Introduction to Data Analytics
MCA-3282
Open Elective – 6th sem B.Tech
Topic - Grouping

Rohini R. Rao
Dept of Computer Applications
Jan 2020
(Slide set 3 out of 5)

Contents

Steps in data analysis projects

Problem Definition

- Objectives
- Deliverables
- Roles and responsibilities
- Current situation
- Timeline
- Costs and benefits

Data Preparation

- Pull together data table
- Categorize the data
- Clean the data
- Remove unnecessary data
- Transform the data
- Partition the data

Implementation of the Analysis

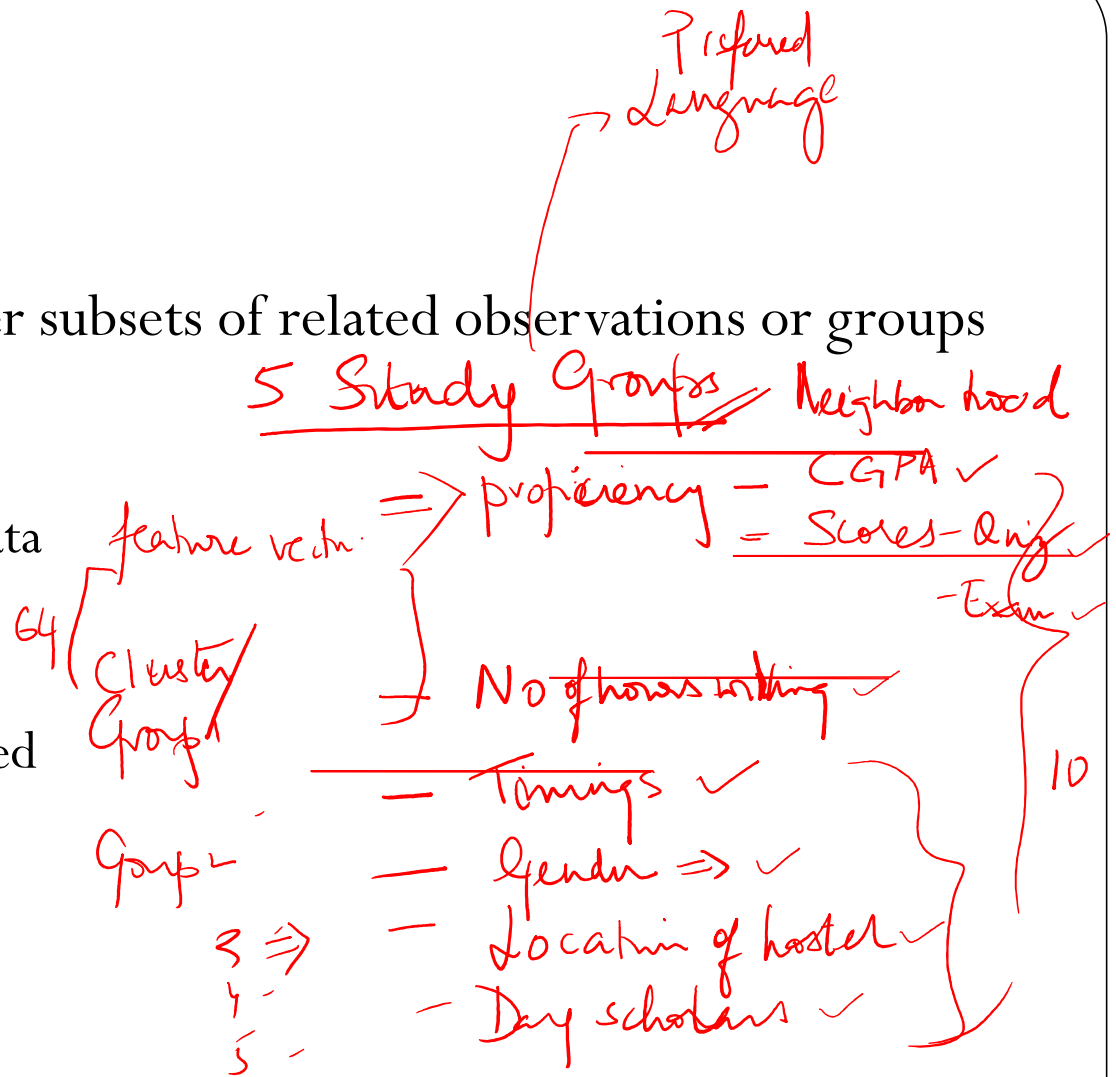
- Summarizing the data
- Finding hidden relationships
- Making prediction

Deployment

- Plan and execute deployment based on the definition in step 1
- Measure and monitor performance
- Review the project

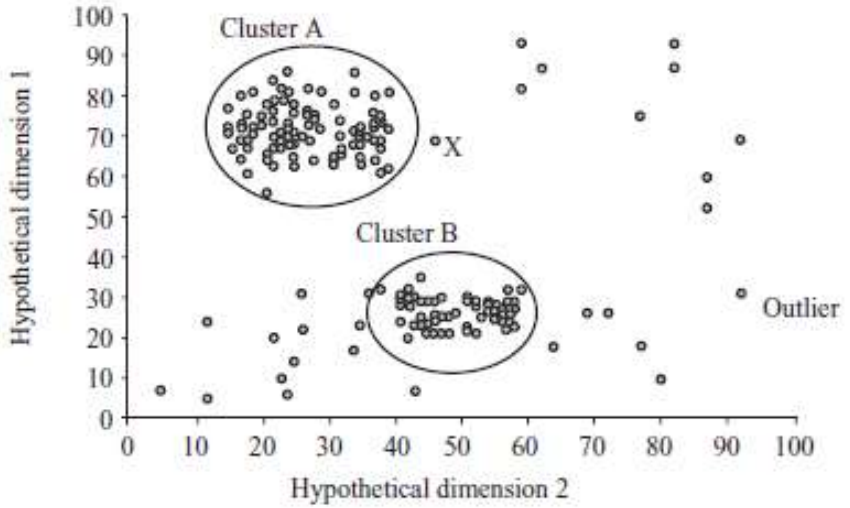
Grouping

- Dividing a data set into smaller subsets of related observations or groups
- **Reasons include**
  - Finding hidden relationships
  - Becoming familiar with the data
  - Segmentation
- **Grouping Approaches**
  - Supervised versus unsupervised
  - Type of variables
  - Data set size limit
  - Interpretable and actionable
  - Overlapping groups
- **Simplest grouping method – Grouping by values or range**
  - Query 1: All cars where Horsepower is greater than or equal to 160 AND Weight is greater than or equal to 4000.



Clustering

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- Unsupervised learning
- Clustering Methods include
  - **Partitioning**
  - **Hierarchical**
  - Others - Density based, grid based, model based etc.



Categorization of Major Clustering Methods

- **Partitioning methods**
  - Given D , a data set of n objects and k, the number of clusters to form
  - Partition algorithm organizes the objects into k partitions where each partition represents a cluster.
  - Centroid Based Technique
    - k_Means Method
    - k-Medoids Method
  - Popular techniques include PAM , CLARA , CLARANS
- **Hierarchical methods**
  - Creates hierarchical decomposition of the given set of data objects
  - Agglomerative or bottom-up approach
  - Divisive or top-down approach
  - Popular techniques include CHAMELEON and BIRCH

Categorization of Major Clustering Methods

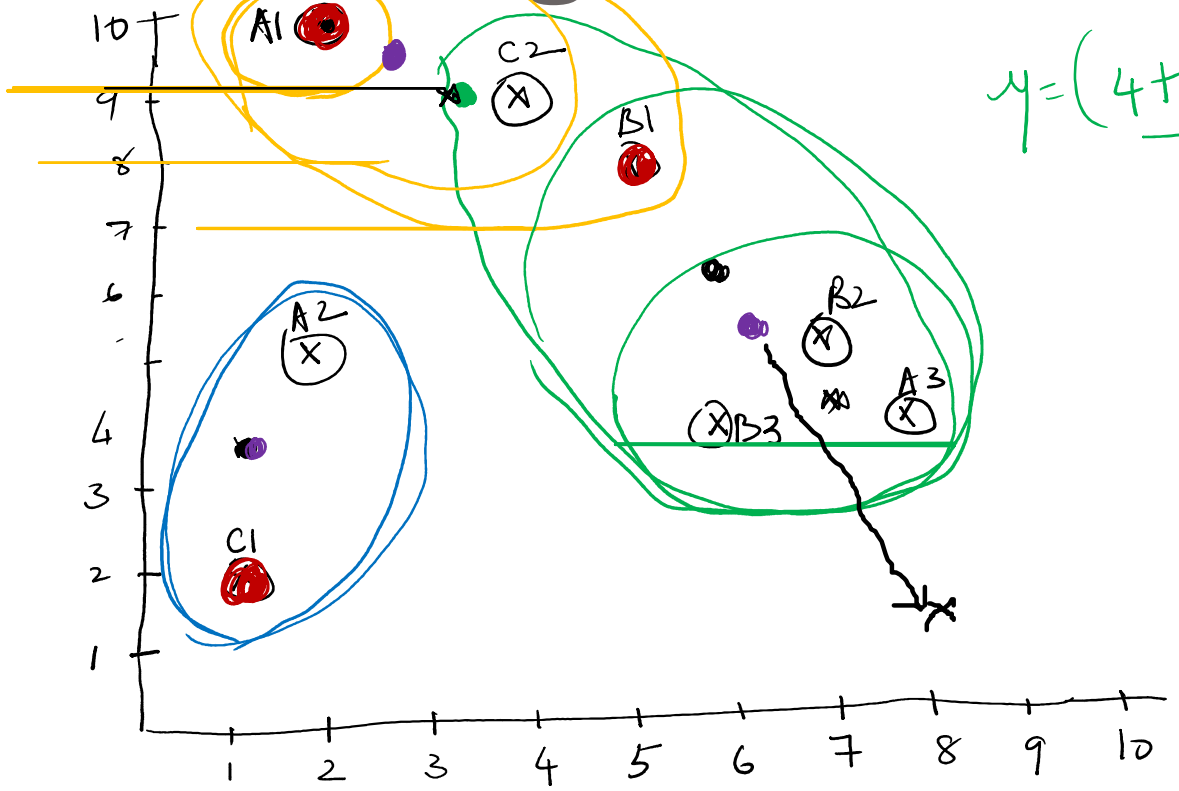
- **Density-based methods**
  - results in non-spherical shaped clusters also
  - Instead continue growing the given cluster as long as the density in the neighborhood exceeds some threshold
  - Popular techniques include DBSCAN , DENCLUE and OPTICS
- **Grid-based methods**
  - Quantize the object space into a finite number of cells that form a grid structure.
  - All clustering performed on the grid structure
  - Fast processing time
  - Popular techniques include STING and WAVECLUSTER
- **Model-based methods**
  - Hypothesize a model for each of the cluster and find the best fit of the data to the given model
  - Popular techniques include EM, COBWEB and SOM

K-Means Clustering

- Choose the number of clusters, k.
- Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.
- Assign each point to the nearest cluster center.
- Recompute the new cluster centers.
- Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).
- *Example:* The data set has three dimensions and the cluster has two points: $X = (x1, x2, x3)$ and $Y = (y1, y2, y3)$. Then the centroid Z becomes $Z = (z1, z2, z3)$, where $z1 = (x1 + y1)/2$ and $z2 = (x2 + y2)/2$ and $z3 = (x3 + y3)/2$.
- Within cluster variation - Square-error criterion $E = \sum_{i=1}^{k} \sum_{p \in Ci} \left| p - m_i \right|^2$
- Advantages:
  - Simplicity and speed
- Disadvantages :
  - Does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.
  - Sensitive to outliers

K-Means Clustering

A1	(2,10)
A2	(2,5)
A3	(8,4)
B1	(5,8)
B2	(7,5)
B3	(6,4)
C1	(1,2)
C2	(4,9)



Seed Values: A1, B1, C1

$x = \frac{(8+5+7+6+4)}{5}$

$y = \frac{(4+8+5+4+9)}{5}$

Iteration 1

Iteration 2

Iter 3 Centre

Iter 4

Cluster 1

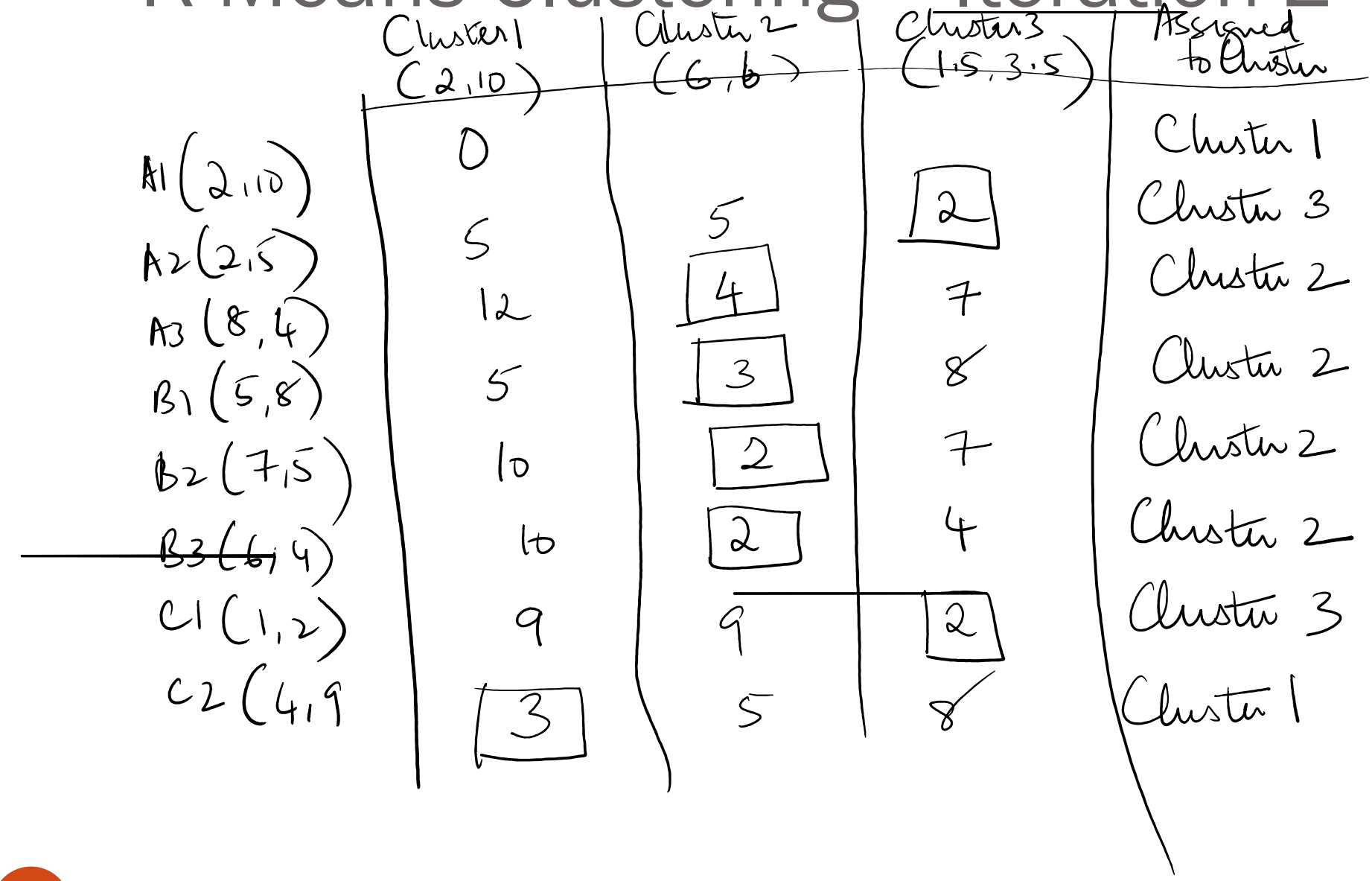
Cluster 2

Cluster 3

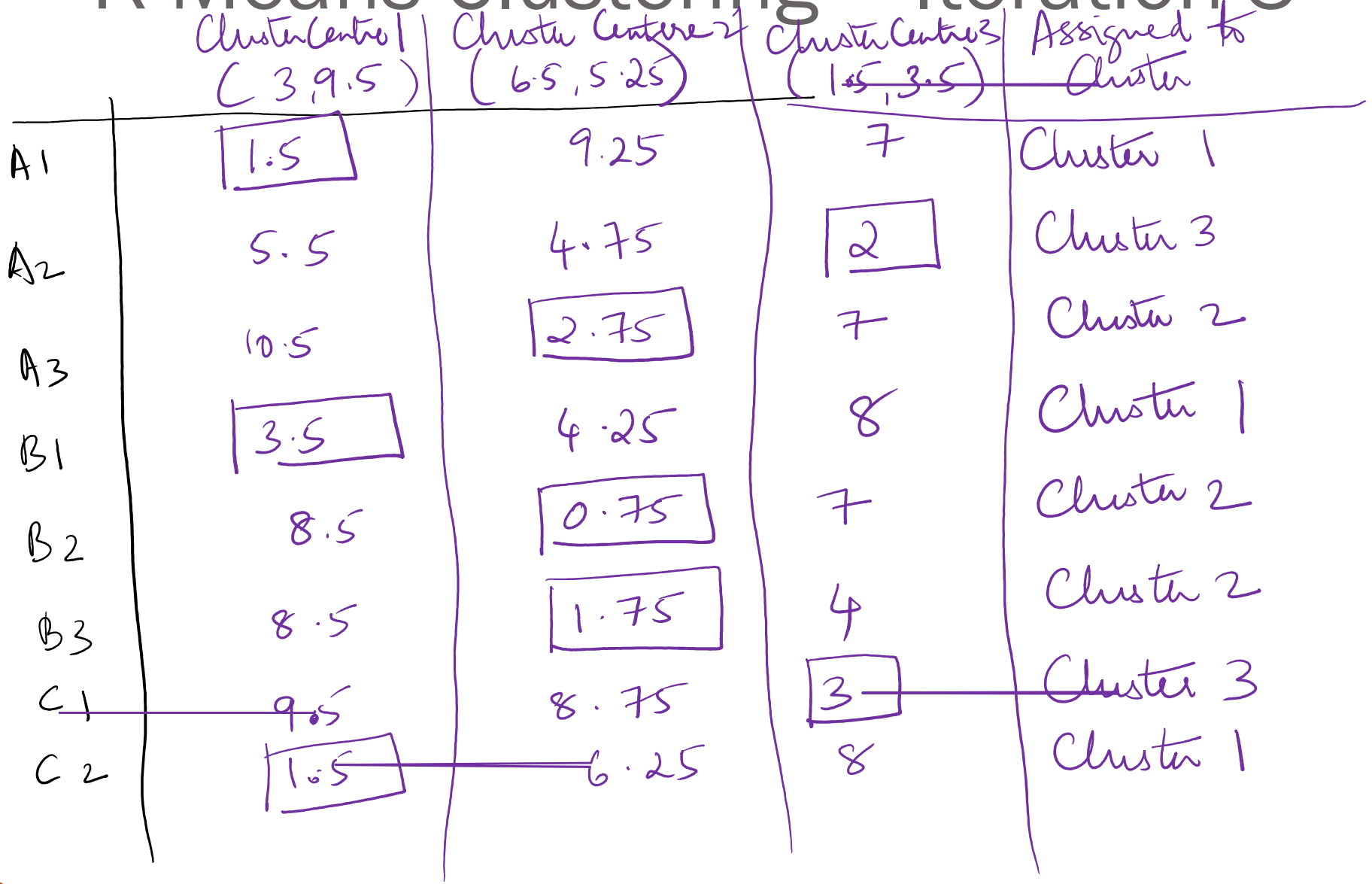
K-Means Clustering – Distance Matrix

DATA POINT	Cluster1 centre (2, 10)	Cluster2 Centre (5, 8)	Cluster3 Centre (1, 2)	Assign to cluster
A1 (2,10)	0	5	9	Cluster 1
A2 (2,5)	5	6	4	Cluster 3
A3 (8,4)	12	7	9	Cluster 2
B1 (5,8)	5	0		Cluster 2
B2 (7,5)	10	5	9	Cluster 2
B3 (6,4)	10	5	7	Cluster 2
C1 (1,2)			0	Cluster 3
C2 (4,9)	3	2	10	Cluster 2

K-Means Clustering – Iteration 2



K-Means Clustering – Iteration 3



Cluster quality measures

- Within cluster variation - Square-error criterion
  - $E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$

Centroid of Cluster 1 = (3.67 , 9) = 12.667

Cluster 2 = (7 , 4.3) = 3.87

Cluster 3 = (1.5 , 3.5) = 8

24.537

Cluster 1 A1(2,10) B1(5,8) C2(4,9)

$(1.67+1)^2 + (2.33)^2 + (0.33)^2$

Cluster 2

Cluster 3

Hierarchical Clustering

- Works by grouping data objects into a tree of clusters
- Can be classified depending on whether decomposition is
  - Agglomerative -bottom-up or Merging
  - Divisive – top down or Splitting
- Quality of clusters not good because once merge or split decision is made – no back tracking
- normally limited to small data sets (often less than 10,000 observations)
- speed to generate the hierarchical tree can be slow for higher numbers of observations

Agglomerative Clustering

Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process hierarchical clustering is this:

1.Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.

2.Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.

3.Compute distances (similarities) between the new cluster and each of the old clusters.

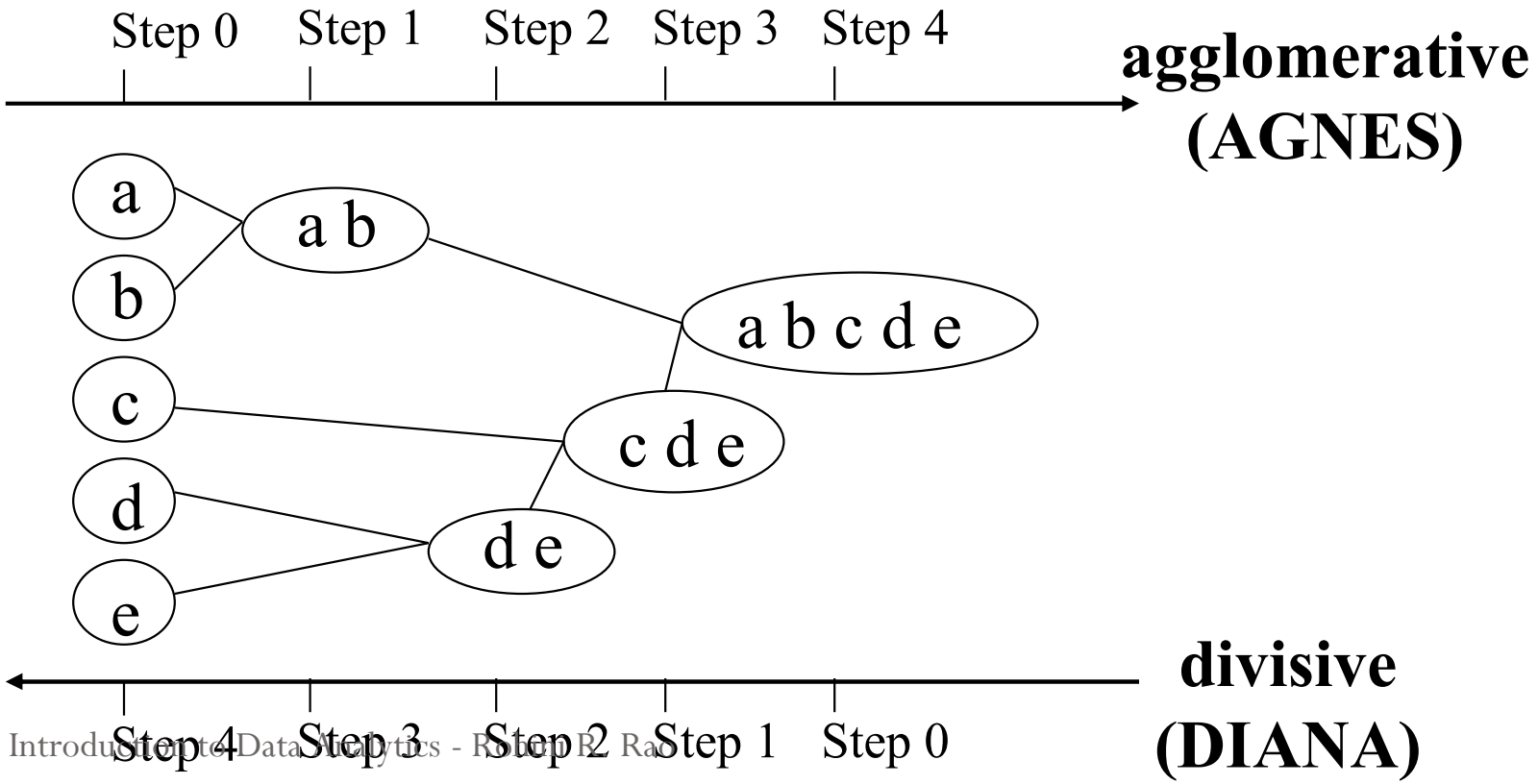
4.Repeat steps 2 and 3 until all items are clustered into a single cluster of size N or until certain termination condition.

• Terminating condition : desired number of clusters or the diameter of each cluster is within a certain threshold.

• Merging Criterion – Single Linkage Approach - Cluster C1 & C2 are merged if any object in C1 and any object in C2 form a minimum euclidean distance between any two objects from different clusters

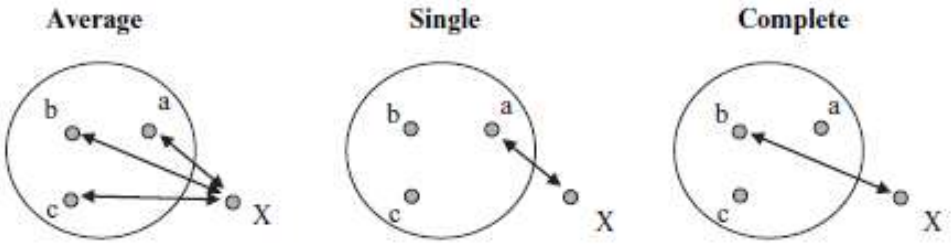
Hierarchical Clustering

- These method does not require the number of clusters ***k*** as an input, but needs a termination condition
- AGNES – Agglomerative Nesting
- DIANA – Divisive Analysis



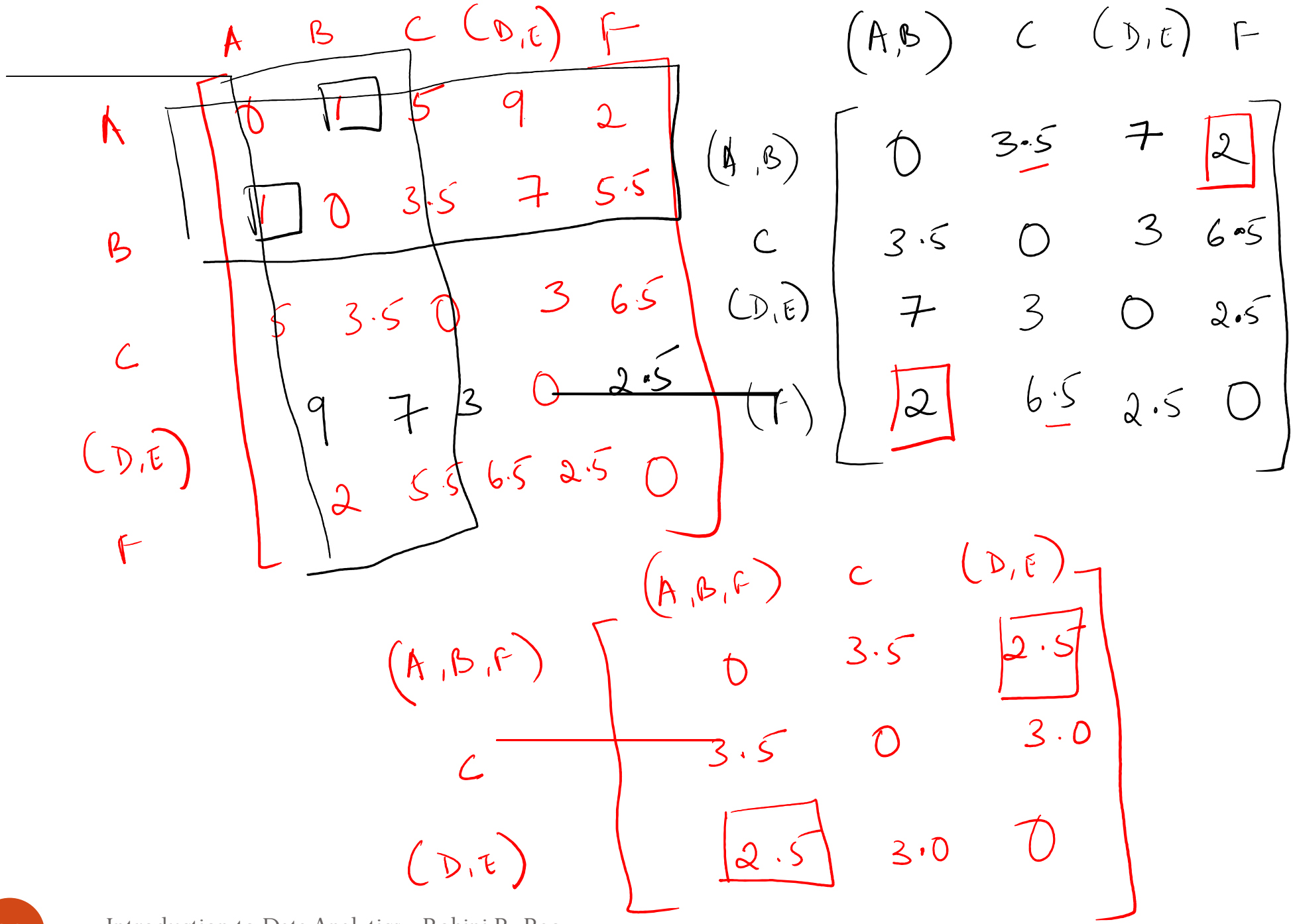
Computing Distances

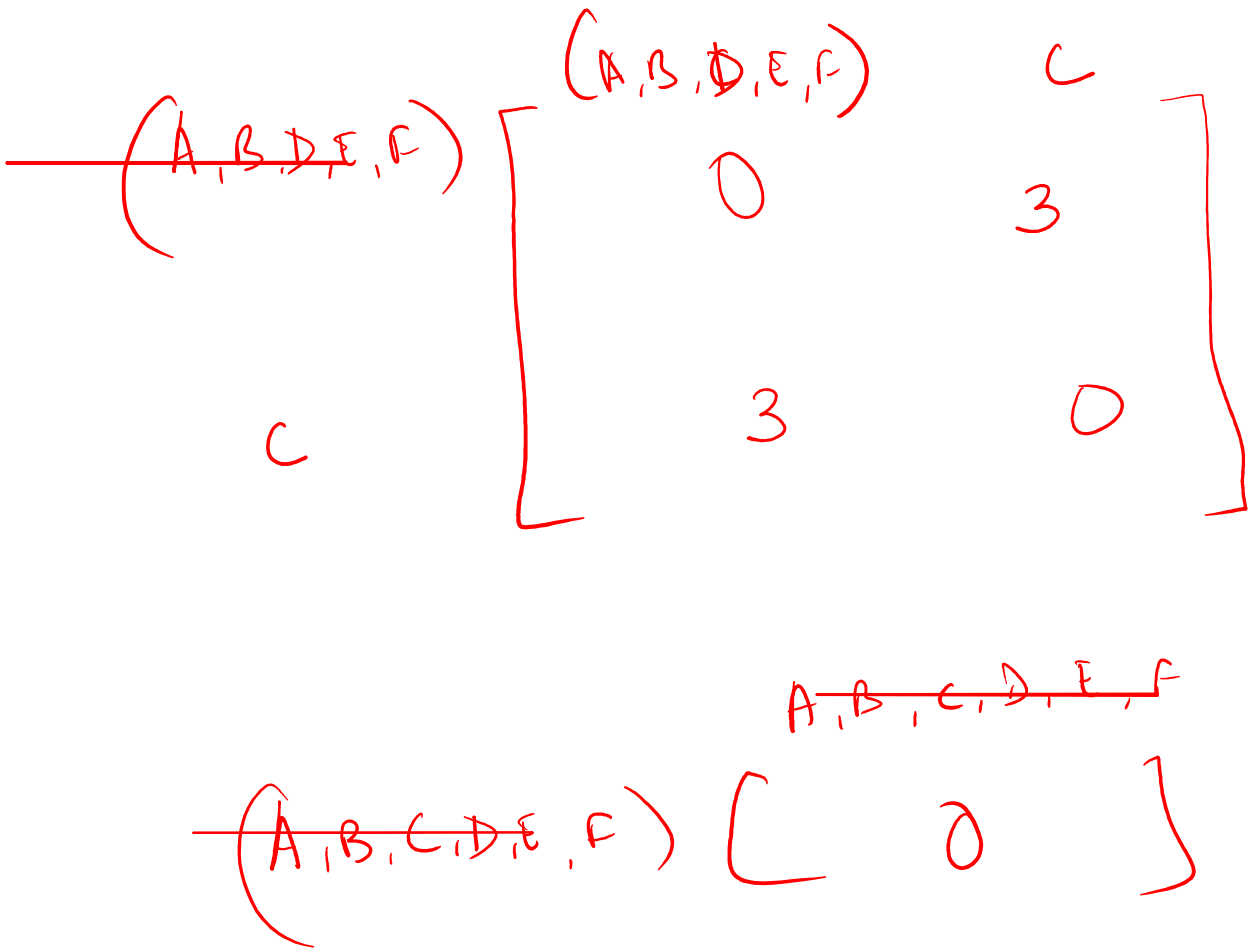
- Single-link clustering
  - •Consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.
- •Complete-link clustering
  - •Consider the distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster.
- •Average-link clustering :
  - • Consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.



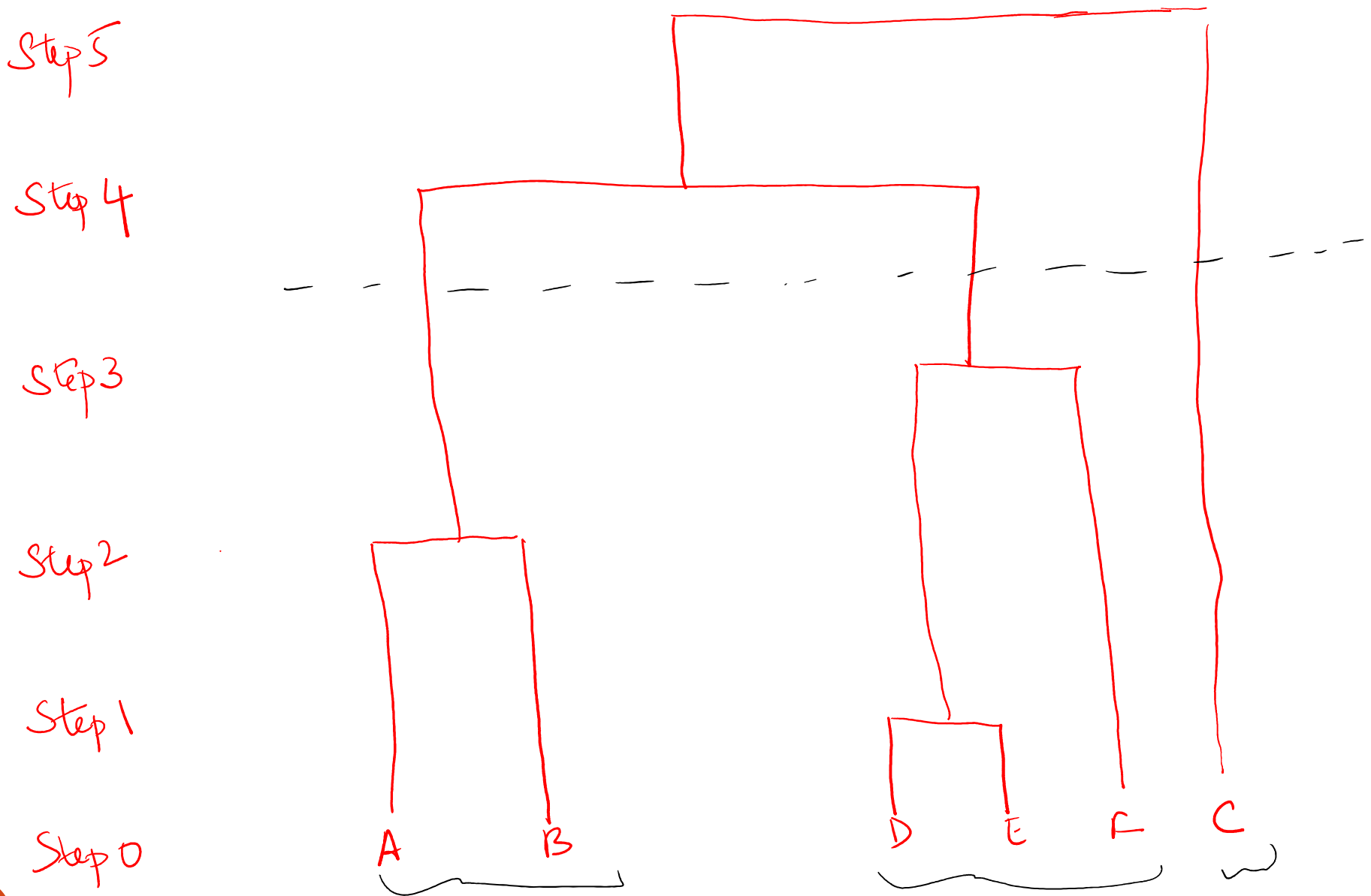
Example- Agglomerative with single link

	A	B	C	D	E	F
A	0	1	5	9	10	2
B	1	0	3.5	8	7	5.5
C	5	3.5	0	3	4	6.5
D	9	8	3	0	0.5	4.5
E	10	7	4	0.5	0	2.5
F	2	5.5	6.5	4.5	2.5	0


A B C (D,E) F
A
B
C
(D,E)
F
0 1 5 9 2
1 0 3.5 7 5.5
5 3.5 0 3 6.5
9 7 3 0 2.5
2 5.5 6.5 2.5 0
(A,B) C (D,E) F
(A,B)
C
(D,E)
F
0 3.5 7 2
3.5 0 3 6.5
7 3 0 2.5
2 6.5 2.5 0
(A,B,F) C (D,E)
(A,B,F)
C
(D,E)
0 3.5 2.5
3.5 0 3.0
2.5 3.0 0



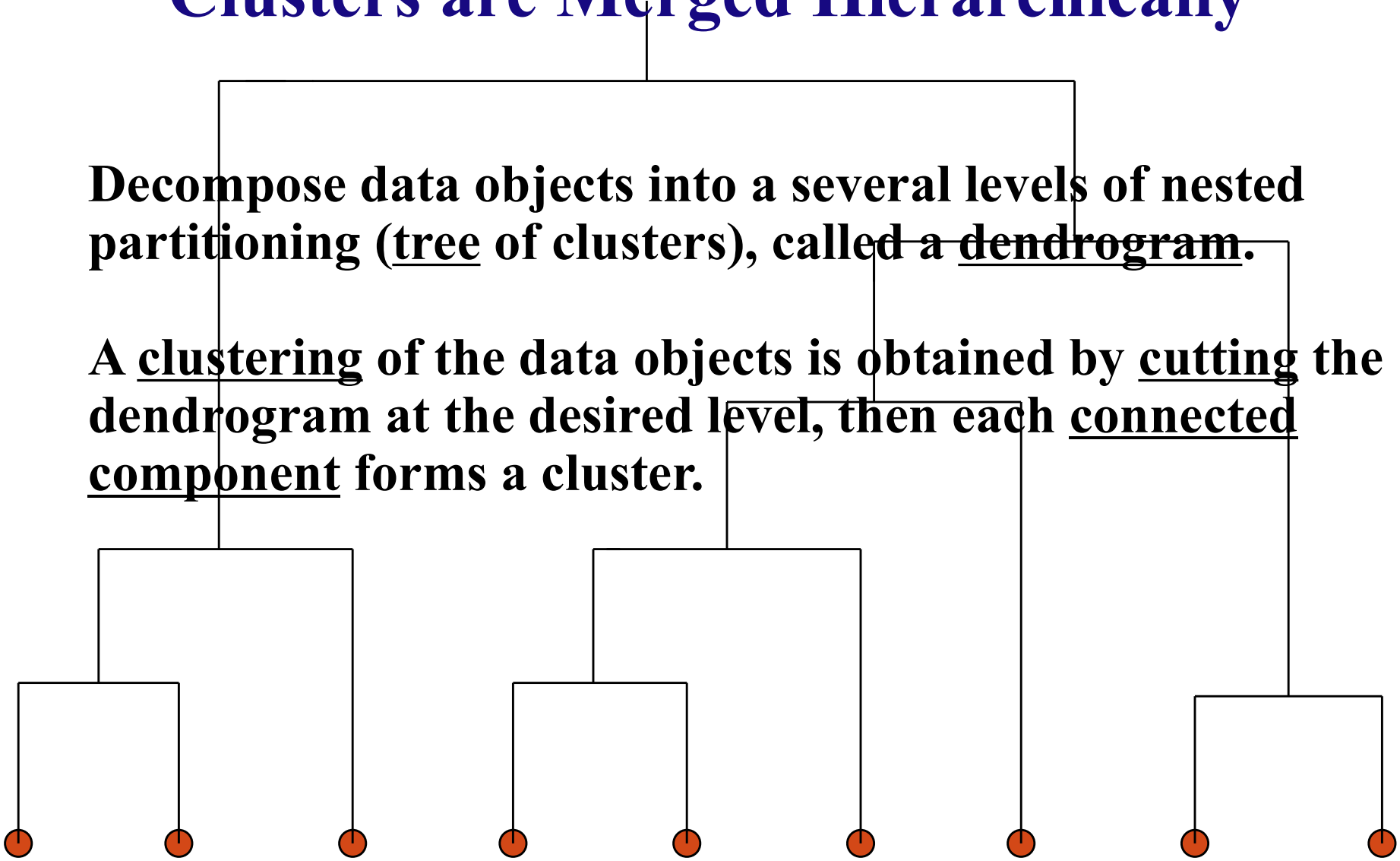
DENDROGRAM



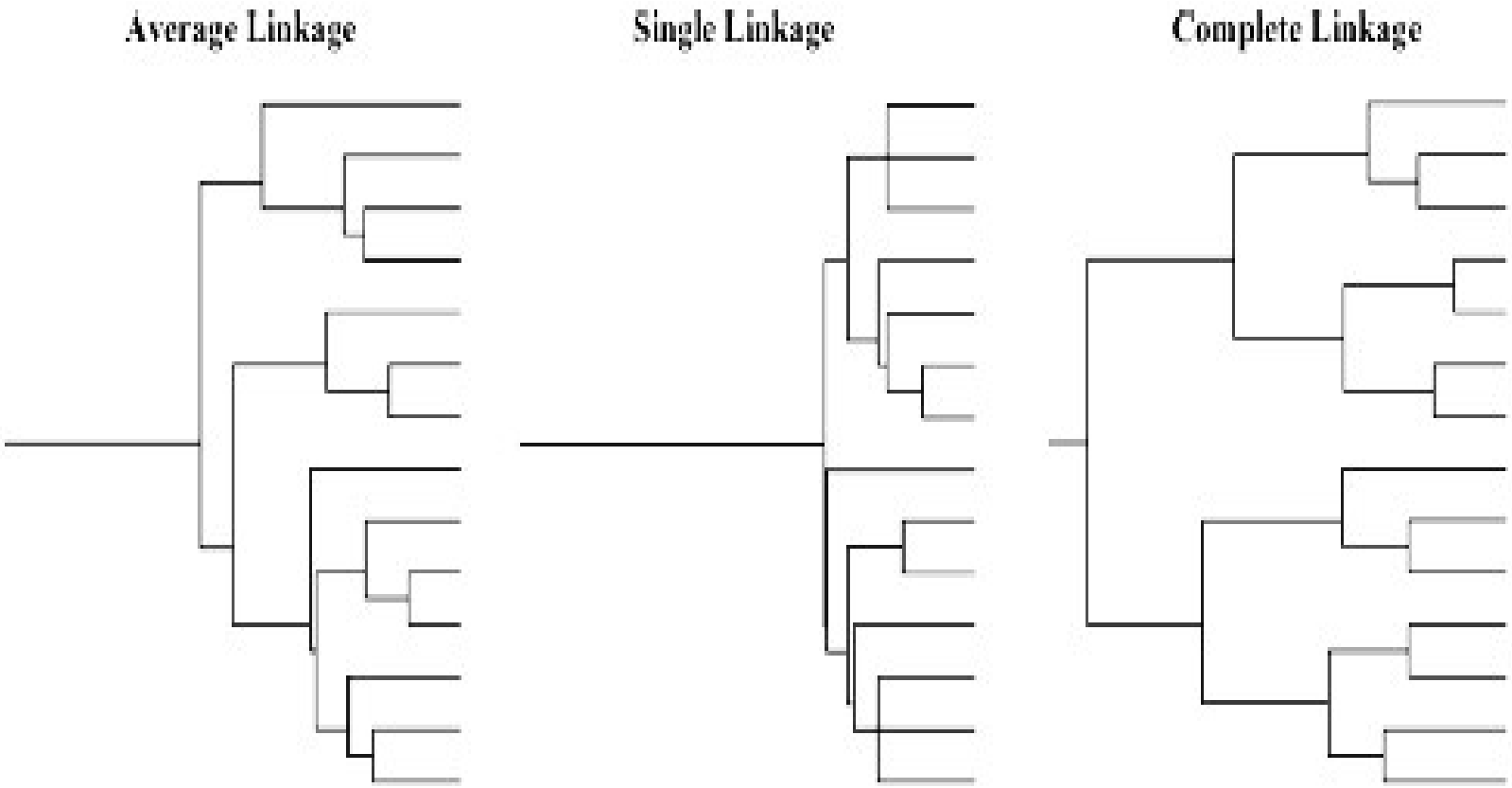
A *Dendrogram* Shows How the Clusters are Merged Hierarchically

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



Clustering under different linkage rules



392 cars – Agglomerative clustering Euclidean distance & average linkage

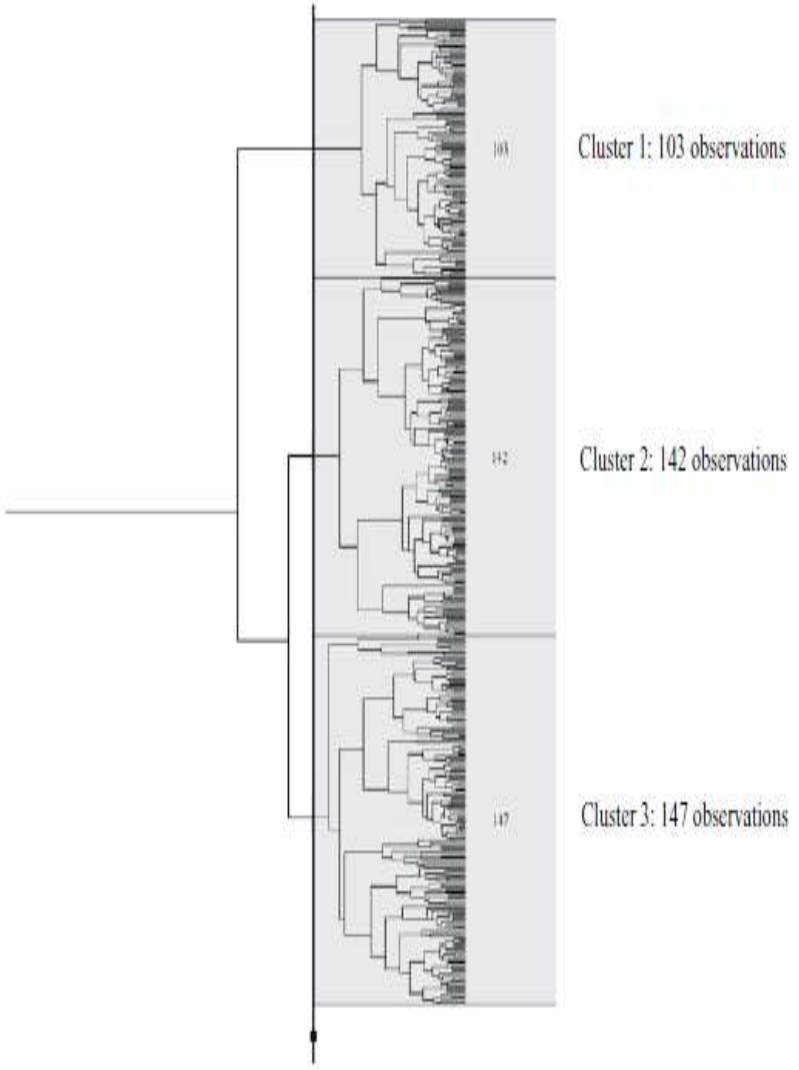
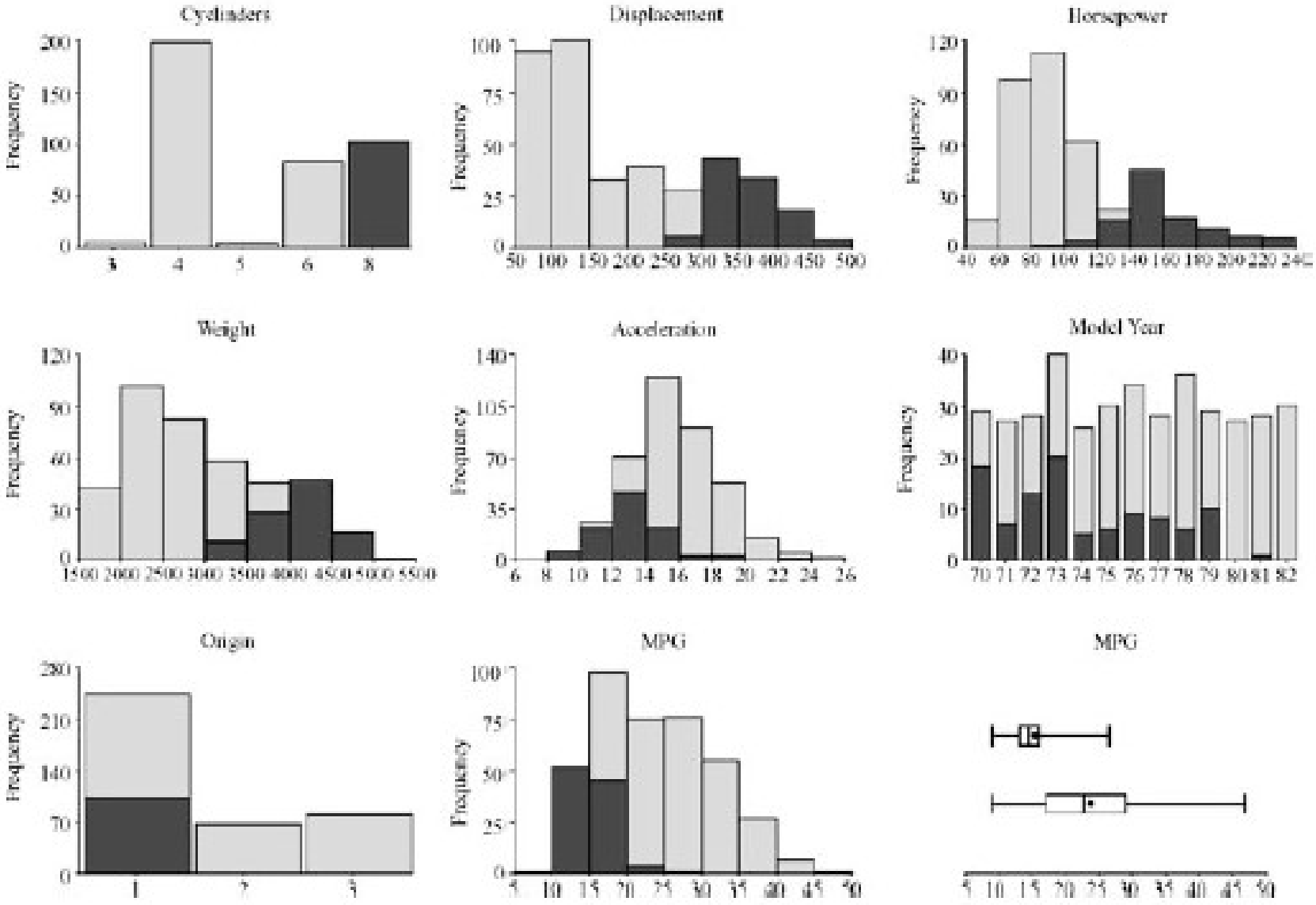


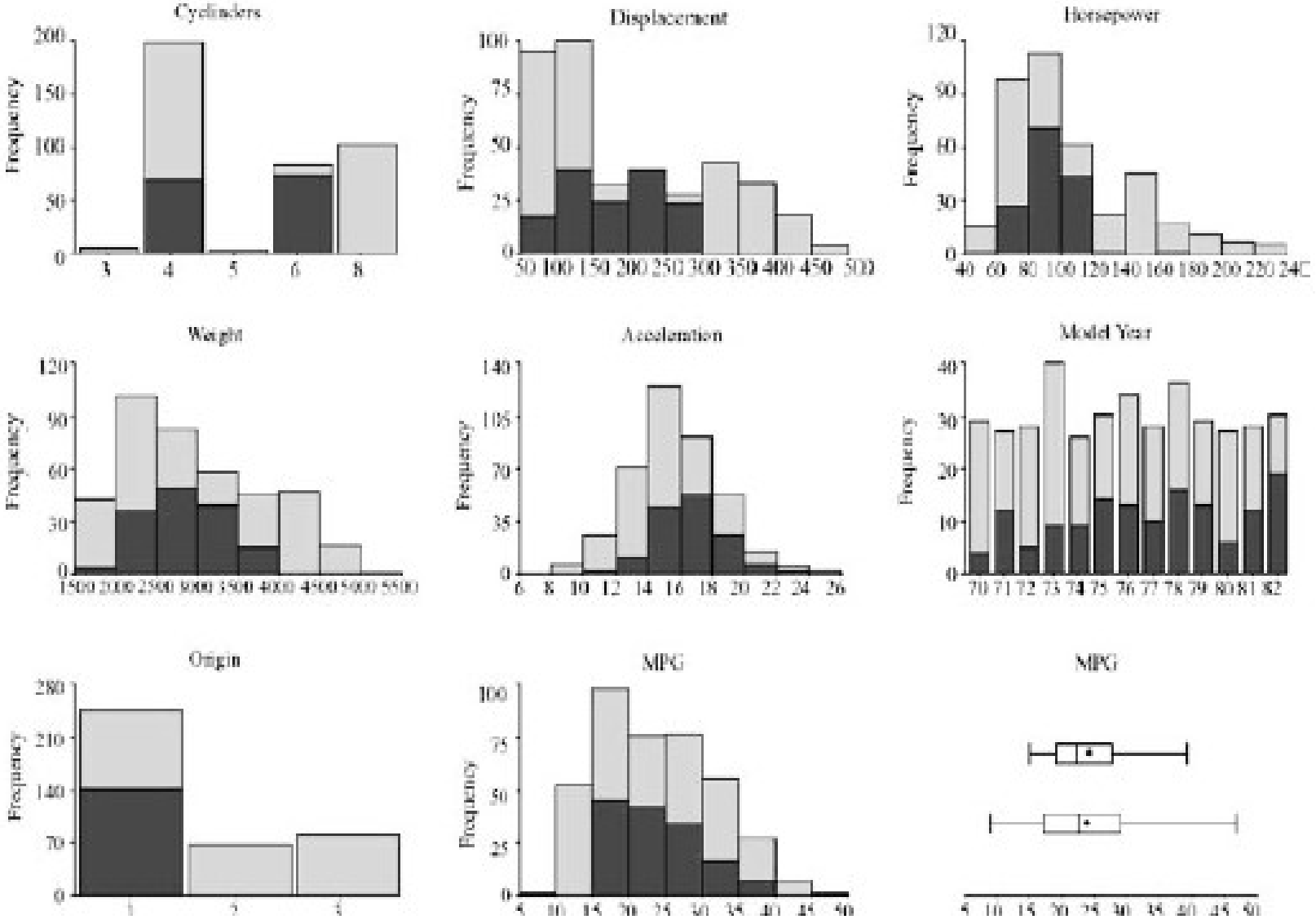
Table 6.8. Table of car observations

Names	Cylinders	Displace-ment	Horse power	Weight	Accele-ration	Model/ Year	Origin	MPG
Chevrolet Chevelle Malibu	8	307	130	3,504	12	1970	1	18
Buick Skylark 320	8	350	165	3,693	11.5	1970	1	15
Plymouth Satellite	8	318	150	3,436	11	1970	1	18
Amc Rebel SST	8	304	150	3,433	12	1970	1	16
Ford Torino	8	302	140	3,449	10.5	1970	1	17
Ford Galaxie 500	8	429	198	4,341	10	1970	1	15
Chevrolet Impala	8	454	220	4,354	9	1970	1	14
Plymouth Fury III	8	440	215	4,312	8.5	1970	1	14

Result - Summary of Cluster 1



Result - Summary of Cluster 2



Clustering the same data set using k-means clustering Euclidean distance

Cluster 1

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
C	8	9	7	8	9
D	6	7	7	7	8
G	7	8	8	6	6
H	8	9	6	5	5
L	2	5	6	8	9
Center (average)	**6.2**	**7.6**	**6.8**	**6.8**	**7.4**

Cluster 2

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
E	1	2	5	3	4
F	3	4	5	3	5
I	2	3	5	6	5
K	3	2	6	5	7
Center (average)	**2.25**	**2.75**	**5.25**	**4.25**	**5.25**

Cluster 3

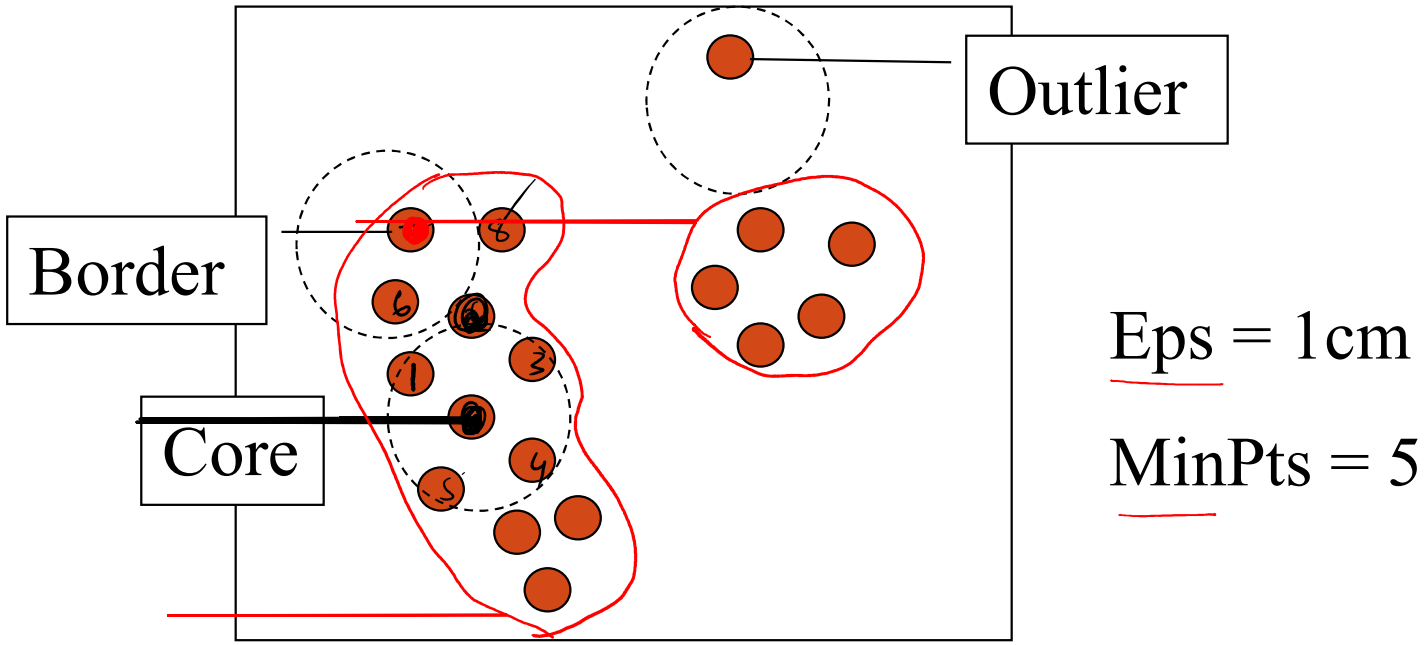
Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	7	8	4	5	2
B	6	8	5	4	2
J	1	2	4	4	2
M	3	5	4	6	3
N	3	5	5	6	3
Center (average)	**4**	**5.6**	**4.4**	**5**	**2.4**

Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



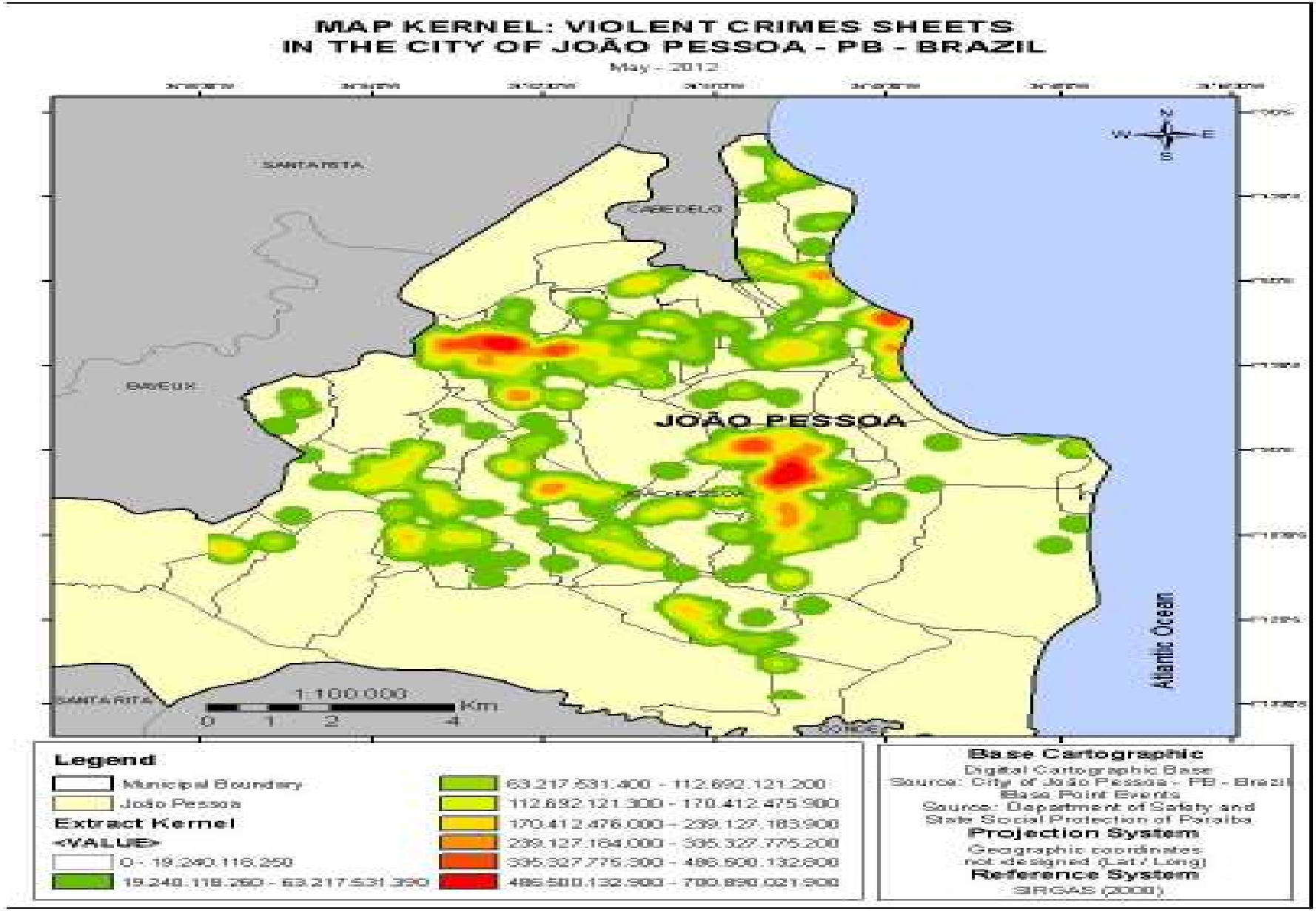
DBSCAN: The Algorithm

- Arbitrary select a point ***p***
- Retrieve all points density-reachable from ***p*** wrt ***Eps*** and ***MinPts***.
- If ***p*** is a core point, a cluster is formed.
- If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

Case study 3- Crime & Hot spot analysis

- Crime hot spots is an area where the number of criminal events or disorder is higher than in any other places particular environments that attract drug trading and crimes in larger-than-expected concentrations, so-called crime generators
- **Hot spots defined by attributes such as**
  - Location – latitude & longitude
  - particular activities such as drug trading, use of firearms , weapons
  - specific concentrations of land uses- hotels , bars
  - interactions between activities and land uses, such as thefts at transit stations or bus stops
- **Techniques that can be used**
  - Partitioning method – k-mode clustering
  - Density based method – Single Kernel Density method
- **Interpretation of clusters**
  - The area of highest crime rate is where many shops, banks and people hang out.
  - Another cluster is the neighborhood of the city, where residents have a high purchasing power
  - the number of crimes using firearms was very large
- Case study

Crime & Hot spot analysis



Frequent Patterns

- Types of Patterns include
  - Temporal Patterns
  - Spatial Patterns
  - Functional Patterns
- Frequent pattern mining searches for recurring relationships in a given sets.
- Frequent Patterns
  - Are patterns that appear in the data set frequently
  - A subsequence that appears frequently is a frequent sequential pattern
  - If a sub structure occurs frequently it is called a frequent structured pattern

Frequent Patterns

IF the customer's age is 18 AND

the customer buys paper AND

the customer buys a hole punch

THEN the customer buys a binder

Market Basket Analysis

- Analyzes customer buying habits by finding associations between different items that customers place in their shopping baskets
- Leads to the discovery of associations and correlations among items in large transactional or relational data sets.
- "*Which groups or sets of items are customers likely to purchase on a given trip to the store ?*"
- Helps in business decision making process, such as
  - catalog design,
  - cross-marketing
  - customer shopping behavior analysis.
- Can be used to plan different store layouts.
  - Place them together
  - Place them far apart
  - Decide which item to put on sale.
- Patterns can be represented in the form of Association rules

Association Rules

- Given a set of records each of which contains some items from a given collection
- Produce dependency rules that will predict occurrence of an item based on occurrence of other items
- Rules discovered
- {Milk} → {Bread}
- {Bread} → {Milk}

1	Bread, Milk
2	Eggs, Bread, Milk
3	Bagels, cream cheese, orange juice
4	Coke, Potato chips
5	Bread, milk, orange juice

Marketing and Sales Promotion:

Let the rule discovered be

$$\{Bagels, \ldots\} \Rightarrow \{Potato\ Chips\}$$

Potato Chips as consequent => Can be used to determine what should be done to boost its sales.

Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.

Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Associative Rules

- Each item has a boolean variable representing the presence or absence of that item.
- Each basket can be represented by a Boolean vector of values assigned to these variables.
- These patterns are then represented as a Boolean Association Rule

  buys(X,"computer") =>buys(X,"antivirus_software ")
  [support = 2%, confidence = 60 %]

- Measures of pattern interestingness can be
- Subjective – useful, previously unknown
- Objective
  - Support of 2 % of all transactions have computers and anti virus software purchased together
  - Confidence of 60 % means that 60 % of all customers who purchased a computer also bought the software.
- Minimum support and confidence set by domain expert

Association Rule Mining

- Steps involved
  - **Find all frequent items sets** – each of these item sets will occur at least as frequently as a pre determined minimum support count
  - **Generate strong association rules from the frequent item sets** – rules should satisfy minimum support and confidence.
  - Association rules can be generated as follows
    - For each frequent itemset l, generate all nonempty subsets of l
    - For every nonempty subset s of l , output the rule *s=> (l-s)*
    - If *support_count(l) / support_count(s) >= min_conf*
    - Where min_conf is the minimum confidence threshold

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

Example 1 – Apriori

Transaction				
T100	I1	I2	I5	
T200	I2	I4		
T300	I2	I3		
T400	I1	I2	I4	
T500	I1	I3		
T600	I2	I3		
T700	I1	I3		
T800	I1	I2	I3	I5
T900	I1	I2	I3	

Basket →

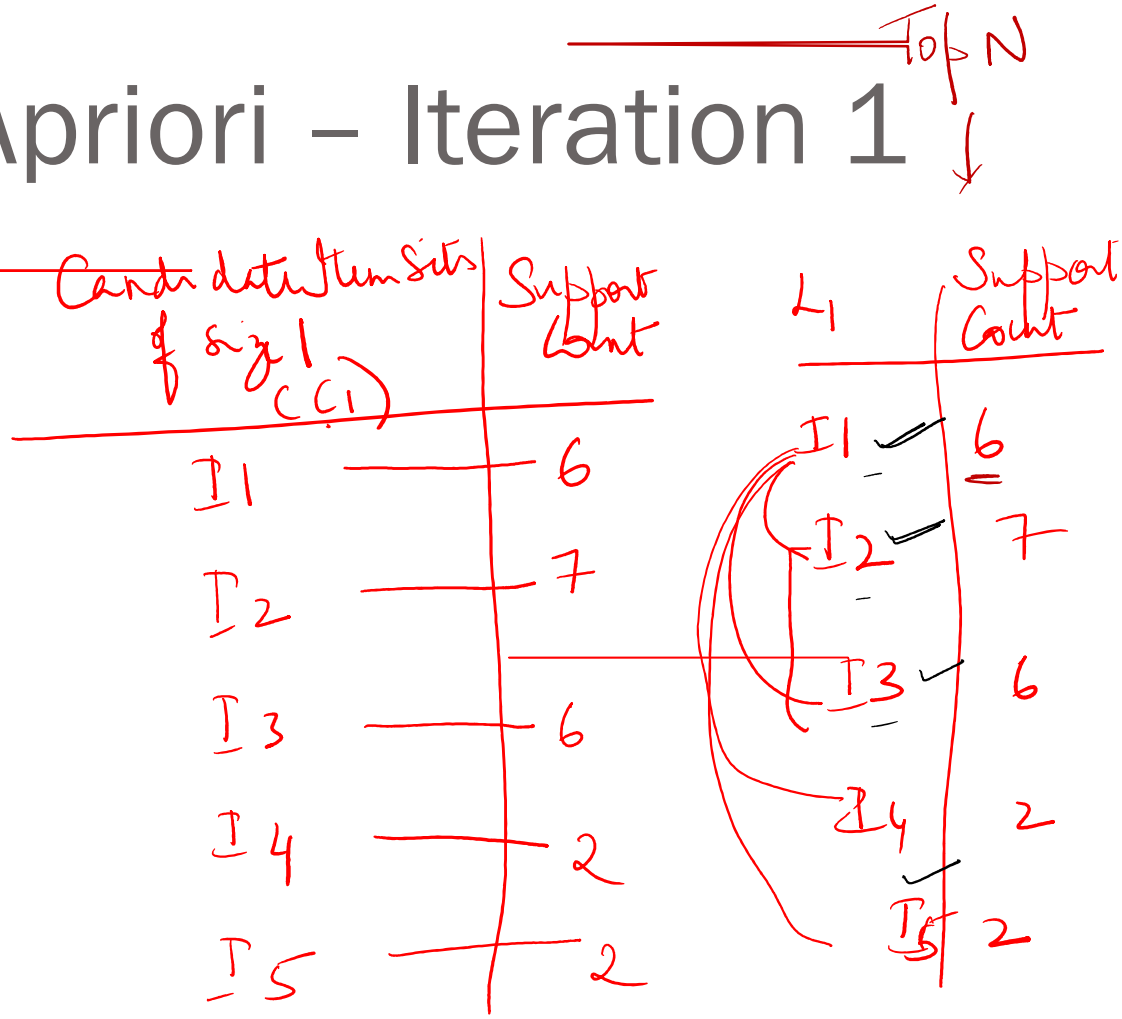
Find all frequent itemsets with minimum support count of 2.

Minimum support(%) = 22.%

Find all association rules with minimum confidence of 70%

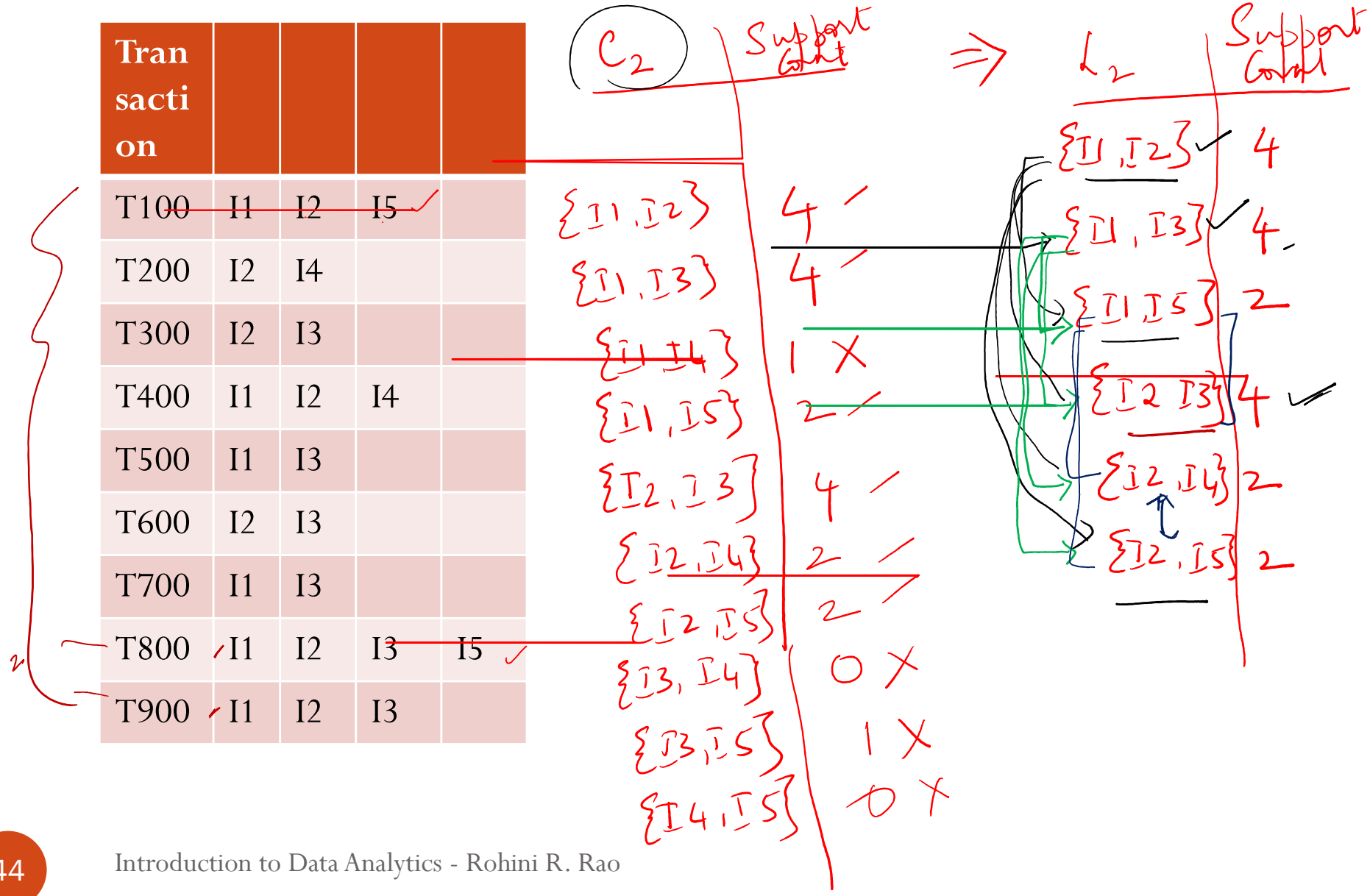
Example 1 – Apriori – Iteration 1

Transaction				
T100	I1	I2	I5	
T200	I2	I4		
T300	I2	I3		
T400	I1	I2	I4	
T500	I1	I3		
T600	I2	I3		
T700	I1	I3		
T800	I1	I2	I3	I5
T900	I1	I2	I3	



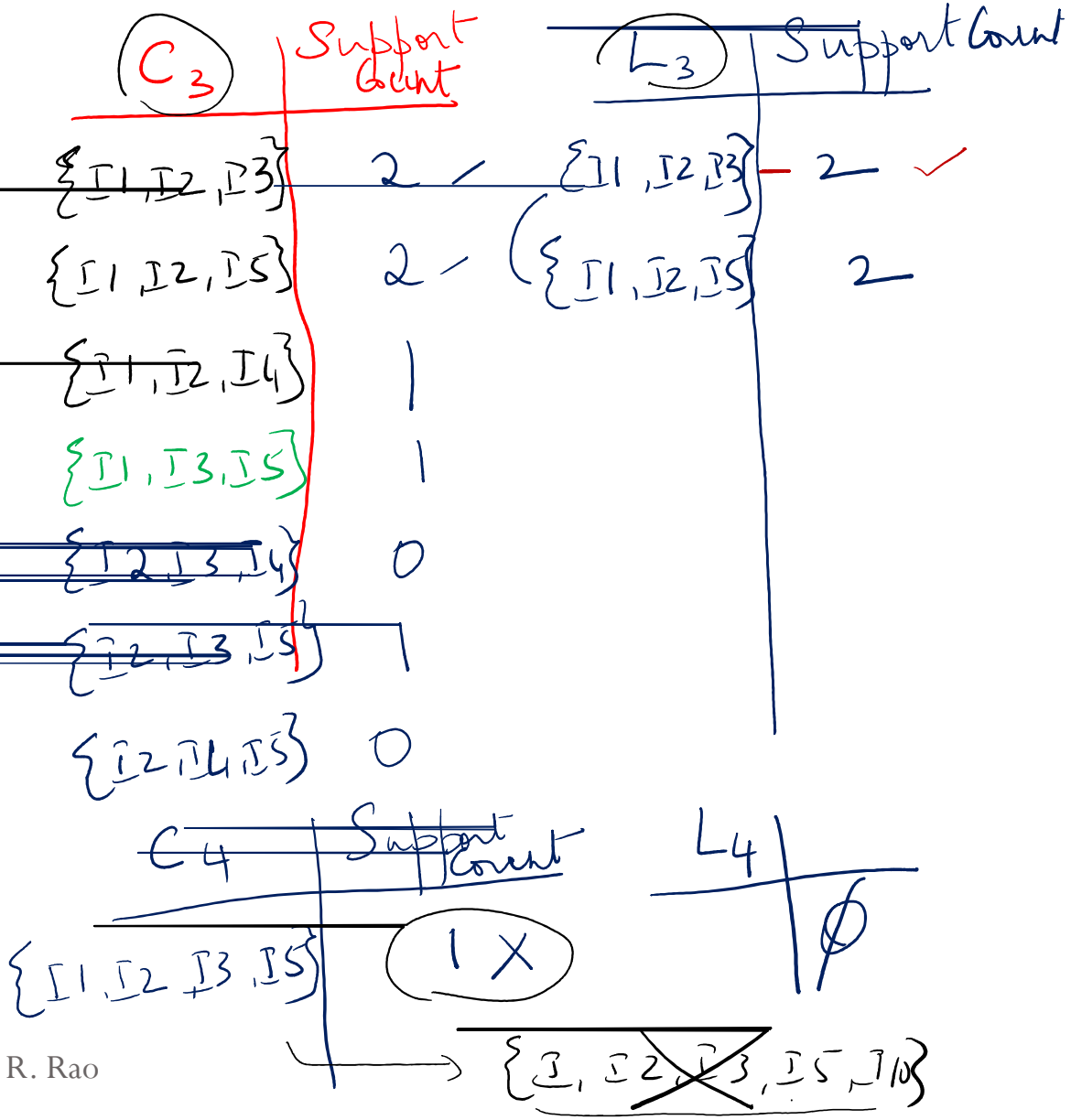
Example 1 – Apriori Iteration 2

Transaction				
T100	I1	I2	I5	
T200	I2	I4		
T300	I2	I3		
T400	I1	I2	I4	
T500	I1	I3		
T600	I2	I3		
T700	I1	I3		
T800	I1	I2	I3	I5
T900	I1	I2	I3	



Example 1 – Apriori Iteration 3

Transaction				
T100	I1	I2	I5	
T200	I2	I4		
T300	I2	I3		
T400	I1	I2	I4	
T500	I1	I3		
T600	I2	I3		
T700	I1	I3		
T800	I1	I2	I3	I5
T900	I1	I2	I3	



Example 1 – Apriori – Association Rules

$A \Rightarrow B$ · Confidence $= \frac{SC(A,B)}{SC(A)}$

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

$\{I1, I2, I3\}$ – Support Count = 2

$\{I1\} \Rightarrow \{I2, I3\}$ Confidence $= 2/6 * 100 = 33\%$

$\{I2\} \Rightarrow \{I1, I3\}$ Confidence $= 2/7 \times 100 = 28.6\%$

$\{I3\} \Rightarrow \{I1, I2\}$ Confidence $= 2/6 * 100 = 33\%$

$\{I1, I2\} \Rightarrow \{I3\}$ Confidence $= 2/4 \times 100 = 50\%$

$\{I1, I3\} \Rightarrow \{I2\}$ Confidence $= 2/4 \times 100 = 50\%$

$\{I2, I3\} \Rightarrow \{I1\}$ Confidence $= 2/4 \times 100 = 50\%$

Example 1 – Apriori – Association Rules

{I1, I2, I5} Support Count = 2

{I1} ⇒ {I2, I5} Confidence = 2/6 × 100 = 33% X

{I2} ⇒ {I1, I5} Confidence = 2/7 × 100 = 28% X

{I5} ⇒ {I1, I2} Confidence = 2/2 × 100 = 100% ✓

{I1, I2} ⇒ {I5} Confidence = 2/4 = 50%

{I1, I5} ⇒ {I2} Confidence = 2/2 × 100 = 100% ✓

{I2, I5} ⇒ {I1} Confidence = 2/2 × 100 = 100% ✓

RESULTS

① $\{I1, I2, I3\}$: 2 ② $\{I1, I2, I5\}$: 2 are frequent item sets

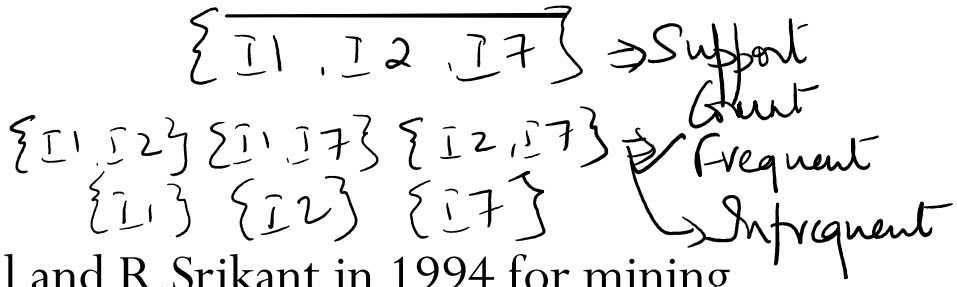
② ASSOCIATION RULES

(a) $\{I5\} \Rightarrow \{I1, I2\}$ [Support = 22% ~~Confidence = 100%~~]

(b) $\{I1, I5\} \Rightarrow \{I2\}$ ~~[Support = 22%]~~ [Confidence = 100%]

(c) $\{I2, I5\} \Rightarrow \{I1\}$ [Support = 22% Confidence = 100%]

Apriori Algorithm



- Seminal algorithm proposed by R. Agarwal and R.Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- Employs an iterative , bottom-up approach known as level-wise search , where k-itemsets are used to explore (k+1) itemsets
- **Based on Apriori property** :
  - **All nonempty subsets of a frequent itemset must also be frequent**
  - **Antimonotone** – if a set cannot pass a test then all its supersets will fail the same test as well
- Two step process of the Apriori algorithm are
- **Join Step**
  - To find L_k , a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k
- **Prune Step**
  - C_k is the superset of L_k, its members may or may not be frequent but all of the frequent k-itemsets are included in C_k
  - A scan of the database to determine the count of each candidate in C_k will result in L_k
  - To reduce the size of C_k the apriori property is used . Any (k-1) itemset that is not frequent cannot be a subset of a frequent k-itemset.

Frequent Item sets

- Let T = { I_1,I_2 , I_3 ….I_m} be a set of items
- Each transaction has a T_id
- An association is an implication of the form
- A=> B where A C T , B C T and A ∏ B = Ø
- Support (A => B) = P (A U B)
- Confidence(A => B) = P (B/A)
- A set of items is called as an item set
- The set { computer , anti_virus} is called 2 item set
- The occurrence frequency of an item set is the number of transactions that contain the item set.
- This is also known as frequency , support count or count of the item set.
- Confidence(A=> B) = P (B/A) = support (A U B) / support (A)

  A ⇒ B, A,B – Set of items.
  - Support : Count (A U B)
  - Confidence : $\frac{\text{support Count (A U B)}}{\text{support Count (A)}}$

Pros & Cons of Apriori

- **Characteristics of Apriori Algorithm:**
  - Simple, user friendly algorithm
  - No. of database scans = k, if (k-1)-large item sets are possible
  - Discovers all possible large item sets
  - Robust algorithm
  - Ranking of rules required
- **Disadvantages of Apriori Algorithm**

In general a data set that contains k items can potentially generate up to 2^K – 1 frequent item sets

- Generation of candidate item sets is expensive (in both space and time)
- Support counting is expensive
- Subset checking (computationally expensive)
- Multiple Database scans (I/O)

Improving the efficiency of Apriori

- **Hash-based technique-**

  - While scanning data base for 1-itemsets L_1 generate all of the n+1 item sets for each transaction

  - Hash ie. map into different buckets of a hash table and increase the bucket count.

  - If the bucket count is below support threshold remove from the candidate set.

- **Transaction reduction**
  - transaction that does not contain any frequent k-item sets cannot contain frequent (k+1) item sets.
  - Remove such transactions or mark them.
- **Partitioning**
  - Transactions in D are divided into n partitions
  - Phase 1 : Find the frequent item sets local to each partition
  - Phase 1 : Combine all local Frequent Item sets to form candidate item set.
  - Phase 2 : Find global frequent item sets among the candidates
  - Phase 2 : Frequent item sets in D
- **Sampling**
  - Search for frequent item sets in S of a given data D
  - Trade off between accuracy and efficiency
  - Lower support thresholds
- **Dynamic Item set Counting**
  - New candidate item sets can be added at any start point unlike Apriori which determines new candidate item sets only immediately before each complete database scan

Partition Algorithm

- Partitions the set of transactions into smaller segments so that they can be accommodated in memory.
- Does only two scans of database
- Phase 1
  - generates n non overlapping partitions of dataset
  - 1st scan generates a set of all potential frequent itemsets
  - Applies local support.
- Phase 2 – Superset of frequent items used to generate actual support count
  - 2nd scan counts actual support
- Theory is that phase 1 frequent item sets can contain false positives, no false negatives

Example 2 – Apriori –

$\sigma = 20\%$ Confidence $= 70\%$

Transaction	A1	A2	A3	A4	A5	A6	A7	A8	A9
1	1	0	0	0	1	1	0	1	0
2	0	1	0	1	0	0	0	1	0
3	0	0	0	1	1	0	1	0	0
4	0	1	1	0	0	0	0	0	0
5	0	0	0	0	1	1	1	0	0
6	0	1	1	1	0	0	0	0	0
7	0	1	0	0	0	1	1	0	1
8	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	1	0
10	0	0	1	0	1	0	1	0	0
11	0	0	1	0	1	0	1	0	0
12	0	0	0	0	1	1	0	1	0
13	0	1	0	1	0	1	1	0	0
14	1	0	1	0	1	0	1	0	0
15	0	1	1	0	0	0	0	0	1

Example 2 - Apriori

- L1

ItemSet	Support Count
A2	6
A3	6
A4	4
A5	8
A6	5
A7	7
A8	4

Example 2- Apriori

- C3

ItemSet	Support Count
{A2,A3,A4}	1
{A2,A3,A5}	0
{A2,A3,A7}	0
{A3,A5,A7}	3
{A3,A5,A6}	0
{A3,A6,A7}	0
{A5,A6,A7}	1

Example 2 - Apriori

- Association Rules – Frequent Itemset{A3,A5,A7}

Example 2 – Partitioning Algorithm

Transaction	A1	A2	A3	A4	A5	A6	A7	A8	A9
1	1	0	0	0	1	1	0	1	0
2	0	1	0	1	0	0	0	1	0
3	0	0	0	1	1	0	1	0	0
4	0	1	1	0	0	0	0	0	0
5	0	0	0	0	1	1	1	0	0
6	0	1	1	1	0	0	0	0	0
7	0	1	0	0	0	1	1	0	1
8	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	1	0
10	0	0	1	0	1	0	1	0	0
11	0	0	1	0	1	0	1	0	0
12	0	0	0	0	1	1	0	1	0
13	0	1	0	1	0	1	1	0	0
14	1	0	1	0	1	0	1	0	0
15	0	1	1	0	0	0	0	0	1

Example 2 – Partitioning Algorithm

Types of patterns mined

- **Multilevel Association rules**
  - buys(X,"computer") => buys(X,"HP-printer")
  - buys(X,"lap-top-computer") => buys(X,"HP-printer")
- **Multidimensional Association rules**
  - age(X,"30..39") Π income(X,"42K..48K") => buys(X,"high resolution TV)
- **Association rules or correlation rules**
  - Strong Gradient relationships – "The average sales from Sony digital camera increases over 16 % when sold together with Sony laptop"
- **Sequential Pattern Mining**
  - which searches for frequent subsequences in a sequence data set.
  - Ex : customer first buys PC followed by digital camera and then memory card.
  - Structured Pattern Mining which looks for frequent substructures
- **Recommender systems**
  - which recommend information items (e.g., books, movies, web pages) that are likely to be of interest to the user based on similar users' patterns.

From Association Mining to Correlation Analysis

- Suppose that 10,000 transactions are analysed, 6000 include computer games, 7500 include videos, 4000 included both games and videos.
- Suppose there is an association rule
  - **buys(X,"computer games") => buys(X,"videos")**
    [support = 40 % , confidence = 66 %]

- Rule is misleading because the probability of purchasing videos is 75% which is even larger than 66 %
- Actually computer games and videos could be negatively associated
- Confidence which is conditional probability does not measure strength of the correlation and implications between antecendent and consequent.
- Even strong association rules can be uninteresting and misleading.
- So support-confidence framework needs be supplemented.

Correlation rules

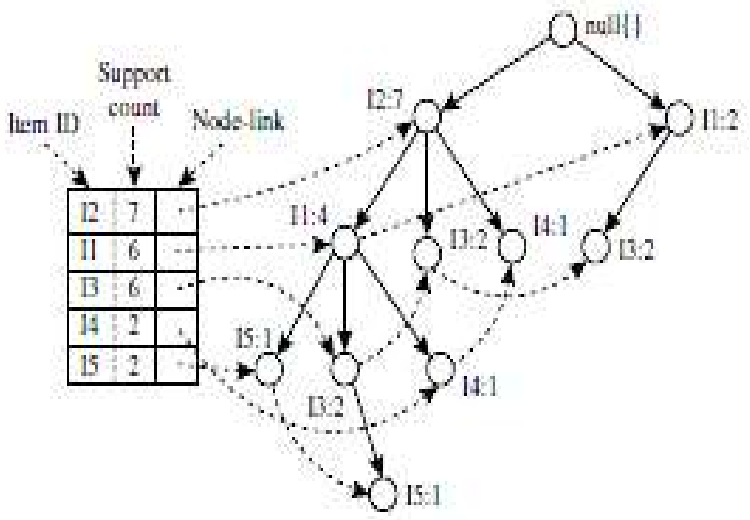
- A => B [support, confidence, correlation]

- The occurrence of itemset A is independent of the occurrence of itemset B if P(AUB) = P(A)P(B) otherwise itemsets A and B are dependent and correlated as events.
- **Lift** :
- between the occurrence of A and B can be measured by computing
  - Lift(A,B) = P(AUB)/P(A)P(B)
  - Also written as Lift(A,B) = conf(A=>B)/ supp(B)
- If the resulting value
  - <1 then the occurrence of A is negatively correlated with occurrence of B
  - >1 then A and B are positively correlated
  - =1 then A and B are independent and there is no correlation between them.

- Ex: P({game,video})/P({game}) * P({video}) = 0.40 / (0.60 * 0.75) = 0.89
- Because it is less than 1 video and game are negatively correlated

Other important Algorithms

- **Frequent Pattern Growth Algorithm**
  - Allows frequent item set discovery without candidate item set generation.
  - Two step approach:
    - Step 1: Build a compact data structure called the FP-tree, built using 2 passes over the data-set.
    - Step 2: Extracts frequent items ets directly from the FP-tree by traversal through FP-Tree
  - Advantages of FP-Growth
    - only 2 passes over data-set
    - no candidate generation
    - much faster than Apriori
  - Disadvantages of FP-Growth
    - FP-Tree may not fit in memory!!
    - FP-Tree is expensive to build



Example 1 – FP Growth

Transaction				
T100	I1	I2	I5	
T200	I2	I4		
T300	I2	I3		
T400	I1	I2	I4	
T500	I1	I3		
T600	I2	I3		
T700	I1	I3		
T800	I1	I2	I3	I5
T900	I1	I2	I3	

Click to add text

Find all frequent itemsets with minimum support count of 2.

Find all association rules with minimum confidence of 70%

Example 1 – FP Growth Tree Construction

Example 1- Mining the FP growth tree

Example 1- Mining the FP growth tree

Associative rules

- **Advantages:**
  - Easy to interpret: The results are presented in the form of a rule that is easily understood.
  - Actionable: It is possible to perform some sort of action based on the rule.
  - Large data sets: It is possible to use this technique with large numbers of observations.
- **Limitations :**
  - Only categorical variables: or to convert continuous variable to categorical variables.
  - Time-consuming: There are ways to make the analysis run faster but they often compromise the final results.
  - Rule prioritization: The method can generate many rules that must be prioritized and interpreted.

Summary
measures of pattern interestingness

Assume that X and Y are items being considered. Let

1. N be the total number of baskets.
2. N_{XY} represent the number of baskets in which X and Y appear together.
3. N_X represent the number of baskets in which X appears.
4. N_Y represent the number of baskets in which Y appears.

$$\text{Support}(X, Y) = \frac{N_{XY}}{N}$$

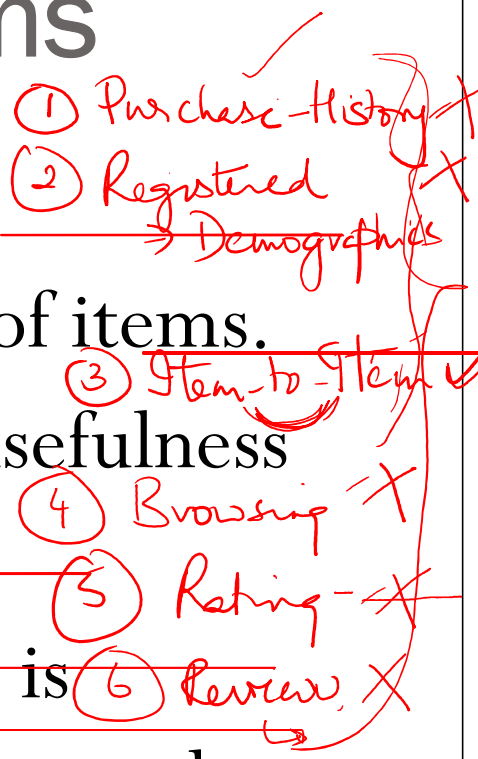
$$\text{Confidence}(X \rightarrow Y) = P(Y \mid X) = \frac{N_{XY}}{N_X}$$

where $P(Y|X)$ is the conditional probability of Y given X.

$$\text{Lift} = \frac{\text{Support}(X,Y)}{\text{Support}(X) \times \text{Support}(Y)} = \frac{N_{XY}}{N_X N_Y}$$

Goal of Recommender Systems

- Personalization of shopping experience
- Considers a set, *C*, of users and a set, *S*, of items.
- *u* is a utility function that measures the usefulness of an item, *s*, to a user, *c*.
- The utility is represented by a rating and is initially defined only for items previously rated by users.
- False positive are irritation to customer
- false negative is loss of revenue



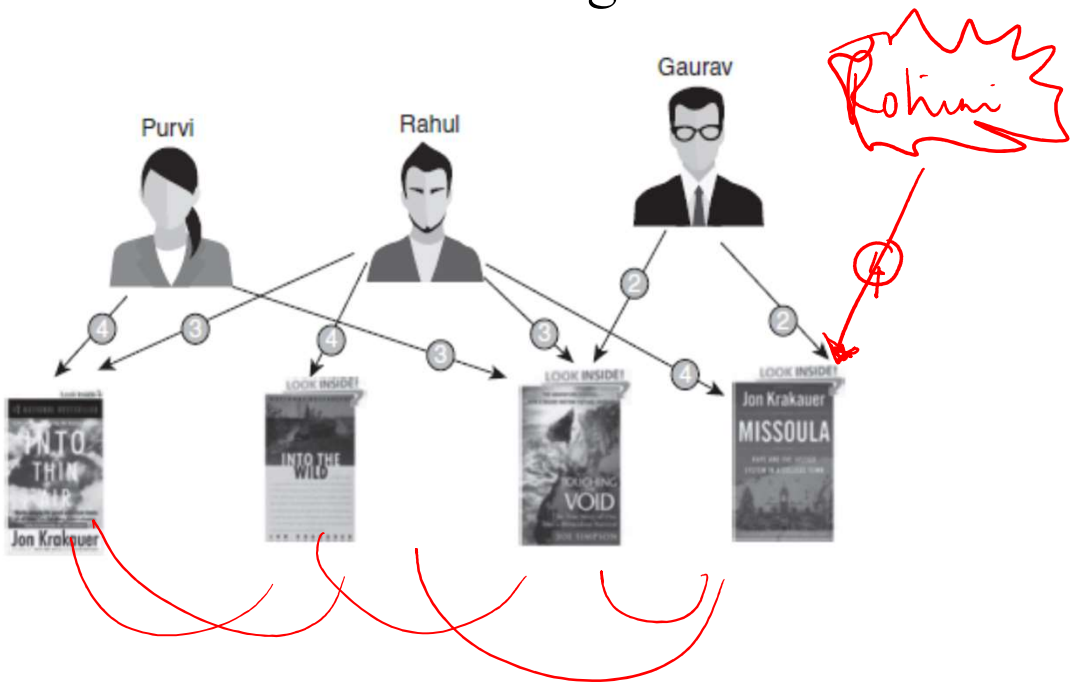
COLD START PROBLEM

Collaborative Filtering

- **Association rules don't take into consideration, the preference or rating given by the customer**
- **Collaborative Filtering**
  - **Take into consideration, what customers bought and how they rated them**
  - **Approaches are**
    - **Content-based approach or Item based similarity**
      - recommends items that are similar to items the user preferred or queried in the past.
    - **Collaborative approach or User based similarity**
      - consider a user's social environment.
      - It recommends items based on the opinions of other customers who have similar tastes or preferences as the user.

Collaborative Filtering

- Notion of Similarity
- If A and B have purchased same products and rated them similarly on common similar scale
- Implication
  - If A has bought a new product and/or rated high recommend the product to B
- **Cold start problem**



Distance Measures for numeric attributes

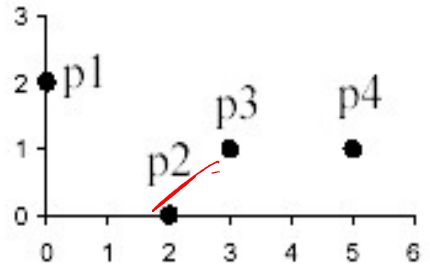
- Euclidean Distance
  - $X_1 = (x_{11}, x_{12}, \ldots x_{1n})$ and $X_2 = (x_{21}, x_{22} .. x_{2n})$ is
  - $dist(X_1, X_2) = \sqrt{\Sigma_{1=1}^{n} (x_{1i} - x_{2i})^2}$

- Manhattan or City Block Distance
  - $dist(i , j) = | x_{i1} - x_{j1} | + | x_{i2} - x_{j2} | + \ldots | x_{in} - x_{jn} |$

- Minkowski Distance
  - $dist(i,j) = (| x_{i1} - x_{j1} |^p + | x_{i2} - x_{j2} |^p + \ldots | x_{in} - x_{jn} |^p)^{1/p}$

If $p = 2$ then Euclidean distance

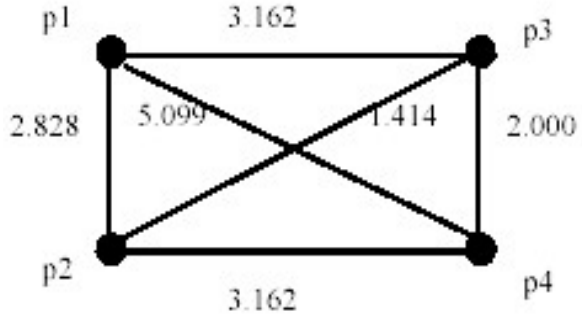
Distance measures for partitioning based clustering



a) points

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

b) data matrix



d) proximity graph

	p1	p2	p3	p4
p1	0.000	2.828	3.162	5.099
p2	2.828	0.000	1.414	3.162
p3	3.162	1.414	0.000	2.000
p4	5.099	3.162	2.000	0.000

c) proximity matrix (Similarity), distance

Four points, their proximity graph, and their corresponding data and proximity (distance) matrices.

Data Matrix

Table 6.4. Three observations with values for five variables

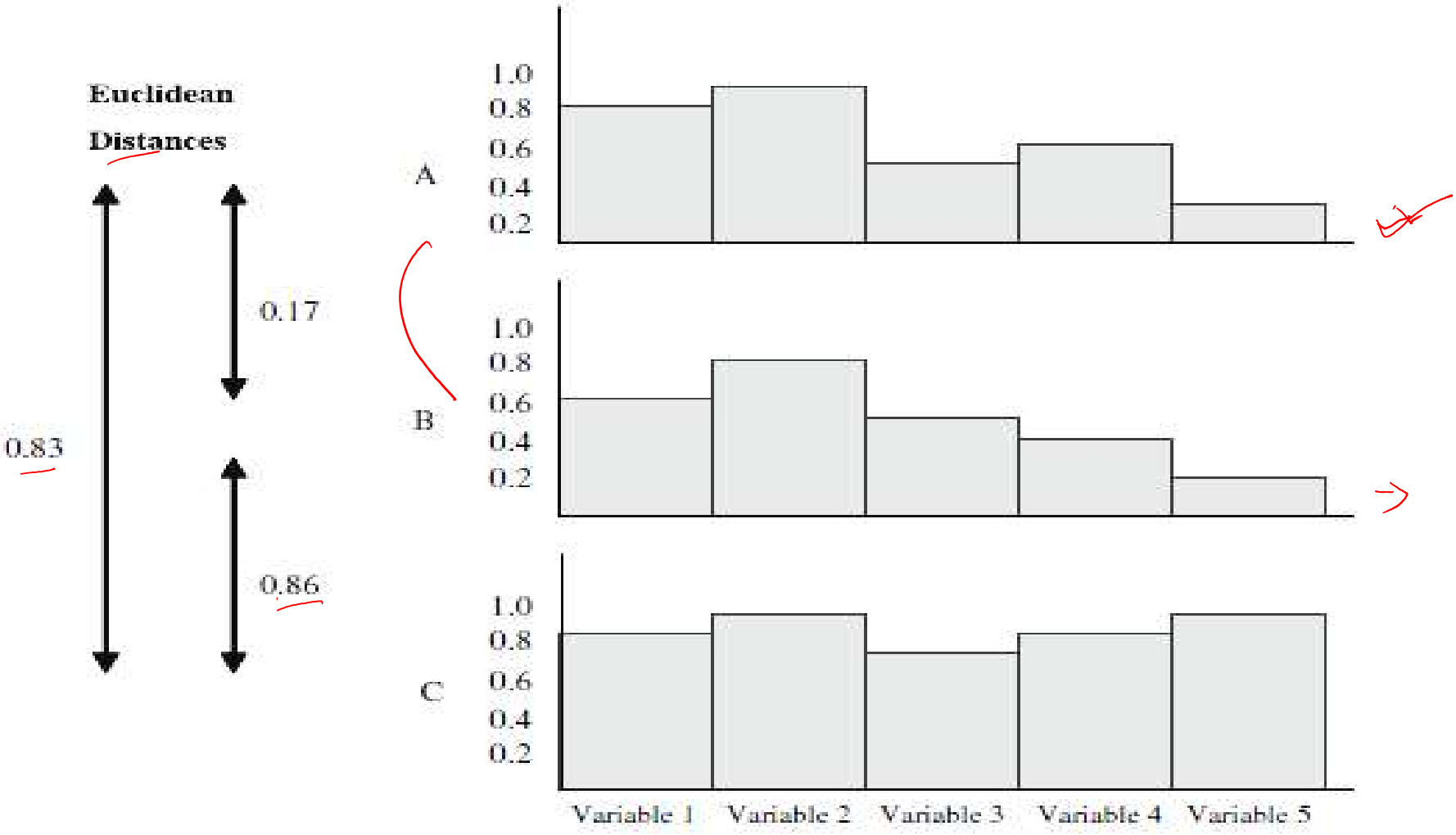
Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

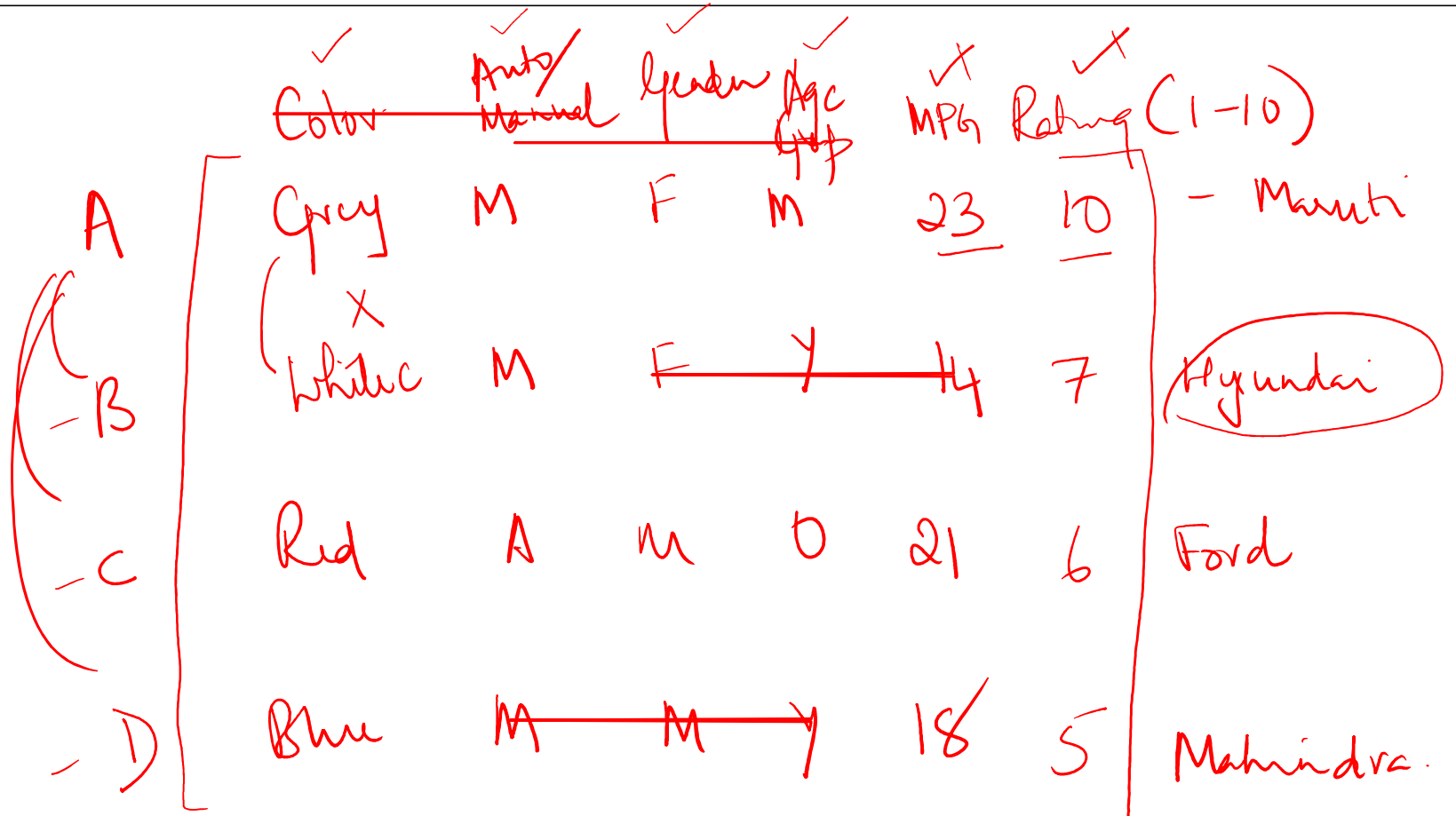
- Three observations
- 5 variables

Manhattan

	A	B	C
A	0	0.3	1.5
B		0	1.6
C			0

Distance between observations with 5 variables





Simple Matching A & B = $\frac{4-(1+1)}{4} = 2/4 = 0.5 \Rightarrow$ ✓

A & C = $4/4 = 1 \Rightarrow$

A & D = $3/4 = 0.75$

Similarity measures between binary vectors

These measures are referred to as similarity coefficients and typically have values between 0 (not at all similar) and 1 (completely similar). The comparison of two binary vectors, ***a*** and ***b***, leads to four quantities:

N_{01} = the number of positions where ***a*** was 0 and ***b*** was 1
N_{10} = the number of positions where ***a*** was 1 and ***b*** was 0
N_{00} = the number of positions where ***a*** was 0 and ***b*** was 0
N_{11} = the number of positions where ***a*** was 1 and ***b*** was 1

Two common similarity coefficients between binary vectors are the simple matching coefficient (SMC) and the Jaccard coefficient.

$$SMC = (N_{11} + N_{00}) / (N_{01} + N_{10} + N_{11} + N_{00})$$

$$Jaccard = N_{11} / (N_{01} + N_{10} + N_{11})$$

For the following two binary vectors, ***a*** and ***b*** we get SMC = 0.7 and Jaccard = 0.

a = 1 0 0 0 0 0 0 0 0 0
b = 0 0 0 0 0 0 1 0 0 1

Nominal Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching
  - m: # of matches, p: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: use a large number of binary variables
  - creating a new binary variable for each of the M nominal states

Cosine Similarity

$$sim(x, y) = \frac{x \cdot y}{||x|| ||y||},$$

- where $||x||$ is the Euclidean norm of vector x
- Conceptually, it is the length of the vector.

- Computes the cosine of the angle between vectors x and y.
- 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match.
- Value close to 1, the smaller the angle and the greater the match between vectors.

Cosine Similarity Example

Document Vector or Term-Frequency Vector

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Document 1 & 2

$$x^t \cdot y = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

$$||x|| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$||y|| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

$$sim(x, y) = 0.94$$

Document 1 & 4 $= \frac{2}{6.48 \times 4.36} = 0.07$

Case study 4- Amazon – recommender systems

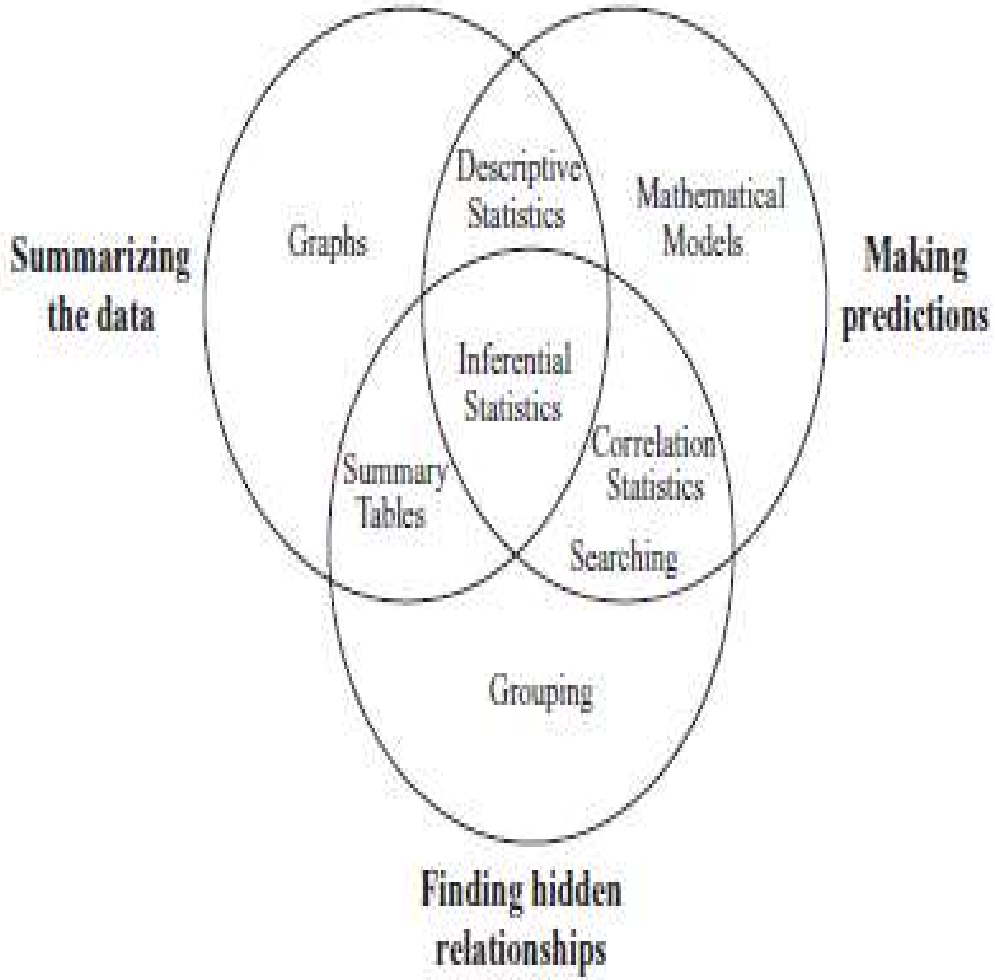
- **Challenges**
  - large retailer might have huge amounts of data, customers and catalog items.
  - Amazon.com has more than 29 million customers and several million catalog items.
  - results set to be returned in real time
  - New customers typically have extremely limited information or product ratings.
  - Older customers can have a glut of information,
  - Customer data is volatile
- **Item-to-Item Collaborative Filtering**
  - scales to massive data sets and produces high-quality recommendations in real time
  - creates the expensive similar-items table offline
  - Rather than matching the user to similar customers, item-to-item collaborative filtering matches each of the user's purchased and rated items to similar items, then combines those similar items into a recommendation list
  - This computation is very quick, depending only on the number of items the user purchased or rated.
  - Works better than search based or collaborative recommendation
- Case study

Summary of grouping methods

	Criteria	Data	Supervised/ unsupervised	Size	Time to compute	Overlapping groups
Agglomerative hierarchical clustering	Distances	Any	Unsupervised	Small	Slow	No
K-means clustering	Distances	Any	Unsupervised	Any	Faster than hierarchical methods	No
Association rules	Categorical values	Categorical	Unsupervised	Large	Dependent on parameters	Yes
Decision trees	Categorical values or ranges	Any	Supervised	Large	Dependent on variables used	No overlapping in terminal nodes

Introduction

- is the science of examining raw data with the purpose of drawing conclusions about that information.
- To make better business decisions
- in the sciences to verify or disprove existing models or theories.
- Consists of
  - exploratory data analysis (EDA)
  - confirmatory data analysis (CDA)
  - Qualitative data analysis (QDA)



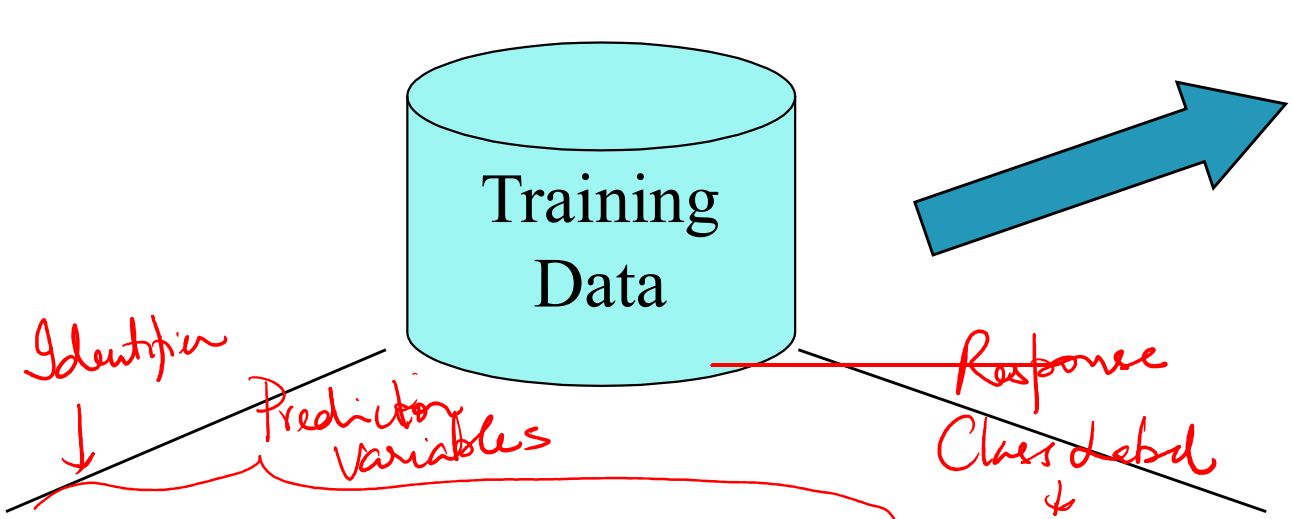
Data Analysis tasks & methods

Introduction - Classification vs. Prediction

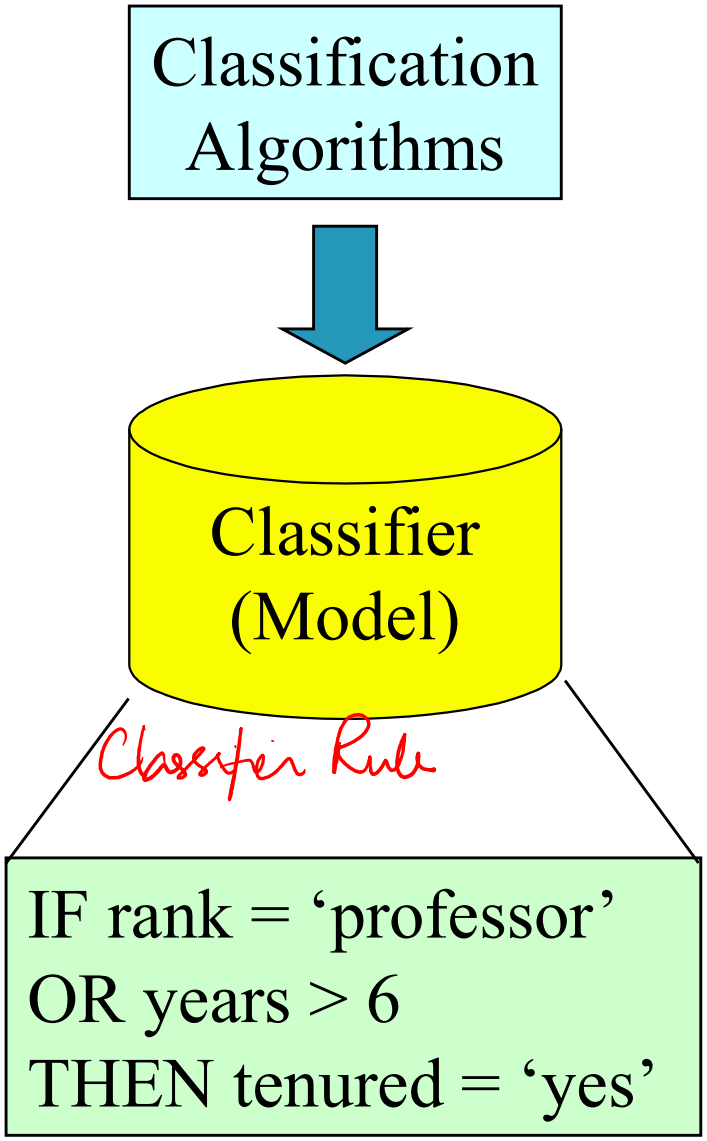
- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
  - Ex – Class labels – buysComputer= "Yes" & "No"
  - Can result in a descriptive model and/or predictive model
- Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values
  - Ex – predicting income on age & qualification
- Typical applications
  - Credit approval
  - Target marketing
  - Medical diagnosis
  - Fraud detection

Phases in Classification

Phase (1): Model Construction or Training Phase

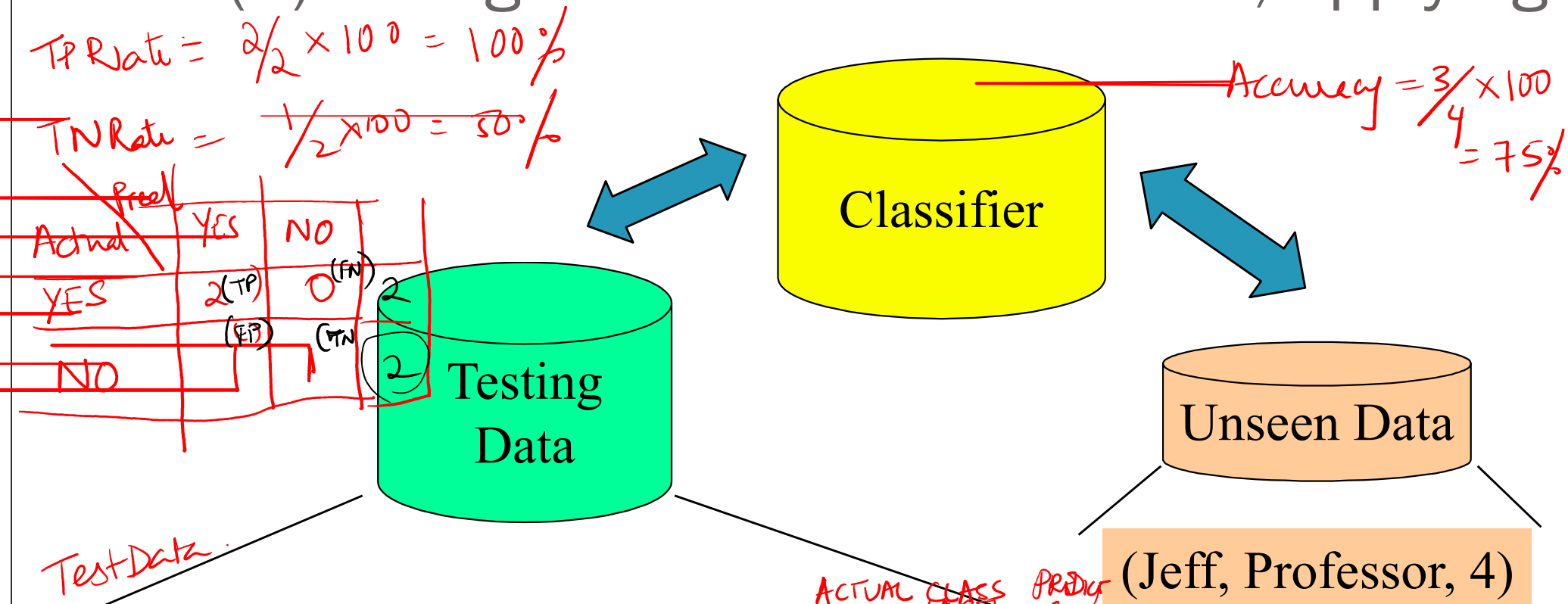


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no



Phases in classification

Phase (2): Using the Model in Prediction, Applying



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

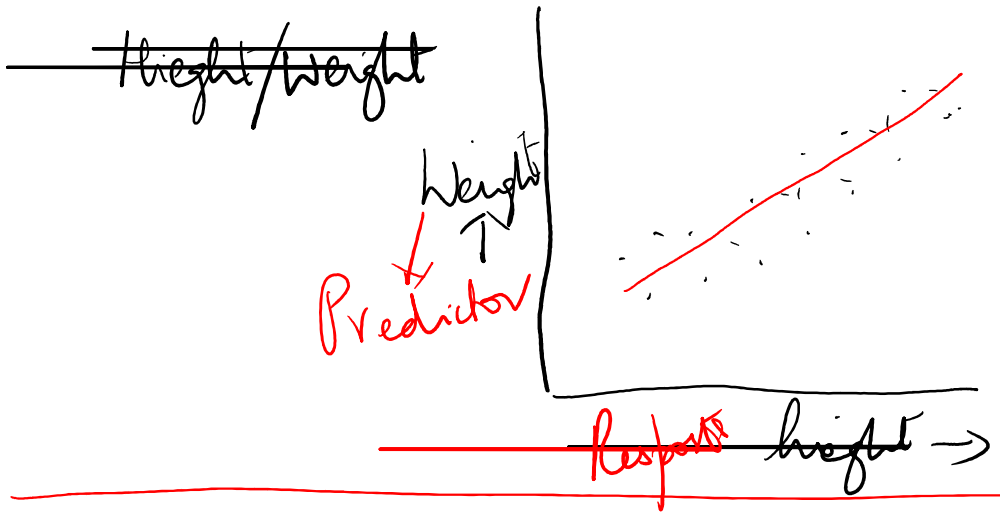
Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Table 7.3. Different classification and regression methods

Classification	~~Regression~~ Prediction
Classification trees	Regression trees
k-Nearest Neighbors	k-Nearest Neighbors
Logistic regression	Linear regressions ⇒
Naïve Bayes classifiers	Neural networks
Neural networks	Nonlinear regression
Rule-based classifiers	Partial least squares
Support vector machines	Support vector machines

Multiple Linear Regression

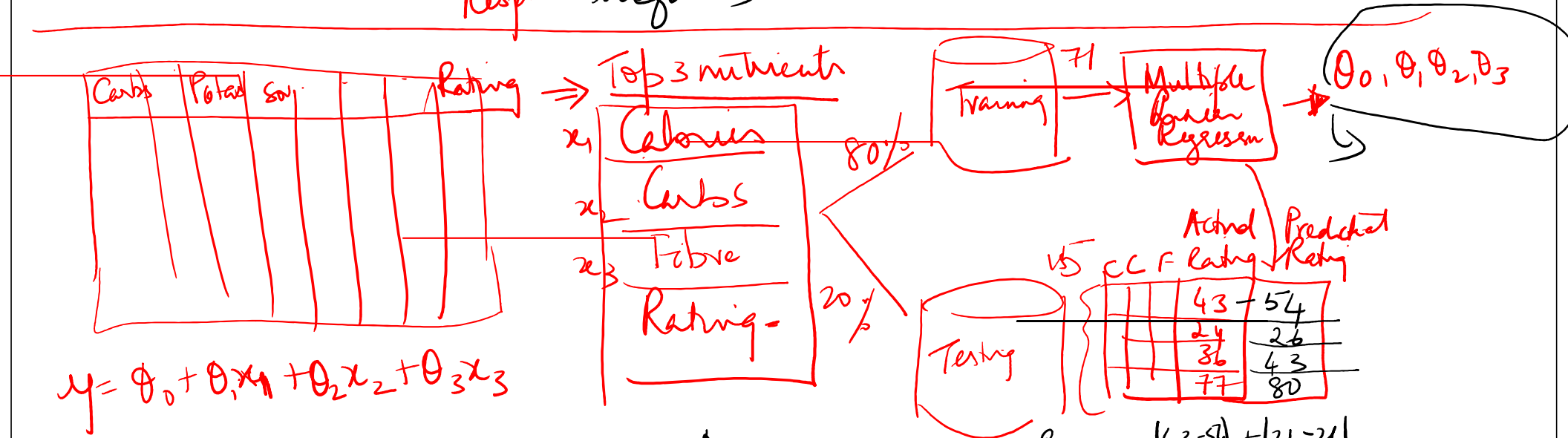


$r = 0.96$

$y = m * x + c$ (slope, y intercept)

$y = \theta_0 + \theta_1 * x$

$y = \beta_0 + \beta_1 * x$



ACCURACY MEASURES? ABS Error $|43-54| + |24-26|$ MSE or RMSE $\sqrt{(\quad)^2 +}$

References

- **Text book**
  - Glenn J. Myatt, Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, John Wiley, November 2006.
- **References**
  - I.H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley, May, 2005.