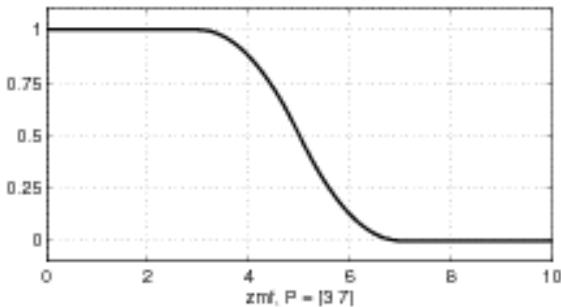
**psigmf**

# Fuzzification

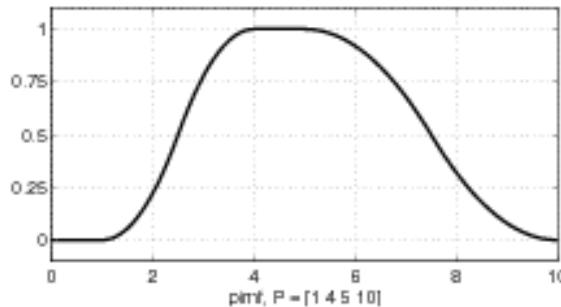
## Membership Functions

**Polynomial based** curves account for several of the membership functions in the toolbox.

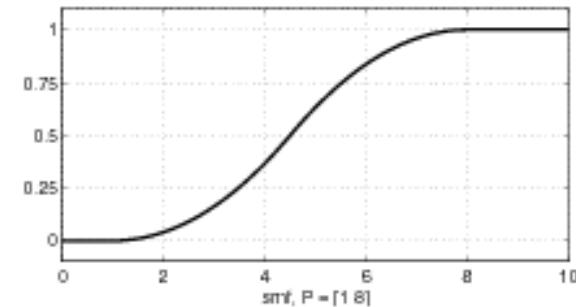
Three related membership functions are the **Z, S, and Pi curves**, all named because of their shape. The function **zmf** is the asymmetrical polynomial curve open to the left, **smf** is the mirror-image function that opens to the right, and **pimf** is zero on both extremes with a rise in the middle.



zmf



pimf



smf

# Unit -5

## GENETIC ALGORITHMS

# Introduction

- After scientists became disillusioned with classical and neo-classical attempts at modeling intelligence, they looked in other directions.
- Two prominent fields arose, connectionism (neural networking, parallel processing) and evolutionary computing.
- It is the latter that this essay deals with - genetic algorithms and genetic programming.

# What is GA

- A **genetic algorithm** (or **GA**) is a search technique used in computing to find true or approximate solutions to optimization and search problems.
- Genetic algorithms are categorized as global search heuristics.
- Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination).

# What is GA

- Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions.
- Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

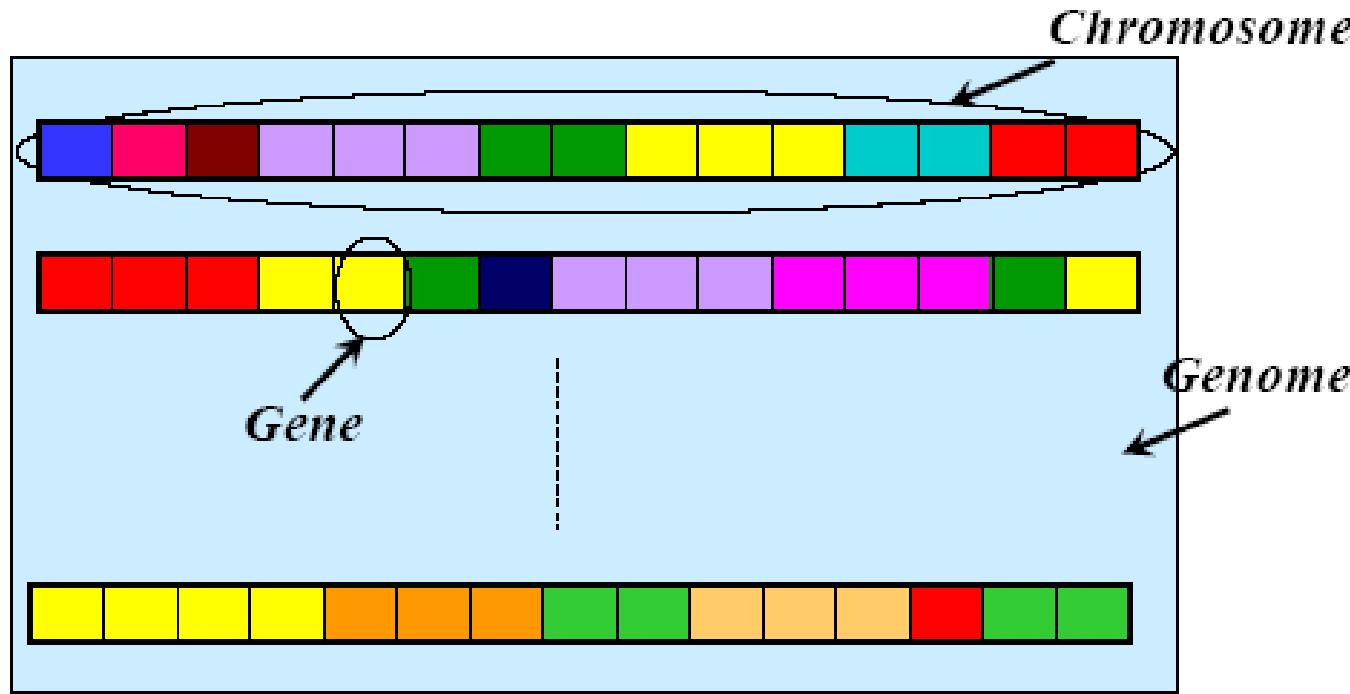
# What is GA

- The new population is then used in the next iteration of the algorithm.
- Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.
- If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

# Key terms

- **Individual** - Any possible solution
- **Population** - Group of all *individuals*
- **Search Space** - All possible solutions to the problem
- **Chromosome** - Blueprint for an *individual*
- **Trait** - Possible aspect (*features*) of an *individual*
- **Allele** - Possible settings of trait (black, blond, etc.)
- **Locus** - The position of a *gene* on the *chromosome*
- **Genome** - Collection of all *chromosomes* for an *individual*

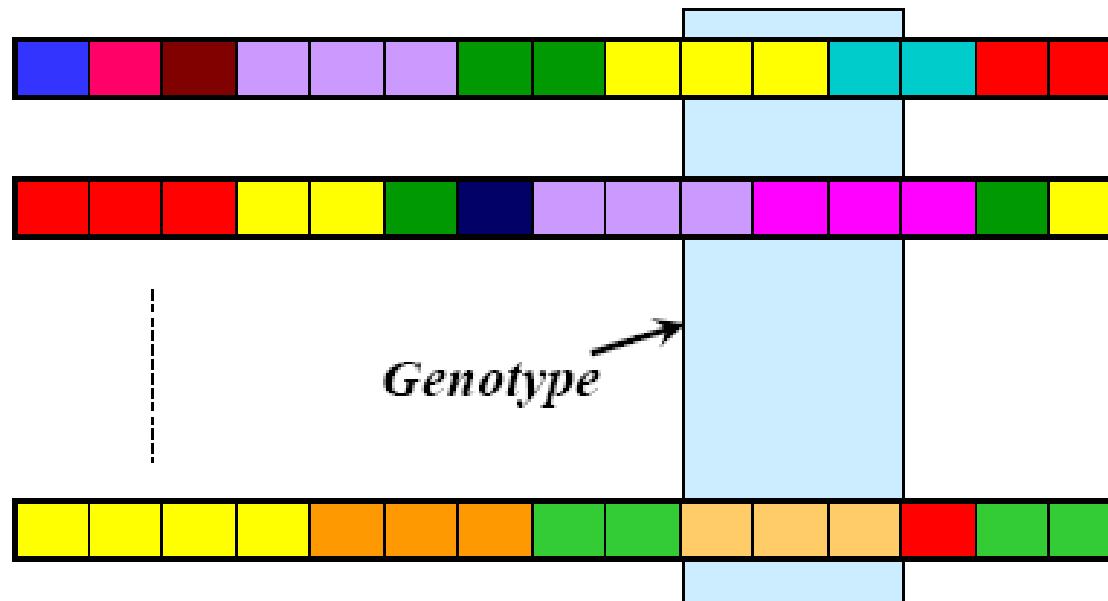
# Chromosome, Genes and Genomes



# Genotype and Phenotype

- ***Genotype:***
  - Particular set of genes in a genome
- ***Phenotype:***
  - Physical characteristic of the genotype  
(smart, beautiful, healthy, etc.)

# Genotype and Phenotype



# GA Requirements

- A typical genetic algorithm requires two things to be defined:
  - a genetic representation of the solution domain, and
  - a fitness function to evaluate the solution domain.
- 
- A standard representation of the solution is as an array of bits. Arrays of other types and structures can be used in essentially the same way.
  - The main property that makes these genetic representations convenient is that their parts are easily aligned due to their fixed size, that facilitates simple crossover operation.
  - Variable length representations may also be used, but crossover implementation is more complex in this case.
  - Tree-like representations are explored in Genetic programming.

# Representation

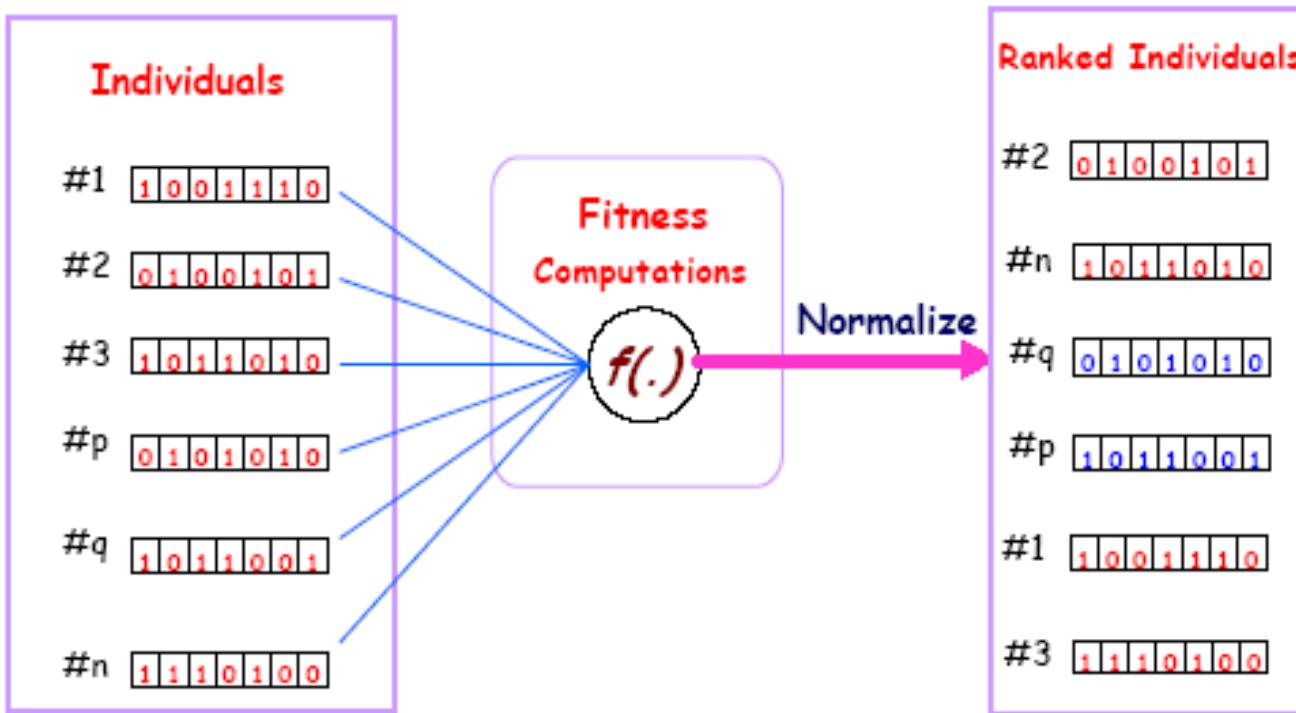
Chromosomes could be:

- Bit strings (0101 ... 1100)
- Real numbers (43.2 -33.1 ... 0.0 89.2)
- Permutations of elements (E11 E3 E7 ... E1 E15)
- Lists of rules (R1 R2 R3 ... R22 R23)
- Program elements (genetic programming)

# GA Requirements

- The fitness function is defined over the genetic representation and measures the *quality* of the represented solution.
- The fitness function is always problem dependent.
- For instance, in the [knapsack problem](#) we want to maximize the total value of objects that we can put in a knapsack of some fixed capacity.
- A representation of a solution might be an array of bits, where each bit represents a different object, and the value of the bit (0 or 1) represents whether or not the object is in the knapsack.
- Not every such representation is valid, as the size of objects may exceed the capacity of the knapsack.
- The *fitness* of the solution is the sum of values of all objects in the knapsack if the representation is valid, or 0 otherwise. In some problems, it is hard or even impossible to define the fitness expression: in these

# A fitness function



# General Algorithm for GA

- **Initialization**
- Initially many individual solutions are randomly generated to form an initial population. The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions.
- Traditionally, the population is generated randomly, covering the entire range of possible solutions (the *search space*).
- Occasionally, the solutions may be "seeded" in areas where optimal solutions are likely to be found.

# General Algorithm for GA

- **Selection**
- During each successive generation, a proportion of the existing population is selected to breed a new generation.
- Individual solutions are selected through a *fitness-based* process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected.
- Certain selection methods rate the fitness of each solution and preferentially select the best solutions. Other methods rate only a random sample of the population, as this process may be very time-consuming.
- Most functions are stochastic and designed so that a small proportion of less fit solutions are selected. This helps keep the diversity of the population large, preventing premature convergence on poor solutions. Popular and well-studied selection methods include roulette wheel selection and tournament selection

# General Algorithm for GA

- In roulette wheel selection, individuals are given a probability of being selected that is directly proportionate to their fitness.
- Two individuals are then chosen randomly based on these probabilities and produce offspring.

# General Algorithm for GA

- These processes ultimately result in the next generation population of chromosomes that is different from the initial generation.
- Generally the average fitness will have increased by this procedure for the population, since only the best organisms from the first generation are selected for breeding, along with a small proportion of less fit solutions, for reasons already mentioned above.

# Evolving Neural Networks

- Evolving the architecture of neural network is slightly more complicated, and there have been several ways of doing it. For small nets, a simple matrix represents which neuron connects which, and then this matrix is, in turn, converted into the necessary 'genes', and various combinations of these are evolved.

# Evolving Neural Networks

- Many would think that a learning function could be evolved via genetic programming. Unfortunately, genetic programming combined with neural networks could be *incredibly* slow, thus impractical.
- As with many problems, you have to constrain what you are attempting to create.
- For example, in 1990, David Chalmers attempted to evolve a function as good as the delta rule.
- He did this by creating a general equation based upon the delta rule with 8 unknowns, which the genetic algorithm then evolved.

# Example

- $f(x) = \{\text{MAX}(x^2): 0 \leq x \leq 32\}$
- Encode Solution: Just use 5 bits (1 or 0).
- Generate initial population.

<b>A</b>	0	1	1	0	1
<b>B</b>	1	1	0	0	0
<b>C</b>	0	1	0	0	0
<b>D</b>	1	0	0	1	1

- Evaluate each solution against objective.

Sol.	String	Fitness	% of Total
<b>A</b>	01101	169	14.4
<b>B</b>	11000	576	49.2
<b>C</b>	01000	64	5.5
<b>D</b>	10011	361	30.9

# Example Cont'd

- Create next generation of solutions
  - Probability of “being a parent” depends on the fitness.
- Ways for parents to create next generation
  - Reproduction
    - Use a string again unmodified.
  - Crossover
    - Cut and paste portions of one string to another.
  - Mutation
    - Randomly flip a bit.
  - COMBINATION of all of the above.

# THANKYOU

# Fuzzy Logic : Introduction

**Debasis Samanta**

IIT Kharagpur

*dsamanta@iitkgp.ac.in*

23.01.2018

# What is Fuzzy logic?

- Fuzzy logic is a mathematical language to **express** something.  
This means it has grammar, syntax, semantic like a language for communication.
- There are some other mathematical languages also known
  - **Relational algebra** (operations on sets)
  - **Boolean algebra** (operations on Boolean variables)
  - **Predicate logic** (operations on well formed formulae (wff), also called predicate propositions)
- **Fuzzy logic deals with Fuzzy set.**

# A brief history of Fuzzy Logic

- First time introduced by **Lotfi Abdelli Zadeh** (1965), University of California, Berkley, USA (1965).



- He is fondly nick-named as **LAZ**

# A brief history of Fuzzy logic



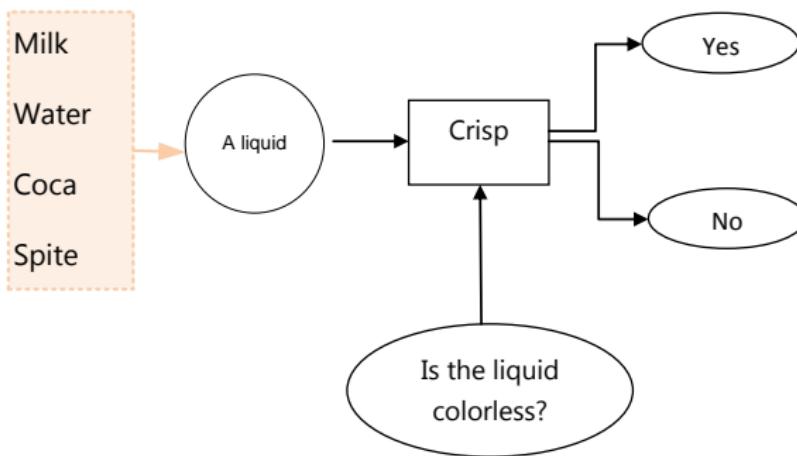
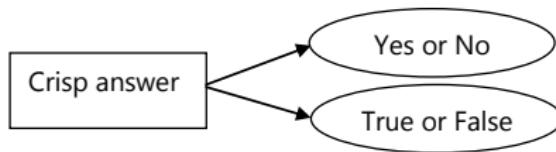
- ① Dictionary meaning of **fuzzy** is not clear, noisy etc.

Example: Is the picture on this slide is fuzzy?

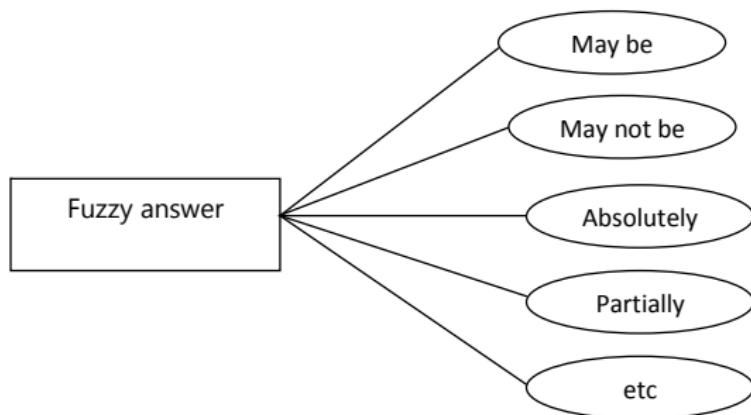
- ② Antonym of fuzzy is **crisp**

Example: Are the chips crisp?

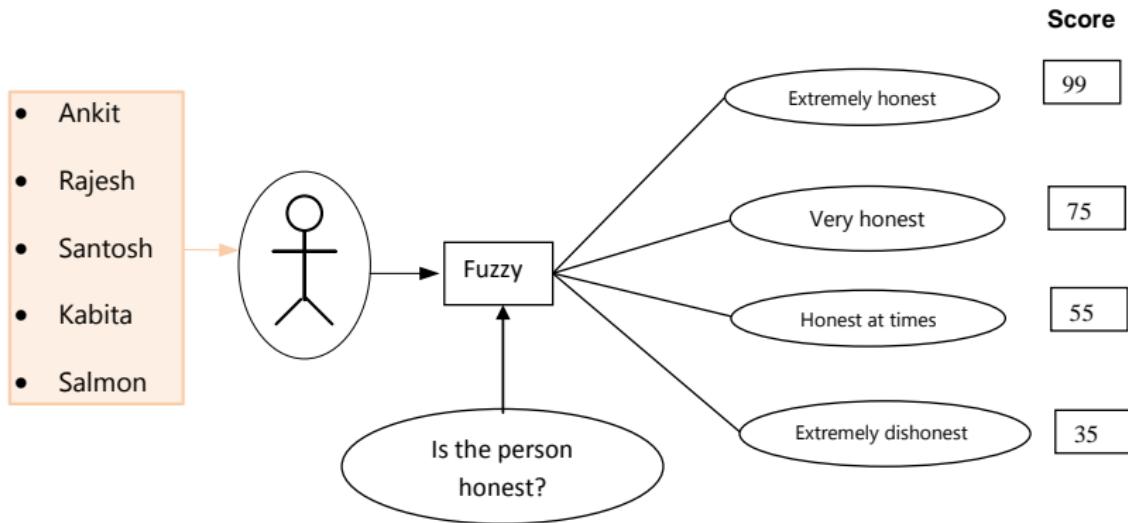
# Example : Fuzzy logic vs. Crisp logic



# Example : Fuzzy logic vs. Crisp logic



# Example : Fuzzy logic vs. Crisp logic

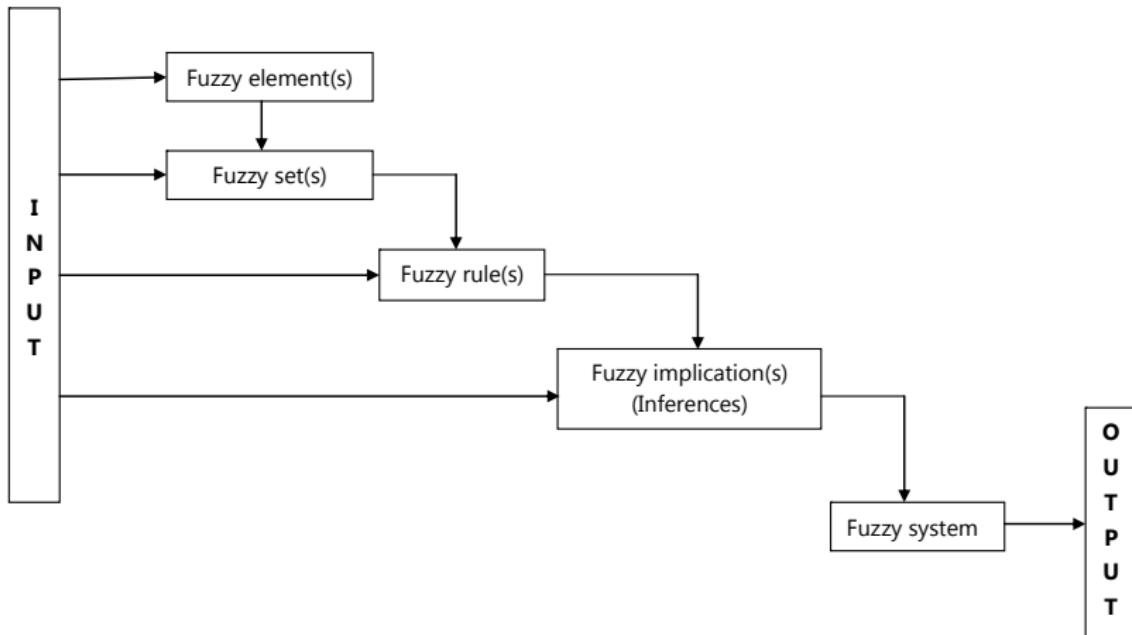


# World is fuzzy!



**Our world is better  
described with  
fuzzily!**

# Concept of fuzzy system



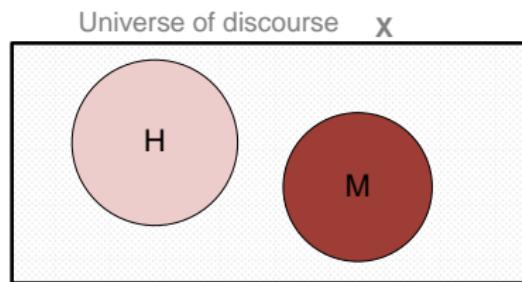
# Concept of fuzzy set

To understand the concept of **fuzzy set** it is better, if we first clear our idea of **crisp set**.

$X$  = The entire population of India.

$H$  = All Hindu population =  $\{ h_1, h_2, h_3, \dots, h_L \}$

$M$  = All Muslim population =  $\{ m_1, m_2, m_3, \dots, m_N \}$



Here, All are the sets of finite numbers of individuals.

Such a set is called **crisp set**.

# Example of fuzzy set

Let us discuss about fuzzy set.

X = All students in IT60108.

S = All **Good students**.

S = { (s, g) | s ∈ X } and g(s) is a measurement of goodness of the student s.

**Example:**

S = { (Rajat, 0.8), (Kabita, 0.7), (Salman, 0.1), (Ankit, 0.9) } etc.

# Fuzzy set vs. Crisp set

Crisp Set	Fuzzy Set
1. $S = \{ s \mid s \in X \}$	1. $F = (s, \mu) \mid s \in X$ and $\mu(s)$ is the degree of $s$ .
2. It is a collection of elements.	2. It is collection of ordered pairs.
3. Inclusion of an element $s \in X$ into $S$ is crisp, that is, has strict boundary <b>yes or no</b> .	3. Inclusion of an element $s \in X$ into $F$ is fuzzy, that is, if present, then with a degree of <b>membership</b> .

# Fuzzy set vs. Crisp set

**Note:** A crisp set is a fuzzy set, but, a fuzzy set is not necessarily a crisp set.

Example:

$$H = \{ (h_1, 1), (h_2, 1), \dots, (h_L, 1) \}$$

$$\text{Person} = \{ (p_1, 1), (p_2, 0), \dots, (p_N, 1) \}$$

In case of a crisp set, the elements are with extreme values of degree of membership namely either 1 or 0.

How to decide the degree of memberships of elements in a fuzzy set?

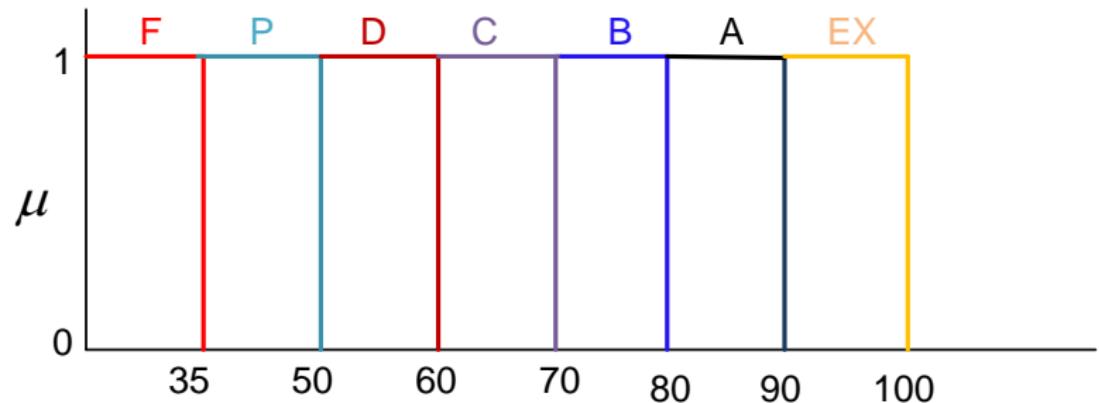
City	Bangalore	Bombay	Hyderabad	Kharagpur	Madras	Delhi
DoM	0.95	0.90	0.80	0.01	0.65	0.75

How the cities of comfort can be judged?

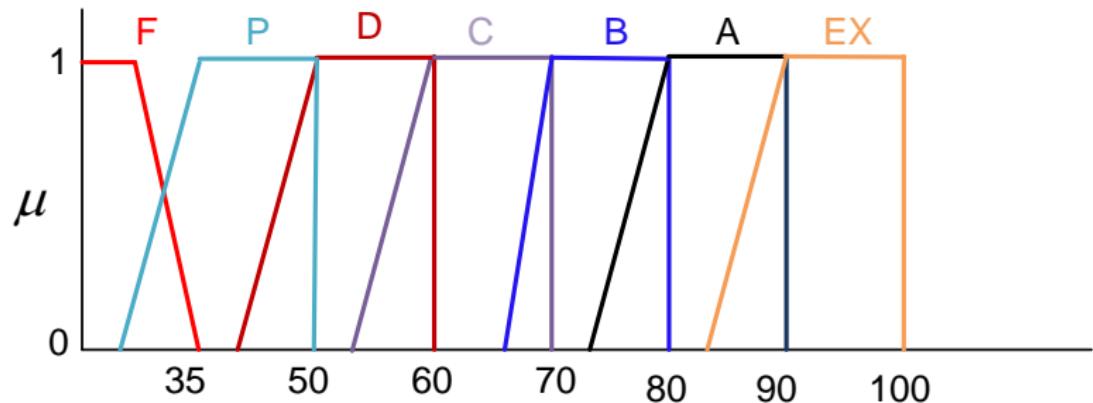
# Example: Course evaluation in a crisp way

- ① EX = Marks  $\geq 90$
- ② A =  $80 \leq \text{Marks} < 90$
- ③ B =  $70 \leq \text{Marks} < 80$
- ④ C =  $60 \leq \text{Marks} < 70$
- ⑤ D =  $50 \leq \text{Marks} < 60$
- ⑥ P =  $35 \leq \text{Marks} < 50$
- ⑦ F = Marks  $< 35$

# Example: Course evaluation in a crisp way



# Example: Course evaluation in a fuzzy way



## Few examples of fuzzy set

- High Temperature
- Low Pressure
- Color of Apple
- Sweetness of Orange
- Weight of Mango

Note: Degree of membership values lie in the range [0...1].

# Some basic terminologies and notations

## Definition 1: Membership function (and Fuzzy set)

If  $X$  is a universe of discourse and  $x \in X$ , then a fuzzy set  $A$  in  $X$  is defined as a set of ordered pairs, that is

$A = \{(x, \mu_A(x)) | x \in X\}$  where  $\mu_A(x)$  is called the **membership function** for the fuzzy set  $A$ .

### Note:

$\mu_A(x)$  map each element of  $X$  onto a membership grade (or membership value) between 0 and 1 (both inclusive).

### Question:

How (and who) decides  $\mu_A(x)$  for a Fuzzy set  $A$  in  $X$ ?

# Some basic terminologies and notations

## Example:

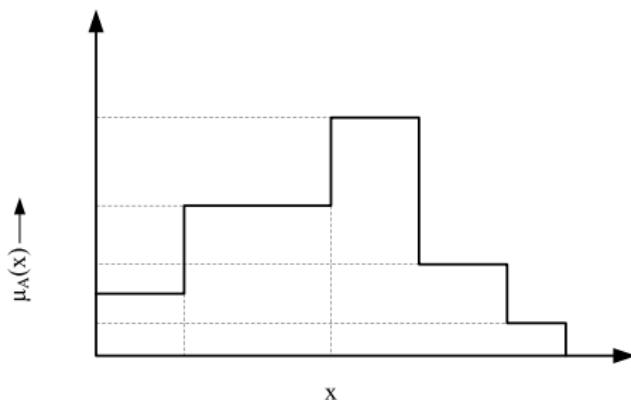
X = All cities in India

A = City of comfort

$A = \{(New\ Delhi, 0.7), (Bangalore, 0.9), (Chennai, 0.8), (Hyderabad, 0.6), (Kolkata, 0.3), (Kharagpur, 0)\}$

# Membership function with discrete membership values

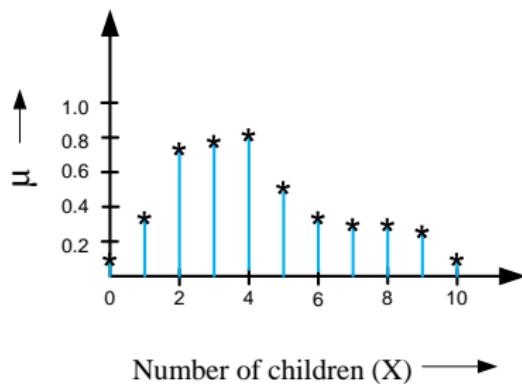
The membership values may be of discrete values.



A fuzzy set with discrete values of  $\mu$

# Membership function with discrete membership values

Either elements or their membership values (or both) also may be of discrete values.



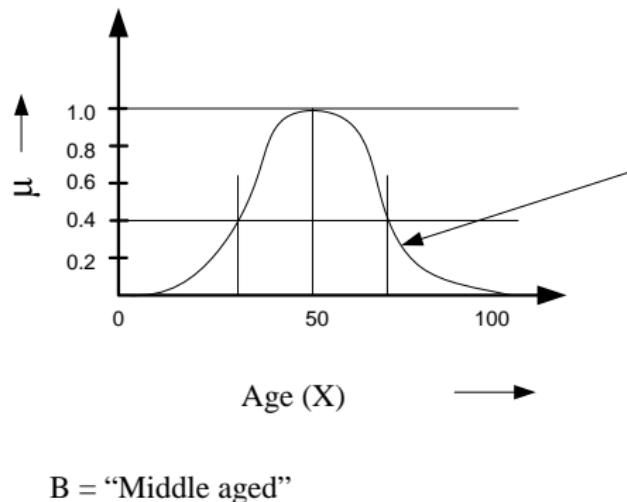
$$A = \{(0,0.1), (1,0.30), (2,0.78), \dots, (10,0.1)\}$$

Note : X = discrete value

How you measure happiness ??

A = "Happy family"

# Membership function with continuous membership values



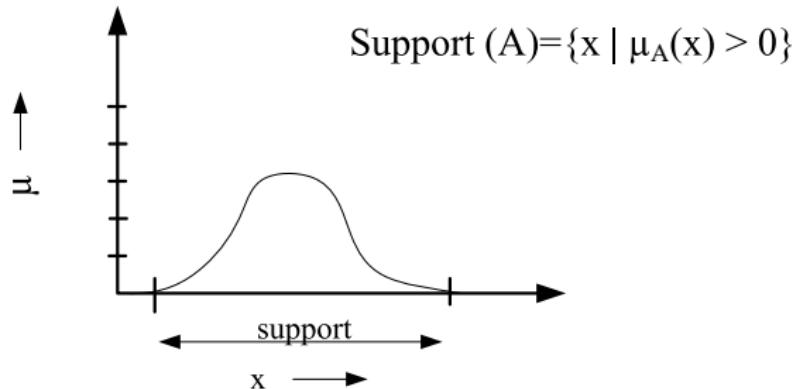
$$\mu_B(x) = \frac{1}{1 + \left(\frac{x-50}{10}\right)^4}$$

$$B = \{(x, \mu_B(x))\}$$

Note :  $x = \text{real value} = \mathbb{R}^+$

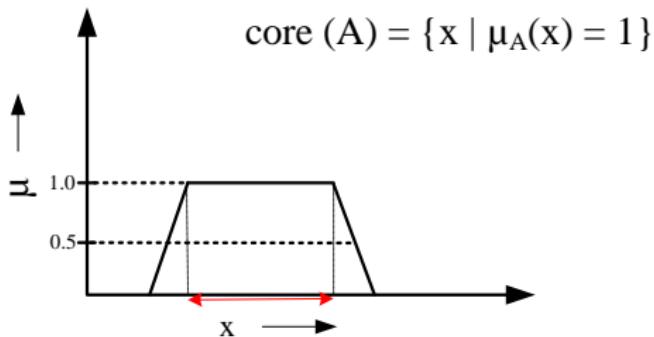
# Fuzzy terminologies: Support

**Support:** The support of a fuzzy set  $A$  is the set of all points  $x \in X$  such that  $\mu_A(x) > 0$



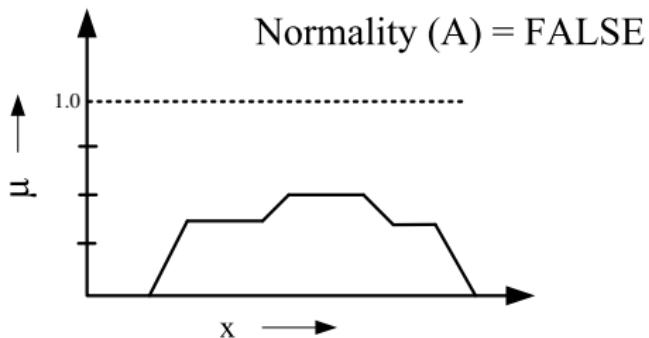
# Fuzzy terminologies: Core

**Core:** The core of a fuzzy set  $A$  is the set of all points  $x$  in  $X$  such that  $\mu_A(x) = 1$



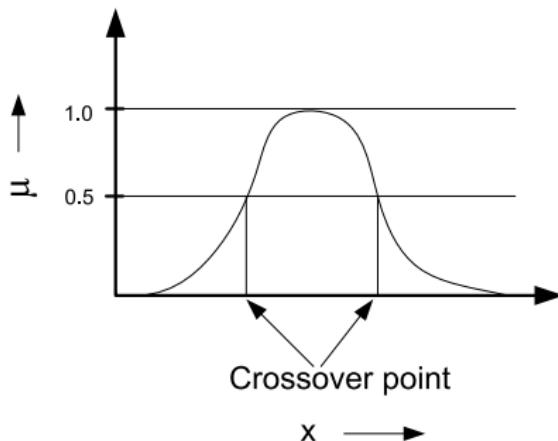
# Fuzzy terminologies: Normality

**Normality** : A fuzzy set  $A$  is normal if its core is non-empty. In other words, we can always find a point  $x \in X$  such that  $\mu_A(x) = 1$ .



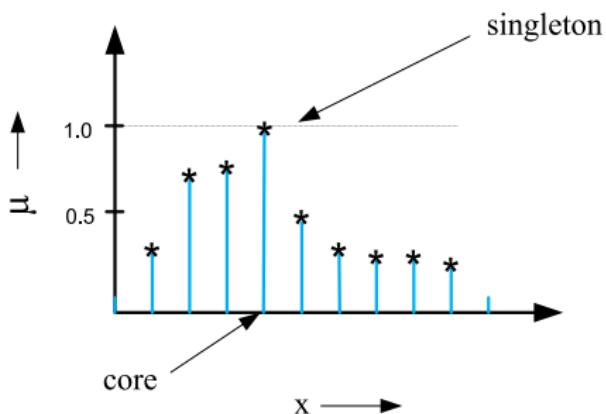
# Fuzzy terminologies: Crossover points

**Crossover point** : A crossover point of a fuzzy set  $A$  is a point  $x \in X$  at which  $\mu_A(x) = 0.5$ . That is  
 $\text{Crossover}(A) = \{x | \mu_A(x) = 0.5\}$ .



# Fuzzy terminologies: Fuzzy Singleton

**Fuzzy Singleton** : A fuzzy set whose support is a single point in  $X$  with  $\mu_A(x) = 1$  is called a fuzzy singleton. That is  $|A| = |\{x \mid \mu_A(x) = 1\}| = 1$ . Following fuzzy set is not a fuzzy singleton.



# Fuzzy terminologies: $\alpha$ -cut and strong $\alpha$ -cut

## $\alpha$ -cut and strong $\alpha$ -cut :

The  $\alpha$ -cut of a fuzzy set  $A$  is a crisp set defined by

$$A_\alpha = \{x \mid \mu_A(x) \geq \alpha\}$$

Strong  $\alpha$ -cut is defined similarly :

$$A_\alpha' = \{x \mid \mu_A(x) > \alpha\}$$

**Note** :  $\text{Support}(A) = A_0'$  and  $\text{Core}(A) = A_1$ .

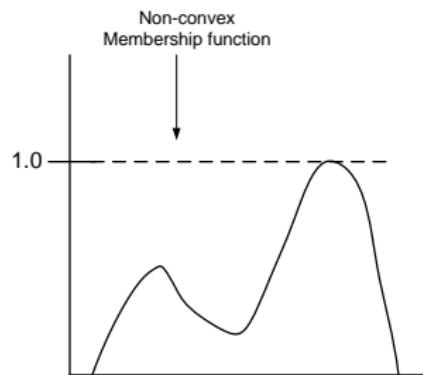
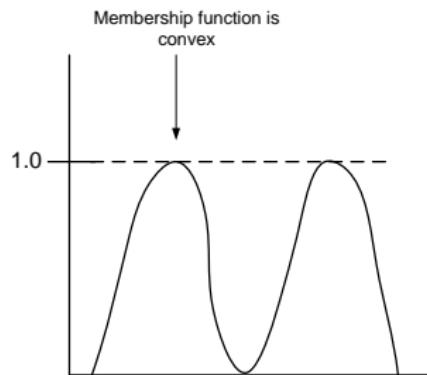
# Fuzzy terminologies: Convexity

**Convexity** : A fuzzy set  $A$  is convex if and only if for any  $x_1$  and  $x_2 \in X$  and any  $\lambda \in [0, 1]$

$$\mu_A(\lambda x_1 + (1 - \lambda)x_2) \geq \min(\mu_A(x_1), \mu_A(x_2))$$

**Note :**

- $A$  is convex if all its  $\alpha$ - level sets are convex.
- Convexity ( $A_\alpha$ )  $\implies A_\alpha$  is composed of a single line segment only.



# Fuzzy terminologies: Bandwidth

## Bandwidth :

For a normal and convex fuzzy set, the bandwidth (or width) is defined as the distance between the two unique crossover points:

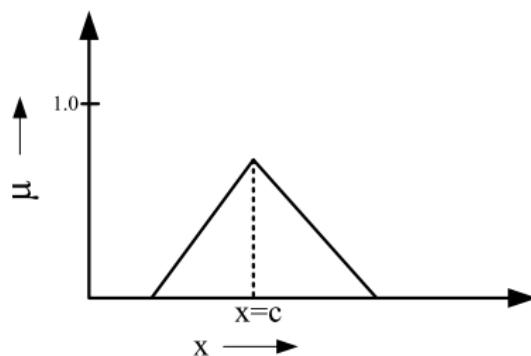
$$\text{Bandwidth}(A) = |x_1 - x_2|$$

where  $\mu_A(x_1) = \mu_A(x_2) = 0.5$

# Fuzzy terminologies: Symmetry

## Symmetry :

A fuzzy set  $A$  is symmetric if its membership function around a certain point  $x = c$ , namely  $\mu_A(x + c) = \mu_A(x - c)$  for all  $x \in X$ .



# Fuzzy terminologies: Open and Closed

A fuzzy set  $A$  is

## Open left

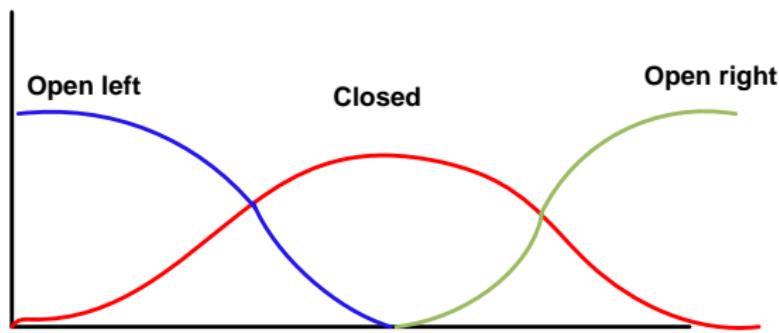
If  $\lim_{x \rightarrow -\infty} \mu_A(x) = 1$  and  $\lim_{x \rightarrow +\infty} \mu_A(x) = 0$

## Open right:

If  $\lim_{x \rightarrow -\infty} \mu_A(x) = 0$  and  $\lim_{x \rightarrow +\infty} \mu_A(x) = 1$

## Closed

If :  $\lim_{x \rightarrow -\infty} \mu_A(x) = \lim_{x \rightarrow +\infty} \mu_A(x) = 0$



# Fuzzy vs. Probability

**Fuzzy** : When we say about certainty of a thing

Example: A patient come to the doctor and he has to diagnose so that medicine can be prescribed.

Doctor prescribed a medicine with certainty 60% that the patient is suffering from flue. So, the disease will be cured with certainty of 60% and uncertainty 40%. Here, in stead of flue, other diseases with some other certainties may be.

**Probability**: When we say about the chance of an event to occur

Example: India will win the T20 tournament with a chance 60% means that out of 100 matches, India own 60 matches.

# Prediction vs. Forecasting

The Fuzzy vs. Probability is analogical to Prediction vs. Forecasting

**Prediction** : When you start guessing about things.

**Forecasting** : When you take the information from the past job and apply it to new job.

**The main difference:**

**Prediction** is based on the **best guess** from experiences.

**Forecasting** is based on **data you have actually recorded and packed from previous job**.

# Fuzzy Membership Functions

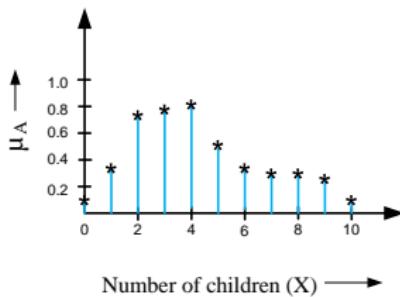
# Fuzzy membership functions

A fuzzy set is completely characterized by its membership function (sometimes abbreviated as  $MF$  and denoted as  $\mu$ ). So, it would be important to learn how a membership function can be expressed (mathematically or otherwise).

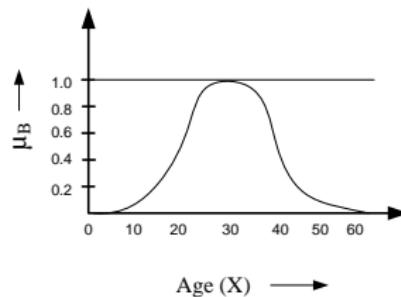
**Note:** A membership function can be on

- (a) a discrete universe of discourse and
- (b) a continuous universe of discourse.

**Example:**



A = Fuzzy set of “Happy family”

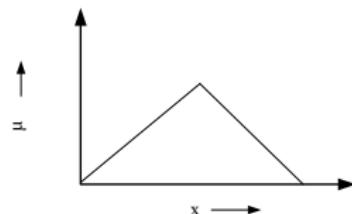


B = “Young age”

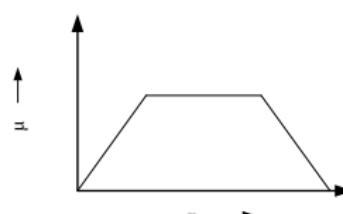
# Fuzzy membership functions

So, membership function on a discrete universe of course is trivial. However, a membership function on a continuous universe of discourse needs a special attention.

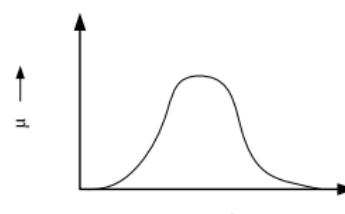
Following figures shows a typical examples of membership functions.



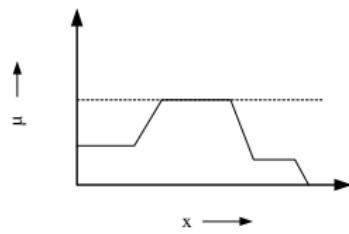
< triangular >



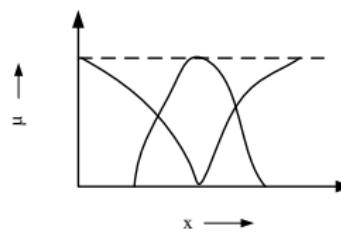
< trapezoidal >



< curve >



< non-uniform >



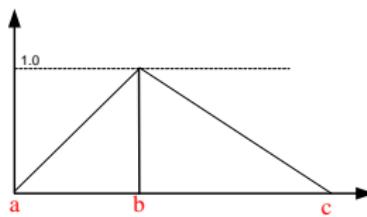
< non-uniform >

# Fuzzy MFs : Formulation and parameterization

In the following, we try to parameterize the different MFs on a continuous universe of discourse.

**Triangular MFs :** A triangular MF is specified by three parameters  $\{a, b, c\}$  and can be formulated as follows.

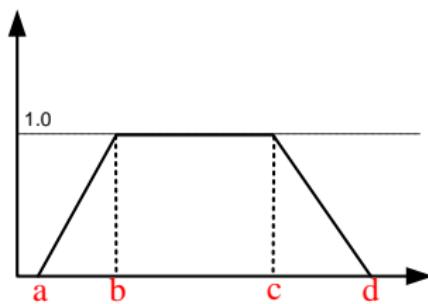
$$triangle(x; a, b, c) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ \frac{c-x}{c-b} & \text{if } b \leq x \leq c \\ 0 & \text{if } c \leq x \end{cases} \quad (1)$$



# Fuzzy MFs: Trapezoidal

A trapezoidal MF is specified by four parameters  $\{a, b, c, d\}$  and can be defined as follows:

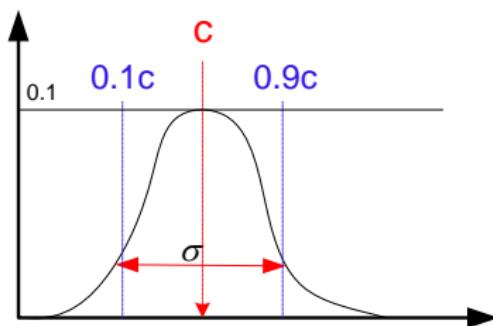
$$trapeziod(x; a, b, c, d) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } b \leq x \leq c \\ \frac{d-x}{d-c} & \text{if } c \leq x \leq d \\ 0 & \text{if } d \leq x \end{cases} \quad (2)$$



# Fuzzy MFs: Gaussian

A Gaussian MF is specified by two parameters  $\{c, \sigma\}$  and can be defined as below:

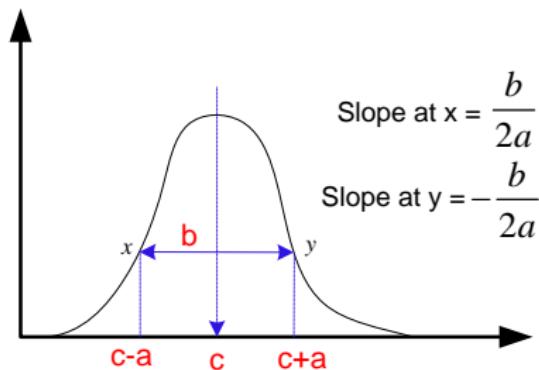
$$\text{gaussian}(x;c,\sigma) = e^{-\frac{1}{2}(\frac{x-c}{\sigma})^2}.$$



# Fuzzy MFs: Generalized bell

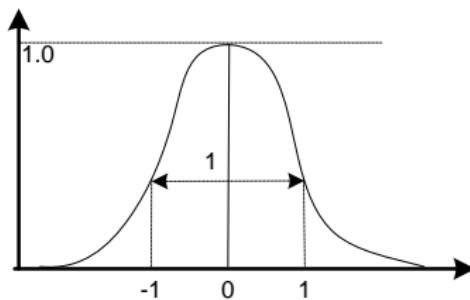
It is also called **Cauchy MF**. A generalized bell MF is specified by three parameters  $\{a, b, c\}$  and is defined as:

$$\text{bell}(x; a, b, c) = \frac{1}{1 + |\frac{x-c}{a}|^{2b}}$$

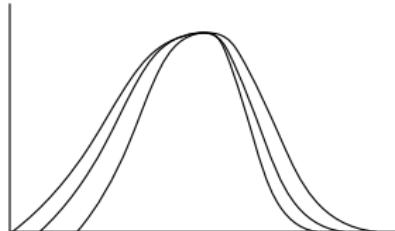


# Example: Generalized bell MFs

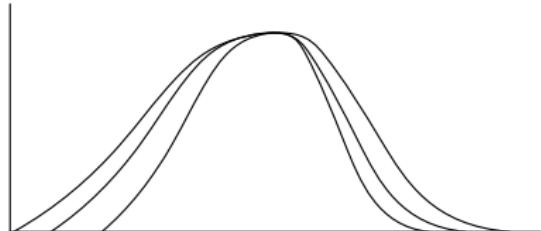
Example:  $\mu(x) = \frac{1}{1+x^2}$  ;  
 $a = b = 1$  and  $c = 0$ ;



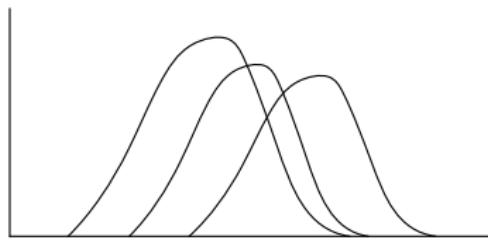
# Generalized bell MFs: Different shapes



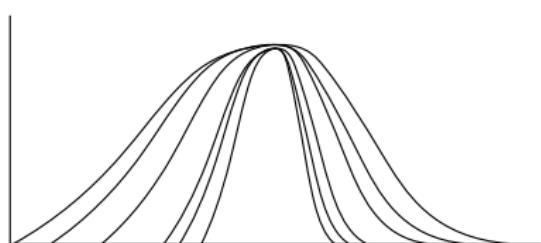
Changing  $a$



Changing  $b$



Changing  $a$

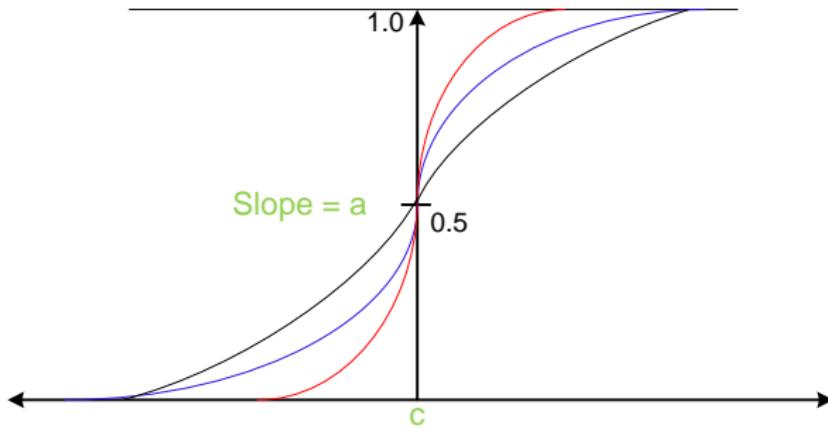


Changing  $a$  and  $b$

# Fuzzy MFs: Sigmoidal MFs

Parameters:  $\{a, c\}$  ; where  $c$  = crossover point and  $a$  = slope at  $c$ ;

$$\text{Sigmoid}(x;a,c) = \frac{1}{1+e^{-[\frac{a}{x-c}]}}$$



# Fuzzy MFs : Example

Example : Consider the following grading system for a course.

Excellent = Marks  $\leq 90$

Very good =  $75 \leq \text{Marks} \leq 90$

Good =  $60 \leq \text{Marks} \leq 75$

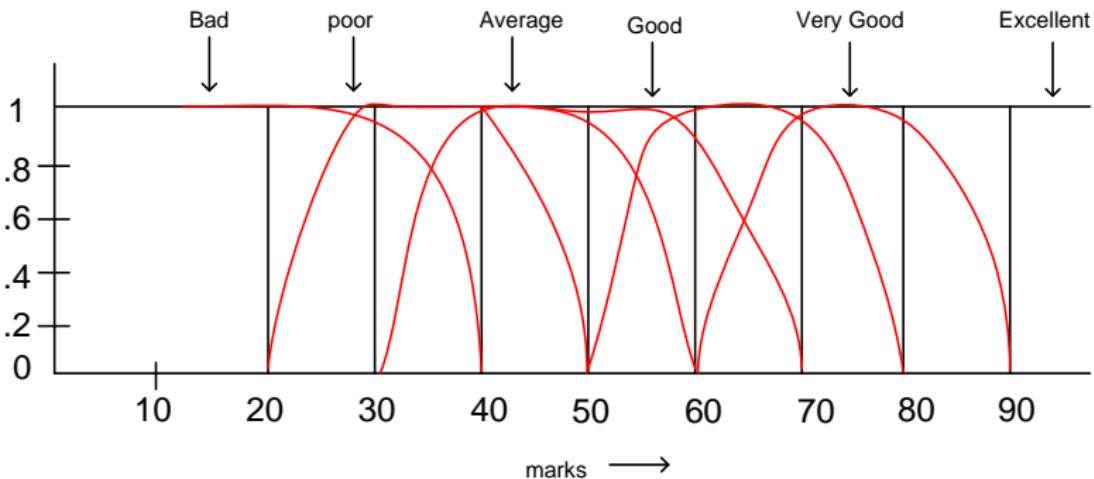
Average =  $50 \leq \text{Marks} \leq 60$

Poor =  $35 \leq \text{Marks} \leq 50$

Bad= Marks  $\leq 35$

# Grading System

A fuzzy implementation will look like the following.



You can decide a standard fuzzy MF for each of the [fuzzy garde](#).

# Operations on Fuzzy Sets

# Basic fuzzy set operations: Union

**Union ( $A \cup B$ ):**

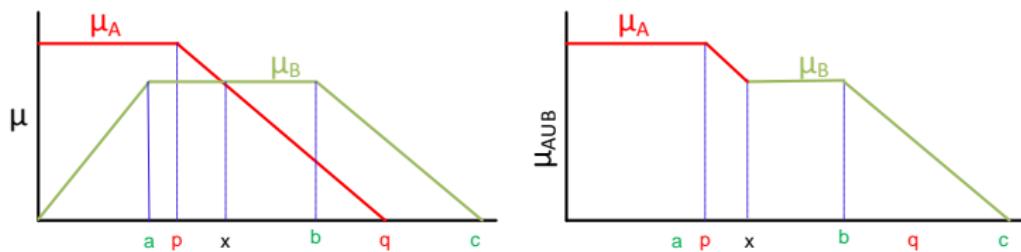
$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}$$

Example:

$A = \{(x_1, 0.5), (x_2, 0.1), (x_3, 0.4)\}$  and

$B = \{(x_1, 0.2), (x_2, 0.3), (x_3, 0.5)\}$ ;

$C = A \cup B = \{(x_1, 0.5), (x_2, 0.3), (x_3, 0.5)\}$



# Basic fuzzy set operations: Intersection

Intersection ( $A \cap B$ ):

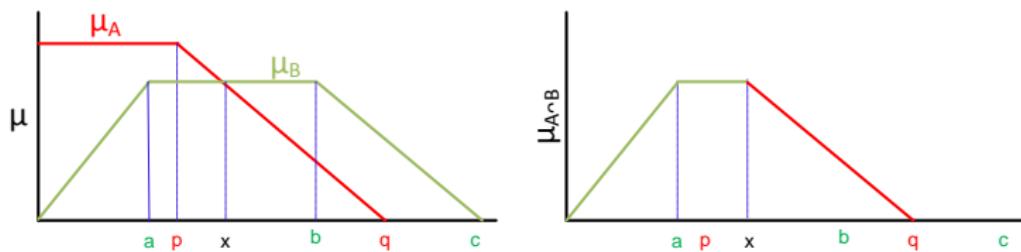
$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$$

Example:

$A = \{(x_1, 0.5), (x_2, 0.1), (x_3, 0.4)\}$  and

$B = \{(x_1, 0.2), (x_2, 0.3), (x_3, 0.5)\}$ ;

$C = A \cap B = \{(x_1, 0.2), (x_2, 0.1), (x_3, 0.4)\}$



# Basic fuzzy set operations: Complement

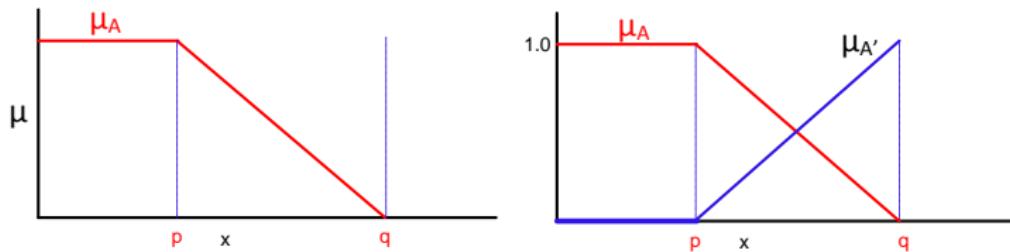
Complement ( $A^C$ ):

$$\mu_{A^C}(x) = 1 - \mu_A(x)$$

Example:

$$A = \{(x_1, 0.5), (x_2, 0.1), (x_3, 0.4)\}$$

$$C = A^C = \{(x_1, 0.5), (x_2, 0.9), (x_3, 0.6)\}$$



# Basic fuzzy set operations: Products

**Algebraic product or Vector product ( $A \bullet B$ ):**

$$\mu_{A \bullet B}(x) = \mu_A(x) \bullet \mu_B(x)$$

**Scalar product ( $\alpha \times A$ ):**

$$\mu_{\alpha A}(x) = \alpha \cdot \mu_A(x)$$

# Basic fuzzy set operations: Sum and Difference

**Sum ( $A + B$ ):**

$$\mu_{A+B}(x) = \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x)$$

**Difference ( $A - B = A \cap B^C$ ):**

$$\mu_{A-B}(x) = \mu_{A \cap B^C}(x)$$

**Disjunctive sum:**  $A \oplus B = (A^C \cap B) \cup (A \cap B^C)$

**Bounded Sum:**  $| A(x) \oplus B(x) |$

$$\mu_{|A(x) \oplus B(x)|} = \min\{1, \mu_A(x) + \mu_B(x)\}$$

**Bounded Difference:**  $| A(x) \ominus B(x) |$

$$\mu_{|A(x) \ominus B(x)|} = \max\{0, \mu_A(x) + \mu_B(x) - 1\}$$

# Basic fuzzy set operations: Equality and Power

**Equality ( $A = B$ ):**

$$\mu_A(x) = \mu_B(x)$$

**Power of a fuzzy set  $A^\alpha$ :**

$$\mu_{A^\alpha}(x) = \{\mu_A(x)\}^\alpha$$

- If  $\alpha < 1$ , then it is called *dilation*
- If  $\alpha > 1$ , then it is called *concentration*

# Basic fuzzy set operations: Cartesian product

Caretsian Product ( $A \times B$ ):

$$\mu_{A \times B}(x, y) = \min\{\mu_A(x), \mu_B(y)\}$$

**Example 3:**

$$A(x) = \{(x_1, 0.2), (x_2, 0.3), (x_3, 0.5), (x_4, 0.6)\}$$

$$B(y) = \{(y_1, 0.8), (y_2, 0.6), (y_3, 0.3)\}$$

$$A \times B = \min\{\mu_A(x), \mu_B(y)\} =$$

	$y_1$	$y_2$	$y_3$
$x_1$	0.2	0.2	0.2
$x_2$	0.3	0.3	0.3
$x_3$	0.5	0.5	0.3
$x_4$	0.6	0.6	0.3

# Properties of fuzzy sets

## Commutativity :

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

## Associativity :

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

## Distributivity :

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

# Properties of fuzzy sets

## Idempotence :

$$A \cup A = A$$

$$A \cap A = \emptyset$$

$$A \cup \emptyset = A$$

$$A \cap \emptyset = \emptyset$$

## Transitivity :

If  $A \subseteq B, B \subseteq C$  then  $A \subseteq C$

## Involution :

$$(A^c)^c = A$$

## De Morgan's law :

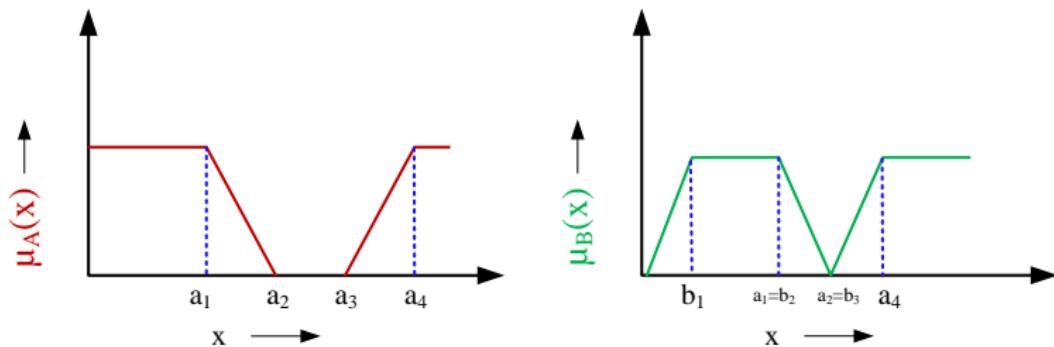
$$(A \cap B)^c = A^c \cup B^c$$

$$(A \cup B)^c = A^c \cap B^c$$

# Few Illustrations on Fuzzy Sets

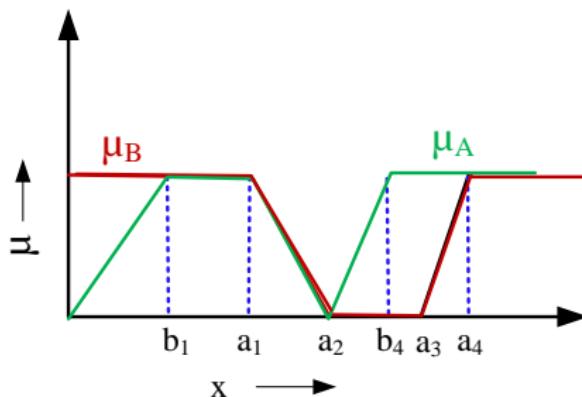
# Example 1: Fuzzy Set Operations

Let A and B are two fuzzy sets defined over a universe of discourse X with membership functions  $\mu_A(x)$  and  $\mu_B(x)$ , respectively. Two MFs  $\mu_A(x)$  and  $\mu_B(x)$  are shown graphically.



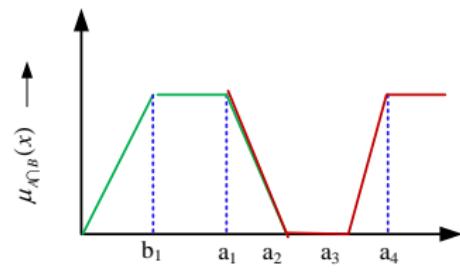
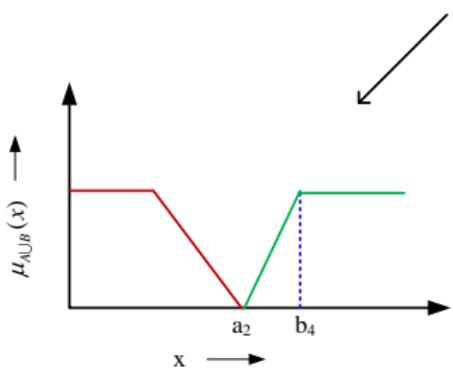
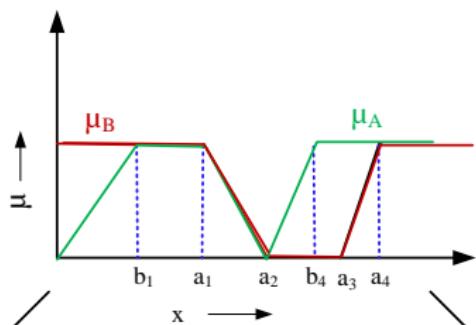
# Example 1: Plotting two sets on the same graph

Let's plot the two membership functions on the same graph



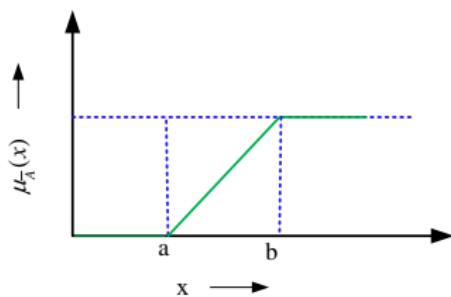
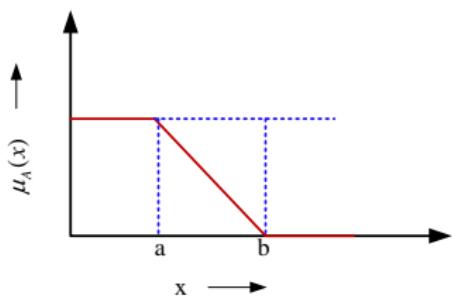
# Example 1: Union and Intersection

The plots of union  $A \cup B$  and intersection  $A \cap B$  are shown in the following.



# Example 1: Intersection

The plots of union  $\mu_{\bar{A}}(x)$  of the fuzzy set  $A$  is shown in the following.



# Fuzzy set operations: Practice

Consider the following two fuzzy sets  $A$  and  $B$  defined over a universe of discourse  $[0,5]$  of real numbers with their membership functions

$$\mu_A(x) = \frac{x}{1+x} \text{ and } \mu_B(x) = 2^{-x}$$

Determine the membership functions of the following and draw them graphically.

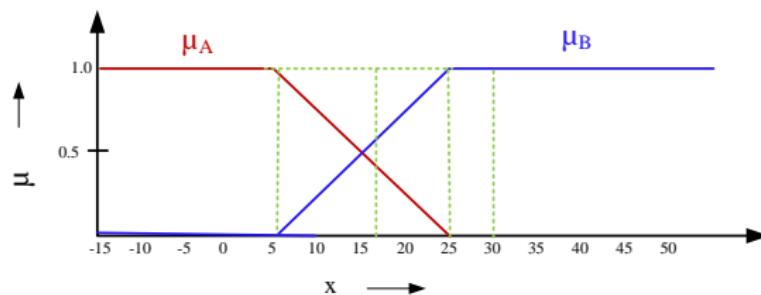
- i.  $\bar{A}$ ,  $\bar{B}$
- ii.  $A \cup B$
- iii.  $A \cap B$
- iv.  $(A \cup B)^c$  [Hint: Use De' Morgan law]

## Example 2: A real-life example

Two fuzzy sets  $A$  and  $B$  with membership functions  $\mu_A(x)$  and  $\mu_B(x)$ , respectively defined as below.

$A = \text{Cold climate}$  with  $\mu_A(x)$  as the MF.

$B = \text{Hot climate}$  with  $\mu_B(x)$  as the M.F.



Here,  $X$  being the universe of discourse representing entire range of temperatures.

## Example 2: A real-life example

What are the fuzzy sets representing the following?

- ① Not cold climate
- ② Not hot climate
- ③ Extreme climate
- ④ Pleasant climate

Note: Note that "Not cold climate"  $\neq$  "Hot climate" and vice-versa.

# Example 2 : A real-life example

Answer would be the following.

## ① Not cold climate

$\overline{A}$  with  $1 - \mu_A(x)$  as the MF.

## ② Not hot climate

$\overline{B}$  with  $1 - \mu_B(x)$  as the MF.

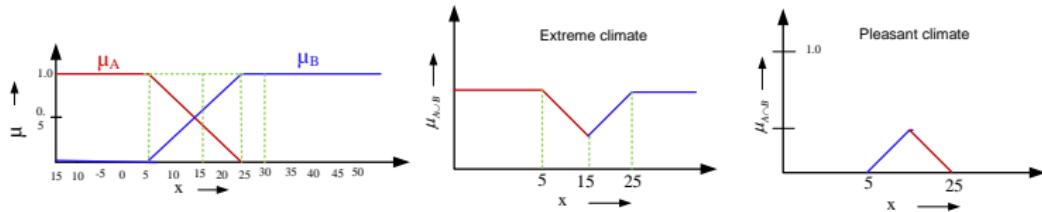
## ③ Extreme climate

$A \cup B$  with  $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$  as the MF.

## ④ Pleasant climate

$A \cap B$  with  $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$  as the MF.

The plot of the MFs of  $A \cup B$  and  $A \cap B$  are shown in the following.



# Few More on Membership Functions

# Generation of MFs

Given a membership function of a fuzzy set representing a **linguistic hedge**, we can derive many more MFs representing several other linguistic hedges using the concept of **Concentration** and **Dilation**.

- **Concentration:**

$$A^k = [\mu_A(x)]^k ; k > 1$$

- **Dilation:**

$$A^k = [\mu_A(x)]^k ; k < 1$$

Example : Age = { Young, Middle-aged, Old }

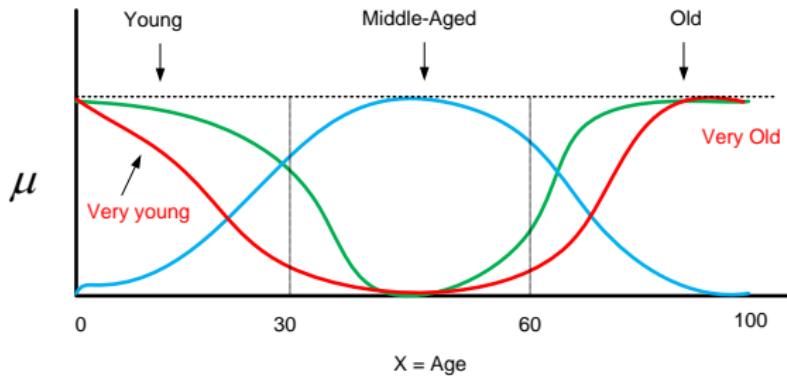
Thus, corresponding to Young, we have : Not young, Very young, Not very young and so on.

Similarly, with Old we can have : old, very old, very very old, extremely old etc.

Thus, **Extremely old** =  $((old)^2)^2$  and so on

Or, **More or less old** =  $A^{0.5} = (old)^{0.5}$

# Linguistic variables and values



$$\mu_{young}(x) = bell(x, 20, 2, 0) = \frac{1}{1 + (\frac{x-20}{2})^4}$$

$$\mu_{old}(x) = bell(x, 30, 3, 100) = \frac{1}{1 + (\frac{x-30}{3})^6}$$

$$\mu_{middle-aged} = bell(x, 30, 60, 50)$$

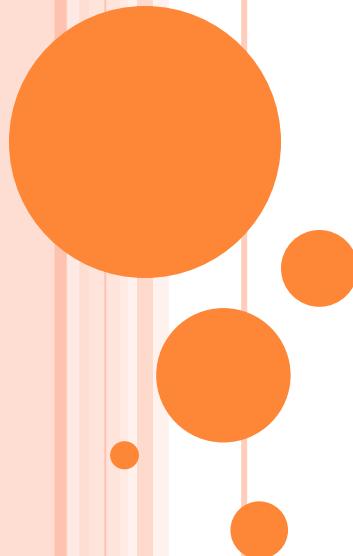
$$\text{Not young} = \overline{\mu_{young}(x)} = 1 - \mu_{young}(x)$$

$$\text{Young but not too young} = \mu_{young}(x) \cap \overline{\mu_{young}(x)}$$

# Any questions??

# FUZZY DECISION TREES (FDT)

DECISION MAKING SUPPORT SYSTEM BASED ON FDT



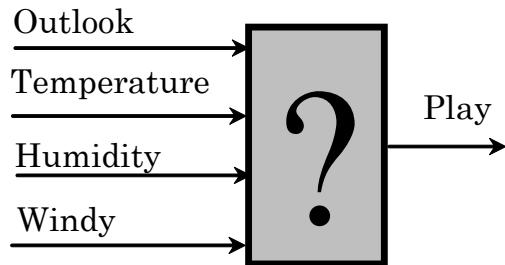
# DECISION SUPPORT SYSTEMS

Decision Support Systems are a specific class of computer-based information systems that support your **decision-making activities**.

A decision support system **analyzes data and provide interactive information support** to professionals during the decision-making process.

Decision making implies selection of the **best decision from a set of possible options**. In some cases, this selection is based on past experience. **Past experience** is used to analyze the situations and the choice made in these situations.

# DECISION MAKING BY ANALYSIS OF PREVIOUS SITUATIONS



Microsoft Access - [Weather-L : Tabulka]

	id	Outlook	Temp	Humidity	Windy	Play
▶	1	sunny	hot	high	false	no
	2	sunny	hot	high	true	no
	3	overcast	hot	high	false	yes
	4	rain	mild	high	false	yes
	5	rain	cool	normal	false	yes
	6	rain	cool	normal	true	no
	7	overcast	cool	normal	true	yes
	8	sunny	mild	high	false	no
	9	sunny	cppl	normal	false	yes
	10	rain	mild	normal	false	yes
	11	sunny	mild	normal	true	yes
	12	overcast	mild	high	true	yes
	13	overcast	hot	normal	false	yes
	14	rain	mild	high	true	no

Our goal is **building a model** for the **recognition** of the new situation:

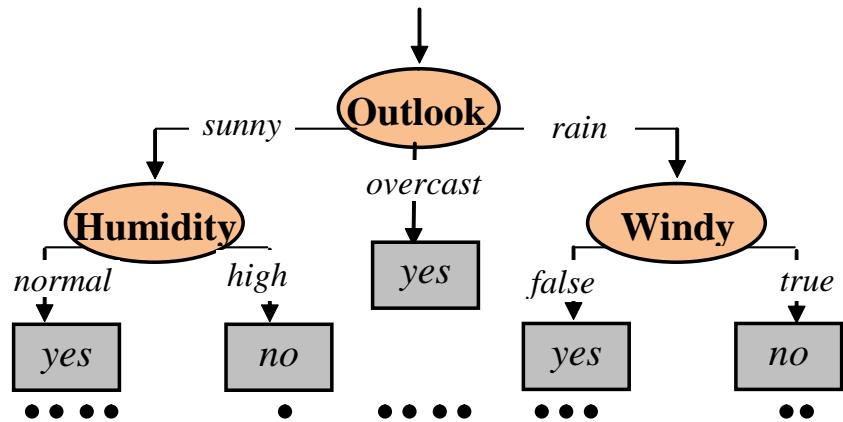
Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	?

# DECISION TREES (1). INTRODUCTION

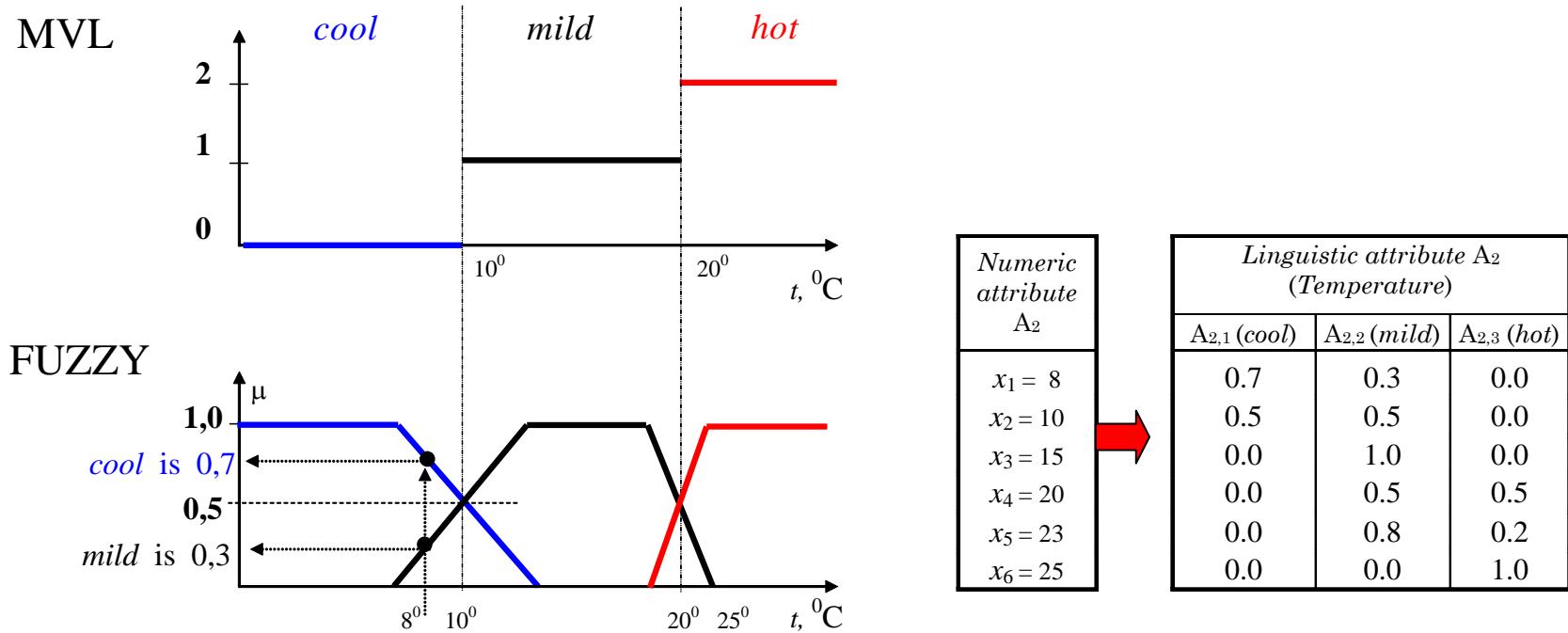
**Decision Tree** is a flow-chart like structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label.

Microsoft Access - [Weather-L : Tabulka]

	id	Outlook	Temp	Humidity	Windy	Play
▶	1	sunny	hot	high	false	no
	2	sunny	hot	high	true	no
	3	overcast	hot	high	false	yes
	4	rain	mild	high	false	yes
	5	rain	cool	normal	false	yes
	6	rain	cool	normal	true	no
	7	overcast	cool	normal	true	yes
	8	sunny	mild	high	false	no
	9	sunny	cppl	normal	false	yes
	10	rain	mild	normal	false	yes
	11	sunny	mild	normal	true	yes
	12	overcast	mild	high	true	yes
	13	overcast	hot	normal	false	yes
	14	rain	mild	high	true	no



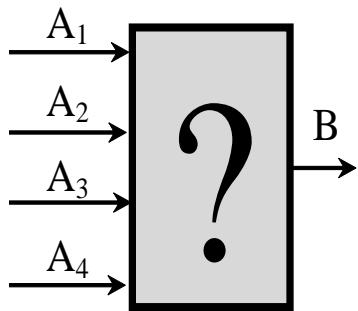
# MVL VS. FUZZY. CLUSTERING



An object  $x$  can belong simultaneously to more than one class and do so to varying degrees called memberships

H.-M. Lee, C.M. Chen, etc, An Efficient Fuzzy Classifier with Feature Selection Based on Fuzzy Entropy, *Journal of IEEE Trans. on Systems, Man and Cybernetics*, Part B, vol.31(3), 2001, pp. 426-432

# FUZZY DATA \*



N	Attribute A <sub>1</sub>			Attribute A <sub>2</sub>			Attribute A <sub>3</sub>		Attribute A <sub>4</sub>		Output B		
	A <sub>11</sub>	A <sub>12</sub>	A <sub>13</sub>	A <sub>21</sub>	A <sub>22</sub>	A <sub>23</sub>	A <sub>31</sub>	A <sub>32</sub>	A <sub>41</sub>	A <sub>42</sub>	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
Cost	Cost (A <sub>1</sub> )=2.5			Cost (A <sub>2</sub> )=2.0			Cost(A <sub>3</sub> )=1.7		Cost(A <sub>4</sub> )=1.8				
1.	0.9	0.1	0.0	1.0	0.0	0.0	0.8	0.2	0.4	0.6	0.0	0.8	0.2
2.	0.8	0.2	0.0	0.6	0.4	0.0	0.0	1.0	0.0	1.0	0.6	0.4	0.0
3.	0.0	0.7	0.3	0.8	0.2	0.0	0.1	0.9	0.2	0.8	0.3	0.6	0.1
4.	0.2	0.7	0.1	0.3	0.7	0.0	0.2	0.8	0.3	0.7	0.9	0.1	0.0
5.	0.0	0.1	0.9	0.7	0.3	0.0	0.5	0.5	0.5	0.5	0.0	0.0	1.0
6.	0.0	0.7	0.3	0.0	0.3	0.7	0.7	0.3	0.4	0.6	0.2	0.0	0.8
7.	0.0	0.3	0.7	0.0	0.0	1.0	0.0	1.0	0.1	0.9	0.0	0.0	1.0
8.	0.0	1.0	0.0	0.0	0.2	0.8	0.2	0.8	0.0	1.0	0.7	0.0	0.3
9.	1.0	0.0	0.0	1.0	0.0	0.0	0.6	0.4	0.7	0.3	0.2	0.8	0.0
10.	0.9	0.1	0.0	0.0	0.3	0.7	0.0	1.0	0.9	0.1	0.0	0.3	0.7
11.	0.7	0.3	0.0	1.0	0.0	0.0	1.0	0.0	0.2	0.8	0.3	0.7	0.0
12.	0.2	0.6	0.2	0.0	1.0	0.0	0.3	0.7	0.3	0.7	0.7	0.2	0.1
13.	0.9	0.1	0.0	0.2	0.8	0.0	0.1	0.9	1.0	0.0	0.0	0.0	1.0
14.	0.0	0.9	0.1	0.0	0.9	0.1	0.1	0.9	0.7	0.3	0.0	0.0	1.0
15.	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.8	0.2	0.0	0.0	1.0
16.	1.0	0.0	0.0	0.5	0.5	0.0	0.0	1.0	0.0	1.0	0.5	0.5	0.0
17.											?	?	?

\*Y. Yuan, M.J. Shaw, Induction of Fuzzy Decision Trees, *Fuzzy Sets and Systems*, 69, 1995, pp.125-139

Measuring the value of each input attribute requires resource costs (money or time):

$\text{Cost (A}_1\text{)}, \text{Cost (A}_2\text{)}, \text{Cost (A}_3\text{)}, \text{Cost (A}_4\text{)} .$

*Our goal* is find a method for transform values of input attributes into the value of output attribute with *minimal resources*:

$\text{sum Cost (A}_i\text{)} \rightarrow \text{minimum}$

# ALGORITHMS ID3 AND C4.5 (BY PROF. ROSS QUINLAN)

We compared the information gain and classical concepts of information theory (information and entropy).

These mathematical expression are similar and common.

Algorithm	Prof. Ross Quinlan	Information theory
ID3	$Gain(A) = \sum_{l=1}^{m_b} -\frac{N_{b_l}}{N} \log_2 \frac{N_{b_l}}{N} - \sum_{j=1}^{m_a} \frac{N_{a_j}}{N} \times \sum_{l=1}^{m_b} -\frac{N_{b_l/a_j}}{N_{a_j}} \log_2 \frac{N_{b_l/a_j}}{N_{a_j}}$	$I(B;A) = H(B) - H(B A)$ (Abs.)
C4.5	$GainRatio(A) = Gain(A) / SplitInfo(A),$ $SplitInfo(A) = \sum_{j=1}^{m_a} -\frac{N_{a_j}}{N} \times \log_2 \frac{N_{a_j}}{N}.$	$s(A_i B) = I(B;A) / H(A)$ (Rel.)

# REVIEW OF INFORMATION ESTIMATION

## Entropy Calculation

$$H = -\sum_{i=1}^{m_i} p_i \times \log_2 p_i$$

Another Entropies:  
Hybrid, Yager, Koufmann, Kosko, ...

$$-\sum_{i=1}^{m_i} \frac{\sum_{j=1}^N \mu_{i,j}}{N} \times \log_2 \frac{\sum_{j=1}^N \mu_{i,j}}{N}$$

$$-\frac{1}{N} \sum_{i=1}^{m_i} \sum_{j=1}^N (\mu_{i,j} \times \log_2 \mu_{i,j} + (1 - \mu_{i,j}) \times \log_2 (1 - \mu_{i,j}))$$

- H.Ichihashi, 1996
- H-M.Lee, 2001
- A.de Luca and S.Termini, 1972
- Y.Yuan and M.Shaw, 1995
- X.Wang etc, 2000

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

### Attribute: Outlook

Values (Outlook) = Sunny, Overcast, Rain

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{\text{Sunny}} \leftarrow [2+, 3-]$$

$$\text{Entropy}(S_{\text{Sunny}}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971$$

$$S_{\text{Overcast}} \leftarrow [4+, 0-]$$

$$\text{Entropy}(S_{\text{Overcast}}) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$

$$S_{\text{Rain}} \leftarrow [3+, 2-]$$

$$\text{Entropy}(S_{\text{Rain}}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

$$\text{Gain}(S, \text{Outlook}) = \text{Entropy}(S) - \sum_{v \in \{\text{Sunny}, \text{Overcast}, \text{Rain}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Outlook})$$

$$= \text{Entropy}(S) - \frac{5}{14} \text{Entropy}(S_{\text{Sunny}}) - \frac{4}{14} \text{Entropy}(S_{\text{Overcast}})$$

$$- \frac{5}{14} \text{Entropy}(S_{\text{Rain}})$$

$$\text{Gain}(S, \text{Outlook}) = 0.94 - \frac{5}{14} 0.971 - \frac{4}{14} 0 - \frac{5}{14} 0.971 = 0.2464$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

### Attribute: Temp

Values (Temp) = Hot, Mild, Cool

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{Hot} \leftarrow [2+, 2-]$$

$$\text{Entropy}(S_{Hot}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1.0$$

$$S_{Mild} \leftarrow [4+, 2-]$$

$$\text{Entropy}(S_{Mild}) = -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0.9183$$

$$S_{Cool} \leftarrow [3+, 1-]$$

$$\text{Entropy}(S_{Cool}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8113$$

$$\text{Gain}(S, \text{Temp}) = \text{Entropy}(S) - \sum_{v \in \{\text{Hot}, \text{Mild}, \text{Cool}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Temp})$$

$$= \text{Entropy}(S) - \frac{4}{14} \text{Entropy}(S_{Hot}) - \frac{6}{14} \text{Entropy}(S_{Mild})$$

$$- \frac{4}{14} \text{Entropy}(S_{Cool})$$

$$\text{Gain}(S, \text{Temp}) = 0.94 - \frac{4}{14} 1.0 - \frac{6}{14} 0.9183 - \frac{4}{14} 0.8113 = 0.289$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

## Attribute: Humidity

Values (Humidity) = High, Normal

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{\text{High}} \leftarrow [3+, 4-]$$

$$\text{Entropy}(S_{\text{High}}) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.9852$$

$$S_{\text{Normal}} \leftarrow [6+, 1-]$$

$$\text{Entropy}(S_{\text{Normal}}) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.5916$$

$$\text{Gain}(S, \text{Humidity}) = \text{Entropy}(S) - \sum_{v \in \{\text{High}, \text{Normal}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Humidity})$$

$$= \text{Entropy}(S) - \frac{7}{14} \text{Entropy}(S_{\text{High}}) - \frac{7}{14} \text{Entropy}(S_{\text{Normal}})$$

$$\text{Gain}(S, \text{Humidity}) = 0.94 - \frac{7}{14} 0.9852 - \frac{7}{14} 0.5916 = 0.1516$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

## Attribute: Wind

Values (Wind) = Strong, Weak

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{Strong} \leftarrow [3+, 3-]$$

$$\text{Entropy}(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [6+, 2-]$$

$$\text{Entropy}(S_{Weak}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.8113$$

$$\text{Gain}(S, \text{Wind}) = \text{Entropy}(S) - \sum_{v \in \{\text{Strong}, \text{Weak}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Wind}) = \text{Entropy}(S) - \frac{6}{14} \text{Entropy}(S_{Strong}) - \frac{8}{14} \text{Entropy}(S_{Weak})$$

$$\text{Gain}(S, \text{Wind}) = 0.94 - \frac{6}{14} \cdot 1.0 - \frac{8}{14} \cdot 0.8113 = 0.0478$$

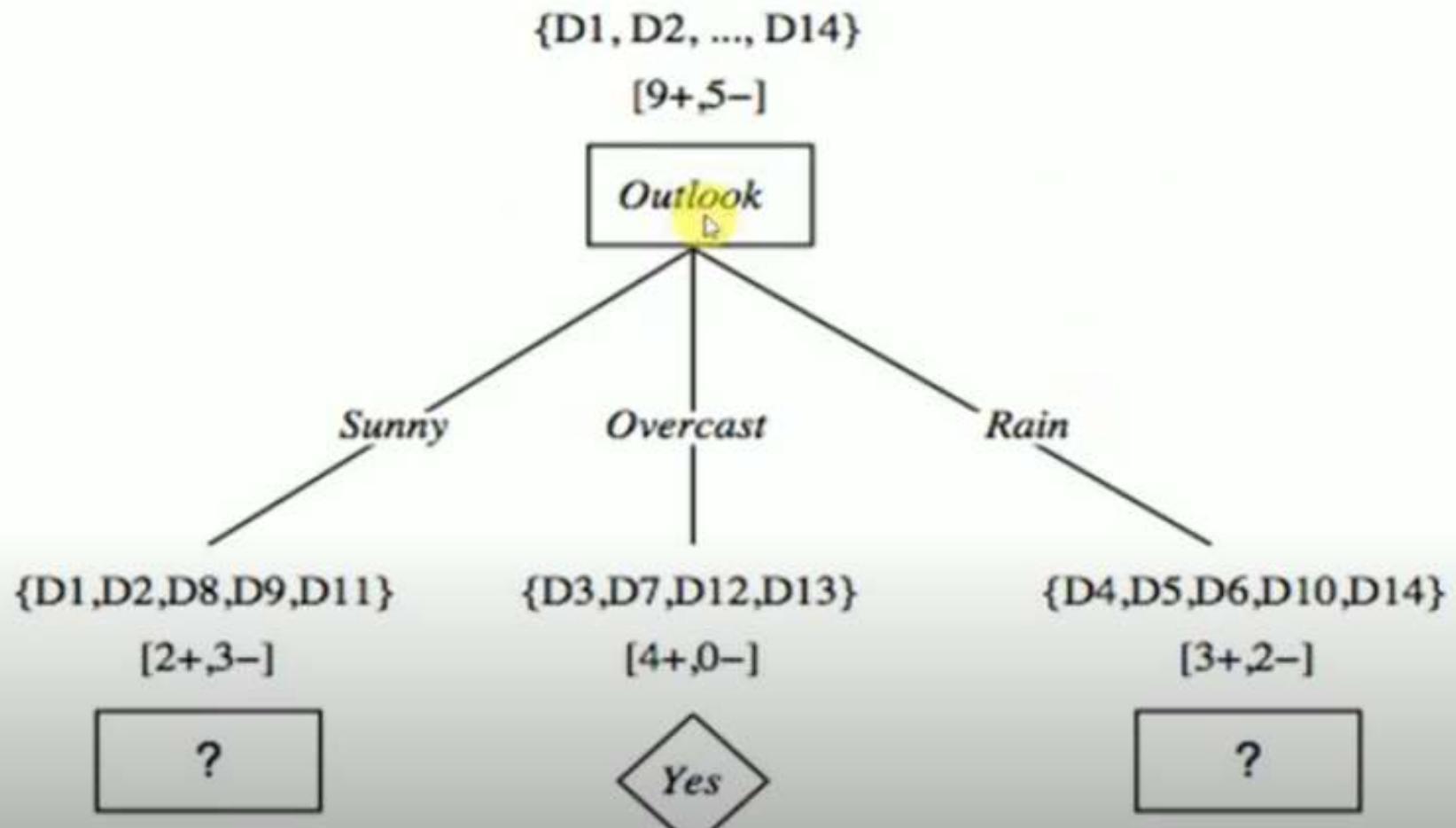
Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Gain(S, Outlook) = 0.2464$$

$$Gain(S, Temp) = 0.0289$$

$$Gain(S, Humidity) = 0.1516$$

$$Gain(S, Wind) = 0.0478$$



Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

### Attribute: Temp

Values (Temp) = Hot, Mild, Cool

$$S_{Sunny} = [2+, 3-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$S_{Hot} \leftarrow [0+, 2-]$$

$$\text{Entropy}(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [1+, 1-]$$

$$\text{Entropy}(S_{Mild}) = 1.0$$

$$S_{Cool} \leftarrow [1+, 0-]$$

$$\text{Entropy}(S_{Cool}) = 0.0$$

$$\text{Gain}(S_{Sunny}, \text{Temp}) = \text{Entropy}(S) - \sum_{v \in \{\text{Hot, Mild, Cool}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S_{Sunny}, \text{Temp})$$

$$= \text{Entropy}(S) - \frac{2}{5} \text{Entropy}(S_{Hot}) - \frac{2}{5} \text{Entropy}(S_{Mild})$$

$$- \frac{1}{5} \text{Entropy}(S_{Cool})$$

$$\text{Gain}(S_{Sunny}, \text{Temp}) = 0.97 - \frac{2}{5} 0.0 - \frac{2}{5} 1 - \frac{1}{5} 0.0 = 0.570$$

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
DI1	Mild	Normal	Strong	Yes

### Attribute: Humidity

Values (Humidity) = High, Normal

$$S_{Sunny} = [2+, 3-]$$

$$\text{Entropy}(S) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$S_{High} \leftarrow [0+, 3-]$$

$$\text{Entropy}(S_{High}) = 0.0$$

$$S_{Normal} \leftarrow [2+, 0-]$$

$$\text{Entropy}(S_{Normal}) = 0.0$$

$$\text{Gain}(S_{Sunny}, \text{Humidity}) = \text{Entropy}(S) - \sum_{v \in \{\text{High, Normal}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S_{Sunny}, \text{Humidity}) = \text{Entropy}(S) - \frac{3}{5} \text{Entropy}(S_{High}) - \frac{2}{5} \text{Entropy}(S_{Normal})$$

$$\text{Gain}(S_{Sunny}, \text{Humidity}) = 0.97 - \frac{3}{5} 0.0 - \frac{2}{5} 0.0 = 0.97$$



Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
DI1	Mild	Normal	Strong	Yes

### Attribute: Wind

Values (Wind) = Strong, Weak

$$S_{Sunny} = [2+, 3-]$$

$$\text{Entropy}(S) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$S_{Strong} \leftarrow [1+, 1-]$$

$$\text{Entropy}(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [1+, 2-]$$

$$\text{Entropy}(S_{Weak}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9183$$

$$Gain(S_{Sunny}, Wind) = \text{Entropy}(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$Gain(S_{Sunny}, Wind) = \text{Entropy}(S) - \frac{2}{5} \text{Entropy}(S_{Strong}) - \frac{3}{5} \text{Entropy}(S_{Weak})$$

$$Gain(S_{Sunny}, Wind) = 0.97 - \frac{2}{5} 1.0 - \frac{3}{5} 0.918 = 0.0192$$

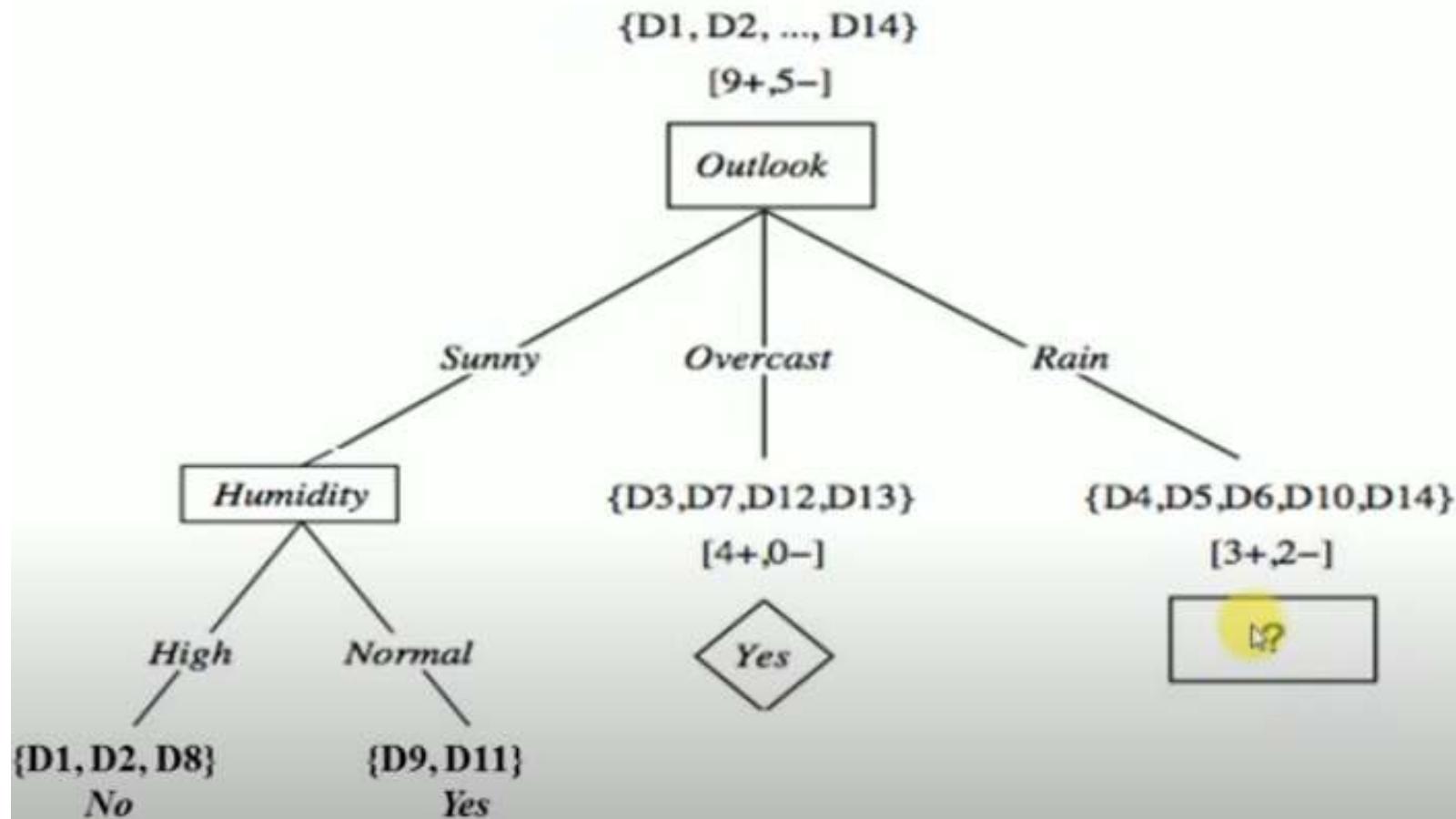
Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

$Gain(S_{sunny}, Temp) = 0.570$



$Gain(S_{sunny}, Humidity) = 0.97$

$Gain(S_{sunny}, Wind) = 0.0192$



Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Day	Temp	Humidity	Wind	Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

Values (Temp) = Hot, Mild, Cool

$$S_{Rain} = [3+, 2-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$S_{Hot} \leftarrow [0+, 0-]$$

$$\text{Entropy}(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [2+, 1-]$$

$$\text{Entropy}(S_{Mild}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183$$

$$S_{Cool} \leftarrow [1+, 1-]$$

$$\text{Entropy}(S_{Cool}) = 1.0$$

$$\text{Gain}(S_{Rain}, Temp) = \text{Entropy}(S) - \sum_{v \in \{\text{Hot, Mild, Cool}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S_{Rain}, Temp)$$

$$= \text{Entropy}(S) - \frac{0}{5} \text{Entropy}(S_{Hot}) - \frac{3}{5} \text{Entropy}(S_{Mild})$$

$$- \frac{2}{5} \text{Entropy}(S_{Cool})$$

$$\text{Gain}(S_{Rain}, Temp) = 0.97 - \frac{0}{5} 0.0 - \frac{3}{5} 0.918 - \frac{2}{5} 1.0 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

### Attribute: Temp

Values (Temp) = Hot, Mild, Cool

$$S_{Rain} = [3+, 2-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$S_{Hot} \leftarrow [0+, 0-]$$

$$\text{Entropy}(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [2+, 1-]$$

$$\text{Entropy}(S_{Mild}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183$$

$$S_{Cool} \leftarrow [1+, 1-]$$

$$\text{Entropy}(S_{Cool}) = 1.0$$

$$\text{Gain}(S_{Rain}, \text{Temp}) = \text{Entropy}(S) - \sum_{v \in \{\text{Hot, Mild, Cool}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S_{Rain}, \text{Temp})$$

$$= \text{Entropy}(S) - \frac{0}{5} \text{Entropy}(S_{Hot}) - \frac{3}{5} \text{Entropy}(S_{Mild})$$

$$- \frac{2}{5} \text{Entropy}(S_{Cool})$$

$$\text{Gain}(S_{Rain}, \text{Temp}) = 0.97 - \frac{0}{0.97} - \frac{3}{0.97} - \frac{2}{0.97} = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

### Attribute: Humidity

Values (Humidity) = High, Normal

$$S_{Rain} = [3+, 2-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$S_{High} \leftarrow [1+, 1-]$$

$$\text{Entropy}(S_{High}) = 1.0$$

$$S_{Normal} \leftarrow [2+, 1-]$$

$$\text{Entropy}(S_{Normal}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183$$

$$\text{Gain}(S_{Rain}, \text{Humidity}) = \text{Entropy}(S) - \sum_{v \in \{\text{High, Normal}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S_{Rain}, \text{Humidity}) = \text{Entropy}(S) - \frac{2}{5} \text{Entropy}(S_{High}) - \frac{3}{5} \text{Entropy}(S_{Normal})$$

$$\text{Gain}(S_{Rain}, \text{Humidity}) = 0.97 - \frac{2}{5} 1.0 - \frac{3}{5} 0.918 = 0.0192$$



Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

### Attribute: Wind

Values (wind) = Strong, Weak

$$S_{Rain} = [3+, 2-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$S_{Strong} \leftarrow [0+, 2-]$$

$$\text{Entropy}(S_{Strong}) = 0.0$$

$$S_{Weak} \leftarrow [3+, 0-]$$

$$\text{Entropy}(S_{Weak}) = 0.0$$

$$\text{Gain}(S_{Rain}, Wind) = \text{Entropy}(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S_{Rain}, Wind) = \text{Entropy}(S) - \frac{2}{5} \text{Entropy}(S_{Strong}) - \frac{3}{5} \text{Entropy}(S_{Weak})$$

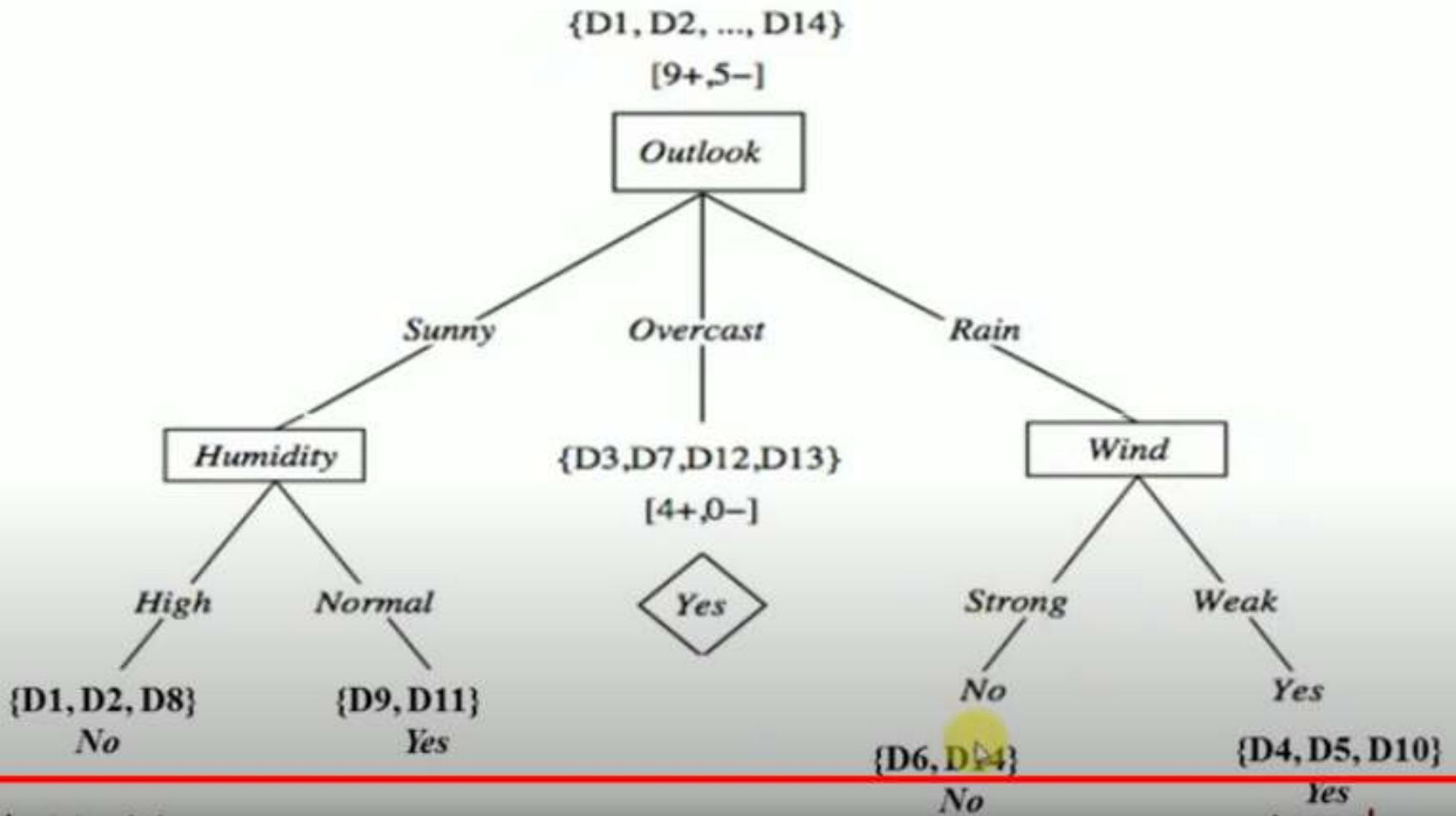
$$\text{Gain}(S_{Rain}, Wind) = 0.97 - \frac{2}{5} 0.0 - \frac{3}{5} 0.0 = 0.97$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

$$Gain(S_{Rain}, Wind) = 0.97$$



---

# Stochastic Search Methods

# Overview

---

- Introduction to stochastic search
- Simulated annealing
- Evolutionary algorithms

# Motivation

---

- Knowledge discovery involves exploration of high-dimensional and multi-modal search spaces
- Finding the global optimum of an objective function with many degrees of freedom and numerous local optima is computationally demanding
- Knowledge discovery systems therefore fundamentally rely on effective and efficient search techniques

# Search techniques

---

- Calculus-based, e.g. gradient methods
- Enumerative, e.g. exhaustive search, dynamic programming
- Stochastic, e.g. Monte Carlo search, tabu search, evolutionary algorithms

# Properties of search techniques

---

- Degree of specialization
- Representation of solutions
- Search operators used to move from one configuration of solutions to the next
- Exploration and exploitation of the search space
- Incorporation of problem-specific knowledge

# Stochastic search

---

- Desired properties of search methods:
  - high probability of finding near-optimal solutions (effectiveness)
  - short processing time (efficiency)
- They are usually conflicting; a compromise is offered by stochastic techniques where certain steps are based on random choice
- Many stochastic search techniques are inspired by processes found in nature

# Inspiration by natural phenomena

---

- Physical and biological processes in nature solve complex search and optimization problems
- Examples:
  - arranging molecules as regular, crystal structures at appropriate temperature reduction
  - creating adaptive, learning organisms through biological evolution

# Nature-inspired methods covered in this presentation

---

- Simulated annealing
- Evolutionary algorithms:
  - evolution strategies
  - genetic algorithms
  - genetic programming

# Simulated annealing: physical background

---

- Annealing: the process of cooling a molten substance; major effect: condensing of matter into a crystalline solid
- Example: hardening of steel by first raising the temperature to the transition to liquid phase and then cooling the steel carefully to allow the molecules to arrange in an ordered lattice pattern

# Simulated annealing: physical background (2)

---

- Annealing can be viewed as an adaptation process optimizing the stability of the final crystalline solid
- The speed of temperature decreasing determines whether or not a state of minimum free energy is reached

# Boltzmann distribution

---

- Probability for the particle system to be in state  $s$  at certain temperature  $T$

$$p_T(s) = \frac{1}{n} \exp\left(\frac{-E(s)}{k \cdot T}\right)$$

$E(s)$  ... free energy

$$n = \sum_{s \in S} \exp\left(\frac{-E(s)}{k \cdot T}\right) \text{ ... normalization}$$

$S$  ... set of all possible system states

$k$  ... Boltzmann constant

# Metropolis algorithm

---

- Stochastic algorithm proposed by Metropolis et al. to simulate the structural evolution of a molten substance for a given temperature
- Assumptions:
  - current system state  $s$
  - temperature  $T$
  - number of equilibration steps  $m$

# Metropolis algorithm (2)

---

- Key step: generate new system state  $s_{\text{new}}$ , evaluate energy difference  $\Delta E = E(s_{\text{new}}) - E(s)$ , and accept the new state with probability depending on  $\Delta E$
- Probability of accepting the new state:

$$p_{\text{accept}} = \begin{cases} 1; & \Delta E < 0 \\ \exp\left(\frac{-\Delta E}{T}\right); & \text{otherwise} \end{cases}$$

# Metropolis algorithm (3)

---

Metropolis( $s, T, m$ );

$i := 0$ ;

while  $i < m$  do

$s_{\text{new}} := \text{Perturb}(s)$ ;

$\Delta E := E(s_{\text{new}}) - E(s)$ ;

if  $(\Delta E < 0)$  or  $(\text{Random}(0,1) < \exp(-\Delta E/T))$

then  $s := s_{\text{new}}$ ;

$i := i + 1$ ;

end\_while;

Return  $s$ ;

# Algorithm Simulated annealing

---

- Starting from a configuration  $s$ , simulate an equilibration process for a fixed temperature  $T$  over  $m$  time steps using  $\text{Metropolis}(s, T, m)$
- Repeat the simulation procedure for decreasing temperatures  $T_{\text{init}} = T_0 > T_1 > \dots > T_{\text{final}}$
- Result: a sequence of annealing configurations with gradually decreasing free energies  
$$E(s_0) \geq \dots \geq E(s_1) \geq \dots \geq E(s_{\text{final}})$$

# Algorithm Simulated annealing (2)

---

```
Simulated_annealing(Tinit, Tfinal, sinit, m, a);  
    T := Tinit;  
    s := sinit;  
    while T > Tfinal do  
        s := Metropolis(s, T, m);  
        T := a · T;  
    end_while;  
    Return s;
```

# Simulated annealing as an optimization process

---

- Solutions to the optimization problem correspond to system states
- System energy corresponds to the objective function
- Searching for a good solution is like finding a system configuration with minimum free energy
- Temperature and equilibration time steps are parameters for controlling the optimization process

# Annealing schedule

---

- A major factor for the optimization process to avoid premature convergence
- Describes how temperature will be decreased and how many iterations will be used during each equilibration phase
- Simple cooling plan:  $T = \alpha \cdot T_i$ , with  $0 < \alpha < 1$ , and fixed number of equilibration steps  $m$

# Algorithm characteristics

---

- At high temperatures almost any new solution is accepted, thus premature convergence towards a specific region can be avoided
- Careful cooling with  $\alpha = 0.8 \dots 0.99$  will lead to asymptotic drift towards  $T_{\text{final}}$
- On its search for optimal solution, the algorithm is capable of escaping from local optima

# Applications and extensions

---

- Initial success in combinatorial optimization, e.g. wire routing and component placement in VLSI design, TSP
- Afterwards adopted as a general-purpose optimization technique and applied in a wide variety of domains
- Variants of the basic algorithm: threshold accepting, parallel simulated annealing, etc., and hybrids, e.g. thermodynamical genetic algorithm

# Evolutionary algorithms (EAs)

---

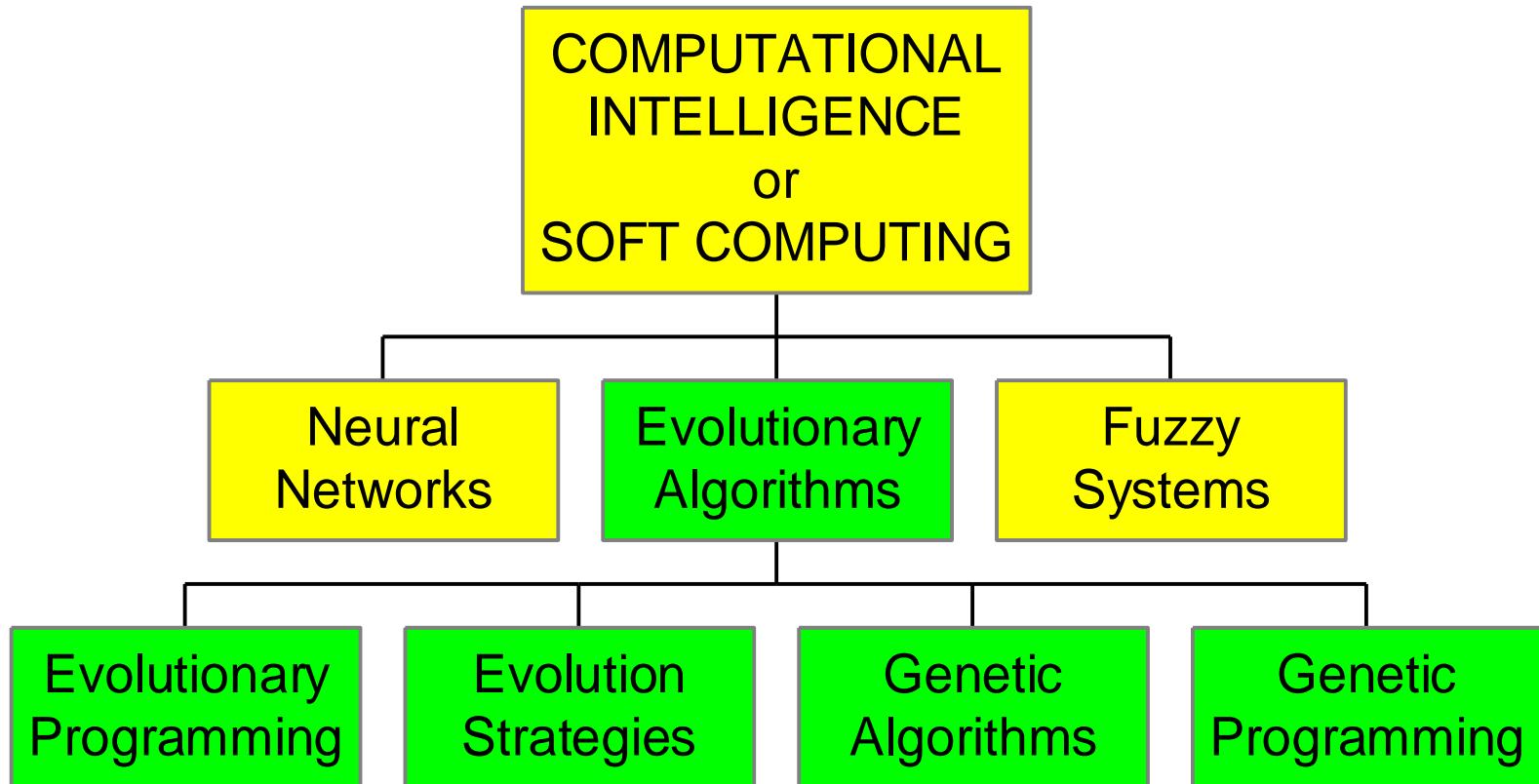
- Simplified models of biological evolution, implementing the principles of Darwinian theory of natural selection (“survival of the fittest”) and genetics
- Stochastic search and optimization algorithms, successful in practice
- Key idea: computer simulated evolution as a problem-solving technique

# Analogy used

---

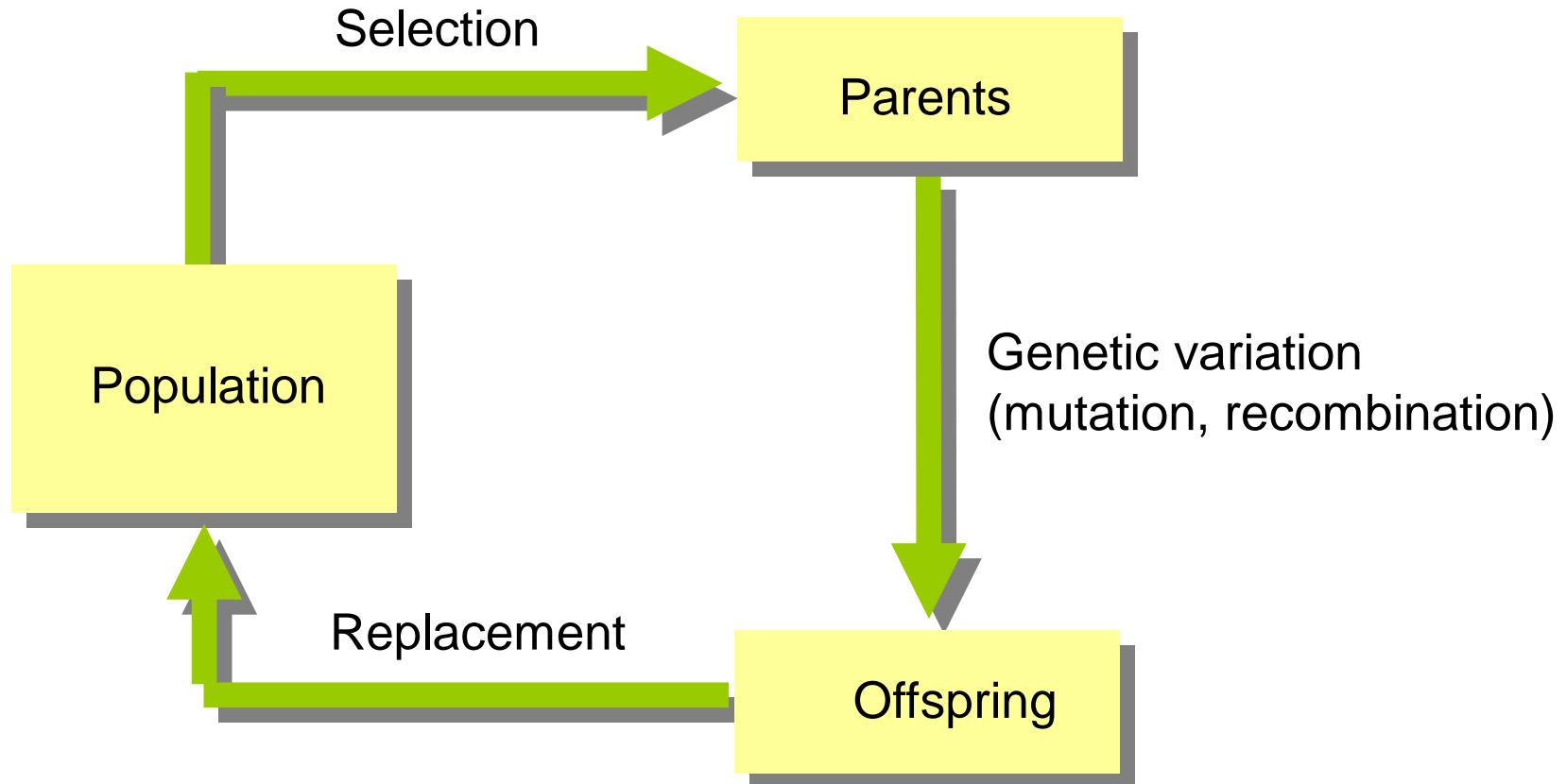
Biological evolution	Computer problem solving
Individual	Solution to a problem
Chromosome	Encoding of a solution
Population	Set of solutions
Crossover, mutation	Search operators
Natural selection	Reuse of good solutions
Fitness	Quality of a solution
Environment	Problem to be solved

# Evolutionary algorithms and soft computing



Source: *EvoNet Flying Circus*

# Evolutionary cycle



Source: *EvoNet Flying Circus*

# Generic Evolutionary algorithm

---

Evolutionary\_algorithm( $t_{\max}$ );

$t := 0;$

Create initial population of individuals;

Evaluate individuals;

result := best\_individual;

while  $t < t_{\max}$  do

$t := t + 1;$

Select better solutions to form new population;

Create their offspring by means of genetic variation;

Evaluate new individuals;

if better solution found then result := best\_individual;

end\_while;

Return result;

# Differences among variants of EAs

---

- Original field of application
- Data structures used to represent solutions
- Realization of selection and variation operators
- Termination criterion

# Evolution strategies (ES)

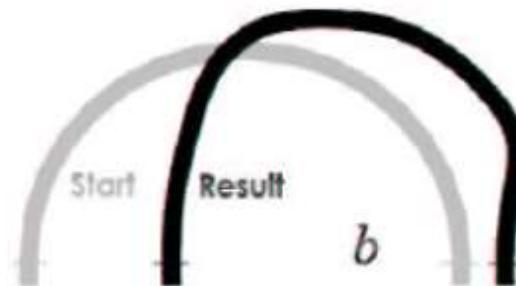
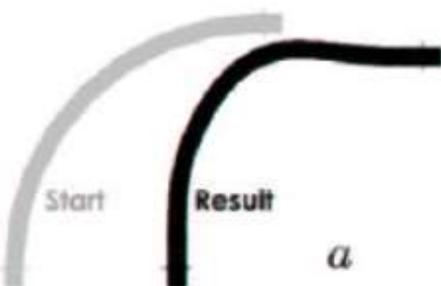
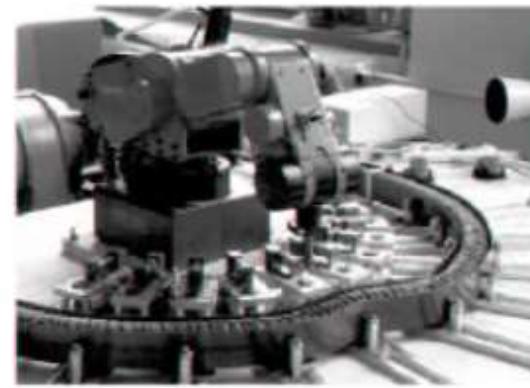
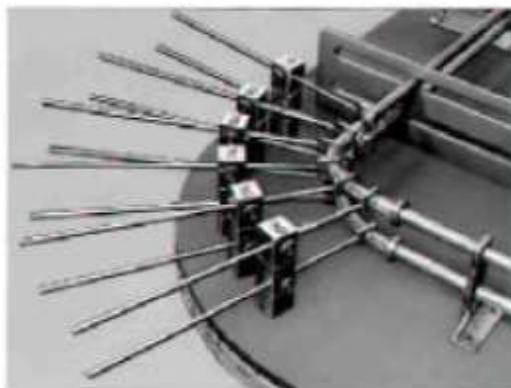
---

- Developed in 1960s and 70s by Ingo Rechenberg and Hans-Paul Schwefel at the Technical University of Berlin
- Originally used as a technique for solving complex optimization problems in engineering design
- Preferred data structures: vectors of real numbers
- Specialty: self-adaptation

# Evolutionary experimentation

---

Pipe-bending experiments (Rechenberg, 1965)



# Algorithm details

---

- Encoding object and strategy parameters:  
 $\mathbf{g} = (\mathbf{p}, \mathbf{s}) = (p_1, p_2, \dots, p_n), (s_1, s_2, \dots, s_n))$   
where  $p_i$  represent problem variables and  $s_i$  mutation variances to be applied to  $p_i$
- Mutation is the major operator for chromosome variation:  
 $\mathbf{g}_{\text{mut}} = (\mathbf{p}_{\text{mut}}, \mathbf{s}_{\text{mut}}) = (\mathbf{p} + N_0(\mathbf{s}), \alpha(\mathbf{s}))$   
 $\mathbf{p}_{\text{mut}} = (p_1 + N_0(s_1), \dots, p_n + N_0(s_n))$   
 $\mathbf{s}_{\text{mut}} = (\alpha(s_1), \dots, \alpha(s_n))$

# Algorithm details (2)

---

- 1/5<sup>th</sup> success rule: Increase mutation strength, if more than 1/5 of offspring are successful, otherwise decrease
- Recombination operators range from swapping respective components between two vectors to component-wise calculation of means

# Algorithm details (3)

---

- Selection schemes
  - $(\mu + \lambda)$ -ES:  $\mu$  parents produce  $\lambda$  offspring,  $\mu$  best out of  $\mu + \lambda$  individuals survive
  - $(\mu, \lambda)$ -ES:  $\mu$  parents produce  $\lambda$  offspring,  $\mu$  best offspring survive
- Originally:  $(1+1)$ -ES
- Advanced techniques: meta-evolution strategies, covariance matrix adaptation ES (CMA-ES)

# Genetic algorithms (GAs)

---

- Developed in 1970s by John Holland at the University of Michigan and popularized as a universal optimization algorithm
- Most remarkable difference between GAs and ES: GAs use string-based, usually binary parameter encoding, resembling discrete nucleotide coding on cellular chromosomes
- Mutation: flipping bits with certain probability
- Recombination performed by crossover

# Crossover operator

---

- Models the breaking of two chromosomes and subsequent crosswise restituation observed on natural genomes during sexual reproduction
- Exchanges information among individuals
- Example: simple (single-point) crossover

Parents	Offspring
1 0 0 1 0 0   1 0 1 0 1 0 1 1	1 0 0 1 0 0 1 1 1 1 0 1 0 0
0 0 1 1 0 1   1 1 1 1 0 1 0 0	0 0 1 1 0 1 1 0 1 0 1 0 1 1

# Selection

---

- Models the principle of “survival of the fittest”
- Traditional approach: fitness proportionate selection performing probabilistic multiplication of individuals with respect to their fitness values
- Implementation: roulette wheel

# Selection (2)

---

- In the population of  $n$  individuals, with the sum of their fitness values  $\Sigma f$  and average fitness  $f_{\text{avg}}$ , the expected number of copies of  $i$ -th individual with fitness  $f_i$  equals to

$$\frac{n \cdot f_i}{\sum f} = \frac{f_i}{f_{\text{avg}}}$$

- Alternative selection schemes: rank-based selection, elitist selection, tournament selection, etc.

# Algorithm extensions

---

- Encoding of solutions: real vectors, permutations, arrays, ...
- Crossover variants: multiple-point crossover, uniform crossover, arithmetic crossover, tailored crossover operators for permutation problems, etc.
- Advanced approaches: meta-GA, parallel GAs, GAs with subjective evaluation of solutions, multi-objective GAs

# Genetic programming (GP)

---

- An extension of genetic algorithms aimed at evolving computer programs using the simulated evolution
- Proposed by John Koza from MIT in 1990s
- Computer programs represented by tree-like symbolic expressions, consisting of functions and terminals
- Crossover: exchange of subtrees between two parent trees

# Genetic programming (2)

---

- Mutation: replacement of a randomly selected subtree with a new, randomly created tree
- Fitness evaluation: program performance in solving the given problem
- GP is a major step towards automatic computer programming, nowadays capable of producing human-competitive solutions in variety of application domains

# Genetic programming (3)

---

- Applications: symbolic regression, process and robotics control, electronic circuit design, signal processing, game playing, evolution of art images and music, etc.
- Main drawback: computational complexity

# Advantages of EAs

---

- Robust and universally applicable
- Besides the solution evaluation, no additional information on solutions and search space properties is required
- As population methods they produce alternative solutions
- Enable incorporation of other techniques (hybridization) and can be parallelized

# Disadvantages of EAs

---

- Suboptimal methodology
- Require tuning of several algorithm parameters
- Computationally expensive

# Conclusion

---

- Stochastic algorithms are becoming increasingly popular in solving complex search and optimization problems in various application domains, including machine learning and data analysis
- A certain degree of randomness, as involved in stochastic algorithms, may help tremendously in improving the ability of a search procedure to discover near-optimal solutions

# Conclusion (2)

---

- Many stochastic methods are inspired by natural phenomena, either by physical or biological processes
- Simulated annealing and evolutionary algorithms discussed in this presentation are two such examples

# Further reading

---

- Corne, D., Dorigo, M. and Glover F. (eds.) (1999): New Ideas in Optimization, McGraw Hill, London
- Eiben, A. E. and Smith, J. E. (2003): Introduction to Evolutionary Computing, Springer, Berlin
- Freitas, A. A. (2002): Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer, Berlin

# Further reading (2)

---

- Jacob, C. (2003): Stochastic Search Methods.  
In: Berthold, M. and Hand, D. J. (eds.)  
Intelligent Data Analysis, Springer, Berlin
- Reeves, C. R. (ed.) (1995): Modern Heuristic  
Techniques for Combinatorial Problems,  
McGraw Hill, London