



ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCE

(Autonomous)

G.B. Nagar, Kalavai - 632506



Big data analytics

UNIT - I

INTRODUCTION TO BIG DATA

We generate 2.5 quintillion bytes of data.

Ninety percentage data created in the last 2 yrs

Strong need for analytical skills and resources

Managing and analyzing these data resources.

Analytics in a Big Data world

Less than 1 MB	(12)	3.7%	$(12/322) * 100$
1.1 to 10 MB	(8)	2.5%	
11 to 100 MB	(14)	4.3%	
101 MB to 1 GB	(50)	15.5%	
1.1 to 10 GB	(59)	18%	
11 to 100 GB	(52)	16%	
101 GB to 1 TB	(59)	18%	
1.1 to 10 TB	(39)	12%	
11 to 100 TB	(15)	4.7%	
101 TB to 1 PB	(6)	1.9%	
1.1 to 10 PB	(2)	0.6%	
11 to 100 PB	(0)	0%	
Over 100 PB	(6)	1.9%	

Data collection from 322 people

Basic Nomenclature

Customer can play different roles

Payer and the end user

**Primary account owner, secondary account owner
main debtor, guarantor**

Target variable needs to be appropriately defined.

customer considered to be a churner or not

a fraudster or not

a responder or not

How should the CLV be appropriately defined?

Analytics process model

- **To define the business problem**
- **All source data need to be identified**
- **OLAP for data analysis (e.g., roll-up, drill down, slicing and dicing)**

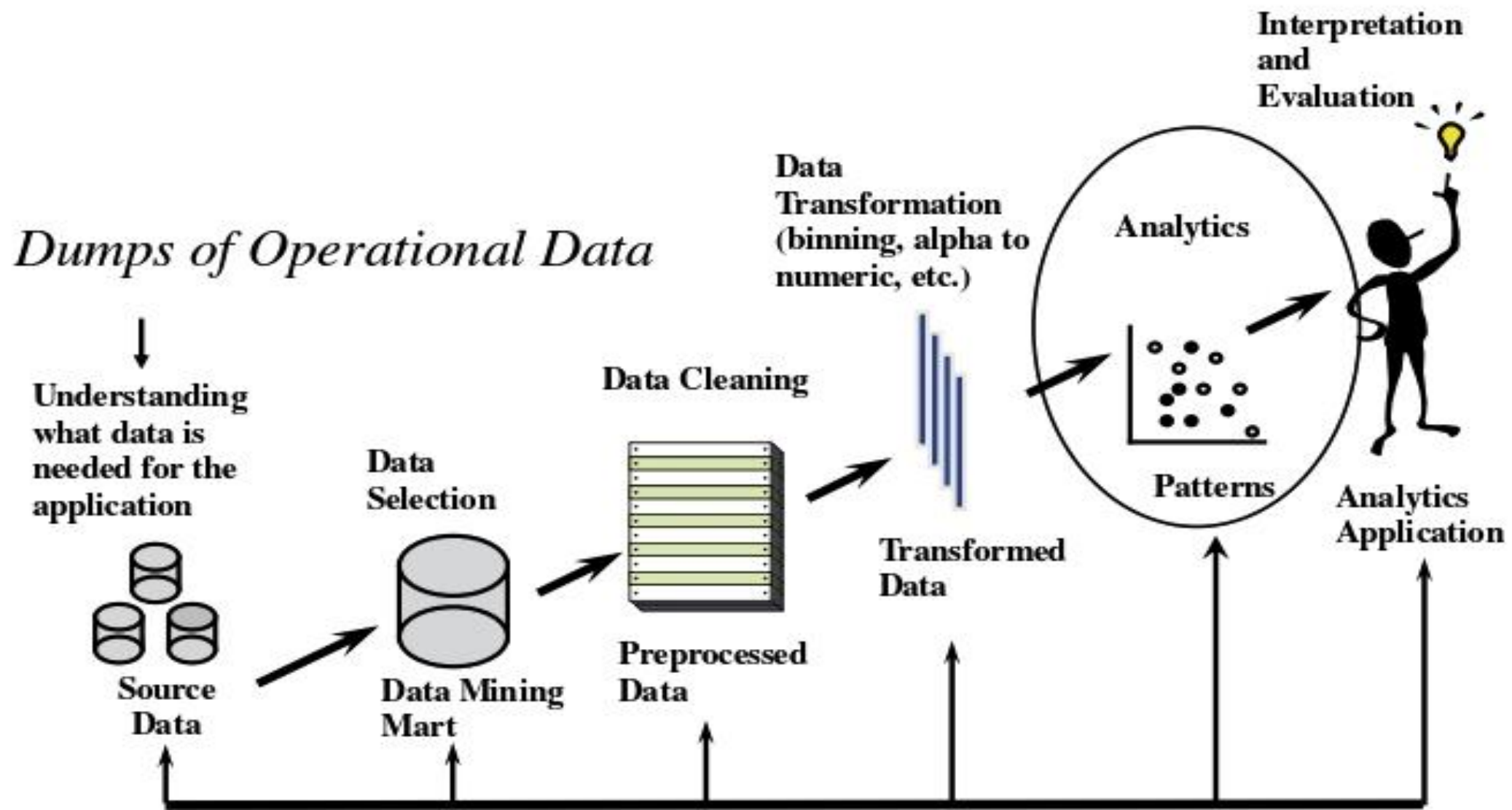


Figure 1.2 The Analytics Process Model

The roll-up operation performs aggregation on a data cube

(DL: 12, New York : 8, TN : 13, Los Angeles : 12)

Total : India : ____ USA: ____

Drill-down is the reverse of roll-up. That means lower level summary to higher level summary.

- India : 25 (DL 12; TN 13)**
- USA: 20 (New York 8 ; Los angels 12)**

**The slice operation performs a selection on one dimension of the given cube,
if we want to make a select where Medal = 12
result : _____**

Dice operation defines a sub-cube by performing a selection on two or more dimensions.

**For example, if we want to make a select where
Medal = 12 or Location = TN**

• result:

Data Cleaning

To get clear all inconsistencies

(missing values, outliers, duplicate data)

Data Transformation

Binning, alphanumeric to numeric coding

Geographical aggregation (bar chart, line chart)

Analytics

Fraud detection

customer segmentation (Basic, Gold, Platinum)

To do churn prediction (Who have left your brand)

Market basket analysis

- **Once the model has been built, it well be interpreted and evaluated by the business experts.**
- **Example :**
 - ✓ **Market basket analysis setting**
 - ✓ **To find frequent sales items are often purchased together. Apple and Orange (Or)
Apple and Lemon**
- **Analytical model has bee approved, if finalized.**

Analytics

- **Analytics is a term that is often used interchangeably with data science, data mining, knowledge discovery.**
- **Predictive analytics : Target variable (Class A or B)**
- **Descriptive analytics : Framing Rules**

Analytics Model Requirements

- **Analytics model requirements, depending on the application area.**
- **Model should actually solve the business problem for which it was developed**
- **Problem to be solved is appropriately defined, qualified, and agreed**
- **To analysis statistical performance**

Analytical models should also be interpretable and justifiable.

Interpretability is understanding the patterns.

Interpretability may depend on the business user's knowledge.

- **Justifiability is prior business knowledge.**
- **Analytical models should also be *operationally efficient*.**
- **Needed to collect the data, preprocess it, evaluate the model and feed its outputs to business**

- **Analytical models should also comply with both local and international regulations and legislation (law)**
- **New regularity development are taking place at various level**

TYPE OF DATA SOURCES

- Data can originate from a variety of diff. sources**
- Transactional data consists of structured, low-level, detailed information**
- (e.g : purchase items, claim amount, cash, payment mode)**
- This type of data is stored in OLTP relational DB.**

- Unstructured data fixed in text document**
- (e.g emails, web pages, claim forms)**
- Source of data is qualitative, expert-based data.**
- Unqualified data are not include analytical process (e.g not maintain minimum value)**
- Poolers (Data warehousing) are more important in the industry**

- Social media data from Facebook, Twitter, and others can be an important source of information.**
- Data gathering respects both local and international privacy regulations.**

SAMPLING

- **To take a subset of past customer data and use that to build an analytical model.**
- **A key requirement for a good sample.**
- **Time aspect becomes important**
- **Sample involves a between lots of data and recent data.**
- **Taken from an average business period.**

- **A sample is taken according to predefined levels.**
- **Sample will contain exactly the same percentages of customer and non-customer as in the original data.**

TYPES OF DATA ELEMENTS

- To appropriately consider the different types of data elements at the start of the analysis.**

Types of data elements can be considered:

Continuous

Categorical

Nominal, Ordinal, Binary

Continuous

- **These are data elements that are defined on an interval that can be limited or unlimited.**

Example :

- **Income**
- **Sales**
- **RFM (Recency, Frequency, Monetary)**

Categorical

- Nominal

These are data elements that can only take on a limited set of values with no meaningful ordering in between. Examples include marital status, profession, purpose of loan.

Ordinal:

These are data elements that can only take on a limited set of values with a meaningful ordering in between.

Examples

Credit rating

Age coded as young, middle aged, and old.

Binary:

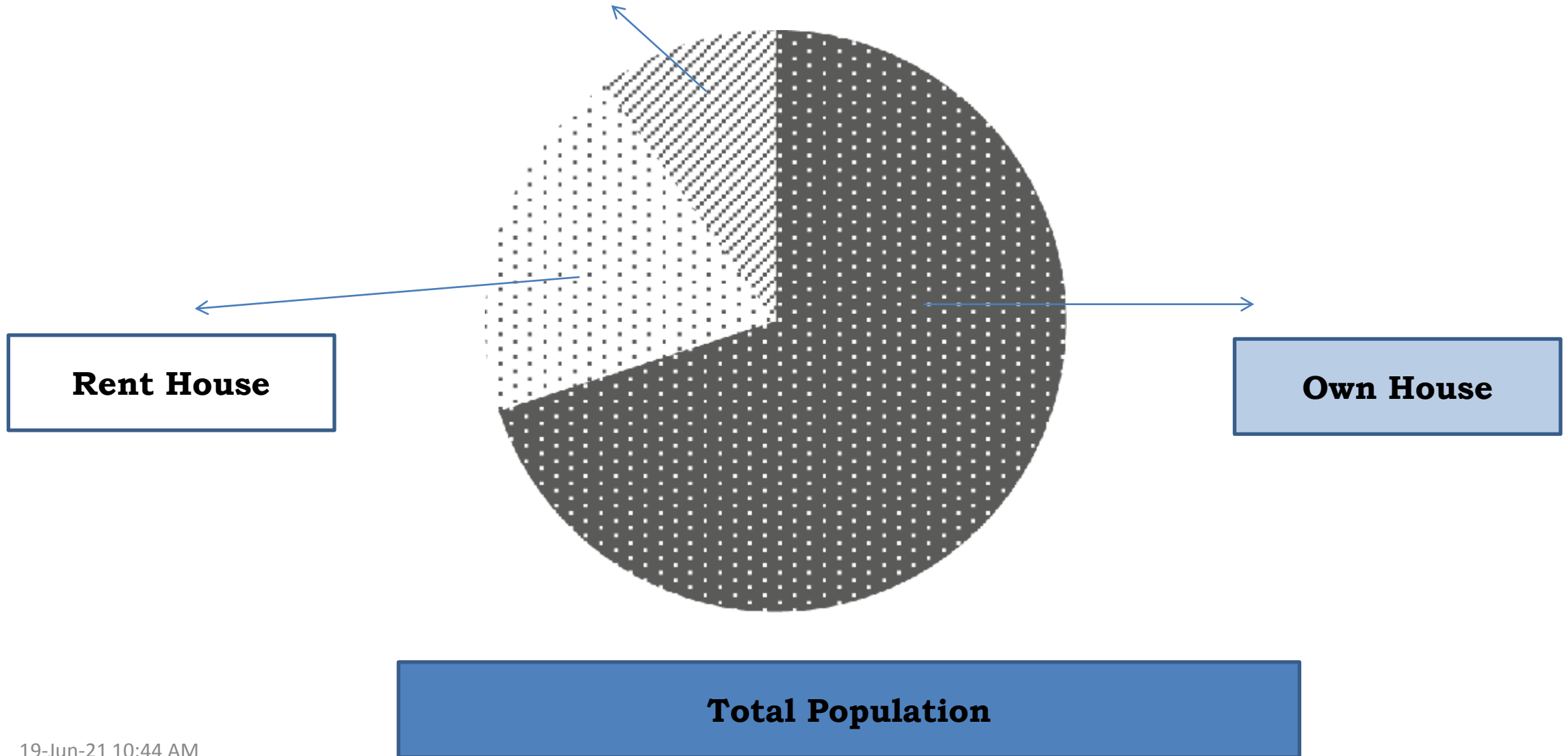
These are data elements that can only take on two values.

Examples include gender, employment status.

Visual Data Exploration and Exploratory Statistical Analysis

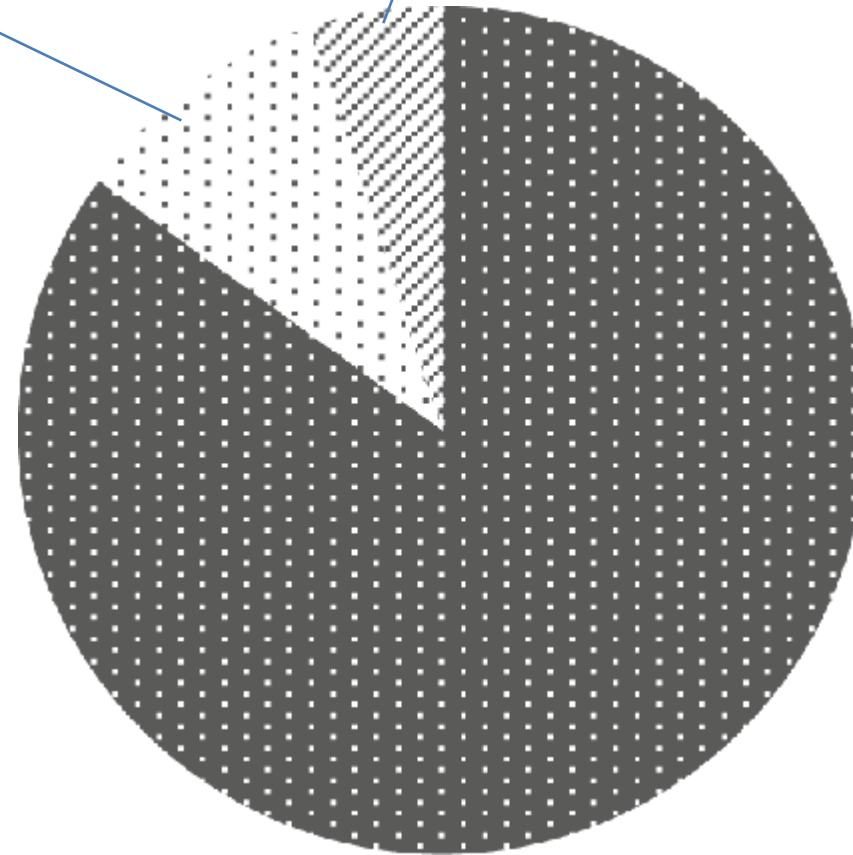
- To know your data in an “informal” way.**
- Different plots/graphs can be used**
- Example is pie charts**
- Represents the portion of the total percent taken
by each value of the variable**

**For free House
(Stay with parents)**



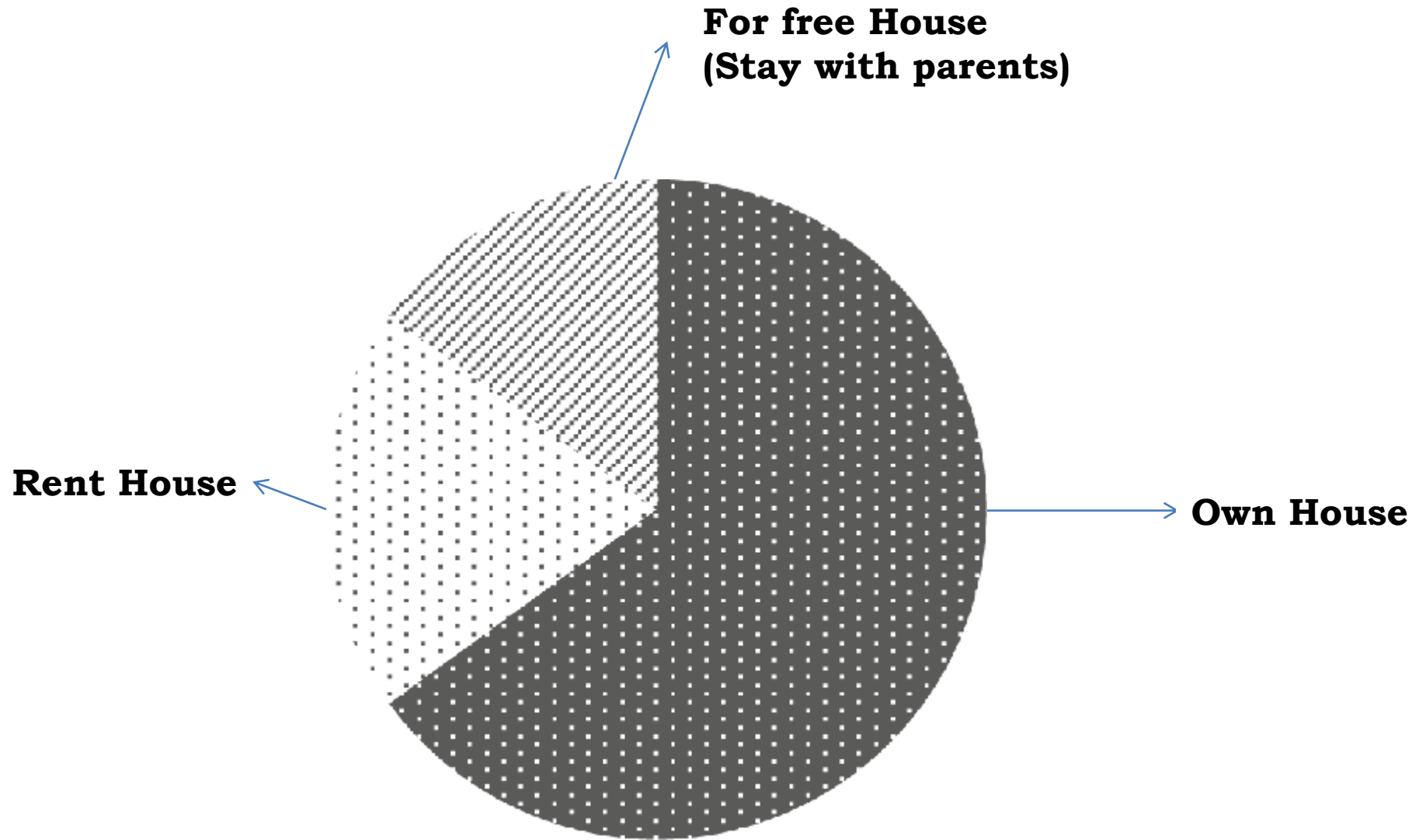
Rent House

**For free House
(Stay with parents)**



Own House

Good Population



Bad Population

- **From chart to analysis more good populations own house property than bad populations**
- **Bar charts represent the frequency of each of the values as bars.**
- **Other handy visual tools are histograms and scatter plots.**

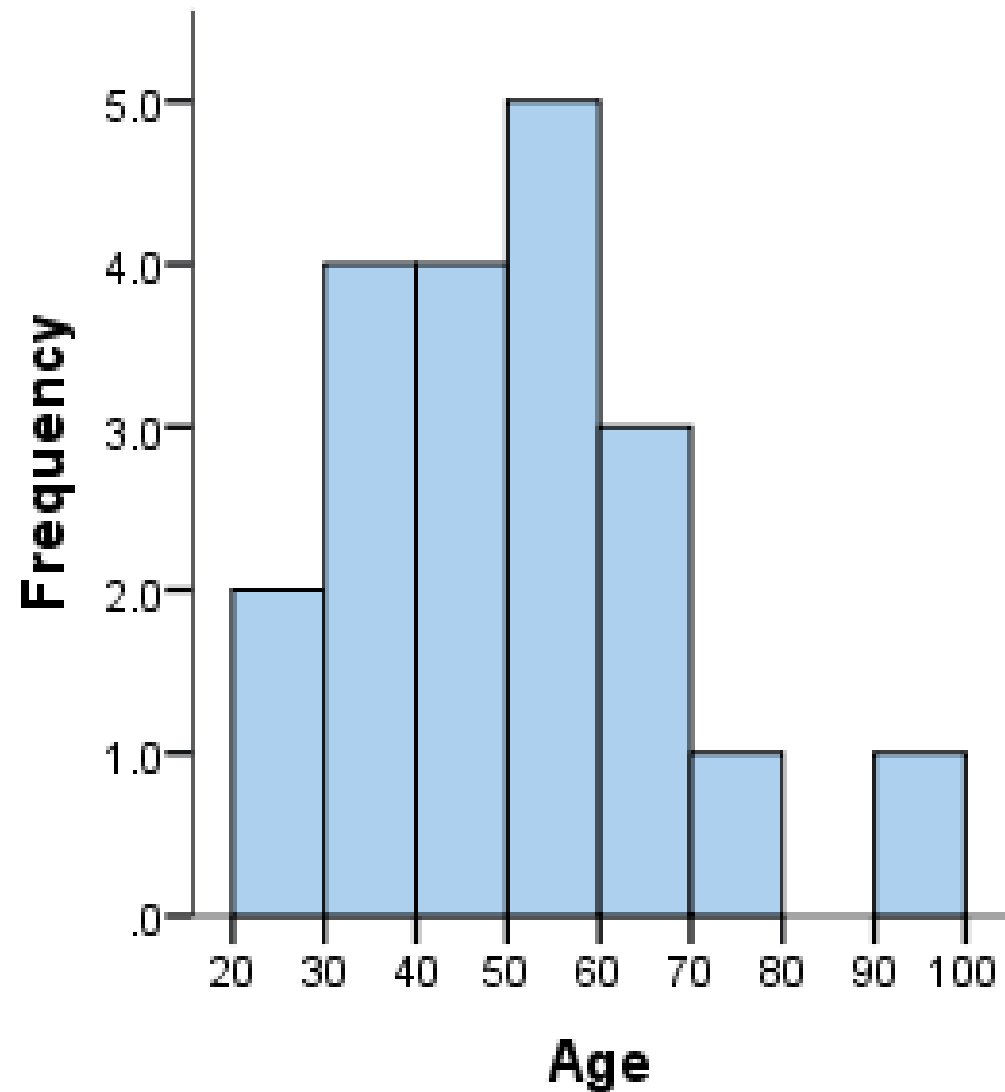
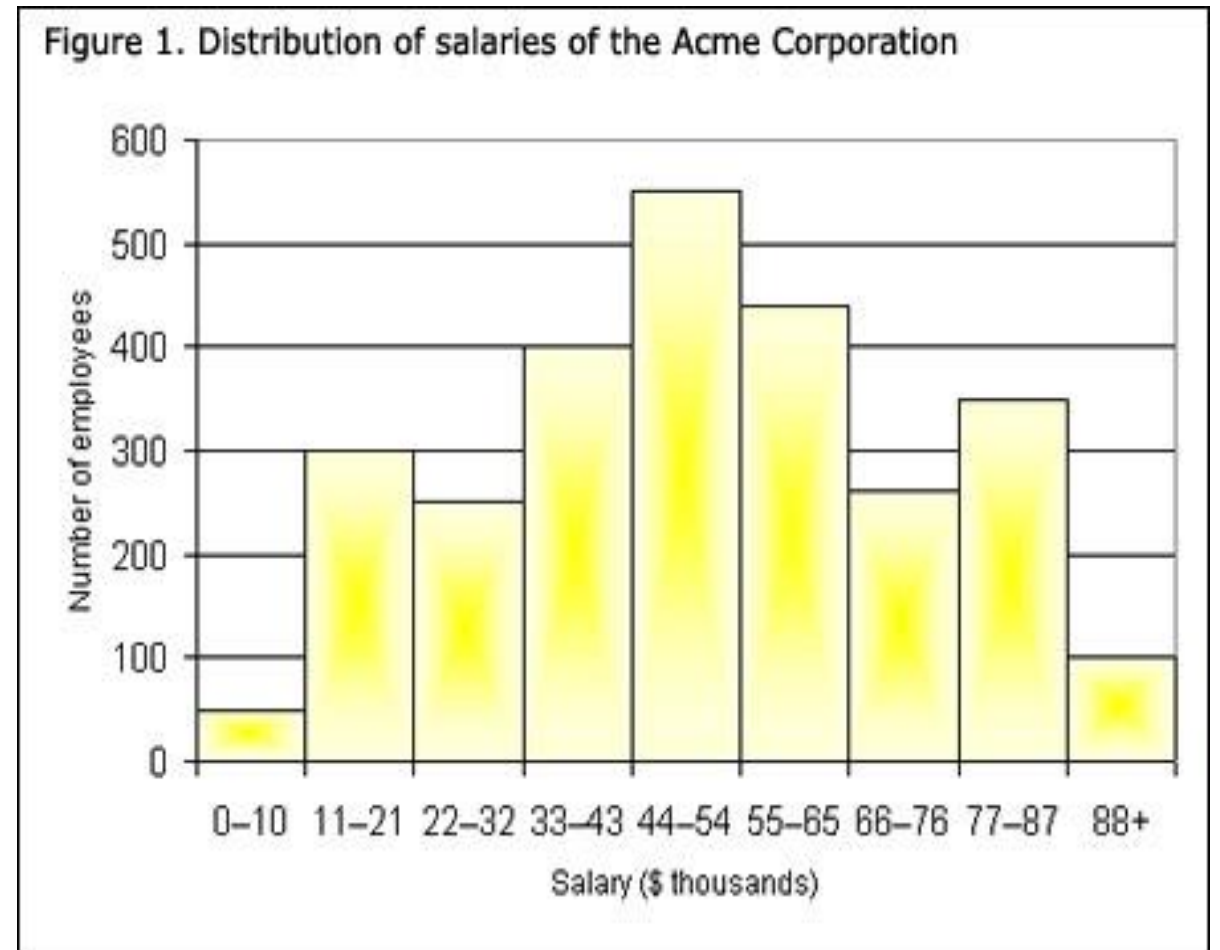


Fig: Histogram



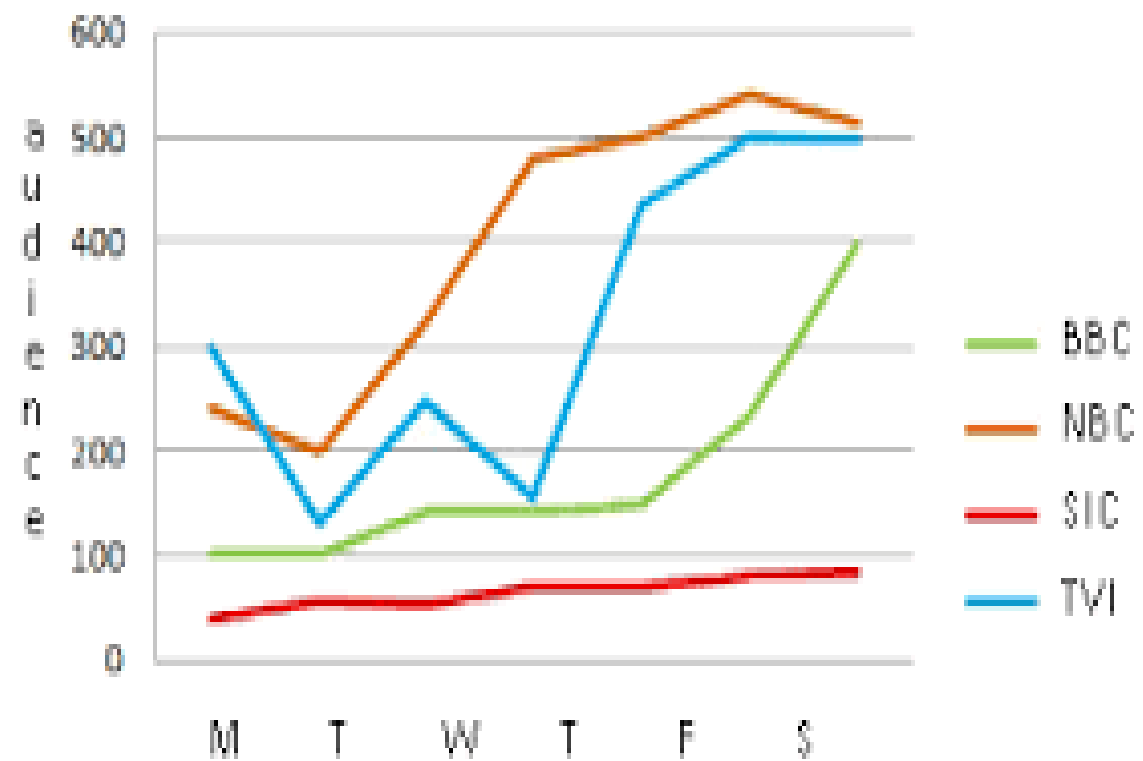


Fig: Scatter plots

- **Visual analysis could be inspecting some basic statistical measurements, such as averages, standard deviations, minimum, maximum, percentile, and confidence intervals..**

MISSING VALUES

- **Missing values can occur because of various reasons. The information can be non-applicable.**
- **For example, a customer decided not to disclose his or her income because of privacy**
- **Missing data can also originate because of an error during merging**

- **Schemes to deal with missing values :**

(i) Replace,

(ii) Delete,

(iii) Keep

Replace :

- **One could assign the missing credit unit scores with the average or median of the known values**
- **To fill the missing values based on the other information available**

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	18000	?	620	‘A’
2	28	12000	Single	?	‘B’
3	22	10000	Single	?	‘B’
4	60	22000	Widowed	700	‘A’
5	48	20000	Married	?	‘B’
6	44	?	?	?	‘B’
7	22	12000	Single	?	‘B’
8	26	15000	Married	350	‘B’
9	34	?	Single	?	‘A’
10	50	21000	Divorced	?	‘B’

Delete :

- **Straightforward option and consists of deleting observations or variables with lots of missing values.**
- **Information is missing at random and has no meaningful interpretation and/or relationship to the target (o/p)**

Keep :

- **Missing values can be meaningful (e.g., a customer did not disclose his or her income because he or she is currently unemployed).**
- **This is clearly related to the target (e.g., Class A or Class B) and needs to be considered as a separate category.**

- **Identify whether missing information is related to the target variable (o/p).**
- **If yes, then we can adopt the keep strategy and make a special category for it.**
- **If not, one can, depending on the no.of observation available, decide to either delete or assign.**

OUTLIER DETECTION AND TREATMENT

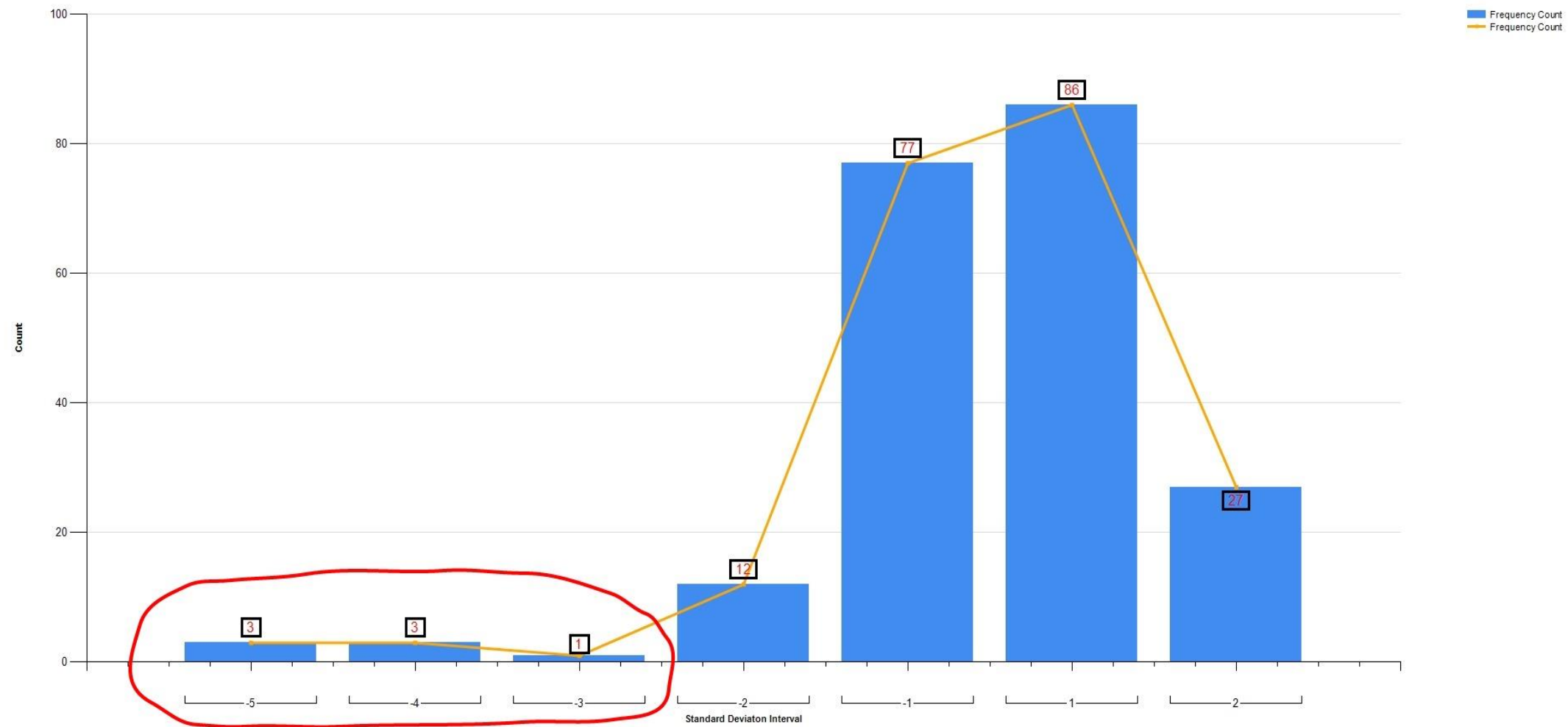
Outliers are extreme observations that are very dissimilar to the rest of the data. Two types

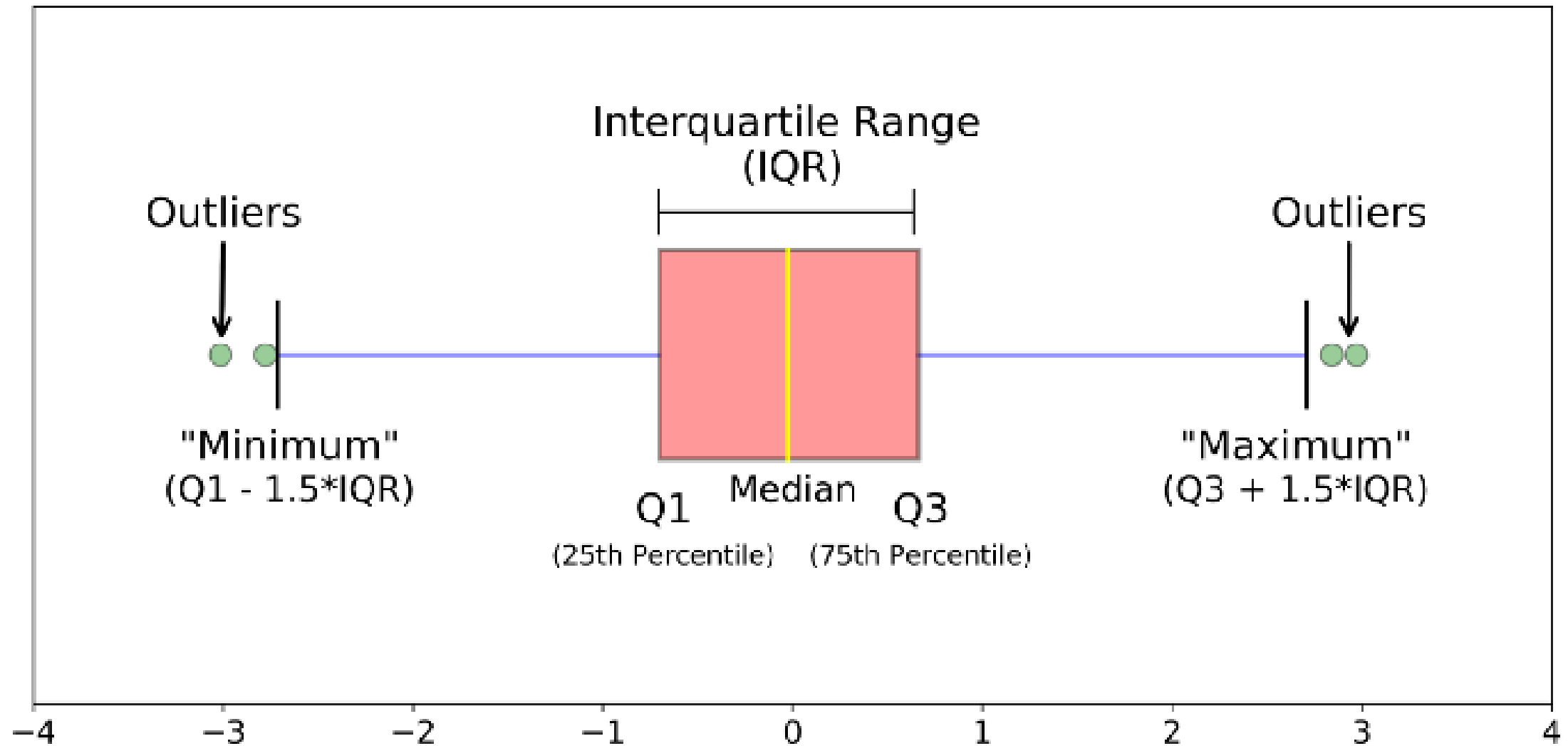
- 1. Valid observations (e.g., salary of boss is Rs. 4000)**
- 2. Invalid observations (e.g., age is 300 years)**

Outliers is to calculate the minimum and maximum values for each of the data elements.

Column Name	Hemopexin	HistogramType	3
Database Name	DataMiningProjects	Schema Name	Health
Table Name	DuchennesTable	DecimalPrecision	38,29

Histogram for the Hemopexin column in DataMiningProjects.Health.DuchennesTable





Expert-based limits based on business knowledge and/or experience can be imposed..

STANDARDIZING DATA

Standardizing data is a data preprocessing activity targeted at scaling variables to a similar range.

Procedure

Min/max standardization

Z-score standardization

Decimal scaling

- min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- z-score normalization

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$

Min/max standardization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

whereby newmax and newmin are the newly imposed maximum and minimum (e.g., 1 and 0).

$$X' = \frac{X - \min_{WPH}}{\max_{WPH} - \min_{WPH}} (\text{new}_{\max} - \text{new}_{\min}) + \text{new}_{\min}$$

To normalize the value 3000 to a new range [0.0, 1.0] the following should be calculated,

$$X' = \frac{3000 - 2400}{3857 - 2400} (1.0 - 0) + 0 = \frac{600}{1457} 1.0 = 0.411$$

So, by min-max normalization, the value 3000 in the WHP metric will be transformed to 0.411.

Z- Score standardization

- **Measuring how many standard deviations an observation lies away from the mean**

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

$$\text{Mean} = (x_1 + x_2 + x_3 + \dots + x_n) / n$$

Standard deviation:

$$\begin{aligned}s_X &= \sqrt{\frac{1}{n-1} \sum_{i=1}^{i=n} (x_i - \bar{X})^2} \\ &= \sqrt{\frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n-1}}\end{aligned}$$

s_X is the sample standard deviation

\bar{X} is the sample mean

n is the number of members of a sample

$x_i, i = 1, \dots, n$ are the members of a sample

Use the **two methods** below to *normalize* the following group of data:

200, 300, 400, 600, 1000

(a) min-max normalization by setting $min = 0$ and $max = 1$

(b) z-score normalization

Answer:

(a) min-max normalization by setting $min = 0$ and $max = 1$

<i>original data</i>	200	300	400	600	1000
$[0,1]$ normalized	0	0.125	0.25	0.5	1

(b) z-score normalization

<i>original data</i>	200	300	400	600	1000
<i>z-score</i>	-1.06	-0.7	-0.35	0.35	1.78

■

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

www.T4Tutorials.com

For Marks as 8:

$$\text{MinMax} = \frac{(V - \text{Min marks})}{\text{Max o marks} - \text{Min marks}} (\text{newMax} - \text{newMin}) + \text{newMin}$$

www.T4Tutorials.com

$$\text{MinMax} = \frac{(8 - 8)}{20 - 8} * (1 - 0) + 0$$

$$\text{MinMax} = \frac{(0)}{12} * 1$$

$$\text{MinMax} = 0$$

www.T4Tutorials.com

For Marks as 10:

$$\text{MinMax} = \frac{(10 - 8)}{20 - 8} * (1 - 0) + 0$$

www.T4Tutorials.com

$$\text{MinMax} = \frac{(2)}{12} * 1$$

$$\text{MinMax} = 0.16$$

Decimal scaling

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A.

$$v' = \frac{v}{10^j}$$

Here j is the smallest integer such that $\max|v'| < 1$.

Example :

A – values range from -986 to 917. Max $|v| = 986$.

$v = -986$ normalize to $v' = -986/1000 = -0.986$

CATEGORIZATION (Ref. Book Page : 24, pdf file page :44)

- **Categorization also known as coarse classification, classing, grouping and binning**
- **Chi-squared analysis is a more sophisticated way to do coarse classification.**
- **Procedure**
 - **Base table, Observation table, independent table**
 - **Compare observation table and independent frequency table**
 - **Conclude by chi-squared value.**

WEIGHTS OF EVIDENCE CODING

The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable. It is generally described as a measure of the separation of good and bad customers. "Bad Customers" refers to the customers who defaulted on a loan. and "Good Customers" refers to the customers who paid back loan..

The WOE is calculated as:

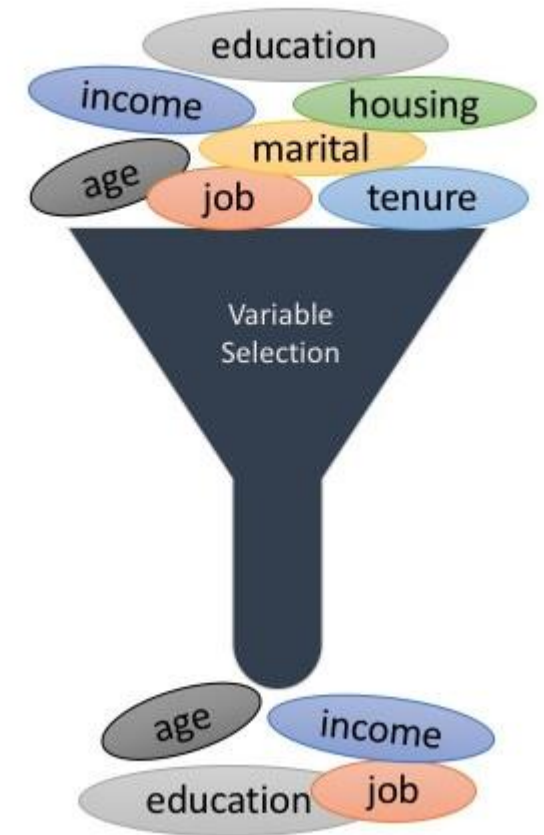
$$\ln(\text{Distr. Good} / \text{Distr. Bad}).$$

Because of the logarithmic transformation, a positive (negative) , WOE means $\text{Distr. Good} > (<) \text{Distr. Bad}$.

Age	Count	Distr.Count	Good	Distr. Good	Bad	Distri.Bad	WOE
Missing	50	2.50%	42	2.33%	8	04.12%	-57.28%
18-22	200	10.00%	152	8.42%	48	24.74%	-107.83%
23-26	300	15.00%	246	13.62%	54	27.84%	-71.47%
27-29	450	22.50%	405	22.43%	45	23.20%	-3.38%
30-35	500	25.00%	475	26.30%	25	12.89%	71.34%
35-44	350	17.50%	339	18.77%	11	05.67%	119.71%
44+	150	7.50%	147	8.14%	3	01.55%	166.08%
	2000		1806		194		

VARIABLE SELECTION

- Many analytical modeling exercises start with tons of variables, of which typically only a few actually contribute to the prediction of the target variable.
- Pearson correlation
- Fisher score
- Information value
- chi-squared analysis



Age	Count	Distr.Count	Good	Distr. Good	Bad	Distri.Bad	WOE	Information (IV)
Missing	50	2.50%	42	2.33%	8	04.12%	-57.28%	0.0103
18-22	200	10.00%	152	8.42%	48	24.74%	-107.83%	0.1760
23-26	300	15.00%	246	13.62%	54	27.84%	-71.47%	0.1016
27-29	450	22.50%	405	22.43%	45	23.20%	-3.38%	0.0003
30-35	500	25.00%	475	26.30%	25	12.89%	71.34%	0.0957
35-44	350	17.50%	339	18.77%	11	05.67%	119.71%	0.1568
44+	150	7.50%	147	8.14%	3	01.55%	166.08%	0.1095
	2000		1806		194			0.6502

Information Value (IV)	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	weak predictor
0.1 to 0.3	medium predictor
0.3 to 0.5	strong predictor
> 0.5	suspicious or too good to be true

SEGMENTATION

- **The data is segmented before the analytical modeling starts.**
- **The segmentation can be conducted using the experience and knowledge from a business expert, or it could be based on statistical analysis using, for example, decision trees**

- **Segmentation is a very useful preprocessing activity**
- **Needs to be careful with it because by segmenting, the number of analytical models to estimate will increase, which will obviously also increase the production, monitoring, and maintenance costs.**

Thank You...



ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCES
(Autonomous)

G.B. Nagar, Kalavai - 632506



Big data analytics

UNIT - II

Predictive Analytics

- The aim is to build an analytical model predicting a target measure of interest
- Types : (i) Regression and (ii) Classification

Regression

- In regression, the target **variable is continuous**. Example: predicting stock prices, loss given default (LGD), and customer lifetime value (CLV).

Regression

- In regression, the target **variable is continuous**. Example: predicting stock prices, loss given default (LGD), and customer lifetime value (CLV).

Classification

- The target is categorical. It can be binary (e.g., credit risk)

TARGET DEFINITION

- **The target variable plays an important role in the learning process**
- **In a customer attrition setting**
ex : easily detected when the customer cancels the contract
- **In credit scoring**
ex : payment 90 days, mortgages 180 days

In fraud detection:

Usually hard to determine because one can never be fully sure that a certain transaction.

The decision is then made based on a legal judgment or a high suspicion by a business expert

In Response modeling

- ***Gross response*** refers to the customers who purchase after having received the marketing message.
- However, it is more interesting to define the target as the ***net response***

Customer lifetime value (CLV) is a continuous target variable

$$CLV = \sum_{i=1}^n \frac{(R_t - C_t)S_t}{(1 + d)^t}$$

where

- ✓ **'n' represents the time horizon considered
(typically 2 to 3) years**
- ✓ **'R_t' the revenue at time 't' (both direct and indirect),**
- ✓ **'C_t' the costs incurred at time 't' (both direct and indirect),**
- ✓ **'S_t' the survival probability at time t,**
- ✓ **d the discounting factor**

Table 3.1 Example CLV Calculation

Month t	Revenue in Month t (R_t)	Cost in Month t (C_t)	Survival Probability in Month t (s_t)	$(R_t - C_t) * s_t / (1 + d)^t$
1	150	5	0.94	135.22
2	100	10	0.92	82.80
3	120	5	0.88	101.20
4	100	0	0.84	84.00
5	130	10	0.82	98.40
6	140	5	0.74	99.90
7	80	15	0.7	45.50
8	100	10	0.68	61.20
9	120	10	0.66	72.60
10	90	20	0.6	42.00
11	100	0	0.55	55.00
12	130	10	0.5	60.00
			CLV	937.82
	Yearly WACC	10%		
	Monthly WACC	1%		

Ac
Go

weighted average cost of capital [WACC]).

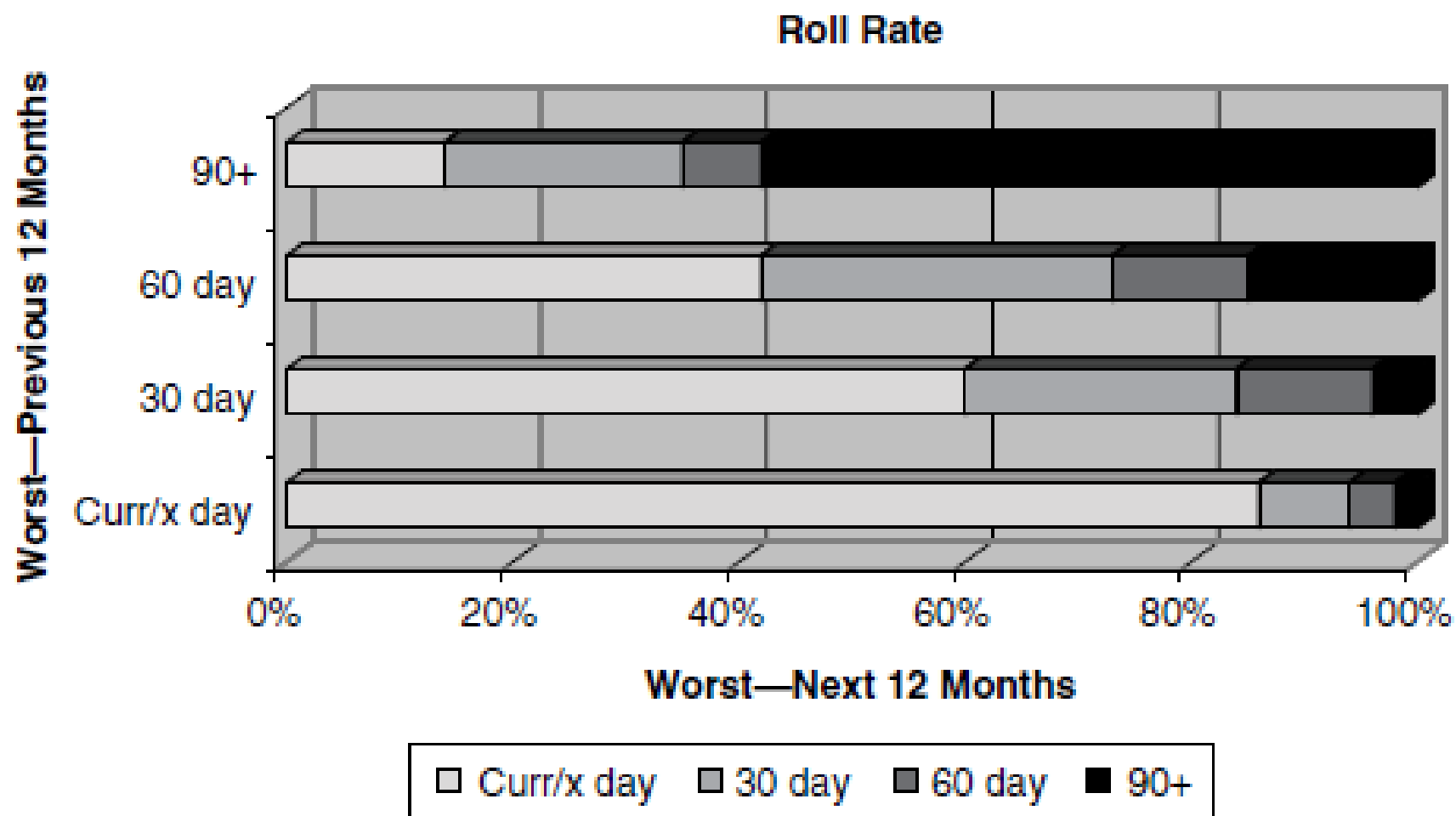


Figure 3.1 Roll Rate Analysis

Source: N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*

Linear Regression

- **Linear regression is a statistical method that analyzes and finds relationships between two variables.**
- **In predictive analytics it can be used to predict a future numerical value of a variable.**

- **Linear regression is a baseline modeling technique to model a continuous target variable.**
- **Example, in a CLV modeling context, a linear regression model can be defined to model CLV in terms of the RFM (recency, frequency, monetary value) predictors as follows:**

$$CLV = \beta_0 + \beta_1 R + \beta_2 F + \beta_3 M$$

DEFINITION OF RFM -

RFM (RECENCY, FREQUENCY, MONETARY) ANALYSIS IS MOST SIMPLE AND PROVEN TECHNIQUE USED BY MARKETING PEOPLE FOR CUSTOMER SEGMENTATION.



HOW RECENTLY (RECENCY)



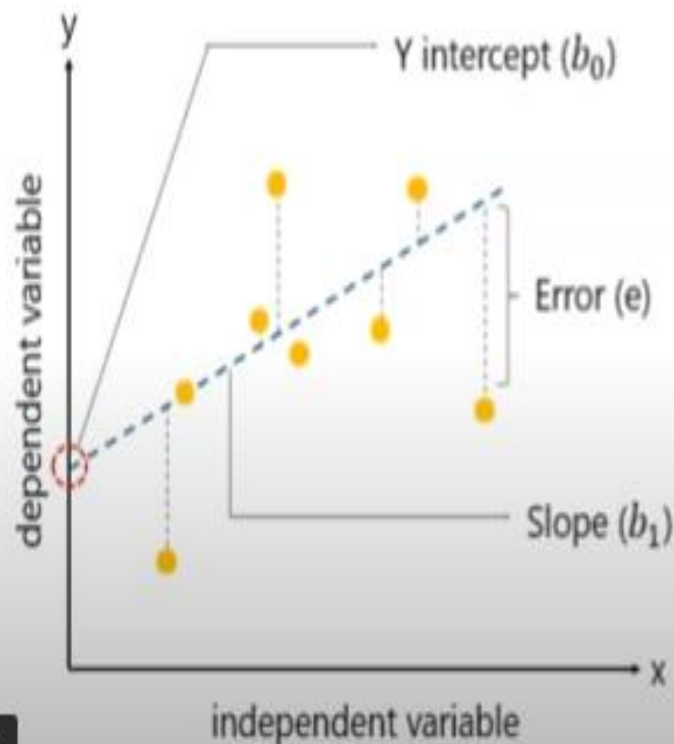
HOW OFTEN (FREQUENCY)



HOW MUCH DID THEY
BUY (MONETARY).

What Is Linear Regression?

Linear Regression is a method to predict dependent variable (Y) based on values of independent variables (X). It can be used for the cases where we want to predict some continuous quantity.



$$Y = b_0 + b_1x + e$$

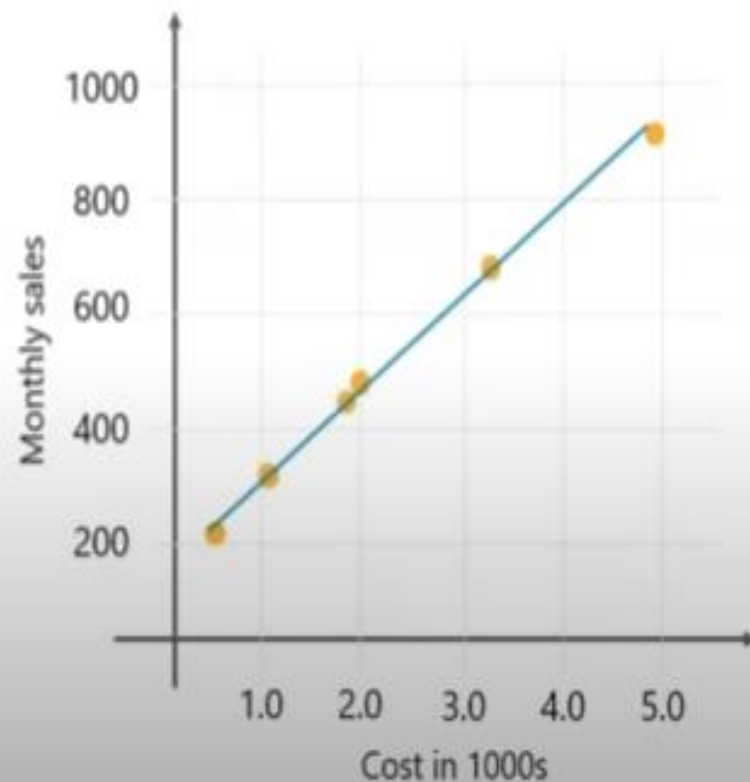
Labels for the equation $Y = b_0 + b_1x + e$:

- Y intercept
- dependent variable
- Slope
- independent variable
- Error

Linear Regression Use Case

To forecast monthly sales by studying the relationship between the monthly e-commerce sales and the online advertising costs.

Monthly sales	Advertising cost In 1000s
200	0.5
900	5
450	1.9
680	3.2
490	2.0
300	1.0



LOGISTIC REGRESSION

- **Consider a classification data set for response modeling**

Table 3.2 Example Classification Data Set

Customer	Age	Income	Gender	...	Response	Y
John	30	1,200	M		No	0
Sarah	25	800	F		Yes	1
Sophie	52	2,200	F		Yes	1
David	48	2,000	M		No	0
Peter	34	1,800	M		Yes	1

- **When modeling the response using linear regression, one gets:**

$$Y = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Gender}$$

Two key problems:

1. The errors/target are not normally distributed.
2. There is no guarantee that the target is between 0 and 1, which would be handy because it can then be interpreted as a probability.

Consider now the following bounding function:

$$f(z) = \frac{1}{1 + e^{-z}}$$

Consider now the following bounding function:

$$f(z) = \frac{1}{1 + e^{-z}}$$

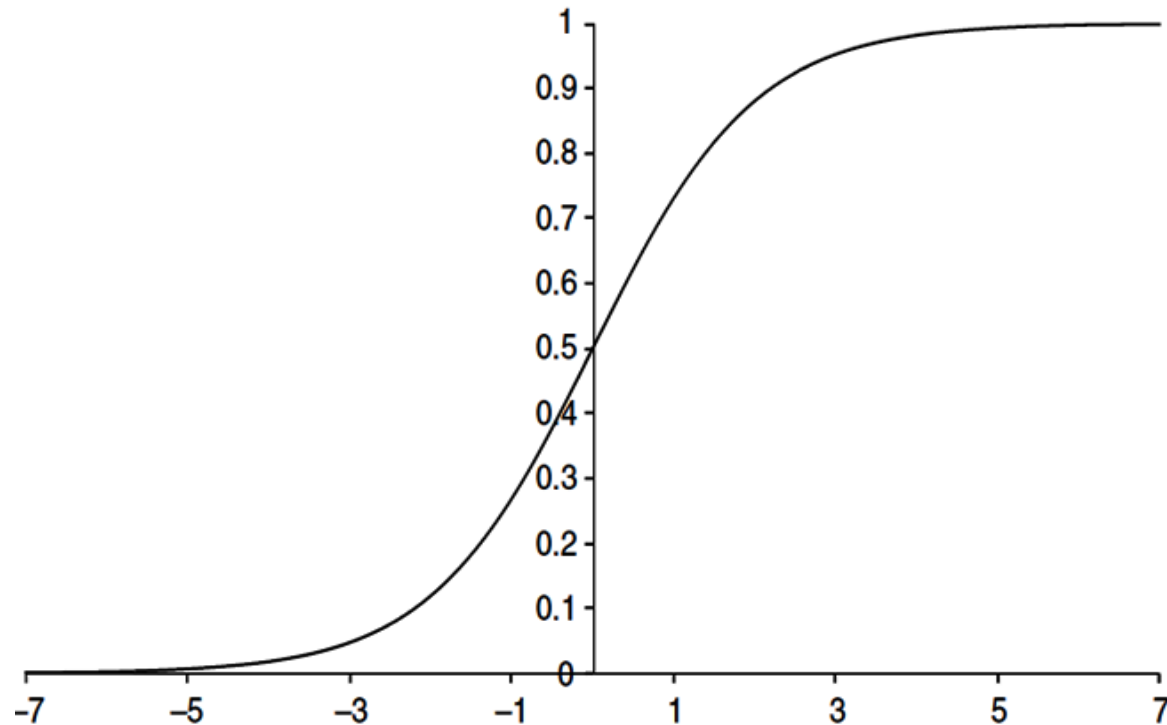
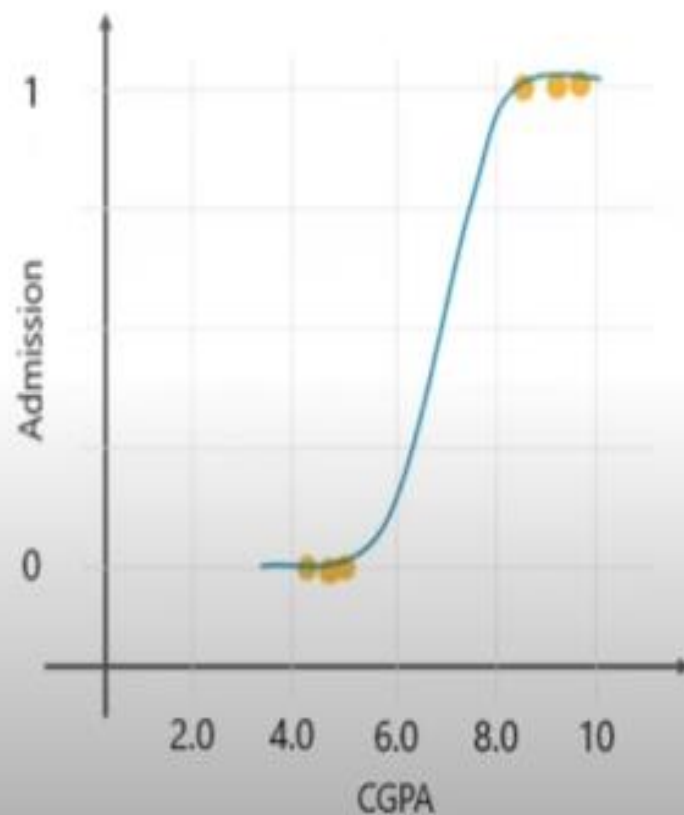


Figure 3.2 Bounding Function for Logistic Regression

Logistic Regression Use Case

To predict if a student will get admitted to a school based on his CGPA.

Admission	CGPA
0	4.2
0	5.1
0	5.5
1	8.2
1	9.0
1	9.1



- **For every possible value of z , the outcome is always between 0 and 1. Hence, by combining the linear regression with the bounding function,**

- we get the following logistic regression model:

$$P(\text{response} = \text{yes} | \text{age}, \text{income}, \text{gender}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{age} + \beta_2 \text{income} + \beta_3 \text{gender})}}$$

- The outcome of the above model is always bounded between 0 and 1, **no matter what values of age, income, and gender** are being used, and can as such be interpreted as a probability.

DECISION TREES

- **Decision trees are recursive partitioning algorithms (RPAs)**

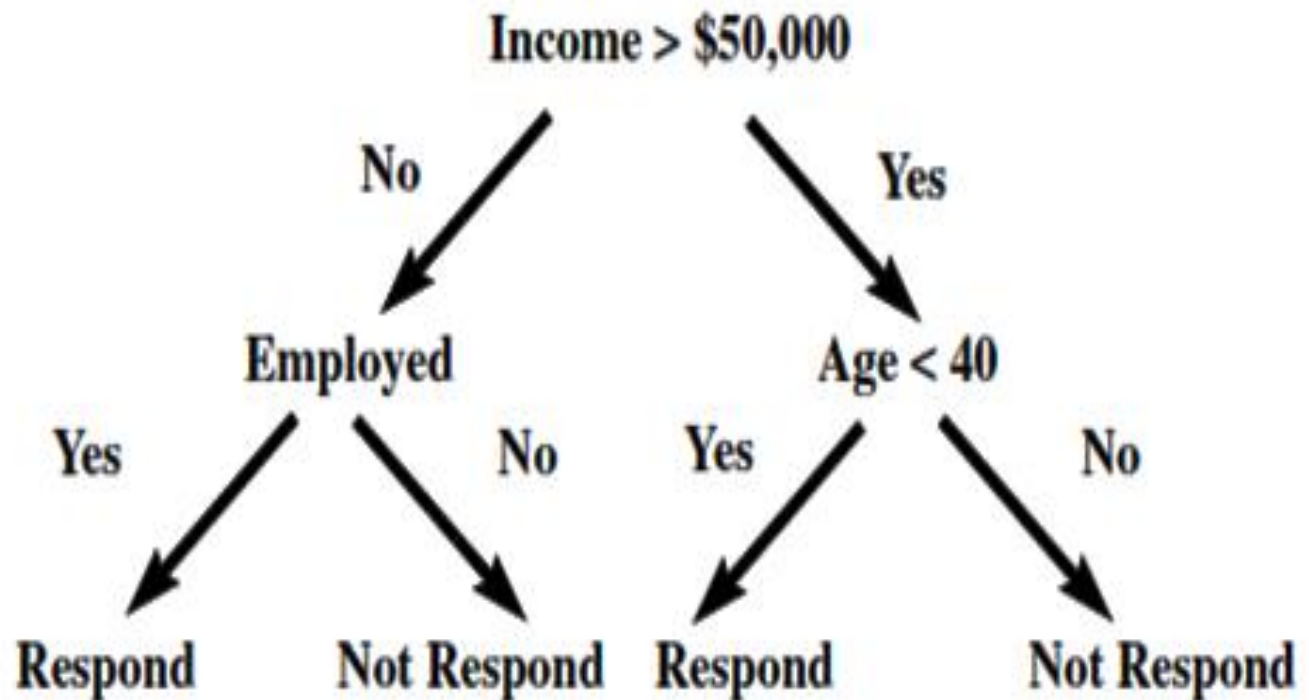
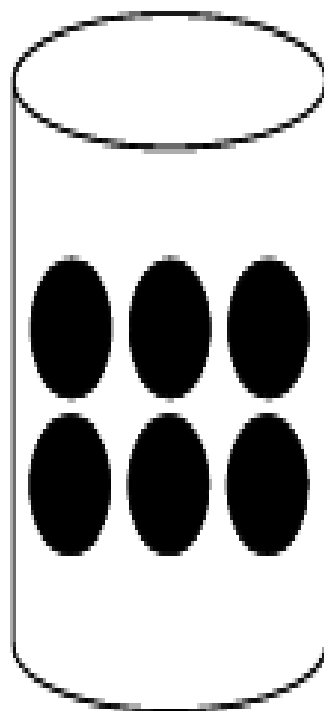


Figure 3.4 Example Decision Tree

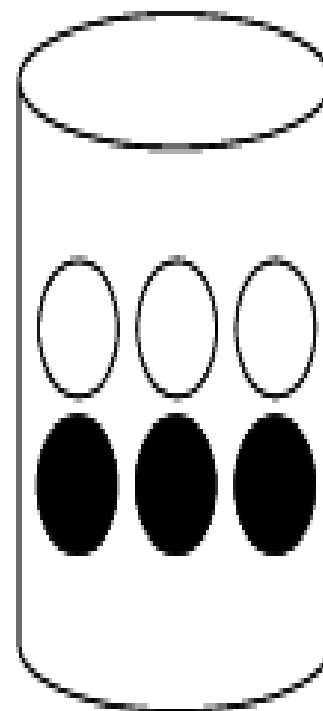
- The top node is the root node specifying a testing condition of which the outcome corresponds to a branch leading up to an internal node.
- The terminal nodes of the tree assign the classifications and are also referred to as the leaf nodes.

- **Splitting decision: Which variable to split at what value (e.g., age < 30 or not, income < 1,000 or not; marital status = married or not)**
- **Stopping decision: When to stop growing a tree?**
- **Assignment decision: What class (e.g., good or bad customer) to assign to a leaf node?**

Minimal Impurity



Maximal Impurity



Minimal Impurity

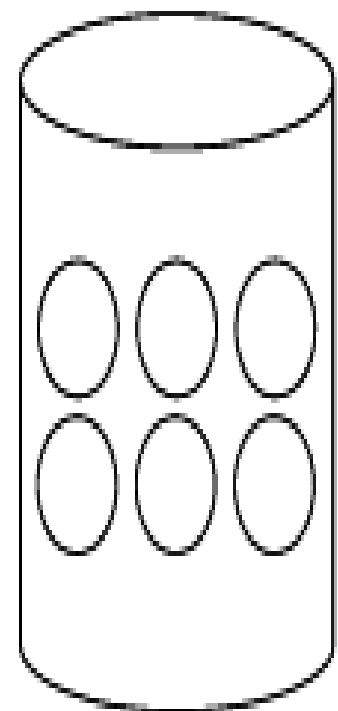


Figure 3.5 Example Data Sets for Calculating Impurity

- **Three data sets, each of which contains good (unfilled circles) and bad (filled circles) customers.**
- **Minimal impurity occurs when all customers are either good or bad.**
- **Maximal impurity occurs when one has the same number of good and bad customers (i.e., the data set in the middle).**
- **Decision trees will now aim at minimizing the impurity in the data.**
- **In order to do so appropriately, one needs a measure to quantify impurity.**

Entropy: $E(S) = -p_G \log_2(p_G) - p_B \log_2(p_B)$

Gini: $Gini(S) = 2p_G p_B$ (CART)

Chi-squared analysis (CHAID)

- **Classification and Regression Tree (CART)**
- **with p_G (p_B) being the shares of good and bad, respectively**
- **Chi-square Automatic Interaction Detector (CHAID)**

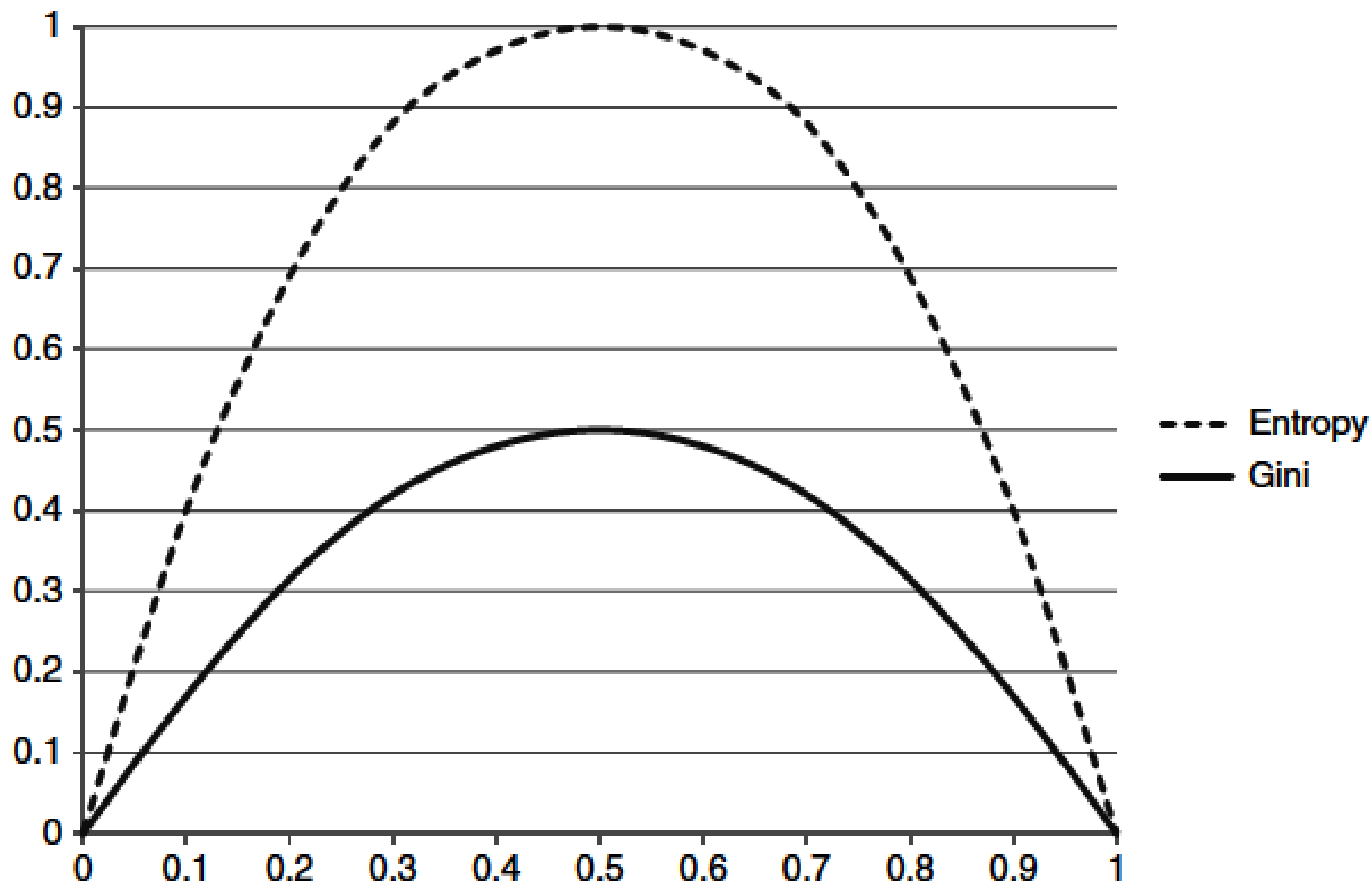


Figure 3.6 Entropy versus Gini

Entropy (Gini) is **minimal when all customers are either good or bad, and **maximal** in the case of the same number of good and bad customers.**

The original data set had maximum entropy. The entropy calculations become:

- Entropy top node = $-1/2 \times \log_2(1/2) - 1/2 \times \log_2(1/2) = 1$
- Entropy left node = $-1/3 \times \log_2(1/3) - 2/3 \times \log_2(2/3) = 0.91$
- Entropy right node = $-1 \times \log_2(1) - 0 \times \log_2(0) = 0$

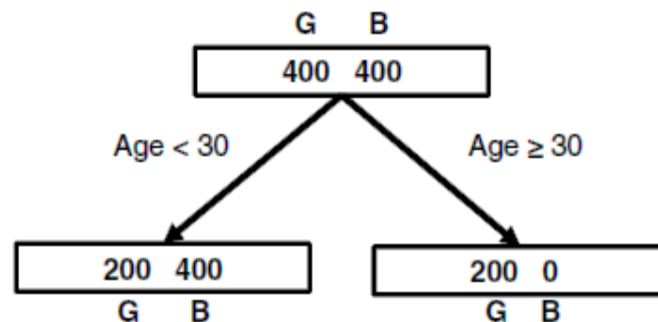


Figure 3.7 Calculating the Entropy for Age Split

The weighted decrease in entropy, also known as the gain, can then be calculated as follows:

Gain =

$$1 - (600/800) \times 0.91 - (200/800) \times 0 = 0.32$$

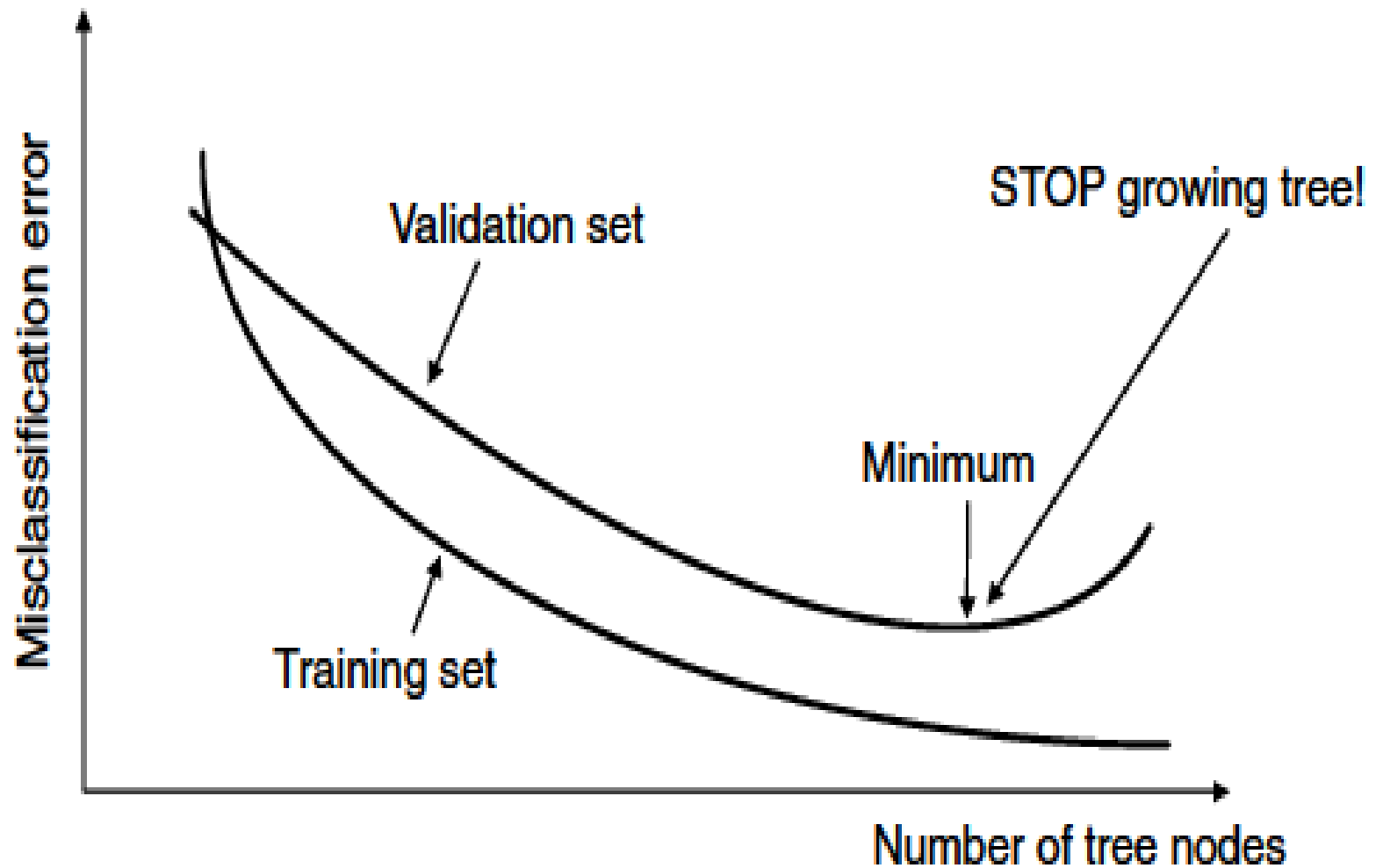


Figure 3.8 Using a Validation Set to Stop Growing a Decision Tree

way to measure impurity in a node is by calculating the mean squared error (MSE) as follows:

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

where n represents the number of observations in a leaf node, Y_i the value of observation i , and \bar{Y} , the average of all values in the leaf node.

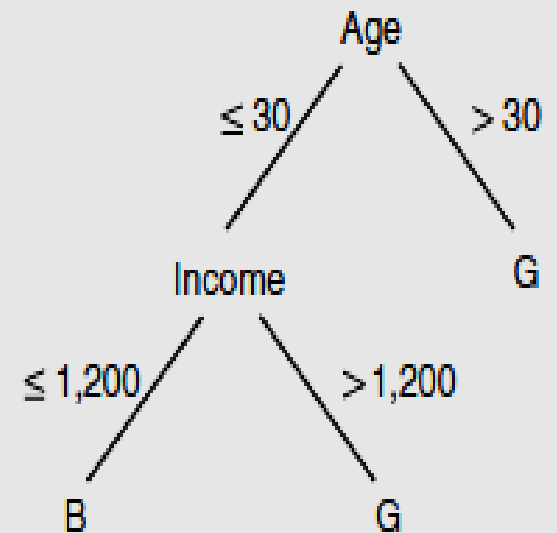
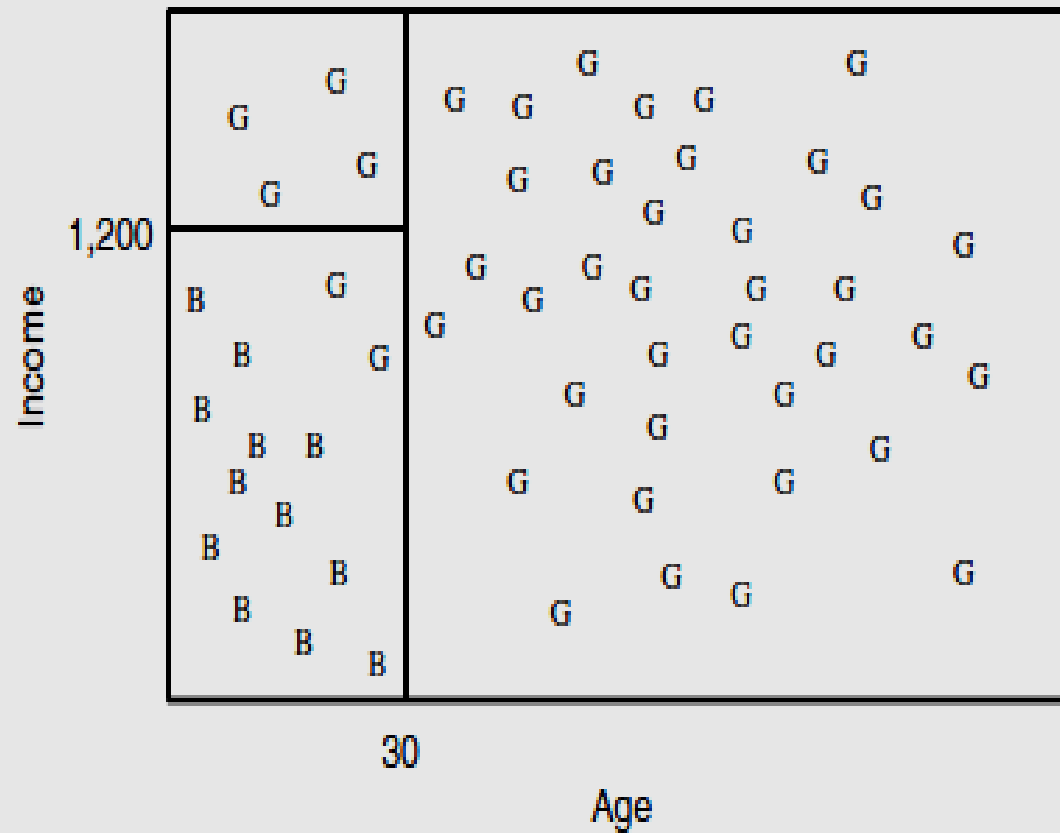


Figure 3.9 Decision Boundary of a Decision Tree

NEURAL NETWORKS

Functioning of the human brain.

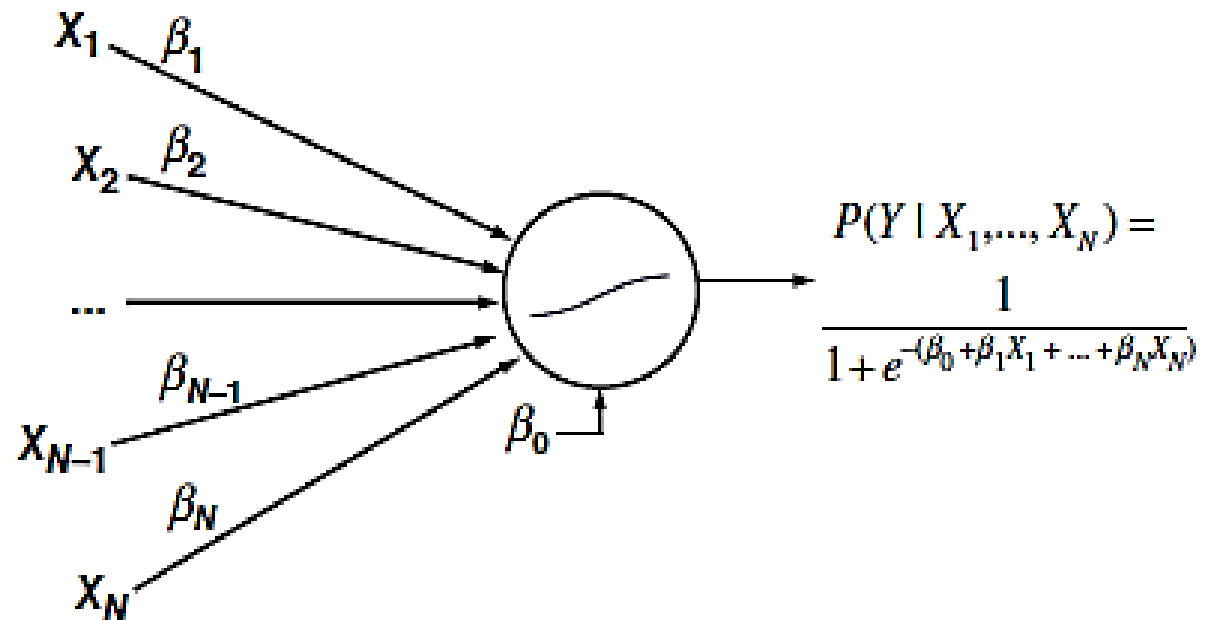


Figure 3.11 Neural Network Representation of Logistic Regression

Neuron in the middle basically performs two operations:

(i) Inputs, (ii) Multiplies them with the weights

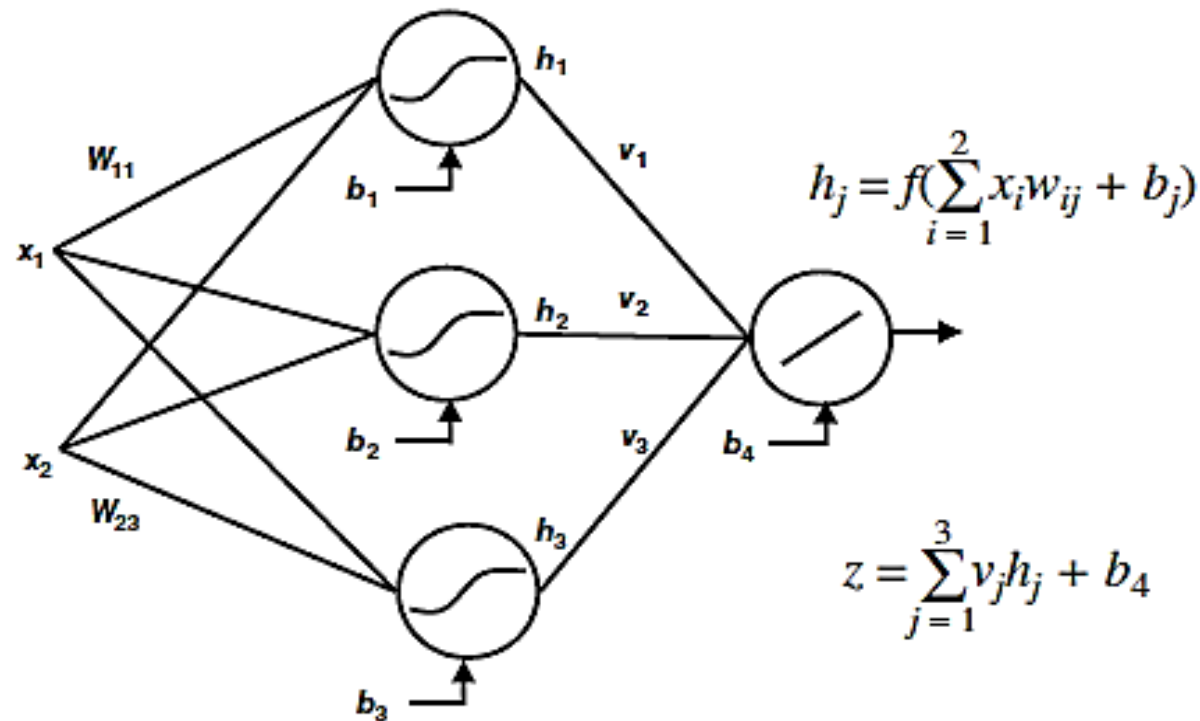


Figure 3.12 A Multilayer Perceptron (MLP) Neural Network

- **MLP with one input layer, one hidden layer, and one output layer.**
- **The hidden layer has a nonlinear transformation function $f(\cdot)$ and the output layer a linear transformation function.**
- **The most popular transformation functions (also called squashing, activation functions) are:**

- Logistic, $f(z) = \frac{1}{1 + e^{-z}}$, ranging between 0 and 1
- Hyperbolic tangent, $f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$, ranging between -1 and +1
- Linear, $f(z) = z$, ranging between $-\infty$ and $+\infty$

- **outputs can then be interpreted as probabilities**
- **The error function can thus have multiple local minima but typically only one global minimum.**

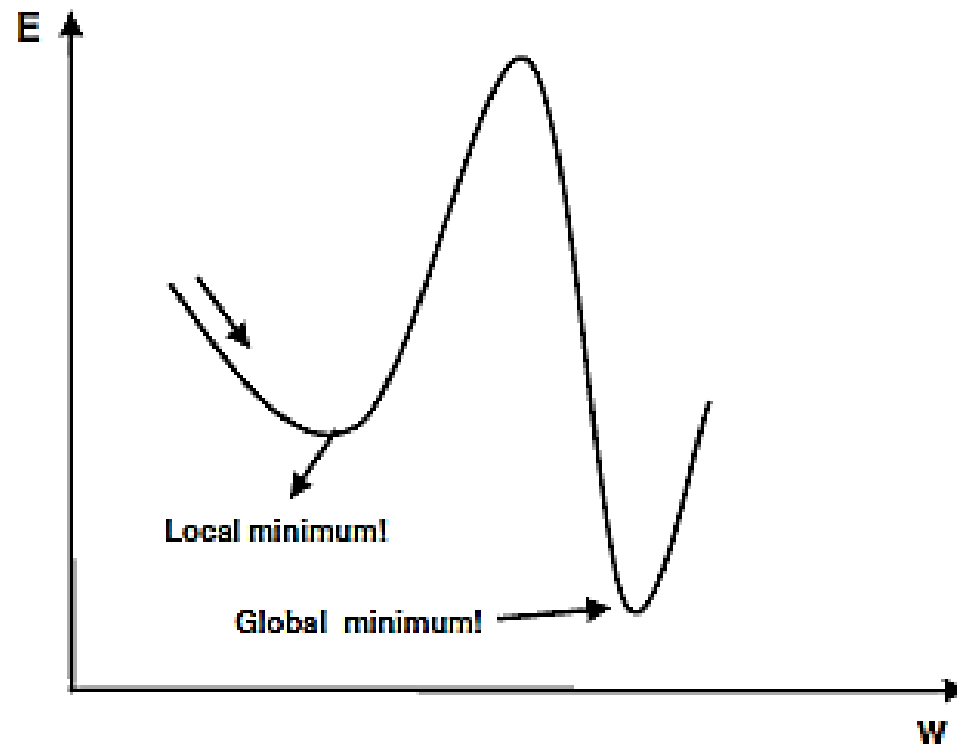


Figure 3.13 Local versus Global Minima

Procedure is as follows:

- **Split the data into a training, validation, and test set.**
- **Vary the number of hidden neurons from 1 to 10 in steps of 1 or more.**
- **Train a neural network on the training set and measure the performance on the validation set**
- **Choose the number of hidden neurons with optimal validation set performance.**
- **Measure the performance on the independent test set.**

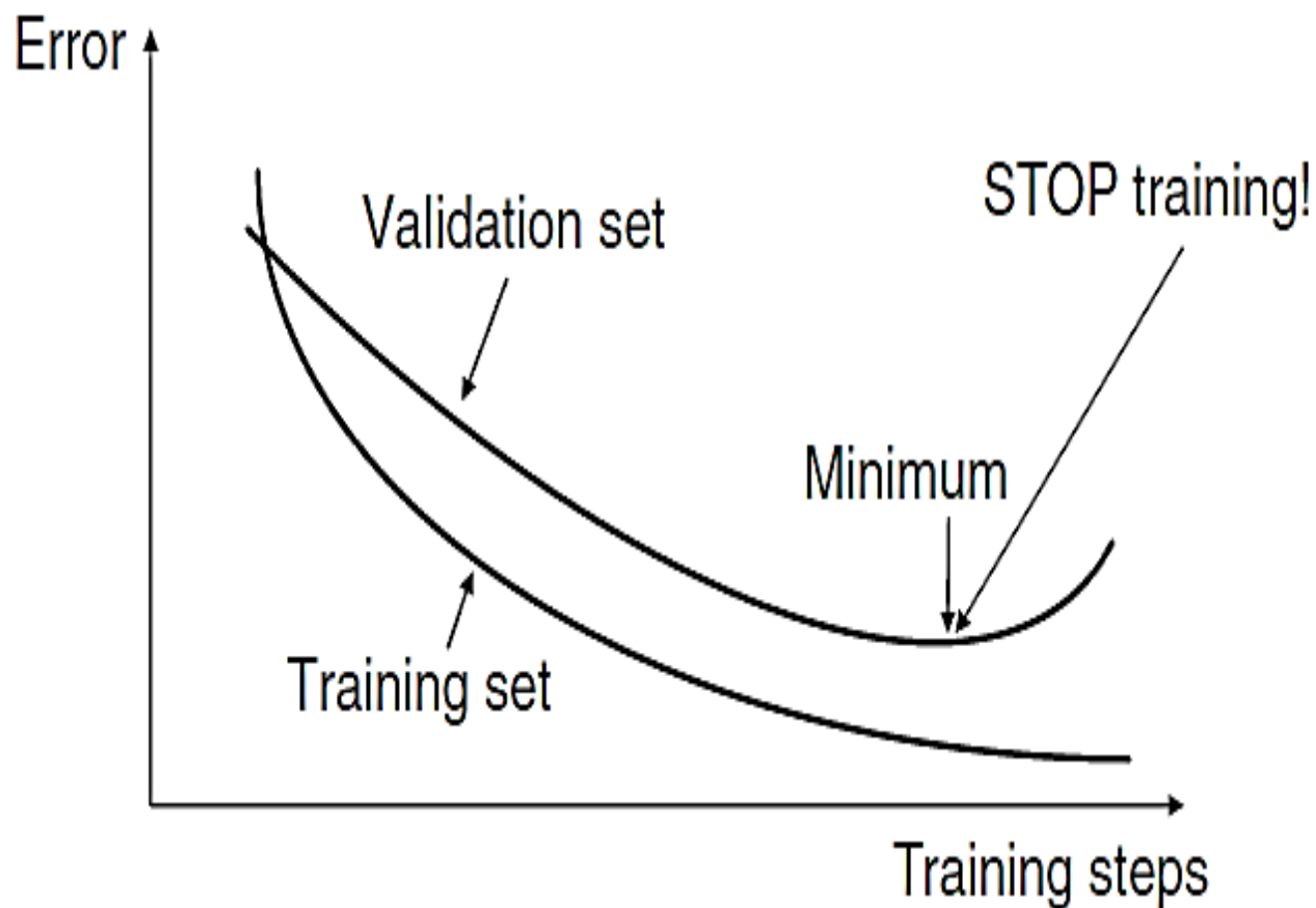


Figure 3.14 Using a Validation Set for Stopping Neural Network Training

A typical five-step approach here could be:

- 1. Train a neural network and prune it as much as possible in terms of connections.**
- 2. Categorize the hidden unit activation values using clustering.**
- 3. Extract rules that describe the network outputs in terms of the categorized hidden unit activation values.**
- 4. Extract rules that describe the categorized hidden unit activation values in terms of the network inputs.**
- 5. Merge the rules obtained in steps 3 and 4 to directly relate the inputs to the outputs.**

Customer	Age	Income	Gender	...	Response
Emma	28	1,000	F		No
Will	44	1,500	M		Yes
Dan	30	1,200	M		No
Bob	58	2,400	M		Yes

Customer	Age	Income	Gender	h1	h2	h3	h1	h2	h3	Response
Emma	28	1,000	F	-1.20	2.34	0.66	1	3	2	No
Will	44	1,500	M	0.78	1.22	0.82	2	3	2	Yes
Dan	30	1,200	M	2.1	-0.18	0.16	3	1	2	No
Bob	58	2,400	M	-0.1	0.8	-2.34	1	2	1	Yes

If $h1 = 1$ and $h2 = 3$, then response = No
 If $h2 = 2$, then response = Yes

If age < 28 and income < 1,000, then $h1 = 1$
 If gender = F, then $h2 = 3$
 If age > 34 and income > 1,500, then $h2 = 2$

If age < 28 and income < 1,000 and gender = F then response = No
 If age > 34 and income > 1,500 then response = Yes

Step 1: Start from original data.

Step 2: Build a neural network
 (e.g, 3 hidden neurons).

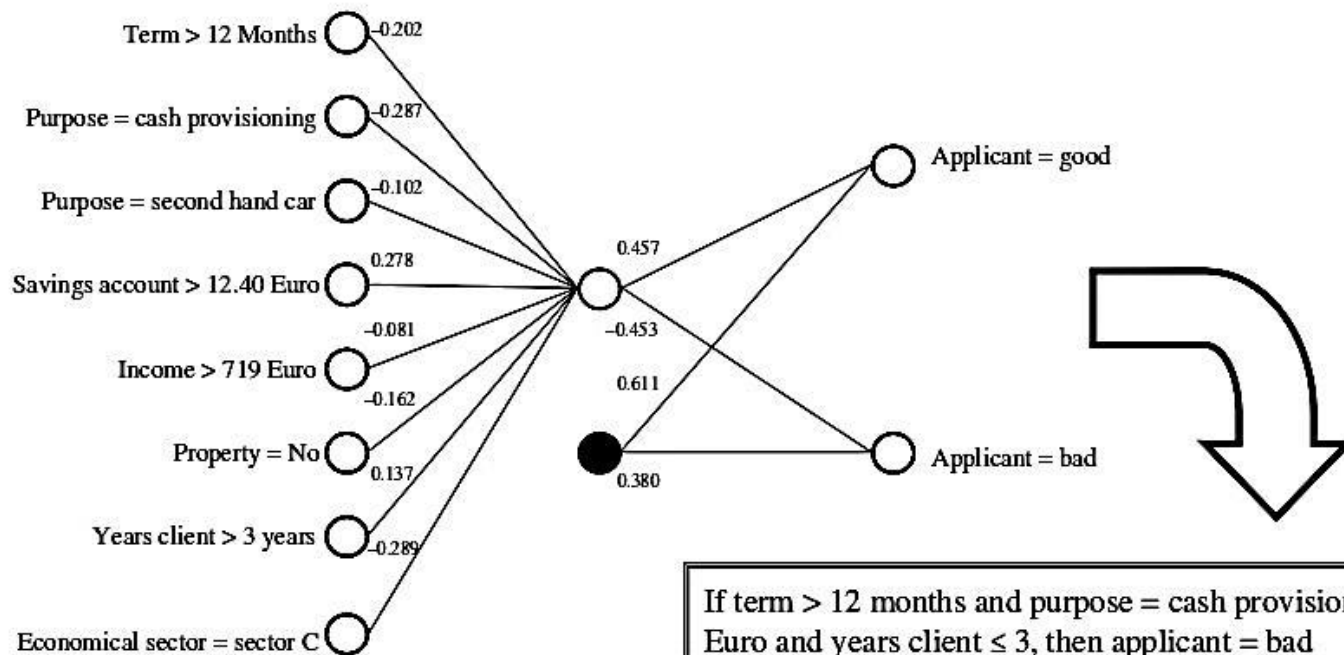
Step 3: Categorize hidden unit activations.

Step 4: Extract rules relating network outputs
 to categorized hidden units.

Step 5: Extract rules relating categorized
 hidden units to inputs.

Step 6: Merge both rule sets

Figure 3.15 Decompositional Approach for Neural Network Rule Extraction



If term > 12 months and purpose = cash provisioning and savings account \leq 12.40 Euro and years client \leq 3, then applicant = bad

If term > 12 months and purpose = cash provisioning and owns property = no and savings account \leq 12.40 Euro and years client \leq 3, then applicant = bad

If purpose = cash provisioning and income > 719 and owns property = no and savings account \leq 12.40 Euro and years client \leq 3, then applicant = bad

If purpose = secondhand car and income > 719 Euro and owns property = no and savings account \leq 12.40 Euro and years client \leq 3, then applicant = bad

If savings account \leq 12.40 Euro and economical sector = sector C, then applicant = bad

Default class: applicant = good

Figure 3.16 Example of Decompositional Neural Network Rule Extraction

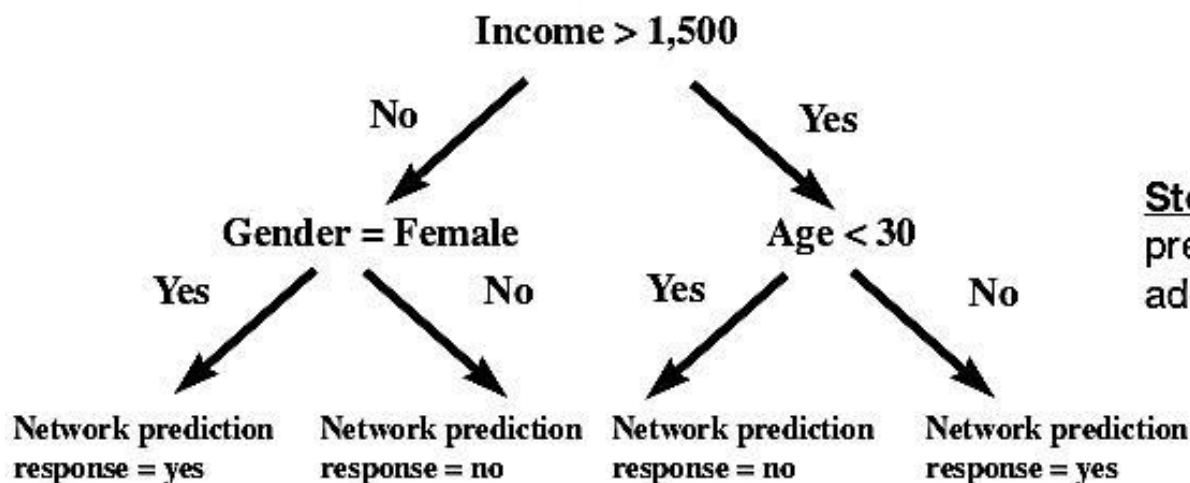
Customer	Age	Income	Gender	...	Response
Emma	28	1,000	F		No
Will	44	1,500	M		Yes
Dan	30	1,200	M		No
Bob	58	2,400	M		Yes

Step 1: Start from original data.

Customer	Age	Income	Gender	Network Prediction	Response
Emma	28	1,000	F	No	No
Will	44	1,500	M	Yes	Yes
Dan	30	1,200	M	Yes	No
Bob	58	2,400	M	Yes	Yes

Step 2: Build a neural network.

Step 3: Get the network predictions and add them to the data set.



Step 4: Extract rules relating network predictions to original inputs. Generate additional data where necessary.

Figure 3.17 Pedagogical Approach for Rule Extraction

		Neural Network Classification	
Rule set classification		Good	Bad
	Good	a	b
	Bad	c	d

Fidelity = $(a + d) / (b + c)$.

- **Target = linear regression (X_1 , X_2 , ... X_N) + neural network (X_1 , X_2 , ... X_N)**
- **Score = logistic regression (X_1 , X_2 , ... X_N) + neural network (X_1 , X_2 , ... X_N)**
- **This setup provides an ideal balance between model interpretability (which comes from the first part) and model performance (which comes from the second part).**

Customer	Age	Income	Gender	...	Response
Emma	28	1,000	F		No
Will	44	1,500	M		Yes
Dan	30	1,200	M		No
Bob	58	2,400	M		Yes

Step 1: Start from original data.

Customer	Age	Income	Gender	...	Response	Logistic Regression Output
Emma	28	1,000	F		No (=0)	0.44
Will	44	1,500	M		Yes (=1)	0.76
Dan	30	1,200	M		No (=0)	0.18
Bob	58	2,400	M		Yes (=1)	0.88

Step 2: Build logistic regression model.

Customer	Age	Income	Gender	...	Response	Logistic Regression Output	Error
Emma	28	1,000	F		No (=0)	0.44	-0.44
Will	44	1,500	M		Yes (=1)	0.76	0.24
Dan	30	1,200	M		No (=0)	0.18	-0.18
Bob	58	2,400	M		Yes (=1)	0.88	0.12

Step 3: Calculate errors from logistic regression model.

Step 4: Build NN predicting errors from logistic regression model.

Customer	Age	Income	Gender	...	Logistic Regression Output	NN Output	Final Output
Bart	28	1,000	F		0.68	-0.32	0.36

Step 5: Score new observations by adding up logistic regression and NN scores.

Figure 3.18 Two-Stage Models

Support Vector machines

The origins of classification SVMs date back to the early dates of linear programming. Consider the following linear program (LP) for classification:

$$\text{mine}_1 + e_2 + \dots + e_{ng} + \dots + e_{nb}$$

subject to

$$w_1x_{i1} + w_2x_{i2} + \dots + w_nx_{in} \geq c - e_i, 1 \leq i \leq n_g$$

$$w_1x_{i1} + w_2x_{i2} + \dots + w_nx_{in} \leq c + e_i, n_g + 1 \leq i \leq n_g + n_b$$

$$e_i \geq 0$$

- **The LP assigns the good customers a score above the cut-off value c ,**
- **The bad customers a score below c .**
- **n_g and n_b represent the number of goods and bads, respectively.**
- **The error variables is e_i**

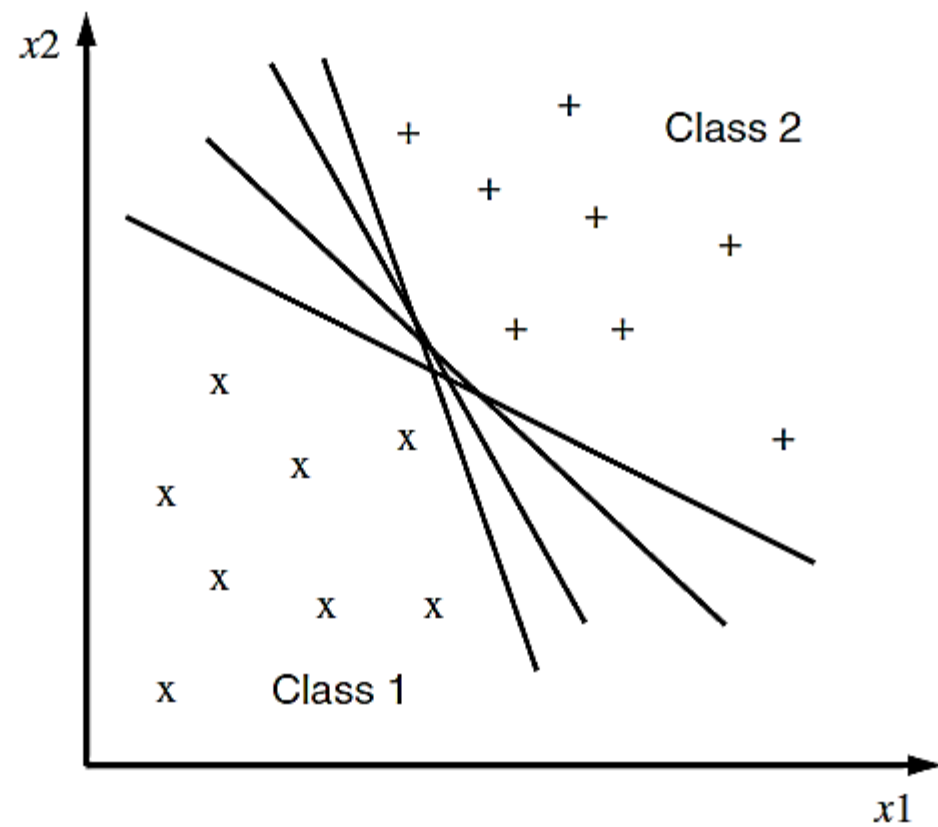


Figure 3.19 Multiple Separating Hyperplanes

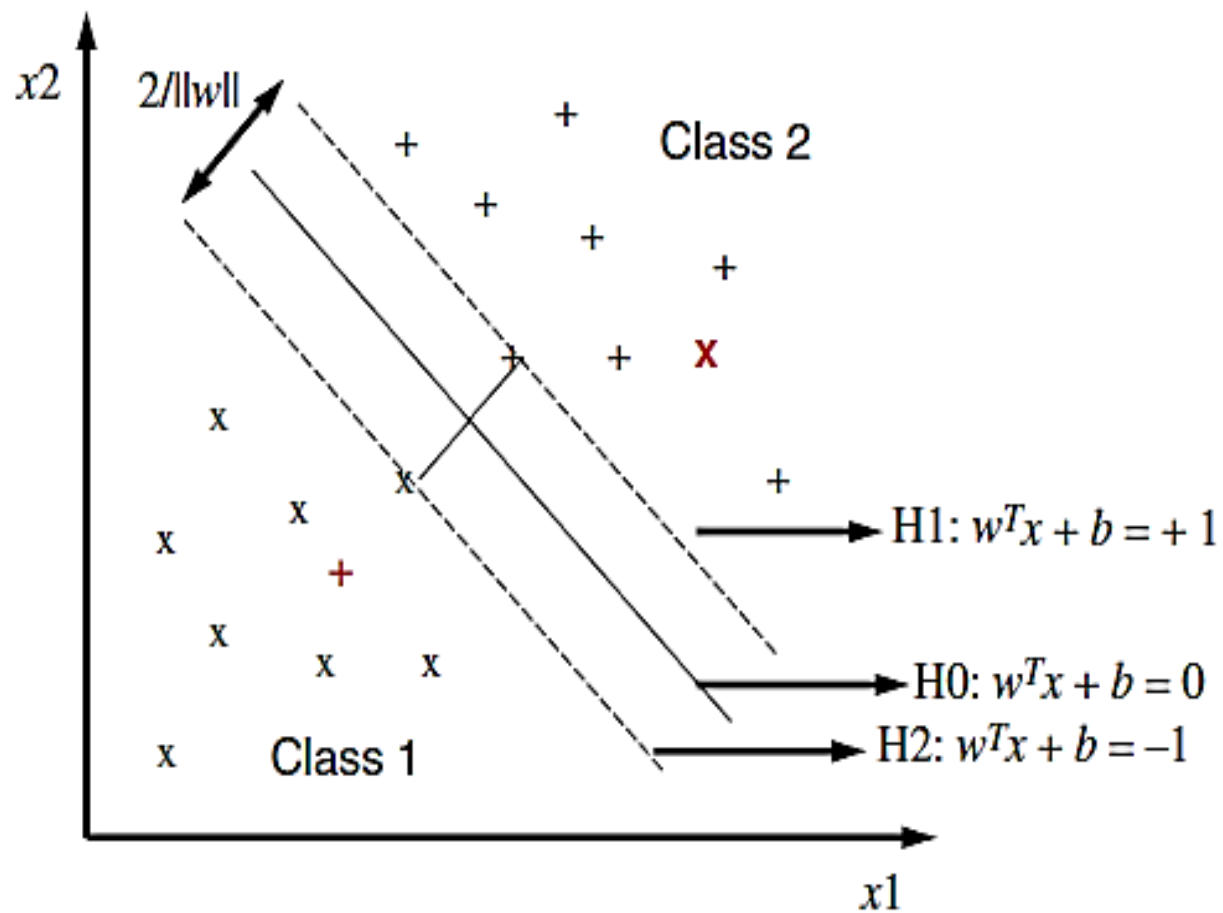


Figure 3.21 SVM Classifier in Case of Overlapping Distributions

- **Two hyperplanes sitting at the edges of both classes and a hyperplane in between, which will serve as the classification boundary.**

from the first hyperplane H_1 to the origin equals $|b-1|/\|w\|$, whereby $\|w\|$ represents the Euclidean norm of w calculated as $\|w\| = \sqrt{w_1^2 + w_2^2}$. Likewise, the perpendicular distance from H_2 to the origin equals $|b+1|/\|w\|$. Hence, the margin between both hyperplanes equals $2/\|w\|$.

Consider a training set: $\{x_k, y_k\}_{k=1}^n$ with $x_k \in R^N$ and $y_k \in \{-1; +1\}$

The goods (e.g., class +1) should be above hyperplane H1, and the
bads (e.g., class -1) below hyperplane H2, which gives:

$$w^T x_k + b \geq 1, \text{ if } y_k = +1$$

$$w^T x_k + b \leq -1, \text{ if } y_k = -1$$

Both can be combined as follows:

$$y_k(w^T x_k + b) \geq 1$$

The optimization problem then becomes:

$$\text{Minimize } \frac{1}{2} \sum_{i=1}^N w_i^2$$

subject to $y_k(w^T x_k + b) \geq 1, k = 1 \dots n$

The most popular are:

- **Linear kernel:**
- **Polynomial kernel**
- **Radial basis function (RBF) kernel**

ENSEMBLE METHODS

- **Ensemble methods aim at estimating multiple analytical models instead of using only one.**
- **The idea here is that multiple models can cover different parts of the data input space and, as such, complement each other's deficiencies.**
- **we will discuss bagging, boosting, and random forests.**

Bagging

- **Bagging (bootstrap aggregating) starts by taking B bootstraps from the underlying sample.**
- **The idea is then to build a classifier (e.g., decision tree) for every bootstrap**
- **To improve the stability and accuracy of machine learning**

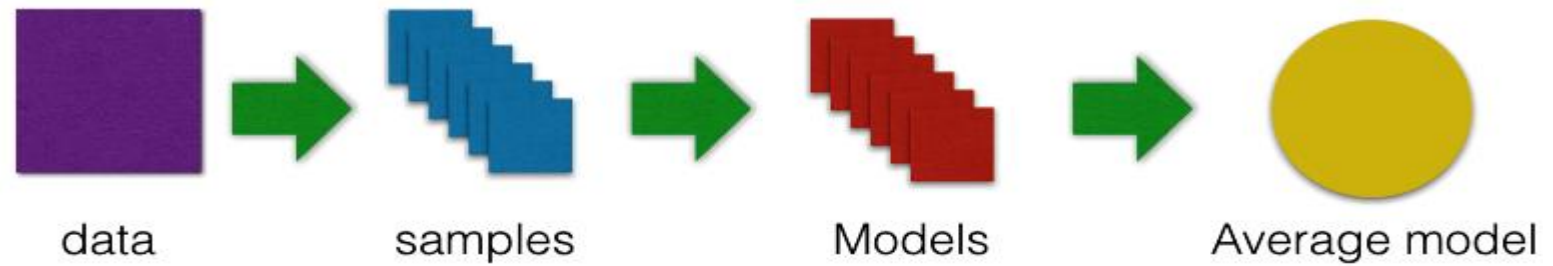
Boosting

- **Boosting works by estimating multiple models using a weighted sample of the data.**
- **Starting from uniform weights, boosting will iteratively reweight the data according to the classification error, whereby misclassified cases get higher weights.**
- **The idea here is that difficult observations should get more attention.**
- **to generate multiple weak learners and combine their predictions to form one strong rule**

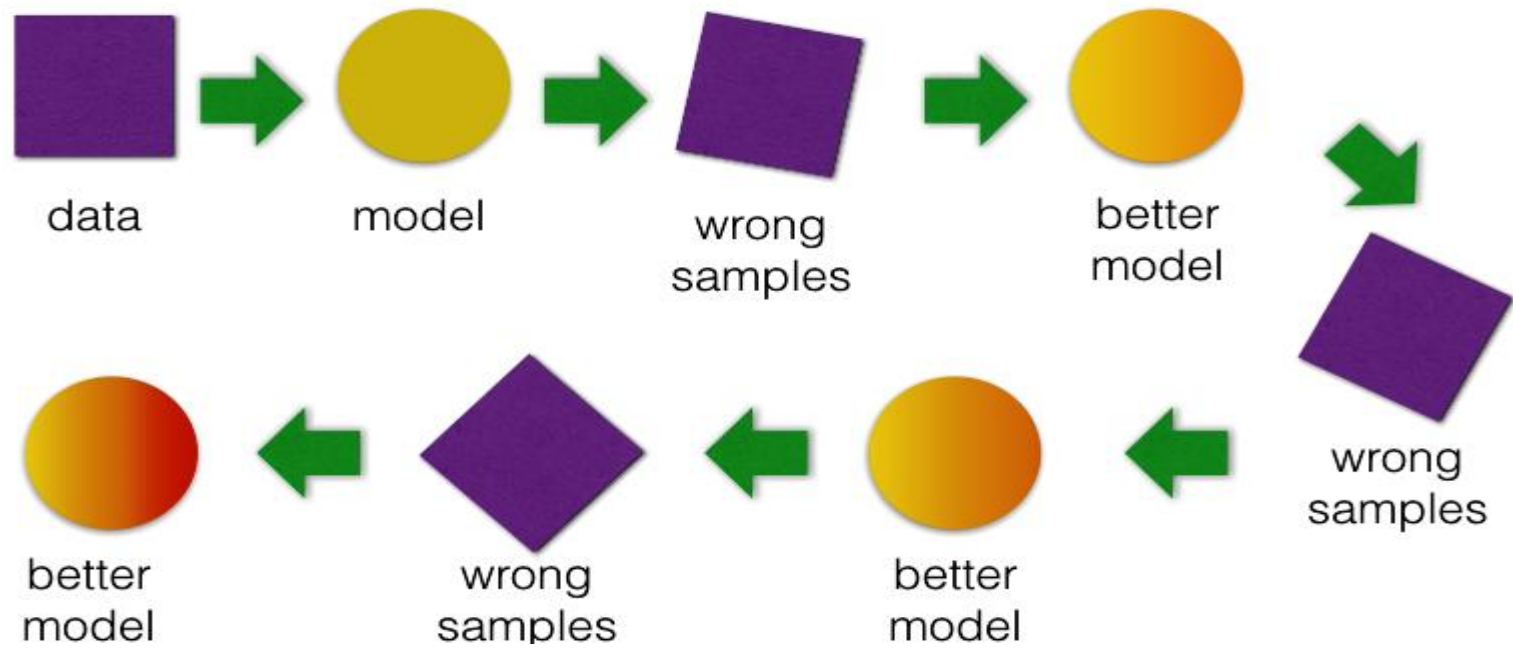
Procedure for boosting

1. Given the following observations: $(x_1, y_1), \dots, (x_n, y_n)$ where x_i is the attribute vector of observation i and $y_i \in \{1, -1\}$
2. Initialize the weights as follows: $W_1(i) = 1/n, i = 1, \dots, n$
3. For $t = 1 \dots T$
 - a. Train a weak classifier (e.g., decision tree) using the weights W_t
 - b. Get weak classifier C_t with classification error ϵ_t
 - c. Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$
 - d. Update the weights as follows:
 - i. $W_{t+1}(i) = \frac{W_t(i)}{Z_t} e^{-\alpha_t}$ if $C_t(x) = y_i$
 - ii. $W_{t+1}(i) = \frac{W_t(i)}{Z_t} e^{\alpha_t}$ if $C_t(x) \neq y_i$
4. Output the final ensemble model: $E(x) = \text{sign} \left(\sum_{t=1}^T (\alpha_t C_t(x)) \right)$

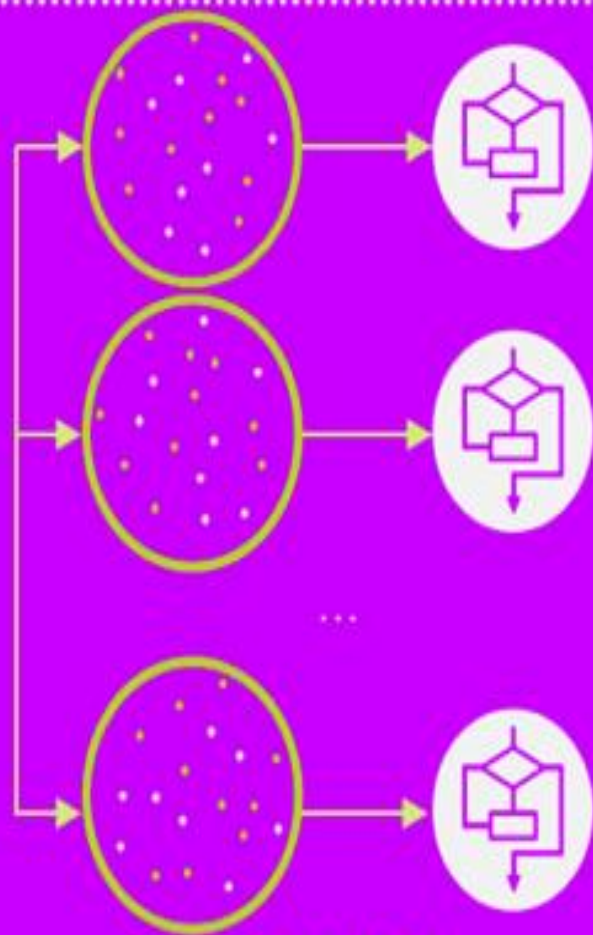
Bagging



Boosting

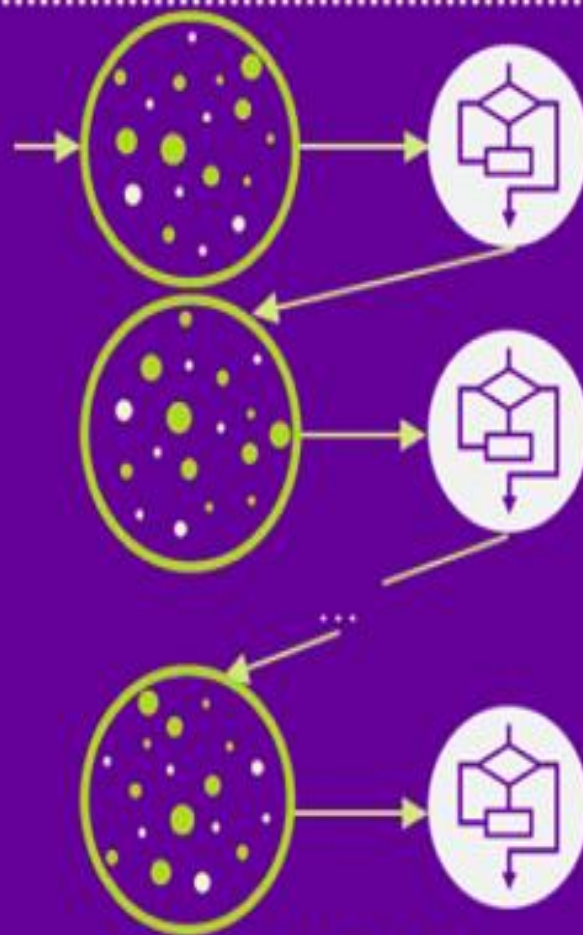


bagging



parallel

boosting



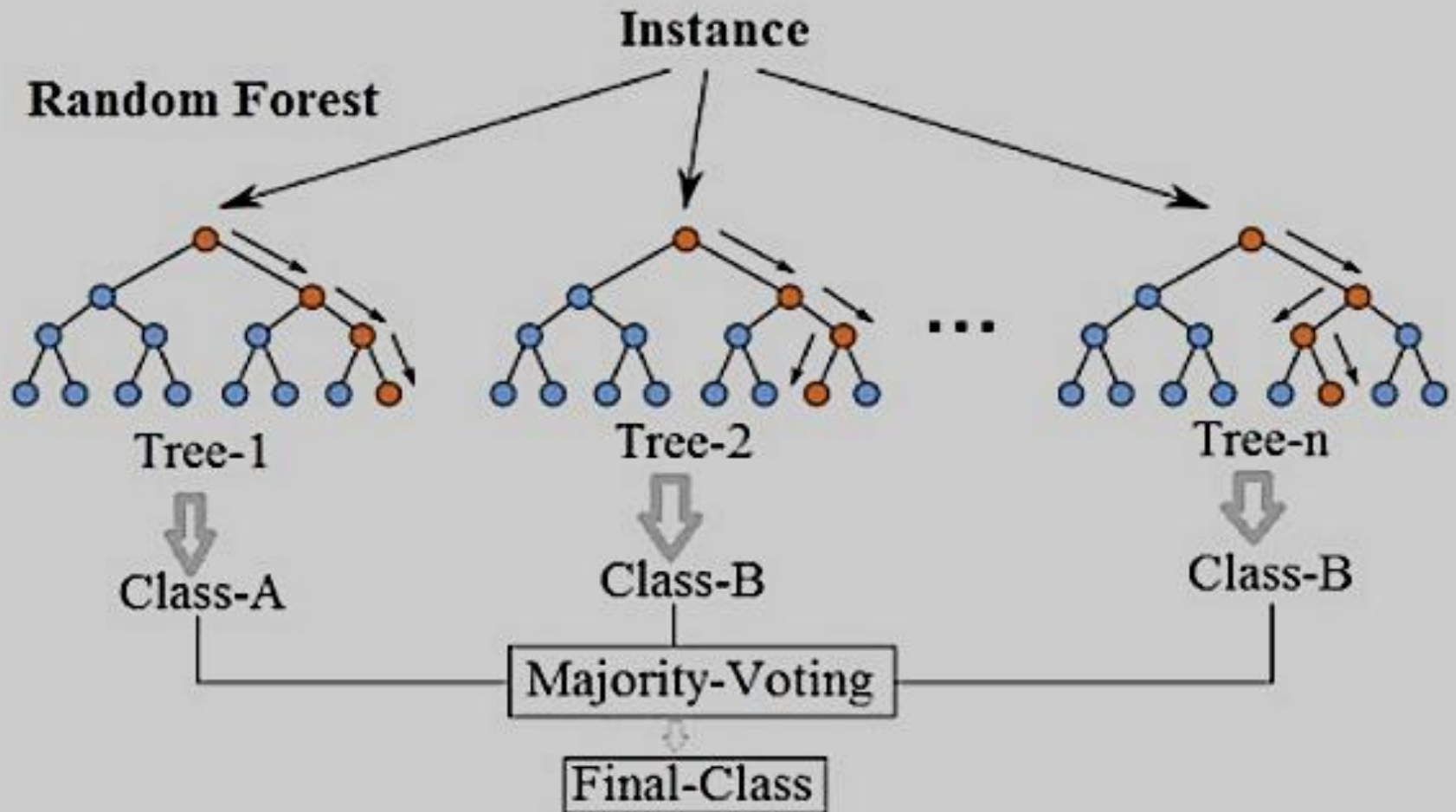
sequential

Random Forests

Forest of decision trees as follows:

- 1. Given a data set with n observations and N inputs**
- 2. m = constant chosen on beforehand**
- 3. For $t = 1, \dots, T$**
 - a. Take a bootstrap sample with n observations**
 - b. Build a decision tree whereby for each node of the tree, randomly choose m inputs on which to base the splitting decision**
 - c. Split on the best of this subset**
 - d. Fully grow each tree without pruning**

Random Forest Simplified



MULTICLASS CLASSIFICATION TECHNIQUES

- **Multiclass Logistic Regression**
- **Multiclass Decision Trees**
- **Multiclass Neural Networks**
- **Multiclass Support Vector Machines**

Multiclass Logistic Regression

- **When estimating a multiclass logistic regression model, one first needs to know whether the target variable is nominal or ordinal.**
- **Nominal targets could be predicting blood type and predicting voting behavior.**
- **Ordinal targets could be predicting credit ratings and predicting income as high, medium, or low.**

For nominal target variables, one of the target classes (say class K) will be chosen as the base class as follows:

$$\frac{P(Y = 1|X_1, \dots, X_N)}{P(Y = K|X_1, \dots, X_N)} = e^{(\beta_0^1 + \beta_1^1 X_1 + \beta_2^1 X_2 + \dots + \beta_N^1 X_N)}$$

$$\frac{P(Y = 2|X_1, \dots, X_N)}{P(Y = K|X_1, \dots, X_N)} = e^{(\beta_0^2 + \beta_1^2 X_1 + \beta_2^2 X_2 + \dots + \beta_N^2 X_N)}$$

...

$$\frac{P(Y = K-1|X_1, \dots, X_N)}{P(Y = K|X_1, \dots, X_N)} = e^{(\beta_0^{K-1} + \beta_1^{K-1} X_1 + \beta_2^{K-1} X_2 + \dots + \beta_N^{K-1} X_N)}$$

Using the fact that all probabilities must sum to 1, one can obtain the following:

$$P(Y = 1|X_1, \dots, X_N) = \frac{e^{(\beta_0^1 + \beta_1^1 X_1 + \beta_2^1 X_2 + \dots + \beta_N^1 X_N)}}{1 + \sum_{k=1}^{K-1} e^{(\beta_0^k + \beta_1^k X_1 + \beta_2^k X_2 + \dots + \beta_N^k X_N)}}$$

$$P(Y = 2|X_1, \dots, X_N) = \frac{e^{(\beta_0^2 + \beta_1^2 X_1 + \beta_2^2 X_2 + \dots + \beta_N^2 X_N)}}{1 + \sum_{k=1}^{K-1} e^{(\beta_0^k + \beta_1^k X_1 + \beta_2^k X_2 + \dots + \beta_N^k X_N)}}$$

$$P(Y = K|X_1, \dots, X_N) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{(\beta_0^k + \beta_1^k X_1 + \beta_2^k X_2 + \dots + \beta_N^k X_N)}}$$

The β parameters are then usually estimated using maximum aposteriori estimation, which is an extension of maximum likelihood estimation. As with binary logistic regression, the procedure comes with standard errors, confidence intervals, and p-values.

In case of ordinal targets, one could estimate a cumulative logistic regression as follows:

$$P(Y \leq 1) = \frac{1}{1 + e^{-\theta_1 + \beta_1 X_1 + \dots + \beta_N X_N}}$$

$$P(Y \leq 2) = \frac{1}{1 + e^{-\theta_2 + \beta_1 X_1 + \dots + \beta_N X_N}}$$

...

$$P(Y \leq K-1) = \frac{1}{1 + e^{-\theta_{K-1} + \beta_1 X_1 + \dots + \beta_N X_N}}$$

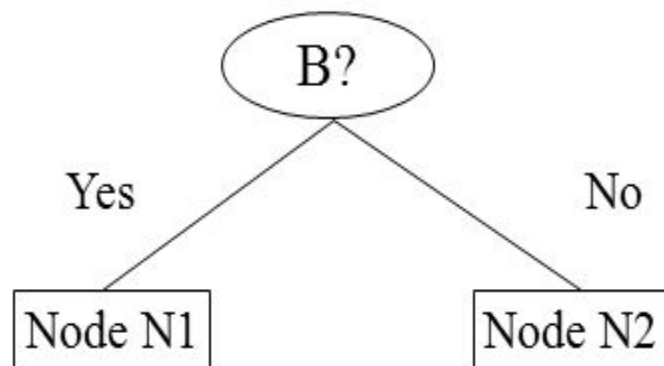
Multiclass Decision Trees

Decision trees can be easily extended to a multiclass setting. For the splitting decision, assuming K classes, the impurity criteria become:

$$Entropy(S) = - \sum_{k=1}^K p_k \log_2(p_k)$$

$$Gini(S) = \sum_{k=1}^K p_k(1 - p_k)$$

- Splits into two partitions
- Effect of Weighing partitions:
 - Larger and Purer Partitions are sought for.



	Parent
C1	6
C2	6
Gini = 0.500	

$$\begin{aligned}
 &\mathbf{Gini(N1)} \\
 &= 1 - (5/6)^2 - (1/6)^2 \\
 &= 0.278
 \end{aligned}$$

$$\begin{aligned}
 &\mathbf{Gini(N2)} \\
 &= 1 - (2/6)^2 - (4/6)^2 \\
 &= 0.444
 \end{aligned}$$

	N1	N2
C1	5	2
C2	1	4
Gini=0.361		

$$\begin{aligned}
 &\mathbf{Gini(Children)} \\
 &= 6/12 * 0.278 + \\
 &\quad 6/12 * 0.444 \\
 &= 0.361
 \end{aligned}$$

$$Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t)$$

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$Entropy = -0 \log 0 - 1 \log 1 = -0 - 0 = 0$$

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$Entropy = - (1/6) \log_2 (1/6) - (5/6) \log_2 (5/6) = 0.65$$

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$Entropy = - (2/6) \log_2 (2/6) - (4/6) \log_2 (4/6) = 0.92$$

Multiclass Neural Networks

- **Multiclass neural network for K classes, is to create K output neurons**
- **An observation is then assigned to the output neuron with the highest activation value**

Multiclass Support Vector Machines

- **Multiclass support vector machine is to map the multiclass classification problem to a set of binary classification problems.**
- **Two well-known schemes here are**
 - one-versus-one**
 - one-versus-all coding**

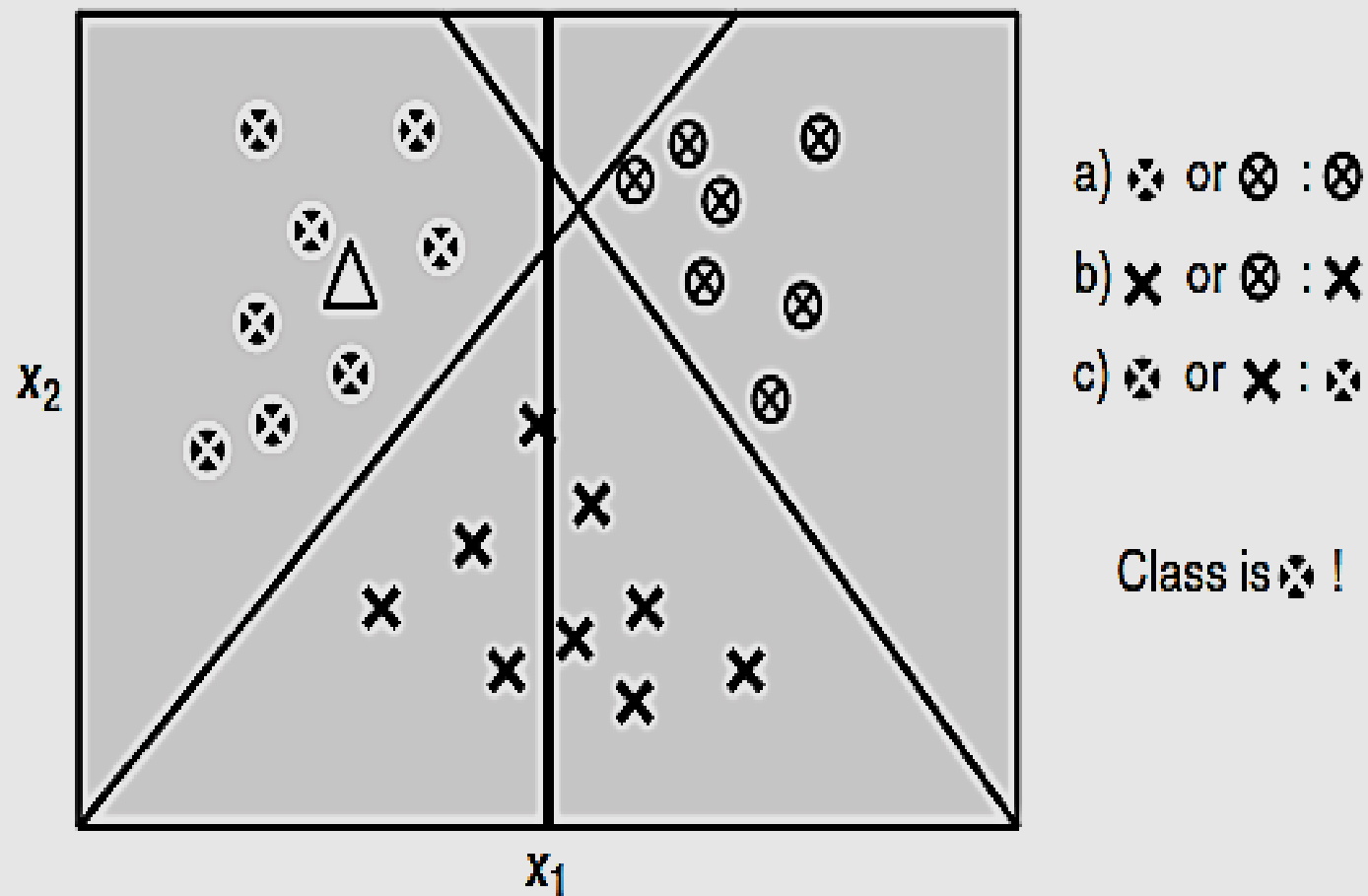
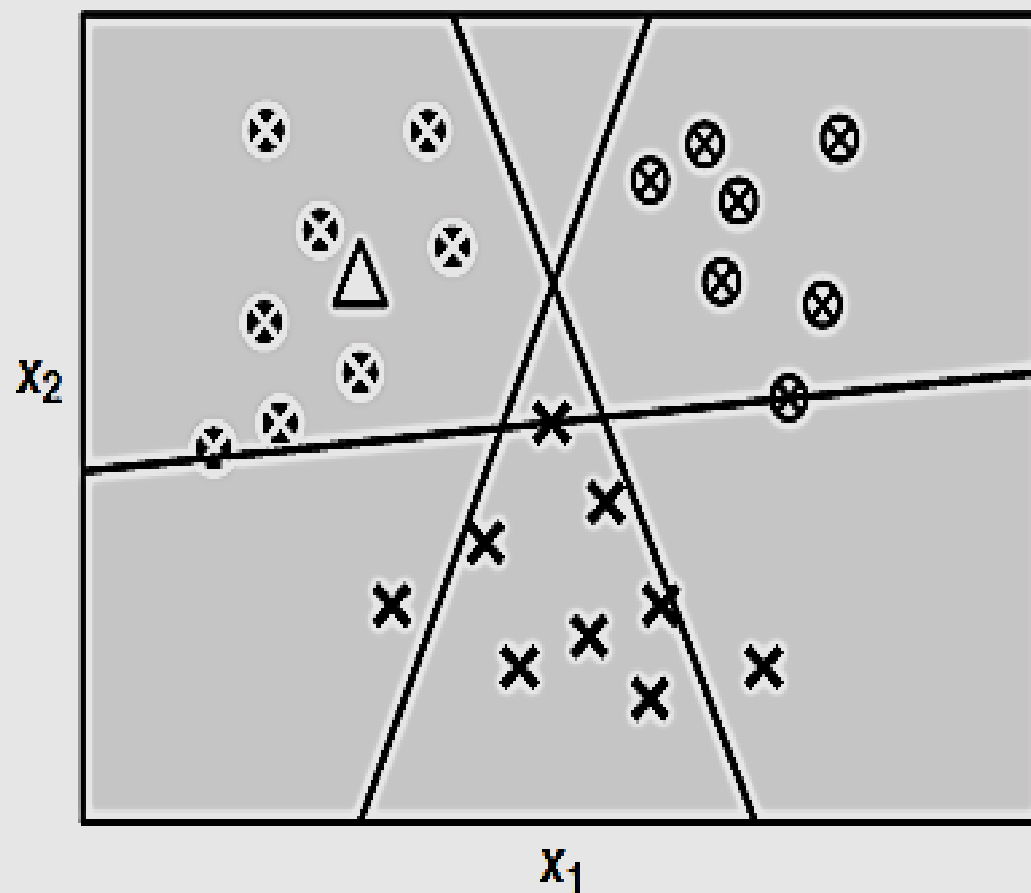


Figure 3.25 One-versus-One Coding for Multiclass Problems



a) ⊗ or other; $p(\otimes) = 0.92$

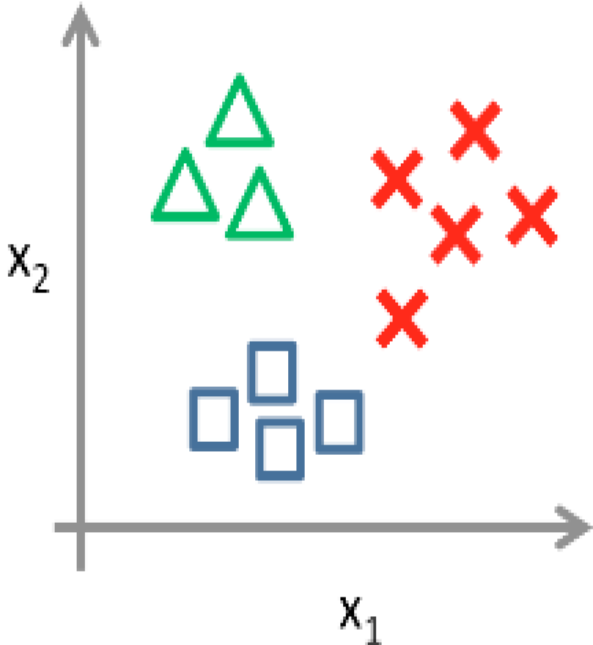
b) ⊗ or other; $p(\otimes) = 0.18$




c) x or other; $p(x) = 0.30$

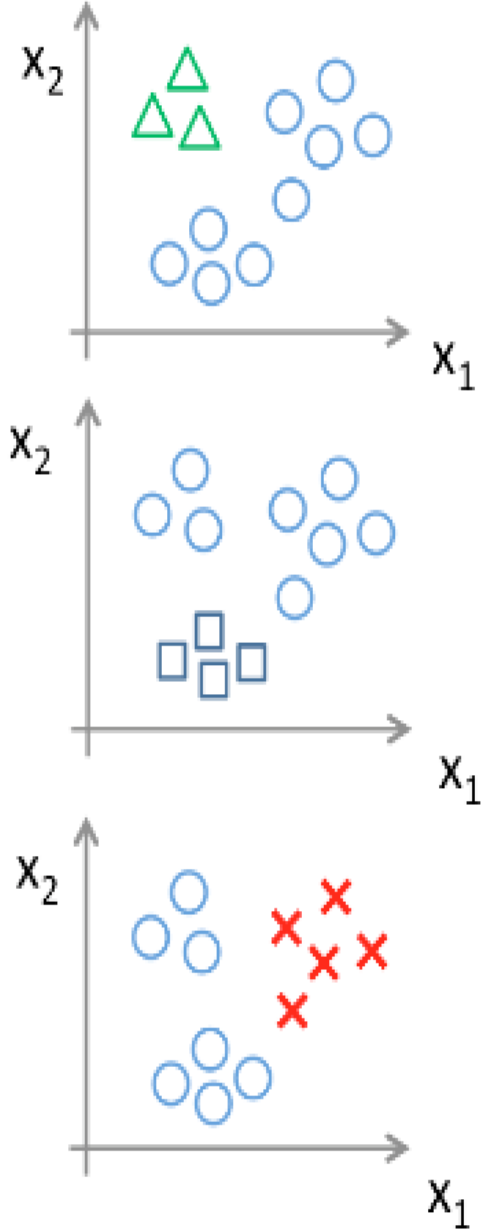
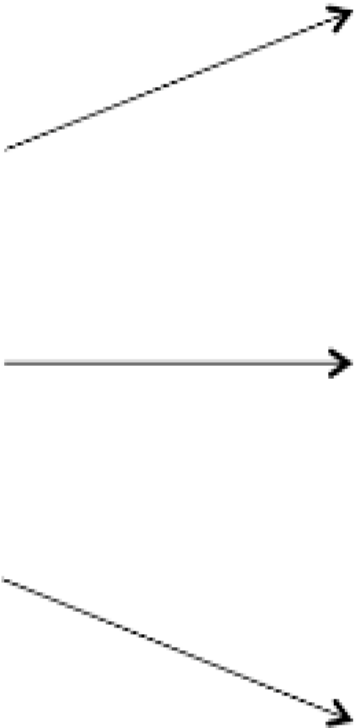
Class is ⊗ !

Figure 3.26 One-versus-All Coding for Multiclass Problems

One-vs-all (one-vs-rest):



- Class 1: 
- Class 2: 
- Class 3: 



EVALUATING PREDICTIVE MODELS

- Splitting Up the Data Set**
- Performance Measures for Classification Models**
- Performance Measures for Regression Models**

Splitting Up the Data Set

- **Which specifies on what part of the data the performance will be measured.**
- **Decision concerns the performance metric.**
- **The data can be split up into a training and a test sample**
- **The training sample (also called development or estimation sample) will be used to build the model,**
- **The test sample (also called the hold out sample) will be used to calculate its performance**

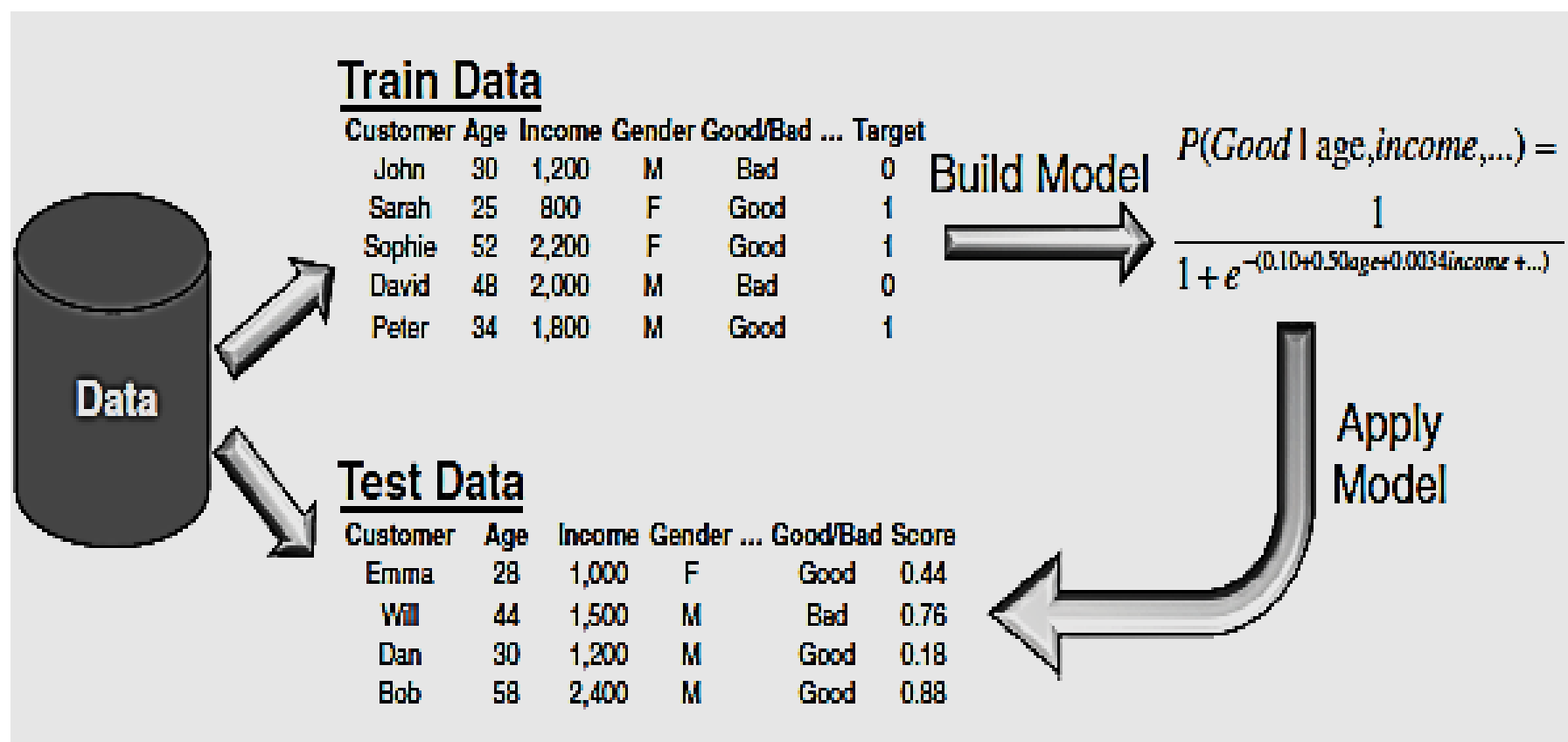


Figure 3.27 Training versus Test Sample Set Up for Performance Estimation

- **Cross- validation, the data is split into K folds.**
- **A model is then trained on $K-1$ training folds and tested on the remaining validation fold.**
- **Validation folds resulting in K performance estimates that can then be averaged.**
- **Cross-validation becomes leave-one-out.**
- **To make sure the good/bad.**

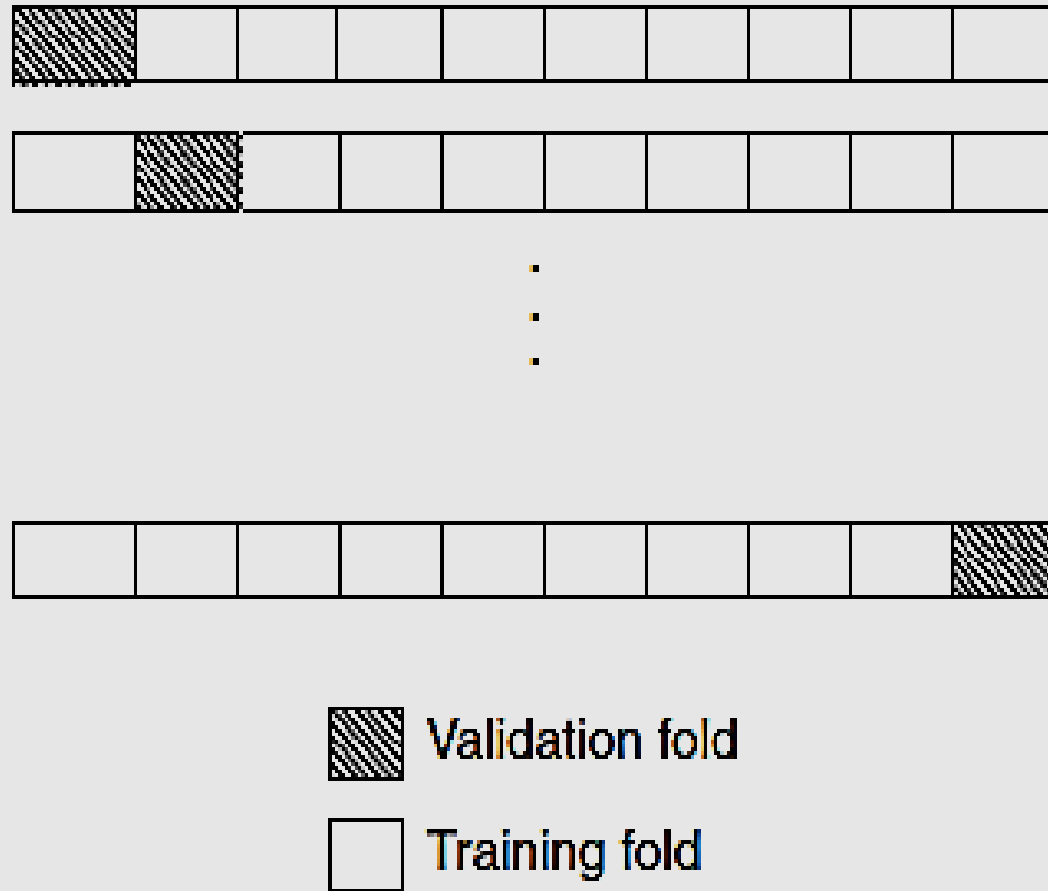


Figure 3.28 Cross-Validation for Performance Measurement

Performance Measures for Classification Models

- Map the scores to predicted classification label by assuming a **default cutoff of 0.5**

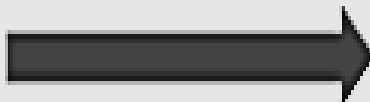
Churn Score				Churn Score Predicted			
John	Yes	0.72	Cutoff = 0.50 	John	Yes	0.72	Yes
Sophie	No	0.56		Sophie	No	0.56	Yes
David	Yes	0.44		David	Yes	0.44	No
Emma	No	0.18		Emma	No	0.18	No
Bob	No	0.36		Bob	No	0.36	No

Figure 3.30 Calculating Predictions Using a Cut-Off

Terms associated with Confusion matrix:

(1) True Positives (TP):

True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True)

Example:

The case where a person is actually having cancer(1) and the model classifying his case as cancer(1) comes under True positive.

2. True Negatives (TN):

True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False)

- ***Example:***

The case where a person NOT having cancer and the model classifying his case as Not cancer comes under True Negatives.

3. False Positives (FP):

False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one. (1)

Example:

A person NOT having cancer and the model classifying his case as cancer comes under False Positives.

4. False Negatives (FN):

False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False).

False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0)

Example:

A person having cancer and the model classifying his case as No-cancer comes under False Negatives.

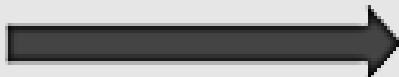
Churn Score				Churn Score Predicted			
John	Yes	0.72	Cutoff = 0.50 	John	Yes	0.72	Yes
Sophie	No	0.56		Sophie	No	0.56	Yes
David	Yes	0.44		David	Yes	0.44	No
Emma	No	0.18		Emma	No	0.18	No
Bob	No	0.36		Bob	No	0.36	No

Figure 3.30 Calculating Predictions Using a Cut-Off

		Actual Status (Churn)	
		Yes	No
Predicated Status	Yes		
	No		

Based upon this matrix, one can now calculate the following performance measures:

- Classification accuracy = $(TP + TN) / (TP + FP + FN + TN) = 3/5$
- Classification error = $(FP + FN) / (TP + FP + FN + TN) = 2/5$
- Sensitivity = $TP / (TP + FN) = 1/2$
- Specificity = $TN / (FP + TN) = 2/3$

Performance Measures for Classification Models

Multiple measures exist to calculate the performance of regression models. A first key metric is the R -squared, defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

whereby y_i is the true value, \hat{y}_i the predicted value, and \bar{y} the average. The R^2 always varies between 0 and 1, and higher values are to be preferred. Two other popular measures are the mean squared error (MSE) and mean absolute deviation (MAD), defined as follows:

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

$$MAD = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

A scatter plot between the predicted and the target values can give a visual representation of model performance (see Figure 3.38). The more the plot approaches a straight line through the origin, the better the performance of the model. It can be summarized by calculating the Pearson correlation coefficient.

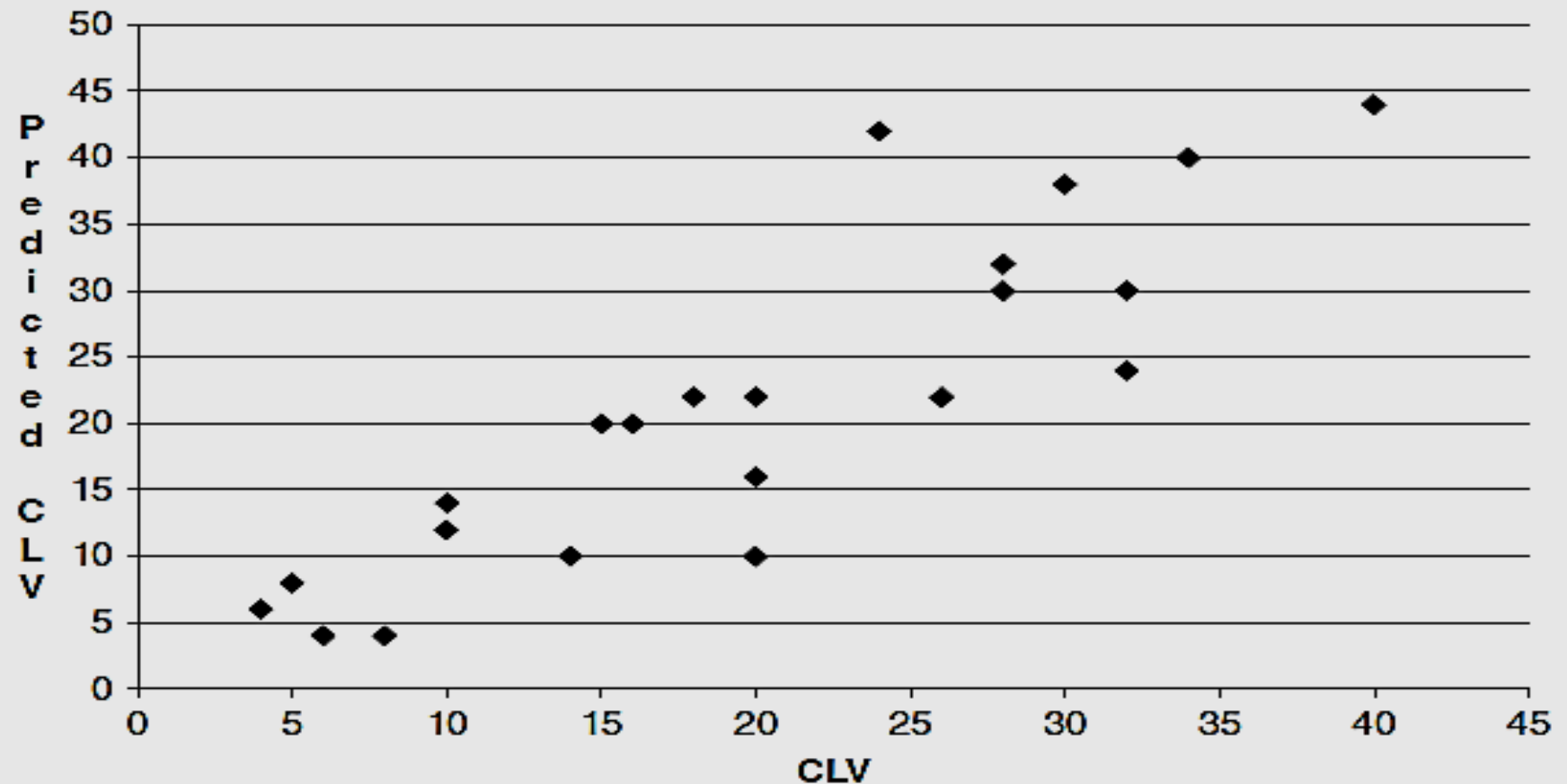


Figure 3.38 Scatter Plot for Measuring Model Performance



ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCES

(Autonomous)

G.B. Nagar, Kalavai - 632506



Big data analytics

Unit - III

DESCRIPTIVE ANALYTICS

- **The aim is to describe patterns of customer behavior.**
- **There is no real target variable available.**
- **Descriptive analytics is often referred to as unsupervised learning because there is no target variable**

Table 4.1 Examples of Descriptive Analytics

Type of Descriptive Analytics	Explanation	Example
Association rules	Detect frequently occurring patterns between items	Detecting what products are frequently purchased together in a supermarket context Detecting what words frequently co-occur in a text document Detecting what elective courses are frequently chosen together in a university setting
Sequence rules	Detect sequences of events	Detecting sequences of purchase behavior in a supermarket context Detecting sequences of web page visits in a web mining context Detecting sequences of words in a text document
Clustering	Detect homogeneous segments of observations	Differentiate between brands in a marketing portfolio Segment customer population for targeted marketing

ASSOCIATION RULES

- **How to mine association rules from data.**
- **Topics are**

Basic setting

Support and Confidence

Association Rule Mining

The Lift Measure

Post Processing Association Rules

Association Rule Extension

Applications of Association Rules

Basic Setting:

- **Association rules typically start from a database of transactions, D**
- **Each transaction consists of a transaction identifier and a set of items ($i_1, i_2, i_3, \dots, i_n$)**
- **The following table an example of a transaction database in a supermarket.**

Table 4.2 Example Transaction Data Set

Transaction Identifier	Items
1	Beer, milk, diapers, baby food
2	Coke, beer, diapers
3	Cigarettes, diapers, baby food
4	Chocolates, diapers, milk, apples
5	Tomatoes, water, apples, beer
6	Spaghetti, diapers, baby food, beer
7	Water, beer, baby food
8	Diapers, baby food, spaghetti
9	Baby food, beer, diapers, milk
10	Apples, wine, baby food

- **An association rule is then an implication of the form $X \Rightarrow Y$, whereby $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. X is referred to as the rule**
- **Example of association rules are:**
 - ✓ **If a customer buys spaghetti, then the customer buys oil in 70% of the cases.**
 - ✓ **If a customer visits web page A, then the customer will visit web page B in 90% of the cases.**
- **The rules measure correlational associations and should not be interpreted in a causal way**

Support and Confidence

- **Support and confidence are two key measures to quantify the strength of an association rule.**
- **The support of an item set is defined as the percentage of total transactions in the database that contains the item set.**
- **Hence, the rule $X \Rightarrow Y$ has support (s) if $100 s \%$ of the transactions in D contain $X \cup Y$.**

- It can be formally defined as follows:

$$\text{support}(X \cup Y) = \frac{\text{number of transactions supporting } (X \cup Y)}{\text{total number of transactions}}$$

The rule $X \Rightarrow Y$

baby food and diapers \Rightarrow beer

has support 3/10 or 30 percent.

A frequent item set is one for which the support is higher than a threshold (minsup) that is typically specified upfront by the business user or data analyst.

- The confidence measures the strength of the association and is defined as the conditional probability of the rule consequent.
- The rule $X \Rightarrow Y$ has confidence (c) if 100c % of the transactions in D that contain X also contain Y .
- It can be formally defined as follows:

$$\text{confidence}(X \rightarrow Y) = P(Y|X) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$$

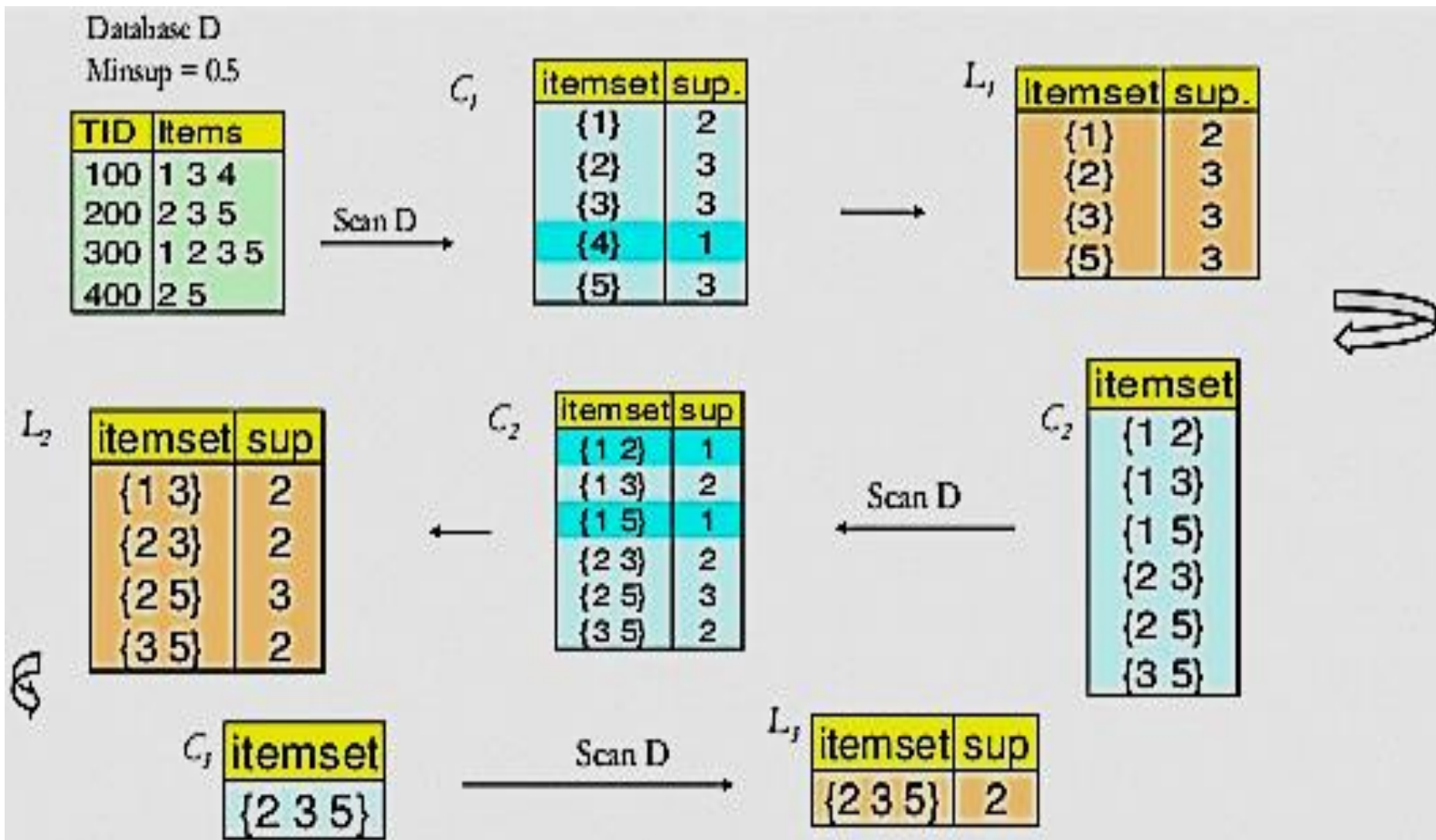
- the association rule **baby food and diapers \Rightarrow beer** has confidence 3/5 or 60 percent.

Association Rule Mining

Mining association rules from data is essentially a two-step process as follows:

- 1. Identification of all item sets having support above minsupport (i.e., “frequent” item sets)**
 - 2. Discovery of all derived association rules having confidence above minconfidence**
- Both minsup and minconf need to be specified beforehand by the data analyst**

- Performed using the Apriori algorithm



Database

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

L_1

Itemsets	Support
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

Minsup = 50%

C_2

Itemsets	Support
{1, 2}	1/4
{1, 3}	2/4
{1, 5}	1/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4

L_2

Itemsets	Support
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4

C_3

Itemsets	Support
{2, 3, 5}	2/4

L_3

Itemsets	Support
{2, 3, 5}	2/4

{1,3} and {2,3} give
{1,2,3}, but because {1,2}
is not frequent, you do not
have to consider it!

Result = { {1},{2},{3},{5},{1,3},{2,3},{2,5},{3,5},{2,3,5} }

Figure 4.1 The Apriori Algorithm

For the frequent item set {baby food, diapers, beer}, the following association rules can be derived:

diapers, beer \Rightarrow baby food [conf = 75%]

baby food, beer \Rightarrow diapers [conf = 75%]

baby food, diapers \Rightarrow beer [conf = 60%]

beer \Rightarrow baby food and diapers [conf = 50%]

baby food \Rightarrow diapers and beer [conf = 43%]

diapers \Rightarrow baby food and beer [conf = 43%]

If the minconf is set to 70 percent, only the first two association rules will be kept for further analysis.

Lift Measure

- The lift, also referred to as the **interestingness** measure, takes this into account by incorporating the prior probability of the rule consequent, as follows:

$$Lift(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X) \cdot support(Y)}$$

Post Processing Association Rules

1. Filter out the trivial rules:

- Contains already known patterns (e.g buying a baby food with beer).**
- This should be done in collaboration with a business expert.**

2. Perform a sensitivity analysis:

- Varying the minimum support and min confidence value**
- (E.g find out the effect of a company's net working capital on its profit margin.)**

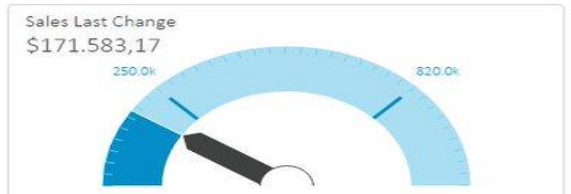
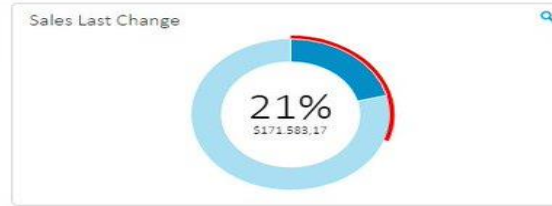
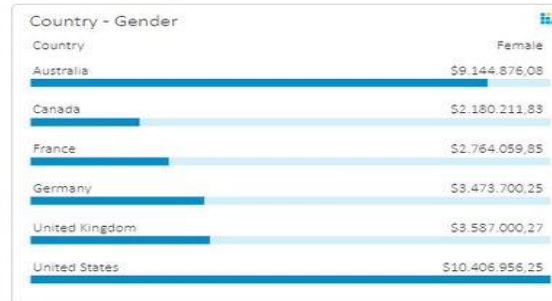
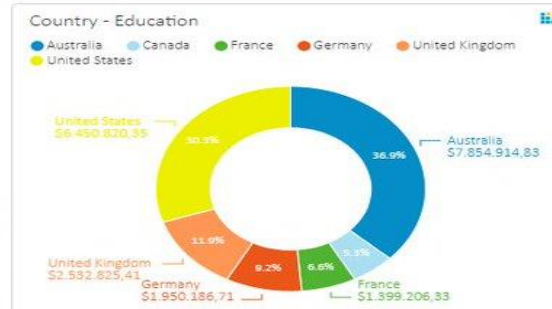
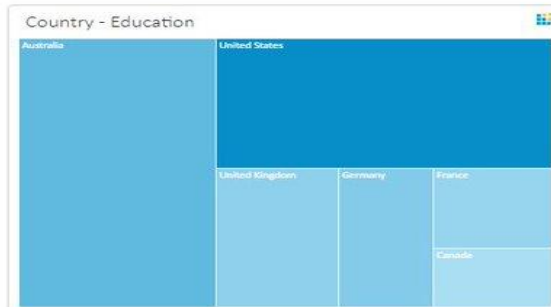
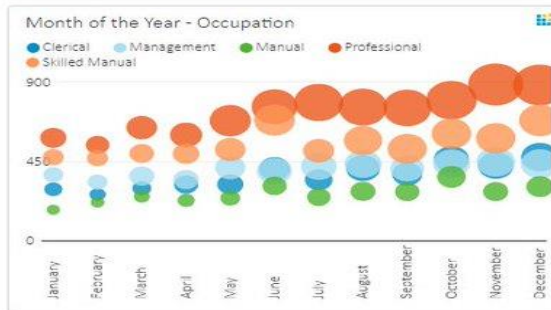
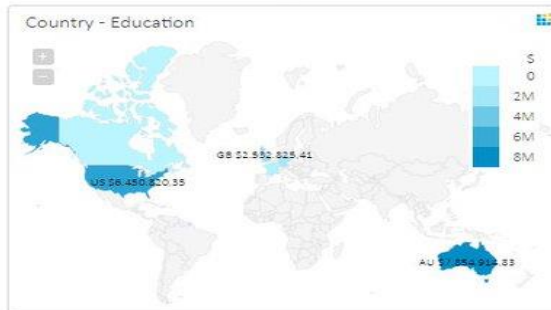
3. Use appropriate visualization facilities

- (E.g : OLAP) to find the unexpected rules that might represent novel and actionable behavior in the data**

4. Measure the economic impact

- (E.g profit, cost) of the association rules.**

Internet Sales by Product Q3



Applications of Association Rules

1. Market basket analysis :

To detect which product or services are frequently purchased together.

Important implications for target marketing

(e.g : next best offer

product bundling

store and shelf layout

catalog design)



? Where should detergents be placed in the Store to maximize their sales?

? Are window cleaning products purchased when detergents and orange juice are bought together?

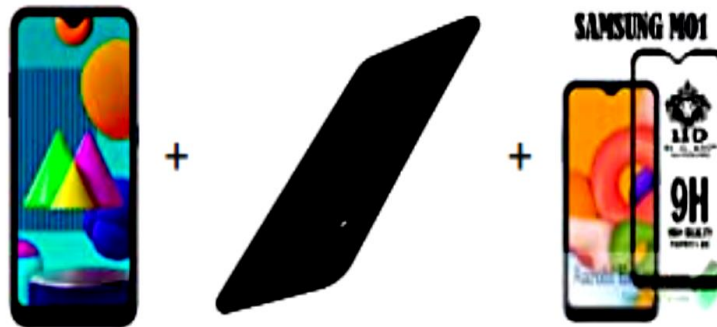
? Is soda typically purchased with bananas? Does the brand of soda make a difference?

? How are the demographics of the neighborhood affecting what customers are buying?

2. Recommender systems :

The systems adopted by companies such as Amazon and flipkart to give a recommendation based on past purchase

Frequently bought together



Total price: ₹ 8,747.00

Add all three to Cart

SEQUENCE RULES:

- **To find the maximal sequences among all sequences.**

Example :

**Home page => Electronics => cameras => Digital
camera => shopping cart => order confirmation
=> return to shopping**

- **Transaction time or sequence field included in the analysis.**

Table 4.4 Example Transactions Data Set for Sequence Rule Mining

Session ID	Page	Sequence
1	A	1
1	B	2
1	C	3
2	B	1
2	C	2
3	A	1
3	C	2
3	D	3
4	A	1
4	B	2
4	D	3
5	D	1
5	C	1
5	A	1

- The letters A, B, C, D refers to web pages.
- A sequential version can then be obtained as follows:

Session 1 : A, B, C

Session 2 : B, C

Session 3 : A, C, D

Session 4 : A, B, D

Session 5 : D, C, A

Sequence rule $A \Rightarrow C$, the support = 2/5 (40%)

Confidence = 2/4 (50%)

SEGMENTATION (Division):

- **To split up a set of customer observations.**
- **Homogeneity within a segment is maximized (cohesive)**
- **Heterogeneity between segment is maximized (separated)**
- **Applications:**
 - **Understanding a customer population**
 - **Efficiently allocating marketing resources**
 - **Differentiating between brands and in a collection**

- **Identifying the most profitable customer**
- **Identifying shopping pattern**
- **Identifying the need for new products**
- **Various types of clustering data can be used**

Demographic

Lifestyle

Attitudinal

Behavioral

Social network

Clustering categorized as either hierarchical or nonhierarchical

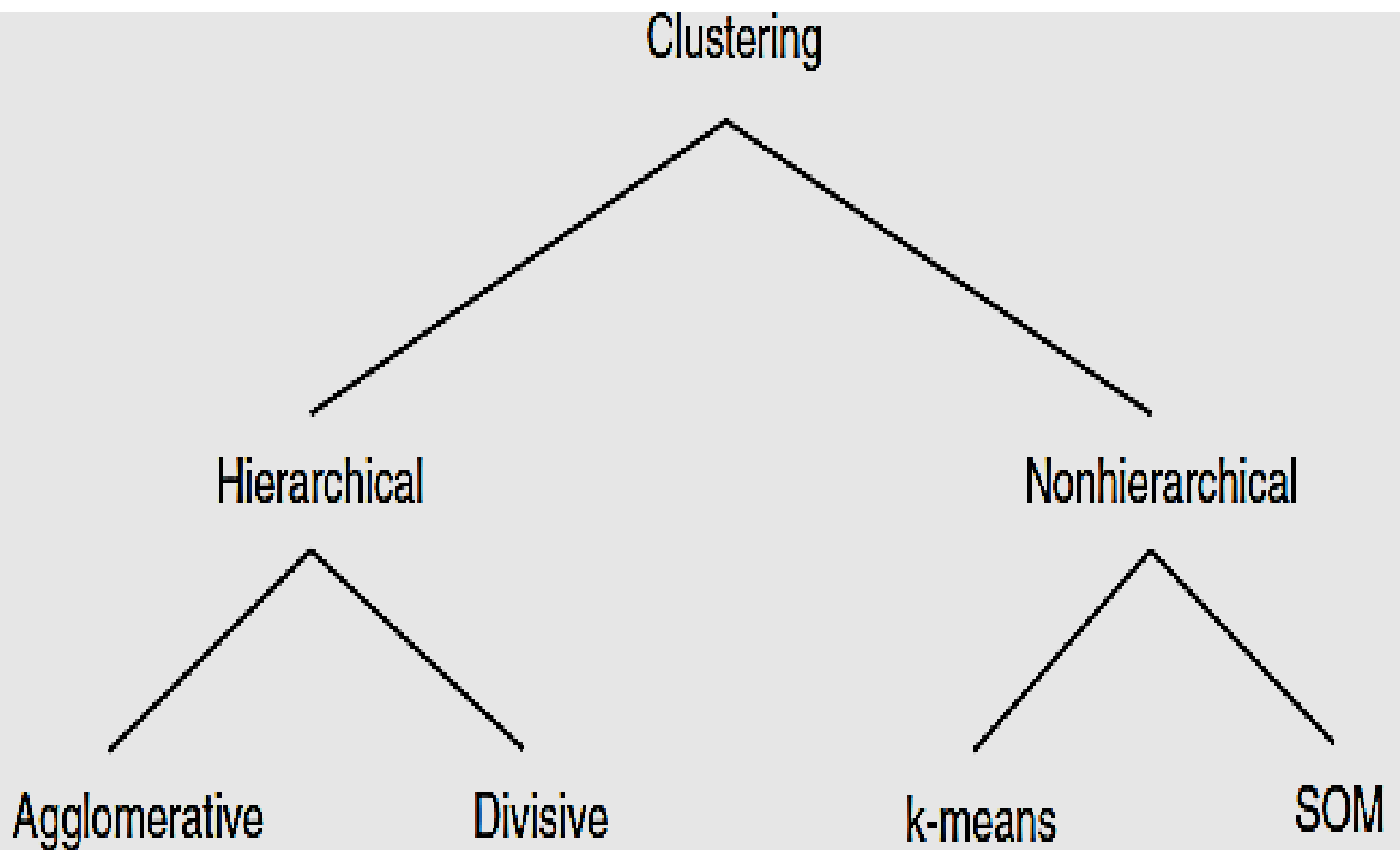


Figure 4.3 Hierarchical versus Nonhierarchical Clustering Techniques

Hierarchical clustering

- **Divisive hierarchical clustering** starts from the whole data set in one cluster, and then breaks this up in each time smaller cluster until one observation per clustering remains.
- **Agglomerative clustering** works the other way around, starting from all observation in one cluster and continuing to merge the ones that are most similar until all observations make up one big cluster.

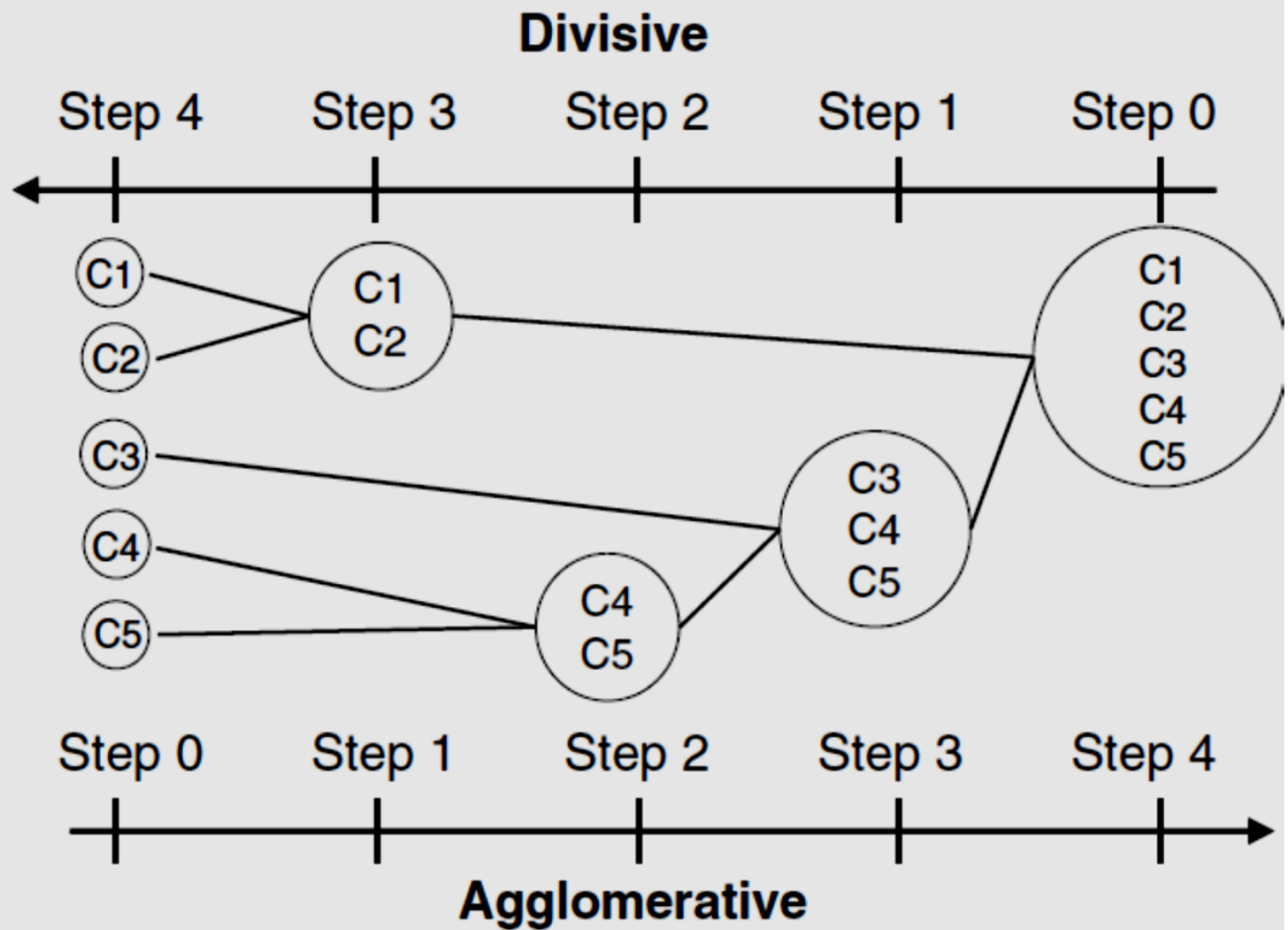


Figure 4.4 Divisive versus Agglomerative Hierarchical Clustering

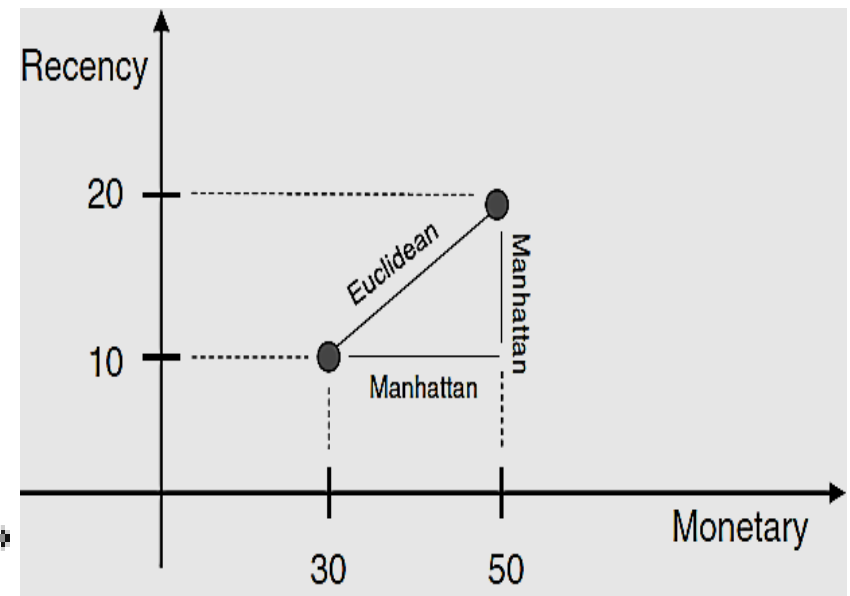
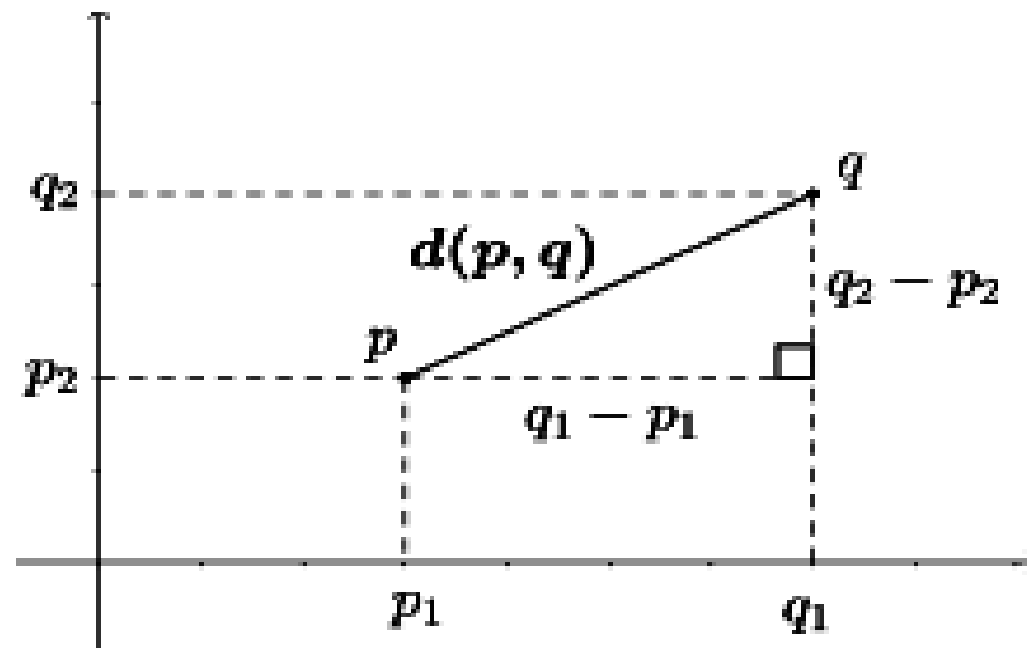
- **Merge or splitting, a similarity rule is needed.**
- **Example :**

Euclidean distance

Euclidean distance is the "ordinary" straight-line distance between two points.

Manhattan

To absolute differences between coordinates of a pair of objects.



Euclidean distance :

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$

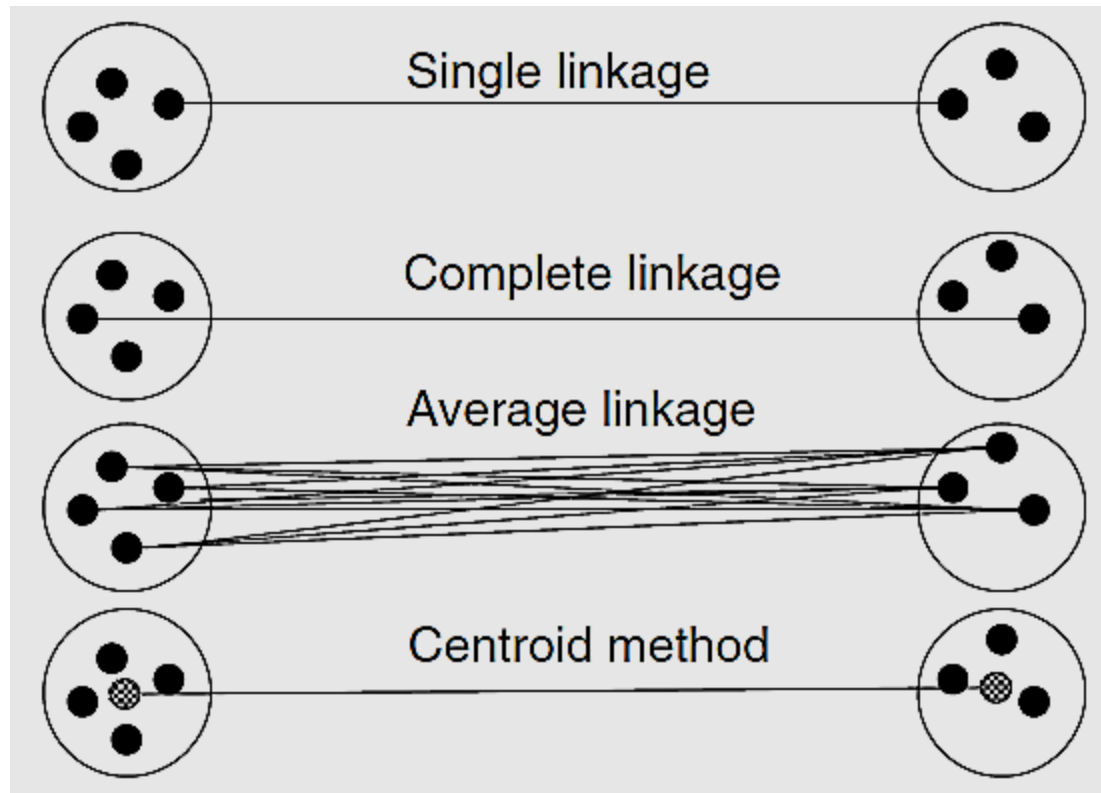
Manhattan distance :

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^n |p_i - q_i|,$$

$$\text{Euclidean: } \sqrt{(50 - 30)^2 + (20 - 10)^2} = 22$$

$$\text{Manhattan: } |50 - 30| + |20 - 10| = 30$$

- **The Euclidean distance will always be shorter than the Manhattan distance.**
- **Calculating Distance between clusters**



- **Single linkage** method defines the distance between two clusters as the shortest possible distance or the distance between the two most similar object
- **The Complete linkage** method defines the distance between two clusters as the biggest distance, or the distance between the two most dissimilar objects.
- The **average linkage** method calculates the average of all possible distances

- **The centroid method** calculates the distance between the centroids of both clusters.
- **Ward's method** merges the pair of clusters that leads to the minimum increase in total within-cluster variance after merging.

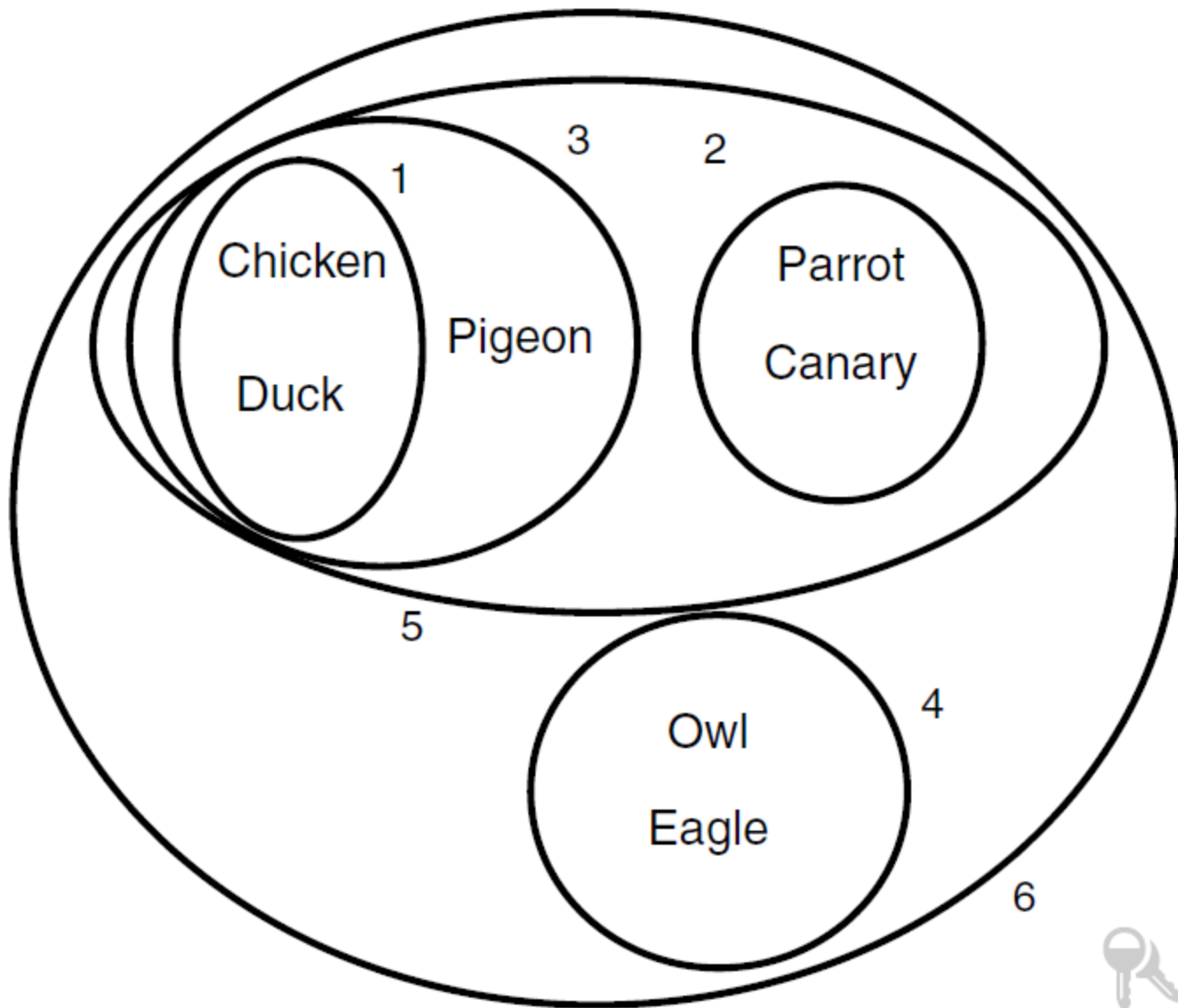


Fig: Clustering Birds. The number indicates the clustering steps

- To decide on the optimal no.of clusters, one could use a **dendrogram** or **scree plot**.
- The dendrogram is a tree-like diagram that records the sequences of merges.
- To find the optimal (best) clustering

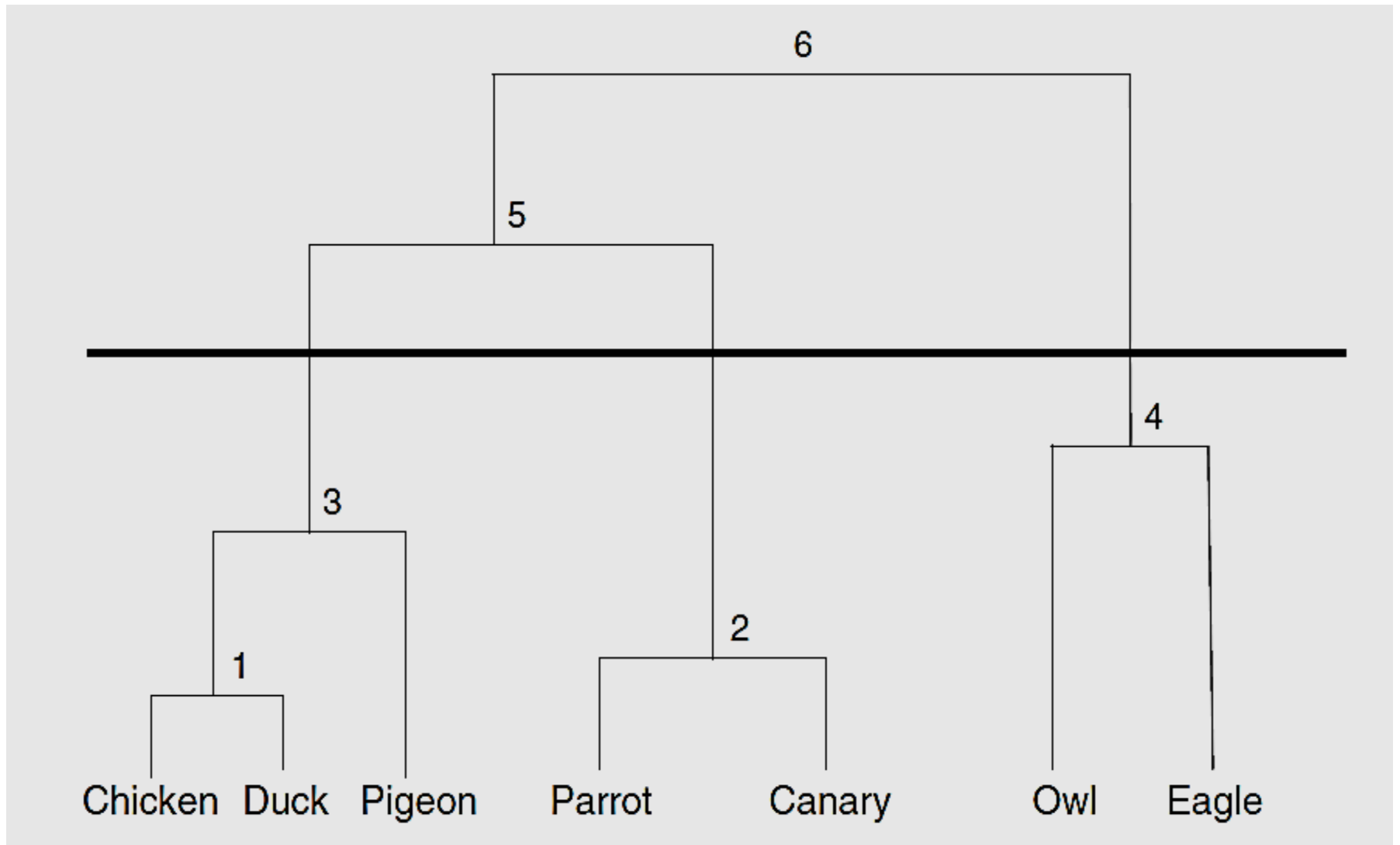
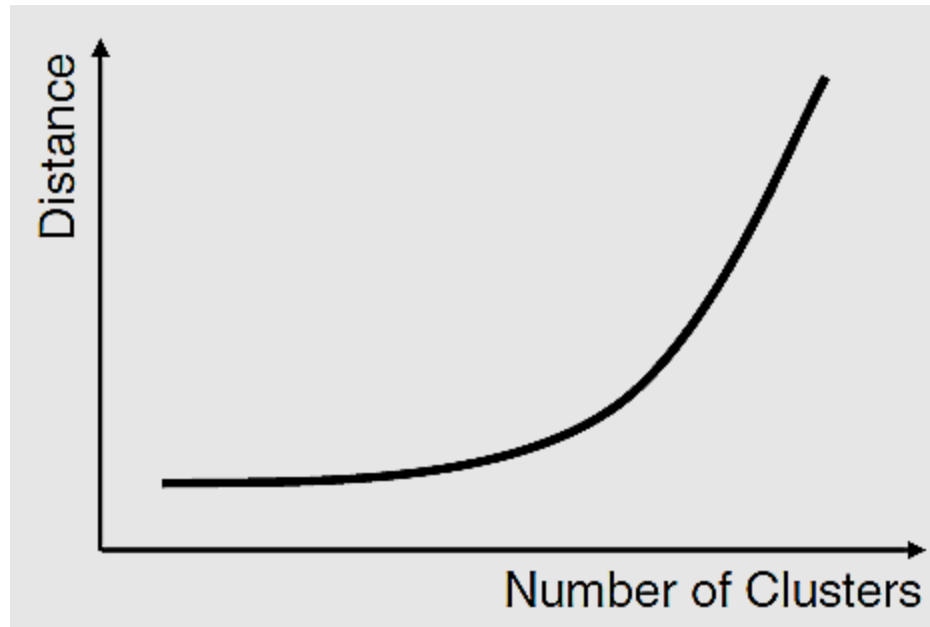


Fig: Dendrogram for Birds. The black line indicates the optimal clustering

- A **scree plot** is a plot of the distance at which clusters are merged.
- The 'elbow point' then indicates the optimal clustering

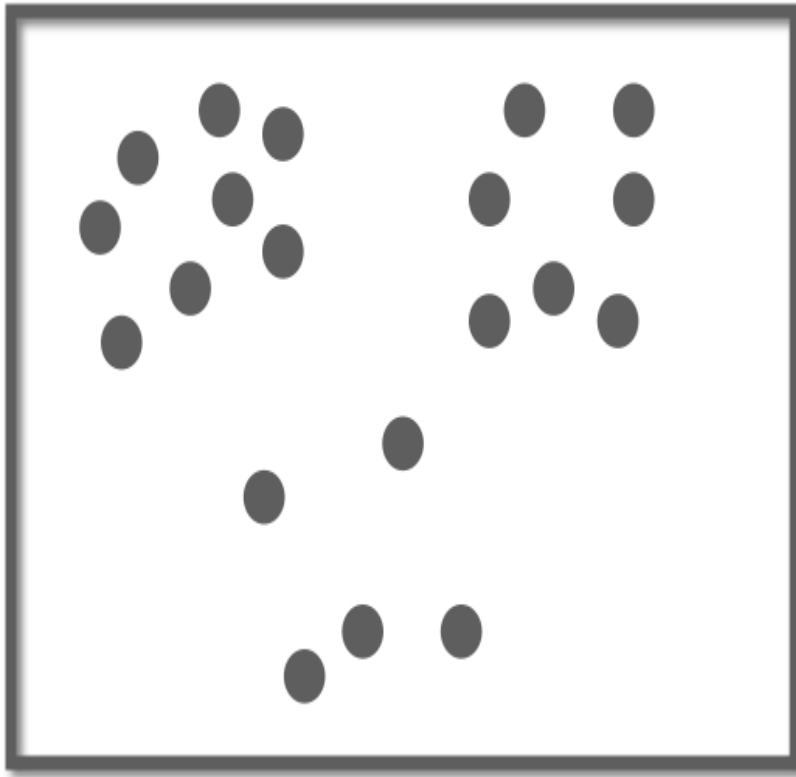


K-Means clustering

K-Means clustering is a nonhierarchical procedure that works along the following steps:

- 1. Select k observation as initial centroids**
- 2. Assign each observation to the cluster that has the closet centroid**
- 3. When all observations have been assigned, recalculate the position of the k centroids.**
- 4. Repeat until the cluster centroids no longer change.**

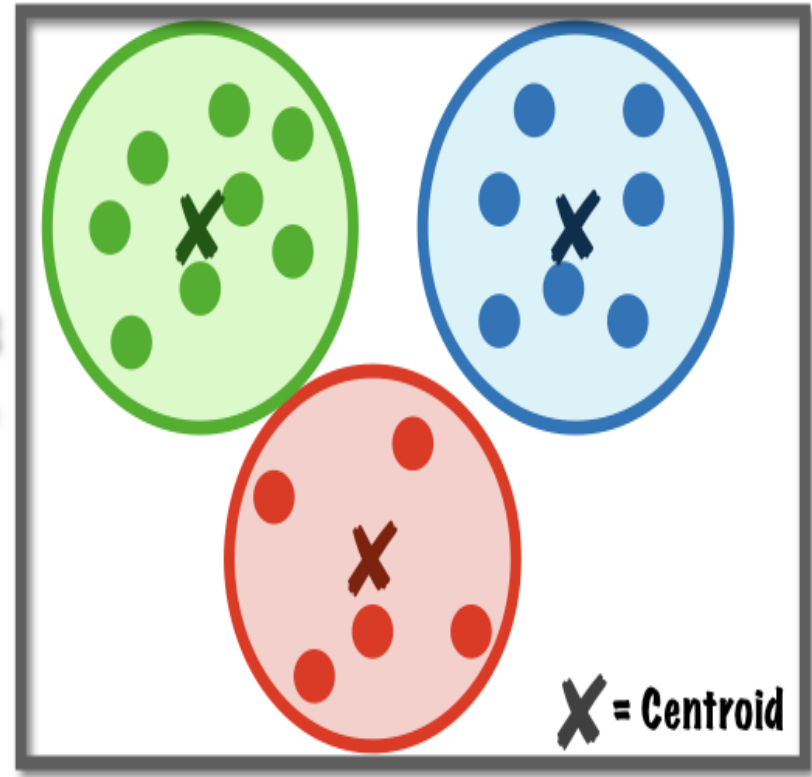
Unlabelled Data



K-means



Labelled Clusters

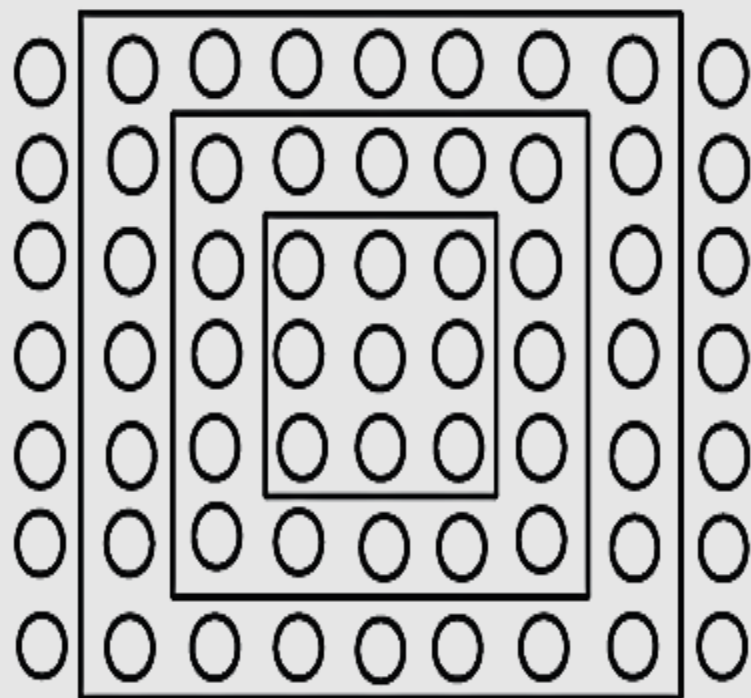


- **To be specified no.of cluster before the start of the analysis**

Self-Organizing Maps

- **Self-organizing map (SOM) is an unsupervised learning algorithm**
- **To visualize and cluster high-dimensional data on a low-dimensional grid of neurons.**
- **The neurons ordered in a two-dimensional rectangular or hexagonal grid**
- **Every neuron has at most eight neighbors, whereas for the later every neuron has at most six neighbors.**

Rectangular SOM Grid



Hexagonal SOM Grid

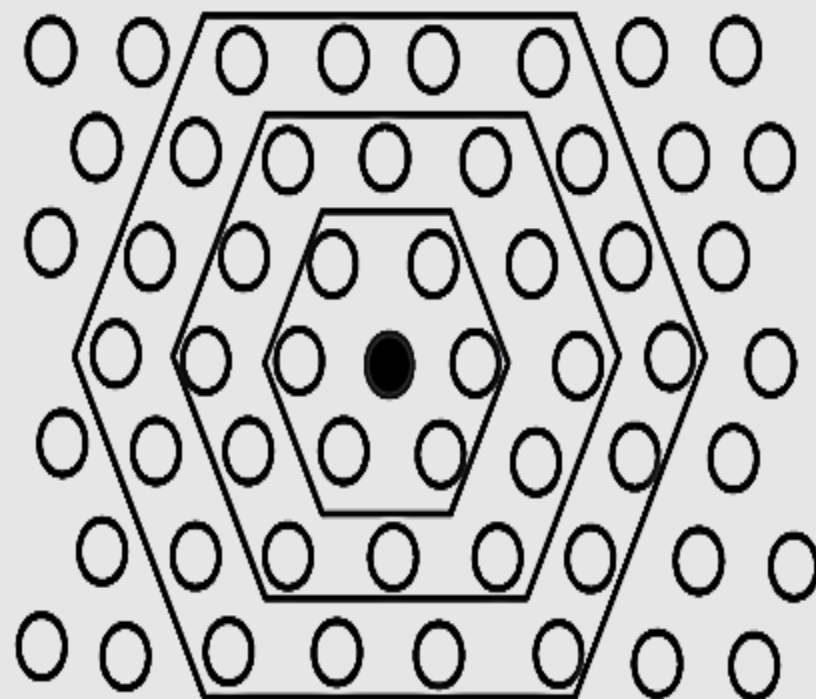


Figure 4.10 Rectangular versus Hexagonal SOM Grid

SURVIVAL ANALYSIS

- **Survival analysis is a set of statistical techniques focusing on the occurrence and timing of events.**

Example

- **Predict when customers churn**
- **Predict when customers make their next purchase**
- **Predict when customers default**
- **Predict when customers pay off their loan early**
- **Predict when customer will visit a website next**

- **A first key problem is censoring (removing)**
- **Censoring refers to the fact that the target time variable is not always known because not all customers may have experienced the event yet at the time of the analysis.**
- **Churn means that customers stop contributing to a facility**

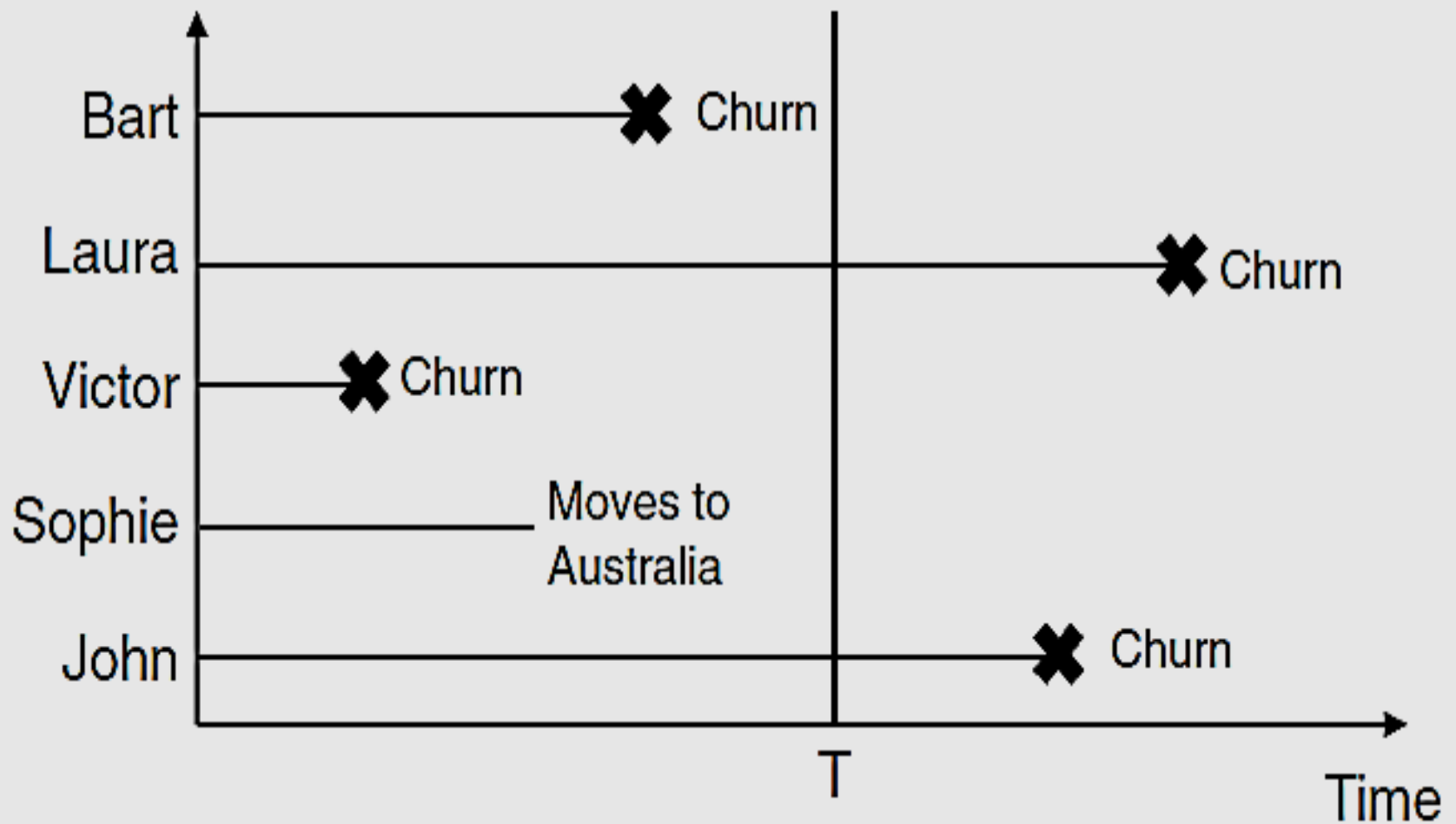
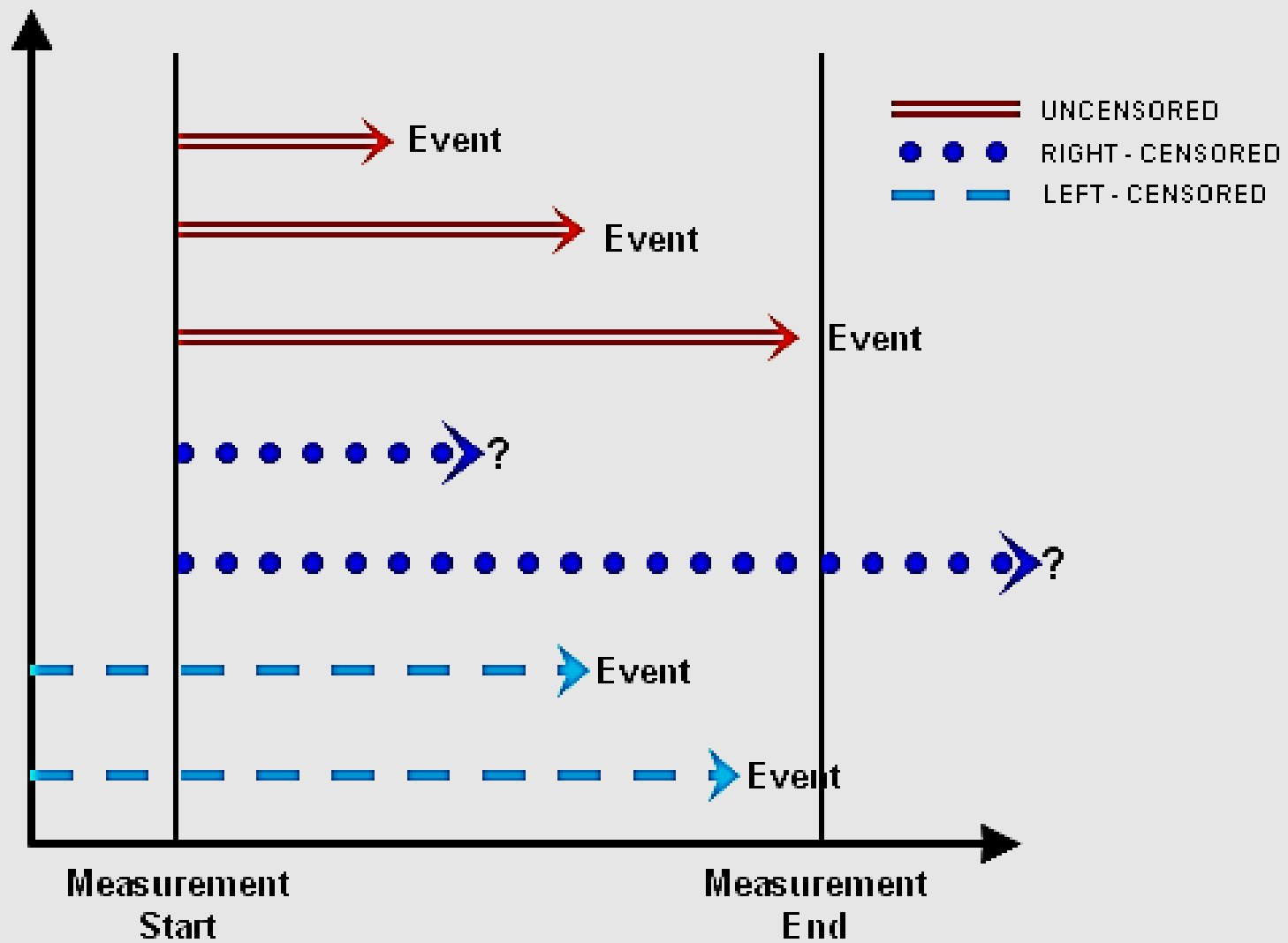


Fig : Examples of Right censoring for churn - prediction



SURVIVAL ANALYSIS MEASUREMENT

- **Survival analysis is a set of statistical techniques focusing on the occurrence and timing of events.**
- **The event time distribution defined as a continuous probability distribution.**

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta T)}{\Delta t}$$

The corresponding cumulative event time distribution is then defined as follows:

$$F(t) = P(T \leq t) = \int_0^t f(u) du$$

Closely related is the survival function:

$$S(t) = 1 - F(t) = P(T > t) = \int_t^{\infty} f(u) du$$

$S(t)$ is a monotonically decreasing function with $S(0) = 1$ and $S(\infty) = 0$.
The following relationships hold:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$



Activate Windows
Go to PC settings to

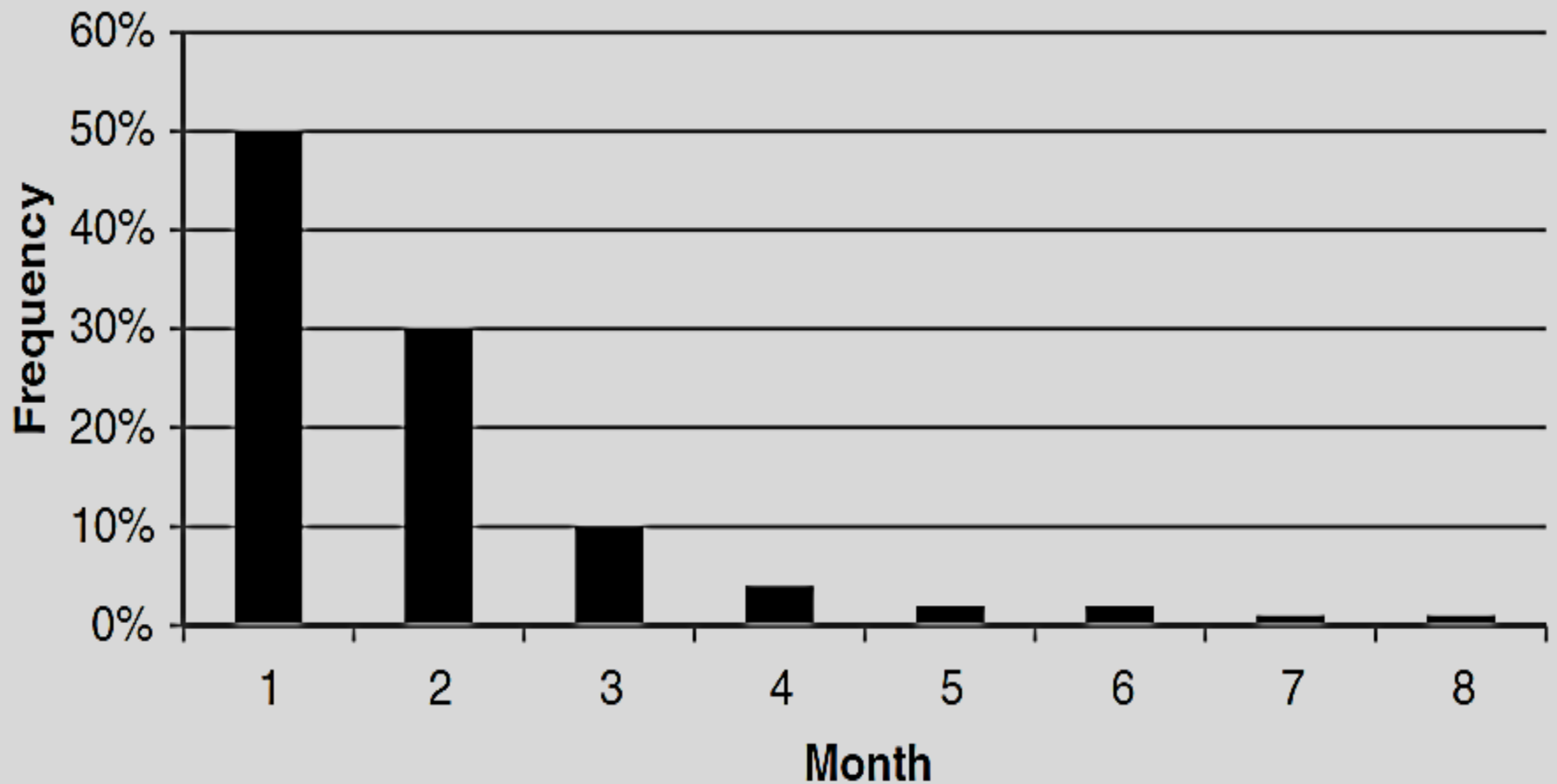


Fig: Example of a Discrete Event Time Distribution

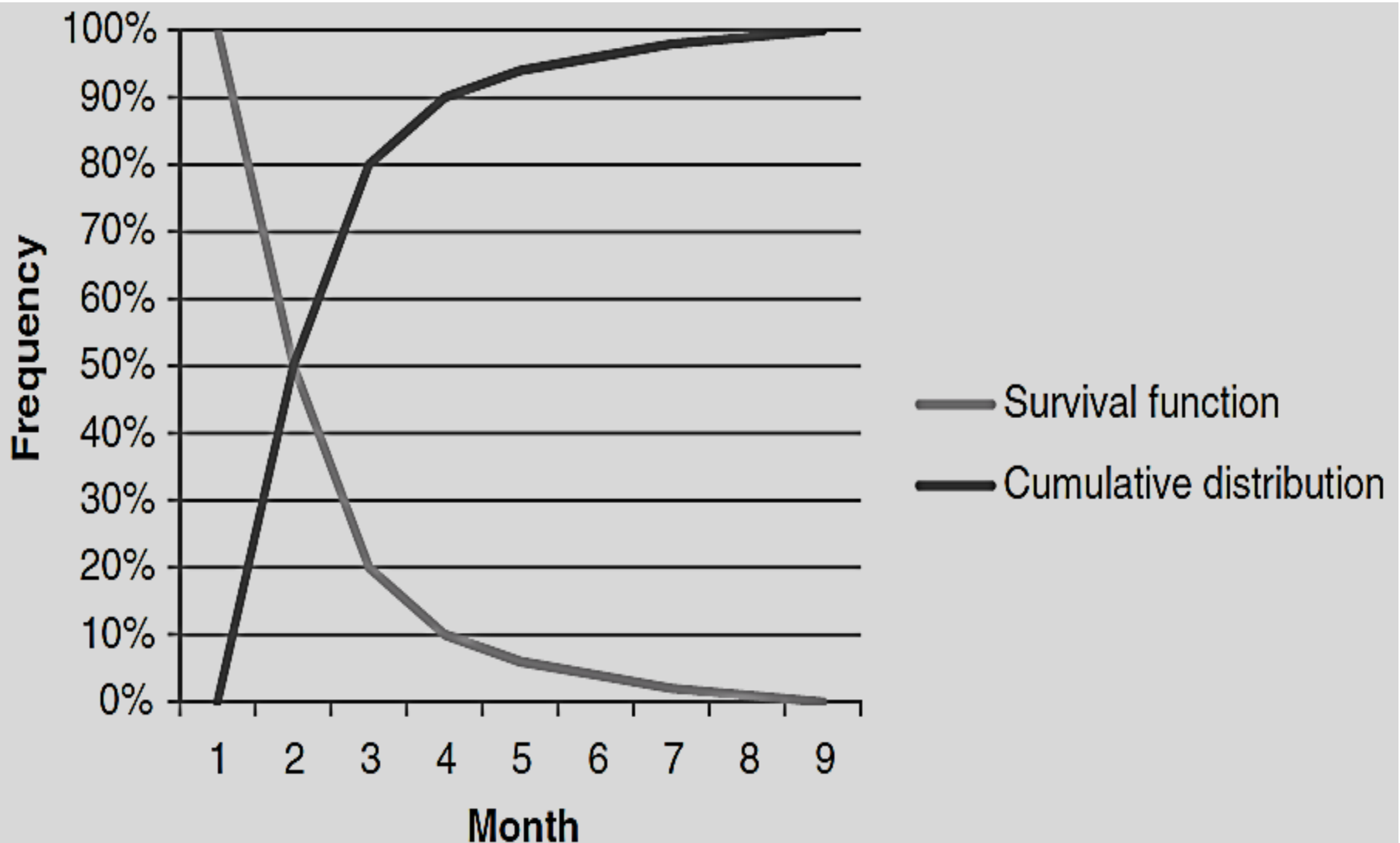


Fig: Cumulative distribution and Survival Function for the Event time Distribution in previous figure

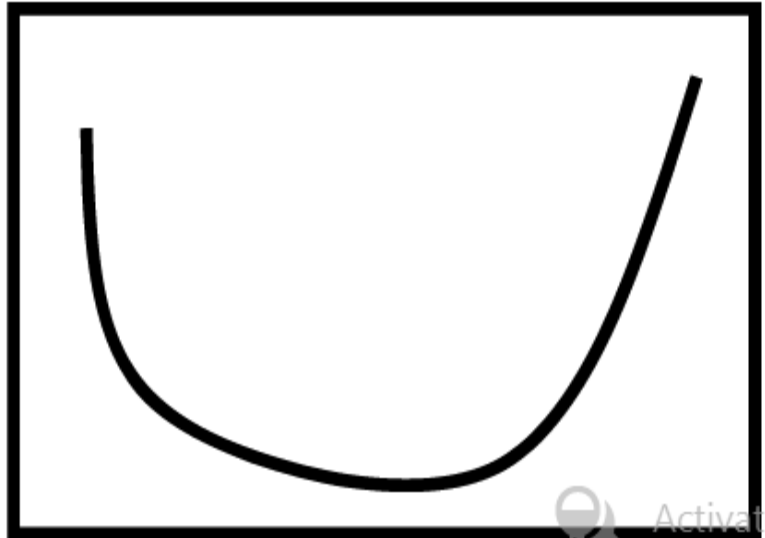
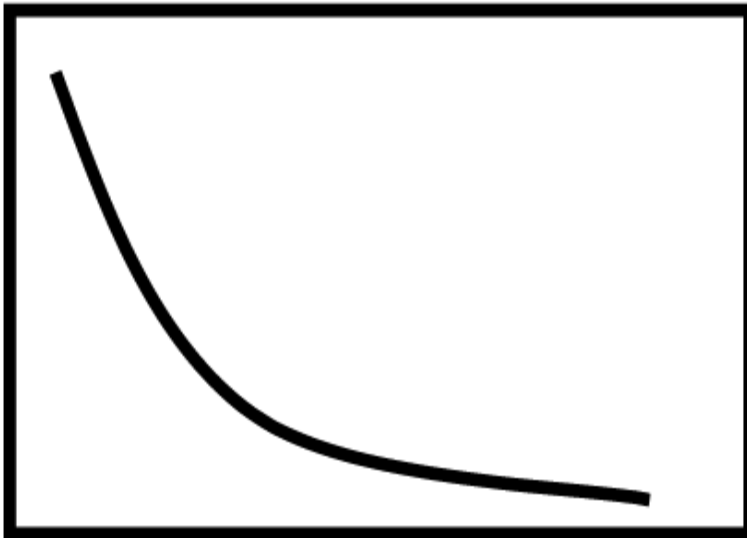
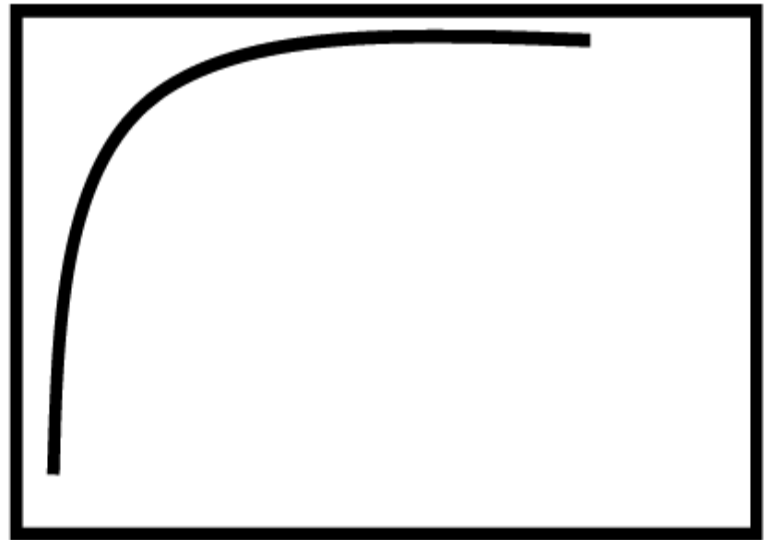
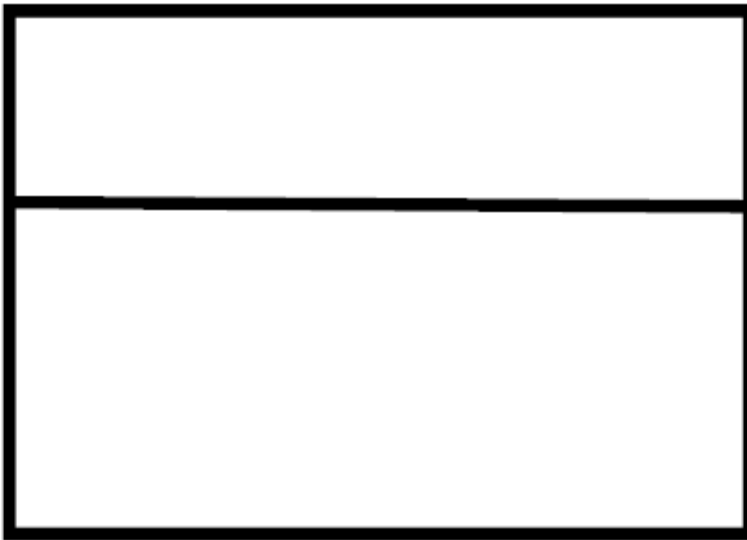


Figure : Example Hazard Shapes

Examples of hazard shapes

- **Constant hazard**, whereby the risk remains the same at all time
- **Increasing hazard**, reflecting an aging effect
- **Decreasing hazard**, reflecting a curing (remedial) effect
- **Convex bathtub shape**, studying human mortality.

The probability density function $f(t)$, survivor function $S(t)$, and the hazard function $h(t)$ are mathematically equivalent ways of describing a continuous probability distribution with the following relationships:

$$h(t) = \frac{f(t)}{S(t)}$$

$$h(t) = -\frac{d \log S(t)}{dt}$$

$$S(t) = \exp \left(-\int_0^t h(u) du \right)$$



PARAMETRIC SURVIVAL ANALYSIS

- **Parametric survival analysis models assume a parametric shape for the event time distribution.**
- **A parametric survival model is a well-recognized statistical technique for exploring the relationship between the survival of a patient,**

A first popular choice

is an exponential distribution, defined as follows:

$$f(t) = \lambda e^{-\lambda t}$$

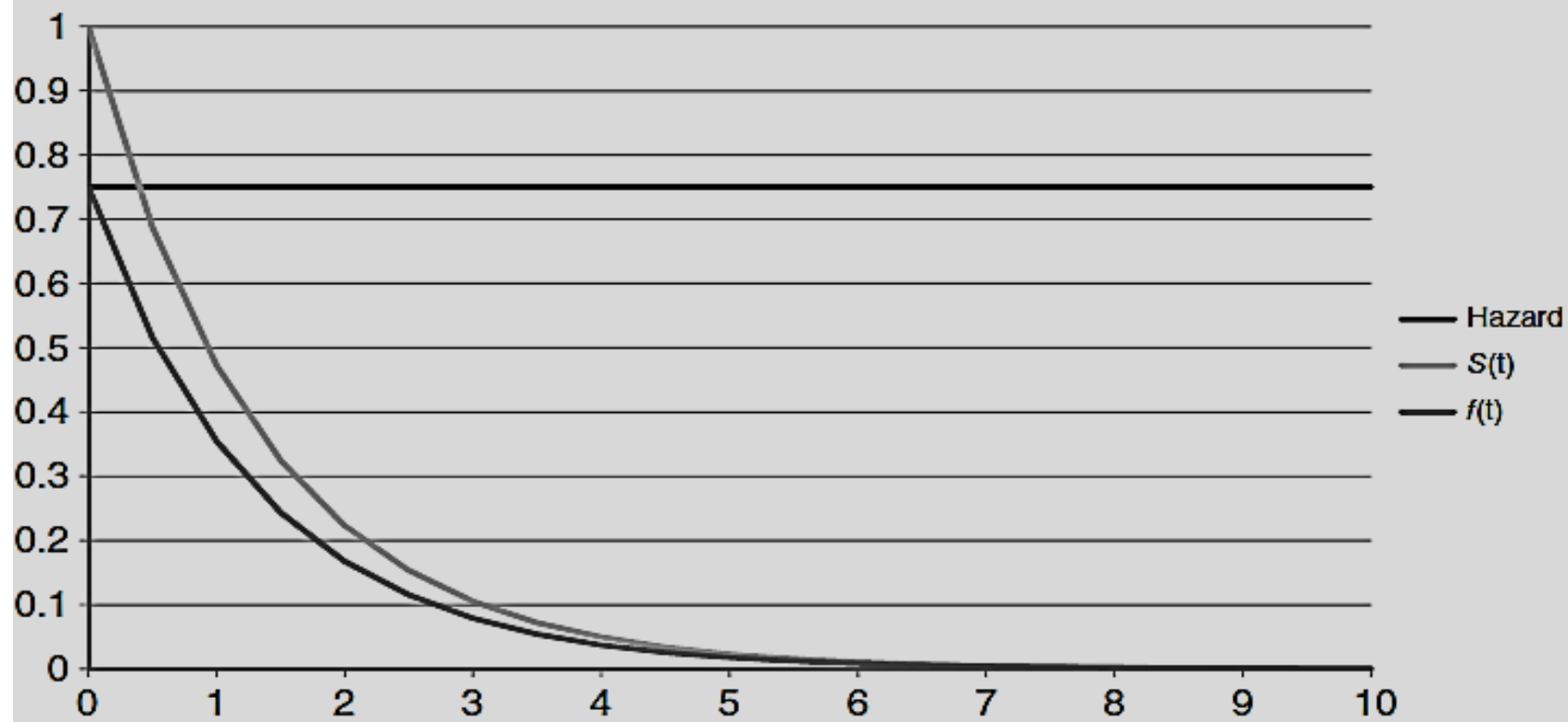
Using the relationships defined earlier, the survival function then becomes:

$$S(t) = e^{-\lambda t}$$

and the hazard rate

$$h(t) = \frac{f(t)}{S(t)} = \lambda$$

$$\log(h(t, x_i)) = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_N x_{iN}$$



Thank You



ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCES

(Autonomous)

G.B. Nagar, Kalavai - 632506



Big data analytics

Unit - IV

SOCIAL NETWORK ANALYTICS

The most popular are:

Facebook

Twitter

Whatsapp

Instagram

Telegram

Google+

LinkedIn

Youtube

Pinterest

Quora

Skype

- **Social networks could be:**

Web pages connected by hyperlinks

Email traffic between people

Research papers connected by citations

Telephone calls between customers.

Banks connected by feel of dependencies

Spread of illness between patients

SOCIAL NETWORK DEFINITIONS

- **A social network consists of both nodes (vertices) and edges.**
- **A node could be defined as customer, household/family, patient, doctor, author, terrorist, web page, etc.,**
- **An edge can be defined as a friend relationship, a call, transmission of disease, reference.**

- **Edges can also be weighted based on interaction frequency, importance of information exchange, intimacy, emotional intensity.**
- **Social networks can be represented as a sociogram.**

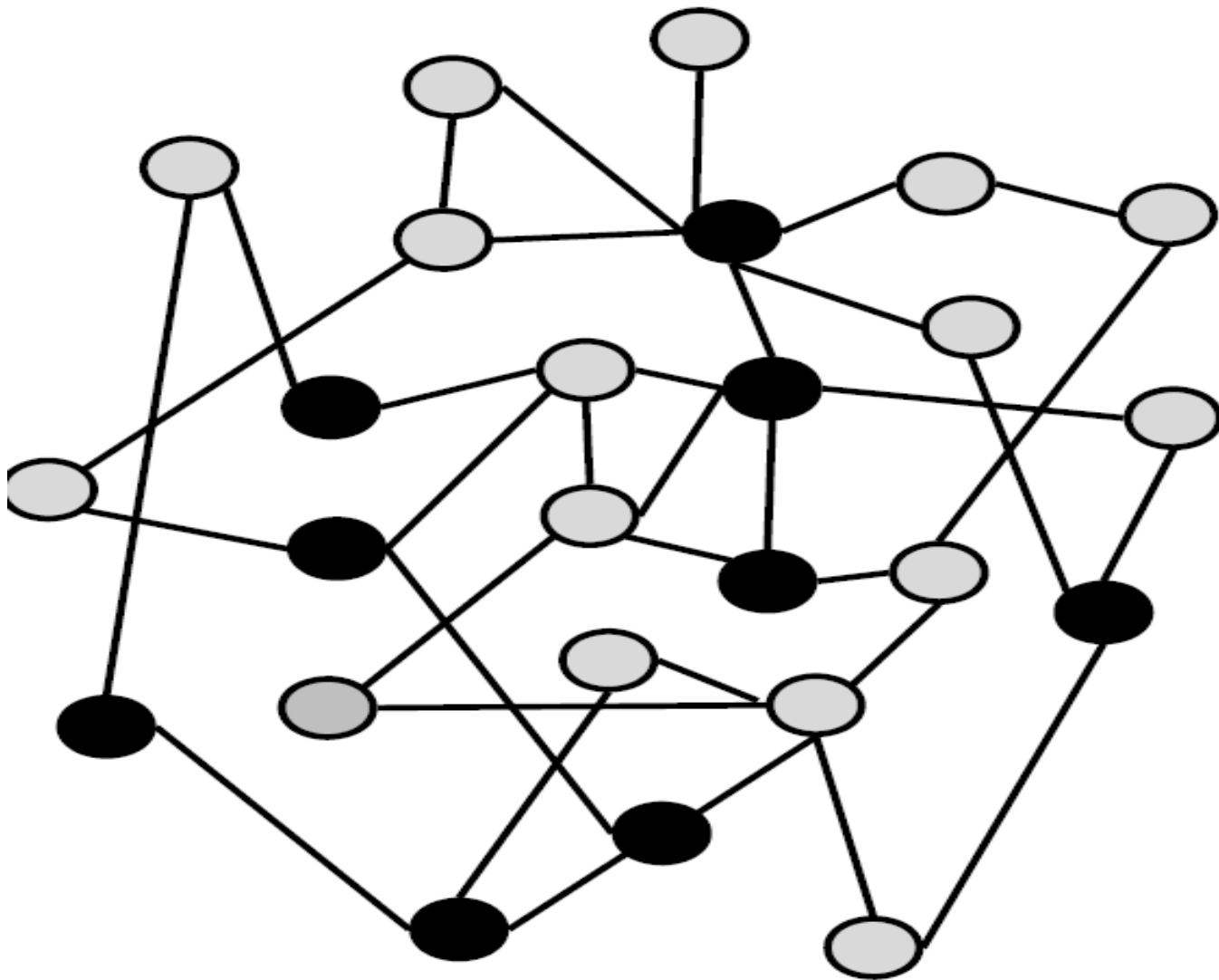


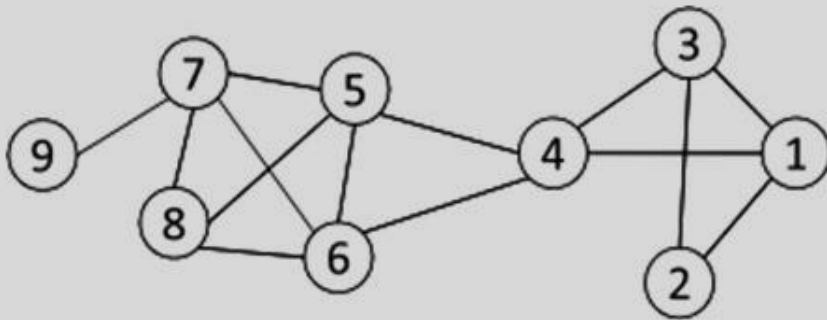
Fig: Example sociogram

- **Sociogram are good for small-scale network**
- **Large-scale networks, represented by matrix.**

Table 6.1 Matrix Representation of a Social Network

	C1	C2	C3	C4
C1	—	1	1	0
C2	1	—	0	1
C3	1	0	—	0
C4	0	1	0	—

- **Graph Representation**



- **Matrix Representation**

Node	1	2	3	4	5	6	7	8	9
1	-	1	1	1	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0
3	1	1	-	1	0	0	0	0	0
4	1	0	1	-	1	1	0	0	0
5	0	0	0	1	-	1	1	1	0
6	0	0	0	1	1	-	1	1	0
7	0	0	0	0	1	1	-	1	1
8	0	0	0	0	1	1	1	-	0
9	0	0	0	0	0	0	1	0	-

SOCIAL NETWORK METRICS

Table 6.2 Network Centrality Measures

Geodesic	Shortest path between two nodes in the network	
Degree	Number of connections of a node (in- versus out-degree if the connections are directed)	
Closeness	The average distance of a node to all other nodes in the network (reciprocal of farness)	$\left[\frac{\sum_{j=1}^g d(n_i n_j)}{g} \right]^{-1}$
Betweenness	Counts the number of times a node or connection lies on the shortest path between any two nodes in the network	$\sum_{j < k} \frac{g_{jk}(n_i)}{g_{jk}}$
Graph theoretic center	The node with the smallest maximum distance to all other nodes in the network	

- Assume a network with 'g' nodes n_i ,
 - ✓ $i = 1, \dots, g$, g_{jk} represents the no.of geodesics from node j to node k,
 - ✓ whereas $g_{jk}(n_i)$ represents the no.of geodesics from node j to node k passing through node n_i
- Metrics can now be illustrated with the well-known Kite network

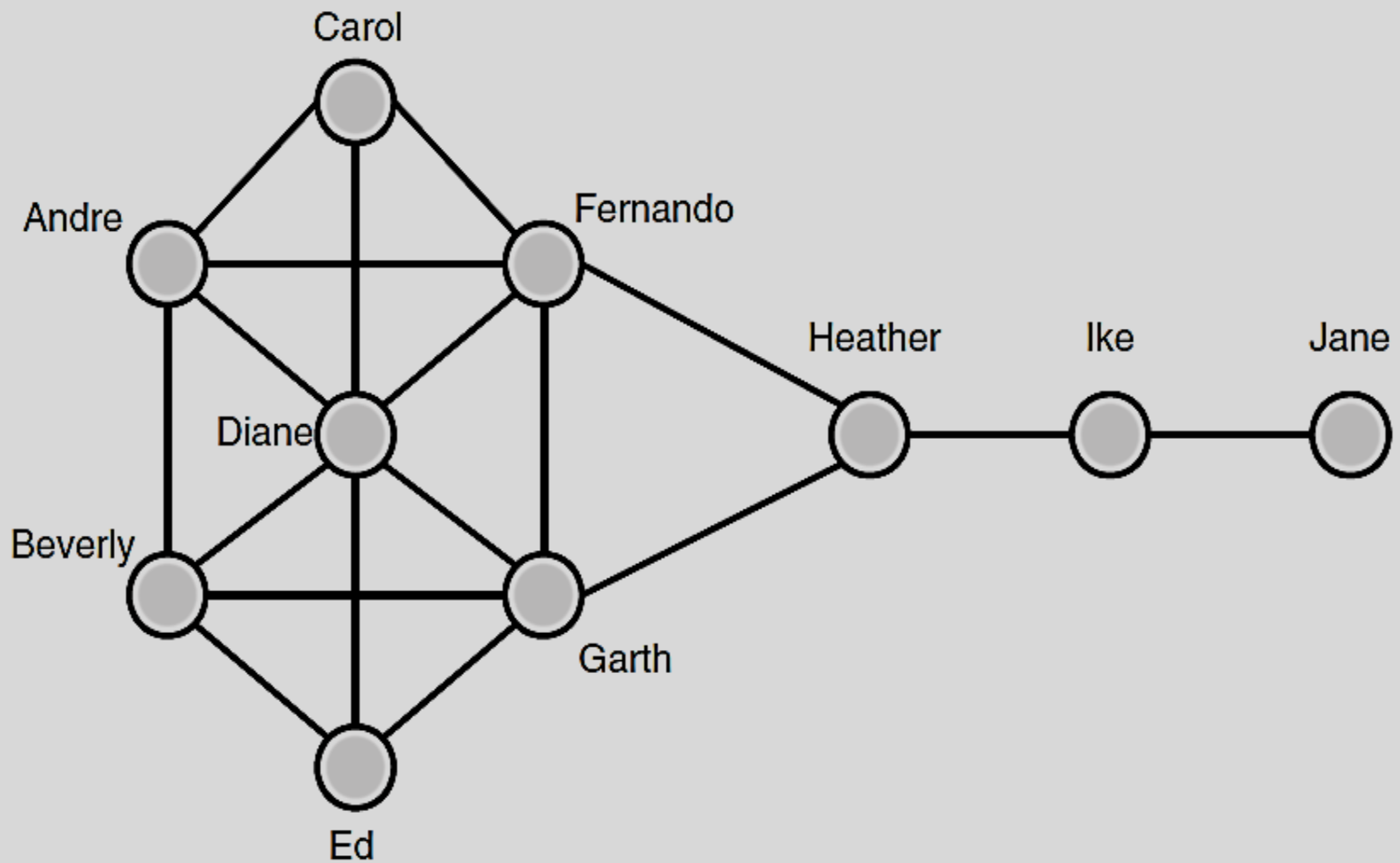


Figure 6.2 The Kite Network

Table 6.3 Centrality Measures for the Kite Network

Degree		Closeness		Betweenness	
6	Diane	0.64	Fernando	14	Heather
5	Fernando	0.64	Garth	8.33	Fernando
5	Garth	0.6	Diane	8.33	Garth
4	Andre	0.6	Heather	8	Ike
4	Beverly	0.53	Andre	3.67	Diane
3	Carol	0.53	Beverly	0.83	Andre
3	Ed	0.5	Carol	0.83	Beverly
3	Heather	0.5	Ed	0	Carol
2	Ike	0.43	Ike	0	Ed
1	Jane	0.31	Jane	0	Jane

A popular technique here is the Girvan-Newman algorithm which works as follows:

- 1. The betweenness of all existing edges in the network is calculated first.**
 - 2. The edge with the highest betweenness is removed**
 - 3. The betweenness of all edges affected by the removal is recalculated**
 - 4. Set 2 and 3 are repeated until no edge remain.**
- The result to provide optimal no.of communities.**

SOCIAL NETWORK LEARNING

- The goal is within-network classification to compute the marginal class memberships probability of a particular node given the other nodes in the network.

Key challenges:

Data are not independent

Identically distributed

An assumption often made in statistical models

A social network learner will usually consist of the following components:

- A local model**
- A network model**
- A collective inferencing procedure**

A local model

This is a model using only node-specific characteristics (e.g : decision tree)

A network model

This is a model that will make use of the connections in the network to do the inferencing

A collective inferencing procedure

To determine how the known on nodes are estimated together

RELATIONL NEIGHBOR CLASSIFIER

- **Use of the assumption, which states that connected nodes have a tendency to belongs to the same class.**
- **The posterior class probability for node ‘n’ to belong to class ‘c’ is then calculate as follows:**

$$P(c|n) = \frac{1}{Z} \sum_{\{n_j \in \text{Neighborhood}_n | \text{class}(n_j)=c\}} w(n, n_j)$$

whereby Neighborhood_n represents the neighborhood of node n , $w(n, n_j)$ the weight of the connection between n and n_j , and Z is a normalization factor to make sure all probabilities sum to one.

For example, consider the network depicted in Figure 6.3, whereby C and NC represent churner and nonchurner nodes, respectively.

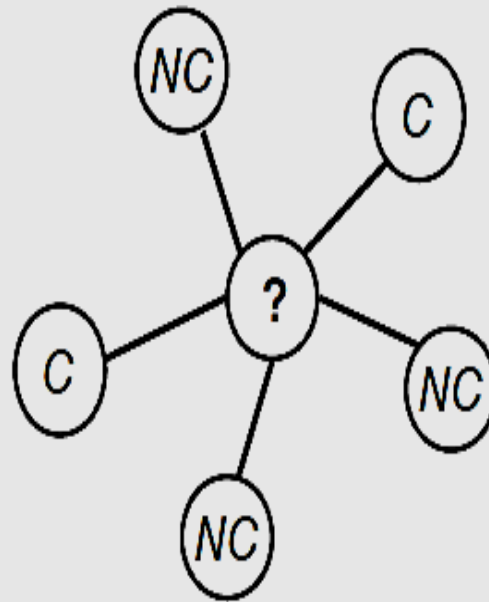


Figure 6.3 Example Social Network for Relational Neighbor Classifier

The calculations then become:

$$P(C|?) = 1/Z(1+1)$$

$$P(NC|?) = 1/Z(1+1+1)$$

Since both probabilities have to sum to 1, Z equals 5, so the probabilities become:

$$P(C|?) = 2/5$$

$$P(NC|?) = 3/5$$

PROBABILISTIC RELATIONAL NEIGHBOR CLASSIFIER

- **A straightforward extension of the relational neighbor classifier, Node 'n', Class 'C'**

$$P(c|n) = \frac{1}{Z} \sum_{\{n_j \in \text{Neighborhood}_n\}} w(n, n_j) P(c|n_j)$$

Note that the summation now ranges over the entire neighborhood of nodes. The probabilities $P(c|n_j)$ can be the result of a local model or of a previously applied network model. Consider the network of Figure 6.4.

The calculations then become:

$$P(C|?) = 1/Z(0.25 + 0.80 + 0.10 + 0.20 + 0.90) = 2.25/Z$$

$$P(NC|?) = 1/Z(0.75 + 0.20 + 0.90 + 0.80 + 0.10) = 2.75/Z$$

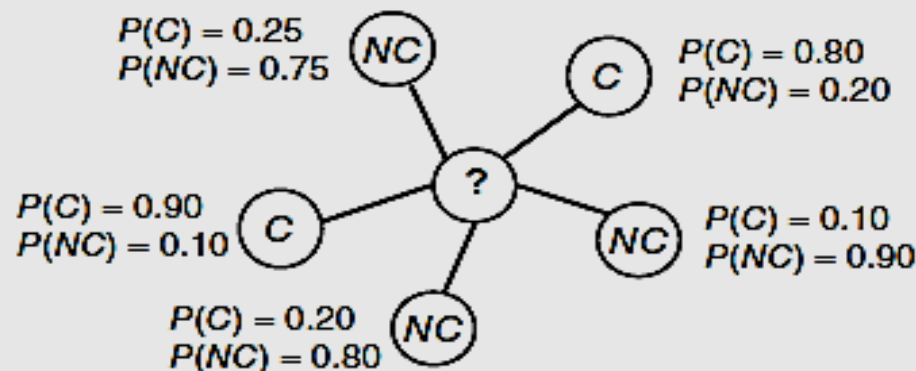


Figure 6.4 Example Social Network for Probabilistic Relational Neighbor Classifier

Since both probabilities have to sum to 1, Z equals 5, so the probabilities become:

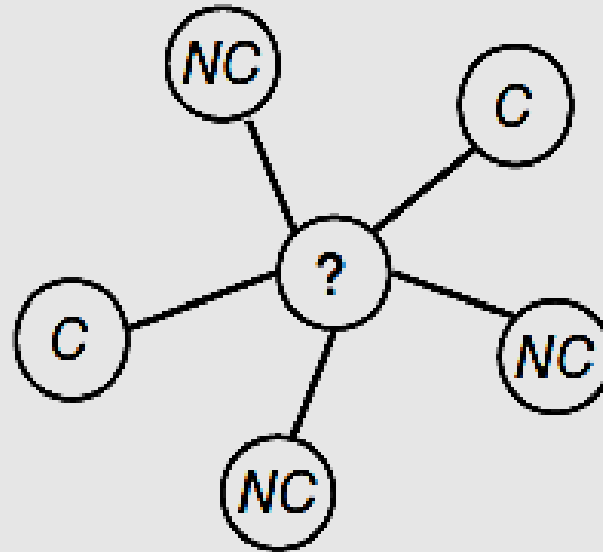
$$P(C|?) = 2.25/5 = 0.45$$

$$P(NC|?) = 2.75/5 = 0.55$$

RELATIONAL LOGISTIC REGRESSION

Basically starts off from a data set with local node-specific characteristics and adds network characteristics to it, as follows:

- **Most frequently occurring class of neighbor (mode-link)**
- **Frequency of the classes of the neighbors (count-link)**
- **Binary indicators indicating class presence (binary-link)**



CID	Age	Income	...	Mode link	Frequency no churn	Frequency churn	Binary no churn	Binary churn
Bart	33	1,000		NC	3	2	1	1

Figure 6.5 Relational Logistic Regression

Local variables				Network variables		
Customer	Age	Recency	Number of contacts	Contacts with churners	Contacts with contacts of churners	Churn
John	35	5	18	3	9	Yes
Sophie	18	10	7	1	6	No
Victor	38	28	11	1	5	No
Laura	44	12	9	0	7	Yes

First-order network variable Second-order network variable

Figure 6.6 Example of Featurization with Features Describing Target Behavior of Neighbors

Customer	Age	Average duration	Average revenue	Promotions	Average age friends	Average duration friends	Average revenue friends	Promotions friends	Churn
John	25	50	123	X	20	55	250	X	Yes
Sophie	35	65	55	Y	18	44	66	Y	No
Victor	50	12	85	None	50	33	50	X, Y	No
Laura	18	66	230	X	65	55	189	X	No

Figure 6.7 Example of Featurization with Features Describing Local Node Behavior of Neighbors

COLLECTIVE INFERENCE

Collective inference procedure infers a set of class labels/probabilities for the unknown nodes by taking into account.

Collective inference procedures are:

- **Gibbs sampling**
- **Iterative classification**
- **Relaxation labeling**
- **Loopy belief propagation**

Gibbs sampling :

- **A sequence of observations which are approximated from a specified multivariate probability distribution**

Iterative classification:

- **The hypothesis underlying this approach is that if two objects are related, inferring something about one object can assist inferences about the other.**
- **We call this approach iterative classification.**

Relaxation labeling :

- **Relaxation labeling is an image treatment methodology.**
- **Its goal is to associate a label to the pixels of a given image or nodes of a given graph**

Loopy belief propagation:

a graph containing loops, despite the fact that the presence of loops does not guarantee convergence.

Thank You...



ADHIPARASAKTHI COLLEGE OF ARTS AND SCIENCES

(Autonomous)

G.B. Nagar, Kalavai - 632506



Big data analytics

Unit - V

BENCHMARKING

- **To compare the output and performance of the analytical model with a reference model or benchmark**
- **To make sure that the current analytical is the optimal one to be used.**
- **Type**

External benchmark

Internal benchmark

Expert-based benchmark

- **Popular agreement statistics for benchmarking are:**
 - **Spearman's rank order correlation**
(ρ - rho to calculate density)
 - **Kendall's**
(τ - tau to calculated by dividing any circle's circumference by its radius.)
 - **Goodman-Kruskal (γ)**

Spearman's rank order correlation (ρ)

- **Measures the degree to which a monotonic relationship between internal score and benchmark**
- **It starts by assigning 1 to the lowest score, 2 to the second lowest score.**
- **Computed as follows:**

$$\rho_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

whereby n is the number of observations and d_i the difference between the scores. Spearman's rank order correlation always ranges between -1 (perfect disagreement) and $+1$ (perfect agreement).

Kendall's (τ)

- **Calculating the concordant and discordant pairs of observations.**
- **Two observations are said to be concordant if the observation that has a higher score assigned by the internal model also has a higher score assigned by the external model.**
- **Calculated as follows:**

$$\tau = \frac{A - B}{\frac{1}{2}n(n-1)},$$

whereby n is the number of observations, A the number of concordant pairs, and B the number of discordant pairs. Note that the denominator gives all possible pairs for n observations. Kendall's τ is 1 for perfect agreement and -1 for perfect disagreement.

- Kendall's basically looks all possible pairs of observations
- Goodman – Kruskal will only consider the united pairs (i.e either concordant or discordant), as follows:

$$\gamma = \frac{A - B}{A + B}$$

The Goodman-Kruskal γ is +1 if there are no discordant pairs (perfect agreement), -1 if there are no concordant pairs (perfect disagreement), and 0 if there are equal numbers of concordant and discordant pairs.

Table 7.9 Example for Calculating Agreement Statistics

Customer	Internal Credit Score	FICO	Rank Internal Score	Rank External Score	d_i
1	20	680	2.5	3	0.25
2	35	580	5	1	16
3	15	640	1	2	1
4	25	720	4	5	1
5	20	700	2.5	4	2.25
				$\sum_{i=1}^n d_i^2$	20.5

For example, consider the example in Table 7.9.

Spearman's rank order correlation then becomes -0.025 . The concordant pairs are as follows: C1,C3; C1,C4; C3,C4; C3,C5; and C4,C5. The discordant pairs are: C1,C2; C2,C3; C3,C4; and C2,C5. The pair C1,C5 is a tie. Kendall's τ thus becomes: $(5 - 4)/10$ or 0.1 and the Goodman-Kruskal γ becomes $(5 - 4)/(5 + 4)$ or 0.11 .

DATA QUALITY

- **Corporate information system consist of many databases linked by real-time and batch data feeds.**
- **The DBs are continuously updated, as are the applications performing data exchange.**
- **High-quality of data in combination with good technology gives added value, whereas poor quality of data with good technology is a big problem.**

- **Poor DQ impacts**

Impact on customer satisfaction

Increases operational expenses

Will lead to lowered employee job satisfaction

- **DQ problems is increase in the size of databases.**

Table 7.10 Data Quality Dimensions

Category	Dimension	Definition: The Extent to Which . . .
Intrinsic	Accuracy	Data are regarded as correct
	Believability	Data are accepted or regarded as true, real, and credible
	Objectivity	Data are unbiased and impartial
	Reputation	Data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	Data are beneficial and provide advantages for their use
	Completeness	Data values are present
	Relevancy	Data are applicable and useful for the task at hand
	Appropriate amount of data	The quantity or volume of available data is appropriate

Representational	Interpretability	Data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	Data are clear without ambiguity and easily comprehended
Accessibility	Accessibility	Data are available or easily and quickly retrieved
	Security	Access to data can be restricted and hence kept secure

- **Accuracy indicates whether the data stored are the correct values.**
- **Example :**
 - **A person birthday is February 27, 1975, for a DB that expects date in USA format, 02/27/1975 is the correct value.**
 - **However, for a DB that expects a INDIAN, the date 02/27/1975 is incorrect; instead of 27/02/1975 is the correct value.**

- **Completeness verifies whether a column of a table has missing values or not.**

Table 7.11 Population Completeness

ID	Name	Surname	Birth Date	Email
1	Monica	Smith	04/10/1978	smith@abc.it
2	Yuki	Tusnoda	04/03/1968	Null ^a
3	Rose	David	02/01/1937	Null ^b
4	John	Edward	14/12/1955	Null ^c

^aNot existing

^bExisting but unknown

^cNot known if existing

- **Tuple 2 : Since the person represented by tuple 2 has no email address, we can say that tuple is incomplete**
- **Tuple 3 : Since the person represented by tuple 3 has an email, but its value is not known, we can say that the tuple is incomplete**
- **Tuple 4: If we do know the person represented by email or not, incompleteness may not be the case.**

- **Believability :**

Which data is regarded as true and creditable

- **Accessibility :**

Refers to how easy the data can be located and retrieved.

- **Consistency (Dependability):**

Consider from various viewpoints

Presence of redundant data (name, address)

City name & zip code

Gender can be encoded male/female, M/F, 0/1

- **Timelines dimension reflects how up-to-date data is with respect to the task for which it is used.**

DQ Problem causes such as :

Multiple data sources :

**The same data may produced duplicates;
a consistency problem**

Subjective Judgment :

Can create data bias; objectivity problem

- **Limited computing facilities:**

Lack of sufficient calculating facilities limits data access; accessibility problem.

- **Size of data :**

Big data can give high response times; accessibility problem

Data quality can be improved through a total data quality management program.

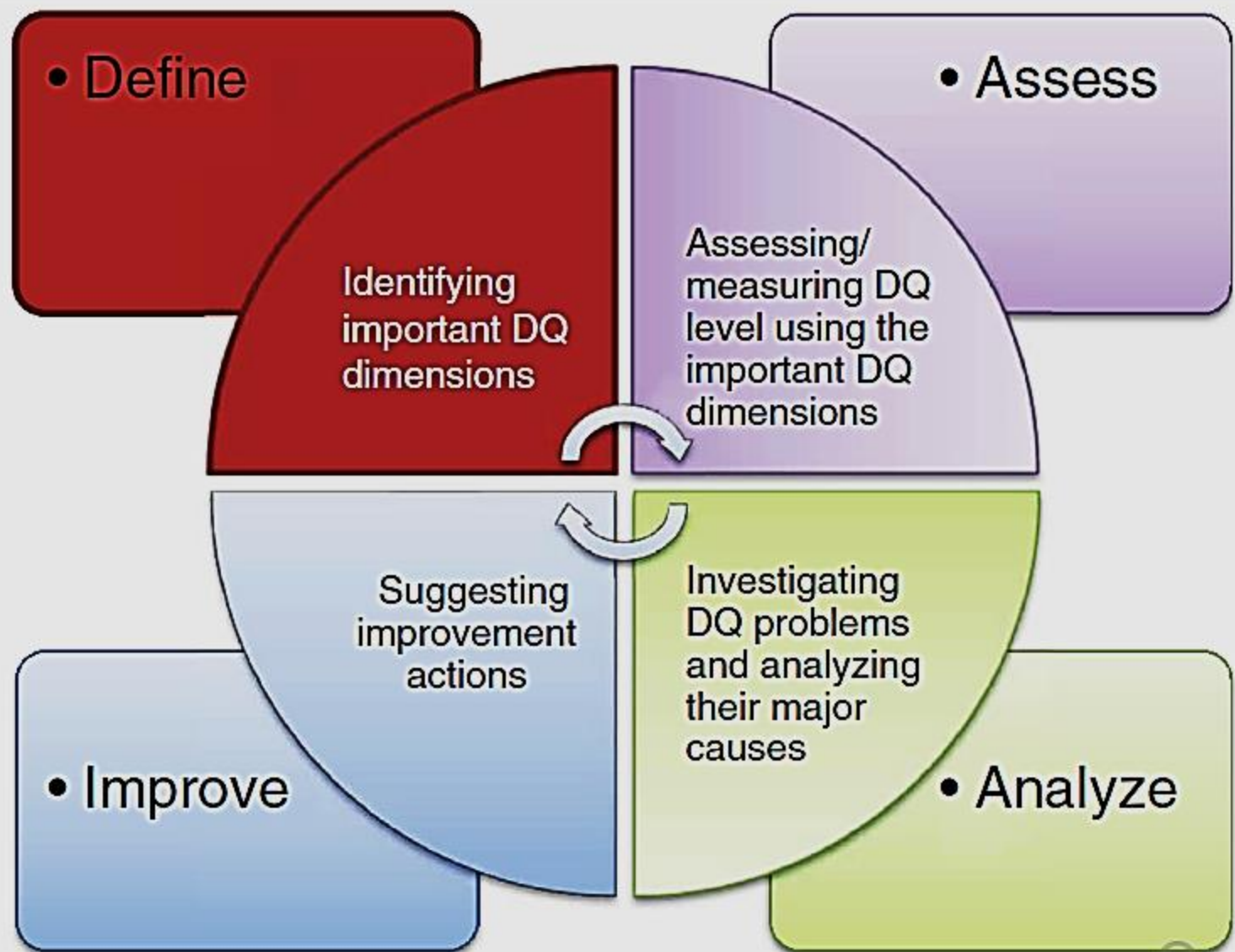


Figure 7.4 Data Quality Management Program

SOFTWARE

- **Different types of software can be used for doing analytics.**
- **Open source s/w and commercial s/w**
- **Open source s/w**

RapidMiner

R

Weka

- **Commercial s/w**

SAS (Statistical Analytical System)

SPSS (Statistical Package for the Social Science)

Matlab

Microsoft – Excel

- **Provide analytical solutions.**
- **Full coverage of the whole range of analytical activities**
- **(Ref. Table 7.12 page 154 in the text book)**

- **The table 7.12 concluded that RapidMiner and R, two open source s/w solutions are the most popular tools for analytics.**
- **Microsoft Excel is still quite popular for doing analytics.**

PRIVACY

Privacy issues can arise by

- **Data about individuals can be collected without these individuals being aware of it.**
- **People may be aware that data is collected about them, but have no say in how the data is being used.**

Privacy as compared to simple data collection and data retrieval from DBs.

- **Data analytics entails the use of massive amounts of data, from several sources.**
- **Information can be used for criminal activities – such as**

Stalking (Irritation)

Kidnapping

Identity theft

Phishing (fraudulent attempt)

scams

Direct marketing by legitimate (genuine)

- **Guidelines on the protection of Privacy and Transborder Flows of Personal Data.**
- **The basic principles are defined to safeguard privacy :**

- 1. Collection limitation principle:**

Data collection should be done lawfully and with knowledge and consent of the data object.

2. Data quality principle :

The data should be relevant for the purpose it is collected for, accurate, complete and up-to-date

3. Purpose specification :

The purpose of the data should be specified before data collection and the use should be limited to these purposes.

4. Use limited principle:

The data should not be used for other purpose than specified.

5. Safety safeguards principles:

The data should be protected against risks of loss, unauthorized access, use, modification or release of data

6. Openness principles:

There should be a policy of honesty about the developments, practices, and policies with respect to personal data.

7. Individual participation principle:

Confirmation whether data exists about him or her, to receive the data

8. Accountability principle:

A data controller can be held accountable for compliance

Zip Code	Age	Gender
83661	26	M
83659	23	M
83645	58	F



Zip Code	Age	Gender
836**	2*	M
836**	2*	M
836**	5*	F

Figure 7.5 Example of Generalization and Suppression to Anonymize Data

Method 1 : Generalization and suppression

Remove information from the quasi-identifiers, until the records are not individually identifiable.

Method 2 : Perturbation (Uneasiness)

Change the data by adding noise, swapping values and etc

MODLE DESIGN AND DOCUMENTATION

- **When was the model designed, and by who?**
- **What is the perimeter of the model?**
(e.g counterparty types, geographical region,
industry sectors)
- **What are the strength and weakness of the model?**

- **What data were used to build the model? How was constructed? What is the time horizon of the sample?**
- **Is human judgment used, and how?**
- **All of this appropriately documented.**
- **Documentation should be transparent and complete.**
- **To keep track of the different version of the documents.**
- **Documentation test, verifies, continue development**

CORPORATE GOVERNANCE

- **Ownership of the analytical models is clearly claimed.**
- **To develop model boards, in terms of their functioning, interpretation and follow-up.**
- **Board of directors and senior management are involved in the implementation and monitoring processes.**
- **Responsible for sound governance of the analytical models.**

- **They should demonstrate active involvement on an ongoing basis, assign clear responsibilities.**
- **Put into place organizational procedures and policies that allow the proper and sound implementation and monitoring.**
- **The outcome of the monitoring and back-testing exercise must be communicated to senior management.**
- **Add a Chief Analytical Officer (CAO) to the board, model development, implementation, monitoring**

WEB ANALYTICS

The measurement, collection, analysis and reporting of internet data for the purpose of understanding and improving web usage.

Web Data Collection

- **Web analytics is to collect data about web visits.**
- **Web server's logging functionality.**

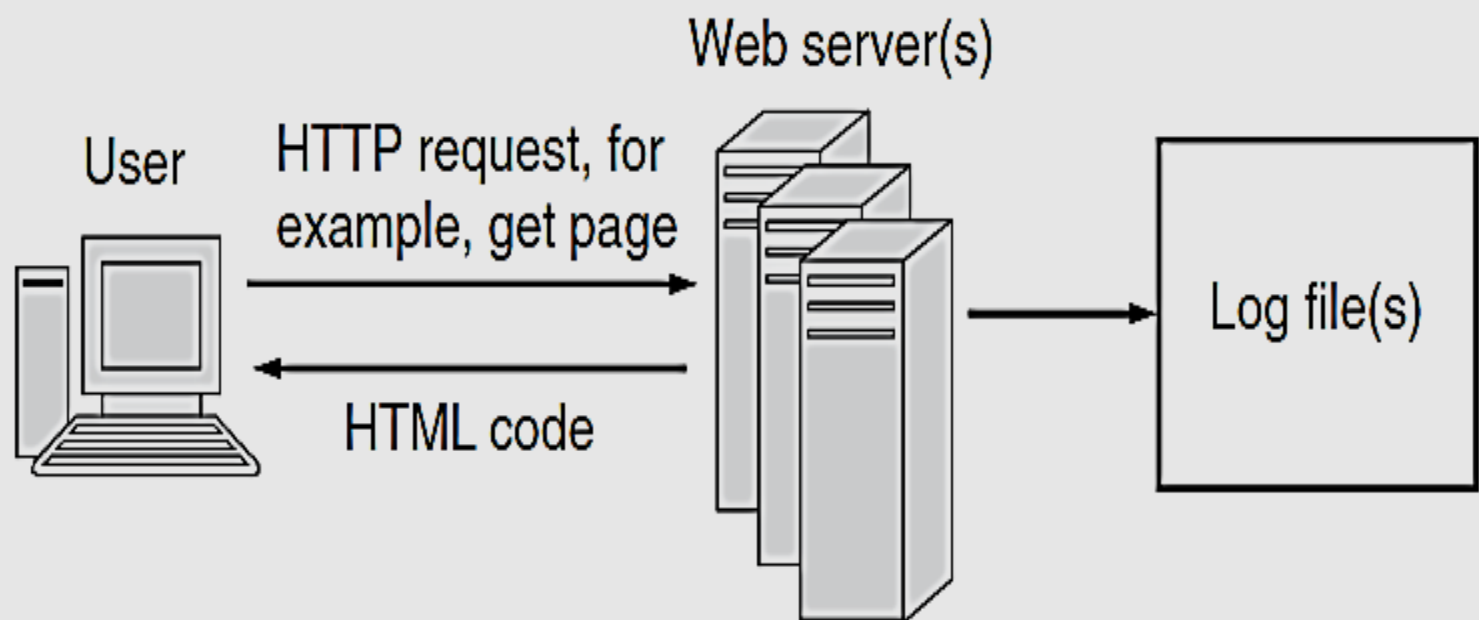


Figure 8.11 Web Server Log Analysis

The data recorded includes:

- **Remote host : IP address or domain name**
- **Remote log name : User name**
- **Date and time**
- **HTTP request method (GET or POST)**
- **Resource requested (Web server)**
- **HTTP status code**

200 range : successful

300 range : redirect

400 range : Client error (404 means not found)

500 range : server error

- **Number of bytes transferred**
- **Referrer : webpage from which user clicked on
link to arrive here**
- **Browser and platform**

- **Cookies can also be used for data collection.**
- **A web server can send to a visitor's web browser**
- **The browser can store on the user's data**
- **Cookies can be set and read by client-side, server-side.**
- **Cookies are used:**

Implementing virtual shopping

Remembering user details

Gathering accurate visitors information

Banner and tracking

- **Another data collection mechanism in web analytics is page tagging.**

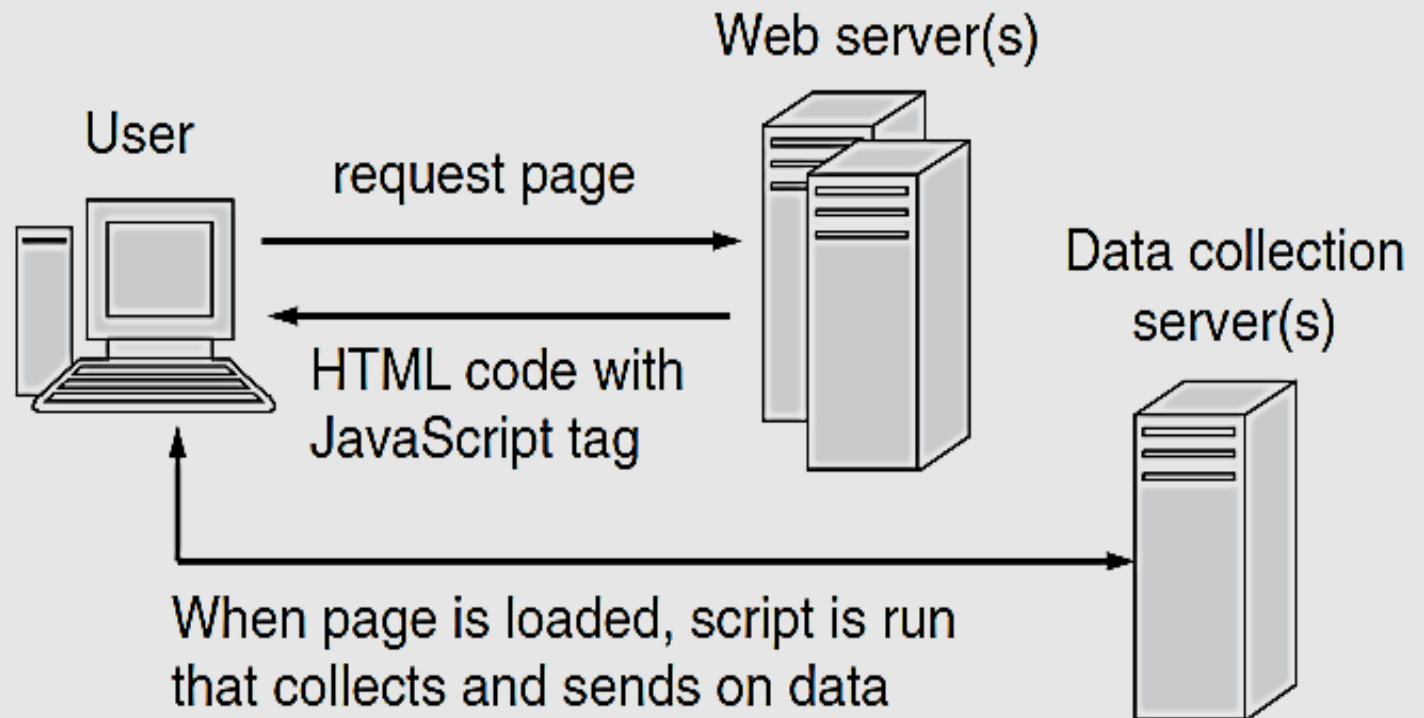


Figure 8.13 Page Tagging

Web KPIs (Key Performance Indicators)

- **Once the data has been collected, it can be analyzed and summarized.**
- **Page views are the no.of times a page was viewed.**
- **Identifying and counting visits or sessions.**
- **Page views per visit**
- **Time on page**
- **Time on site (visit duration)**

Bounce rate:

- **Bounce rate of the site:**

Ratio of single page view visits

- **Bounce rate of a specific page:**

Single page view visits of that page over no.of visits where that page was the entry page

Measures:

- **Most viewed pages**
- **Top entry pages**

Measures:

- **Most viewed pages**
- **Top entry pages**
- **Top exit pages**
- **Top destinations (exit links)**

The conversion rate is then defined as percentage of visits or of unique visitors for which we observed the action.

order received

lead collected

newsletter sign up

sales price

revenue

ROI

Checkout process:

- **Cart rejection rate =**

$$1 - \frac{\text{no.of people who start checkout}}{\text{total Add to Cart clicks}}$$

- **Checkout reject rate =**

$$1 - \frac{\text{no.of people who complete checkout}}{\text{no.of people who start checkout}}$$

Turning Web KPIs into Actionable Insights

- Compared in time to see whether there are any significant changes.
- To verify whether there is an upward/down-ward trend, or any seasonality or daily/ weekly/ monthly patterns to observe

<u>KPI</u>	<u>This week</u>	<u>Last week</u>	<u>Percent change</u>
Conversion rate	1.6%	2.0%	-20% ▼
...			

Figure 8.15 Monitoring the Conversion Rate

One could segment bounce/ conversion rates by:

- **Top five referrers**
- **Search traffic or not**
- **Geographic region**
- **Acquisition (Gaining) strategy**
- **Site search usage**

How much is the search function used?

What keywords are used most?

- **Site search quality**

Calculate bounce rate for site search (% search exits)

Navigation Analysis

- **To understand how users navigate (cross) through the website.**
- **Path analysis gives awareness into frequent navigation patterns.**
- **It analyzes, from a given page, which other pages a group of users visit next in 'x' percent of the times.**

Search Engine Marketing analytics

- **Used to measure the efficiency of search engine marketing.**

- **Types**

Search engine optimization (SEO)

Pay per click (PPC)

SEO

- **To improve organic (living) search results in search engine without paying for it (like Google, Yahoo)**

SEO efforts:

- **Inclusion ratio =**
no.of pages indexed/no.of pages on your website
- **Robot/crawl statistics report.**
- **Track inbound links**
- **Google webmaster tools that show, for the most popular search keywords.**
- **Track the ranking for your top keywords**
- **See whether keywords link to your most important pages.**

PPC

- **One pays a search engine for a link to the website to appear in the search results.**

PPC efforts :

- **Report that differentiate bid terms Vs search terms.**
- **Analysis additional data obtained about ad impressions, clicks, cost.**
- **Keywords position reports**

Example Applications

- **Credit Risk Modeling**
- **Fraud Detection**
- **Recommender System**
- **Web Analytics**

(Kindly refer the book page 161 to 194)

Thank You...