Decision Trees, Regression, Artificial Neural Networks, Cluster Analysis, Association Rule Mining

Mahesh G. Huddar

Asst. Professor

Dept. of CSE, HIT, Nidasoshi

Decision Trees

- Decision trees are a simple way to guide one’s path to a decision.
- The decision may be a simple binary one, whether to approve a loan or not. Or it may be a complex multivalued decision, as to what may be the diagnosis for a particular sickness.
- Decision trees are hierarchically branched structures that help one come to a decision based on asking certain questions in a particular sequence.
- Decision trees are one of the most widely used techniques for classification.
- A good decision tree should be short and ask only a few meaningful questions.
- Decision trees can generate knowledge from a few test instances that can then be applied to a broad population.
- Decision trees are used mostly to answer relatively simple binary decisions.

Caselet: Predicting Heart Attacks Using Decision Trees

- *study was done at UC San Diego concerning heart disease patient data. The patients were diagnosed with a heart attack from chest pain, diagnosed by EKG, high enzyme levels in their heart muscles, and so on. The objective was to predict which of these patients was at risk of dying from a second heart attack within the next 30 days. The prediction would determine the treatment plan, such as whether to keep the patient in intensive care or not. For each patient, more than 100 variables were collected, including demographics, medical history, and lab data. Using that data and the CART algorithm, a decision tree was constructed.*
- *The decision tree showed that if blood pressure (BP) was low (≤90), the chance of another heart attack was very high (70 percent).*
- *If the patient's BP was OK, the next question to ask was the patient's age. If the age was low (≤62), then the patient's survival was almost guaranteed (98 percent).*
- *If the age was higher, then the next question to ask was about sinus problems.*
- *If their sinus was OK, the chances of survival were 89 percent. Otherwise, the chance of survival dropped to 50 percent. This decision tree predicts 86.5 percent of the cases correctly.*

Decision Tree Problem

- Imagine a conversation between a doctor and a patient. The doctor asks questions to determine the cause of the ailment. The doctor would continue to ask questions, till he or she is able to arrive at a reasonable decision. If nothing seems plausible, he or she might recommend some tests to generate more data and options.
- This is how experts in any field solve problems. They use decision trees or decision rules. For every question they ask, the potential answers create separate branches for further questioning. For each branch, the expert would know how to proceed ahead. The process continues until the end of the tree is reached, which means a leaf node is reached.
- Human experts learn from past experiences or data points. Similarly, a machine can be trained to learn from the past data points and extract some knowledge or rules from it.

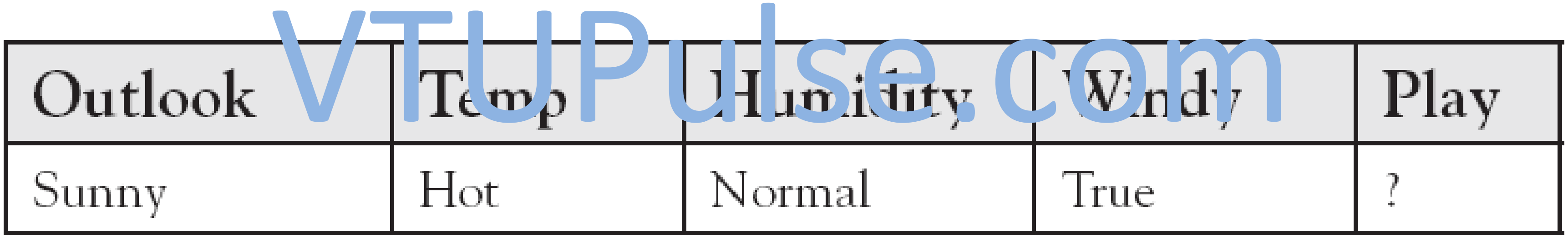
Decision Tree Problem

A decision tree would have a predictive accuracy based on how often it makes correct decisions.

1. The more training data is provided, the more accurate its knowledge extraction will be, and thus, it will make more accurate decisions.
2. The more variables the tree can choose from, the tree will come out better with higher accuracy.
3. In addition, a good decision tree should also be frugal so that it takes the least number of questions, and thus, the least amount of effort, to get to the right decision.

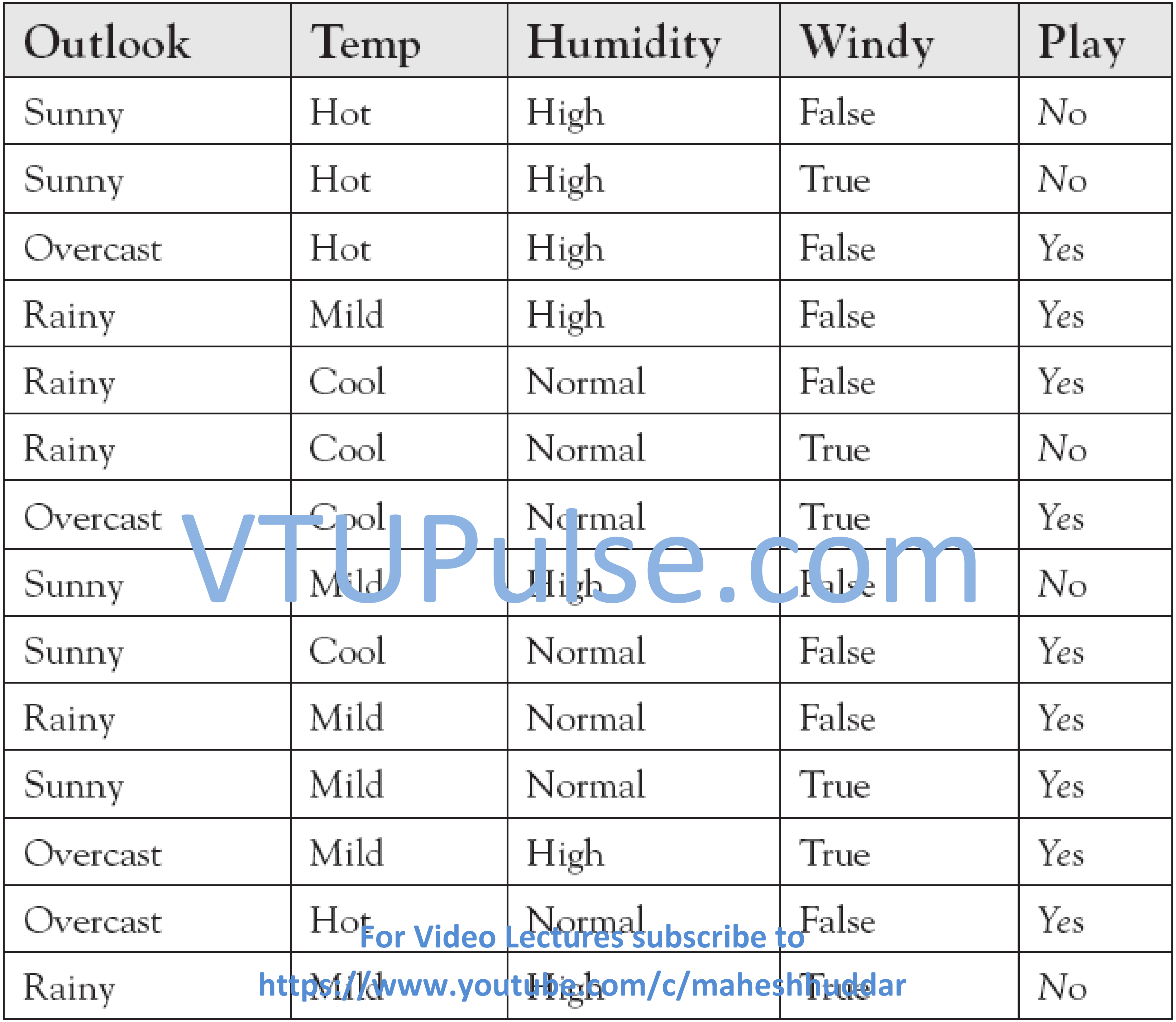
Decision Tree Exercise

- Here is an exercise to create a decision tree that helps make decisions about approving the play of an outdoor game. The objective is to predict the *play* decision given the atmospheric conditions out there. The decision is: Should the game be allowed or not? Here is the decision problem.



Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	Normal	True	?

- To answer that question, one should look at past experience and see what decision was made in a similar instance, if such an instance exists. One could look up the database of past decisions to find the answer and try to come to an answer.



Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Decision Tree Exercise

If there were a row for *sunny/hot/normal/windy* condition in the data table, it would match the current problem, and the decision from that situation could be used to answer the problem today. However, there is no such past instance in this case. There are three disadvantages of looking up the data table:

1. As mentioned earlier, how to decide if there is not a row that corresponds to the exact situation today? If there is no matching instance available in the database, the past experience cannot guide the decision.
2. Searching through the entire past database may be time consuming, depending on the number of variables and the organization of the database.
3. What if the data values are not available for all the variables? In this instance, if the data for humidity variable was not available, looking up the past data would not help.

Decision Tree Construction

Determining root node of the tree

In this example, there are four choices of questions based on the four variables: what is the outlook, what is the temperature, what is the humidity, and what is the wind speed?

1. A criterion should be used by which one of these questions gives the most insight about the situation.
2. The criterion of frugality is a good one, that is, which question will provide us the shortest ultimate tree?
3. Another way to look at this is that if one is allowed to ask only one question, which one would one ask?
4. **The most important question should be the one that, by itself, helps make the most correct decisions with the fewest errors.**

Decision Tree Construction

- Start with any variable, in this case outlook. It can take three values: sunny, overcast, and rainy.
- Start with the sunny value of outlook. There are five instances where the outlook is sunny. In two of the five instances, the *play* decision was *yes*, and in the other three, the decision was *no*. Thus, if the decision rule was that outlook: sunny → no, then three out of five decisions would be correct, while two out of five such decisions would be incorrect. There are two errors out of five. This can be recorded in Row 1.

Attribute	Rules	Error	Total Error
Outlook	Sunny → No	2/5	

Decision Tree Construction

- There are four instances where the outlook is overcast. In all four out of four instances, the *play* decision was *yes*. Thus, if the decision rule was that outlook: overcast → yes, then four out of four decisions would be correct, while none of the decisions would be incorrect. There are zero errors out of four. This can be recorded in the next row.

Attribute	Rules	Error	Total Error
Outlook	Sunny → No	2/5	
	Overcast → Yes	0/4	

Decision Tree Construction

- There are five instances where the outlook is rainy. In three out of five instances, the *play* decision was *yes*, and in the other three, the decision was *no*. Thus, if the decision rule was that outlook: rainy → yes, then three out of five decisions would be correct, while two out of five decisions would be incorrect. There will be two errors out of five. This can be recorded in next row.

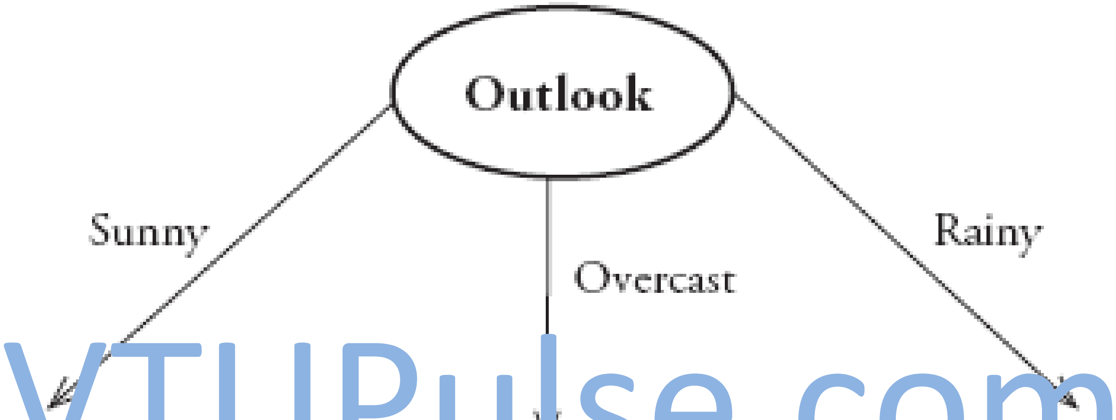
Attribute	Rules	Error	Total Error
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	

Decision Tree Construction

- A similar analysis can be done for the other three variables. At the end of that analytical exercise, the following error table will be constructed.

Attribute	Rules	Error	Total Error
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	hot → No	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	high → No	3/7	4/14
	Normal → Yes	1/7	
Windy	False → Yes	2/8	5/14
	True → No	3/6	

Decision Tree Construction

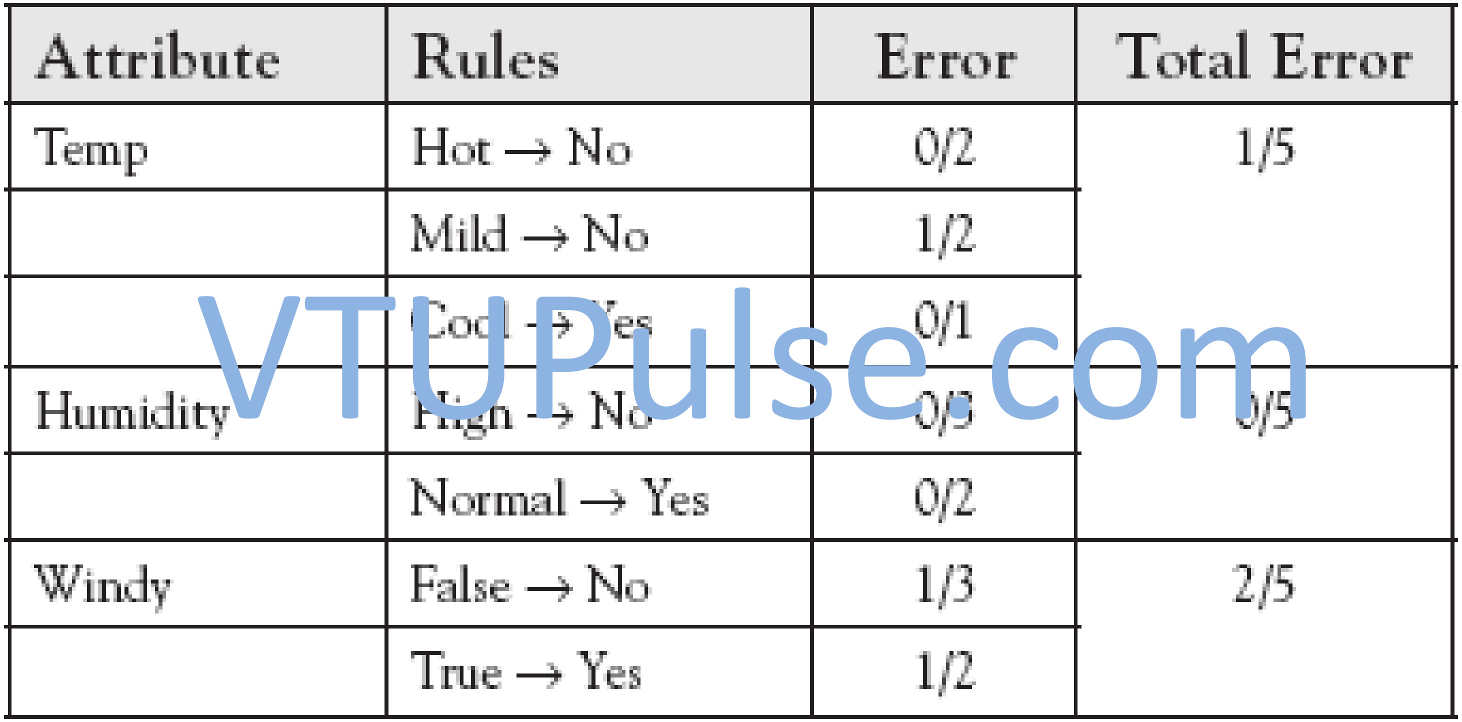


Temp	Humidity	Windy	Play
Hot	High	False	No
Hot	High	True	No
Mild	High	False	No
Cool	Normal	False	Yes
Mild	Normal	True	Yes

YES

Temp	Humidity	Windy	Play
Mild	High	False	Yes
Cool	Normal	False	Yes
Cool	Normal	True	No
Mild	Normal	False	Yes
Mild	High	True	No

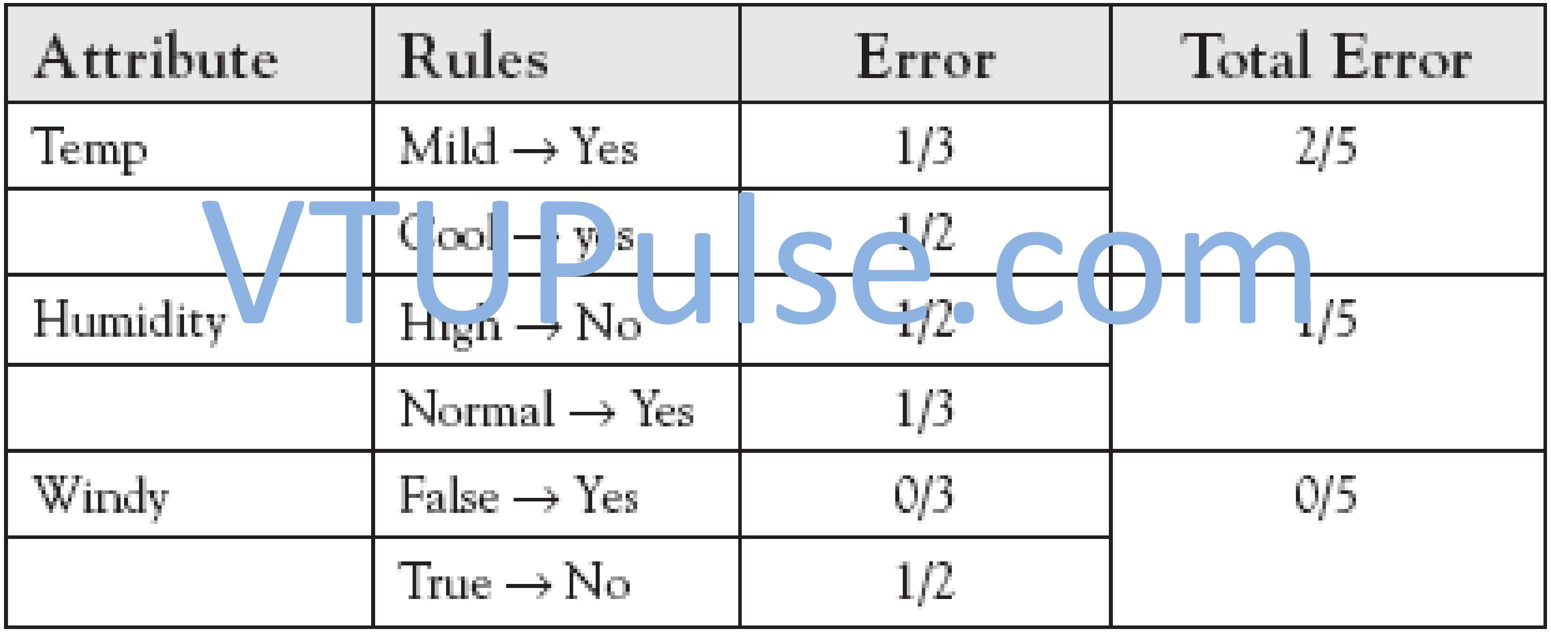
Decision Tree Construction



Attribute	Rules	Error	Total Error
Temp	Hot → No	0/2	1/5
	Mild → No	1/2	
	Cool → Yes	0/1	
Humidity	High → No	0/3	0/5
	Normal → Yes	0/2	
Windy	False → No	1/3	2/5
	True → Yes	1/2	

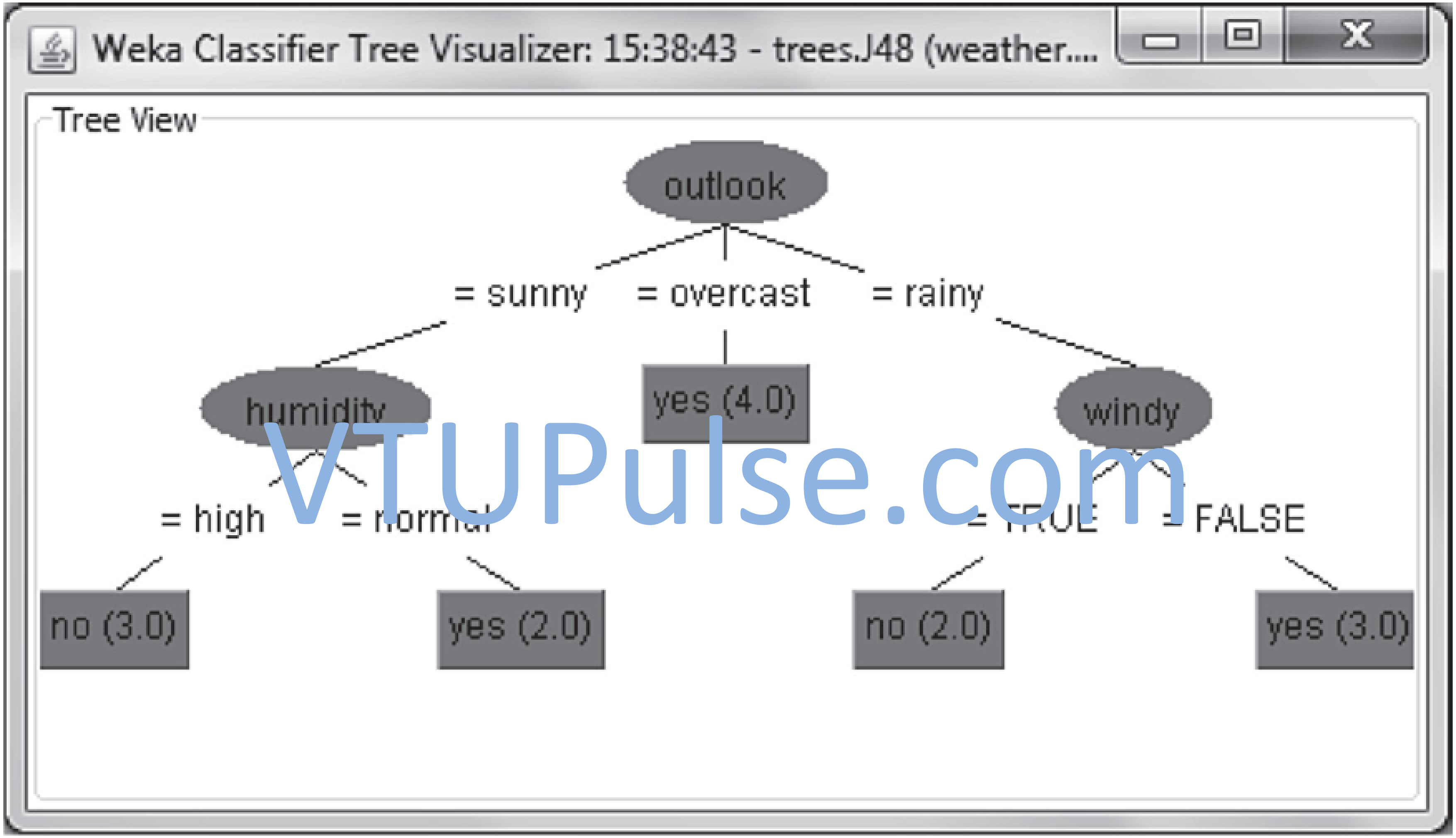
For Video Lectures subscribe to
https://www.youtube.com/c/maheshhuddar

Decision Tree Construction

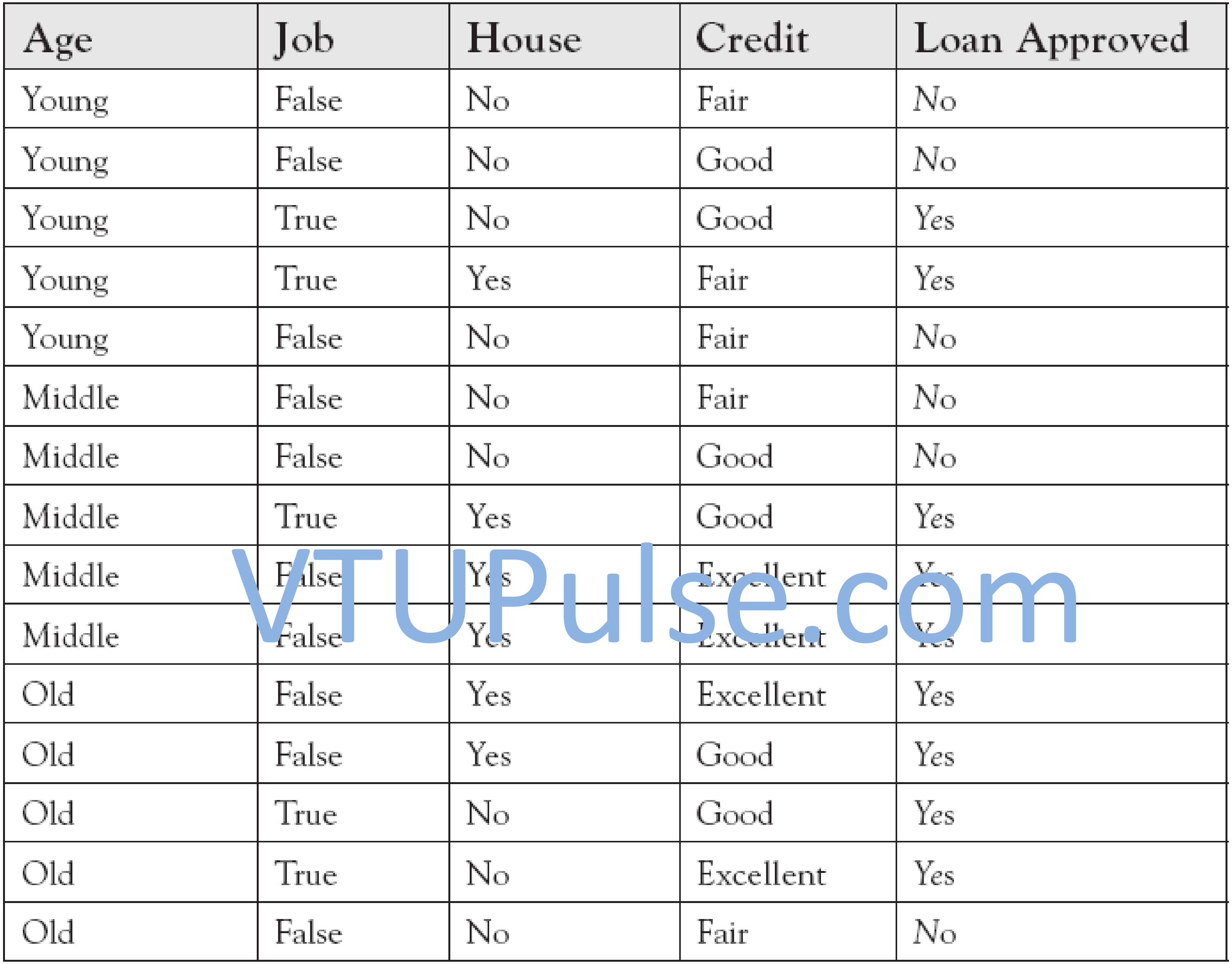


Attribute	Rules	Error	Total Error
Temp	Mild → Yes	1/3	2/5
	Cool → yes	1/2	
Humidity	High → No	1/2	1/5
	Normal → Yes	1/3	
Windy	False → Yes	0/3	0/5
	True → No	1/2	

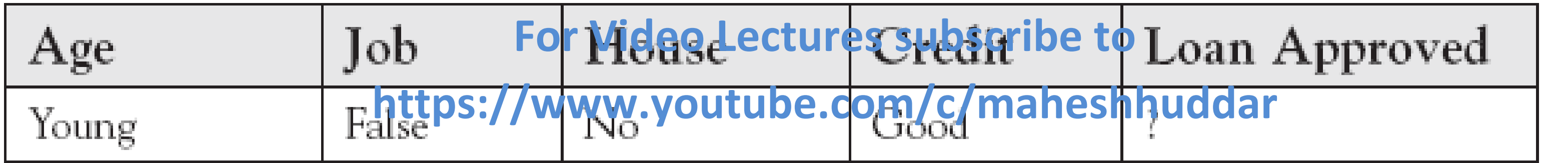
Decision Tree Construction



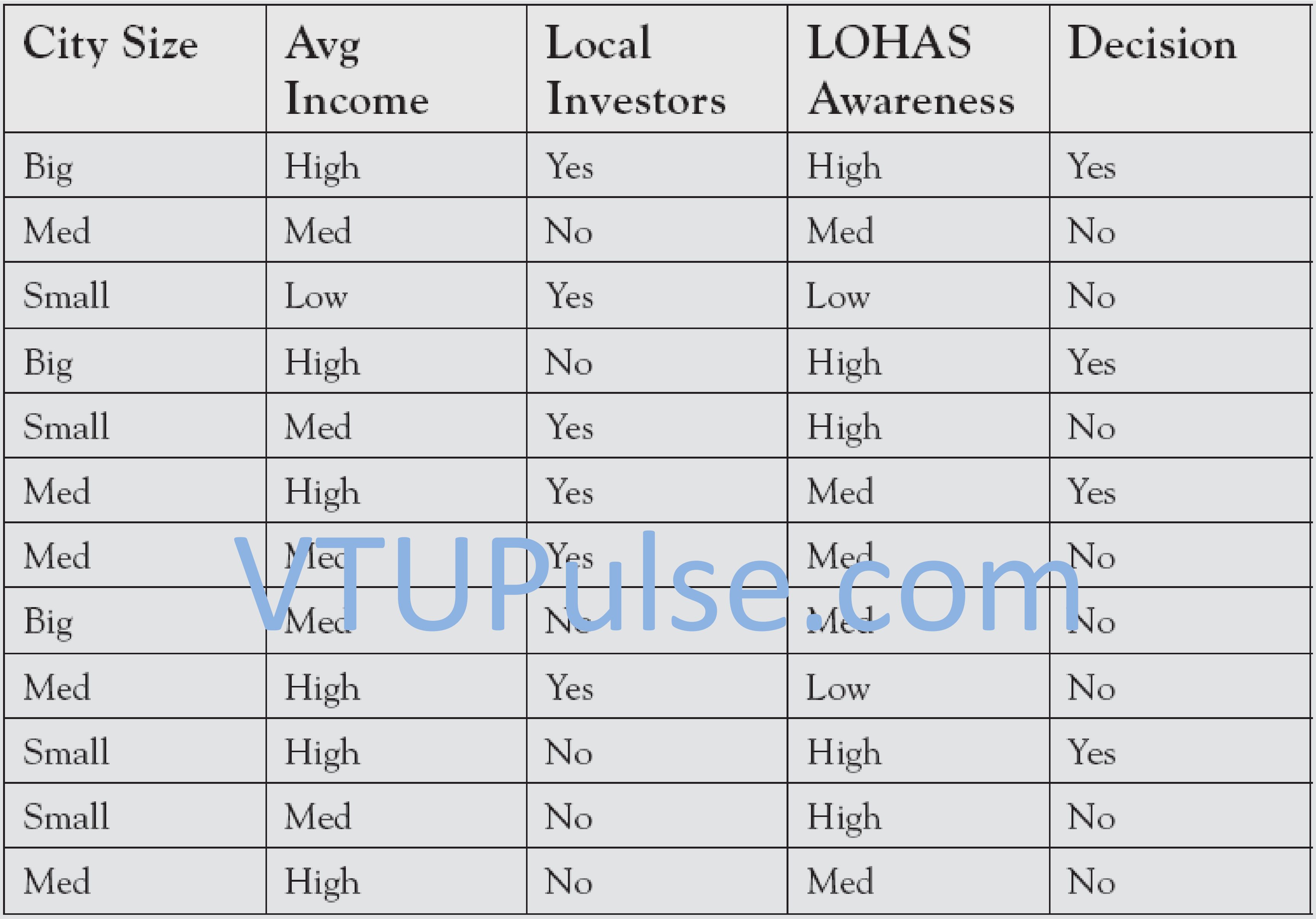
Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	Normal	True	?



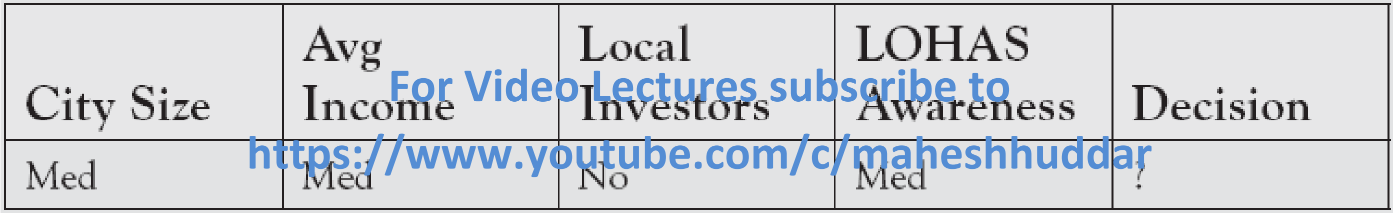
Age	Job	House	Credit	Loan Approved
Young	False	No	Fair	No
Young	False	No	Good	No
Young	True	No	Good	Yes
Young	True	Yes	Fair	Yes
Young	False	No	Fair	No
Middle	False	No	Fair	No
Middle	False	No	Good	No
Middle	True	Yes	Good	Yes
Middle	False	Yes	Excellent	Yes
Middle	False	Yes	Excellent	Yes
Old	False	Yes	Excellent	Yes
Old	False	Yes	Good	Yes
Old	True	No	Good	Yes
Old	True	No	Excellent	Yes
Old	False	No	Fair	No



Age	Job	House	Credit	Loan Approved
Young	False	No	Good	?



City Size	Avg Income	Local Investors	LOHAS Awareness	Decision
Big	High	Yes	High	Yes
Med	Med	No	Med	No
Small	Low	Yes	Low	No
Big	High	No	High	Yes
Small	Med	Yes	High	No
Med	High	Yes	Med	Yes
Med	Med	Yes	Med	No
Big	Med	No	Med	No
Med	High	Yes	Low	No
Small	High	No	High	Yes
Small	Med	No	High	No
Med	High	No	Med	No

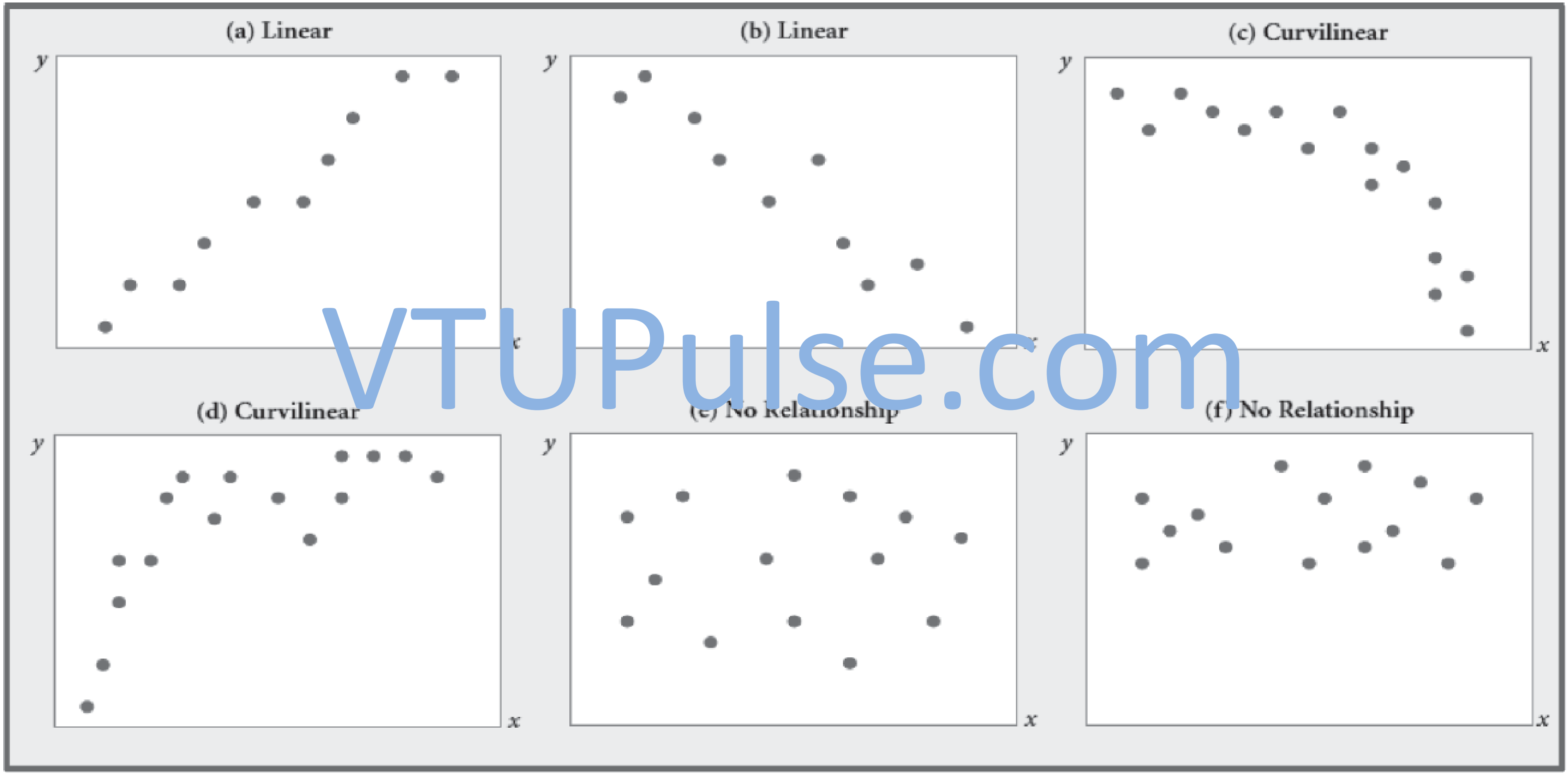


City Size	Avg Income	Local Investors	LOHAS Awareness	Decision
Med	Med	No	Med	?

Regression

- Regression is a well-known statistical technique to model the predictive relationship between several independent variables (DVs) and one dependent variable.
- The objective is to find the best-fitting curve for a dependent variable in a multidimensional space, with each independent variable being a dimension.
- The curve could be a straight line, or it could be a nonlinear curve.
- The quality of fit of the curve to the data can be measured by a coefficient of correlation (*r*), which is the square root of the amount of variance explained by the curve.

Visual Look at Relationships



Regression

The key steps for regression are simple:

1. List all the variables available for making the model.
2. Establish a dependent variable of interest.
3. Examine visual (if possible) relationships between variables of interest.
4. Find a way to predict dependent variable using the other variables.

Correlations and Relationships

- Statistical relationships are about which elements of data hang together, and which ones hang separately.
- It is about categorizing variables that have a relationship with one another, and categorizing variables that are distinct and unrelated to other variables.
- It is about describing significant positive relationships and significant negative differences.

Correlations and Relationships

- Given two numeric variables *x* and *y*, the coefficient of correlation *r* is mathematically computed by the following equation. $\bar{x}$ is the mean of *x*, and $\bar{y}$ is the mean of *y*.

$$r = \frac{[(x - \bar{x})][(y - \bar{y})]}{\sqrt{[(x - \bar{x})^2][(y - \bar{y})^2]}}$$

Correlations

- The first and foremost measure of the strength of a relationship is co-relation (or correlation). The strength of a correlation is a quantitative measure that is measured in a normalized range between 0 (zero) and 1
- A correlation of 1 indicates a perfect relationship, where the two variables are in perfect sync.
- A correlation of 0 indicates that there is no relationship between the variables.

Relationships

- The relationship can be positive, or it can be an inverse relationship, that is, the variables may move together in the same direction or in the opposite direction.
- Therefore, a good measure of correlation is the correlation coefficient, which is the square root of correlation.
- This coefficient, called *r*, can thus range from −1 to +1. An r value of 0 signifies no relationship.
- An *r* value of 1 shows perfect relationship in the same direction, and an *r* value of −1 shows a perfect relationship but moving in opposite directions.

Introduction to Regression Analysis

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

Independent variable: the variable used to explain the dependent variable

Introduction to Regression Analysis

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

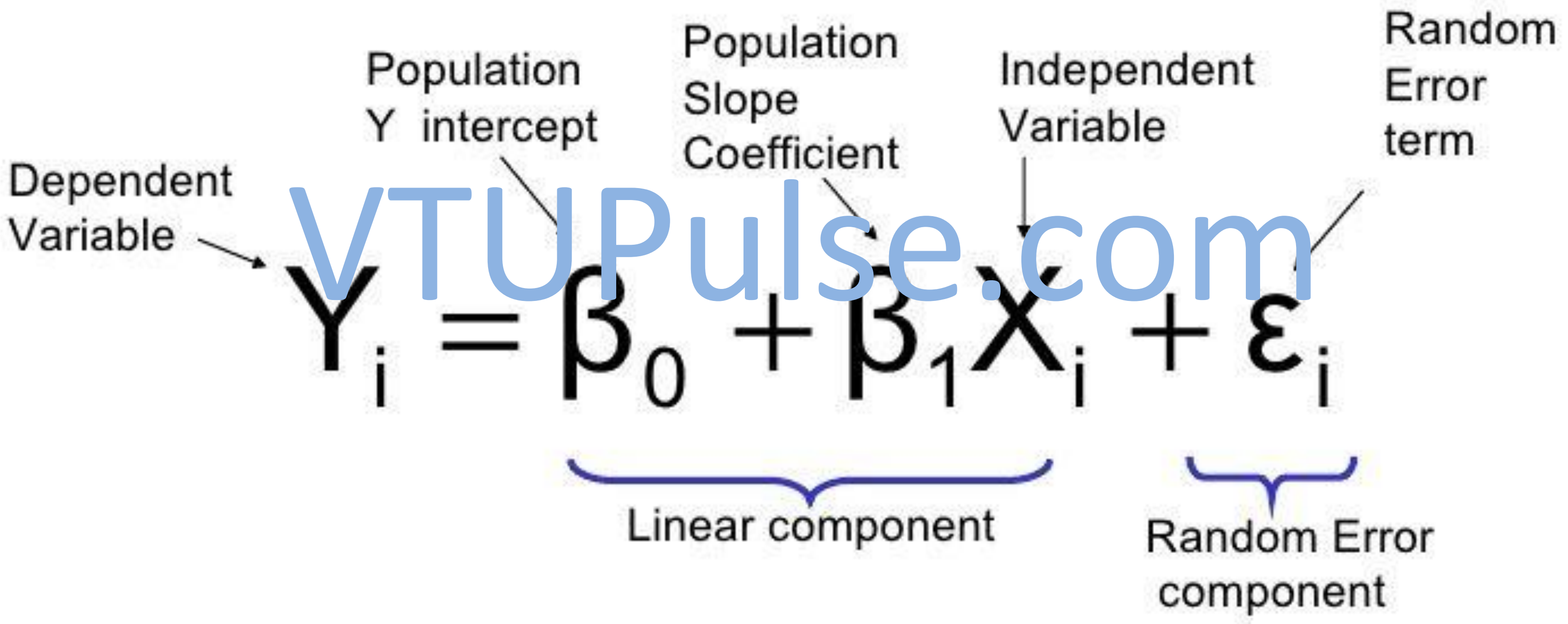
Independent variable: the variable used to explain the dependent variable

Regression Exercise

- The regression model is described as a linear equation that follows. *y* is the dependent variable, that is, the variable being predicted. *x* is the independent variable, or the predictor variable. There could be many predictor variables (such as *x*1, *x*2, . . .) in a regression equation. However, there can be only one dependent variable (*y*) in the regression equation.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Simple Linear Regression Model



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

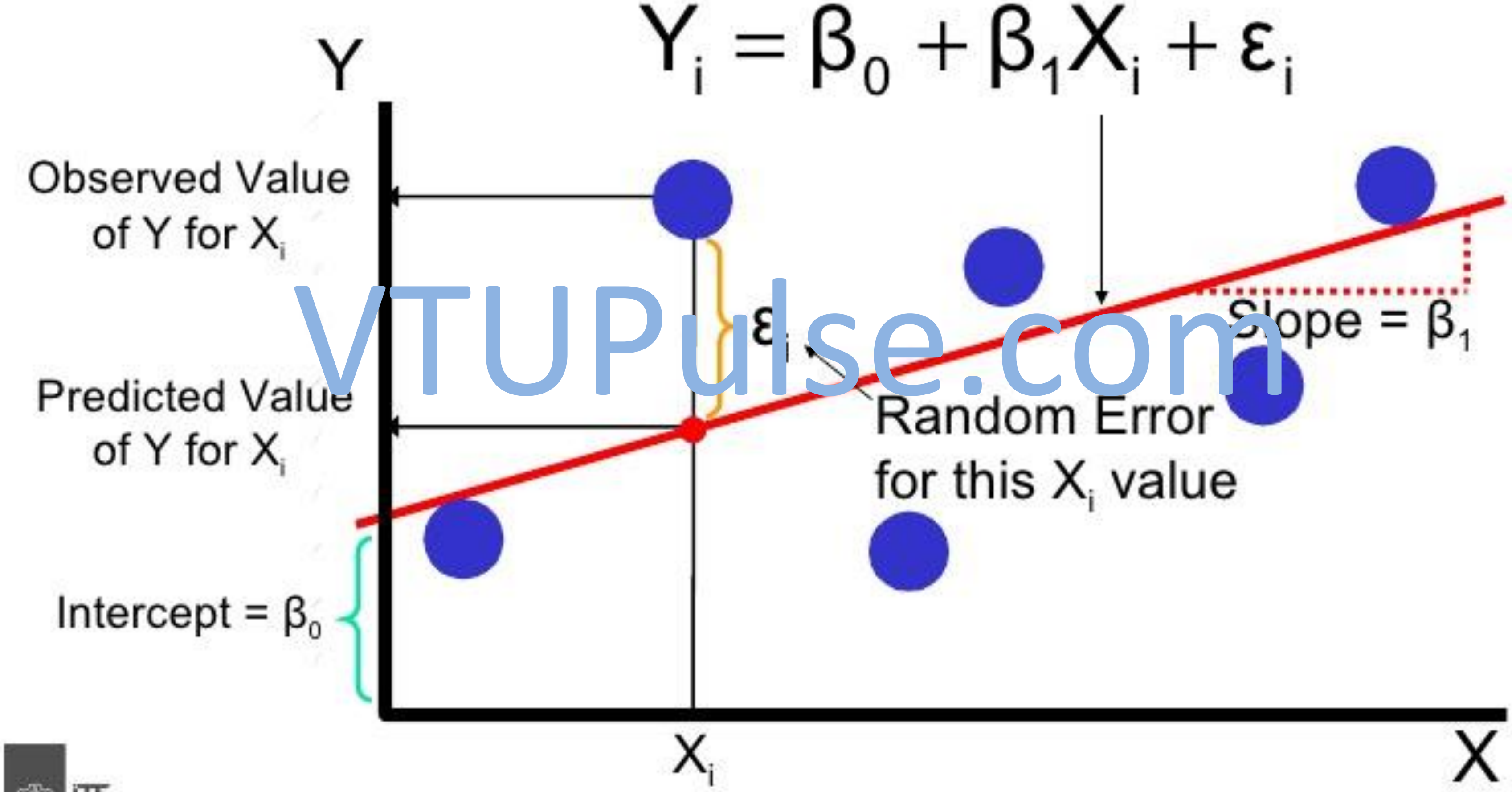
Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple Linear Regression Model

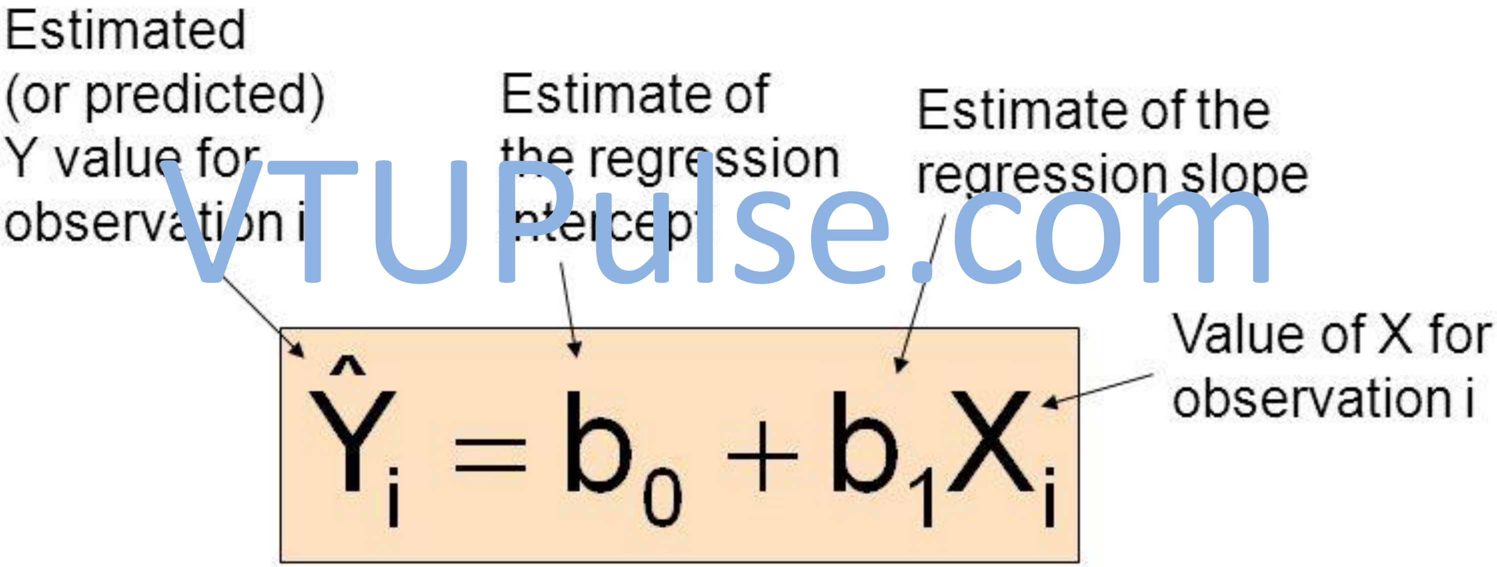
~~(continued)~~

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$


Y
Observed Value of Y for X_i
ε_i
Predicted Value of Y for X_i
Random Error for this X_i value
Slope = β_1
Intercept = β_0
X_i
X

Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line



Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

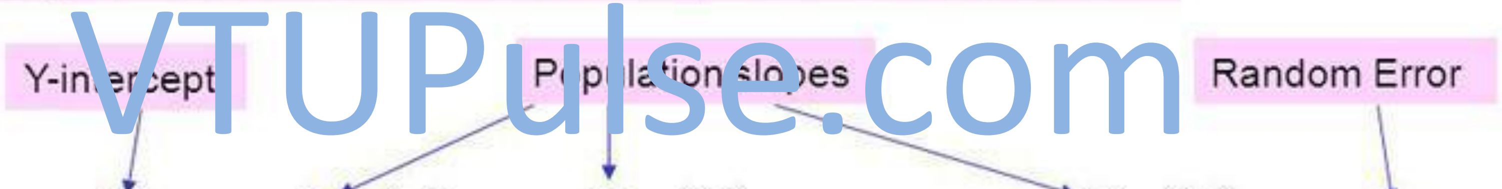
$$\hat{Y}_i = b_0 + b_1 X_i$$

The individual random error terms e_i have a mean of zero

The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

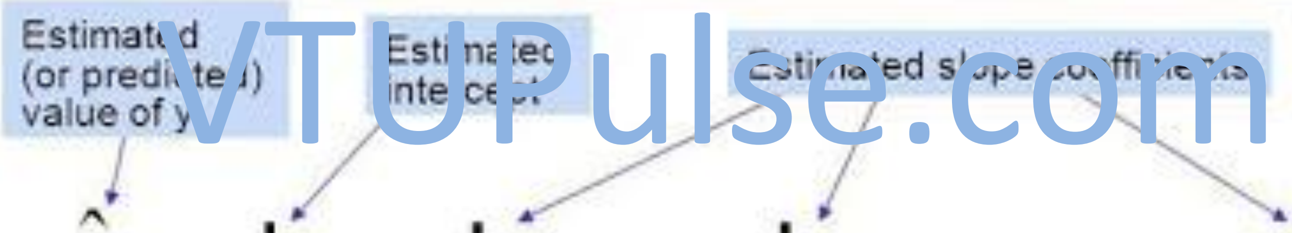


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:



Estimated (or predicted) value of y

Estimated intercept

Estimated slope coefficients

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}$$

In this chapter we will always use a computer to obtain the regression slope coefficients and other regression summary measures.

A simple example of a regression equation would be to predict glucose level from the age.

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

A simple example of a regression equation would be to predict glucose level from the age.

Step 1: *Make a chart of your data, filling in the columns in the same way as you would fill in the chart if you were finding the Pearson's Correlation Coefficient*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	**247**	**486**	**20485**	**11409**	**40022**

A simple example of a regression equation would be to predict glucose level from the age.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

A simple example of a regression equation would be to predict glucose level from the age.

Step 2: Use the following equations to find a and b.

Find a:

$((486 \times 11{,}409) - ((247 \times 20{,}485)) / 6\ (11{,}409) - 247^2)$

$= 484979 / 7445$

$= \mathbf{65.14}$

Find b:

$(6(20{,}485) - (247 \times 486)) / (6\ (11409) - 247^2)$

$= (122{,}910 - 120{,}042) / 68{,}454 - 247^2$

$= 2{,}868 / 7{,}445$

$= \mathbf{.385225}$

A simple example of a regression equation would be to predict glucose level from the age.

- **Step 3:** *Insert the values into the equation*.

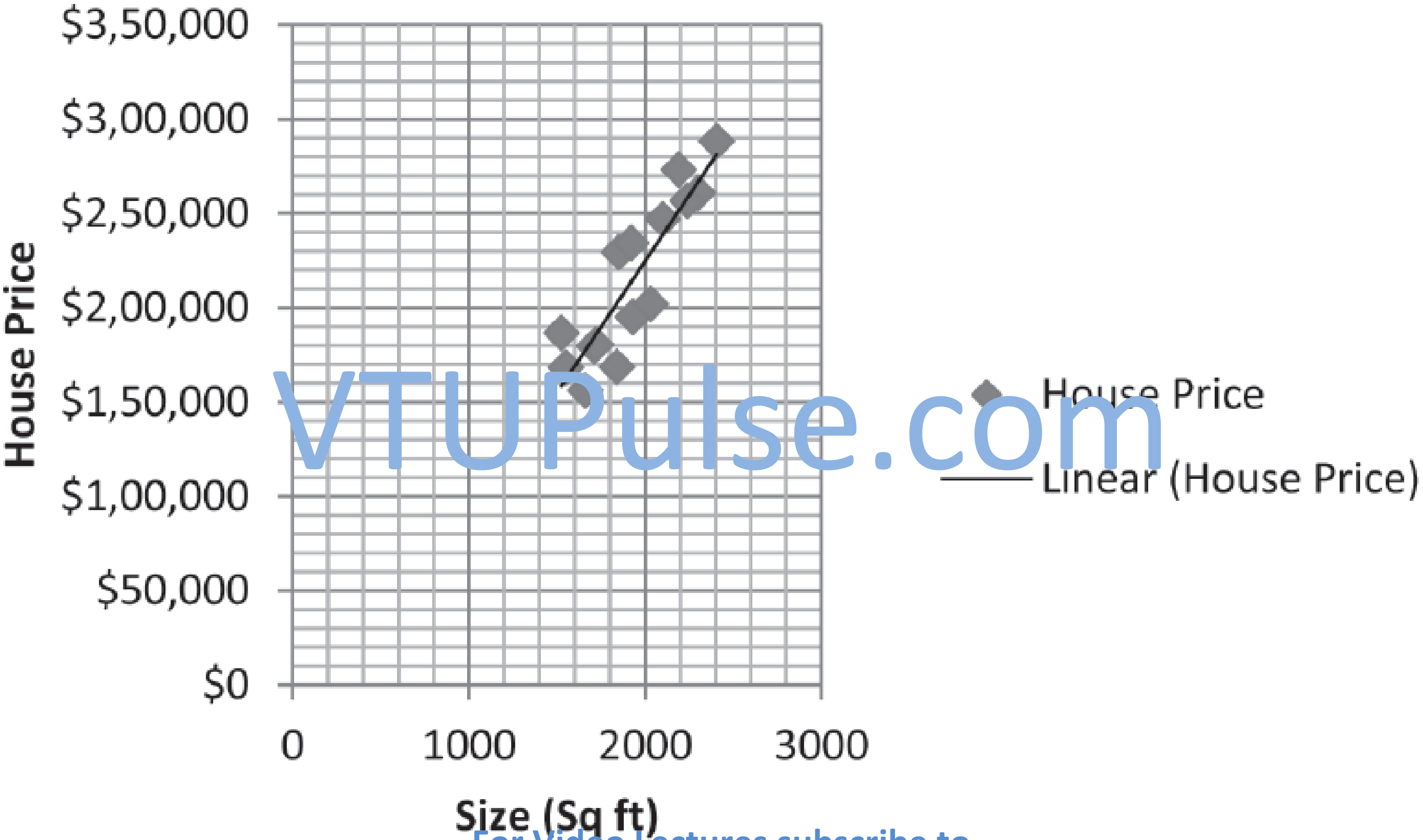
y’ = a + bx

y’ = 65.14 + .385225x

A simple example of a regression equation would be to predict a house price from the size of the house. Here are sample house data:

House Price	Size (sqft)
$229,500	1,850
$273,300	2,190
$247,000	2,100
$195,100	1,930
$261,000	2,300
$179,700	1,710
$168,500	1,550
$234,400	1,920
$168,800	1,840
$180,400	1,720
$156,200	1,660
$288,350	2,405
$186,750	1,525
$202,100	2,030
$256,800	2,240

Scatter plot and regression equation between House price and house size



- Visually, one can see a positive correlation between house price and size (sqft). However, the relationship is not perfect. Running a regression model between the two variables produces the following output (truncated).

Regression Statistics	
Multiple r	0.891
r2	0.794
	Coefficients
Intercept	-54,191
Size (sqft)	139.48

$$\text{House Price (\$)} = 139.48 \times \text{Size (sqft)} - 54{,}191$$

Linear Regression 2 independent variable

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Linear Regression 2 independent variable

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

Predict the value of Y give X1 and X2

Y	X1	X2
-3.7	3	8
3.5	4	5
2.5	5	7
11.5	6	3
5.7	2	1
?	3	2

Predict the value of Y give X1 and X2

- b0 = 2.799
- b1 = 2.288
- b2 = -2.671
- Y = 2.799 + 2.288 x1 – 2.671x2
- Y = ?

A simple example of a regression equation would be to predict a house price from the size of the house. Here are sample house data:, with room extra variable

House Price	Size (sqft)	# Rooms
$229,500	1,850	4
$273,300	2,190	5
$247,000	2,100	4
$195,100	1,930	3
$261,000	2,300	4
$179,700	1,710	2
$168,500	1,550	2
$234,400	1,920	4
$168,800	1,840	2
$180,400	1,720	2
$156,200	1,660	2
$288,350	2,405	5
$186,750	1,525	3
$202,100	2,030	2
$256,800	2,240	4

Regression Statistics	
Multiple r	0.984
r2	0.968
	Coefficients
Intercept	12,923
Size(sqft)	65.60
Rooms	23,613

$$\text{House Price (\$)} = 65.6 \times \text{Size (sqft)} + 23{,}613 \times \text{Rooms} + 12{,}924$$

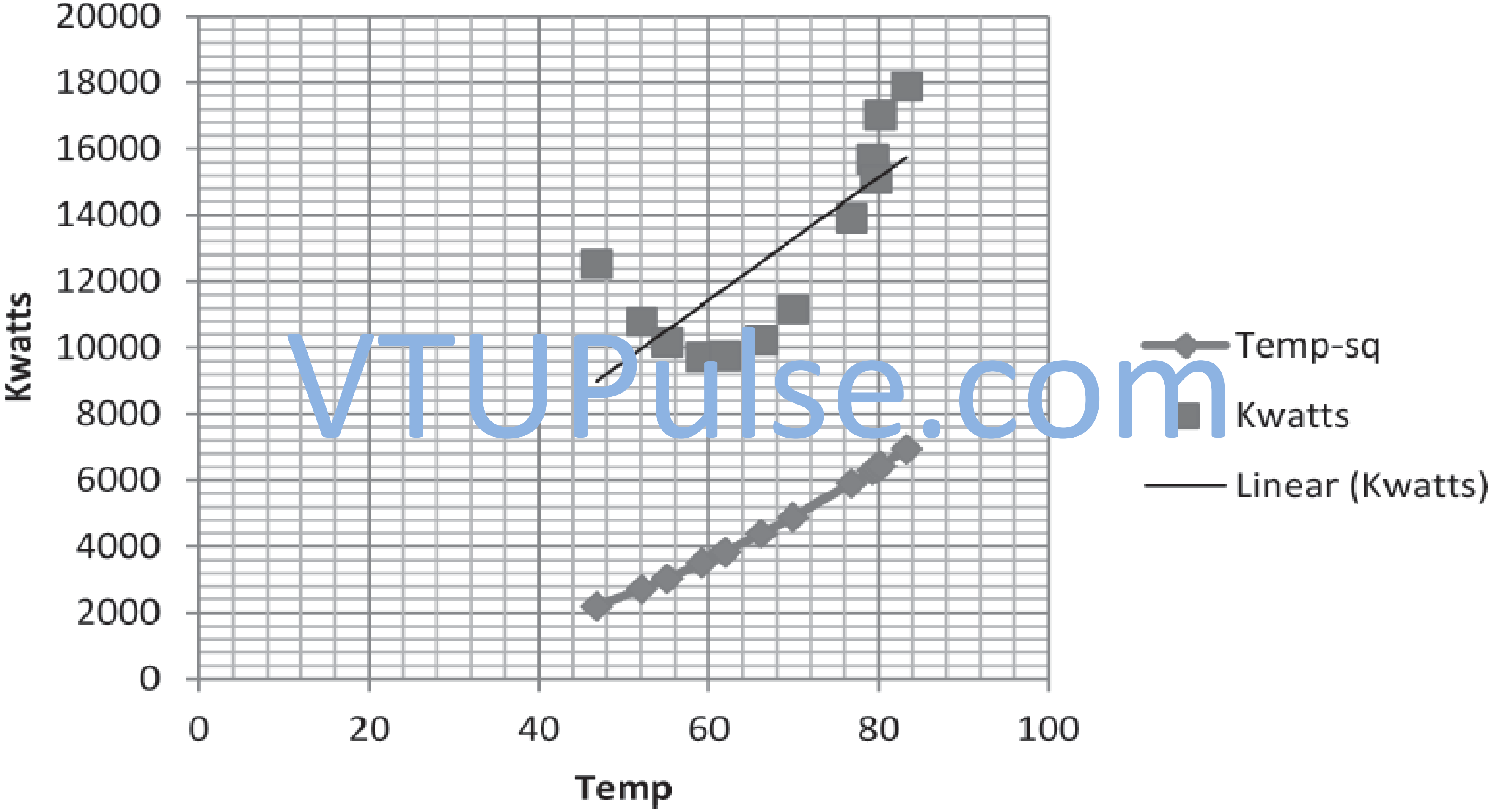
Nonlinear Regression Exercise

- The relationship between the variables may also be curvilinear. For example, given past data from electricity consumption (kWh) and temperature (temp), the objective is to predict the electrical consumption from the temperature value. Here are a dozen past observations.

Nonlinear Regression Exercise

KWatts	Temp (F)
12,530	46.8
10,800	52.1
10,180	55.1
9,730	59.2
9,750	61.9
10,230	66.2
11,160	69.9
13,910	76.8
15,690	79.3
15,110	79.7
17,020	80.2
17,880	83.3

Nonlinear Regression Exercise



(a) Kwatts and temp, and (b) kwatts and temp-sq

Regression Statistics	
Multiple r	0.992
r2	0.984
	Coefficients
Intercept	67,245
Temp (F)	–1,911
Temp-sq	15.87

Energy Consumption = 15.87 × temp-sq – 1,911 × Temp + 67,245

Cluster Analysis

- Cluster analysis is used for automatic identification of natural groupings of things. It is also known as the **segmentation technique**.
- In this technique, data instances that are **similar to (or near)** each other are categorized into one cluster.
- Similarly, data instances that are **very different (or far away)** from each other are moved into different clusters.
- Clustering is an unsupervised learning technique as there is no output or dependent variable for which a right or wrong answer can be computed.
- The correct number of clusters or the definition of those clusters is not known ahead of time.
- Clustering techniques can only suggest to the user how many clusters would make sense from the characteristics of the data.

Cluster Analysis

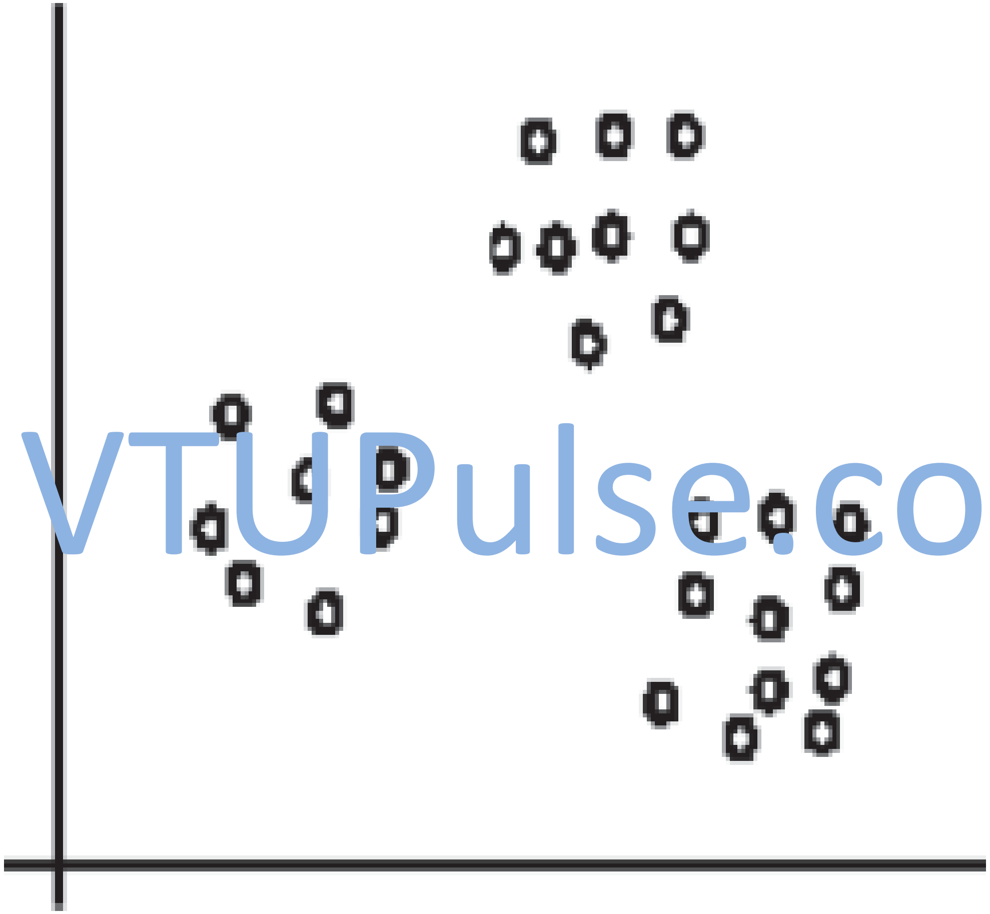


Figure 8.1

Applications of Cluster Analysis

- *Market segmentation:* Categorizing customers according to their similarities, for example by their common wants and needs and propensity to pay, can help with targeted marketing.
- *Product portfolio:* People of similar sizes can be grouped together to make small, medium, and large sizes for clothing items.
- *Text mining:* Clustering can help organize a given collection of text documents according to their content similarities into clusters of related topics.

Clustering Techniques

- Cluster analysis is a machine-learning technique. The quality of a clustering result depends on the *algorithm*, the *distance* function, and the *application*.
- First, consider the distance function. Most cluster analysis methods use a distance measure to calculate the closeness between pairs of items.
- There are two major measures of distances: **Euclidian distance** is the most intuitive measure.
- The other popular measure is the **Manhattan (rectilinear) distance**, where one can go only in orthogonal directions.
- In either case, the key objective of the clustering algorithm is the same:
  - Interclusters distance ⇒ maximized
  - Intraclusters distance ⇒ minimized

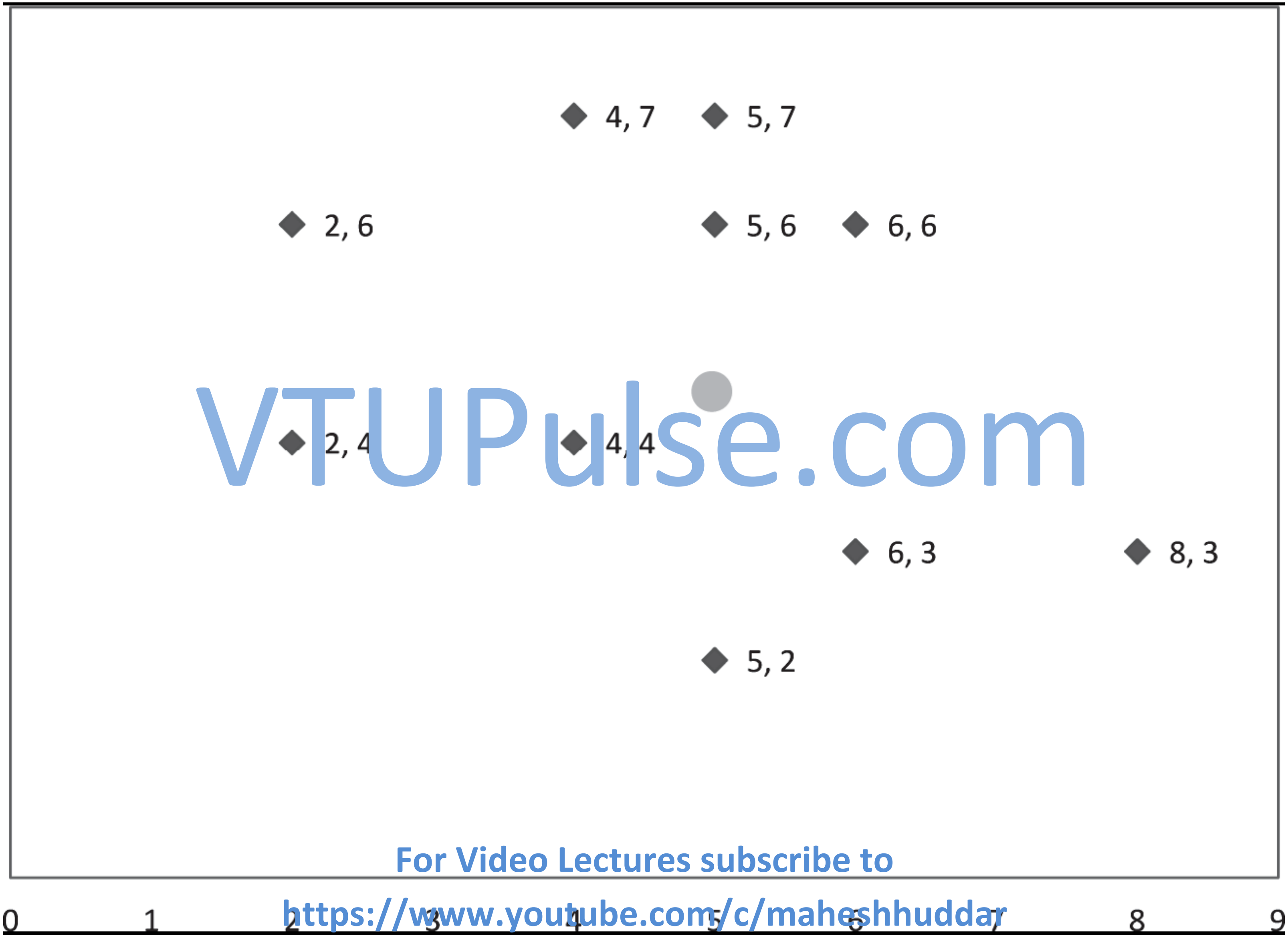
Here is the generic pseudocode for clustering

1. *Pick an arbitrary number of groups/segments to be created.*
2. *Start with some initial randomly chosen center values for groups.*
3. *Classify instances to closest groups*
4. *Compute new values for the group centers.*
5. *Repeat Steps 3 and 4 till groups converge.*
6. *If clusters are not satisfactory, go to Step 1 and pick a different number of groups/segments.*

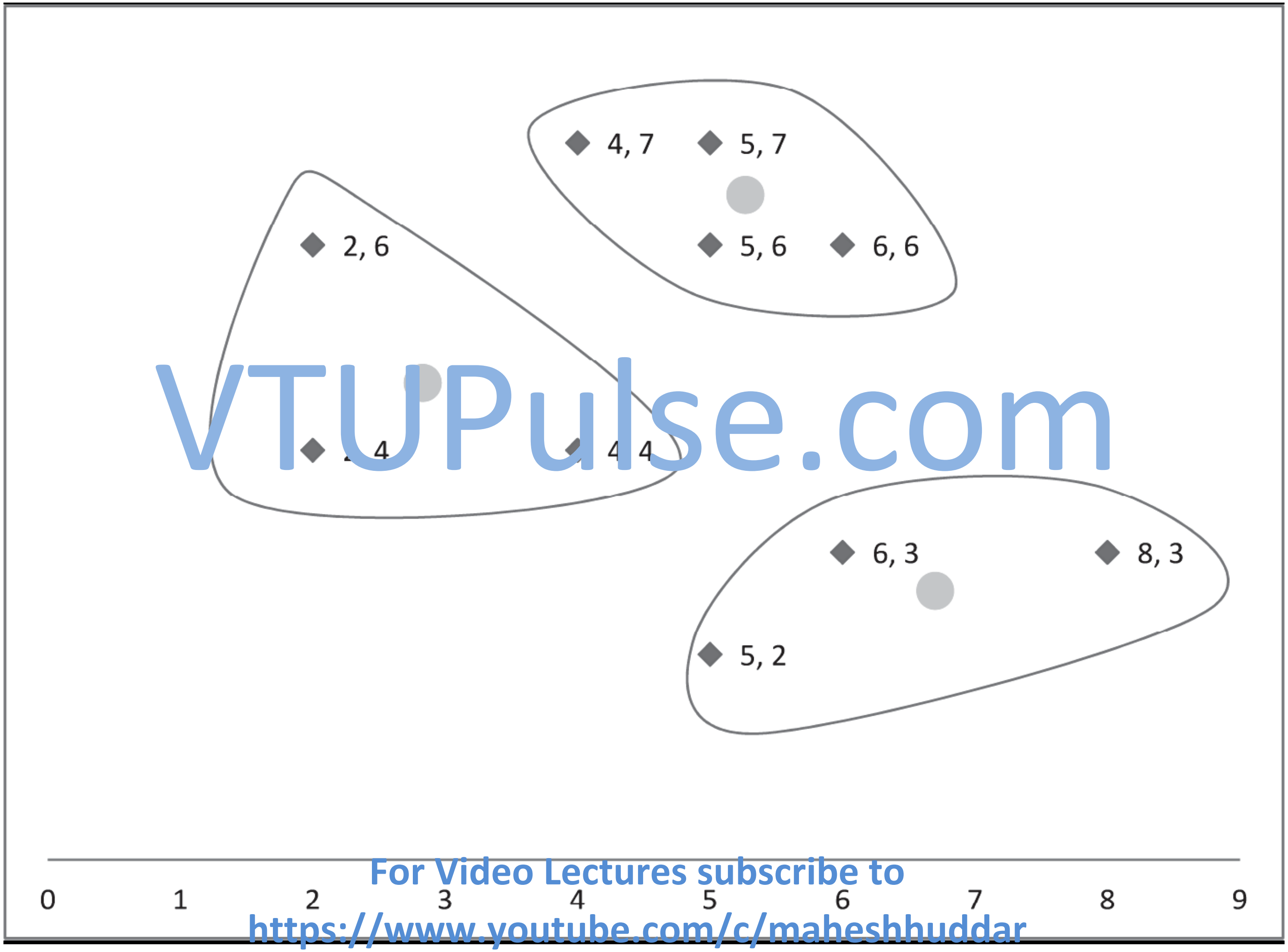
Clustering Exercise

X	Y
2	4
2	6
5	6
4	7
8	3
6	6
5	2
5	7
6	3
4	4

Clustering Exercise



Clustering Exercise

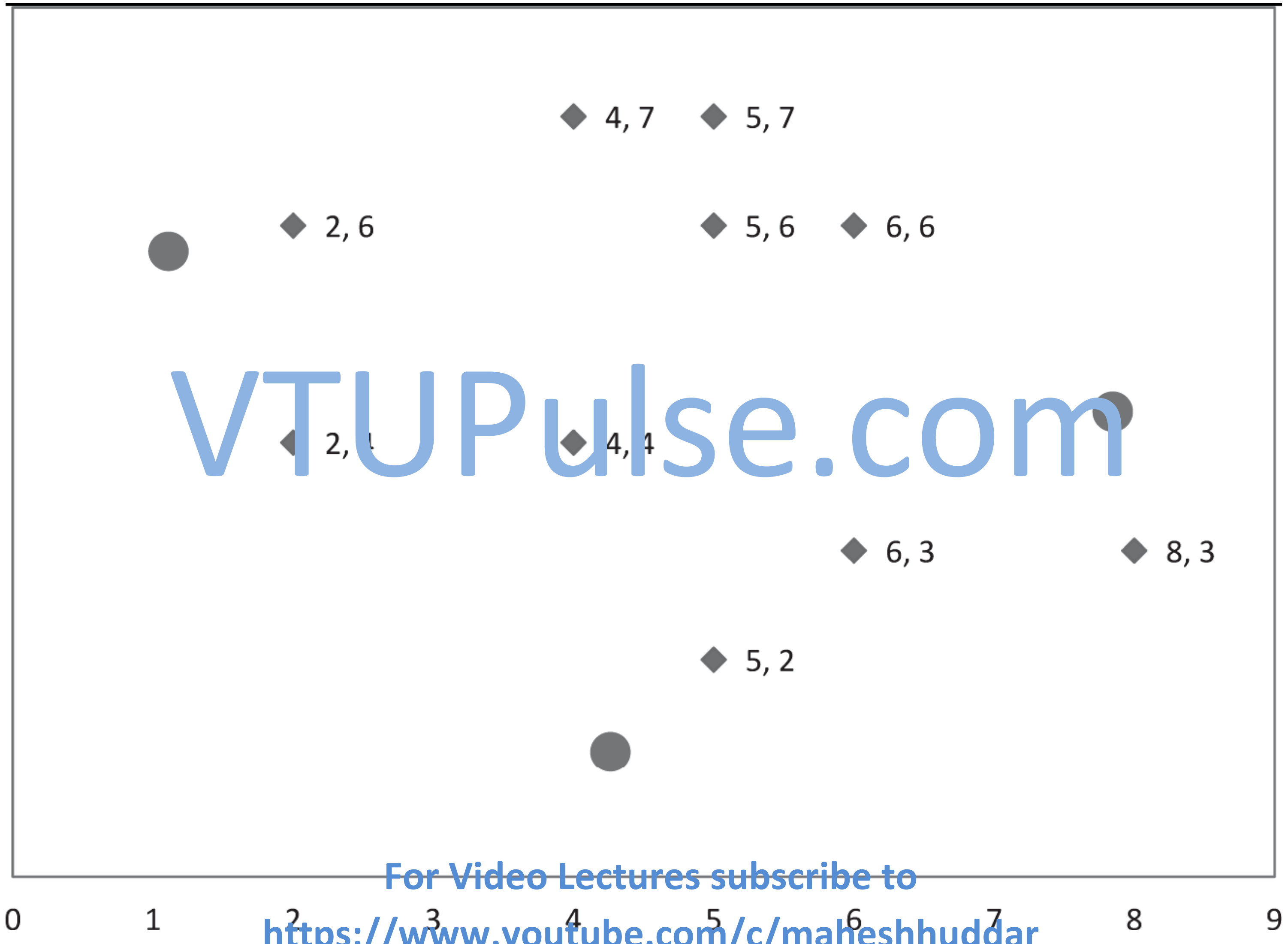


K-Means Algorithm for Clustering

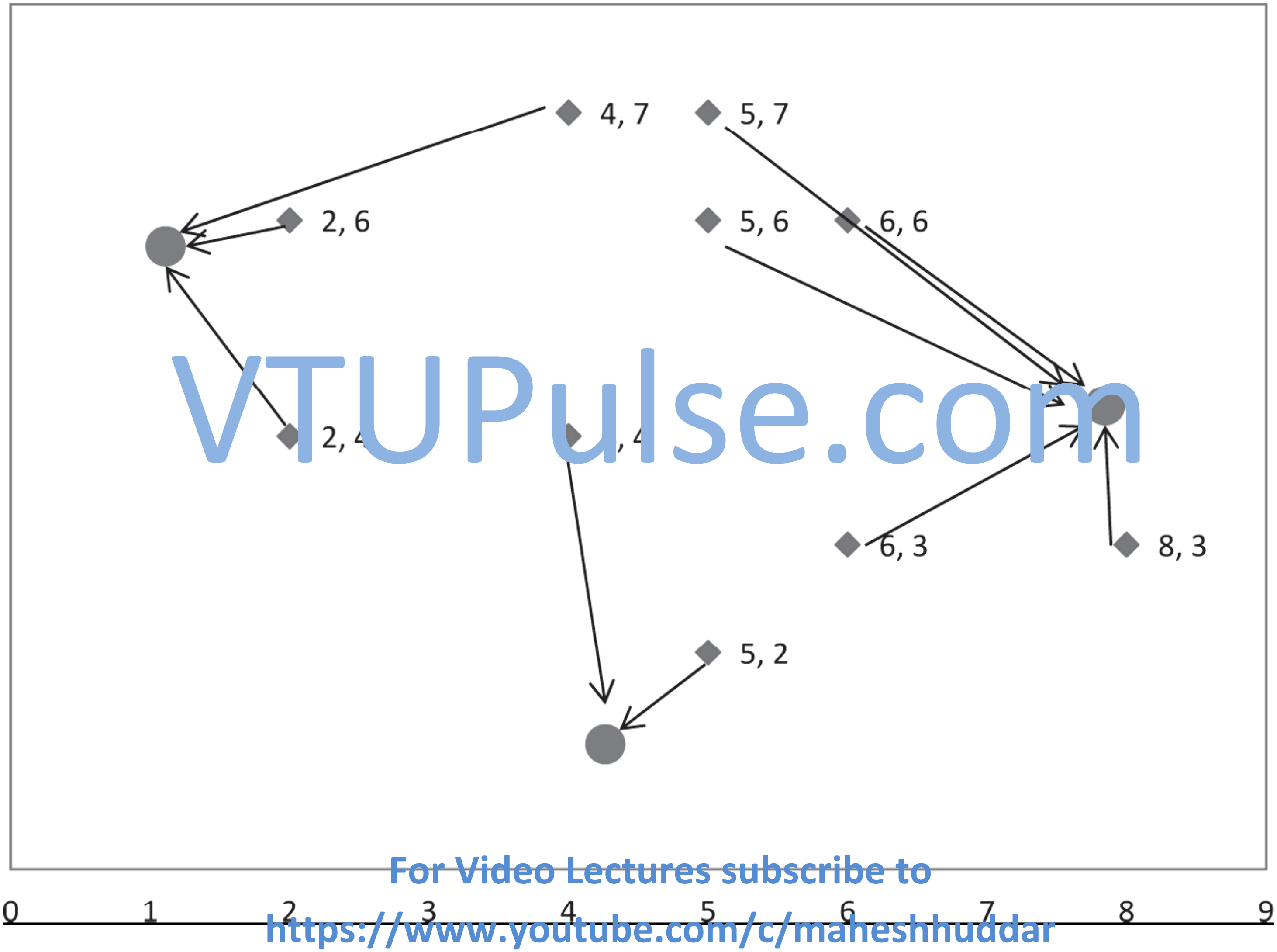
Here is the pseudocode for implementing a K-means algorithm.

```
Input: Algorithm K-Means (K number of clusters, D list of data points)
1. Choose K number of random data points as initial centroids (cluster centers).
2. Repeat till cluster centers stabilize:
    a. Allocate each point in D to the nearest of Kth centroids.
    b. Compute centroid for the cluster using all points in the cluster.
```

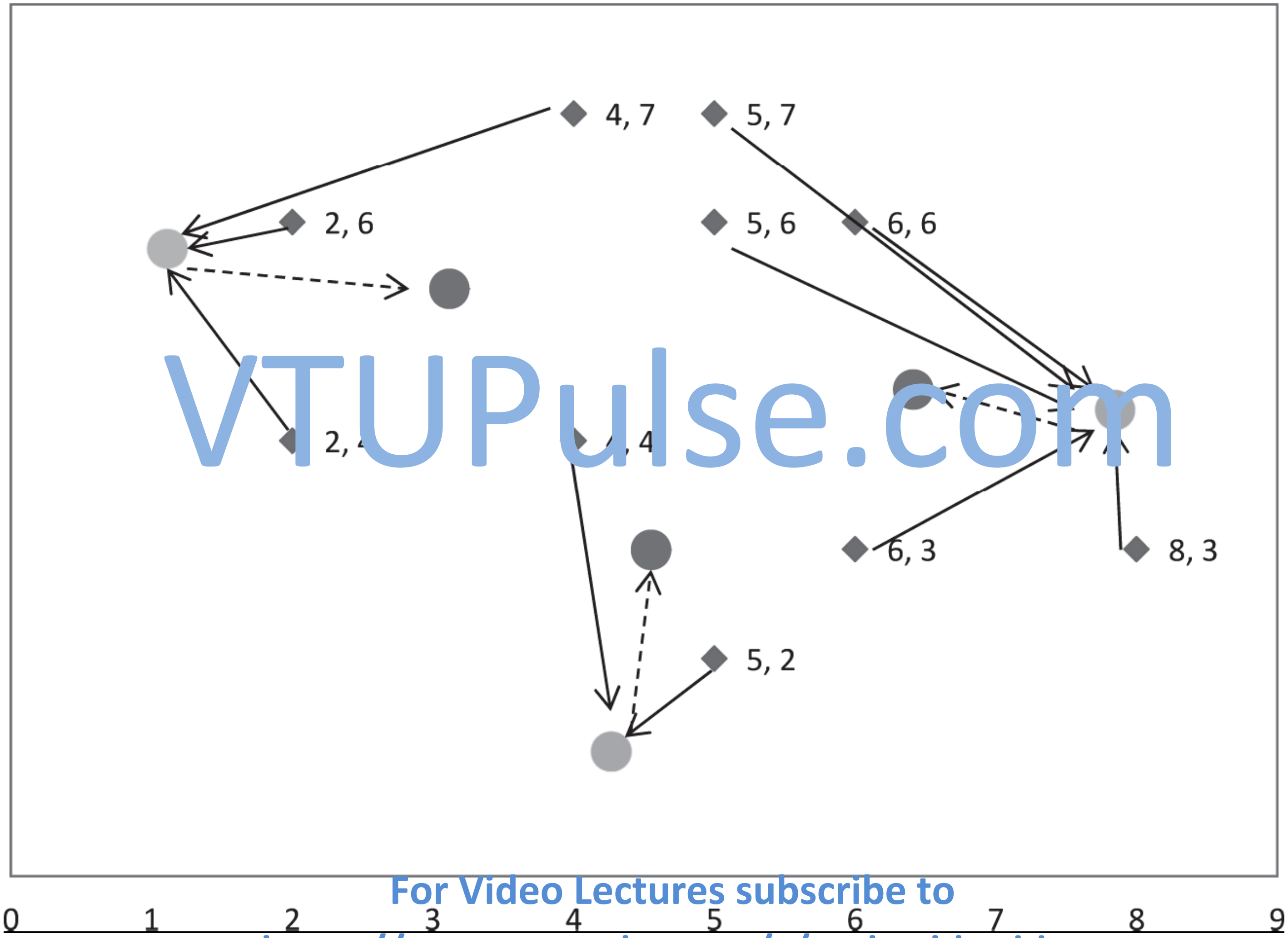
K-Means Algorithm for Clustering



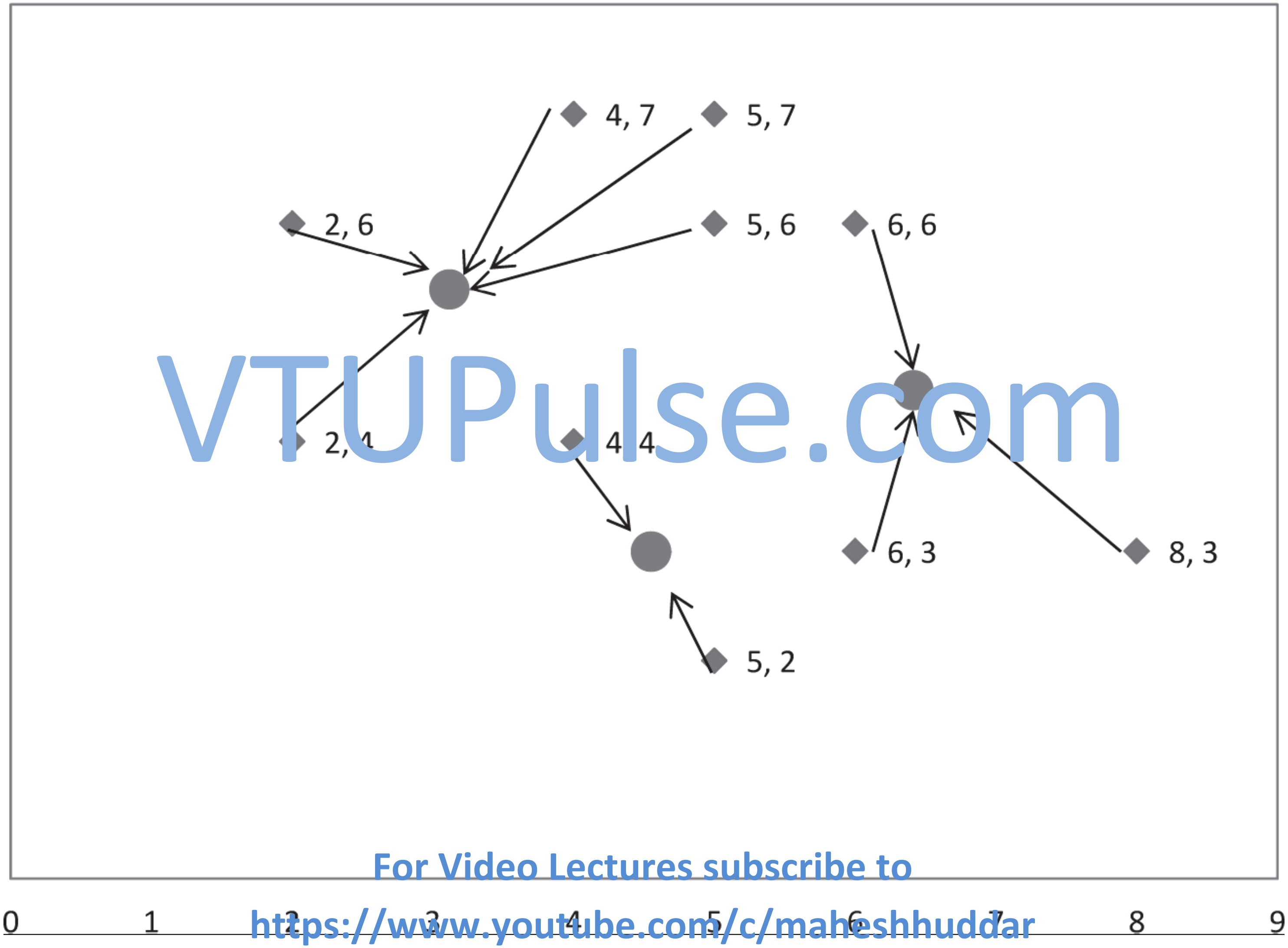
K-Means Algorithm for Clustering



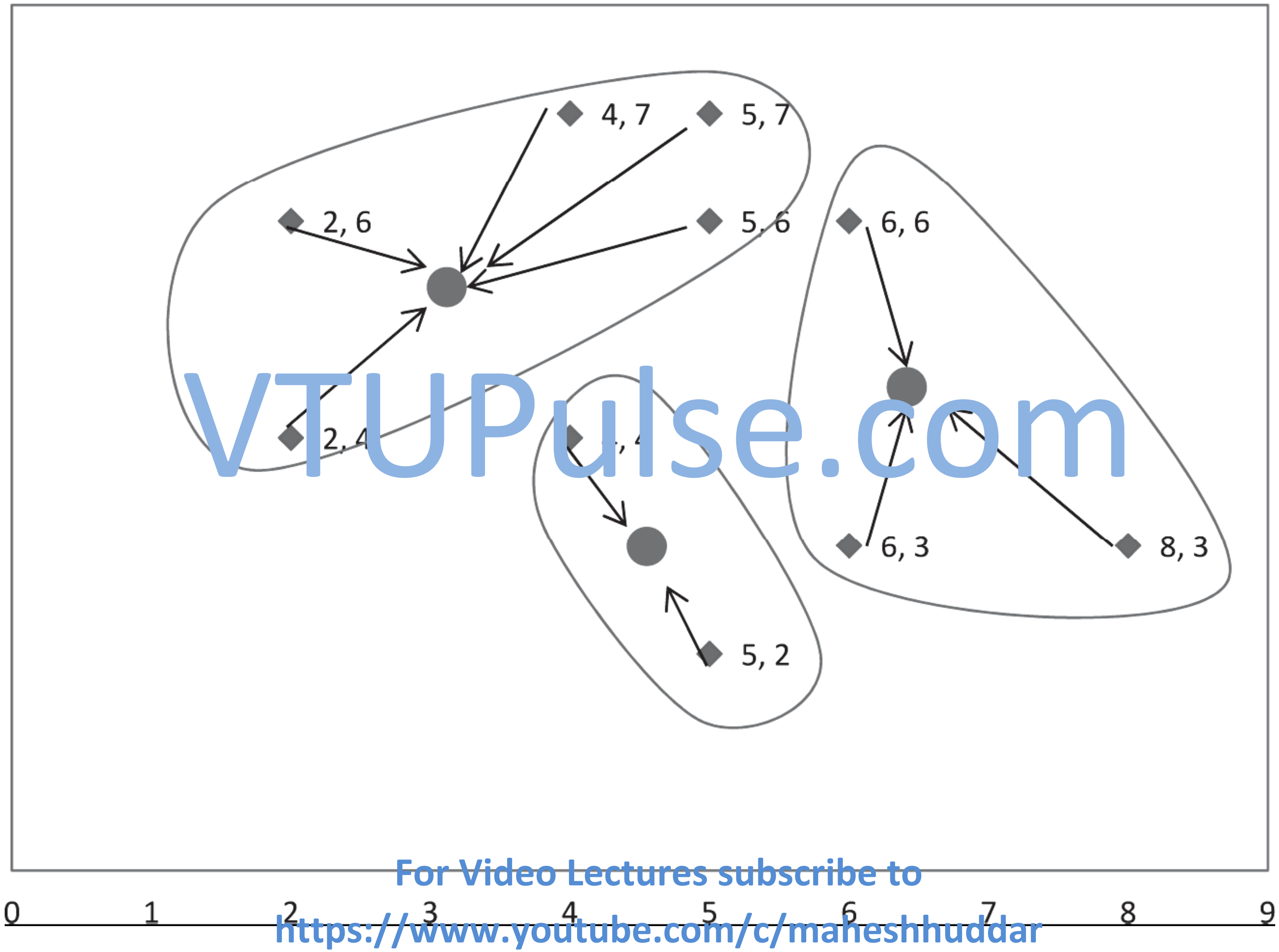
K-Means Algorithm for Clustering



K-Means Algorithm for Clustering



K-Means Algorithm for Clustering



Advantages and Disadvantages of K-Means Algorithm

There are many advantages of **K-Means Algorithm**

1. K-means algorithm is simple, easy to understand, and easy to implement.
2. It is also efficient, in which the time taken to cluster K-means rises linearly with the number of data points.
3. No other clustering algorithm performs better than K-means, in general.

There are many disadvantages of **K-Means Algorithm**

1. The user needs to specify an initial value of K.
2. The process of finding the clusters may not converge.
3. It is not suitable for discovering clusters that are not hyper ellipsoids or hyper spheres).

Data about height and weight for a few volunteers is available. Create a set of clusters for the following data, to decide how many sizes of Tshirts should be ordered

Height	Weight
71	165
68	165
72	180
67	113
72	178
62	101
70	150
69	172
72	185
63	149
69	132
61	115

Artificial Neural Networks

- Artificial neural networks (ANNs) are inspired by the information processing model of the mind/brain.
- The human brain consists of billions of neurons that link with one another in an intricate pattern.
- Every neuron receives information from many other neurons, processes it, gets excited or not, and passes its state information to other neurons.

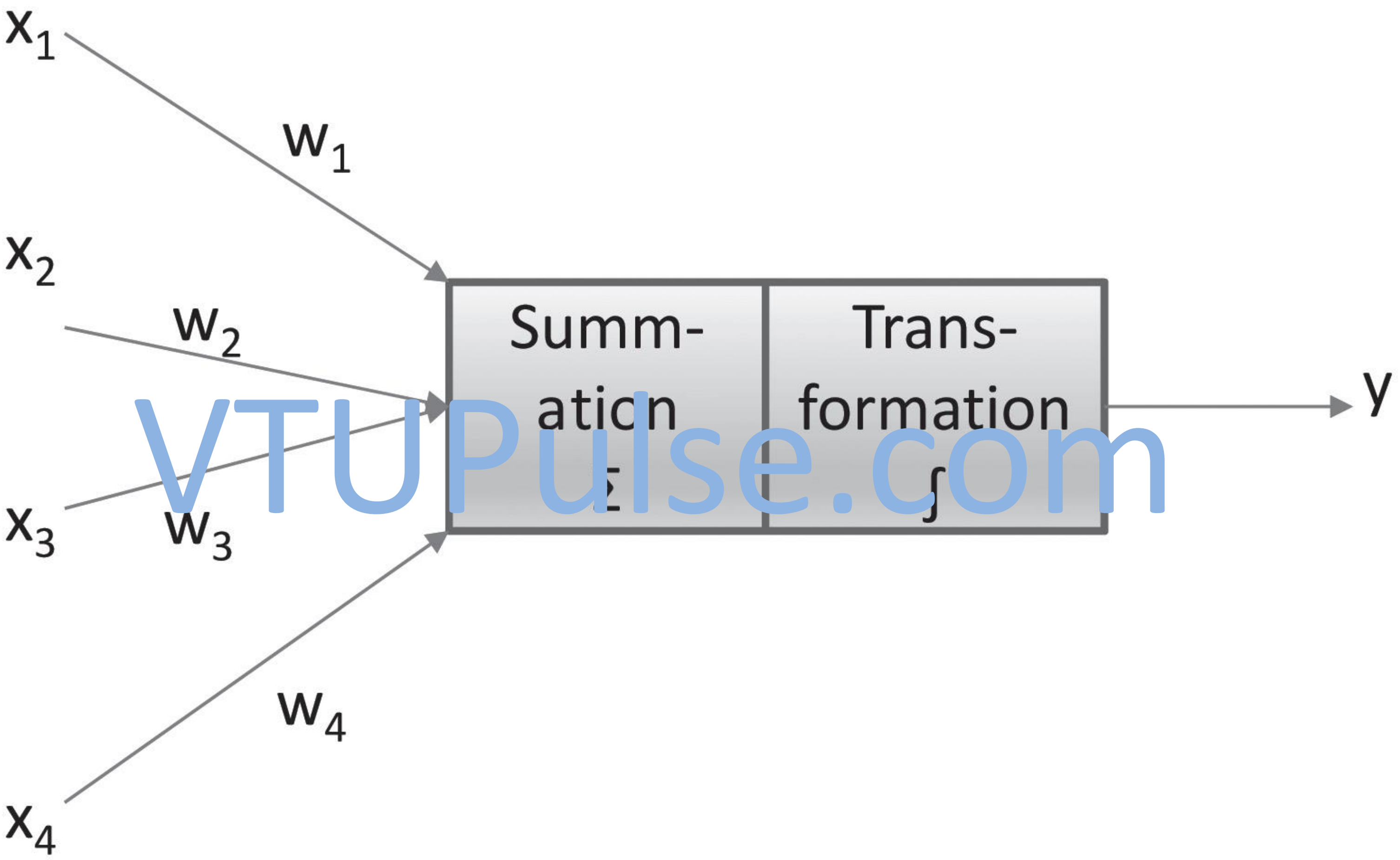
Inputs

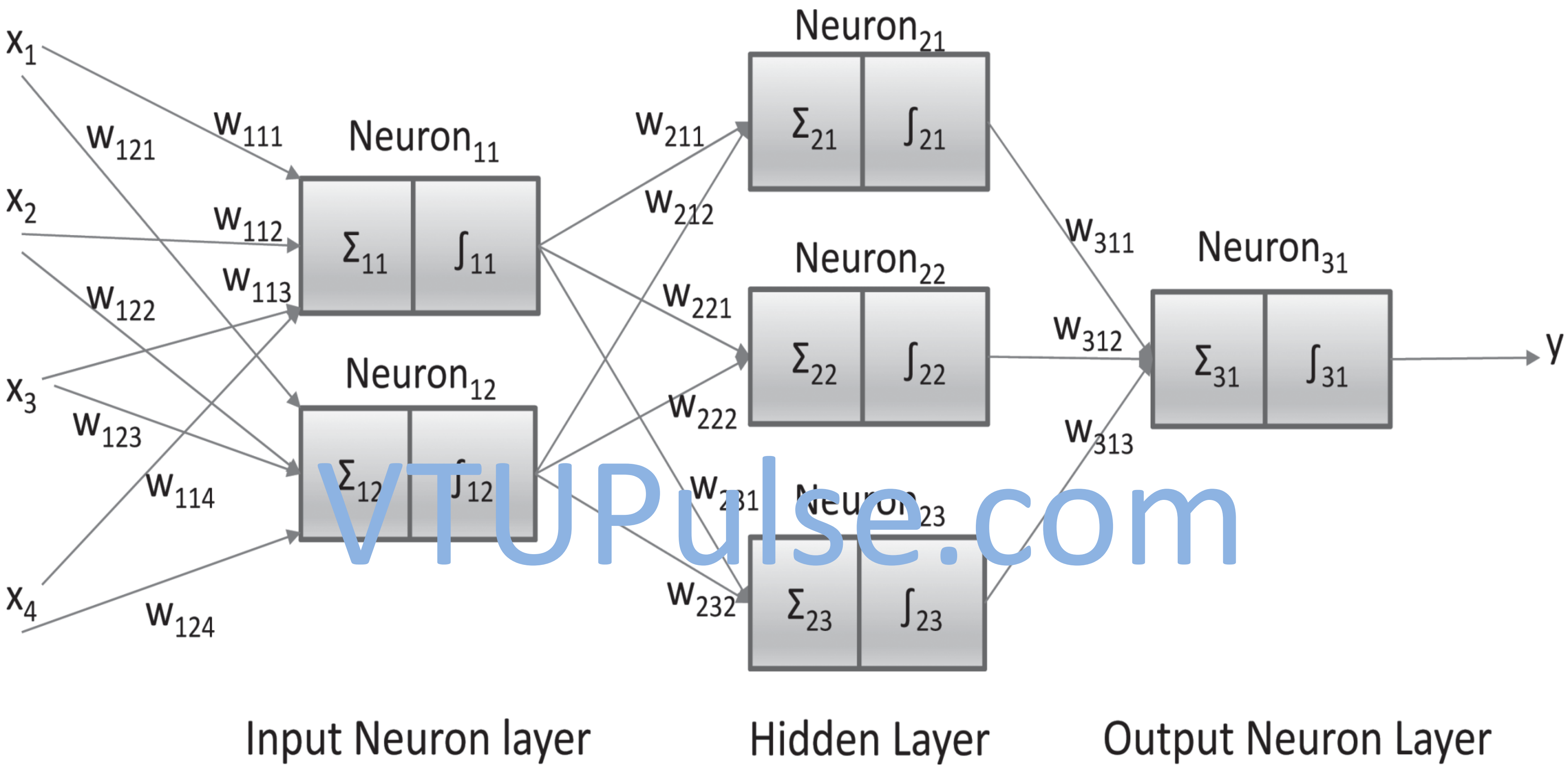
Artificial Neural Network (ANN)

Outputs

Business Applications of ANN

1. They are used in stock price prediction where the rules of the game are extremely complicated, and a lot of data needs to be processed very quickly.
2. They are used for character recognition, as in recognizing handwritten text, or damaged or mangled text. They are used in recognizing fingerprints. These are complicated patterns and are unique for each person. Layers of neurons can progressively clarify the pattern.
3. They are also used in traditional classification problems, such as approving a loan application.

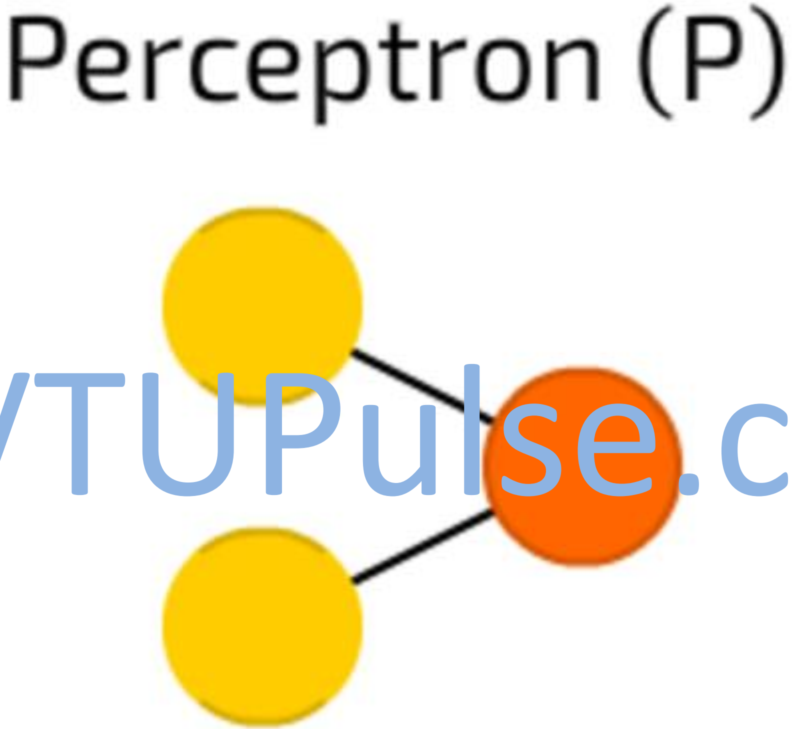

x1
w1
x2
w2
x3
w3
x4
w4
Summ-
ation
Σ
Trans-
formation
∫
y


x_1
x_2
x_3
x_4
w_{111}
w_{112}
w_{113}
w_{114}
w_{121}
w_{122}
w_{123}
w_{124}
Neuron$_{11}$
Σ_{11}
$\int_{11}$
Neuron$_{12}$
Σ_{12}
$\int_{12}$
w_{211}
w_{212}
w_{221}
w_{222}
w_{231}
w_{232}
Neuron$_{21}$
Σ_{21}
$\int_{21}$
Neuron$_{22}$
Σ_{22}
$\int_{22}$
Neuron$_{23}$
Σ_{23}
$\int_{23}$
w_{311}
w_{312}
w_{313}
Neuron$_{31}$
Σ_{31}
$\int_{31}$
y
Input Neuron layer
Hidden Layer
Output Neuron Layer

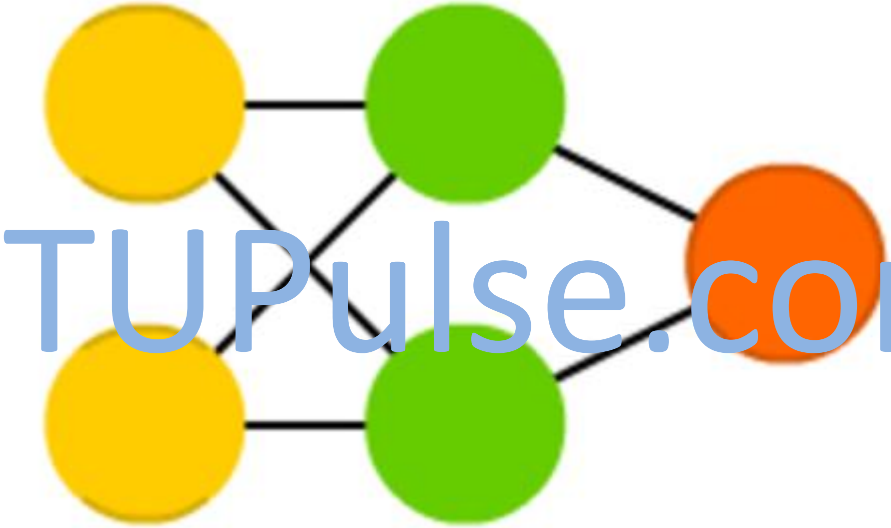
Developing an ANN

The steps required to build an ANN are as follows:

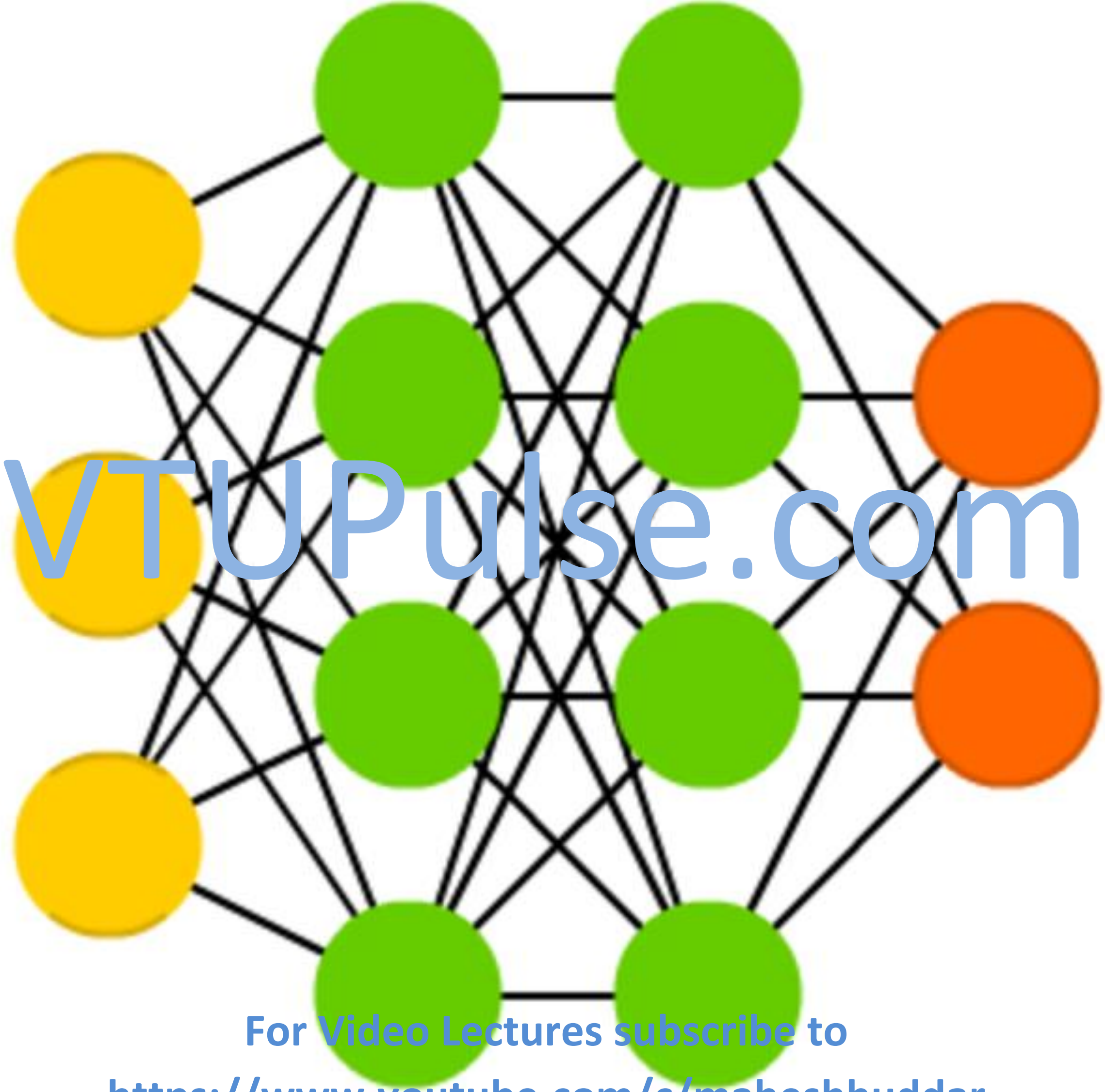
1. Gather data: Divide into training data and test data. The training data needs to be further divided into training data, validation data and testing data.
2. Select the network architecture, such as feedforward network.
3. Select the algorithm, such as Multilayer Perception
4. Set network parameters.
5. Train the ANN with training data.
6. Validate the model with validation data.
7. Freeze the weights and other parameters.
8. Test the trained network with test data.
9. Deploy the ANN when it achieves good predictive accuracy.


Perceptron (P)

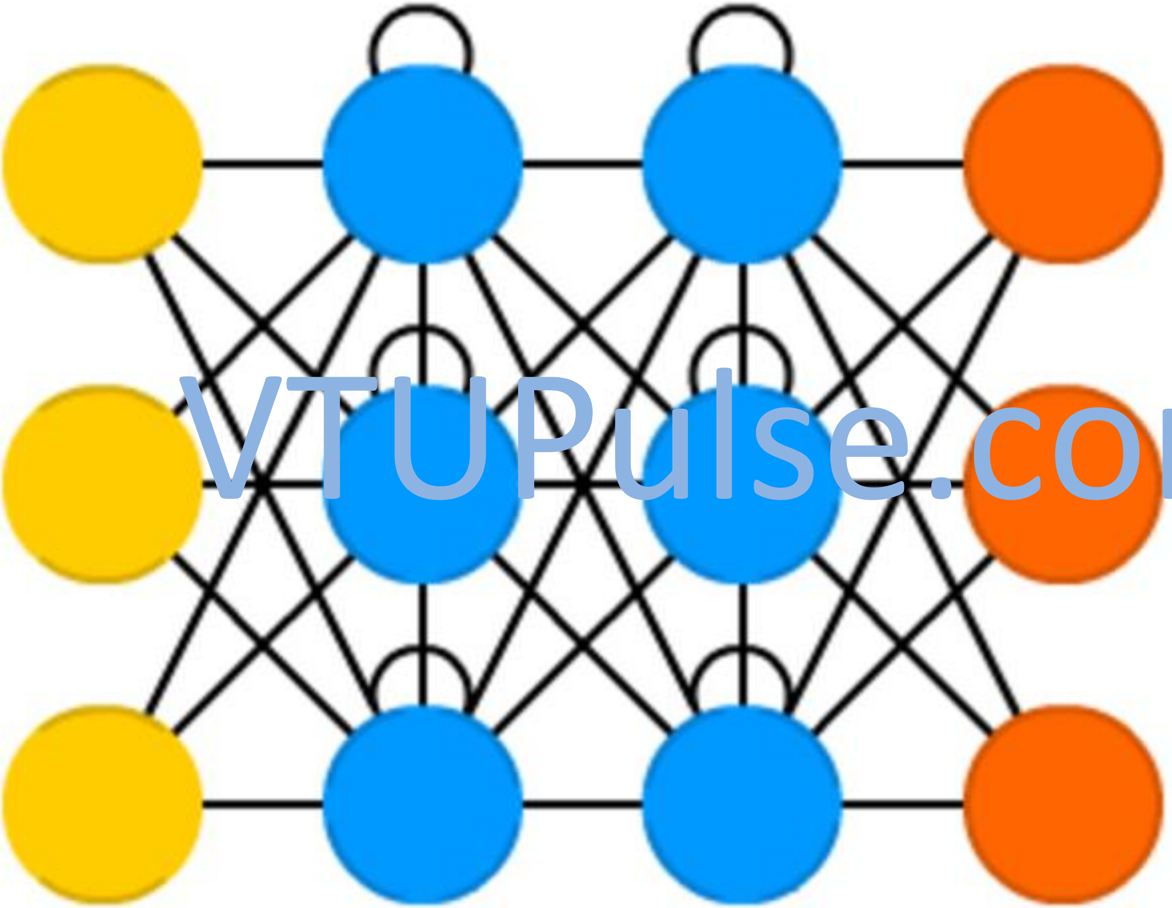
Feed Forward (FF)



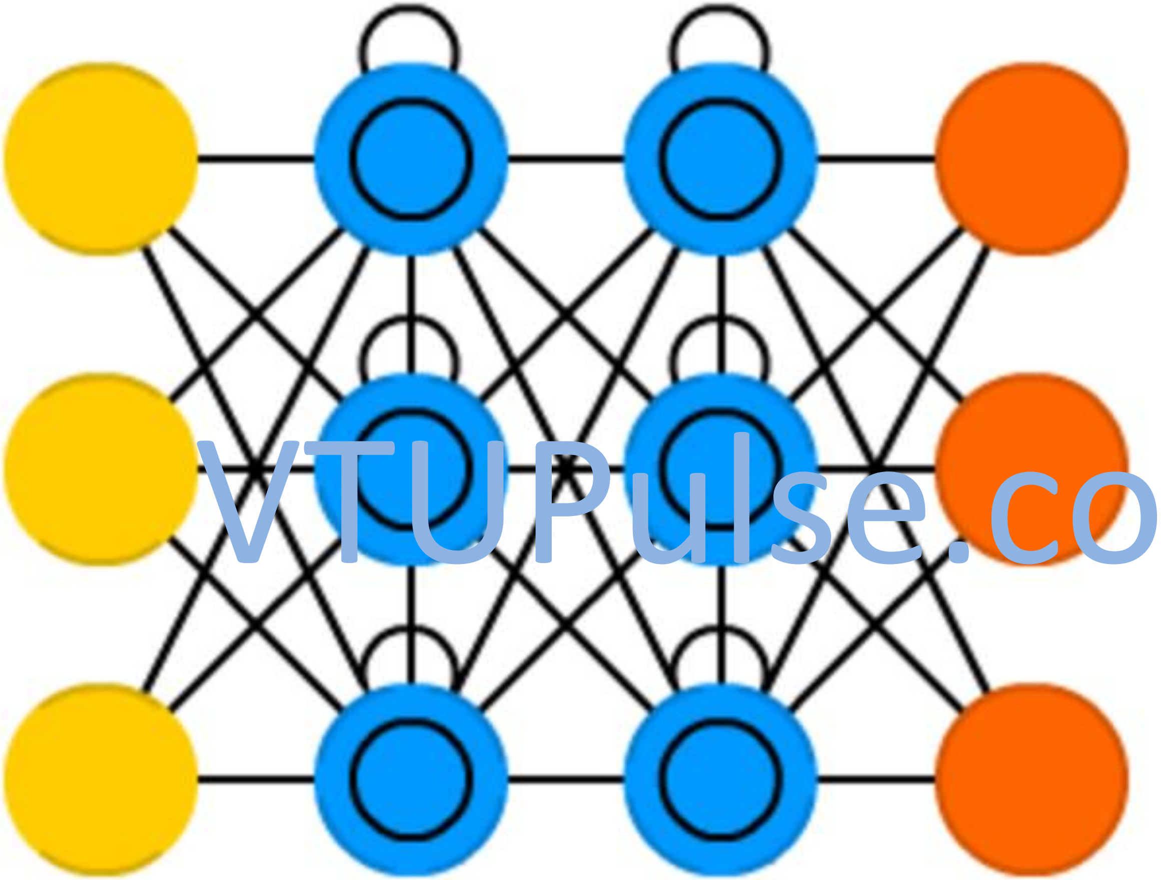
Deep Feed Forward (DFF)



Recurrent Neural Network (RNN)



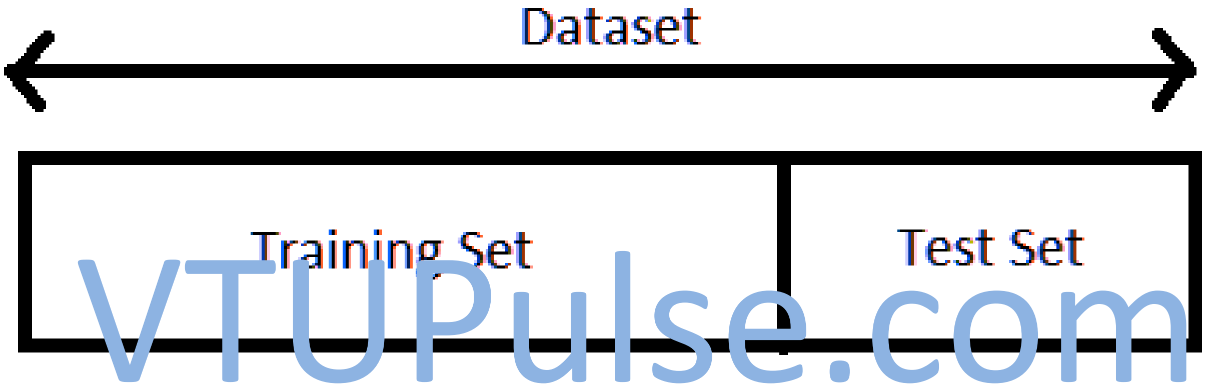
Long / Short Term Memory (LSTM)



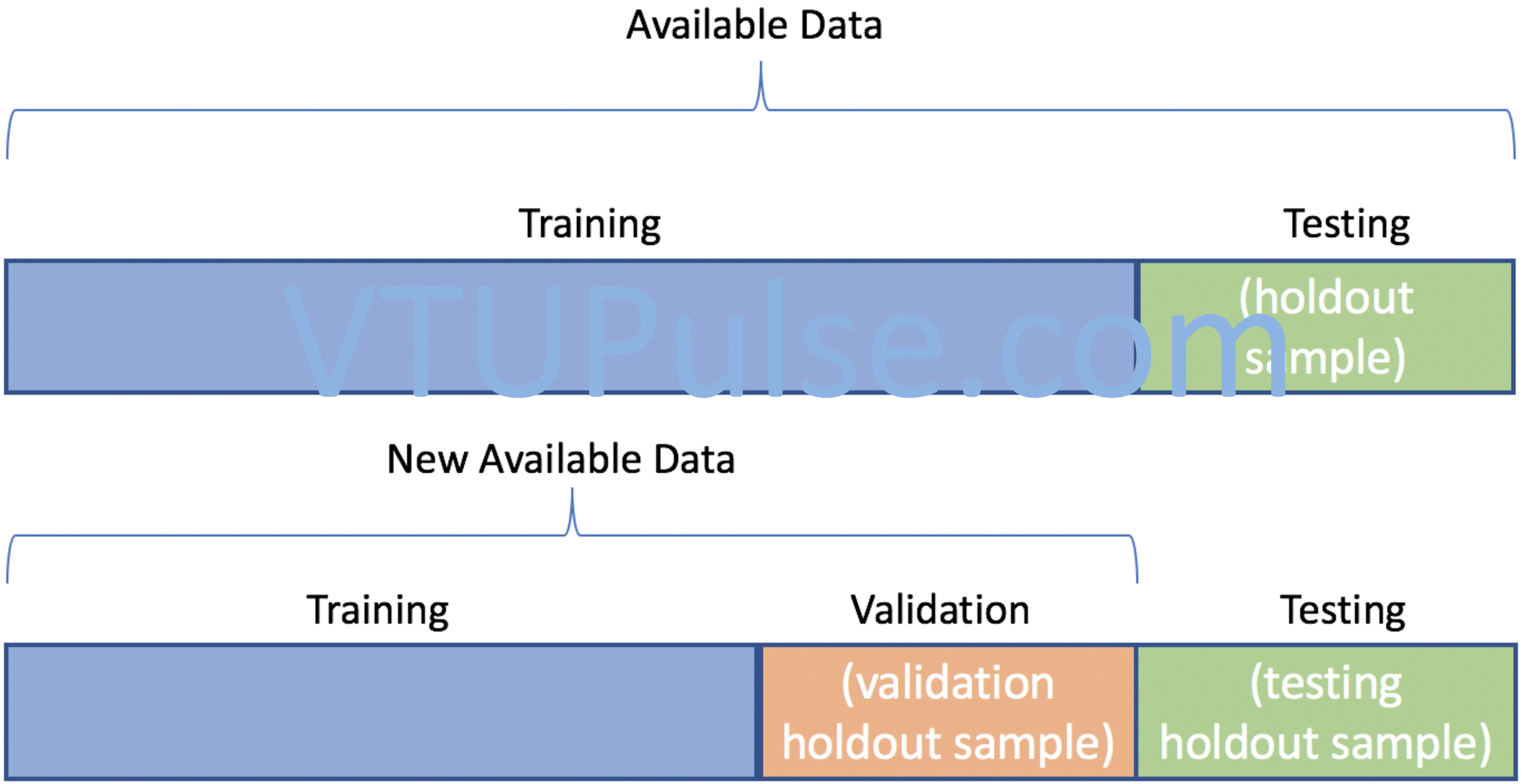
Training an ANN: Training data is split into three parts

Training set	This data set is used to adjust the weights on the neural network (~ 60%).
Validation set	This data set is used to minimize overfitting and verifying accuracy (~ 20%).
Testing set	This data set is used only for testing the final solution in order to confirm the actual predictive power of the network (~ 20%).
k-fold cross-validation	approach means that the data is divided into k equal pieces, and the learning process is repeated k-times with each pieces becoming the training set. This process leads to less bias and more accuracy, but is more time consuming.

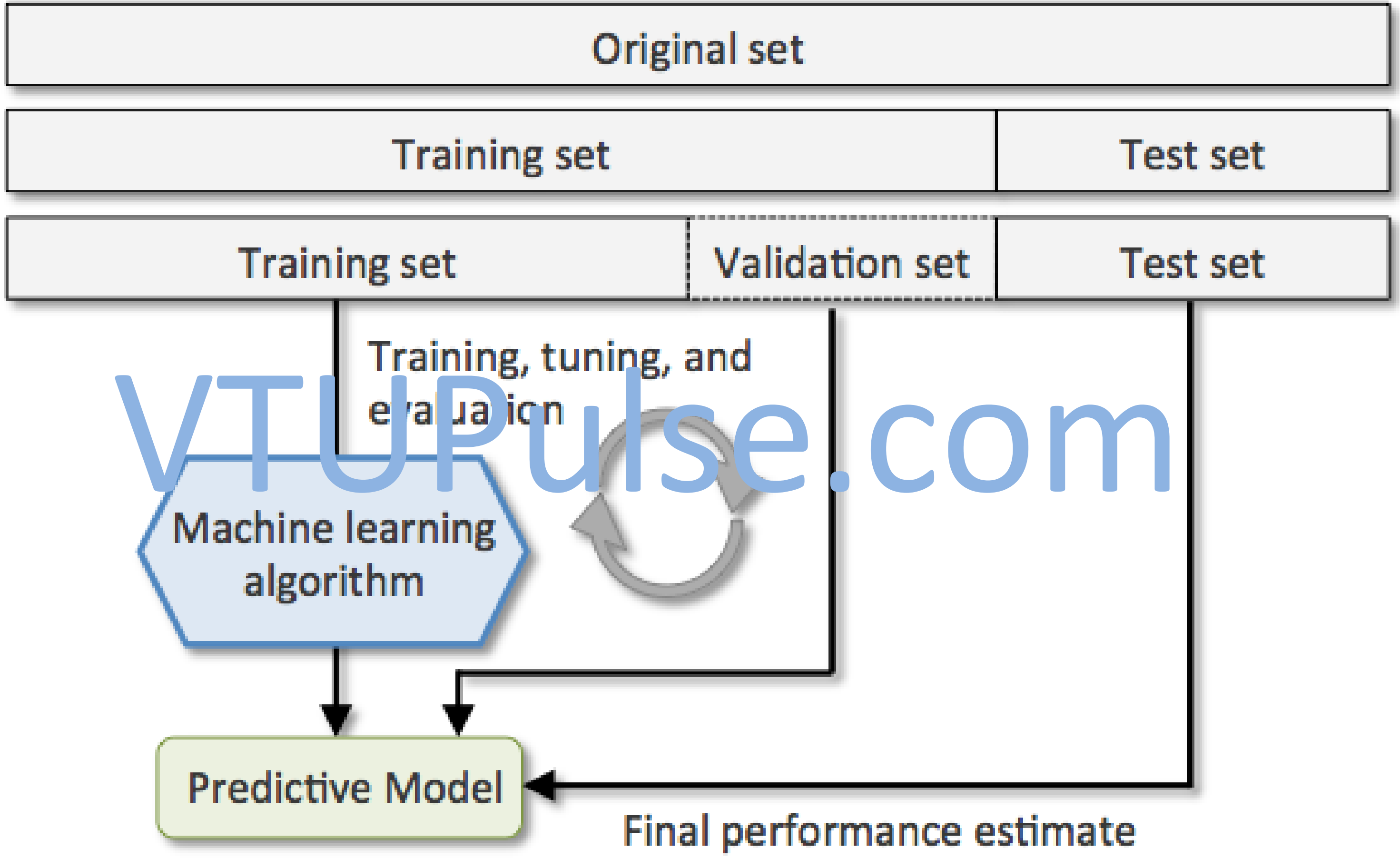
Train Test Distribution



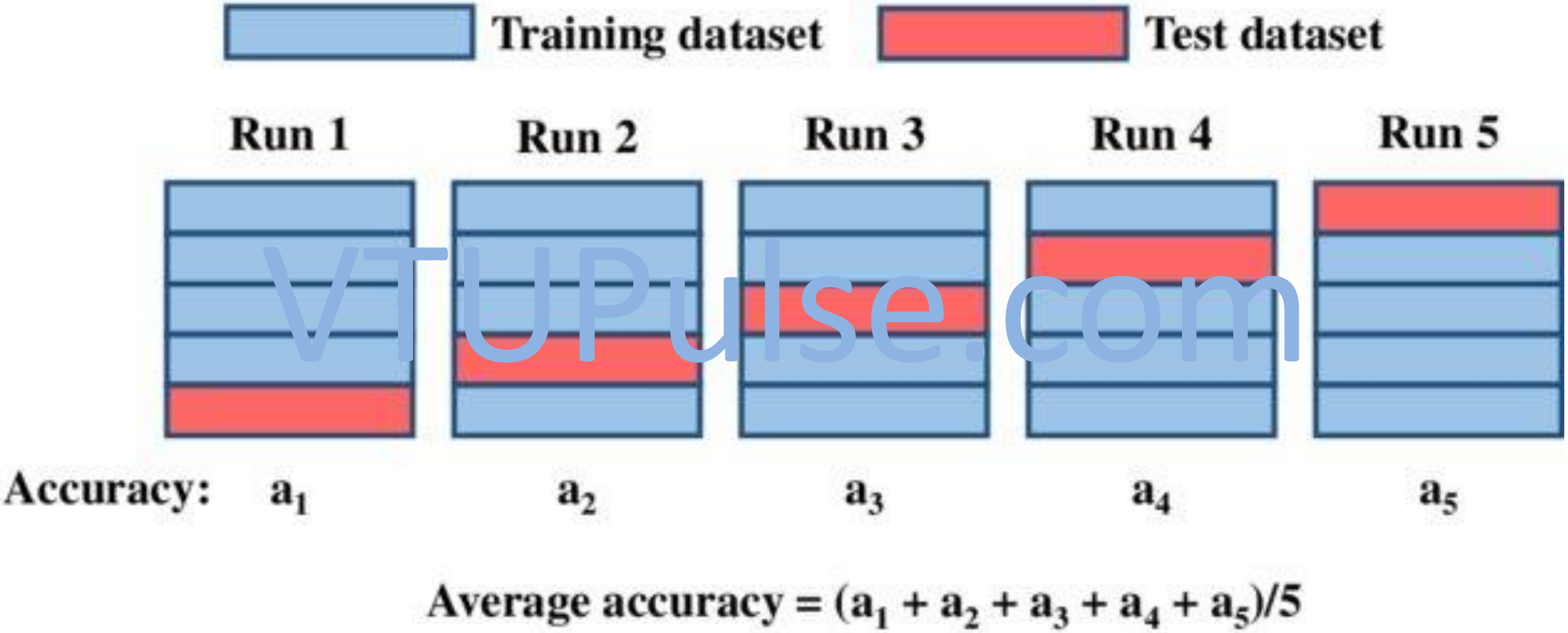
Train Test Validation



Train Test Validation


Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

K-Fold Cross Validation



Advantages of Using ANNs

1. ANNs impose very little restrictions on their use.
2. There is no need to program ANN neural networks, as they learn from examples. They get better with use, without much programing effort.
3. ANN can handle a variety of problem types, including classification, clustering, associations, and so on.
4. ANNs are tolerant of data quality issues, and they do not restrict the data to follow strict normality and/or independence assumptions.
5. ANN can handle both numerical and categorical variables.
6. ANNs can be much faster than other techniques.
7. Most importantly, ANN usually provide better results (prediction and/or clustering) compared to statistical counterparts, once they have been trained enough.

Disadvantages of Using ANNs

1. They are deemed to be black-box solutions, lacking explainability.
2. Optimal design of ANN is still an art: It requires expertise and extensive experimentation.
3. It can be difficult to handle a large number of variables (especially the rich nominal attributes) with an ANN.
4. It takes large data sets to train an ANN.

Association Rule Mining

- Association rule mining is a popular, unsupervised learning technique, used in business to help identify shopping patterns.
- It is also known as market basket analysis. It helps find interesting relationships (affinities) between variables (items or events).
- Thus, it can help cross-sell related items and increase the size of a sale.

Association Rule Mining

- All data used in this technique is categorical.
- There is no dependent variable. It uses machine-learning algorithms. The fascinating “relationship between sales of diapers and beers” is how it is often explained in popular literature.
- This technique accepts as input the raw point-of-sale transaction data.
- The output produced is the description of the most frequent affinities among items.
- An example of an association rule would be, “a Customer who bought a laptop computer and virus protection software also bought an extended service plan 70 percent of the time.”

Business Applications of Association Rules

- In business environments a pattern or knowledge can be used for many purposes.
- In sales and marketing, it is used for e-commerce site design, online advertising optimization, product pricing, and sales/promotion configurations. This analysis can suggest not to put one item on sale at a time, and instead to create a bundle of products promoted as a package to sell other non-selling items.
- In retail environments, it can be used for store design. Strongly associated items can be kept close tougher for customer convenience. Or they could be placed far from each other so that the customer has to walk the aisles and by doing so is potentially exposed to other items.
- In medicine, this technique can be used for relationships between symptoms and illnesses; diagnosis and patient characteristics / treatments; genes and their functions; and so on.

Representing Association Rules

- A generic rule is represented between a set X and Y: X ⇒ Y [S%, C%]
- **X, Y**: products and/or services
- **X:** Left-hand-side (LHS or Antecedent)
- **Y:** Right-hand-side (RHS or Consequent)
- **S:** Support: how often **X** and **Y** go together in the total transaction set
- **C:** Confidence: how often **Y** goes together with **X**
- *Example*: {Laptop Computer, Antivirus Software} ⇒ {Extended Service Plan} [30%, 70%]

Representing Association Rules

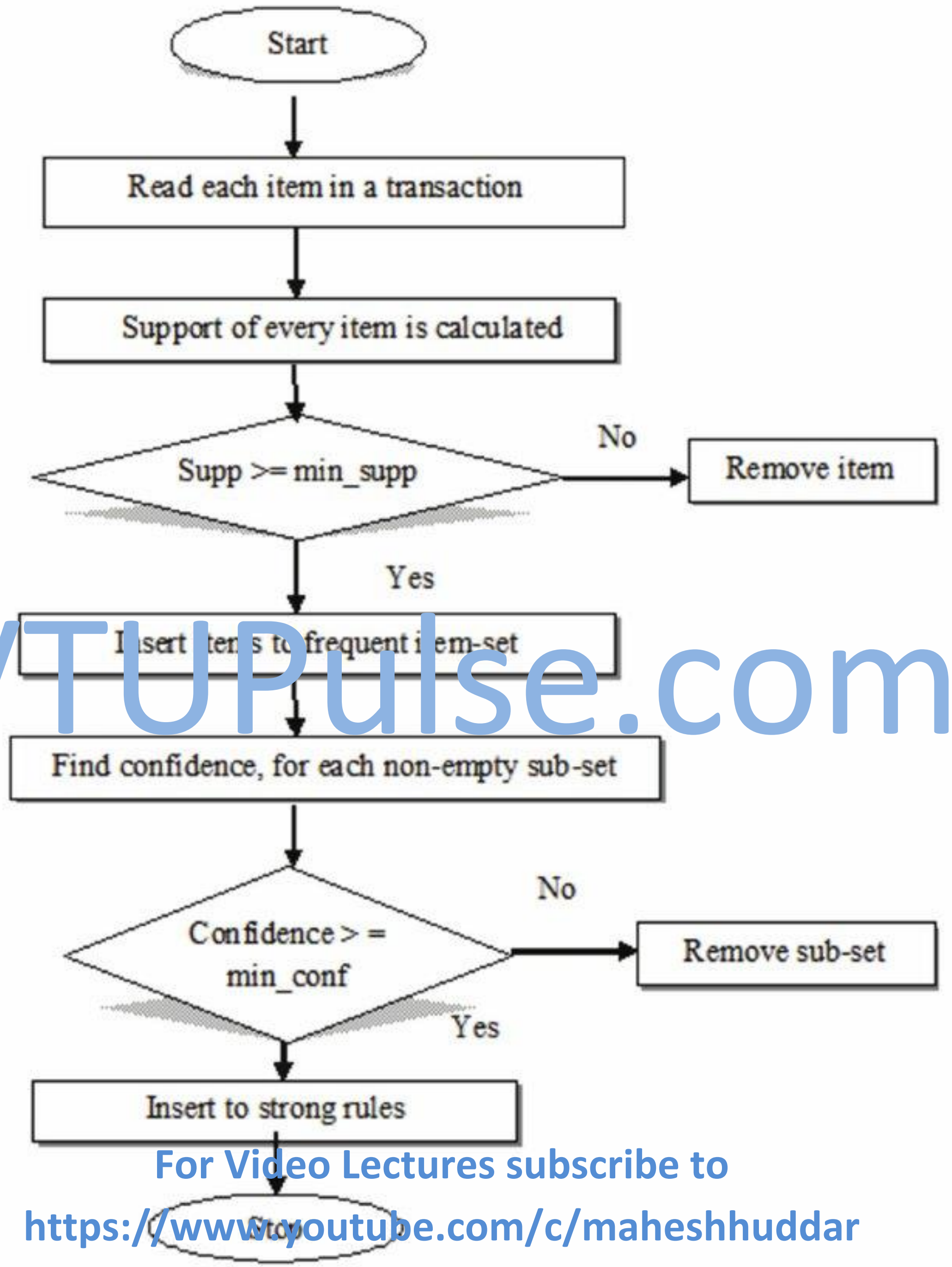
$$\begin{tikzcd}
 & Support = \frac{frq(X,Y)}{N} \\
Rule : X \Rightarrow Y \arrow[ur] \arrow[dr] & \\
 & Confidence = \frac{frq(X,Y)}{frq(X)}
\end{tikzcd}$$

Algorithms for Association Rule

- Not all association rules are interesting and useful, only those that are strong rules and also those that occur frequently.
- In association rule mining, the goal is to find all rules that satisfy the user-specified *minimum support* and *minimum confidence*.
- The resulting sets of rules are all the same irrespective of the algorithm used, that is, given a transaction data set T, a minimum support and a minimum confidence, the set of association rules existing in T is *uniquely determined*.
- The most popular algorithms are Apriori, Eclat, and FP-growth, along with various derivatives and hybrids of the three.

Apriori Algorithm

- This is the most popular algorithm used for association rule mining.
- A frequent itemset is an itemset whose support is greater than or equal to minimum support threshold.
- The Apriori property is a downward closure property, which means that any subsets of a frequent itemset are also frequent itemsets.
- Thus, if (A,B,C,D) is a frequent itemset, then any subset such as (A,B,C) or (B,D) are also frequent itemsets.
- This uses a bottom-up approach; and the size of frequent subsets is gradually increased, from one-item subsets to two-item subsets, then three-item subsets, and so on.
- Groups of candidates at each level are tested against the data for minimum support.


Start
Read each item in a transaction
Support of every item is calculated
Supp >= min_supp
No
Remove item
Yes
Insert items to frequent item-set
Find confidence, for each non-empty sub-set
Confidence > = min_conf
No
Remove sub-set
Yes
Insert to strong rules
Stop

Association Rules Exercise

- Here are a dozen sales transactions.
- There are six products being sold: Milk, Bread, Butter, Eggs, Cookies, and Ketchup.
- Transaction #1 sold Milk, Eggs, Bread, and Butter. Transaction #2 sold Milk, Butter, Egg, and Ketchup. And so on.
- The objective is to use this transaction data to find affinities between products, that is, which products sell together often.
- The support level will be set at 33 percent; the confidence level will be set at 50 percent. That means that we have decided to consider rules from only those itemsets that occur at least 33 percent of the time in the total set of transactions.
- Confidence level means that within those itemsets, the rules of the form X → Y should be such that there is at least 50 percent chance of Y occurring based on X occurring.

Association Rules Exercise

Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

Association Rules Exercise

- First step is to compute 1-item itemsets, that is, how often any product sells.

1-item Sets	Freq
Milk	9
Bread	10
Butter	10
Egg	3
Ketchup	3
Cookies	5

- If itemsets that occur 4 or more times out of 12 are selected, which corresponds to meeting a minimum support level of 33 percent (4 out of 12). Only 4 items make the cut.

Association Rules Exercise

Frequent 1-item Sets	Freq
Milk	9
Bread	10
Butter	10
Cookies	5

Association Rules Exercise

- The second step is to go for the next level of itemsets using items selected earlier: 2-item itemsets.

2-item Sets	Freq
Milk, Bread	7
Milk, Butter	7
Milk, Cookies	3
Bread, Butter	9
Butter, Cookies	3
Bread, Cookies	4

Association Rules Exercise

- Thus, (Milk, Bread) sell 7 times out of 12. (Milk, Butter) sell together 7 times, (Bread, Butter) sell together 9 times, and (Bread, Cookies) sell 4 times. However, only 4 of these transactions meet the minimum support level of 33 percent.

Frequent 2-item Sets	Freq
Milk, Bread	7
Milk, Butter	7
Bread, Butter	9
Bread, Cookies	4

Association Rules Exercise

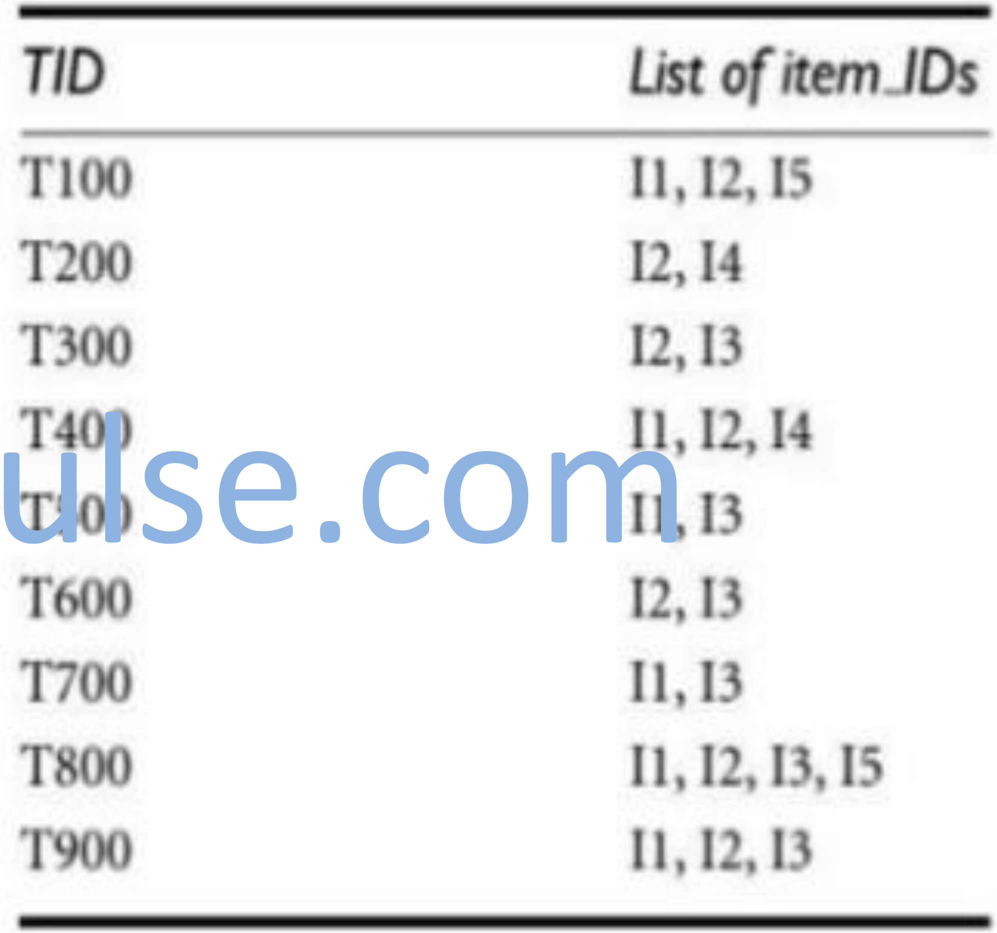
- The next step is to go for the next higher level of itemsets: 3-item itemsets.

3-item Sets	Freq
Milk, Bread, Butter	6
Milk, Bread, Cookies	1
Bread, Butter, Cookies	3
Milk, Butter, Cookies	2

- Again, only a subset of them meets the minimum support requirements.

Frequent 3-item Sets	Freq
Milk, Bread, Butter	6

- Consider the following transaction database:
- min_sup=2



TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3