

The background is a light blue gradient. In the top-left and bottom-right corners, there are several realistic water droplets of various sizes, some overlapping. The text is centered in the middle of the slide.

UNIT-5

TRANSFORM SUPERSTEP

UNIVARIATE ANALYSIS

- UNIVARIATE ANALYSIS IS THE SIMPLEST FORM OF ANALYZING DATA. UNI MEANS “ONE,” SO YOUR DATA HAS ONLY ONE VARIABLE.
- IT DOESN'T DEAL WITH CAUSES OR RELATIONSHIPS, AND ITS MAIN PURPOSE IS TO DESCRIBE THE DATA AND FIND PATTERNS THAT EXIST WITHIN IT.
- IT TAKES DATA, SUMMARIZES THAT DATA, AND FINDS PATTERNS IN THE DATA.

**Heights
(in cm)**

164

167.3

170

174.2

178

180

186

UNIVARIATE ANALYSIS

- THE PATTERNS FOUND IN UNIVARIATE DATA INCLUDE CENTRAL TENDENCY (MEAN, MODE, AND MEDIAN) AND DISPERSION, RANGE, VARIANCE, MAXIMUM, MINIMUM, QUARTILES (INCLUDING THE INTERQUARTILE RANGE), AND STANDARD DEVIATION.
- YOU CAN USE FREQUENCY DISTRIBUTION TABLES, FREQUENCY POLYGONS, HISTOGRAMS, BAR CHARTS, OR PIE CHARTS FOR DESCRIBING DATA USING A UNIVARIATE APPROACH.

BIVARIATE ANALYSIS

- THIS TYPE OF DATA INVOLVES TWO DIFFERENT VARIABLES. THE ANALYSIS OF THIS TYPE OF DATA DEALS WITH CAUSES AND RELATIONSHIPS AND THE ANALYSIS IS DONE TO FIND OUT THE RELATIONSHIP AMONG THE TWO VARIABLES. EXAMPLE OF BIVARIATE DATA CAN BE TEMPERATURE AND ICE CREAM SALES IN SUMMER SEASON.

TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

MULTIVARIATE ANALYSIS

- WHEN THE DATA INVOLVES THREE OR MORE VARIABLES, IT IS CATEGORIZED UNDER MULTIVARIATE.
- EXAMPLE OF THIS TYPE OF DATA IS SUPPOSE AN ADVERTISER WANTS TO COMPARE THE POPULARITY OF FOUR ADVERTISEMENTS ON A WEBSITE, THEN THEIR CLICK RATES COULD BE MEASURED FOR BOTH MEN AND WOMEN AND RELATIONSHIPS BETWEEN VARIABLES CAN THEN BE EXAMINED.

MULTIVARIATE ANALYSIS

- IT IS SIMILAR TO BIVARIATE BUT CONTAINS MORE THAN ONE DEPENDENT VARIABLE.
- SOME OF THE TECHNIQUES ARE REGRESSION ANALYSIS, PATH ANALYSIS, FACTOR ANALYSIS AND MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA).

LINEAR REGRESSION

- Linear regression is a type of statistical analysis used to predict the relationship between two variables. it assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship.
- The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

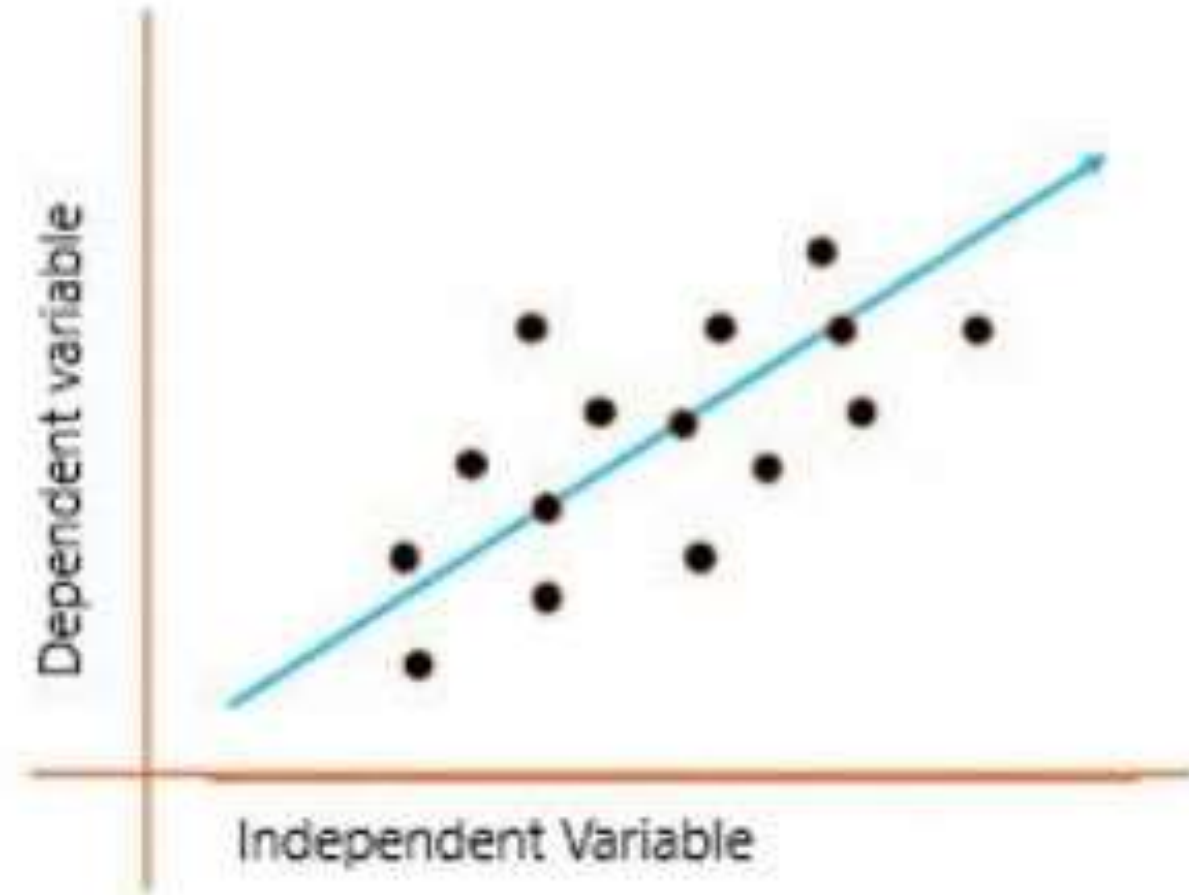



REAL-LIFE APPLICATIONS OF LINEAR REGRESSION

- BUSINESS
 - DEMAND FORECASTING
 - MEDICAL
- 

SIMPLE LINEAR REGRESSION


- In a simple linear regression, there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables.
- Linear regression shows the linear relationship between the independent(predictor) variable i.e. x-axis and the dependent(output) variable i.e. y-axis, called linear regression.





The graph above presents the linear relationship between the output(y) and predictor(X) variables.

The blue line is referred to as the *best-fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.



EQUATION FOR THE LINEAR REGRESSION LINE

- $Y = A + BX$, WHERE, X = EXPLANATORY VARIABLE
- Y = DEPENDENT VARIABLE
- B = SLOPE OF THE LINE
- A = INTERCEPT (THE VALUE OF Y WHEN $X = 0$)

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

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

Linear component

Random Error component

The diagram illustrates the components of a simple linear regression model. The equation $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ is centered on a white background. Arrows point from descriptive labels to each term: 'Dependent Variable' points to Y_i , 'Population Y intercept' points to β_0 , 'Population Slope Coefficient' points to β_1 , 'Independent Variable' points to X_i , and 'Random Error term' points to ε_i . Below the equation, two blue curly braces group the terms into 'Linear component' (under $\beta_0 + \beta_1 X_i$) and 'Random Error component' (under ε_i).

RANSAC LINEAR REGRESSION

- RANSAC IS AN ACRONYM FOR RANDOM SAMPLE CONSENSUS. WHAT THIS ALGORITHM DOES IS FIT A REGRESSION MODEL ON A SUBSET OF DATA THAT THE ALGORITHM JUDGES AS INLIERS WHILE REMOVING OUTLIERS. THIS NATURALLY IMPROVES THE FIT OF THE MODEL DUE TO THE REMOVAL OF SOME DATA POINTS.

RANSAC LINEAR REGRESSION

- AN ADVANTAGE OF RANSAC IS ITS ABILITY TO DO ROBUST ESTIMATION OF THE MODEL PARAMETERS, I.E., IT CAN ESTIMATE THE PARAMETERS WITH A HIGH DEGREE OF ACCURACY, EVEN WHEN A SIGNIFICANT NUMBER OF OUTLIERS IS PRESENT IN THE DATA SET.

HOUGH TRANSFORM

- THE HOUGH TRANSFORM IS A MATHEMATICAL TECHNIQUE USED IN COMPUTER VISION AND IMAGE ANALYSIS TO DETECT SIMPLE GEOMETRIC SHAPES LIKE **LINES, CIRCLES, AND ELLIPSES**.
- THE BASIC IDEA BEHIND THE HOUGH TRANSFORM IS TO REPRESENT LINES OR CURVES IN AN IMAGE AS POINTS IN A PARAMETER SPACE. THE HOUGH TRANSFORM IS A POWERFUL TOOL FOR LINE AND CURVE DETECTION IN IMAGES.

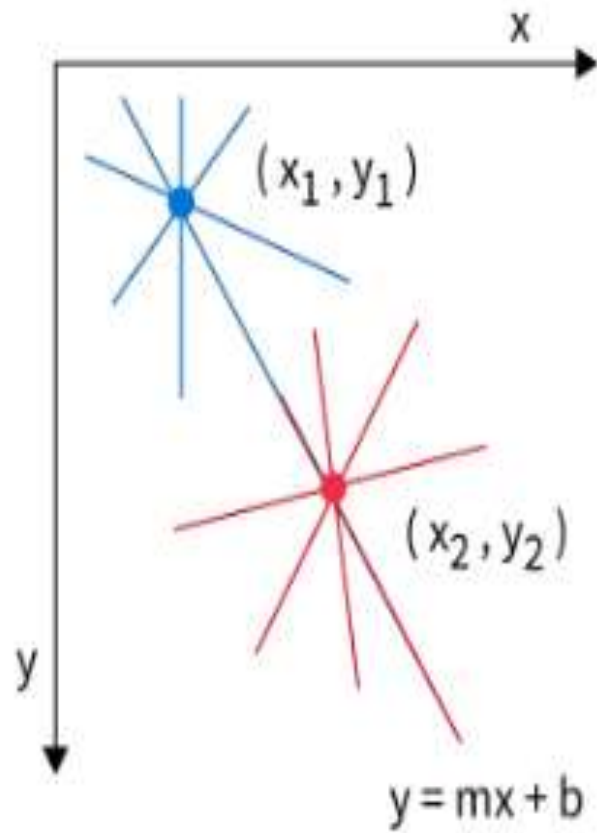
HOUGH TRANSFORM

- ITS ABILITY TO **HANDLE BROKEN AND INCOMPLETE LINES** MAKES IT A VALUABLE ADDITION TO ANY IMAGE PROCESSING TOOLKIT.

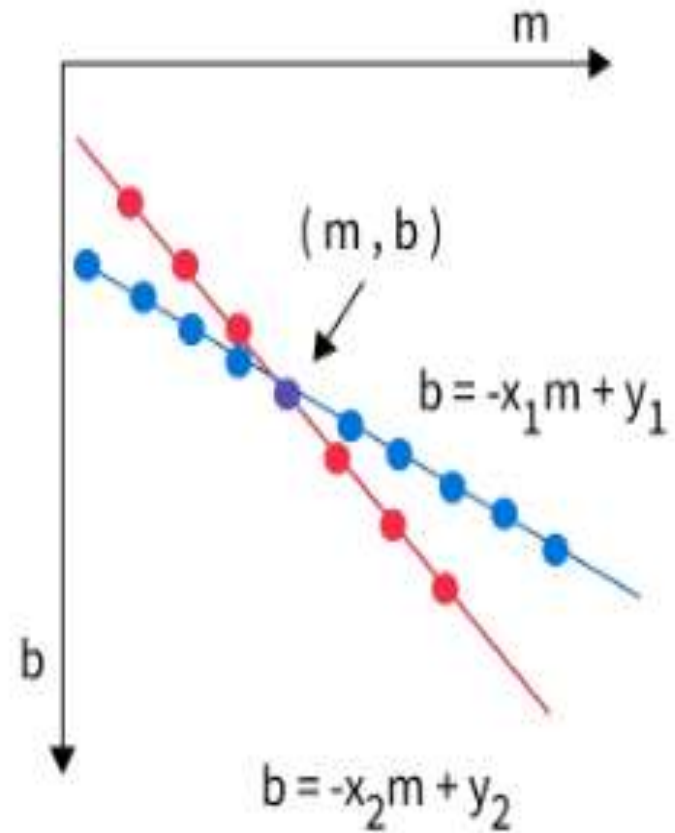
HISTORY

- THE HOUGH TRANSFORM WAS FIRST PROPOSED BY **PAUL HOUGH IN 1962** AS A METHOD FOR DETECTING LINES IN IMAGES. IT WAS LATER EXTENDED TO DETECT OTHER SHAPES LIKE CIRCLES AND ELLIPSES.

Image Space



Parameter Space



WHY IS HOUGH TRANSFORM NEEDED?

- THE HOUGH TRANSFORM CAN DETECT THESE SHAPES BY TRANSFORMING THE IMAGE SPACE INTO A PARAMETER SPACE WHERE THE SHAPES CAN BE MORE EASILY IDENTIFIED.

HOW DOES IT WORK?

- THE HOUGH TRANSFORM IN IMAGE PROCESSING WORKS BY TRANSFORMING THE IMAGE SPACE INTO A PARAMETER SPACE. FOR EXAMPLE, IN THE CASE OF DETECTING LINES IN IMAGES,
- THE IMAGE SPACE IS TRANSFORMED INTO A PARAMETER SPACE CONSISTING OF TWO PARAMETERS: **THE SLOPE AND THE Y-INTERCEPT OF THE LINE.**

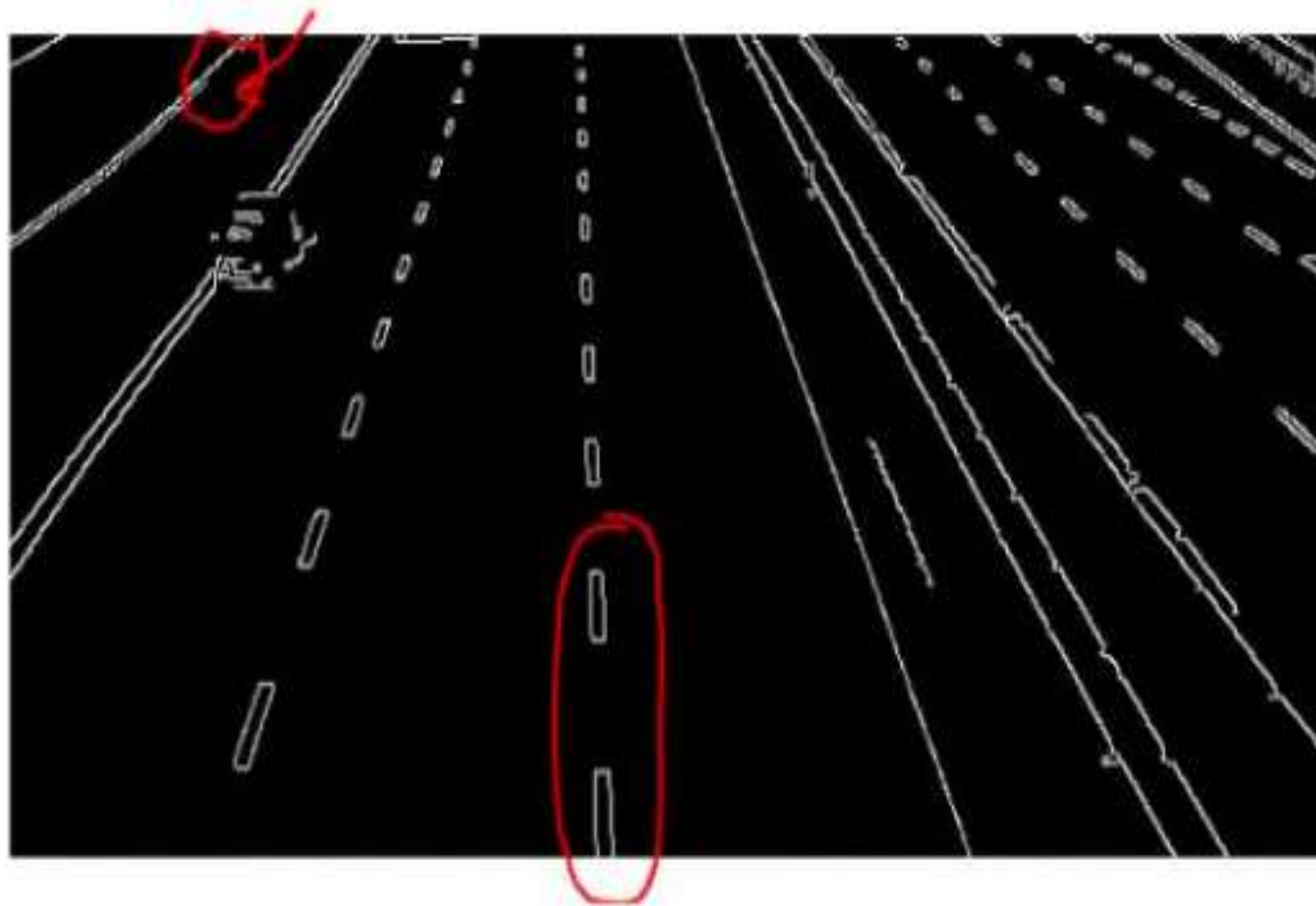
HOW DOES IT WORK?

- EACH PIXEL IN THE IMAGE SPACE IS THEN MAPPED TO A CURVE IN THE PARAMETER SPACE THAT REPRESENTS ALL THE POSSIBLE LINES THAT COULD PASS THROUGH THAT PIXEL. THE CURVES IN THE PARAMETER SPACE ARE THEN ANALYZED TO DETECT THE PRESENCE OF LINES IN THE IMAGE.

Original Image of the lane



Image after applying edge detection technique. Red circles show that the line is breaking there.

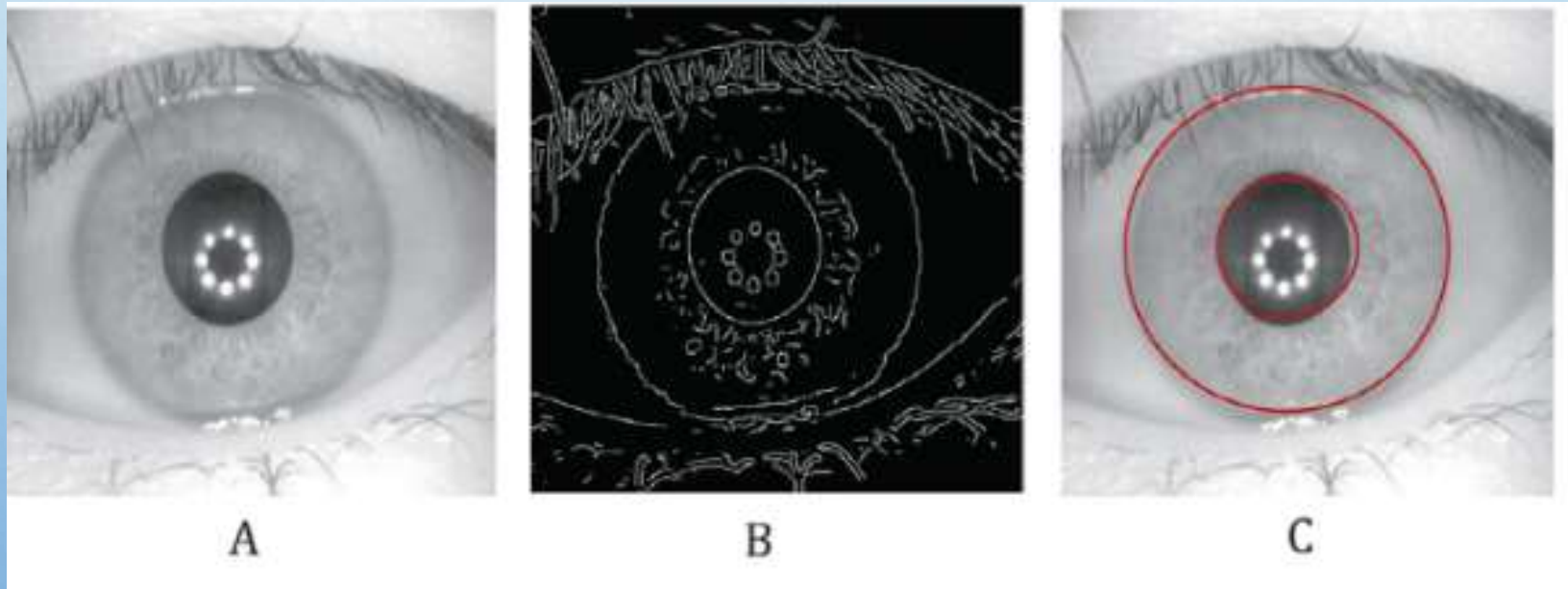


After using Hough Transform



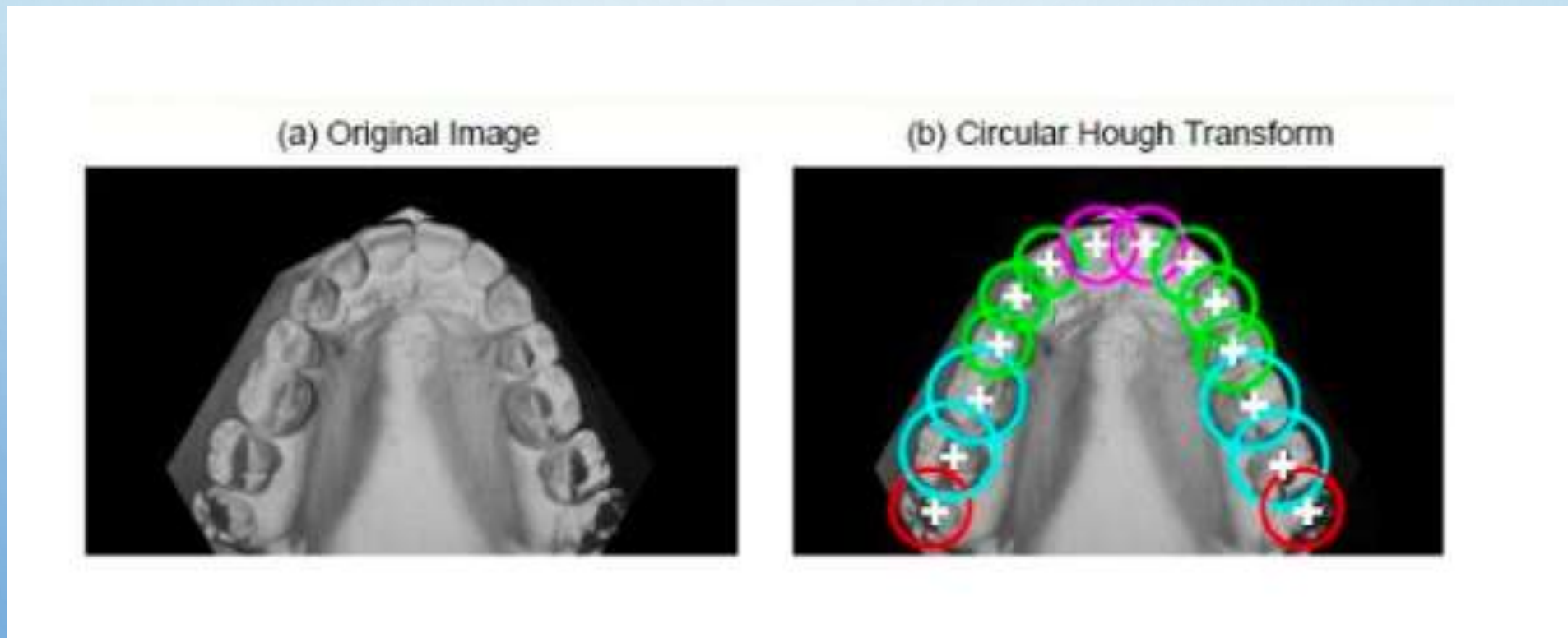
APPLICATIONS

- **BIOMETRIC AND MAN-MACHINE INTERACTION**



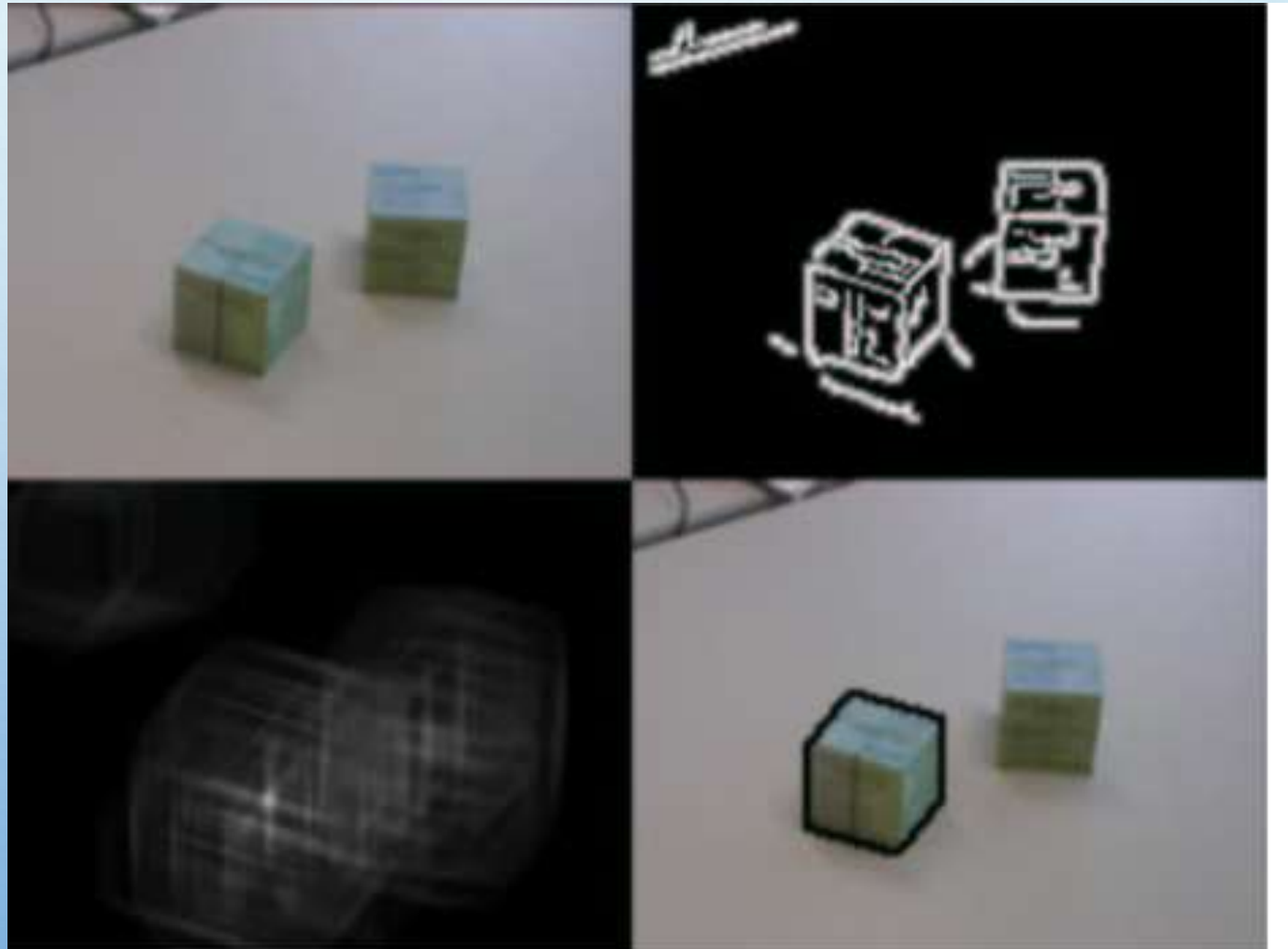
APPLICATIONS

- MEDICAL APPLICATION



APPLICATIONS

- OBJECT RECOGNITION



LOGISTIC REGRESSION

- LOGISTIC REGRESSION IS THE TECHNIQUE TO FIND RELATIONSHIPS BETWEEN A SET OF INPUT VARIABLES AND AN OUTPUT VARIABLE (JUST LIKE ANY REGRESSION), BUT THE OUTPUT VARIABLE, IN THIS CASE, IS A BINARY OUTCOME (THINK OF 0/1 OR YES/NO).

EXAMPLE OF LOGISTIC REGRESSION

- A LOGISTIC REGRESSION MODEL CAN BE BUILT TO DETERMINE IF A PERSON WILL OR WILL NOT PURCHASE A NEW AUTOMOBILE IN THE NEXT 12 MONTHS. THE TRAINING SET COULD INCLUDE INPUT VARIABLES FOR A PERSON'S AGE, INCOME, AND GENDER AS WELL AS THE AGE OF AN EXISTING AUTOMOBILE.



LOGISTIC REGRESSION IN REAL LIFE APPLICATIONS

- MEDICAL
 - FINANCE
 - MARKETING
 - ENGINEERING
- 

SIMPLE LOGISTIC REGRESSION

- SIMPLE LOGISTIC REGRESSION IS ANALOGOUS TO LINEAR REGRESSION, EXCEPT THAT THE DEPENDENT VARIABLE IS NOMINAL, NOT A MEASUREMENT.
- WHEN THE OUTCOME VARIABLE IS CATEGORICAL IN NATURE, LOGISTIC REGRESSION CAN BE USED TO PREDICT THE LIKELIHOOD OF AN OUTCOME BASED ON THE INPUT VARIABLES

EXAMPLE OF SIMPLE LOGISTIC REGRESSION

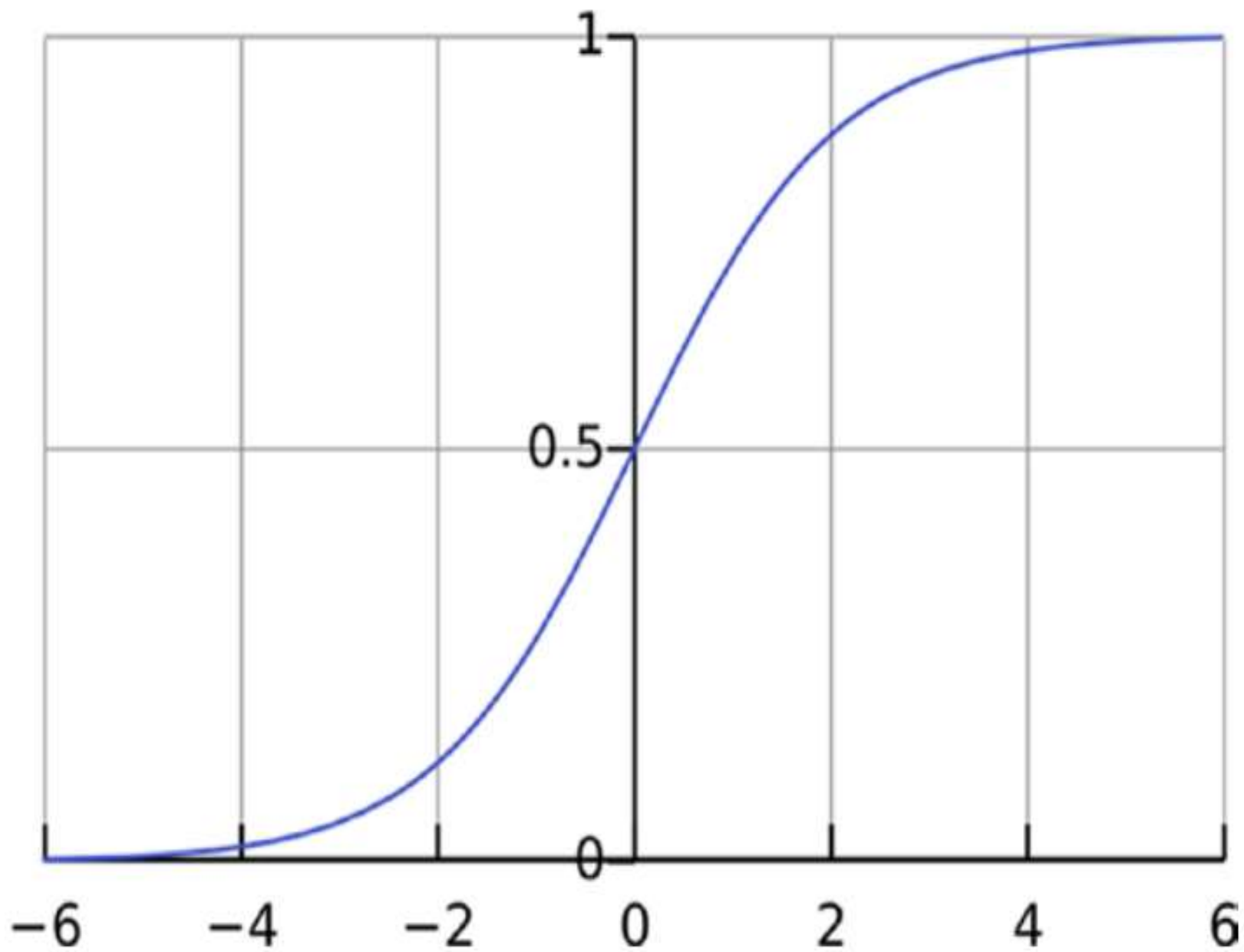


STUDY OF A TRAFFIC JAM AT A CERTAIN LOCATION IN LONDON

- USING A BINARY VARIABLE THE OUTPUT IS A CATEGORICAL: YES OR NO. HENCE, IS THERE A TRAFFIC JAM? YES OR NO?
- THE PROBABILITY OF OCCURRENCE OF TRAFFIC JAMS CAN BE DEPENDENT ON ATTRIBUTES SUCH AS WEATHER CONDITION, DAY OF THE WEEK AND MONTH, TIME OF DAY, NUMBER OF VEHICLES, ETC.

STUDY OF A TRAFFIC JAM AT A CERTAIN LOCATION IN LONDON

- USING LOGISTIC REGRESSION, YOU CAN FIND THE BEST-FITTING MODEL THAT EXPLAINS THE RELATIONSHIP BETWEEN INDEPENDENT ATTRIBUTES AND TRAFFIC JAM OCCURRENCE RATES AND PREDICTS PROBABILITY OF JAM OCCURRENCE.
- THIS PROCESS IS CALLED BINARY LOGISTIC REGRESSION. THE STATE OF THE TRAFFIC CHANGES FOR NO = ZERO TO YES = ONE.



Logistics Regression is based on the logistics function $f(y)$, as given in the equation below,

$$f(y) = \frac{e^y}{1 + e^y} \quad \text{for } -\infty < y < \infty$$

Note that as $y \rightarrow \infty$, $f(y) \rightarrow 1$, and as $y \rightarrow -\infty$, $f(y) \rightarrow 0$.

MULTINOMIAL LOGISTIC REGRESSION

- MULTINOMIAL LOGISTIC REGRESSION (OFTEN JUST CALLED 'MULTINOMIAL REGRESSION') IS USED TO PREDICT A NOMINAL DEPENDENT VARIABLE GIVEN ONE OR MORE INDEPENDENT VARIABLES.
- IT IS SOMETIMES CONSIDERED AN EXTENSION OF BINOMIAL LOGISTIC REGRESSION TO ALLOW FOR A DEPENDENT VARIABLE WITH MORE THAN TWO CATEGORIES.

EXAMPLE OF MULTINOMIAL LOGISTIC REGRESSION



EXAMPLE OF MULTINOMIAL LOGISTIC REGRESSION

Which type of drink consumers prefer based on location in the UK and age (i.e., the dependent variable would be "type of drink", with four categories – Coffee, Soft Drink, Tea and Water – and your independent variables would be the nominal variable, "location in UK", assessed using three categories – London, South UK and North UK – and the continuous variable, "age", measured in years).

ORDINAL LOGISTIC REGRESSION

- ORDINAL LOGISTIC REGRESSION (OFTEN JUST CALLED 'ORDINAL REGRESSION') IS USED TO PREDICT AN ORDINAL DEPENDENT VARIABLE GIVEN ONE OR MORE INDEPENDENT VARIABLES.
- AS WITH OTHER TYPES OF REGRESSION, ORDINAL REGRESSION CAN ALSO USE INTERACTIONS BETWEEN INDEPENDENT VARIABLES TO PREDICT THE DEPENDENT VARIABLE.

EXAMPLE OF ORDINAL LOGISTIC REGRESSION

- FOR EXAMPLE, YOU COULD USE ORDINAL REGRESSION TO PREDICT THE BELIEF THAT "TAX IS TOO HIGH" (YOUR ORDINAL DEPENDENT VARIABLE, MEASURED ON A 4-POINT LIKERT ITEM FROM "STRONGLY DISAGREE" TO "STRONGLY AGREE"), BASED ON TWO INDEPENDENT VARIABLES: "AGE" AND "INCOME".

EXAMPLE OF ORDINAL LOGISTIC REGRESSION

- YOU COULD USE ORDINAL REGRESSION TO DETERMINE WHETHER A NUMBER OF INDEPENDENT VARIABLES, SUCH AS "AGE", "GENDER", "LEVEL OF PHYSICAL ACTIVITY" (AMONGST OTHERS), PREDICT THE ORDINAL DEPENDENT VARIABLE, "OBESITY", WHERE OBESITY IS MEASURED USING THREE ORDERED CATEGORIES: "NORMAL", "OVERWEIGHT" AND "OBESE".

BUSINESS PROBLEM

- RATE A STUDENT ON A SCALE FROM 1 TO 5, TO DETERMINE IF HE OR SHE HAS THE REQUISITE QUALIFICATIONS (“PRESTIGE”) TO JOIN THE UNIVERSITY.
- CRITERIA USED IS PRESTIGE

BUSINESS PROBLEM

```
import sys
import os
import pandas as pd
import statsmodels.api as sm
import pylab as pl
import numpy as np

if sys.platform == 'linux':
    Base=os.path.expanduser('~') + '/VKHCG'
else:
    Base='C:/VKHCG'

Retrieve StudentData.
```



```
sFileName=Base + '/01-Vermeulen/00-RawData/StudentData.csv'
StudentFrame = pd.read_csv(sFileName,header=0)
StudentFrame.columns = ["sname", "gre", "gpa", "prestige","admit","QR","VR",
,"gpatrue"]
StudentSelect=StudentFrame[["admit", "gre", "gpa", "prestige"]]
print('Record select:',StudentSelect.shape[0])
df=StudentSelect
I add the following two lines if you want to speed-up the processing, but
it does make the predictions less accurate. So use it if you want.
#df=StudentSelect.drop_duplicates(subset=None, keep='first', inplace=False)
#print('Records Unique:', df.shape[0])
```


Here are the columns for the data set:

```
print(df.columns)
```

Here is a description of the data profile:

```
print(df.describe())
```

```
print(pd.crosstab(df['admit'], df['prestige'], rownames=['admit']))
```

	prestige			
	1	2	3	4
admit				
0	16810	15810	15810	7905
1	0	1000	1000	500

There is clearly a criterion related to prestige. A student requires a minimum of Prestige = 2.

MINIMUM CRITERIA PRESTIGE=2

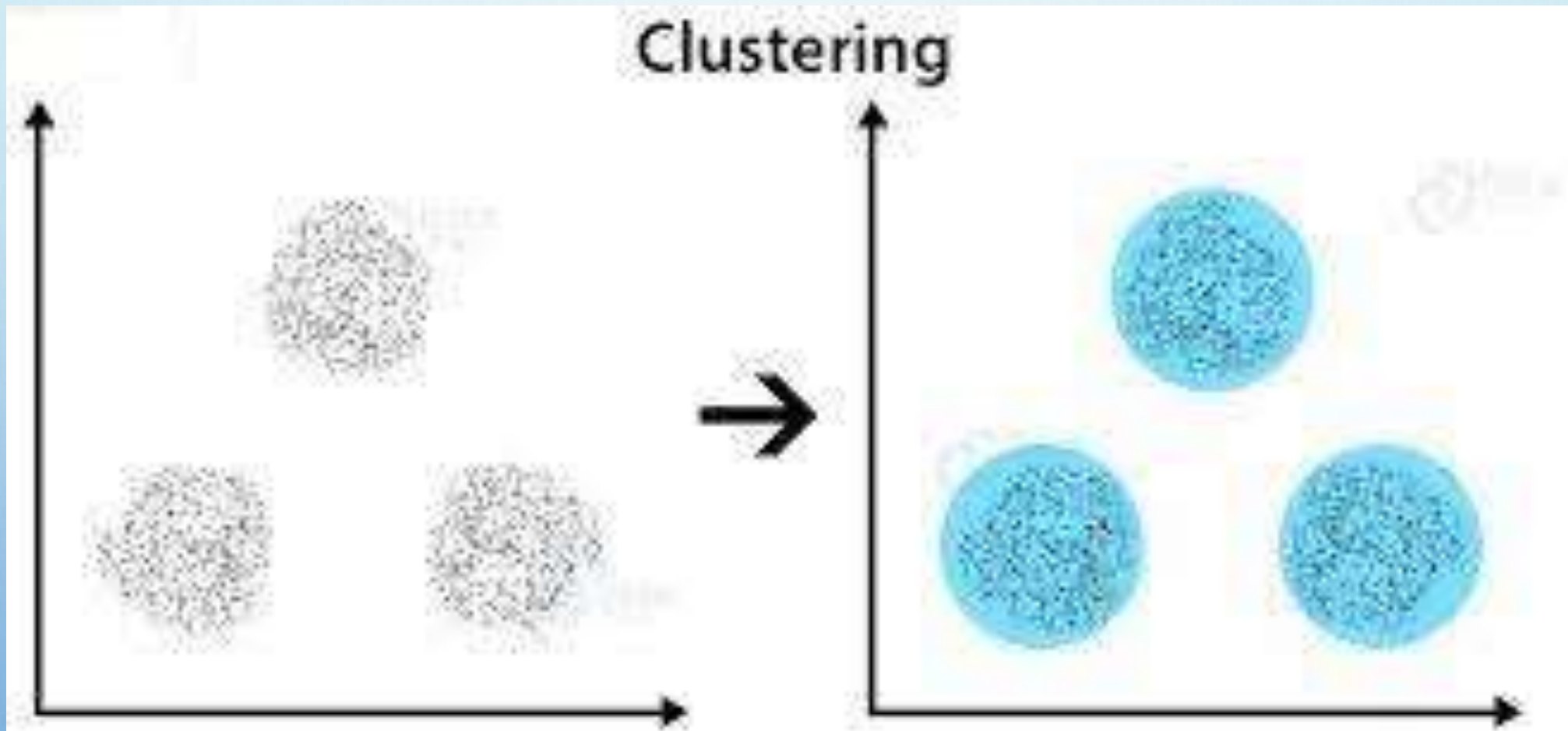
```
df.hist()
```

```
pl.tight_layout()  
pl.show()
```

CLUSTERING TECHNIQUES

- CLUSTERING IS THE USE OF UNSUPERVISED TECHNIQUES FOR GROUPING SIMILAR OBJECTS. IN MACHINE LEARNING, UNSUPERVISED REFERS TO THE PROBLEM OF FINDING HIDDEN STRUCTURE WITHIN UNLABELLED DATA.
- IN CLUSTERING, THERE ARE NO PREDICTIONS MADE. RATHER, CLUSTERING METHODS FIND THE SIMILARITIES BETWEEN OBJECTS ACCORDING TO THE OBJECT ATTRIBUTES AND GROUP THE SIMILAR OBJECTS INTO CLUSTERS. CLUSTERING TECHNIQUES ARE UTILIZED IN MARKETING, ECONOMICS, AND VARIOUS BRANCHES OF SCIENCE.

Clustering

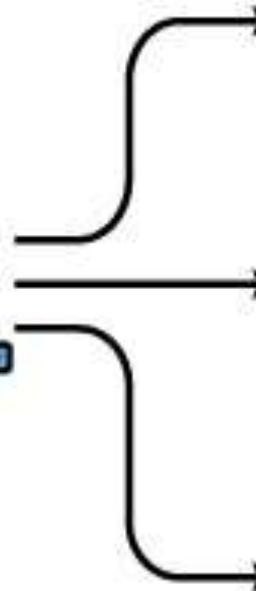




Raw Data



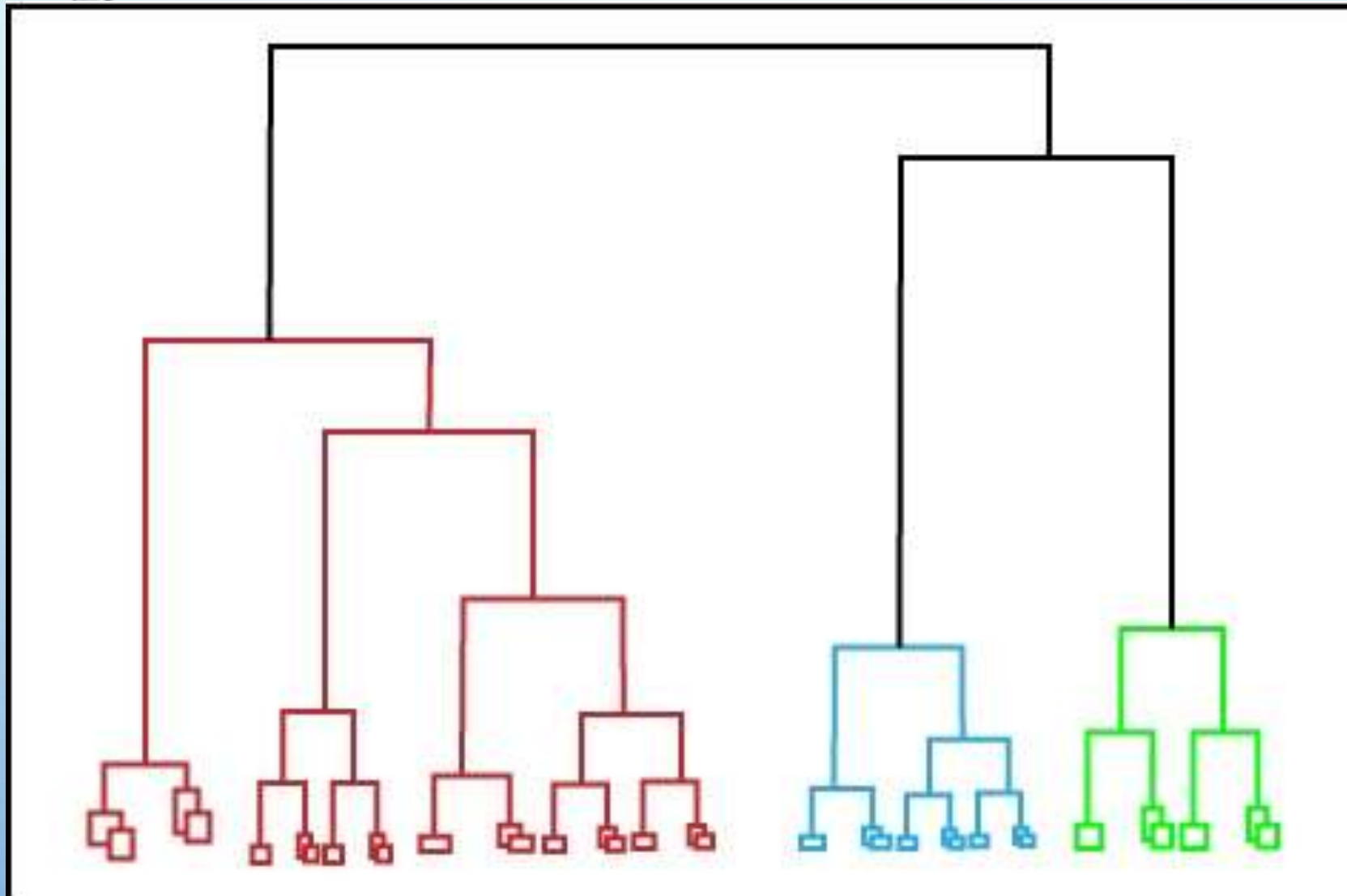
Algorithm



Output

HIERARCHICAL CLUSTERING

- HIERARCHICAL CLUSTERING IS A METHOD OF CLUSTER ANALYSIS WHEREBY YOU BUILD A HIERARCHY OF CLUSTERS. THIS WORKS WELL FOR DATA SETS THAT ARE COMPLEX .
- ALSO CALLED HIERARCHICAL CLUSTER ANALYSIS OR HCA IS AN UNSUPERVISED CLUSTERING ALGORITHM WHICH INVOLVES CREATING CLUSTERS THAT HAVE PREDOMINANT ORDERING FROM TOP TO BOTTOM.



AGGLOMERATIVE CLUSTERING

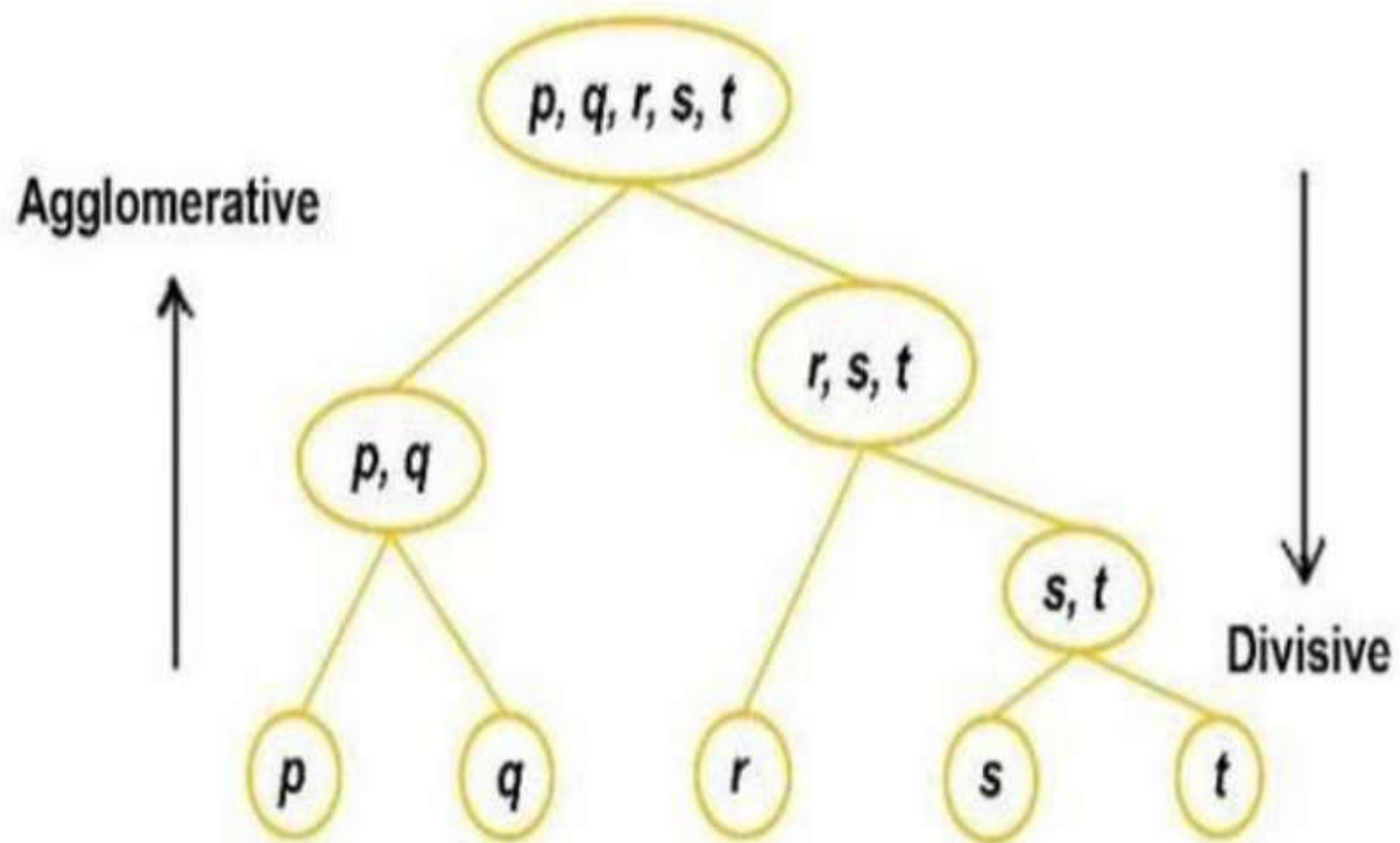
- IT'S ALSO KNOWN AS AGNES (AGGLOMERATIVE NESTING). IT'S A “BOTTOM-UP” APPROACH: EACH OBSERVATION STARTS IN ITS OWN CLUSTER, AND PAIRS OF CLUSTERS ARE MERGED AS ONE MOVES UP THE HIERARCHY.

HOW DOES IT WORK?

1. MAKE EACH DATA POINT A SINGLE-POINT CLUSTER → FORMS N CLUSTERS
2. TAKE THE TWO CLOSEST DATA POINTS AND MAKE THEM ONE CLUSTER → FORMS $N-1$ CLUSTERS
3. TAKE THE TWO CLOSEST CLUSTERS AND MAKE THEM ONE CLUSTER → FORMS $N-2$ CLUSTERS.
4. REPEAT STEP-3 UNTIL YOU ARE LEFT WITH ONLY ONE CLUSTER.

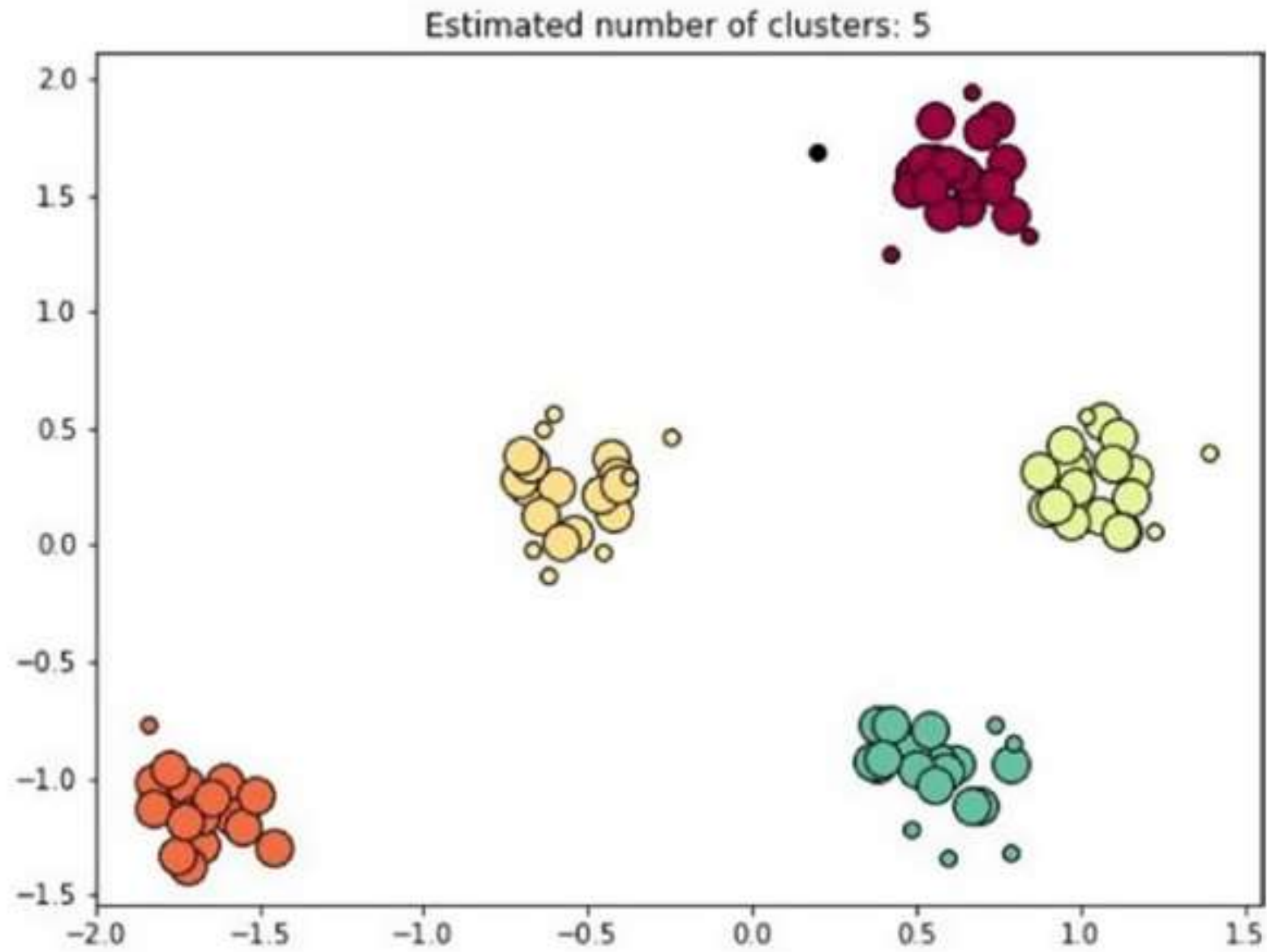
DIVISIVE CLUSTERING

- IN DIVISIVE OR DIANA (DIVISIVE ANALYSIS CLUSTERING) IS A TOP-DOWN CLUSTERING METHOD WHERE WE ASSIGN ALL OF THE OBSERVATIONS TO A SINGLE CLUSTER AND THEN PARTITION THE CLUSTER TO TWO LEAST SIMILAR CLUSTERS.
- FINALLY, WE PROCEED RECURSIVELY ON EACH CLUSTER UNTIL THERE IS ONE CLUSTER FOR EACH OBSERVATION.



PARTITIONAL CLUSTERING

- . PARTITIONAL CLUSTERING DECOMPOSES A DATA SET INTO A SET OF DISJOINT CLUSTERS.
- IT CLASSIFIES THE DATA INTO K GROUPS BY SATISFYING THE FOLLOWING REQUIREMENTS:
- (1) EACH GROUP CONTAINS AT LEAST ONE POINT,
- (2) EACH POINT BELONGS TO EXACTLY ONE GROUP



Partitional clustering

DENSITY-BASED CLUSTERING

- IN DENSITY-BASED CLUSTERING, AN AREA OF HIGHER DENSITY IS SEPARATED FROM THE REMAINDER OF THE DATA SET. DATA ENTRIES IN SPARSE AREAS ARE PLACED IN SEPARATE CLUSTERS. THESE CLUSTERS ARE CONSIDERED TO BE NOISE, OUTLIERS, AND BORDER DATA ENTRIES


ANOVA

- THE ONE-WAY ANALYSIS OF VARIANCE (ANOVA) TEST IS USED TO DETERMINE WHETHER THE MEAN OF MORE THAN TWO GROUPS OF DATA SETS IS SIGNIFICANTLY DIFFERENT FROM EACH DATA SET.



CONTT..

THE CORE OF THIS TECHNIQUE LIES IN ASSESSING WHETHER ALL THE GROUPS ARE IN FACT PART OF ONE LARGER POPULATION OR A COMPLETELY DIFFERENT POPULATION WITH DIFFERENT CHARACTERISTICS.



EXAMPLE OF ANOVA

A BOGOF (BUY-ONE-GET-ONE-FREE) CAMPAIGN IS EXECUTED ON 5 GROUPS OF 100 CUSTOMERS EACH.

EACH GROUP IS DIFFERENT IN TERMS OF ITS DEMOGRAPHIC ATTRIBUTES. WE WOULD LIKE TO DETERMINE WHETHER THESE FIVE RESPOND DIFFERENTLY TO THE CAMPAIGN. THIS WOULD HELP US OPTIMIZE THE RIGHT CAMPAIGN FOR THE RIGHT DEMOGRAPHIC GROUP, INCREASE THE RESPONSE RATE, AND REDUCE THE COST OF THE CAMPAIGN

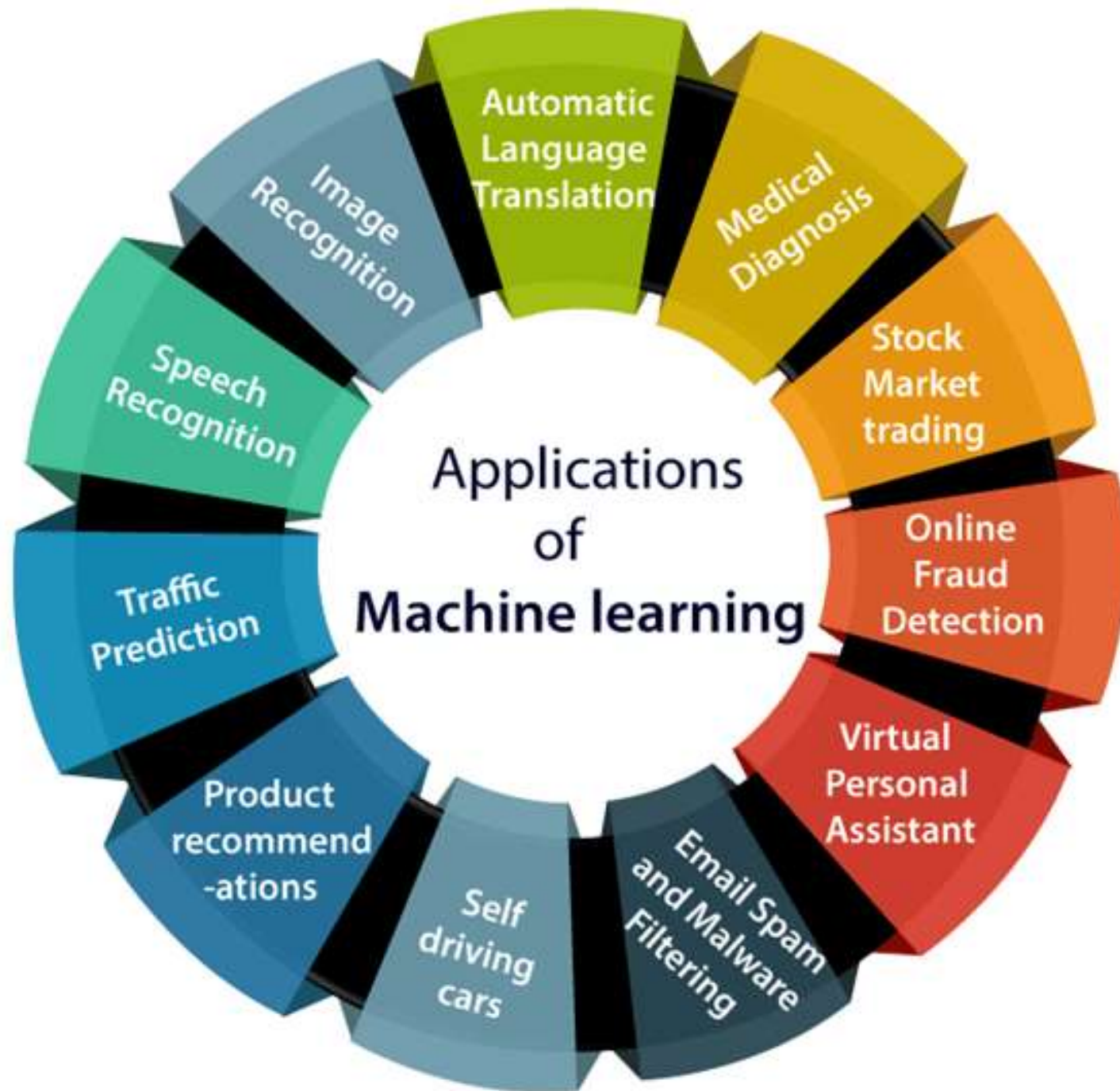
TYPES OF ANOVA


- ONE WAY ANOVA
- WITH A ONE-WAY, YOU HAVE ONE INDEPENDENT VARIABLE AFFECTING A DEPENDENT VARIABLE.
- TWO WAY ANOVA
- WITH A TWO-WAY ANOVA, THERE ARE TWO INDEPENDENTS. FOR EXAMPLE, A TWO-WAY ANOVA ALLOWS A COMPANY TO COMPARE WORKER PRODUCTIVITY BASED ON TWO INDEPENDENT VARIABLES, SUCH AS SALARY AND SKILL SET.

SUPPORT VECTOR MACHINES

INTRODUCTION TO MACHINE LEARNING


- MACHINE LEARNING IS A BRANCH OF ARTIFICIAL INTELLIGENCE (AI) AND COMPUTER SCIENCE WHICH FOCUSES ON THE USE OF DATA AND ALGORITHMS TO IMITATE THE WAY THAT HUMANS LEARN, GRADUALLY IMPROVING ITS ACCURACY.





Supervised learning: Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well-labelled.

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

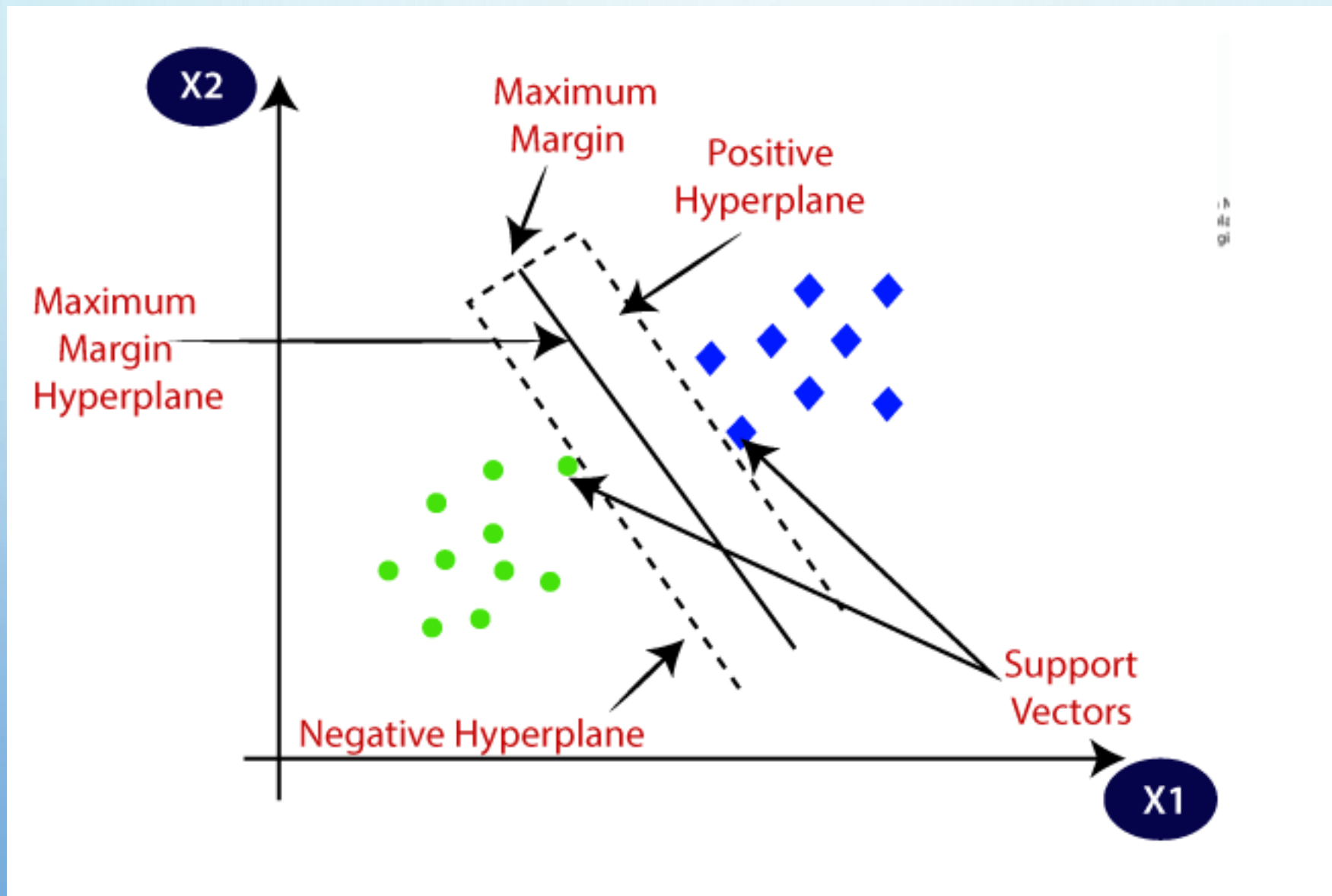


SVM(SUPPORT VECTOR MACHINES)

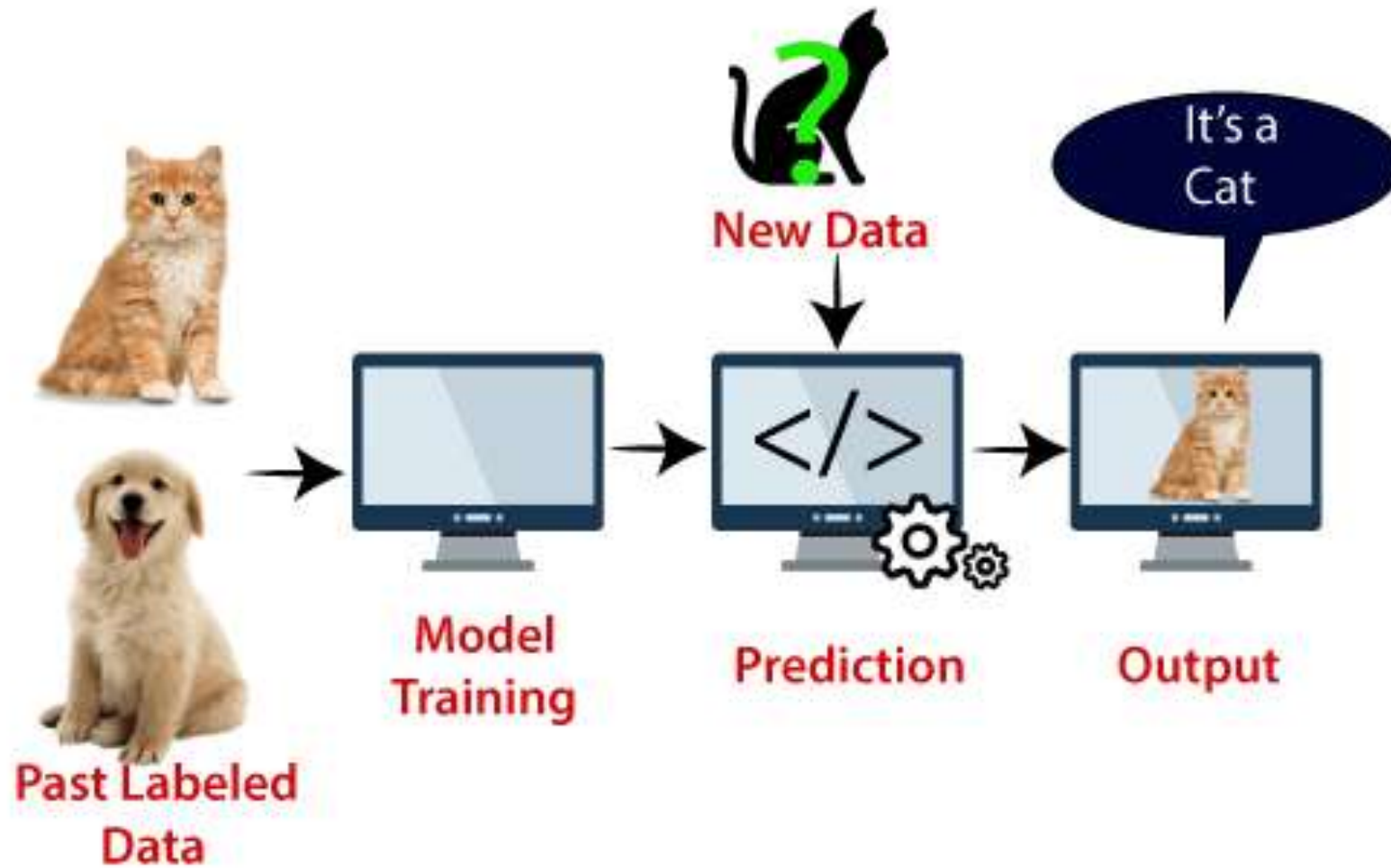
- SUPPORT VECTOR MACHINE OR SVM IS ONE OF THE MOST POPULAR SUPERVISED LEARNING ALGORITHMS, WHICH IS USED FOR CLASSIFICATION AS WELL AS REGRESSION PROBLEMS.

GOAL OF SVM

- THE GOAL OF THE SVM ALGORITHM IS TO CREATE THE BEST LINE OR DECISION BOUNDARY THAT CAN SEGREGATE N-DIMENSIONAL SPACE INTO CLASSES SO THAT WE CAN EASILY PUT THE NEW DATA POINT IN THE CORRECT CATEGORY IN THE FUTURE.
- THIS BEST DECISION BOUNDARY IS CALLED A HYPERPLANE.



EXAMPLE OF SVM



The background is a light blue gradient with several realistic water droplets of various sizes scattered around the edges. The text is centered in the upper half of the image.

**SVM ALGORITHM CAN BE USED FOR FACE
DETECTION, IMAGE CLASSIFICATION, TEXT
CATEGORIZATION, ETC.**

TYPES OF SVM

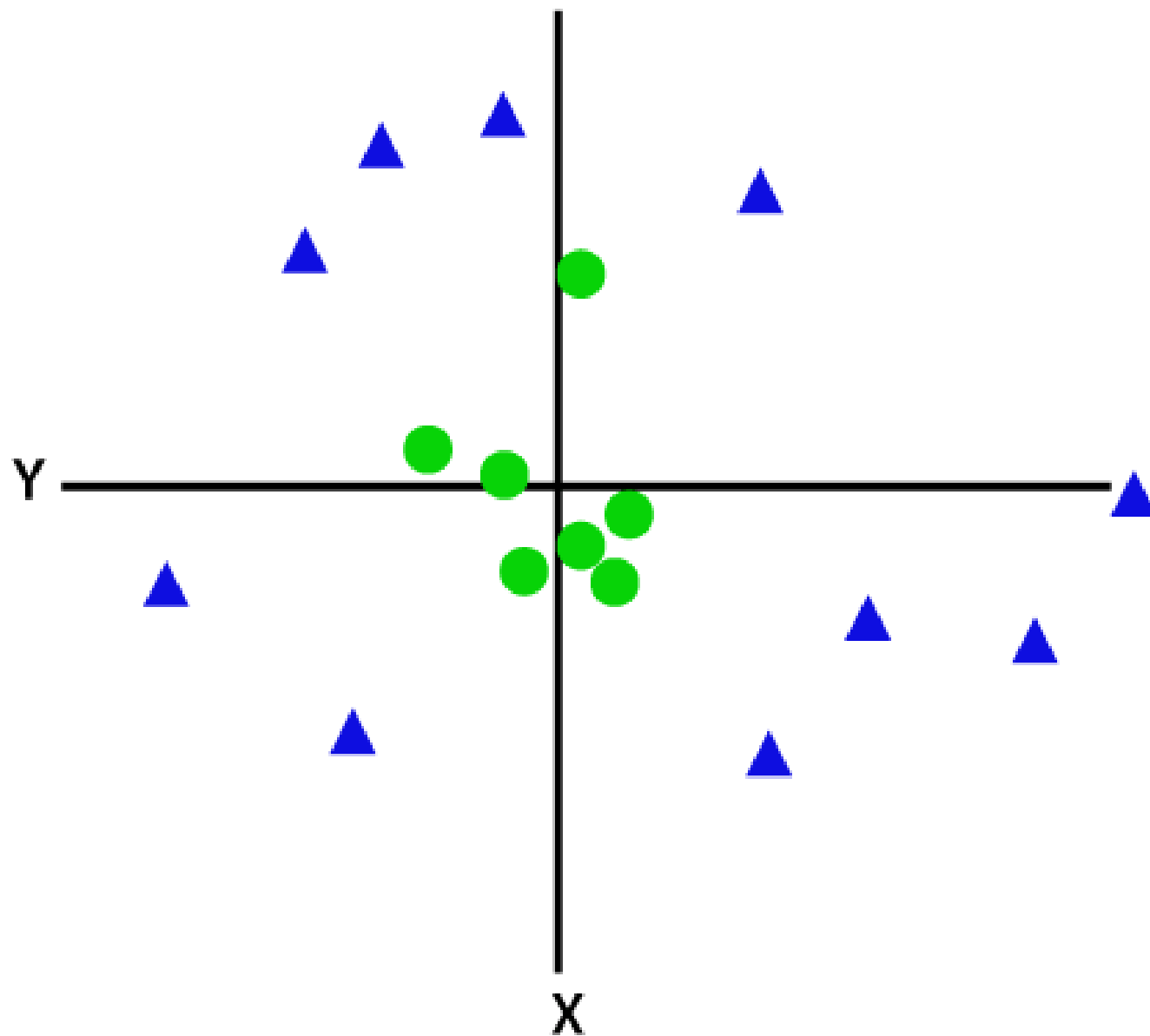
- LINEAR SVM
- NON-LINEAR SVM

LINEAR SVM

- LINEAR SVM IS USED FOR LINEARLY SEPARABLE DATA, WHICH MEANS IF A DATASET CAN BE CLASSIFIED INTO TWO CLASSES BY USING A SINGLE STRAIGHT LINE, THEN SUCH DATA IS TERMED AS LINEARLY SEPARABLE DATA, AND CLASSIFIER IS USED CALLED AS LINEAR SVM CLASSIFIER.

NON-LINEAR SVM

- NON-LINEAR SVM IS USED FOR NON-LINEARLY SEPARATED DATA, WHICH MEANS IF A DATASET CANNOT BE CLASSIFIED BY USING A STRAIGHT LINE, THEN SUCH DATA IS TERMED AS NON-LINEAR DATA AND CLASSIFIER USED IS CALLED AS NON-LINEAR SVM CLASSIFIER.



SUPPORT VECTOR NETWORKS

- THE SUPPORT VECTOR NETWORK IS THE ENSEMBLE OF A NETWORK OF SUPPORT VECTOR MACHINES THAT TOGETHER CLASSIFY THE SAME DATA SET, BY USING DIFFERENT PARAMETERS .

SUPPORT VECTOR CLUSTERING

- SUPPORT VECTOR CLUSTERING IS USED WHERE THE DATA POINTS ARE CLASSIFIED INTO CLUSTERS, WITH SUPPORT VECTOR MACHINES PERFORMING THE CLASSIFICATION AT THE CLUSTER LEVEL.

SUPPORT VECTOR GRID

- THE SUPPORT VECTOR GRID (SVG) IS AN SVC OF AN SVN OR AN SVN OF AN SVC.
- IT USES SVMS TO HANDLE SMALLER CLUSTERS OF THE DATA, TO APPLY SPECIFIC TRANSFORM STEPS.

PRINCIPAL COMPONENT ANALYSIS

- DIMENSION (VARIABLE) REDUCTION TECHNIQUES AIM TO REDUCE A DATA SET WITH HIGHER DIMENSION TO ONE OF LOWER DIMENSION, WITHOUT THE LOSS OF FEATURES OF INFORMATION THAT ARE CONVEYED BY THE DATA SET.

VARIABLE REDUCTION TECHNIQUES

- FACTOR ANALYSIS
- CONJOINT ANALYSIS

FACTOR ANALYSIS

- THE CRUX OF PCA LIES IN MEASURING THE DATA FROM THE PERSPECTIVE OF A PRINCIPAL COMPONENT. A PRINCIPAL COMPONENT OF A DATA SET IS THE DIRECTION WITH THE LARGEST VARIANCE.
- THE HIGHEST VARIANCE AXIS OR, IN OTHER WORDS, THE DIRECTION THAT MOST DEFINES THE DATA.

FACTOR ANALYSIS

- PCA IS FUNDAMENTALLY A DIMENSIONALITY REDUCTION ALGORITHM, BUT IT IS JUST AS USEFUL AS A TOOL FOR
- VISUALIZATION
- NOISE FILTERING
- FEATURE EXTRACTION
- ENGINEERING.

FACTOR ANALYSIS

- FACTOR ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS IDENTIFY PATTERNS IN THE CORRELATIONS BETWEEN VARIABLES. THESE PATTERNS ARE USED TO INFER THE EXISTENCE OF UNDERLYING LATENT VARIABLES IN THE DATA. THESE LATENT VARIABLES ARE OFTEN REFERRED TO AS FACTORS, COMPONENTS, AND DIMENSIONS.
- THE MOST WELL-KNOWN APPLICATION OF THESE TECHNIQUES IS IN IDENTIFYING DIMENSIONS OF PERSONALITY IN PSYCHOLOGY.

	<i>Professional Boxing</i>	<i>This Week</i>	<i>Today</i>	<i>World of Sport</i>	<i>Grandstand</i>	<i>Line-Up</i>	<i>Match of the Day</i>	<i>Panorama</i>	<i>Rugby Special</i>	<i>24 Hours</i>
<i>Professional Boxing</i>		.1	.1	.5	.5	.1	.5	.2	.3	.1
<i>This Week</i>	.1		.3	.1	.1	.2	.1	.4	.1	.4
<i>Today</i>	.1	.3		.1	.1	.2	.0	.2	.1	.2
<i>World of Sport</i>	.5	.1	.1		.6	.1	.6	.2	.3	.1
<i>Grandstand</i>	.5	.1	.1	.6		.1	.6	.2	.3	.1
<i>Line-Up</i>	.1	.2	.2	.1	.1		.0	.2	.1	.3
<i>Match of the Day</i>	.5	.1	.0	.6	.6	.0		.1	.3	.1
<i>Panorama</i>	.2	.4	.2	.2	.2	.2	.1		.1	.5
<i>Rugby Special</i>	.3	.1	.1	.3	.3	.1	.3	.1		.1
<i>24 Hours</i>	.1	.4	.2	.1	.1	.3	.1	.5	.1	

EXAMPLE OF FACTOR AND PRINCIPAL COMPONENT ANALYSIS

- THE TABLE BELOW SHOWS A *CORRELATION MATRIX* OF THE CORRELATIONS BETWEEN VIEWING OF TV PROGRAMS IN THE U.K. IN THE 1970S.
- THE CORRELATION OF .5 BETWEEN *WORLD OF SPORT* AND *PROFESSIONAL BOXING* IS HIGHER THAN THE CORRELATION OF .1 BETWEEN *TODAY* AND *PROFESSIONAL BOXING*.

CONJOINT ANALYSIS

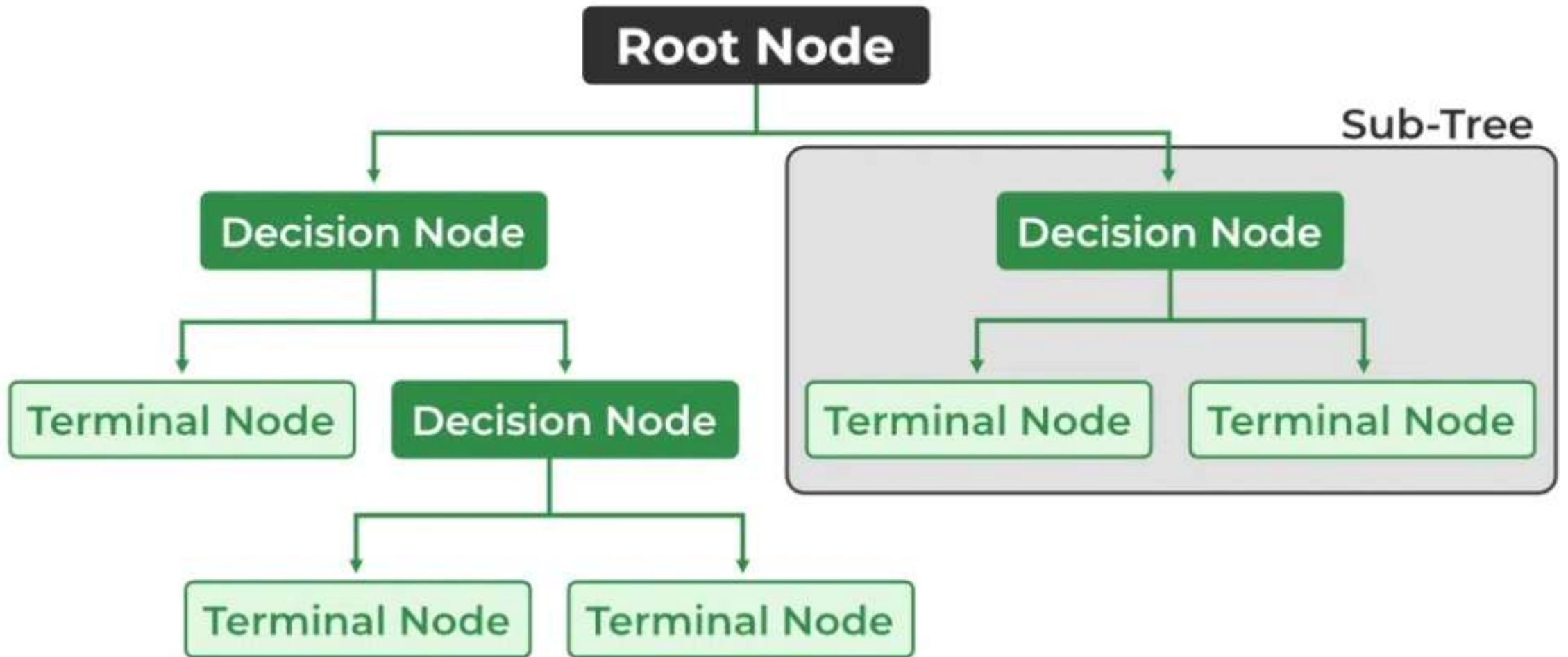
- CONJOINT ANALYSIS IS WIDELY USED IN MARKET RESEARCH TO IDENTIFY CUSTOMERS' PREFERENCE FOR VARIOUS ATTRIBUTES THAT MAKE UP A PRODUCT. THE ATTRIBUTES CAN BE VARIOUS FEATURES, SUCH AS SIZE, COLOR, USABILITY, PRICE, ETC.
- SUCH ANALYSIS IS A HIGHLY USED TECHNIQUE IN NEW PRODUCT DESIGN OR PRICING STRATEGIES.

EXAMPLE

- THE DATA IS A RANKING OF THREE DIFFERENT FEATURES (TV SIZE, TV TYPE, TV COLOR).
- TV SIZE OPTIONS ARE 42", 47", OR 60".
- TV TYPE OPTIONS ARE LCD OR PLASMA.
- TV COLOR OPTIONS ARE RED, BLUE, OR PINK.

DECISION TREES

- A DECISION TREE IS ONE OF THE MOST POWERFUL TOOLS OF SUPERVISED LEARNING ALGORITHMS USED FOR BOTH CLASSIFICATION AND REGRESSION TASKS.
- IT BUILDS A FLOWCHART-LIKE TREE STRUCTURE WHERE EACH INTERNAL NODE DENOTES A TEST ON AN ATTRIBUTE, EACH BRANCH REPRESENTS AN OUTCOME OF THE TEST, AND EACH LEAF NODE (TERMINAL NODE) HOLDS A CLASS LABEL.

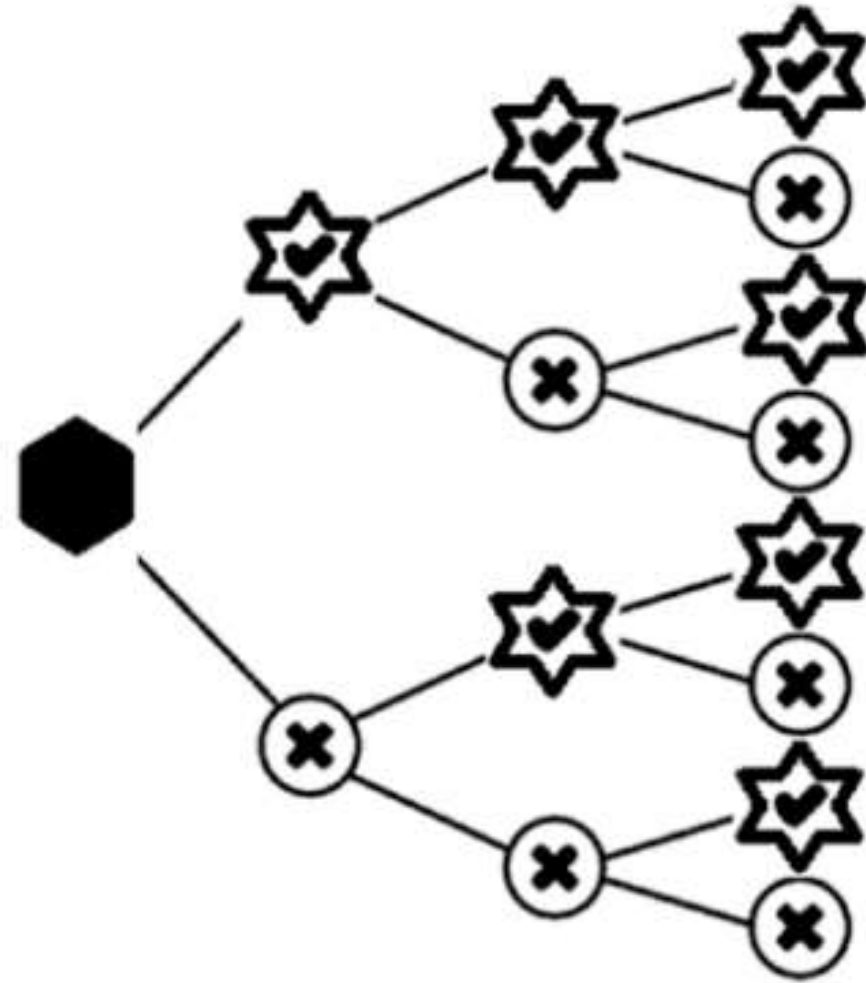


VARIANCE

- **VARIANCE:** VARIANCE MEASURES HOW MUCH THE PREDICTED AND THE TARGET VARIABLES VARY IN DIFFERENT SAMPLES OF A DATASET. IT IS USED FOR REGRESSION PROBLEMS IN DECISION TREES. **MEAN SQUARED ERROR, MEAN ABSOLUTE ERROR** ARE USED TO MEASURE THE VARIANCE FOR THE REGRESSION TASKS IN THE DECISION TREE.

PRUNING

- THE PROCESS OF REMOVING BRANCHES FROM THE TREE THAT DO NOT PROVIDE ANY ADDITIONAL INFORMATION OR LEAD TO OVERFITTING.



Simple decision tree

ENTROPY AND GINI IMPURITY OR INDEX:

- ENTROPY IS THE MEASURE OF THE DEGREE OF RANDOMNESS OR UNCERTAINTY IN THE DATASET.
- GINI IMPURITY IS A SCORE THAT EVALUATES HOW ACCURATE A SPLIT IS AMONG THE CLASSIFIED GROUPS. THE GINI IMPURITY EVALUATES A SCORE IN THE RANGE BETWEEN 0 AND 1, WHERE 0 IS WHEN ALL OBSERVATIONS BELONG TO ONE CLASS, AND 1 IS A RANDOM DISTRIBUTION OF THE ELEMENTS WITHIN CLASSES.

HOW DOES THE DECISION TREE ALGORITHM WORK?

- THE DECISION TREE OPERATES BY ANALYZING THE DATA SET TO PREDICT ITS CLASSIFICATION.
- STEP-1: BEGIN THE TREE WITH THE ROOT NODE, SAYS S , WHICH CONTAINS THE COMPLETE DATASET.
- STEP-2: FIND THE BEST ATTRIBUTE IN THE DATASET USING ATTRIBUTE SELECTION MEASURE (ASM).

CONTT..

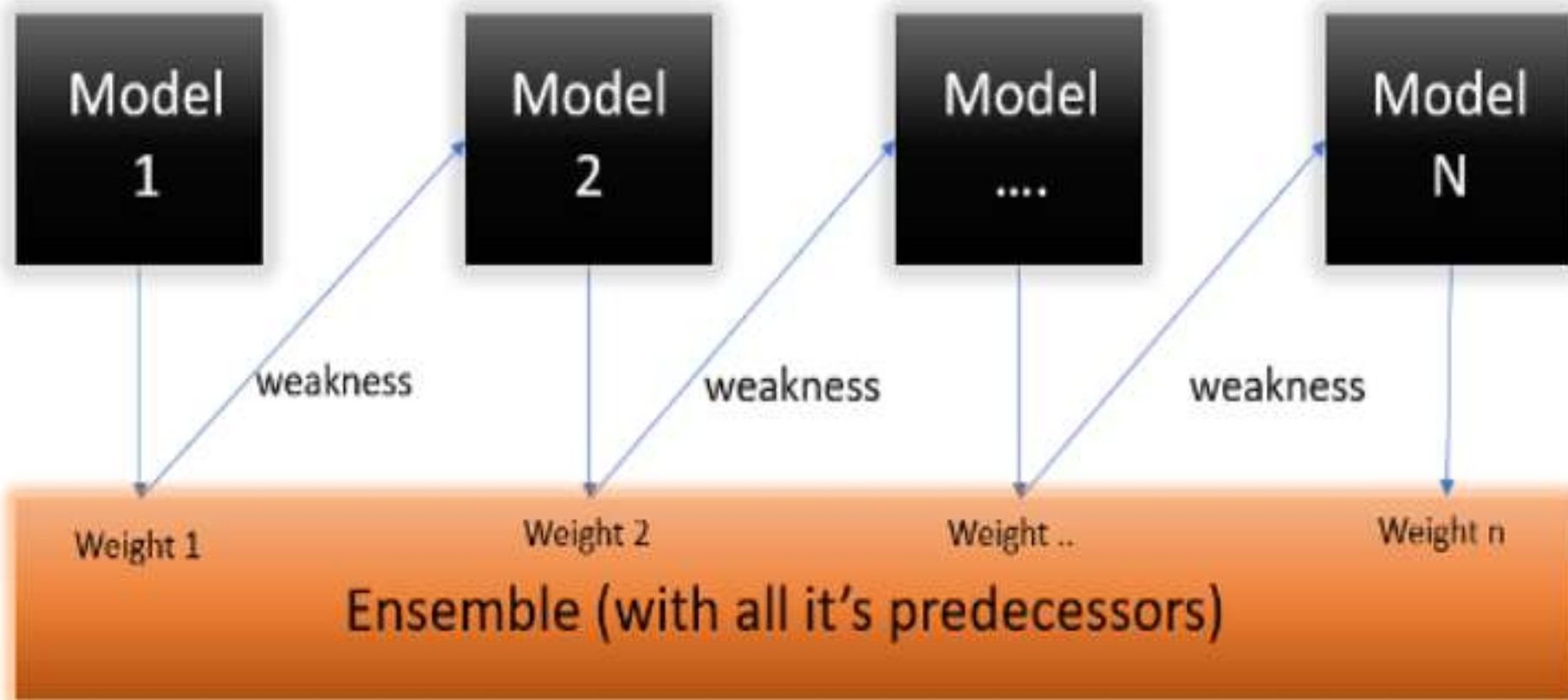
- STEP-3: DIVIDE THE S INTO SUBSETS THAT CONTAINS POSSIBLE VALUES FOR THE BEST ATTRIBUTES.
- STEP-4: GENERATE THE DECISION TREE NODE, WHICH CONTAINS THE BEST ATTRIBUTE.
- STEP-5: RECURSIVELY MAKE NEW DECISION TREES USING THE SUBSETS OF THE DATASET CREATED IN STEP -3. CONTINUE THIS PROCESS UNTIL A STAGE IS REACHED WHERE YOU CANNOT FURTHER CLASSIFY THE NODES AND CALLED THE FINAL NODE AS A LEAF NODE CLASSIFICATION AND REGRESSION TREE ALGORITHM.

ADABOOST ALGORITHM FOR DECISION TREES

- THE PRINCIPLE BEHIND BOOSTING ALGORITHMS IS THAT WE FIRST BUILD A MODEL ON THE TRAINING DATASET AND THEN BUILD A SECOND MODEL TO RECTIFY THE ERRORS PRESENT IN THE FIRST MODEL.
- THIS PROCEDURE IS CONTINUED UNTIL AND UNLESS THE ERRORS ARE MINIMIZED AND THE DATASET IS PREDICTED CORRECTLY.

ADABOOST ALGORITHM FOR DECISION TREES

- WHAT THIS ALGORITHM DOES IS THAT IT BUILDS A MODEL AND GIVES EQUAL WEIGHTS TO ALL THE DATA POINTS. IT THEN ASSIGNS HIGHER WEIGHTS TO POINTS THAT ARE WRONGLY CLASSIFIED. NOW ALL THE POINTS WITH HIGHER WEIGHTS ARE GIVEN MORE IMPORTANCE IN THE NEXT MODEL. IT WILL KEEP TRAINING MODELS UNTIL AND UNLESS A LOWER ERROR IS RECEIVED.



UNDERSTANDING THE WORKING OF THE ADABOOST ALGORITHM

- STEP 1: ASSIGNING WEIGHTS

The formula to calculate the sample weights is:

$$w(x_i, y_i) = \frac{1}{N}, \quad i = 1, 2, \dots, n$$

Where N is the total number of data points

Here since we have 5 data points, the sample weights assigned will be 1/5.

Row No.	Gender	Age	Income	Illness	Sample Weights
1	Male	41	40000	Yes	1/5
2	Male	54	30000	No	1/5
3	Female	42	25000	No	1/5
4	Female	40	60000	Yes	1/5
5	Male	46	50000	Yes	1/5

STEP 2: CLASSIFY THE SAMPLES

- WE START BY SEEING HOW WELL “*GENDER*” CLASSIFIES THE SAMPLES AND WILL SEE HOW THE VARIABLES (AGE, INCOME) CLASSIFY THE SAMPLES.

STEP 3: CALCULATE THE INFLUENCE

- WE'LL NOW CALCULATE THE “**AMOUNT OF SAY**” OR “**IMPORTANCE**” OR “**INFLUENCE**” FOR THIS CLASSIFIER IN CLASSIFYING THE DATA POINTS USING THIS FORMULA:

$$\frac{1}{2} \log \frac{1 - Total\ Error}{Total\ Error}$$

$$\alpha = \frac{1}{2} \log_e \left(\frac{1 - \frac{1}{5}}{\frac{1}{5}} \right)$$

$$\alpha = \frac{1}{2} \log_e \left(\frac{0.8}{0.2} \right)$$

$$\alpha = \frac{1}{2} \log_e(4) = \frac{1}{2} * (1.38)$$

$$\alpha = 0.69$$

Note: Total error will always be between 0 and 1.

STEP 4: CALCULATE TE AND PERFORMANCE

- THE WRONG PREDICTIONS WILL BE GIVEN MORE WEIGHT, WHEREAS THE CORRECT PREDICTIONS WEIGHTS WILL BE DECREASED. NOW WHEN WE BUILD OUR NEXT MODEL AFTER UPDATING THE WEIGHTS, MORE PREFERENCE WILL BE GIVEN TO THE POINTS WITH HIGHER WEIGHTS.

$$\text{New sample weight} = \text{old weight} * e^{\pm \text{Amount of say } (\alpha)}$$

The amount of, say (alpha) will be *negative* when the sample is **correctly classified**.

The amount of, say (alpha) will be *positive* when the sample is **miss-classified**.

New weights for *correctly classified* samples are:

$$\text{New sample weight} = \frac{1}{5} * \exp(-0.69)$$

$$\text{New sample weight} = 0.2 * 0.502 = 0.1004$$

For *wrongly classified* samples, the updated weights will be:

$$\text{New sample weight} = \frac{1}{5} * \exp(0.69)$$

$$\text{New sample weight} = 0.2 * 1.994 = 0.3988$$

Row No.	Gender	Age	Income	Illness	Sample Weights	New Sample Weights
1	Male	41	40000	Yes	1/5	0.1004
2	Male	54	30000	No	1/5	0.1004
3	Female	42	25000	No	1/5	0.1004
4	Female	40	60000	Yes	1/5	0.3988
5	Male	46	50000	Yes	1/5	0.1004

Row No.	Gender	Age	Income	Illness	Sample Weights	New Sample Weights
1	Male	41	40000	Yes	1/5	$0.1004/0.8004 = 0.1254$
2	Male	54	30000	No	1/5	$0.1004/0.8004 = 0.1254$
3	Female	42	25000	No	1/5	$0.1004/0.8004 = 0.1254$
4	Female	40	60000	Yes	1/5	$0.3988/0.8004 = 0.4982$
5	Male	46	50000	Yes	1/5	$0.1004/0.8004 = 0.1254$

STEP 5: DECREASE ERRORS

- NOW, WE NEED TO MAKE A NEW DATASET TO SEE IF THE ERRORS DECREASED OR NOT. FOR THIS, WE WILL REMOVE THE “SAMPLE WEIGHTS” AND “NEW SAMPLE WEIGHTS” COLUMNS AND THEN, BASED ON THE “NEW SAMPLE WEIGHTS,” DIVIDE OUR DATA POINTS INTO BUCKETS.

Row No.	Gender	Age	Income	Illness	New Sample Weights	Buckets
1	Male	41	40000	Yes	$0.1004/0.8004=0.1254$	0 to 0.1254
2	Male	54	30000	No	$0.1004/0.8004=0.1254$	0.1254 to 0.2508
3	Female	42	25000	No	$0.1004/0.8004=0.1254$	0.2508 to 0.3762
4	Female	40	60000	Yes	$0.3988/0.8004=0.4982$	0.3762 to 0.8744
5	Male	46	50000	Yes	$0.1004/0.8004=0.1254$	0.8744 to 0.9998

The background is a light blue gradient. There are several realistic-looking water droplets of various sizes in the corners: top-left, top-right, and bottom-right. The text is centered in the upper half of the image.

**STEP-6:CREATE THE NEW DATASET AND REPEAT THE
PREVIOUS STEPS**

DATA MINING

- DATA MINING IS PROCESSING DATA TO PINPOINT PATTERNS AND ESTABLISH RELATIONSHIPS BETWEEN DATA ENTITIES

ASSOCIATION PATTERNS

- PATTERN ASSOCIATIONS SIMPLY DISCOVER THE CORRELATION OF OCCASIONS IN THE DATA.
- CORRELATION IS ONLY A RELATIONSHIP OR INDICATION OF BEHAVIOR BETWEEN TWO DATA SETS. THE RELATIONSHIP IS NOT A CAUSE-DRIVEN ACTION.

```
import pandas as pd
df1 = pd.DataFrame({'A': range(8), 'B': [2*i for i in range(8)]})
df2 = pd.DataFrame({'A': range(8), 'B': [-2*i for i in range(8)]})
```



```
print('Positive Data Set')  
print(df1)  
print('Negative Data Set')  
print(df2)
```

Here are your results.

```
print('Results')  
print('Correlation Positive:', df1['A'].corr(df1['B']))  
print('Correlation Negative:', df2['A'].corr(df2['B']))
```


IF CORRELATION IS $+1$, THIS MEANS THERE IS A 100% CORRELATION BETWEEN THE TWO VALUES, HENCE THEY CHANGE AT THE SAME RATE.

IF IT IS -1 , THIS MEANS THERE IS A 100% NEGATIVE CORRELATION, INDICATING THAT THE TWO VALUES HAVE A RELATIONSHIP IF ONE INCREASES WHILE THE OTHER DECREASES.

CLASSIFICATION PATTERNS

- DATA CLASSIFICATION IS THE PROCESS OF CONSOLIDATING DATA INTO CATEGORIES, FOR ITS MOST EFFECTIVE AND EFFICIENT USE BY THE DATA PROCESSING.

CLUSTERING PATTERNS

- CLUSTERING IS THE DISCOVERY AND LABELING OF GROUPS OF SPECIFICS NOT PREVIOUSLY KNOWN.

BAYESIAN CLASSIFICATION

- NAÏVE BAYESIAN CLASSIFICATION IS A SUPERVISED LEARNING TECHNIQUE AND A STATISTICAL CLASSIFICATION METHOD. BAYES THEOREM IS USED IN DECISION-MAKING AND USES THE KNOWLEDGE OF PRIOR EVENTS TO PREDICT FUTURE EVENTS.

WHY IS IT CALLED NAÏVE BAYES?

- **NAÏVE:** IT IS CALLED NAÏVE BECAUSE IT ASSUMES THAT THE OCCURRENCE OF A CERTAIN FEATURE IS INDEPENDENT OF THE OCCURRENCE OF OTHER FEATURES.
- **BAYES:** IT IS CALLED BAYES BECAUSE IT DEPENDS ON THE PRINCIPLE OF [BAYES' THEOREM](#).

BAYES THEOREM

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

WORKING OF NAÏVE BAYES' CLASSIFIER:

- CONVERT THE GIVEN DATASET INTO FREQUENCY TABLES.
- GENERATE LIKELIHOOD TABLE BY FINDING THE PROBABILITIES OF GIVEN FEATURES.
- NOW, USE BAYES THEOREM TO CALCULATE THE POSTERIOR PROBABILITY.

PROBLEM: IF THE WEATHER IS SUNNY, THEN THE PLAYER **SHOULD PLAY OR NOT?**

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

Frequency table for the weather conditions.

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Likelihood table weather condition:

Weather	No	Yes	
Overcast	0	5	$5/14 = 0.35$
Rainy	2	2	$4/14 = 0.29$
Sunny	2	3	$5/14 = 0.35$
All	$4/14 = 0.29$	$10/14 = 0.71$	

APPLYING BAYES'THEOREM:

- $P(\text{YES} \mid \text{SUNNY}) = P(\text{SUNNY} \mid \text{YES}) * P(\text{YES}) / P(\text{SUNNY})$
- $P(\text{SUNNY} \mid \text{YES}) = 3/10 = 0.3$
- $P(\text{SUNNY}) = 0.35$
- $P(\text{YES}) = 0.71$
- SO $P(\text{YES} \mid \text{SUNNY}) = 0.3 * 0.71 / 0.35 = \mathbf{0.60}$

APPLYING BAYES' THEOREM:

- $P(\text{NO} \mid \text{SUNNY}) = P(\text{SUNNY} \mid \text{NO}) * P(\text{NO}) / P(\text{SUNNY})$
- $P(\text{SUNNY} \mid \text{NO}) = 2/4 = 0.5$
- $P(\text{NO}) = 0.29$
- $P(\text{SUNNY}) = 0.35$
- SO $P(\text{NO} \mid \text{SUNNY}) = 0.5 * 0.29 / 0.35 = \mathbf{0.41}$
- SO AS WE CAN SEE FROM THE ABOVE CALCULATION THAT $P(\text{YES} \mid \text{SUNNY}) > P(\text{NO} \mid \text{SUNNY})$
- **HENCE ON A SUNNY DAY, PLAYER CAN PLAY THE GAME.**

PATTERN RECOGNITION

- PATTERN RECOGNITION IDENTIFIES REGULARITIES AND IRREGULARITIES IN DATA SETS. THE MOST COMMON APPLICATION OF THIS IS IN TEXT ANALYSIS, TO FIND COMPLEX PATTERNS IN THE DATA.
- EXAMPLE: TO EXTRACT A TEXT FILES FROM A COMMON 20 NEWSGROUPS DATA SET AND THEN CREATE CATEGORIES OF TEXT THAT WAS FOUND TOGETHER IN THE SAME DOCUMENT. THIS WILL PROVIDE YOU WITH THE MOST COMMON WORD THAT IS USED IN THE NEWSGROUPS.

PATTERN RECOGNITION

```
print("Loading 20 newsgroups dataset for categories:")  
print(categories)
```

PATTERN RECOGNITION

You must now load the training data.

```
data = fetch_20newsgroups(subset='train', categories=categories)
print("%d documents" % len(data filenames))
print("%d categories" % len(data.target_names))
```

You now have to define a pipeline, combining a text feature extractor with a simple classifier.

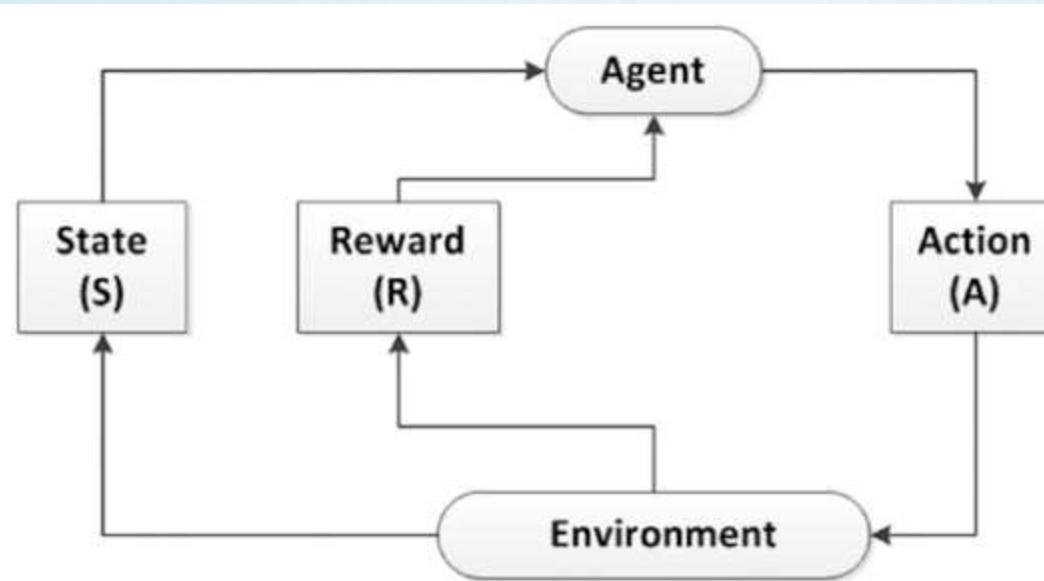
```
pipeline = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf', SGDClassifier()),
])
```


REINFORCEMENT LEARNING

- REINFORCEMENT LEARNING IS AN AREA OF MACHINE LEARNING. IT IS ABOUT TAKING SUITABLE ACTION TO MAXIMIZE REWARD IN A PARTICULAR SITUATION.
- THIS IS USED IN SEVERAL DIFFERENT AREAS, SUCH AS GAME THEORY, OPERATIONS RESEARCH, SIMULATION-BASED OPTIMIZATION, MULTI-AGENT SYSTEMS, STATISTICS, AND GENETIC ALGORITHMS.
- REINFORCEMENT LEARNING IS AN AUTONOMOUS, SELF-TEACHING SYSTEM THAT ESSENTIALLY LEARNS BY TRIAL AND ERROR.



The robot learns by trying all the possible paths and then choosing the path which gives him the reward with the least hurdles. Each right step will give the robot a reward and each wrong step will subtract the reward of the robot.



. *Reinforced learning diagram*

COMPUTER VISION

- COMPUTER VISION IS A COMPLEX FEATURE EXTRACTION AREA, BUT ONCE YOU HAVE THE FEATURES EXPOSED. IT SIMPLY BECOMES A MATRIX OF VALUES.

```
import matplotlib.pyplot as plt
from PIL import Image
import numpy as np
sPicNameIn='C:/VKHCG/01-Vermeulen/00-RawData/AudiR8.png'
imageIn = Image.open(sPicNameIn)
fig1=plt.figure(figsize=(10, 10))
fig1.suptitle('Audi R8', fontsize=20)
imgplot = plt.imshow(imageIn)
plt.show()
```

You should see a car.

```
imagewidth, imageheight = imageIn.size  
imageMatrix=np.asarray(imageIn)  
pixelscnt = (imagewidth * imageheight)  
print('Pixels:', pixelscnt)  
print('Size:', imagewidth, ' x', imageheight,)  
print(imageMatrix)
```

This is what your computer sees!

NATURAL LANGUAGE PROCESSING

- **NATURAL LANGUAGE PROCESSING** ENABLES MACHINES TO UNDERSTAND AND RESPOND TO TEXT OR VOICE DATA.
- THIS IS A WIDELY USED TECHNOLOGY FOR PERSONAL ASSISTANTS THAT ARE USED IN VARIOUS BUSINESS FIELDS/AREAS.
- NLP IS USED IN A WIDE RANGE OF APPLICATIONS, INCLUDING MACHINE TRANSLATION, SENTIMENT ANALYSIS, SPEECH RECOGNITION, CHATBOTS, AND TEXT CLASSIFICATION.

WORKING OF NATURAL LANGUAGE PROCESSING (NLP)

- SPEECH RECOGNITION—THE TRANSLATION OF SPOKEN LANGUAGE INTO TEXT.
- NATURAL LANGUAGE UNDERSTANDING (NLU) —THE COMPUTER'S ABILITY TO UNDERSTAND WHAT WE SAY.
- NATURAL LANGUAGE GENERATION (NLG) —THE GENERATION OF NATURAL LANGUAGE BY A COMPUTER.

TEXT-BASED

```
import nltk  
nltk.download()
```

You will see a program that enables you to download several text libraries, which will assist the process to perform text analysis against any text you submit for analysis.

The basic principle is that the library matches your text against the text stored in the data libraries and will return the correct matching text analysis.

```
from nltk.tokenize import sent_tokenize, word_tokenize

Txt = "Good Day Mr. Vermeulen,\n\nhow are you doing today?\n\nThe weather is great, and Data Science is awesome.\n\nYou are doing well!"

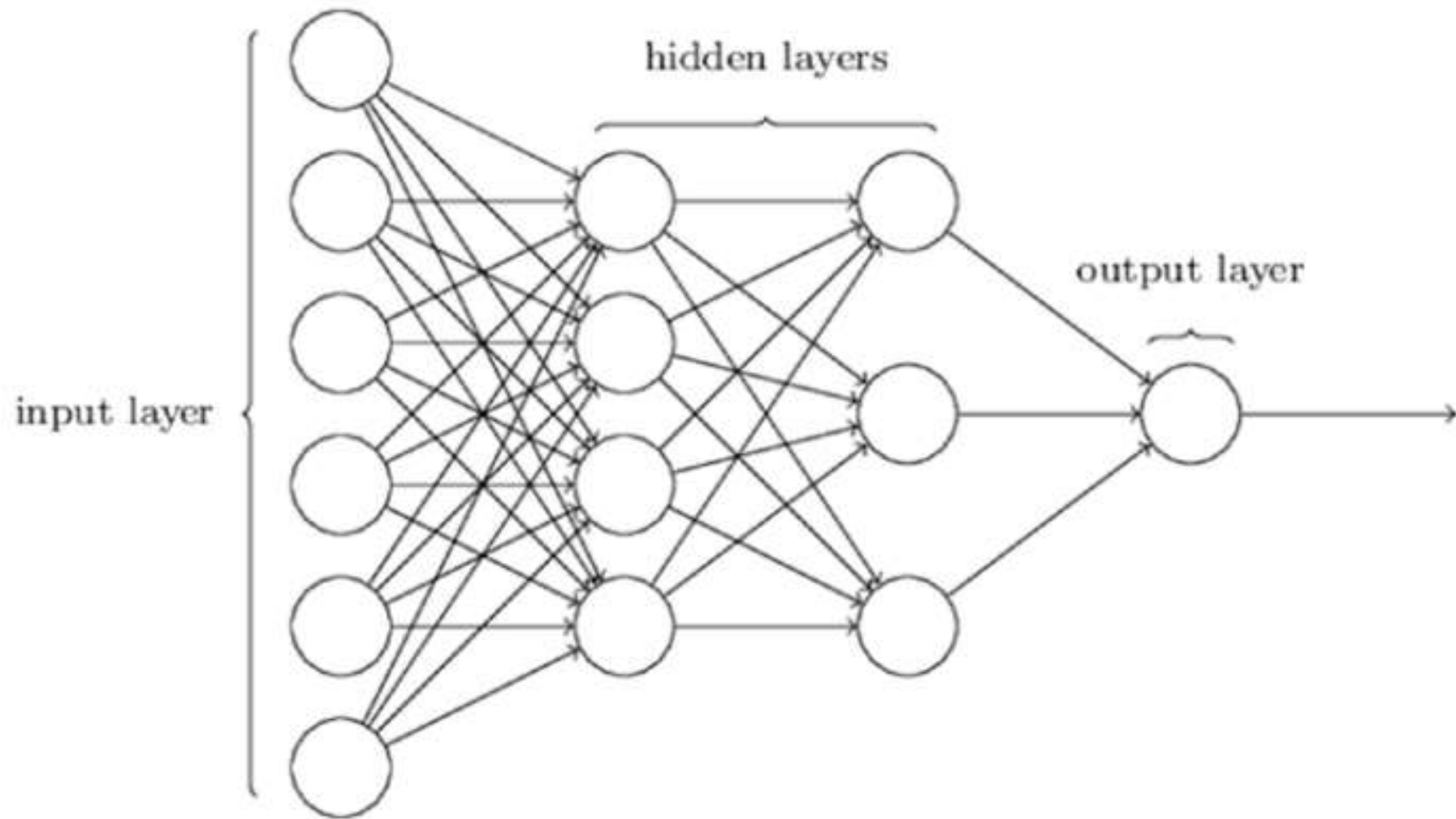
print(Txt, '\n')

print('Identify sentences')
print(sent_tokenize(Txt), '\n')

print('Identify Word')
print(word_tokenize(Txt))
```

NEURAL NETWORKS

- NEURAL NETWORKS (ALSO KNOWN AS ARTIFICIAL NEURAL NETWORKS) ARE INSPIRED BY THE HUMAN NERVOUS SYSTEM..THEY SOLVE VARIOUS REAL-TIME TASKS BECAUSE OF ITS ABILITY TO PERFORM COMPUTATIONS QUICKLY AND ITS FAST RESPONSES.
- NEURAL NETWORKS ARE USED TO FIND PATTERNS IN EXTREMELY COMPLEX DATA AND, THUS, DELIVER FORECASTS AND CLASSIFY DATA POINTS. NEURAL NETWORKS ARE USUALLY ORGANIZED IN LAYERS. LAYERS ARE MADE UP OF A NUMBER OF INTERCONNECTED “NODES.”



General artificial neural network



TYPES OF TASKS THAT CAN BE SOLVED USING AN ARTIFICIAL NEURAL NETWORK INCLUDE CLASSIFICATION PROBLEMS, PATTERN MATCHING, DATA CLUSTERING, ETC.

TYPES OF NEURAL NETWORK IN MACHINE LEARNING

- **ANN**
- ANN IS ALSO KNOWN AS AN ARTIFICIAL NEURAL NETWORK. IT IS A FEED-FORWARD NEURAL NETWORK BECAUSE THE INPUTS ARE SENT IN THE FORWARD DIRECTION.
- IT IS COMPARATIVELY LESS POWERFUL THAN CNN AND RNN.

CNN


- CONVOLUTIONAL NEURAL NETWORKS IS MAINLY USED FOR IMAGE DATA. IT IS USED FOR COMPUTER VISION. SOME OF THE REAL-LIFE APPLICATIONS ARE OBJECT DETECTION IN AUTONOMOUS VEHICLES.
- IT IS MORE POWERFUL THAN BOTH ANN AND RNN.

RNN

- IT IS ALSO KNOWN AS RECURRENT NEURAL NETWORKS. IT IS USED TO PROCESS AND INTERPRET TIME SERIES DATA. IN THIS TYPE OF MODEL, THE OUTPUT FROM A PROCESSING NODE IS FED BACK INTO NODES IN THE SAME OR PREVIOUS LAYERS.



TYPES OF LEARNING IN NEURAL NETWORKS

- **SUPERVISED LEARNING**
 - **UNSUPERVISED LEARNING**
 - **REINFORCEMENT LEARNING**
- 

REGULARIZATION STRENGTH

- REGULARIZATION STRENGTH IS THE PARAMETER THAT PREVENTS OVERFITTING OF THE NEURAL NETWORK.
- THE COMMON NAME FOR THIS SETTING IS THE EPSILON PARAMETER, ALSO KNOWN AS THE LEARNING RATE.

GRADIENT DESCENT

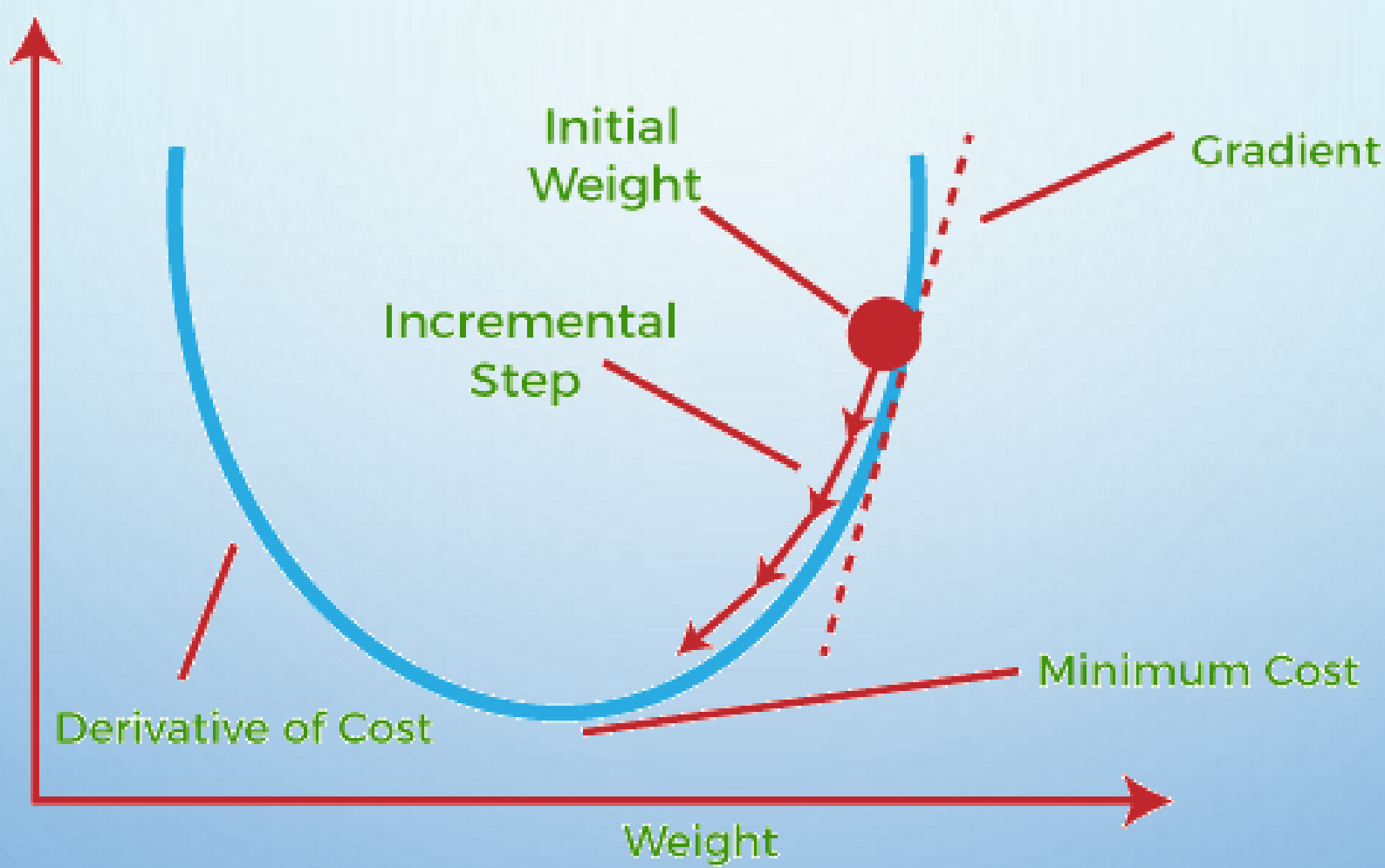
- GRADIENT DESCENT IS KNOWN AS ONE OF THE MOST COMMONLY USED OPTIMIZATION ALGORITHMS TO TRAIN MACHINE LEARNING MODELS BY MEANS OF MINIMIZING ERRORS BETWEEN ACTUAL AND EXPECTED RESULTS. FURTHER, GRADIENT DESCENT IS ALSO USED TO TRAIN NEURAL NETWORKS.
- IN MACHINE LEARNING, OPTIMIZATION IS THE TASK OF MINIMIZING THE COST FUNCTION PARAMETERIZED BY THE MODEL'S PARAMETERS.

LOCAL MINIMUM AND LOCAL MAXIMUM

- IF WE MOVE TOWARDS A NEGATIVE GRADIENT OR AWAY FROM THE GRADIENT OF THE FUNCTION AT THE CURRENT POINT, IT WILL GIVE THE **LOCAL MINIMUM** OF THAT FUNCTION.
- WHENEVER WE MOVE TOWARDS A POSITIVE GRADIENT OR TOWARDS THE GRADIENT OF THE FUNCTION AT THE CURRENT POINT, WE WILL GET THE **LOCAL MAXIMUM** OF THAT FUNCTION.

STEPS OF GRADIENT DESCENT TO MINIMIZE COST FUNCTION

- CALCULATES THE FIRST-ORDER DERIVATIVE OF THE FUNCTION TO COMPUTE THE GRADIENT OR SLOPE OF THAT FUNCTION.
- MOVE AWAY FROM THE DIRECTION OF THE GRADIENT, WHICH MEANS SLOPE INCREASED FROM THE CURRENT POINT BY ALPHA TIMES, WHERE ALPHA IS DEFINED AS LEARNING RATE.

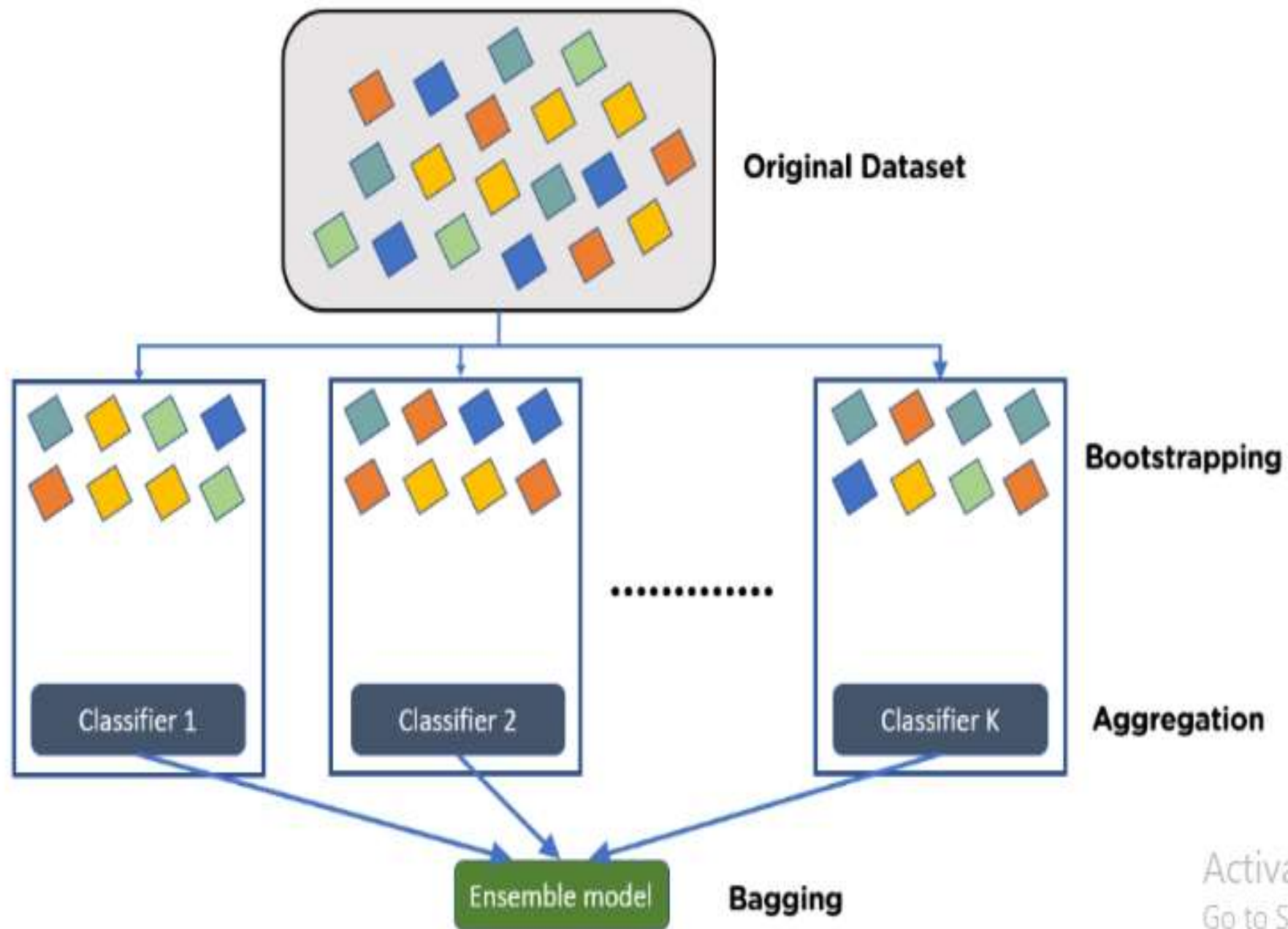


ENSEMBLE LEARNING

- ENSEMBLE LEARNING IS A WIDELY-USED AND PREFERRED MACHINE LEARNING TECHNIQUE IN WHICH MULTIPLE INDIVIDUAL MODELS, OFTEN CALLED BASE MODELS, ARE COMBINED TO PRODUCE AN EFFECTIVE OPTIMAL PREDICTION MODEL.
- THE RANDOM FOREST ALGORITHM IS AN EXAMPLE OF ENSEMBLE LEARNING.

BAGGING DATA

- BAGGING, ALSO KNOWN AS BOOTSTRAP AGGREGATING, IS AN ENSEMBLE LEARNING TECHNIQUE THAT HELPS TO IMPROVE THE PERFORMANCE AND ACCURACY OF MACHINE LEARNING ALGORITHMS.
- REDUCES THE VARIANCE OF A PREDICTION MODEL. BAGGING AVOIDS OVERFITTING OF DATA AND IS USED FOR BOTH REGRESSION AND CLASSIFICATION MODELS, SPECIFICALLY FOR DECISION TREE ALGORITHMS.



STEPS TO PERFORM BAGGING

- CONSIDER THERE ARE N OBSERVATIONS AND M FEATURES IN THE TRAINING SET. YOU NEED TO SELECT A RANDOM SAMPLE FROM THE TRAINING DATASET WITHOUT REPLACEMENT.
- A SUBSET OF M FEATURES IS CHOSEN RANDOMLY TO CREATE A MODEL USING SAMPLE OBSERVATIONS

STEPS TO PERFORM BAGGING

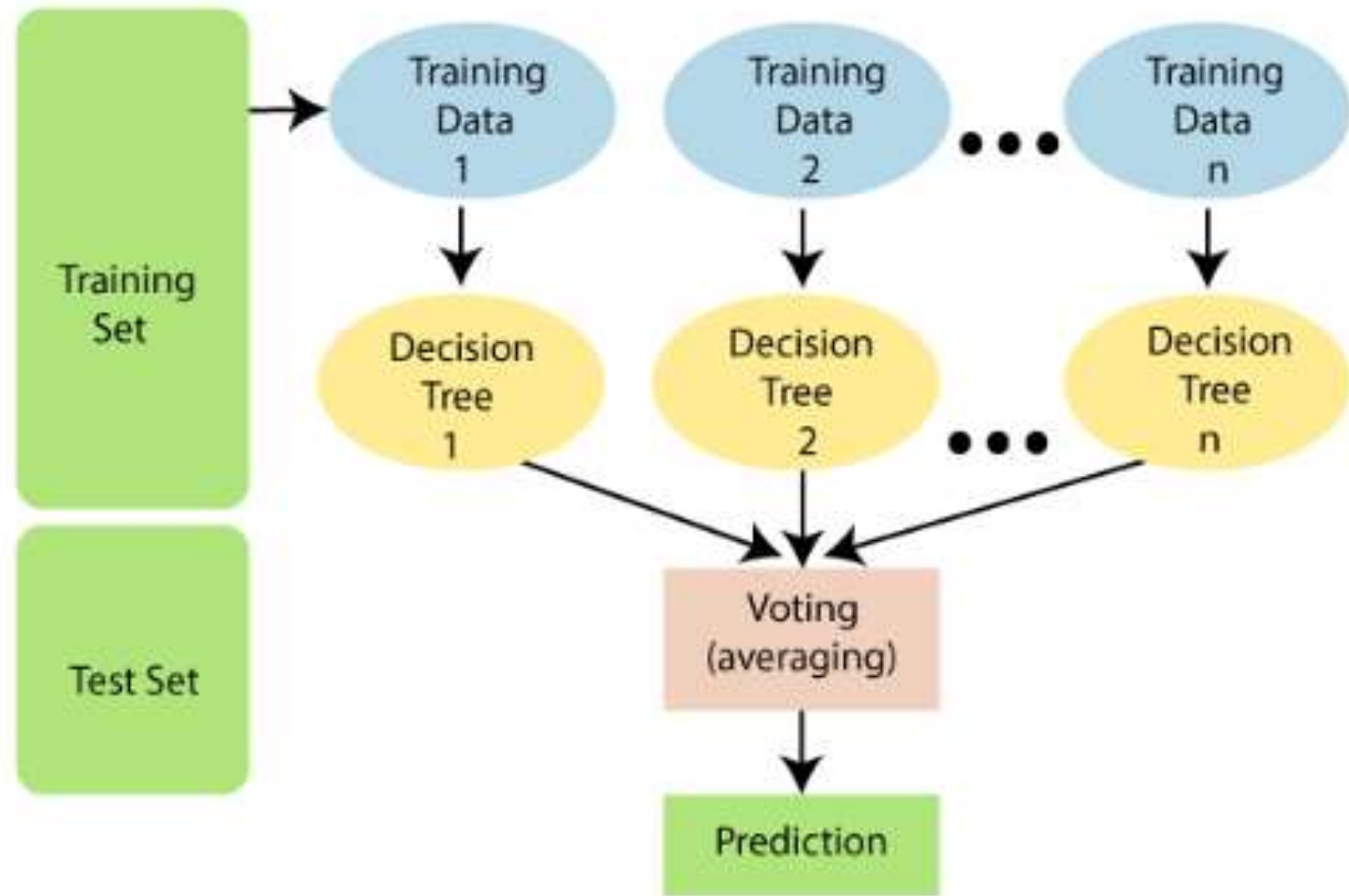
- THE FEATURE OFFERING THE BEST SPLIT OUT OF THE LOT IS USED TO SPLIT THE NODES.
- THE TREE IS GROWN, SO YOU HAVE THE BEST ROOT NODES.
- THE ABOVE STEPS ARE REPEATED N TIMES. IT AGGREGATES THE OUTPUT OF INDIVIDUAL DECISION TREES TO GIVE THE BEST PREDICTION.

RANDOM FOREST

- IT IS A **SUPERVISED LEARNING TECHNIQUE** THAT CONSTRUCTS AN ENSEMBLE OF DECISION-TREE CLASSIFIERS AND USES RANDOM SELECTION TO CREATE MULTIPLE FORESTS FROM WHICH THE FINAL PREDICTION IS MADE.
- THE RANDOM FOREST HAS BEEN UTILISED IN VARIOUS APPLICATIONS, RANGING FROM **HEALTHCARE** TO FINANCE.

BASIC CONCEPT

- ***"RANDOM FOREST IS A CLASSIFIER THAT CONTAINS A NUMBER OF DECISION TREES ON VARIOUS SUBSETS OF THE GIVEN DATASET AND TAKES THE AVERAGE TO IMPROVE THE PREDICTIVE ACCURACY OF THAT DATASET."***
- **THE GREATER NUMBER OF TREES IN THE FOREST LEADS TO HIGHER ACCURACY AND PREVENTS THE PROBLEM OF OVERFITTING.**



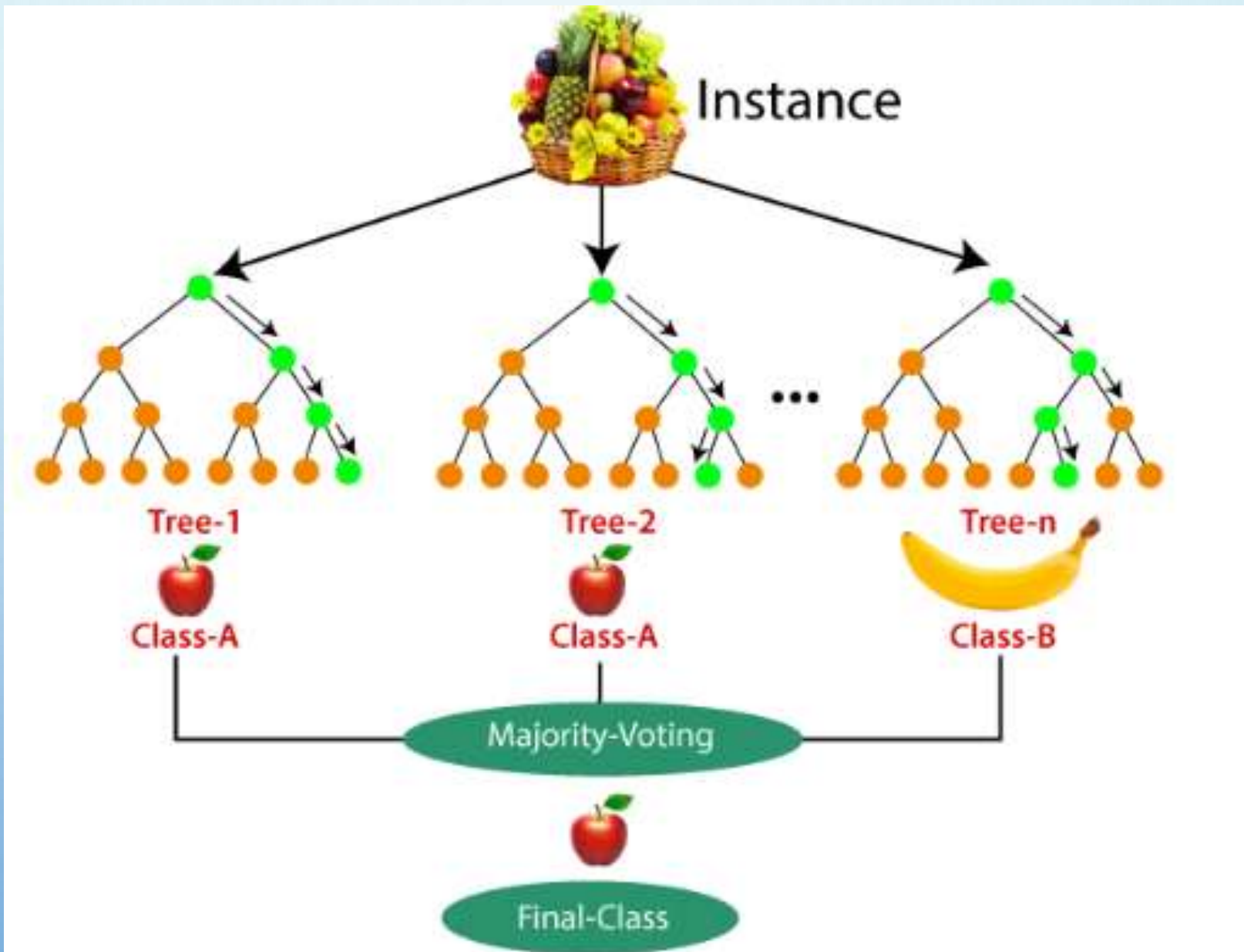
WHY USE RANDOM FOREST?

- IT TAKES LESS TRAINING TIME AS COMPARED TO OTHER ALGORITHMS.
- IT PREDICTS OUTPUT WITH HIGH ACCURACY, EVEN FOR THE LARGE DATASET IT RUNS EFFICIENTLY.
- IT CAN ALSO MAINTAIN ACCURACY WHEN A LARGE PROPORTION OF DATA IS MISSING.

STEPS OF RANDOM FOREST ALGORITHM


- **STEP-1:** SELECT RANDOM K DATA POINTS FROM THE TRAINING SET.
- **STEP-2:** BUILD THE DECISION TREES ASSOCIATED WITH THE SELECTED DATA POINTS (SUBSETS).
- **STEP-3:** CHOOSE THE NUMBER N FOR DECISION TREES THAT YOU WANT TO BUILD.
- **STEP-4:** REPEAT STEP 1 & 2.
- **STEP-5:** FOR NEW DATA POINTS, FIND THE PREDICTIONS OF EACH DECISION TREE, AND ASSIGN THE NEW DATA POINTS TO THE CATEGORY THAT WINS THE MAJORITY VOTES.

EXAMPLE OF RANDOM FOREST





APPLICATIONS OF RANDOM FOREST

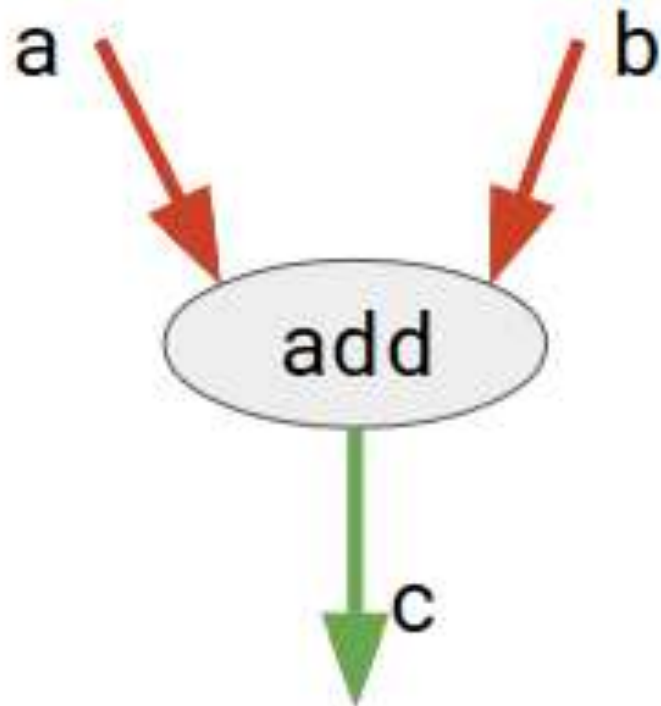
- IMAGE RECOGNITION
 - DISEASE DIAGNOSIS
 - TEXT ANALYSIS
 - RECOMMENDATION SYSTEMS
- 

LIMITATIONS OF RANDOM FOREST

- MAY NOT WORK WELL ON TOO SMALL DATASETS.
- RANDOM FORESTS CAN ALSO BE SLOW TO TRAIN AND PREDICT AS THE NUMBER OF TREES IN THE FOREST INCREASES.

TENSOR FLOW

- **TENSORFLOW** IS BASICALLY A SOFTWARE LIBRARY FOR NUMERICAL COMPUTATION USING **DATA FLOW GRAPHS** WHERE:
- **NODES** IN THE GRAPH REPRESENT MATHEMATICAL OPERATIONS.
- **EDGES** IN THE GRAPH REPRESENT THE MULTIDIMENSIONAL DATA ARRAYS (CALLED **TENSORS**) COMMUNICATED BETWEEN THEM.



Here, **add** is a node which represents addition operation. **a** and **b** are input tensors and **c** is the resultant tensor.

INSTALLING TENSORFLOW

```
import tensorflow as tf
```

THE COMPUTATIONAL GRAPH

- A **COMPUTATIONAL GRAPH** IS NOTHING BUT A SERIES OF TENSORFLOW OPERATIONS ARRANGED INTO A GRAPH OF NODES.

```
# importing tensorflow
import tensorflow as tf

# creating nodes in computation graph
node1 = tf.constant(3, dtype=tf.int32)
node2 = tf.constant(5, dtype=tf.int32)
node3 = tf.add(node1, node2)

# create tensorflow session object
sess = tf.compat.v1.Session()

# evaluating node3 and printing the result
print("sum of node1 and node2 is :",sess.run(node3))
# closing the session
sess.close()
```

Sum of node1 and node2 is: 8

In order to run the computational graph, we need to create a **session**.

VARIABLES IN TENSORFLOW

- TENSORFLOW HAS **VARIABLE** NODES TOO WHICH CAN HOLD VARIABLE DATA. THEY ARE MAINLY USED TO HOLD AND UPDATE PARAMETERS OF A TRAINING MODEL.

```
b = tf.Variable(2.5, name='b')  
c = tf.Variable(10.0, name='c')
```


PLACEHOLDERS IN TENSORFLOW

- A GRAPH CAN BE PARAMETERIZED TO ACCEPT EXTERNAL INPUTS, KNOWN AS **PLACEHOLDERS**.

```
# importing tensorflow
import tensorflow as tf

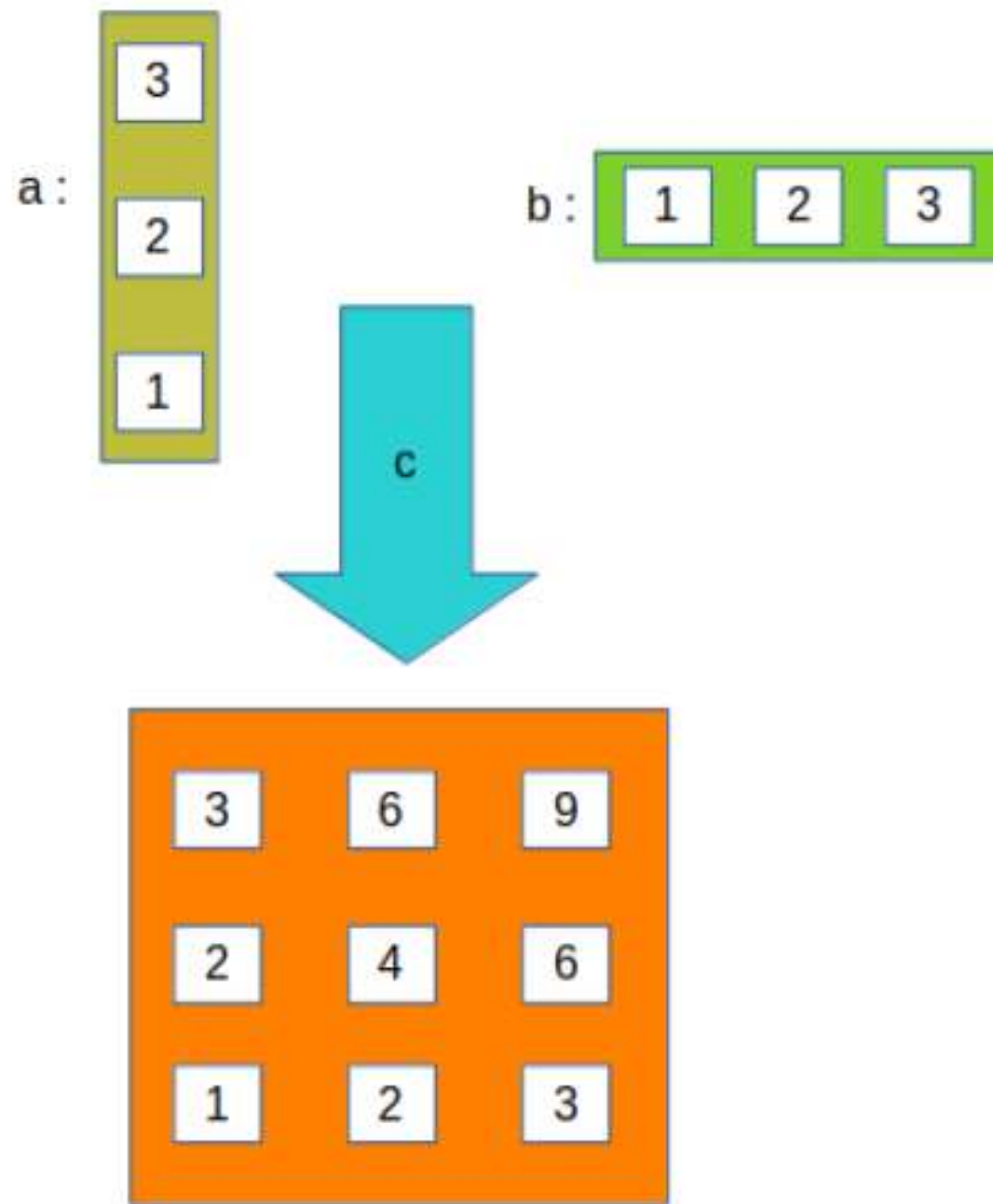
# creating nodes in computation graph
a = tf.placeholder(tf.int32, shape=(3,1))
b = tf.placeholder(tf.int32, shape=(1,3))
c = tf.matmul(a,b)

# running computation graph
with tf.Session() as sess:
    print(sess.run(c, feed_dict={a:[[3],[2],[1]], b:[[1,2,3]]}))
```

OUTPUT


```
[[3 6 9]  
 [2 4 6]  
 [1 2 3]]
```

- After sess.run:





REAL-WORLD USES OF TENSORFLOW

- BASKET ANALYSIS: WHAT DO PEOPLE BUY? WHAT DO THEY BUY TOGETHER?
 - FOREX TRADING: PROVIDING RECOMMENDATIONS ON WHEN TO PURCHASE FOREX FOR COMPANY REQUIREMENTS.
 - COMMODITIES TRADING: BUYING AND SELLING FUTURES.
- 

The background is a light blue gradient. In the top-left corner, there are several water droplets of varying sizes, some overlapping. In the top-right corner, there is one small droplet. In the bottom-right corner, there is a cluster of droplets, including a large one and several smaller ones. In the bottom-center, there are two small droplets.

UNIT-5 PART-2

ORGANIZE AND REPORT SUPERSTEPS

ORGANIZE SUPERSTEP

- THE ORGANIZE SUPERSTEP TAKES THE COMPLETE DATA WAREHOUSE YOU BUILT AT THE END OF THE TRANSFORM SUPERSTEP AND SUBSECTIONS IT INTO BUSINESS-SPECIFIC DATA MARTS. A DATA MART IS THE ACCESS LAYER OF THE DATA WAREHOUSE ENVIRONMENT BUILT TO EXPOSE DATA TO THE USERS.
- THE DATA MART IS A SUBSET OF THE DATA WAREHOUSE AND IS GENERALLY ORIENTED TO A SPECIFIC BUSINESS GROUP.

DATA MART VS. DATA WAREHOUSE VS DATA LAKE

- A DATA WAREHOUSE IS A SYSTEM THAT AGGREGATES DATA FROM MULTIPLE SOURCES INTO A SINGLE, CENTRAL, CONSISTENT DATA STORE TO SUPPORT DATA MINING, ARTIFICIAL INTELLIGENCE (AI), AND MACHINE LEARNING.
- A DATA MART (AS NOTED ABOVE) IS A FOCUSED VERSION OF A DATA WAREHOUSE THAT CONTAINS A SMALLER SUBSET OF DATA IMPORTANT TO AND NEEDED BY A SINGLE TEAM OR A SELECT GROUP OF USERS WITHIN AN ORGANIZATION.

DATA MART VS. DATA WAREHOUSE VS DATA LAKE

- A DATA LAKE, TOO, IS A REPOSITORY FOR DATA. A DATA LAKE PROVIDES MASSIVE STORAGE OF UNSTRUCTURED OR RAW DATA FED VIA MULTIPLE SOURCES, BUT THE INFORMATION HAS NOT YET BEEN PROCESSED OR PREPARED FOR ANALYSIS.

HORIZONTAL STYLE

- HORIZONTAL STYLE SLICING SELECTS THE SUBSET OF ROWS FROM THE POPULATION WHILE PRESERVING THE COLUMNS.


```
# -*- coding: utf-8 -*-  
#####  
import sys  
import os  
import pandas as pd  
import sqlite3 as sq  
#####  
  
    If sys.platform == 'linux' or sys.platform == 'darwin':  
  
        Base=os.path.expanduser('~') + '/VKHCG'  
else:  
    Base='C:/VKHCG'  
print('#####')  
print('Working Base :',Base, ' using ', sys.platform)  
print('#####')
```


Company='01-Vermeulen'

#####

sDataWarehouseDir=Base + '/99-DW'

```
if not os.path.exists(sDataWarehouseDir):  
    os.makedirs(sDataWarehouseDir)
```

```
#####  
sDatabaseName=sDataWarehouseDir + '/datawarehouse.db'  
conn1 = sq.connect(sDatabaseName)  
#####  
sDatabaseName=sDataWarehouseDir + '/datamart.db'  
conn2 = sq.connect(sDatabaseName)  
#####
```

Load the complete BMI data set from the data warehouse.

The next query loads all the data into memory, and that means you will have the complete data set ready in memory

```
print('#####')  
sTable = 'Dim-BMI'  
print('Loading :',sDatabaseName,' Table:',sTable)  
sSQL="SELECT * FROM [Dim-BMI];"  
PersonFrame0=pd.read_sql_query(sSQL, conn1)
```

LOADING THE HORIZONTAL DATA SLICE FOR BMI

```
print('#####')  
sTable = 'Dim-BMI'  
print('Loading :',sDatabaseName,' Table:',sTable)  
sSQL="SELECT PersonID,\n      Height,\n      Weight,\n      bmi,\n      Indicator\
```


LOADING THE HORIZONTAL DATA SLICE FOR BMI

```
FROM [Dim-BMI]\
WHERE \
Height > 1.5 \
and Indicator = 1\
ORDER BY \
    Height,\
    Weight;"
PersonFrame1=pd.read_sql_query(sSQL, conn1)
#####
DimPerson=PersonFrame1
DimPersonIndex=DimPerson.set_index(['PersonID'],inplace=False)
#####
```


**STORE THE HORIZONTAL DATA SLICE FOR BMI INTO
THE DATA WAREHOUSE.**

```

sTable = 'Dim-BMI'
print('\n#####')
print('Storing :',sDatabaseName,'\n Table:',sTable)
print('\n#####')
DimPersonIndex.to_sql(sTable, conn2, if_exists="replace")
#####
print('#####')
sTable = 'Dim-BMI'
print('Loading :',sDatabaseName,' Table:',sTable)
sSQL="SELECT * FROM [Dim-BMI];"
PersonFrame2=pd.read_sql_query(sSQL, conn2)

```

You can show your results by printing the following code. You can see the improvement you achieved.

```

print('Full Data Set (Rows):', PersonFrame0.shape[0])
print('Full Data Set (Columns):', PersonFrame0.shape[1])
print('Horizontal Data Set (Rows):', PersonFrame2.shape[0])
print('Horizontal Data Set (Columns):', PersonFrame2.shape[1])

```

#####

Full Data Set (Rows): 1080

Full Data Set (Columns): 5

#####

Horizontal Data Set (Rows): 194

Horizontal Data Set (Columns): 5

#####

VERTICAL STYLE

- THE VERTICAL-STYLE SLICING SELECTS THE SUBSET OF COLUMNS FROM THE POPULATION, WHILE PRESERVING THE ROWS.
- THE USE OF VERTICAL-STYLE DATA SLICING IS COMMON IN SYSTEMS IN WHICH SPECIFIC DATA COLUMNS MAY NOT BE SHOWN TO EVERYBODY, OWING TO SECURITY OR PRIVACY REGULATIONS

ISLAND STYLE

- PERFORMING ISLAND-STYLE SLICING OR SUBSETTING OF THE DATA WAREHOUSE IS ACHIEVED BY APPLYING A COMBINATION OF HORIZONTAL- AND VERTICAL-STYLE SLICING. THIS GENERATES A SUBSET OF SPECIFIC ROWS AND SPECIFIC COLUMNS REDUCED AT THE SAME TIME.

Secure Vault Style

SECURE VAULT STYLE

- THE SECURE VAULT IS A VERSION OF ONE OF THE HORIZONTAL, VERTICAL, OR ISLAND SLICING TECHNIQUES, BUT THE OUTCOME IS ALSO ATTACHED TO THE PERSON WHO PERFORMS THE QUERY. THIS IS COMMON IN MULTI-SECURITY ENVIRONMENTS, WHERE DIFFERENT USERS ARE ALLOWED TO SEE DIFFERENT DATA SETS

ASSOCIATION RULE MINING

- ASSOCIATION RULE LEARNING IS A RULE-BASED MACHINE-LEARNING METHOD FOR DISCOVERING INTERESTING RELATIONS BETWEEN VARIABLES IN LARGE DATABASES, SIMILAR TO THE DATA YOU WILL FIND IN A DATA LAKE.

EXAMPLE

- SHOPPING BASKET ANALYSIS TOOL IN MICROSOFT EXCEL, WHICH ANALYZES TRANSACTION DATA CONTAINED IN A SPREADSHEET AND PERFORMS MARKET BASKET ANALYSIS.
- A TRANSACTION ID MUST RELATE TO THE ITEMS TO BE ANALYZED. THE SHOPPING BASKET ANALYSIS TOOL THEN CREATES TWO WORKSHEETS:

EXAMPLE

- THE SHOPPING BASKET ITEM GROUPS WORKSHEET, WHICH LISTS ITEMS THAT ARE FREQUENTLY PURCHASED TOGETHER,
- AND THE SHOPPING BASKET RULES WORKSHEET SHOWS HOW ITEMS ARE RELATED (FOR EXAMPLE, PURCHASERS OF PRODUCT A ARE LIKELY TO BUY PRODUCT B).

HOW DOES MARKET BASKET ANALYSIS WORK?

- MARKET BASKET ANALYSIS IS MODELLED ON ASSOCIATION RULE MINING, I.E., THE IF {}, THEN {} CONSTRUCT. FOR EXAMPLE, IF A CUSTOMER BUYS BREAD, THEN HE IS LIKELY TO BUY BUTTER AS WELL.
- ASSOCIATION RULES ARE USUALLY REPRESENTED AS: {BREAD} -> {BUTTER}

ANTECEDANTS AND CONSEQUENT

- **ANTECEDENT:** ITEMS OR 'ITEMSETS' FOUND WITHIN THE DATA ARE ANTECEDENTS. IN SIMPLER WORDS, IT'S THE IF COMPONENT, WRITTEN ON THE LEFT-HAND SIDE. IN THE ABOVE EXAMPLE, BREAD IS THE ANTECEDENT.
- **CONSEQUENT:** A CONSEQUENT IS AN ITEM OR SET OF ITEMS FOUND IN COMBINATION WITH THE ANTECEDENT. IT'S THE THEN COMPONENT, WRITTEN ON THE RIGHT-HAND SIDE. IN THE ABOVE EXAMPLE, BUTTER IS THE CONSEQUENT.

Types of Market Basket Analysis

01


Descriptive Market
Basket Analysis

Predictive Market
Basket Analysis

02


03

Differential Market
Basket Analysis



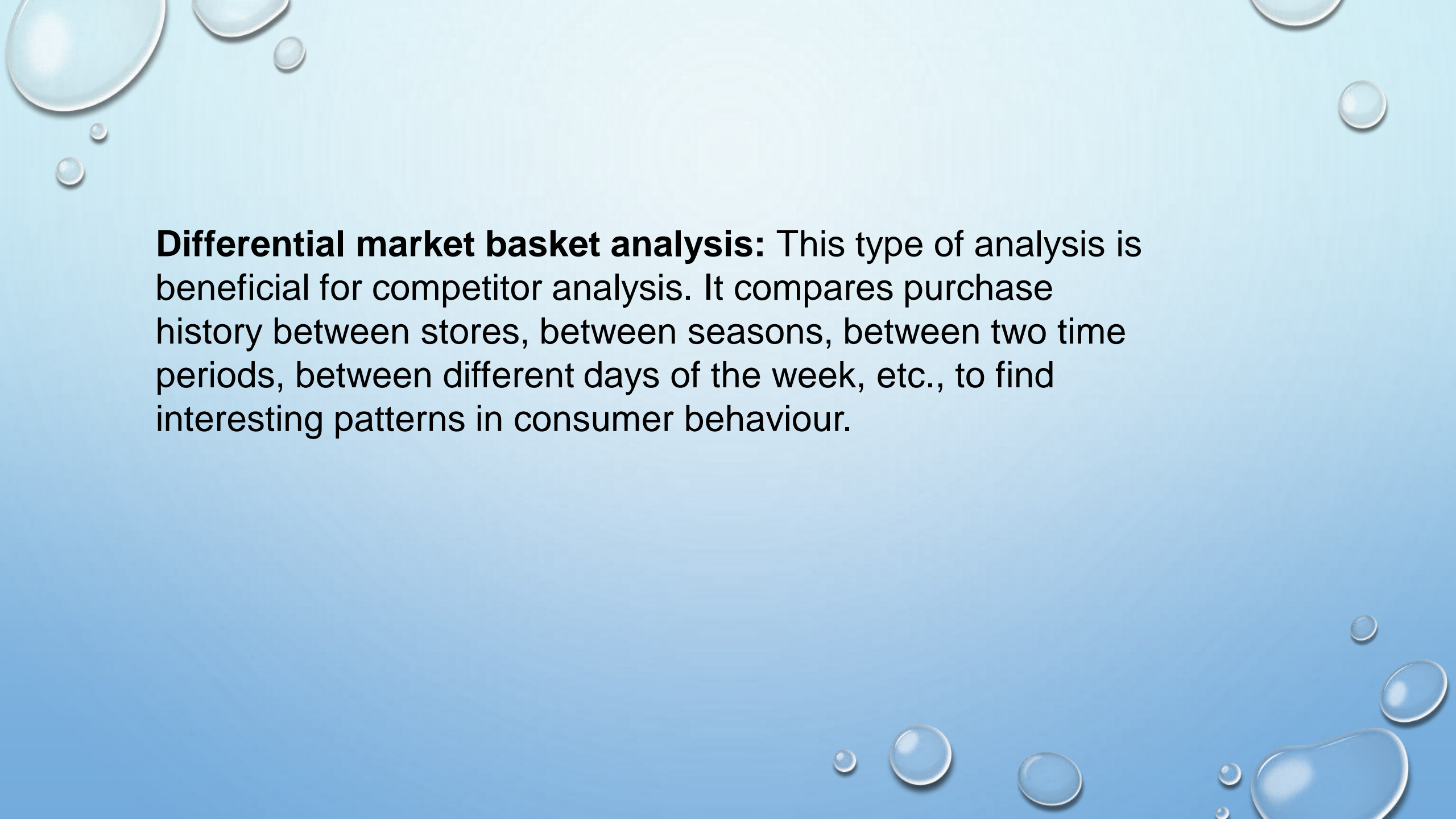
Descriptive market basket analysis: This type only derives insights from past data and is the most frequently used approach.

The analysis here does not make any predictions but rates the association between products using statistical techniques.



Predictive market basket analysis: This type uses supervised learning models like classification and regression. It essentially aims to mimic the market to analyze what causes what to happen.

For example, buying an extended warranty is more likely to follow the purchase of an iPhone.



Differential market basket analysis: This type of analysis is beneficial for competitor analysis. It compares purchase history between stores, between seasons, between two time periods, between different days of the week, etc., to find interesting patterns in consumer behaviour.

Market Basket Analysis Examples

Retail

01

Telecom

02

IBFS

03

Medicine

04

Benefits of Market Basket Analysis



APRIORI ALGORITHM

- THE APRIORI ALGORITHM IN DATA MINING IS A POPULAR ALGORITHM USED FOR FINDING FREQUENT ITEMSETS IN A DATASET.

- THE APRIORI ALGORITHM WAS DEVELOPED BY R. AGRAWAL AND R. SRIKANT IN 1994.
- THE APRIORI ALGORITHM IS USED TO IMPLEMENT FREQUENT PATTERN MINING (FPM).
- FREQUENT PATTERN MINING INVOLVES FINDING SETS OF ITEMS OR ITEMSETS THAT OCCUR TOGETHER FREQUENTLY IN A DATASET.

APRIORI ALGORITHM

- FOR EXAMPLE, IN SALES TRANSACTIONS OF A RETAIL STORE, AN ITEMSET CAN BE REFERRED TO AS PRODUCTS PURCHASED TOGETHER, SUCH AS BREAD AND MILK, WHICH WOULD BE A 2-ITEM SET.
- FOR INSTANCE, THE ALGORITHM MIGHT DISCOVER THAT CUSTOMERS WHO PURCHASE BREAD AND MILK TOGETHER OFTEN ALSO PURCHASE EGGS. THIS INFORMATION CAN BE USED TO RECOMMEND EGGS TO CUSTOMERS WHO PURCHASE BREAD AND MILK IN THE FUTURE.

WHY IS AN APRIORI ALGORITHM SO CALLED?

- THE APRIORI ALGORITHM IS CALLED "APRIORI" BECAUSE IT USES PRIOR KNOWLEDGE ABOUT THE FREQUENT ITEMSETS. THE ALGORITHM USES THE CONCEPT OF "APRIORI PROPERTY," WHICH STATES THAT IF AN ITEMSET IS FREQUENT, THEN ALL OF ITS SUBSETS MUST ALSO BE FREQUENT.

APRIORI PROPERTY

- IF AN ITEMSET APPEARS FREQUENTLY ENOUGH IN THE DATASET TO BE CONSIDERED SIGNIFICANT, THEN ALL OF ITS SUBSETS MUST ALSO APPEAR FREQUENTLY ENOUGH TO BE SIGNIFICANT.
- FOR EXAMPLE, IF THE ITEMSET $\{A, B, C\}$ FREQUENTLY APPEARS IN A DATASET, THEN THE SUBSETS $\{A, B\}$, $\{A, C\}$, $\{B, C\}$, $\{A\}$, $\{B\}$, AND $\{C\}$ MUST ALSO APPEAR FREQUENTLY IN THE DATASET.

APRIORI ALGORITHM COMPONENTS

- **SUPPORT**
- **LIFT**
- **CONFIDENCE**

SUPPORT

- IN THE APRIORI ALGORITHM, SUPPORT REFERS TO THE FREQUENCY OR OCCURRENCE OF AN ITEM SET IN A DATASET.

$$\text{Support}(A) = \frac{\text{Number of Transactions in which } A \text{ occurs}}{\text{Number of all Transactions}}$$

CONFIDENCE

- IN THE APRIORI ALGORITHM, CONFIDENCE IS ALSO A MEASURE OF THE STRENGTH OF THE ASSOCIATION BETWEEN TWO ITEMS IN AN ITEMSET.

$$\text{confidence}(A \Rightarrow B) = P(B/A) = \frac{\text{sup}(A \cup B)}{\text{sup}(A)}$$

LIFT

- IT IS DEFINED AS THE RATIO OF THE SUPPORT OF THE TWO ITEMS OCCURRING TOGETHER TO THE SUPPORT OF THE INDIVIDUAL ITEMS MULTIPLIED TOGETHER. LIFT FOR ANY TWO ITEMS CAN BE CALCULATED USING THE BELOW FORMULA –

$$Lift(A \rightarrow B) = \frac{Support(A \text{ and } B)}{Support(A) * Support(B)}$$

STEPS IN APRIORI ALGORITHM

- **DEFINE MINIMUM SUPPORT THRESHOLD** - THIS IS THE MINIMUM NUMBER OF TIMES AN ITEM SET MUST APPEAR IN THE DATASET TO BE CONSIDERED AS FREQUENT.
- **GENERATE A LIST OF FREQUENT 1-ITEM SETS** - SCAN THE ENTIRE DATASET TO IDENTIFY THE ITEMS THAT MEET THE MINIMUM SUPPORT THRESHOLD. THESE ITEM SETS ARE KNOWN AS FREQUENT 1-ITEM SETS.

CONTT..

- **GENERATE CANDIDATE ITEM SETS** - IN THIS STEP, THE ALGORITHM GENERATES A LIST OF CANDIDATE ITEM SETS OF LENGTH $K+1$ FROM THE FREQUENT K -ITEM SETS IDENTIFIED IN THE PREVIOUS STEP.
- **COUNT THE SUPPORT OF EACH CANDIDATE ITEM SET** - SCAN THE DATASET AGAIN TO COUNT THE NUMBER OF TIMES EACH CANDIDATE ITEM SET APPEARS IN THE DATASET.
- **PRUNE THE CANDIDATE ITEM SETS** - REMOVE THE ITEM SETS THAT DO NOT MEET THE MINIMUM SUPPORT THRESHOLD.
- REPEAT STEPS 3-5 UNTIL NO MORE FREQUENT ITEM SETS CAN BE GENERATED.

CONTT..

- **GENERATE ASSOCIATION RULES** - ONCE THE FREQUENT ITEM SETS HAVE BEEN IDENTIFIED, THE ALGORITHM GENERATES ASSOCIATION RULES FROM THEM. ASSOCIATION RULES ARE RULES OF FORM $A \rightarrow B$, WHERE A AND B ARE ITEM SETS. THE RULE INDICATES THAT IF A TRANSACTION CONTAINS A, IT IS ALSO LIKELY TO CONTAIN B.
- **EVALUATE THE ASSOCIATION RULES** - FINALLY, THE ASSOCIATION RULES ARE EVALUATED BASED ON METRICS SUCH AS CONFIDENCE AND LIFT.

EXAMPLE OF APRIORI ALGORITHM

TID	Items
T1	{milk, bread}
T2	{bread, sugar}
T3	{bread, butter}
T4	{milk, bread, sugar}
T5	{milk, bread, butter}
T6	{milk, bread, butter}
T7	{milk, sugar}
T8	{milk, sugar}
T9	{sugar, butter}
T10	{milk, sugar, butter}

T11

{milk, bread, butter}

Item	Support (Frequency)
milk	8
bread	7
sugar	5
butter	7

Candidate Item Sets	Support (Frequency)
{milk, bread}	5
{milk, sugar}	3
{milk, butter}	5
{bread, sugar}	2
{bread, butter}	3
{sugar, butter}	2

Candidate Item Sets	Support (Frequency)
{milk, bread, sugar}	1
{milk, bread, butter}	3
{milk, sugar, butter}	1

Candidate Item Sets	Support (Frequency)
{milk, bread}	{butter} (Confidence - 60%) ¹
{bread, butter}	{milk} (Confidence - 100%)
{milk, butter}	{bread} (Confidence - 60%)

- Based on association rules mentioned in the above table, we can recommend products to the customer or optimize product placement in retail stores.

REPORT SUPERSTEP

- THE REPORT SUPERSTEP IS THE STEP IN THE ECOSYSTEM THAT ENHANCES THE DATA SCIENCE FINDINGS WITH THE ART OF STORYTELLING AND DATA VISUALIZATION.
- THE MOST IMPORTANT STEP IN ANY ANALYSIS IS THE SUMMARY OF THE RESULTS.

REPORT SUPERSTEP

- WHAT DIFFERENTIATES GOOD DATA SCIENTISTS FROM THE BEST DATA SCIENTISTS ARE NOT THE ALGORITHMS OR DATA ENGINEERING; IT IS THE ABILITY OF THE DATA SCIENTIST TO APPLY THE CONTEXT OF HIS FINDINGS TO THE CUSTOMER.
- DATA VISUALIZATIONS CAN BE USED TO EFFECTIVELY COMMUNICATE THE RESULTS OF ANALYSES AND GUIDE DECISION MAKING.

APPROPRIATE VISUALIZATION

- IT IS TRUE THAT A PICTURE TELLS A THOUSAND WORDS. BUT IN DATA SCIENCE, YOU ONLY WANT YOUR VISUALIZATIONS TO TELL ONE STORY: THE FINDINGS OF THE DATA SCIENCE YOU PREPARED.
- BEWARE OF A LOT OF COLORS AND UNCLEAR GRAPH FORMAT CHOICES. YOU WILL LOSE THE MESSAGE! KEEP IT SIMPLE AND TO THE POINT.

ELIMINATE CLUTTER

- THE BIGGEST TASK OF A DATA SCIENTIST IS TO ELIMINATE CLUTTER IN THE DATA SETS. THERE ARE VARIOUS ALGORITHMS, SUCH AS PRINCIPAL COMPONENT ANALYSIS (PCA), MULTICOLLINEARITY USING THE VARIANCE INFLATION FACTOR TO ELIMINATE DIMENSIONS AND IMPUTE OR ELIMINATE MISSING VALUES.

FREYTAG'S PYRAMID

- FREYTAG USED THESE FIVE PARTS TO ANALYZE THE STRUCTURE:
- EXPOSITION,
- RISING ACTION,
- CLIMAX,
- FALLING ACTION,
- RESOLUTION



FREYTAG'S PYRAMID

- THIS IS USED BY WRITERS OF BOOKS AND SCREENPLAYS AS THE BASIC FRAMEWORK OF ANY STORY.

EXPOSITION AND RISING ACTION

- EXPOSITION IS THE PORTION OF A STORY THAT INTRODUCES IMPORTANT BACKGROUND INFORMATION TO THE AUDIENCE. IN DATA SCIENCE, YOU TELL THE BACKGROUND OF THE INVESTIGATION YOU PERFORMED.
- RISING ACTION REFERS TO A SERIES OF EVENTS THAT BUILD TOWARD THE POINT OF GREATEST INTEREST. IN DATA SCIENCE, YOU POINT OUT THE IMPORTANT FINDINGS OR RESULTS. KEEP IT SIMPLE AND TO THE POINT.

CLIMAX AND FALLING ACTION

- THE CLIMAX IS THE TURNING POINT THAT DETERMINES A GOOD OR BAD OUTCOME FOR THE STORY'S CHARACTERS. IN DATA SCIENCE, YOU SHOW HOW YOUR SOLUTION OR FINDINGS WILL CHANGE THE OUTCOME OF THE WORK YOU PERFORMED.
- DURING THE FALLING ACTION, THE CONFLICT BETWEEN WHAT OCCURRED BEFORE AND AFTER THE CLIMAX TAKES PLACE. IN DATA SCIENCE, YOU PROVE THAT AFTER YOUR SUGGESTION HAS BEEN IMPLEMENTED IN A PILOT, THE SAME TECHNIQUES CAN BE USED TO FIND THE ISSUES NOW PROVING THAT THE ISSUES CAN INEVITABLY BE RESOLVED

RESOLUTION

- RESOLUTION IS THE OUTCOME OF THE STORY. IN DATA SCIENCE, YOU PRODUCE THE SOLUTION AND MAKE THE IMPROVEMENTS PERMANENT.

GRAPHICS

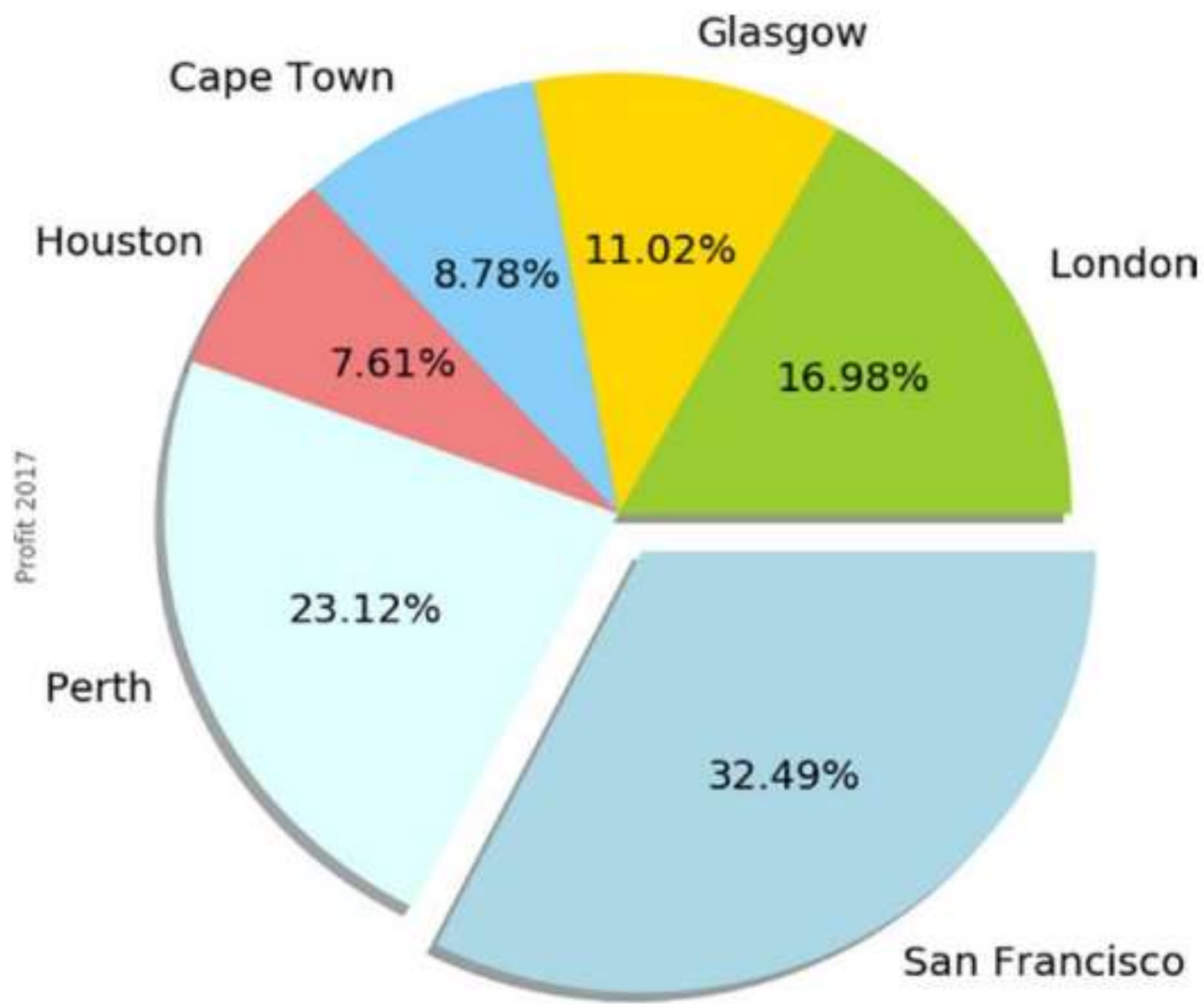
- PIE GRAPH
- LINE GRAPH
- BAR GRAPH
- AREA GRAPH
- SCATTER GRAPH
- HEX BIN GRAPH
- KERNEL DENSITY ESTIMATION (KDE) GRAPH

GRAPHICS

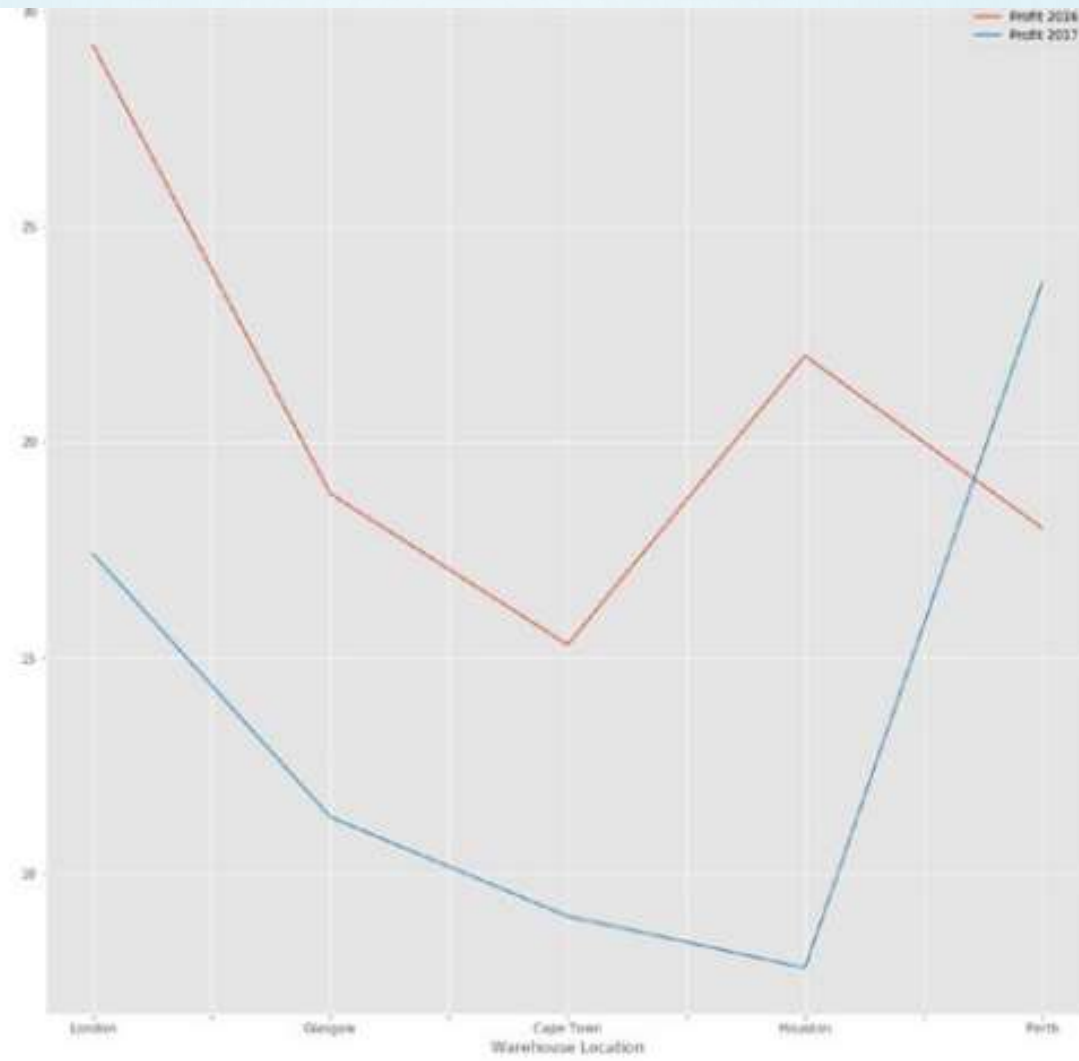
- SCATTER MATRIX GRAPH
- ANDREWS' CURVES
- PARALLEL COORDINATES
- RADVIZ METHOD
- LAG PLOT
- AUTOCORRELATION PLOT

GRAPHICS

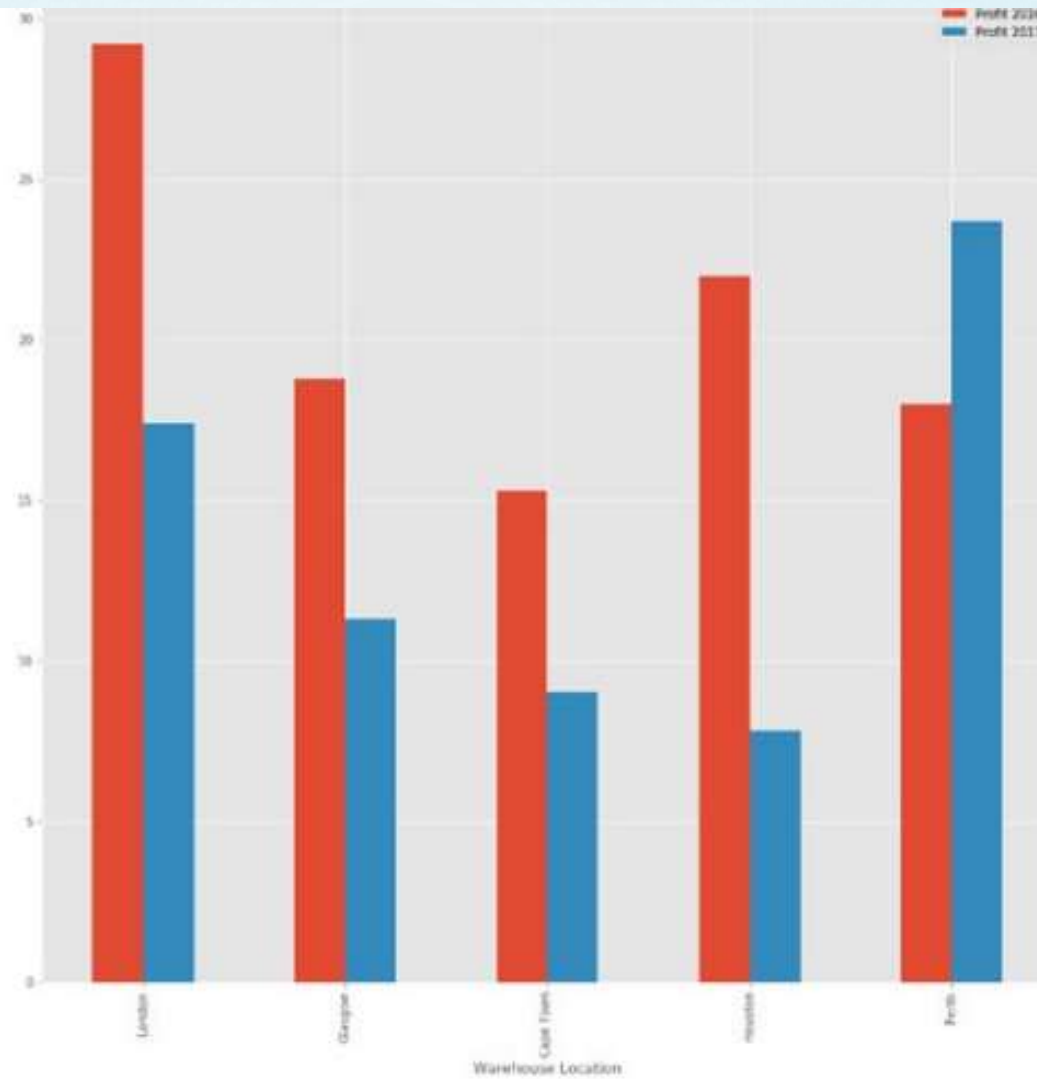
- BOOTSTRAP PLOT
- CONTOUR GRAPHS
- 3D GRAPHS



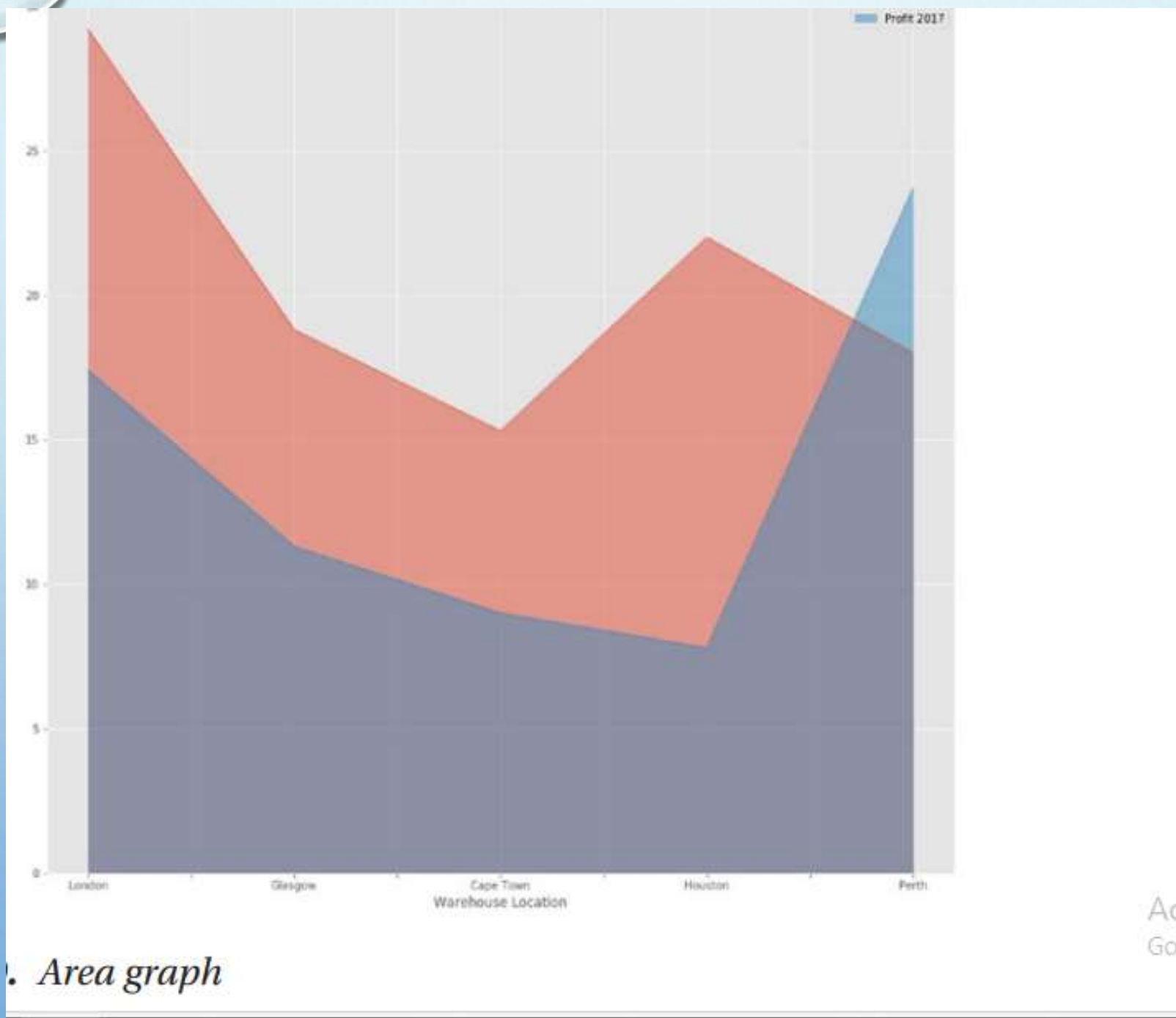
Pie graph



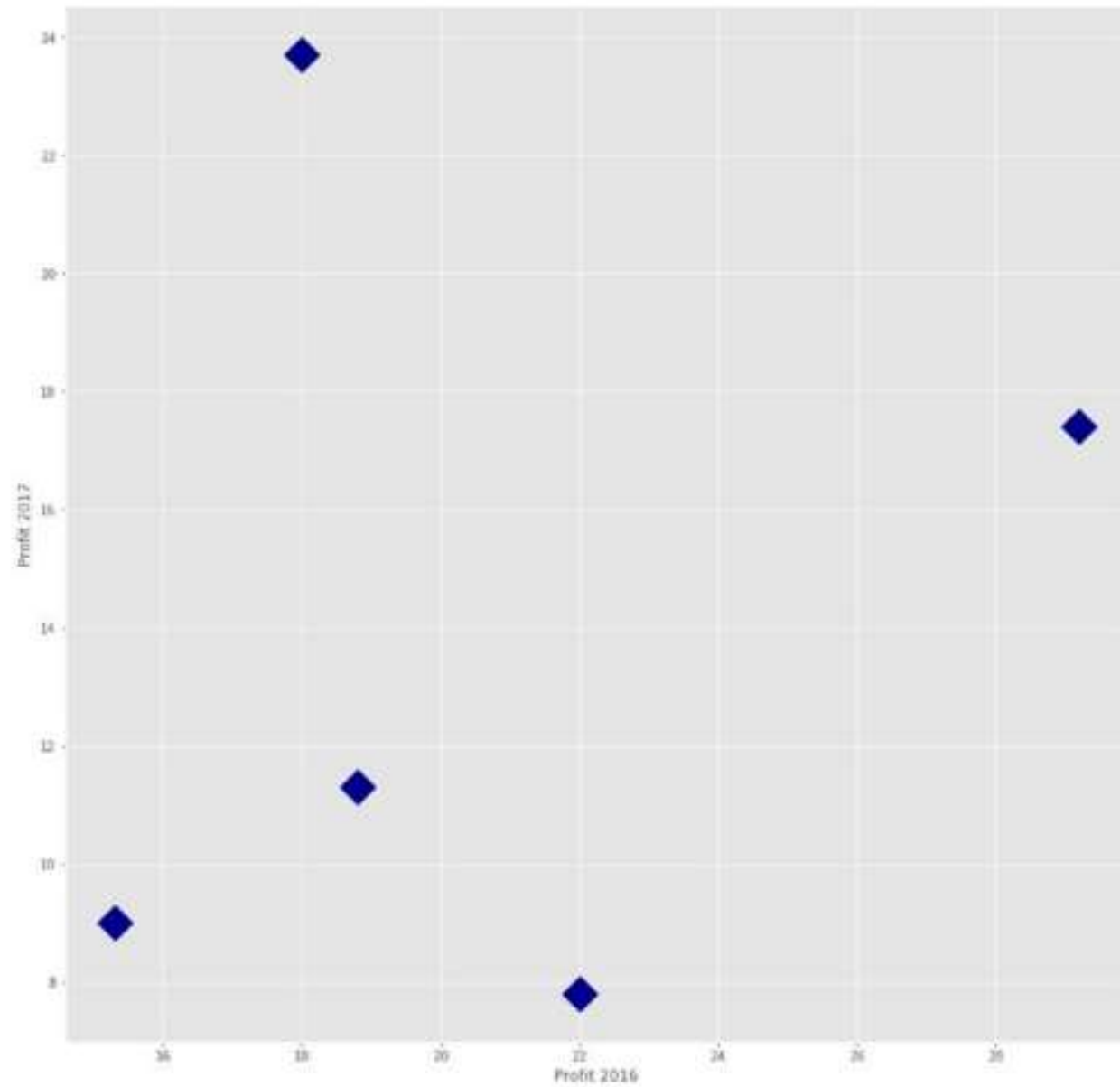
Line graph



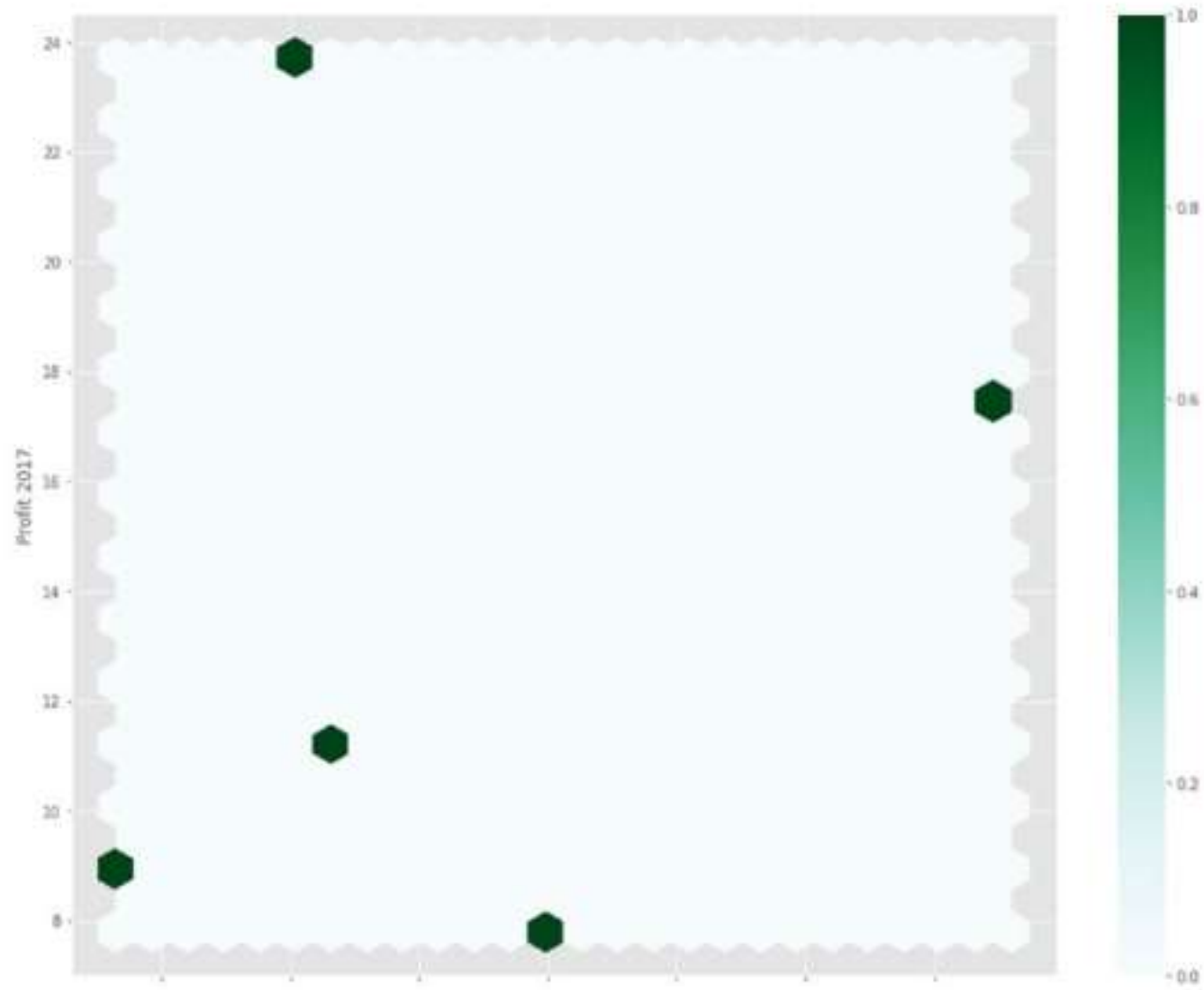
Bar graph



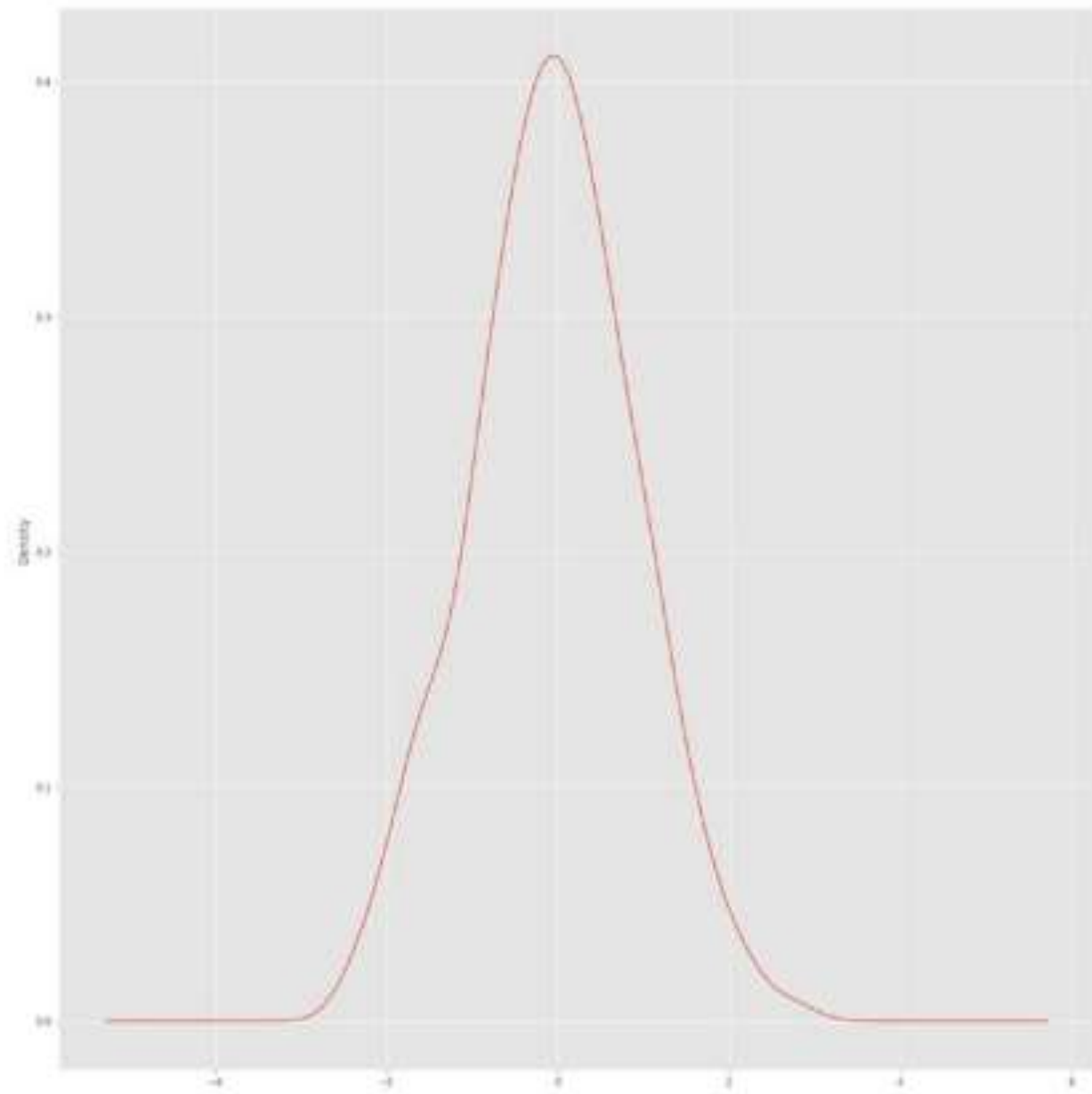
1. Area graph



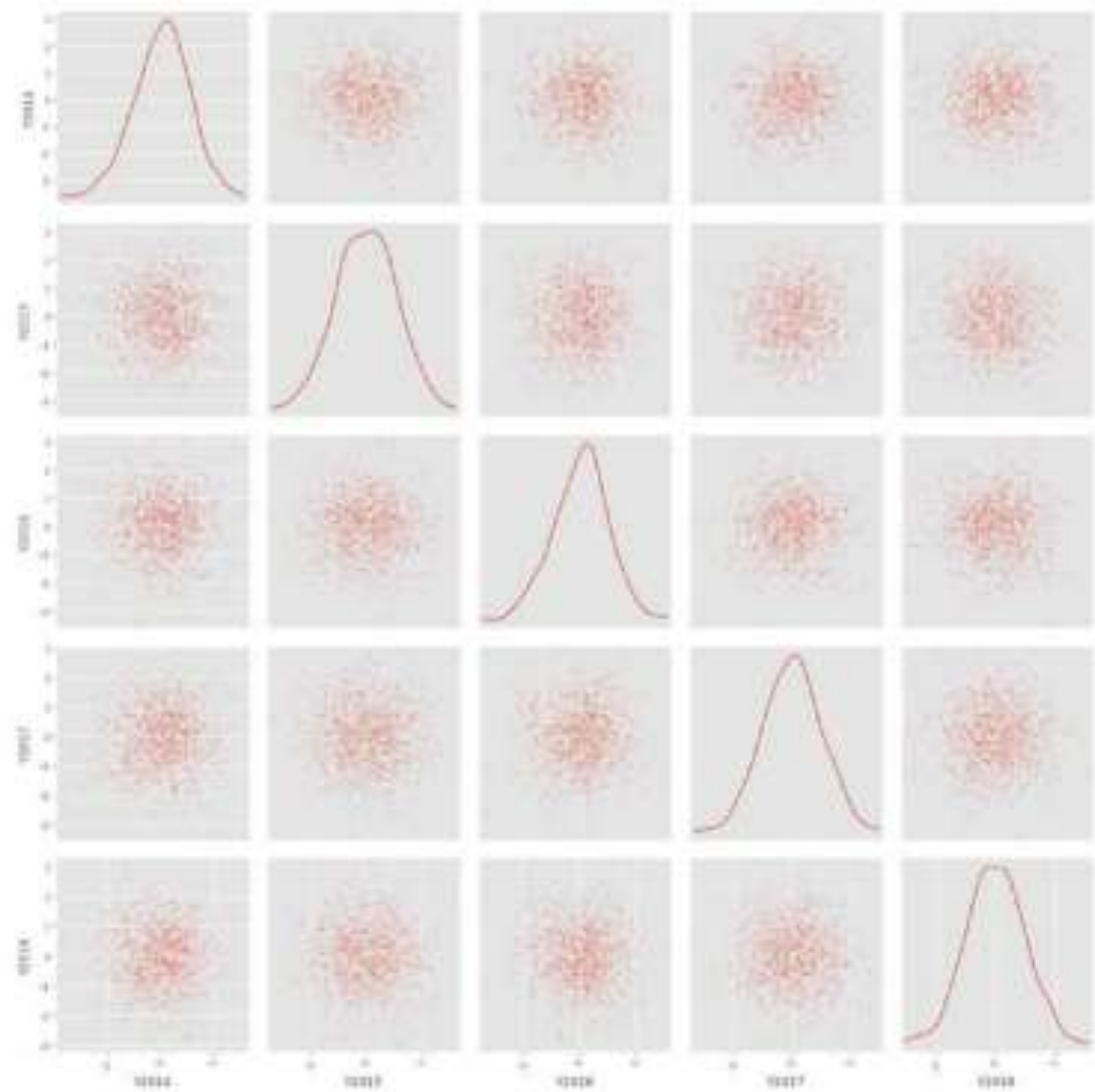
Scatter graph



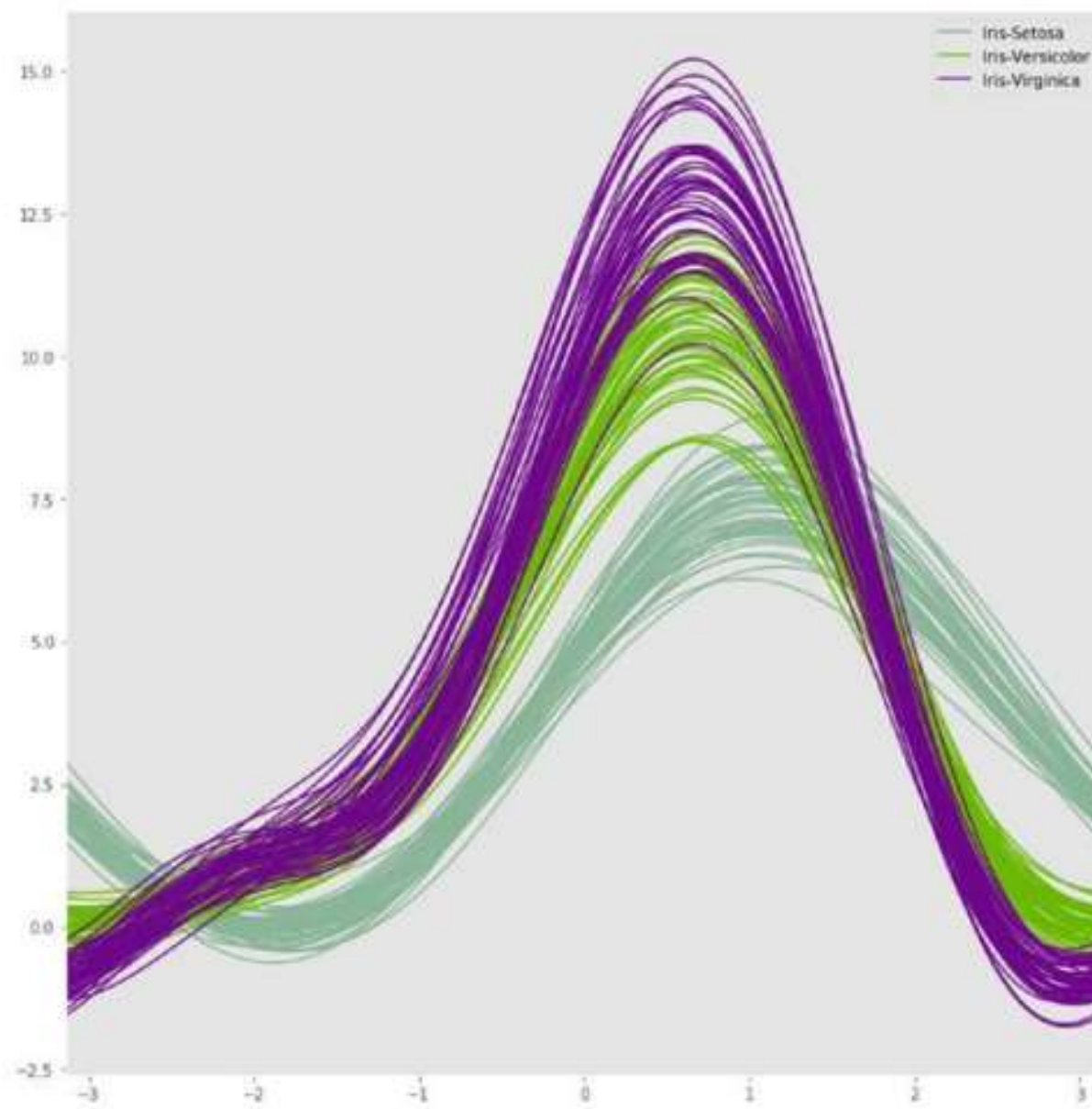
Hex bin graph for farm



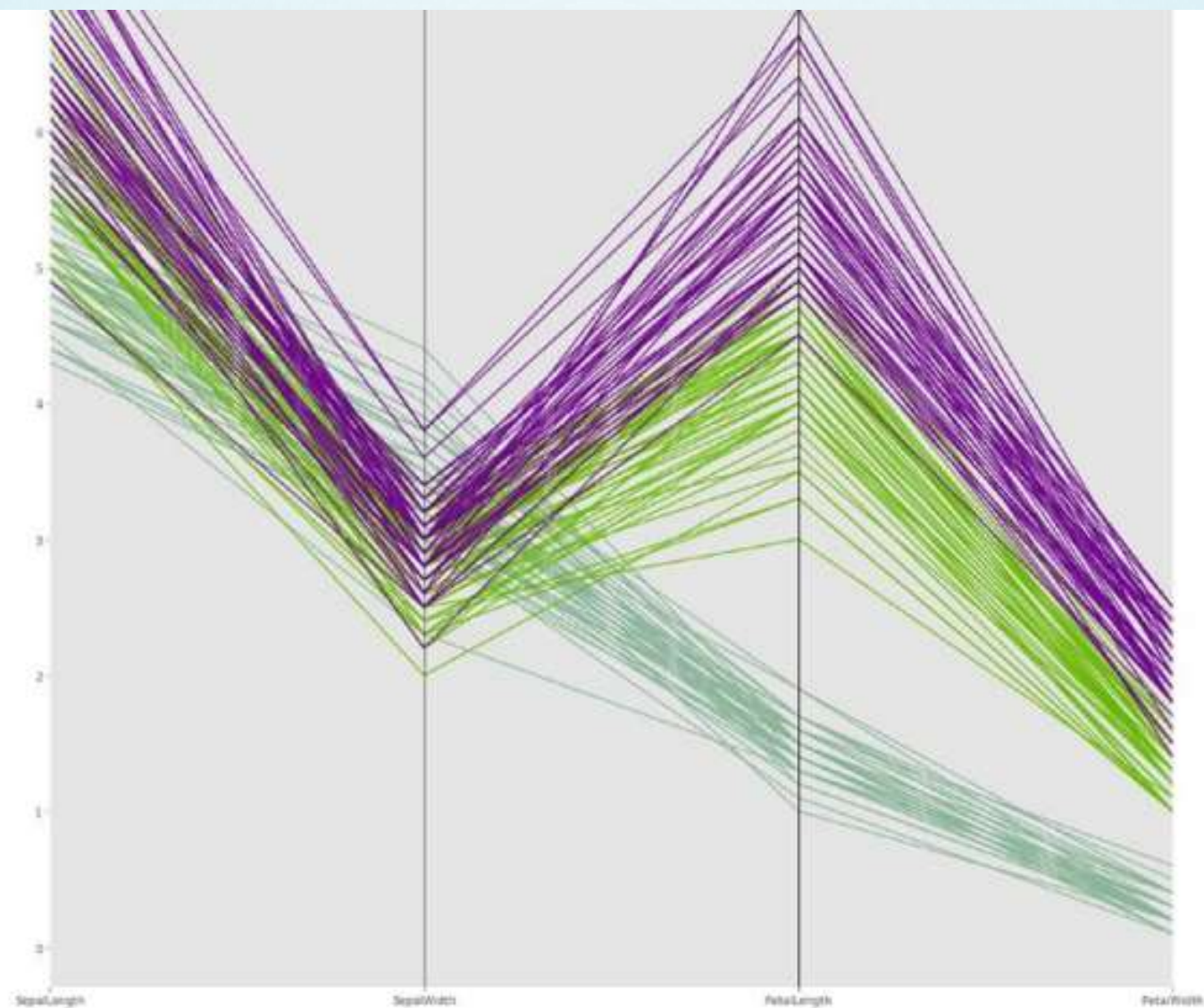
Kernel density estimation graph



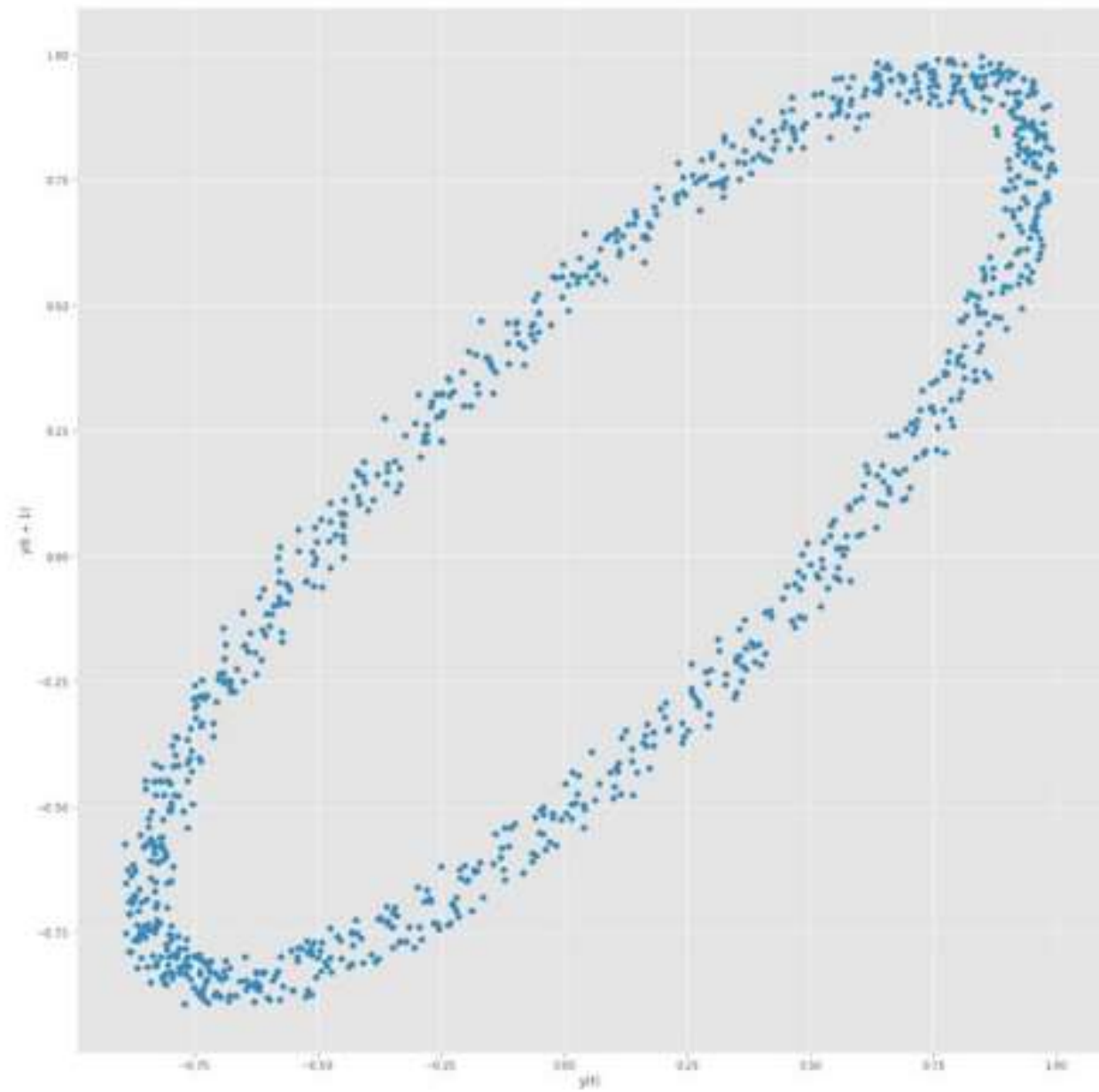
Scatter matrix graph



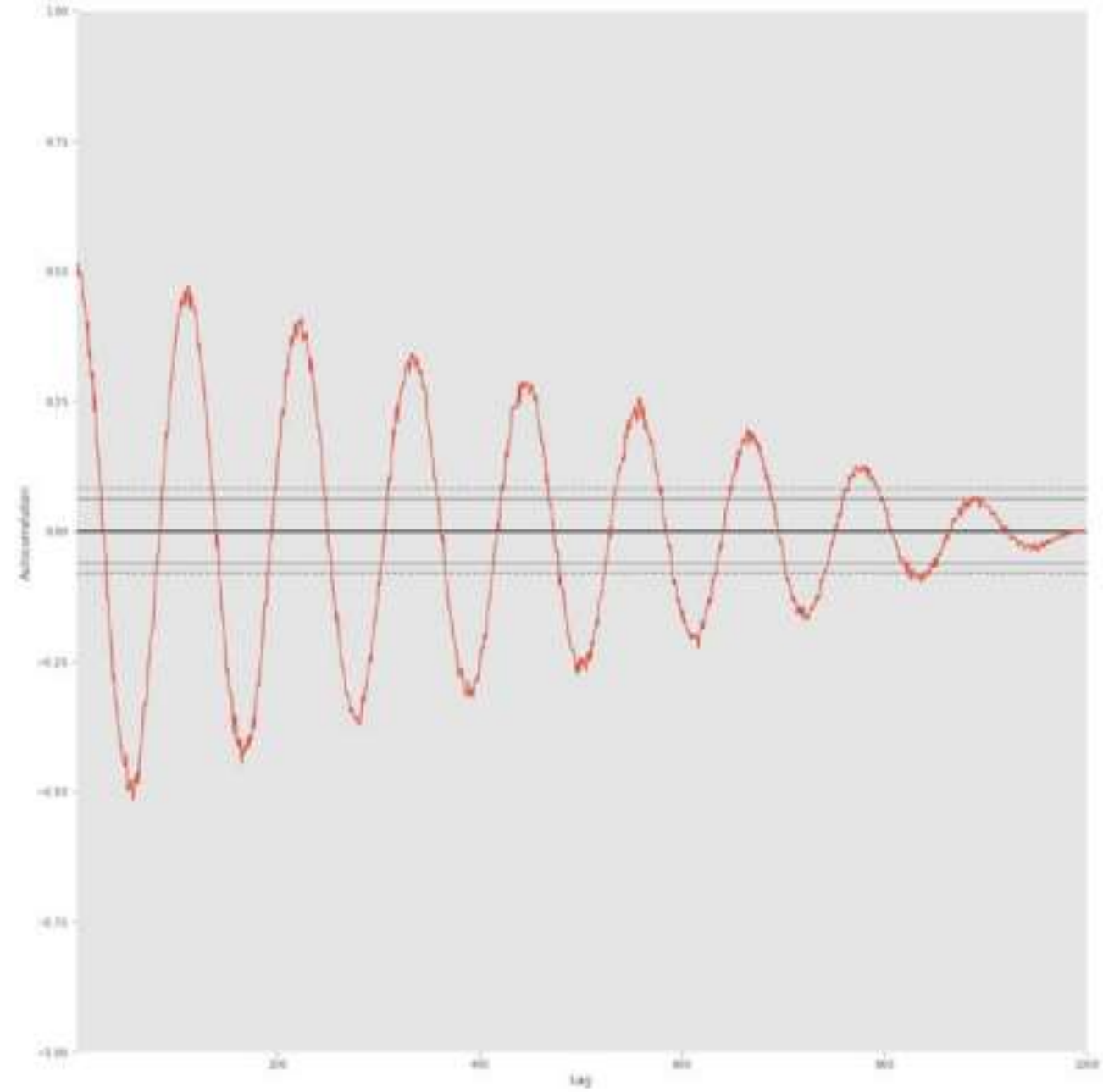
Graph produced using Andrews' curves



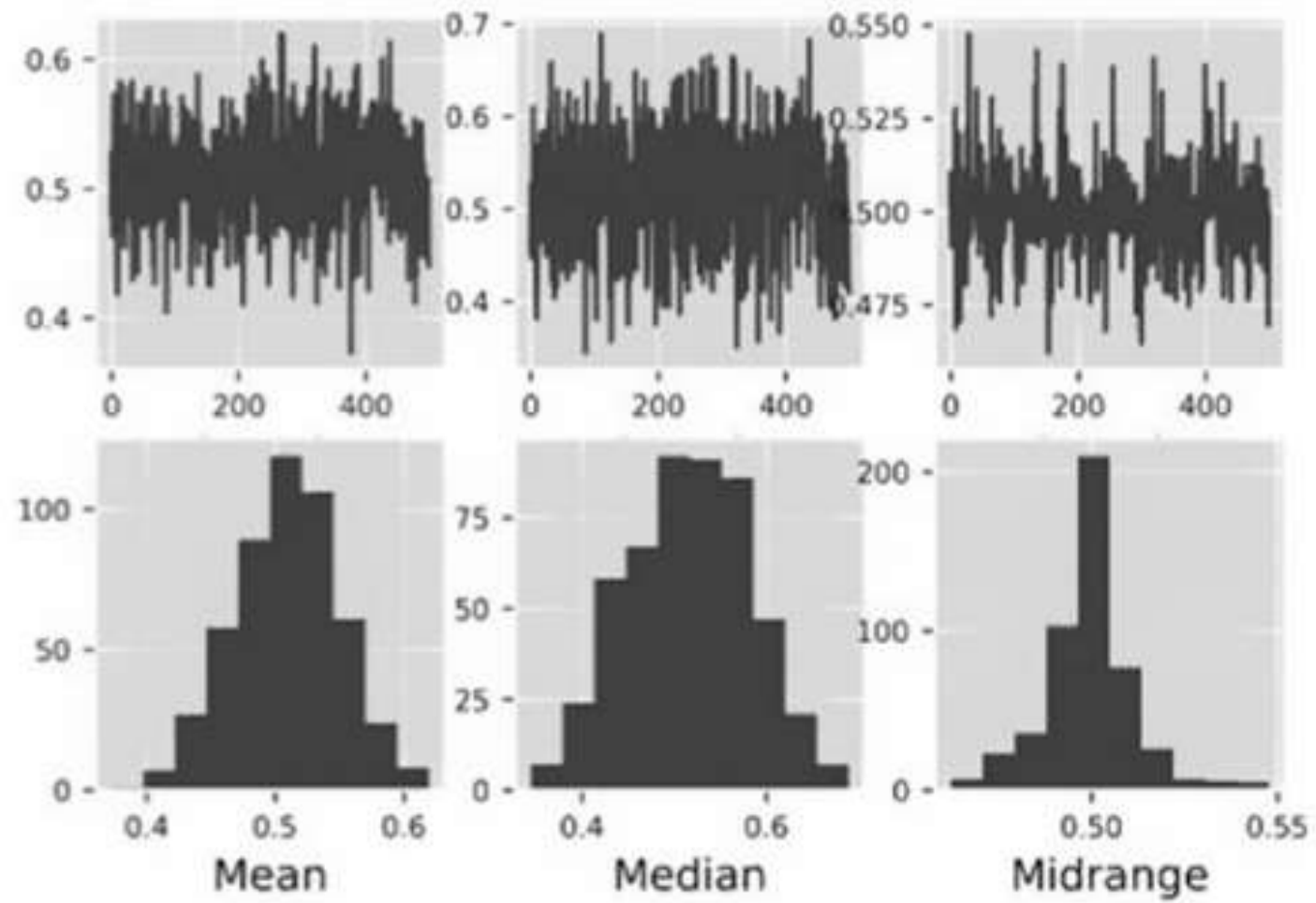
Graph produced using parallel coordinates



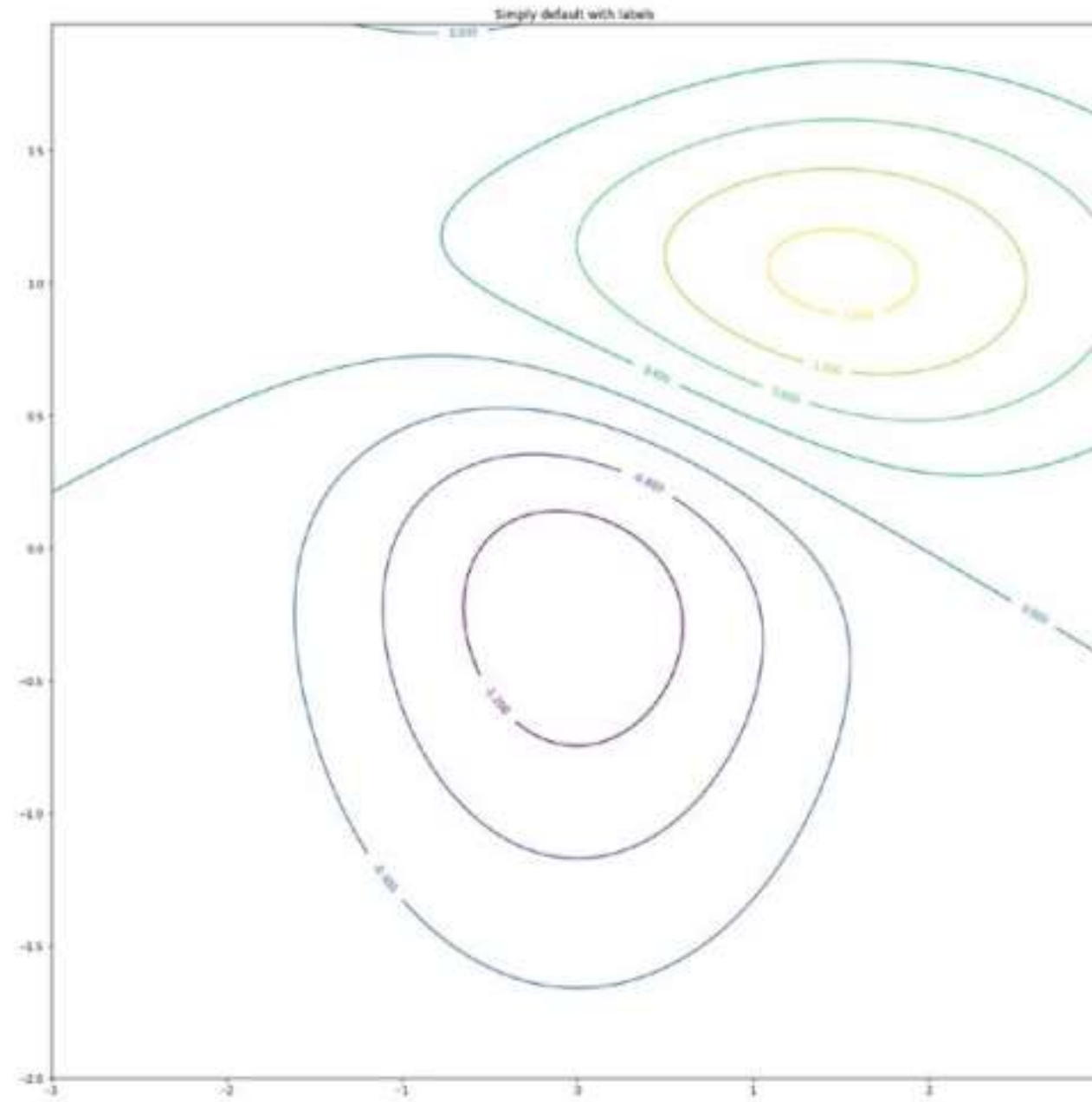
Lag plot graph



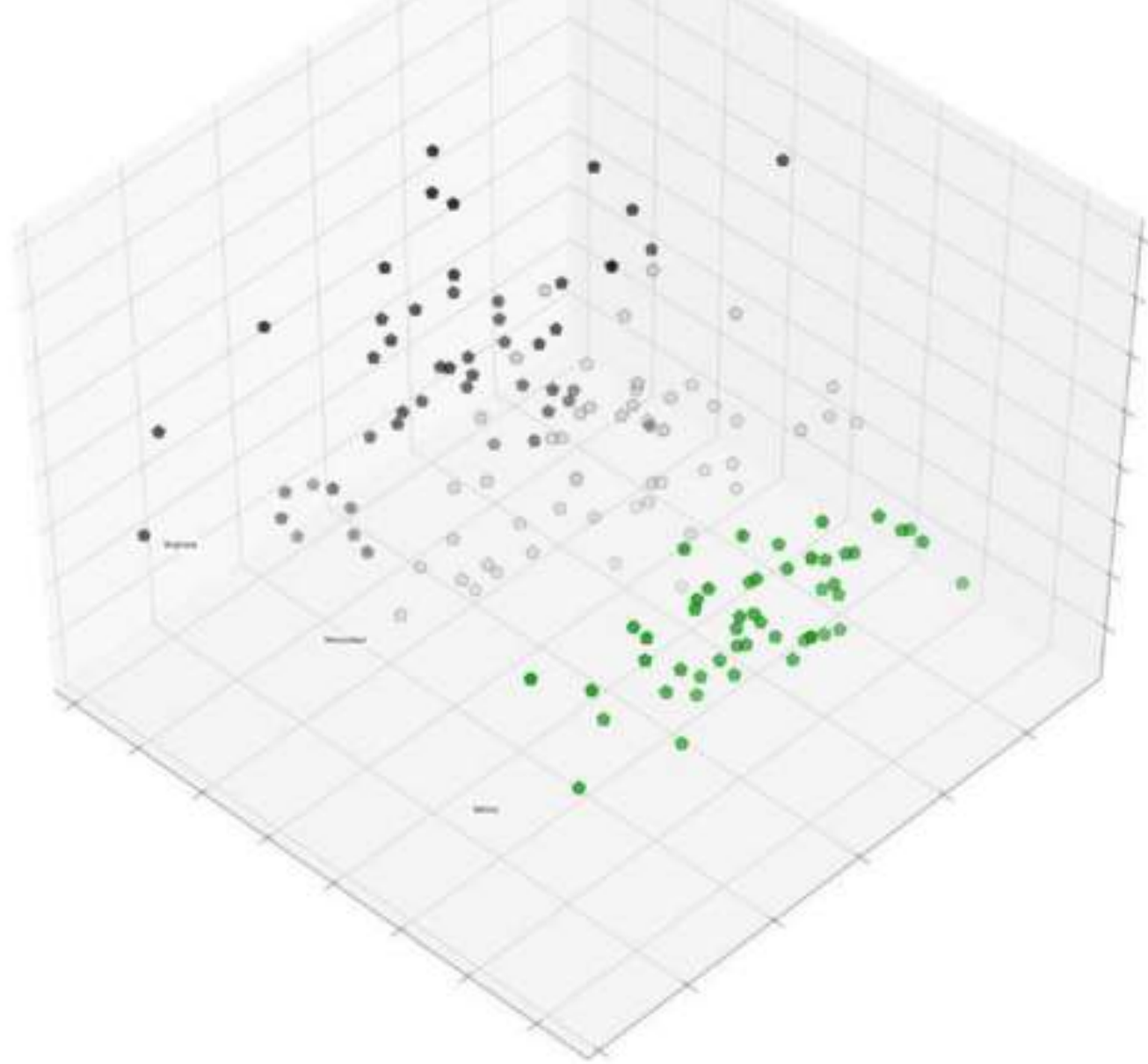
Autocorrelation plot graph



Bootstrap plot graph



• ~ Contour plot graph



Three-dimensional graph

PICTURES

- PICTURES ARE AN INTERESTING DATA SCIENCE SPECIALTY. THE PROCESSING OF MOVIES AND PICTURES IS A SCIENCE ON ITS OWN.

CHANNELS OF IMAGES


- THE INTERESTING FACT ABOUT ANY PICTURE IS THAT IT IS A COMPLEX DATA SET IN EVERY IMAGE. PICTURES ARE BUILT USING MANY LAYERS OR CHANNELS THAT ASSISTS THE VISUALIZATION TOOLS TO RENDER THE REQUIRED IMAGE.


```
import sys
import os
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
```



The image module in matplotlib library is used for working with images in Python.

The image module also includes two useful methods which are `imread` which is used to read images and `imshow` which is used to display the image.






```
import sys
import os
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
sPicName='<u>/content/1.png</u>'
t=0
img=mpimg.imread(sPicName)
print('Size:', img.shape)
plt.figure(figsize=(10, 10))
t+=1
sTitle= '(' + str(t) + ') Original'
plt.title(sTitle)
plt.imshow(img)
plt.show()
for c in range(img.shape[2]):
    t+=1
    plt.figure(figsize=(10, 10))
    sTitle= '(' + str(t) + ') Channel: ' + str(c)
    plt.title(sTitle)
    lum_img = img[:, :, c]
    plt.imshow(lum_img)
    plt.show()
```



THE SHAPE OF AN IMAGE IS ACCESSED BY `IMG.SHAPE`. IT RETURNS A TUPLE OF THE NUMBER OF ROWS, COLUMNS, AND CHANNELS (IF THE IMAGE IS COLOR):

IF AN IMAGE IS GRAYSCALE, THE TUPLE RETURNED CONTAINS ONLY THE NUMBER OF ROWS AND COLUMNS, SO IT IS A GOOD METHOD TO CHECK WHETHER THE LOADED IMAGE IS GRAYSCALE OR COLOR.



(2) Channel: 0

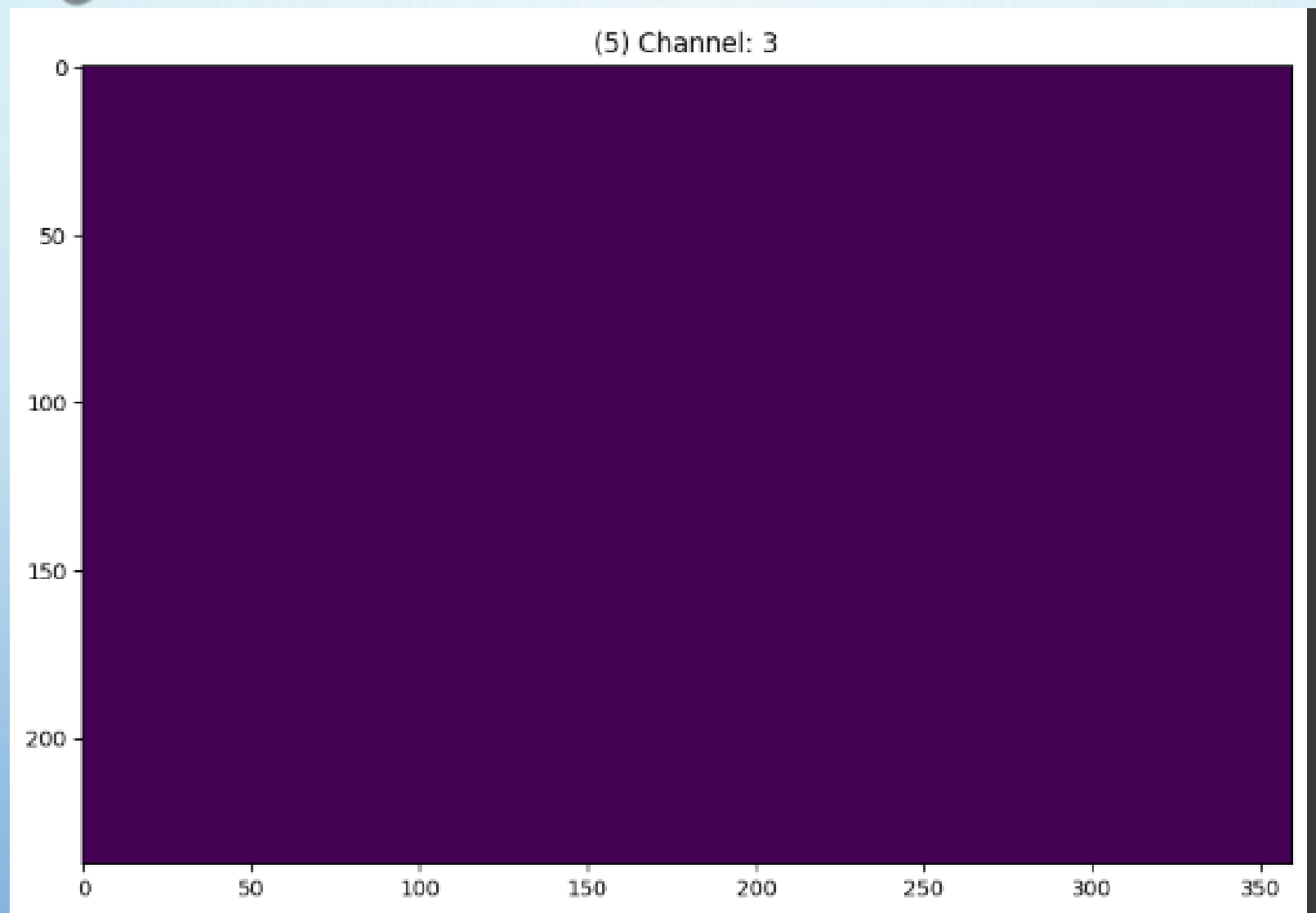


(3) Channel: 1



(4) Channel: 2





OBSERVATIONS

- YOU CAN CLEARLY SEE THAT THE IMAGE IS OF THE FOLLOWING SIZE: SIZE: (238, 360, 4)
- NUMBER OF CHANNELS = 4
- THIS MEANS YOU HAVE 238×360 PIXELS PER CHANNEL.
- THAT IS A TOTAL OF 85,680 PIXELS PER LAYER.

CUTTING THE EDGES

- ONE OF THE MOST COMMON TECHNIQUES THAT MOST DATA SCIENCE PROJECTS REQUIRE IS THE DETERMINATION OF THE EDGE OF AN ITEM'S IMAGE. THIS IS USEFUL IN AREAS SUCH AS ROBOTICS OBJECT SELECTION AND FACE RECOGNITION.

```
import matplotlib.pyplot as plt
from PIL import Image
```

```
sPicNameIn='/content/1.png'
sPicNameOut='/content/audi.png'
imageIn = Image.open(sPicNameIn)
fig1=plt.figure(figsize=(10, 10))
fig1.suptitle('Audi R8', fontsize=20)
imgplot = plt.imshow(imageIn)
mask=imageIn.convert("L")
th=49
```

```
imageOut = mask.point(lambda i: i < th and 255)
imageOut.save(sPicNameOut)
```

```
imageTest = Image.open(sPicNameOut)
fig2=plt.figure(figsize=(10, 10))
fig2.suptitle('Audi R8 Edge', fontsize=20)
imgplot = plt.imshow(imageTest)
```



Audi R8 Edge



ONE SIZE DOES NOT FIT ALL

- THE IMAGES WE GET TO PROCESS ARE MOSTLY OF DIFFERENT SIZES AND QUALITY. YOU WILL HAVE TO SIZE IMAGES TO SPECIFIC SIZES FOR MOST OF YOUR DATA SCIENCE.

The background is a light blue gradient with several realistic water droplets of various sizes scattered around the edges. The droplets have highlights and shadows, giving them a 3D appearance.

**WHAT HAPPENS TO AN IMAGE IF YOU REDUCE A
PIXEL QUALITY**

```
import sys
import os
import matplotlib.pyplot as plt
from PIL import Image
sPicName='<u>/content/1.png</u>'
nSize=4
img = Image.open(sPicName)
plt.figure(figsize=(nSize, nSize))
sTitle='Unchanges'
plt.title(sTitle)
imgplot = plt.imshow(img)
```

Unchanges



The background is a light blue gradient. There are several realistic-looking water droplets of various sizes in the corners. Top-left: a large droplet and a small one. Top-right: a medium droplet. Bottom-left: a small droplet. Bottom-right: a cluster of droplets including a large one, a medium one, and several small ones.

YOU NOW APPLY A THUMBNAIL FUNCTION THAT
CREATES A 64×64 PIXEL THUMBNAIL IMAGE.

APPLY A THUMBNAIL FUNCTION


```
#thumbnail

img.thumbnail((64, 64), Image.ANTIALIAS)
# resizes image in-place
plt.figure(figsize=(nSize, nSize))
sTitle='Resized'
plt.title(sTitle)
imgplot = plt.imshow(img)
plt.figure(figsize=(nSize, nSize))
sTitle='Resized with Bi-Cubic'
plt.title(sTitle)
imgplot = plt.imshow(img, interpolation="bicubic")
print('### Done!! #####')
```



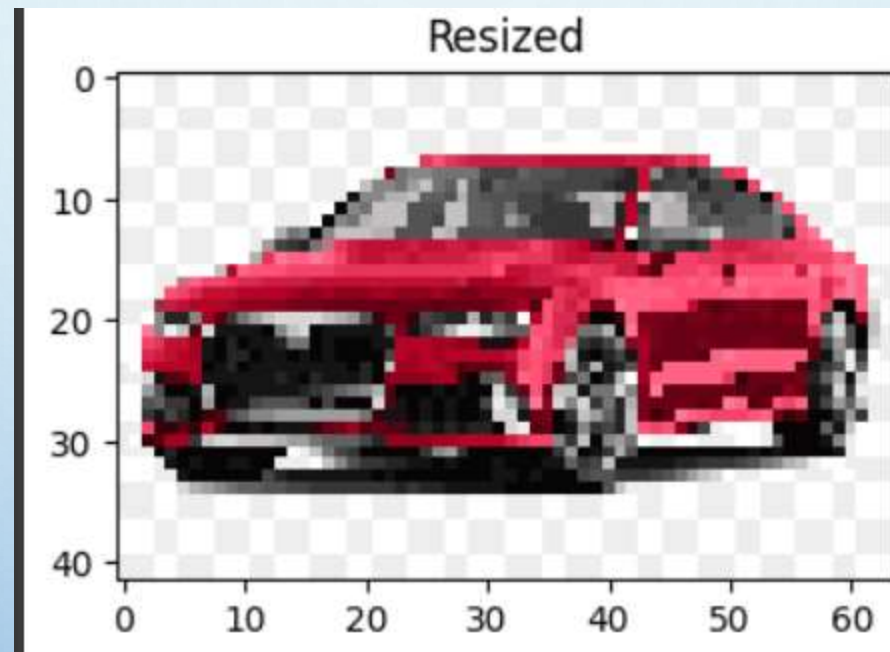

INTERPOLATION IN PYTHON IS A TECHNIQUE USED TO
ESTIMATE UNKNOWN DATA POINTS BETWEEN TWO
KNOWN DATA POINTS.

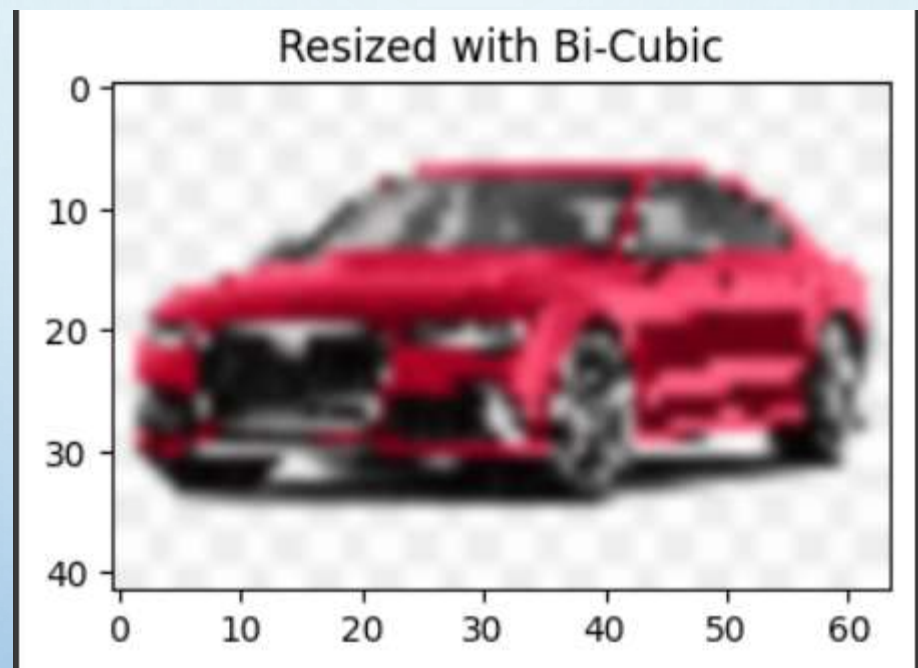
IN PYTHON, INTERPOLATION IS A TECHNIQUE MOSTLY USED
TO IMPUTE MISSING VALUES IN THE DATA FRAME OR SERIES
WHILE PREPROCESSING DATA.





**BICUBIC INTERPOLATION DETERMINES THE PIXEL
VALUE FROM THE WEIGHTED AVERAGE OF THE 16
CLOSEST NEIGHBORING PIXELS.**





The background is a light blue gradient. In the top-left and bottom-right corners, there are several realistic water droplets of various sizes, some overlapping. The word "THANKS" is centered in the middle of the image.

THANKS