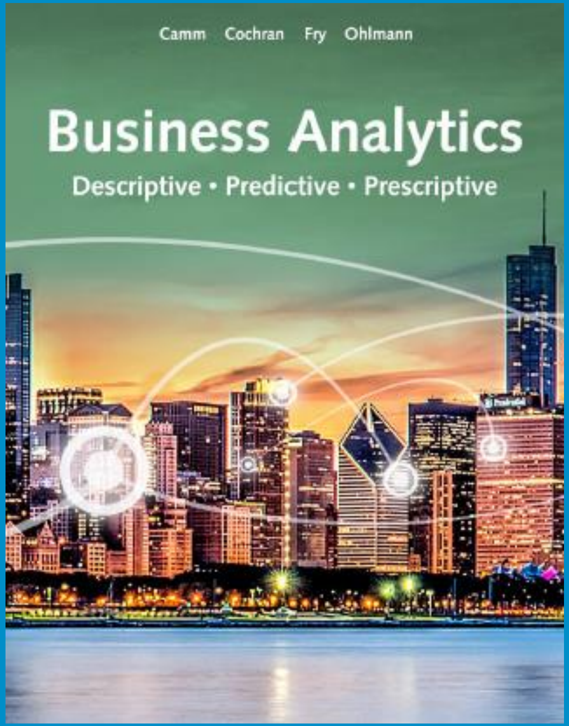
MENG 3065 - MODULE 5

Unsupervised Learning (Business Analytics – Chapter 3 - Descriptive Data Mining)



These slides has been extracted, modified and updated from original slides of :

HUMBER

WE ARE HUMBER

Outline

- What is Unsupervised Learning?
- Types of Unsupervised Learning
- Clustering Algorithms
- Dimensionality Reduction
- Applications of Unsupervised Learning
- Challenges and Considerations
- Conclusion

Remember: Supervised Learning

- **Goal:** Predict a single “target” or “outcome” variable
- Training data, where **target value is known**
- Score to data where value is not known
- **Methods:** Classification and Prediction
- **Examples:** Purchase/no purchase, fraud/no fraud, creditworthy/not creditworthy…

Unsupervised Learning

- **Goal:** Segment data into meaningful segments; detect patterns
- There is **no target (outcome) variable** to predict or classify
- **Methods:** Clustering, Association rules, data reduction & exploration, visualization
- **Example:** “If X was purchased, Y was also purchased”

Unsupervised Learning (Cont’d)

- Unsupervised learning is a machine learning paradigm where the model learns patterns and structures in data without explicit supervision or labeled targets.
- In other words, it's about finding hidden patterns, grouping similar data points, or reducing the complexity of data without predefined categories.

Class Activity

Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.

1. Deciding whether to issue a loan to an applicant based on demographic and financial data
2. In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions (leveraging patterns and similarities in customer purchase history).
3. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.
4. Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.
5. Estimating the repair time required for an aircraft based on a trouble ticket.
6. Automated sorting of mail by zip code scanning.

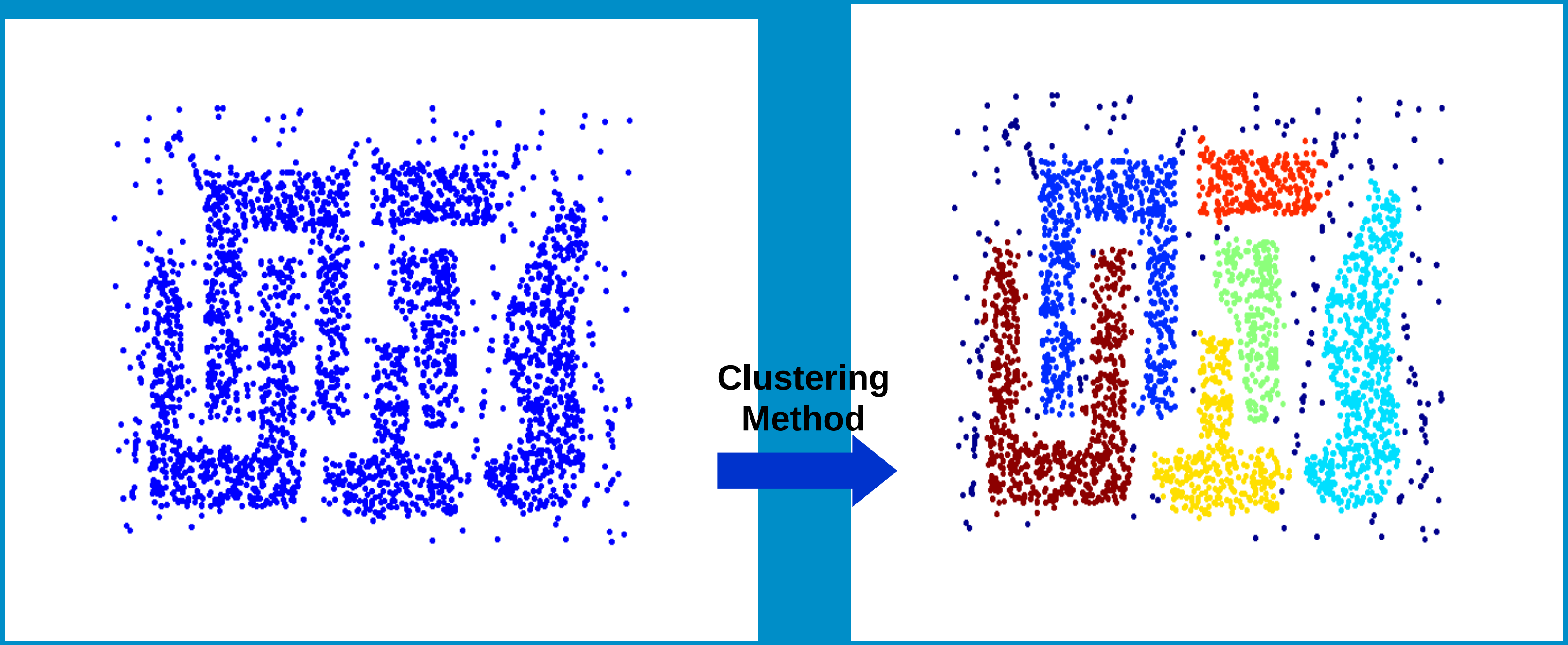
Types of Unsupervised Learning

- **Clustering**: Grouping similar data points together.
- **Dimensionality Reduction**: Reducing the number of features while preserving meaningful information.
- **Generative Models**: Learning the underlying data distribution for generating new data samples.

Clustering

Clustering

- **Cluster detection:** Measure similarity among data objects and group them into clusters accordingly

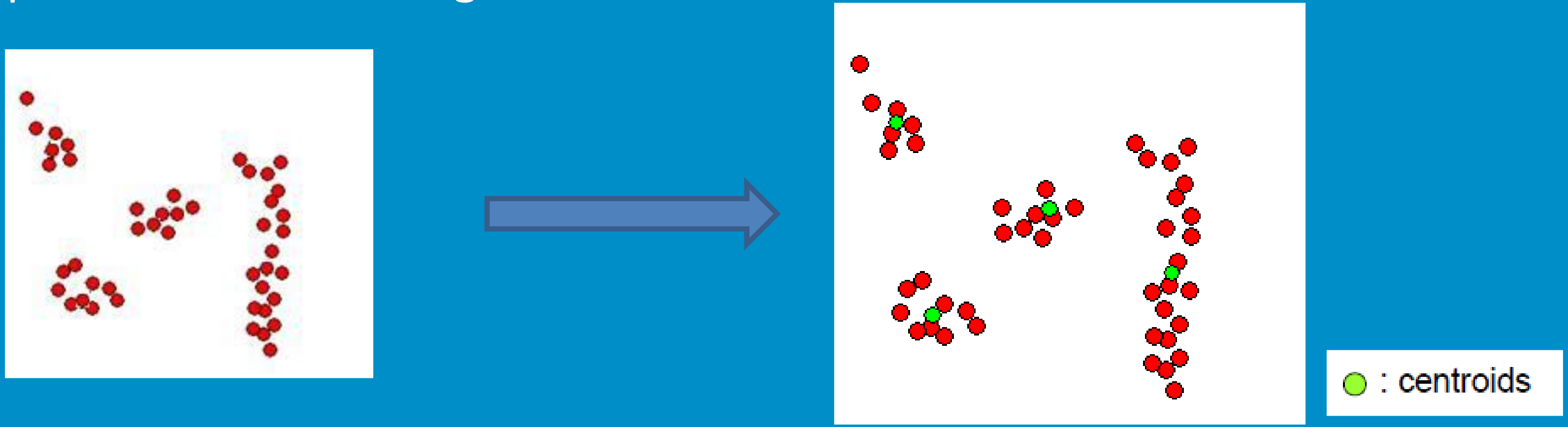


Input data points

Cluster Memberships of Data Points

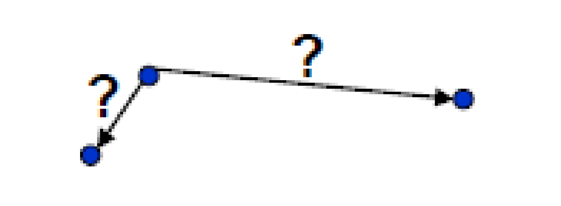
Clustering Definition

- **What is cluster detection?**
- **Cluster:** a group of objects known as **members**
- The center of a cluster is known as the **centroid**
- Members of a cluster are **similar to each other**
- Members of different clusters are different
- **Clustering** is a process of discovering clusters

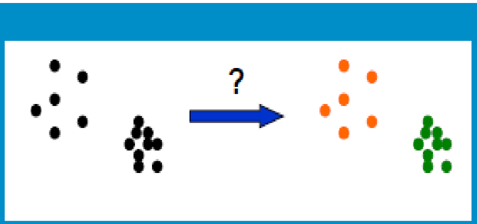


Clustering Elements

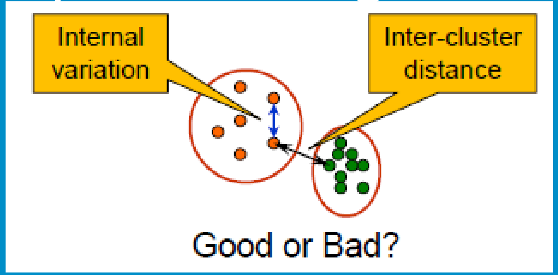
- **A sensible measure** for similarity
  - e.g. Euclidean distance



- **An effective and efficient clustering algorithm**
  - e.g. K-means



- **A goodness-of-fit** function for evaluating the quality of resulting clusters
  - e.g. Error Sum of Squares (SSE)



Clustering Algorithms

- **Connectivity models**: based on distance connectivity.
  - e.g. hierarchical clustering
- **Centroid models:** represents each cluster by a single mean vector.
  - e.g. the k-means algorithm
- **Distribution models:** clusters are modeled using statistical distributions
  - e.g. multivariate normal distributions used by the expectation-maximization algorithm.
- **Density models:** defines clusters as connected dense regions in the data space.
  - e.g. DBSCAN and OPTICS

Clustering Algorithms Examples

- K-Means: Divides data into 'k' clusters based on similarity.
- Hierarchical Clustering: Builds a hierarchy of clusters, often represented as a tree (dendrogram).
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Clusters based on data density.
- Gaussian Mixture Models (GMM): Assumes data is generated from a mixture of Gaussian distributions.

Clustering Algorithms Examples-Extended

- **K-Means**: Divides data into 'k' clusters based on similarity. Used in:
  - Customer segmentation: Grouping customers with similar buying behavior.
  - Image compression: Reducing image size by clustering similar pixel values.
- **Hierarchical Clustering**: Builds a hierarchy of clusters, often represented as a tree (dendrogram). Used in:
  - Biology: Taxonomy and evolutionary relationships.
  - Document clustering: Organizing documents into topics or themes.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Clusters based on data density. Used in:
  - Anomaly detection: Identifying outliers in sensor data.
  - Geospatial data: Identifying hotspots in disease spread.
- **Gaussian Mixture Models (GMM)**: Assumes data is generated from a mixture of Gaussian distributions. Used in:
  - Image segmentation: Dividing an image into regions with similar pixel statistics.
  - Speech recognition: Modeling phonemes with Gaussian distributions.

Cluster Analysis

- Measuring Similarity Between Observations
- Hierarchical Clustering
- *k*-Means Clustering
- Hierarchical Clustering versus *k*-Means Clustering

Cluster Analysis (Slide 1 of 21)

- Goal of clustering is to segment observations into similar groups based on observed variables.
- Can be employed during the data-preparation step to identify variables or observations that can be aggregated or removed from consideration.
- Commonly used in marketing to divide customers into different homogenous groups; known as **market segmentation**.
- Used to identify outliers.

Cluster Analysis (Slide 2 of 21)

- Clustering methods:
  - Bottom-up **hierarchical clustering** starts with each observation belonging to its own cluster and then sequentially merges the most similar clusters to create a series of nested clusters.
  - ***k*-means clustering** assigns each observation to one of *k* clusters in a manner such that the observations assigned to the same cluster are as similar as possible.
- Both methods depend on how two observations are similar—hence, we have to measure similarity between observations.

Cluster Analysis (Slide 3 of 21)

Measuring Similarity Between Observations:

When observations include numeric variables, **Euclidean distance** is the most common method to measure dissimilarity between observations.

Let observations $u = (u_1, u_2, \ldots, u_q)$ and $v = (v_1, v_2, \ldots, v_q)$ each comprise measurements of *q* variables.

The Euclidean distance between observations *u* and *v* is:

$$d_{uv} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_q - v_q)^2}$$

Cluster Analysis (Slide 4 of 21)

Measuring Similarity Between Observations:

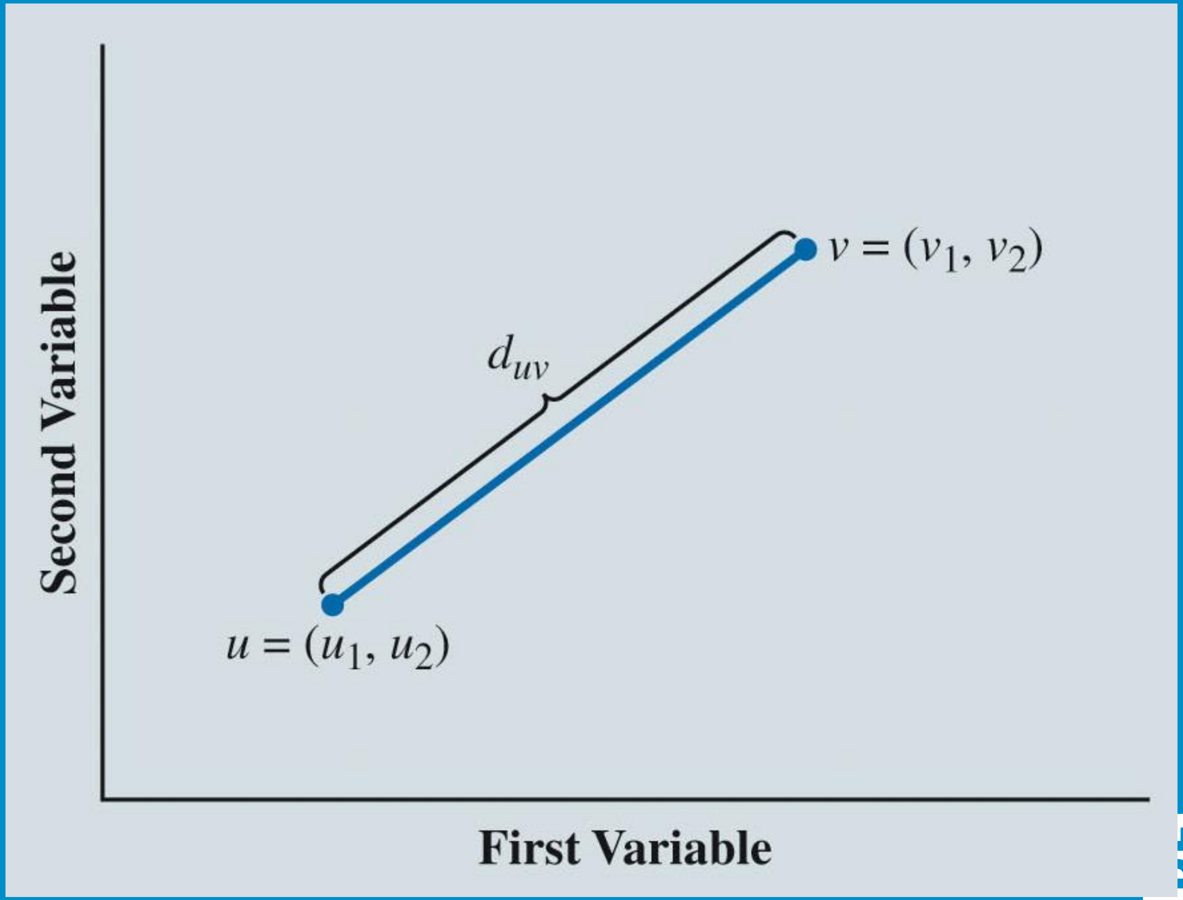
Illustration:

- KTC is a financial advising company that provides personalized financial advice to its clients.
- KTC would like to segment its customers into several groups (or clusters) so that the customers within a group are similar and dissimilar with respect to key characteristics.
- For each customer, KTC has an observation of seven variables: Age, Female, Income, Married, Children, Car Loan, Mortgage.

  Example: The observation *u* = (61, 0, 57881, 1, 2, 0, 0) corresponds to a 61-year-old male with an annual income of $57,881, married with two children, but no car loan and no mortgage.

Cluster Analysis (Slide 5 of 21)

Figure 5.1: Euclidean Distance



Euclidean distance becomes smaller as a pair of observations become more similar with respect to their variable values.

Cluster Analysis (Slide 6 of 21)

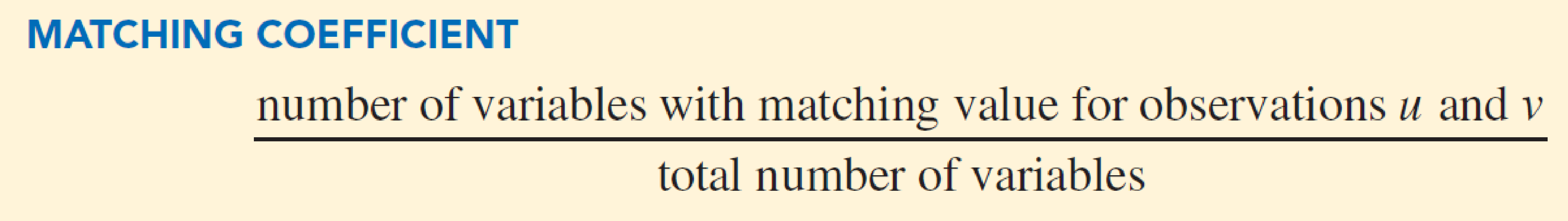
- Euclidean distance is highly influenced by the scale on which variables are measured:
  - Common to standardize the units of each variable j of each observation *u.*

Example: u_j, the value of variable j, in observation u, is replaced with its z-score z_j.

- The conversion to z-scores also makes it easier to identify outlier measurements, which can distort the Euclidean distance between observations.

Cluster Analysis (Slide 7 of 21)

- When clustering observations solely on the basis of categorical variables encoded as 0–1, a better measure of similarity between two observations can be achieved by counting the number of variables with matching values.
- The simplest overlap measure is called the **matching coefficient** and is computed as:

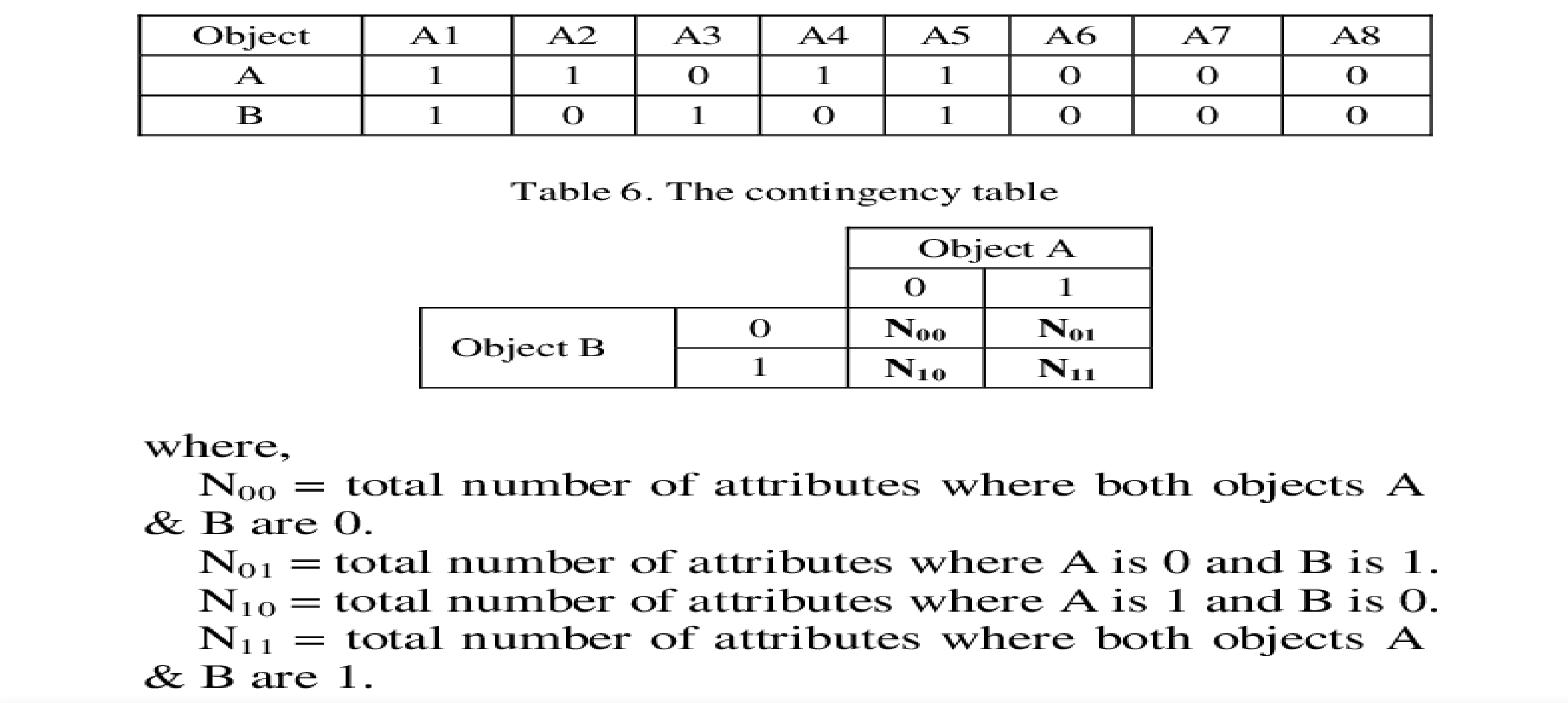


MATCHING COEFFICIENT

$$\frac{\text{number of variables with matching value for observations } u \text{ and } v}{\text{total number of variables}}$$

Example – Matching Coefficient

Table: Two sample objects described using binary attributes



Object	A1	A2	A3	A4	A5	A6	A7	A8
A	1	1	0	1	1	0	0	0
B	1	0	1	0	1	0	0	0

Table 6. The contingency table

		Object A	
		0	1
Object B	0	N_{00}	N_{01}
	1	N_{10}	N_{11}

where,

N_{00} = total number of attributes where both objects A & B are 0.

N_{01} = total number of attributes where A is 0 and B is 1.

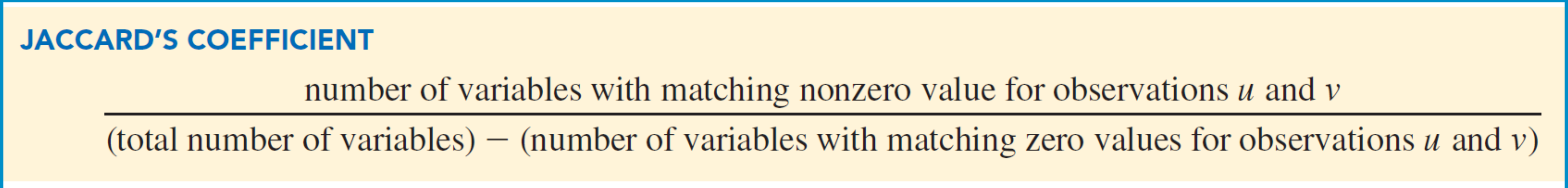
N_{10} = total number of attributes where A is 1 and B is 0.

N_{11} = total number of attributes where both objects A & B are 1.

Cluster Analysis (Slide 8 of 21)

- A weakness of the matching coefficient is that if two observations both have a 0 entry for a categorical variable, this is counted as a sign of similarity between the two observations.

- To avoid misstating similarity due to the absence of a feature, a similarity measure called **Jaccard's coefficient** does not count matching zero entries and is computer as:



JACCARD'S COEFFICIENT

$$\frac{\text{number of variables with matching nonzero value for observations } u \text{ and } v}{(\text{total number of variables}) - (\text{number of variables with matching zero values for observations } u \text{ and } v)}$$

Cluster Analysis (Slide 9 of 21)

Table 5.1: Comparison of Similarity Matrixes for Observations with Binary Variables

Observation	Female	Married	Loan	Mortgage
1	1	0	0	0
2	0	1	1	1
3	1	1	1	0
4	1	1	0	0
5	1	1	0	0

Cluster Analysis (Slide 10 of 21)

Table 5.1: Comparison of Similarity Matrixes for Observations with Binary Variables (cont.)

- Similarity Matrix Based on Matching Coefficient:

Observation	1	2	3	4	5
1	1				
2	0	1			
3	0.5	0.5	1		
4	0.75	0.25	0.75	1	
5	0.75	0.25	0.75	1	1

Cluster Analysis (Slide 11 of 21)

Table 5.1: Comparison of Similarity Matrixes for Observations with Binary Variables (cont.)

- Similarity Matrix Based on Jaccard's Coefficient:

Observation	1	2	3	4	5
1	1				
2	0	1			
3	0.333	0.5	1		
4	0.5	0.25	0.667	1	
5	0.5	0.25	0.667	1	1

Cluster Analysis (Slide 12 of 21)

Hierarchical Clustering:

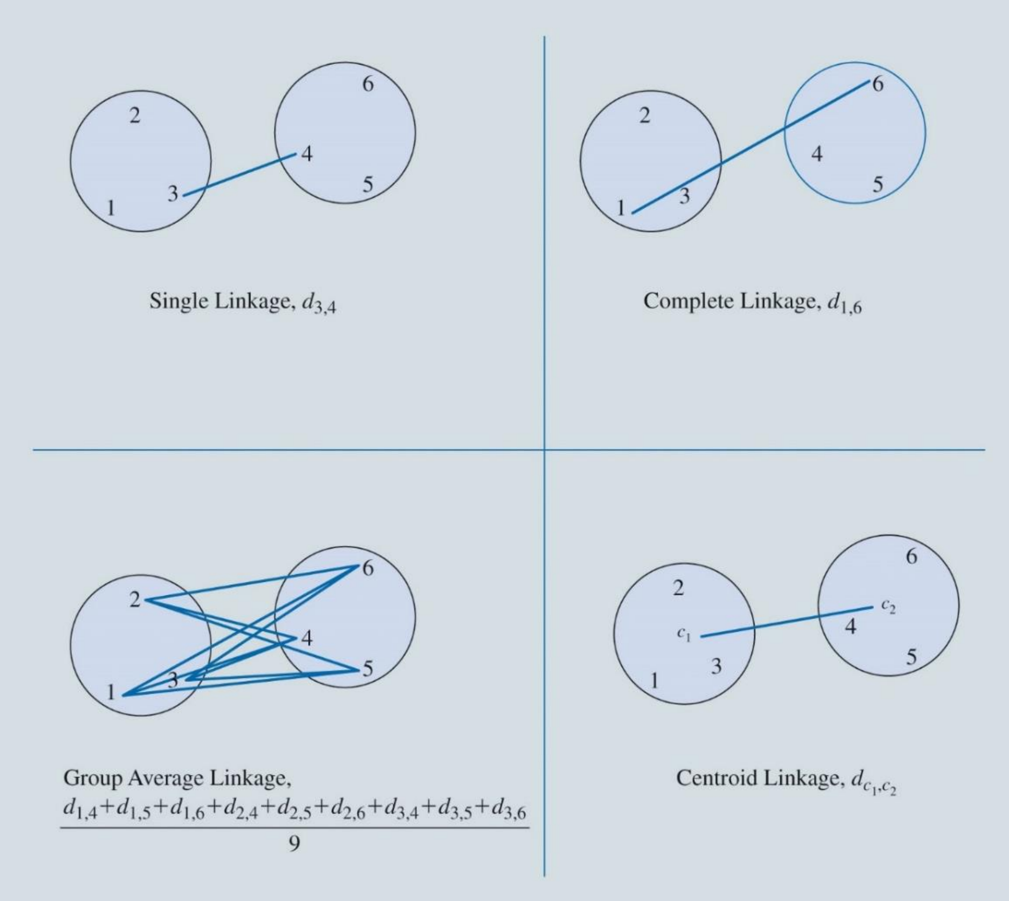
- Determines the similarity of two clusters by considering the similarity between the observations composing either cluster.
- Starts with each observation in its own cluster and then iteratively combines the two clusters that are the most similar into a single cluster.
- Given a way to measure similarity between observations, there are several clustering method alternatives for comparing observations in two clusters to obtain a cluster similarity measure:
  - Single linkage.
  - Complete linkage.
  - Group average linkage.
  - Median linkage.
  - Centroid linkage.

Cluster Analysis (Slide 13 of 21)

- **Single linkage**: The similarity between two clusters is defined by the similarity of the pair of observations (one from each cluster) that are the most similar.
- **Complete linkage**: This clustering method defines the similarity between two clusters as the similarity of the pair of observations (one from each cluster) that are the most different.
- **Group Average linkage**: Defines the similarity between two clusters to be the average similarity computed over all pairs of observations between the two clusters.
- **Median linkage**: Analogous to group average linkage except that it uses the median of the similarities computed between all pairs of observations between the two clusters.
- **Centroid linkage** uses the averaging concept of cluster centroids to define between-cluster similarity.

Cluster Analysis (Slide 14 of 21)

Figure 5.2: Measuring Similarity Between Clusters

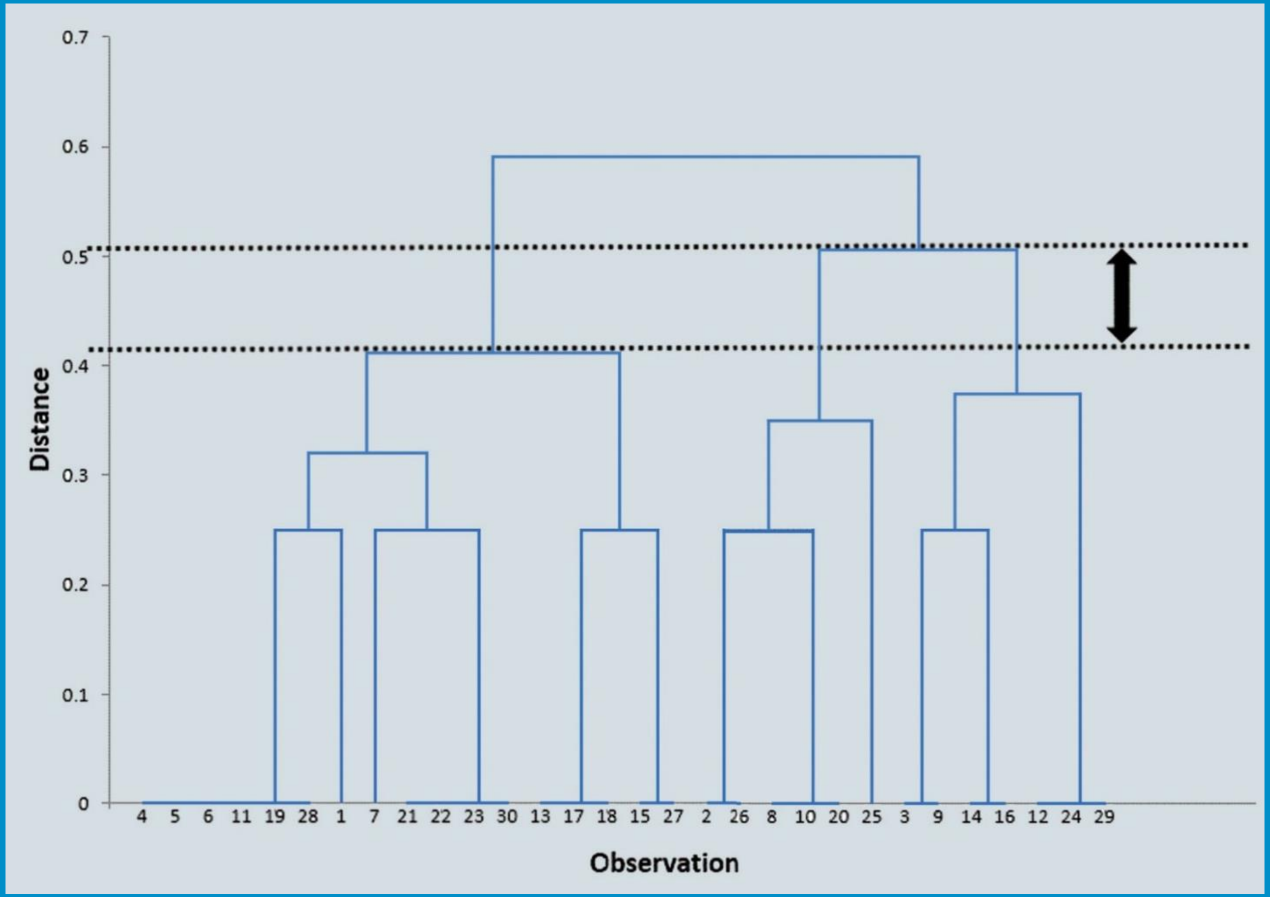

1
2
3
4
5
6
Single Linkage, $d_{3,4}$
1
2
3
4
5
6
Complete Linkage, $d_{1,6}$
1
2
3
4
5
6
Group Average Linkage,
$\frac{d_{1,4}+d_{1,5}+d_{1,6}+d_{2,4}+d_{2,5}+d_{2,6}+d_{3,4}+d_{3,5}+d_{3,6}}{9}$
1
2
3
c_1
4
5
6
c_2
Centroid Linkage, d_{c_1,c_2}

Cluster Analysis (Slide 15 of 21)

- **In hierarchical clustering, we have different methods to decide how to merge clusters**
  - **Ward's method** merges two clusters such that the dissimilarity of the observations with the resulting single cluster increases as little as possible.
  - When **McQuitty's method** considers merging two clusters A and B, the dissimilarity of the resulting cluster AB to any other cluster C is calculated as: ((dissimilarity between A and C) + (dissimilarity between B and C)) divided by 2).
- A **dendrogram** is a chart that depicts the set of nested clusters resulting at each step of aggregation.

Cluster Analysis (Slide 16 of 21)

Figure 5.3: Dendrogram for KTC Using Matching Coefficients and Group Average Linkage



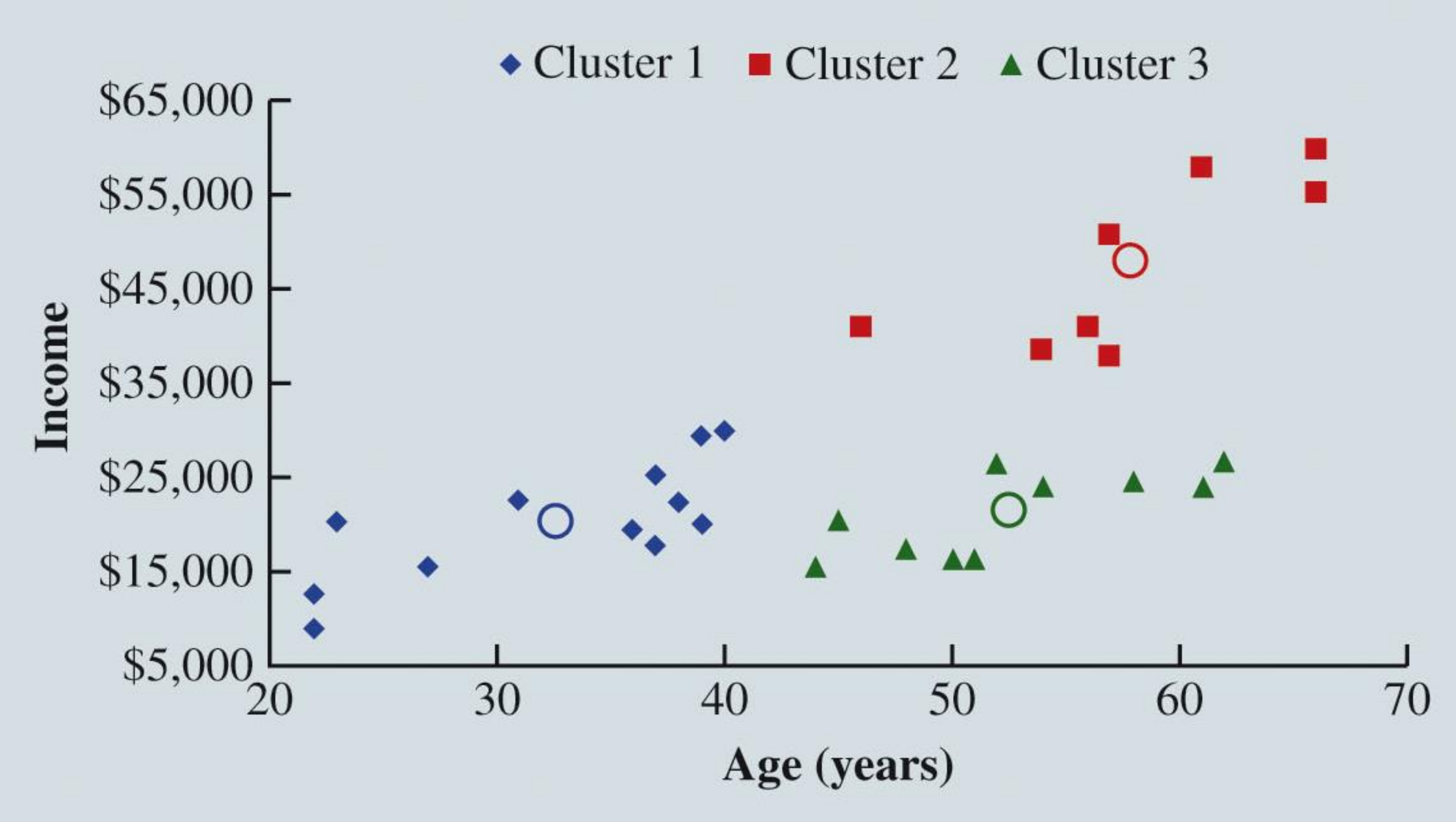
Cluster Analysis (Slide 17 of 21)

k-Means Clustering:

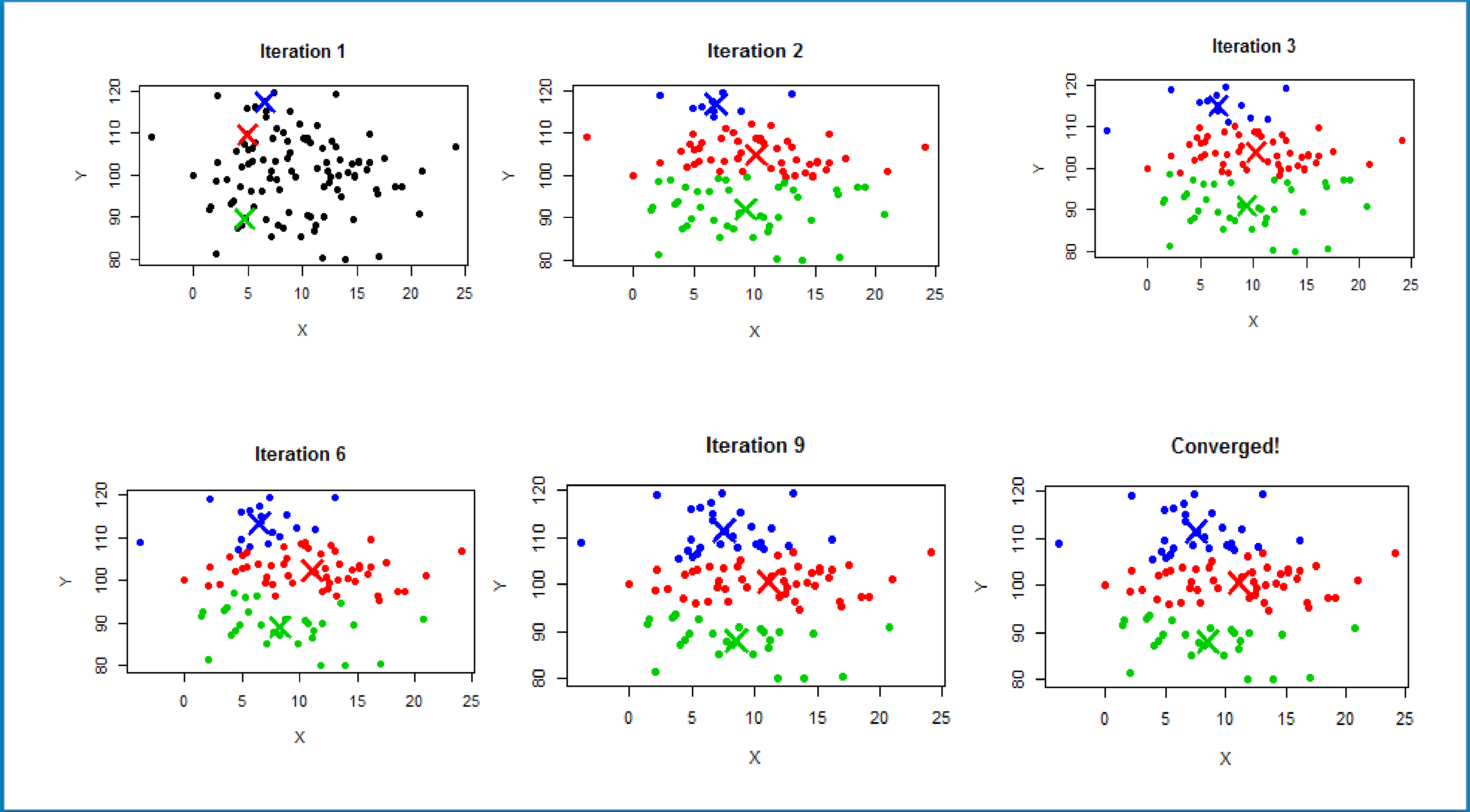
- Given a value of *k*, the *k*-means algorithm randomly assigns each observation to one of the *k* clusters.
- After all observations have been assigned to a cluster, the resulting cluster centroids are calculated.
- Using the updated cluster centroids, all observations are reassigned to the cluster with the closest centroid.
- The algorithm repeats this process (calculate cluster centroid, assign observation to cluster with nearest centroid) until there is no change in the clusters or a specified maximum number of iterations is reached

Cluster Analysis (Slide 18 of 21)

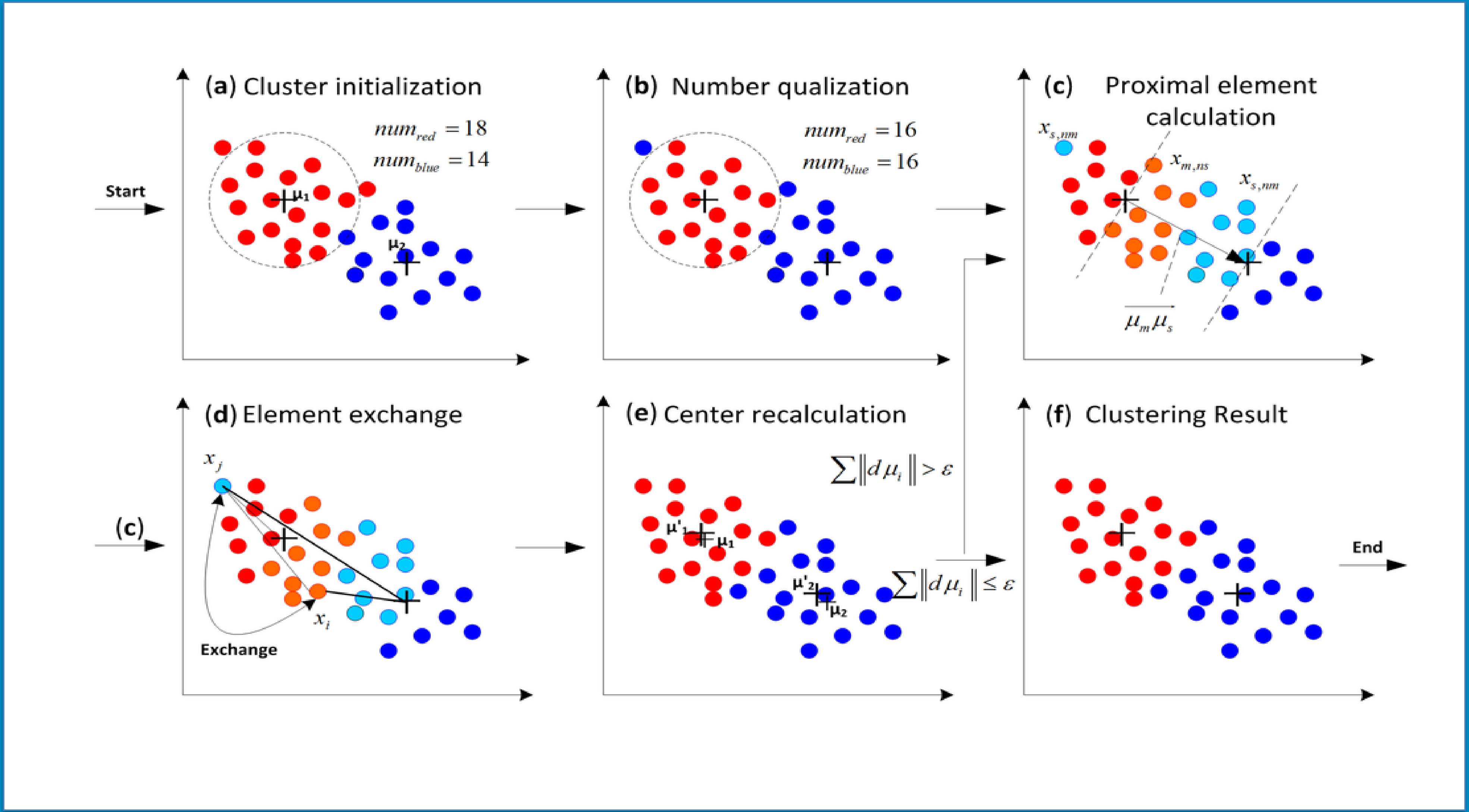
Figure 5.4: Clustering Observations by Age and Income Using *k*-Means Clustering with *k* = 3



K-Means Iteration



K-means iterations

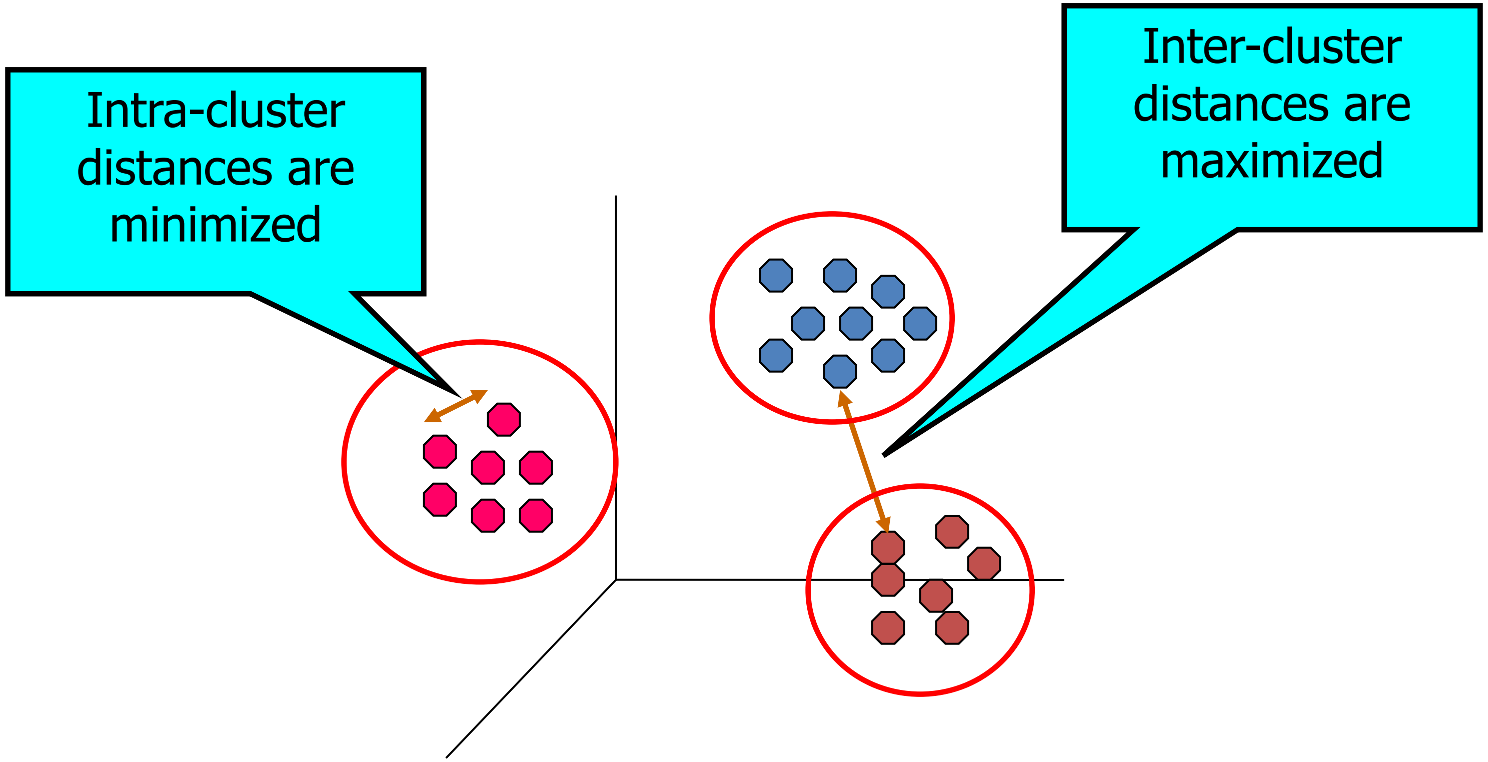


WE ARE HUMBER

Determining the number of Clusters K

- With the preceding algorithm, K clusters can be defined in a given dataset, but what value of K should be selected?
- The value of K can be chosen based on a reasonable guess or some predefined requirements,
- There is a coefficient that could be computed to determines a reasonable optimal of K which is called Within Sum Square (WSS),
- WSS is the sum of the squares of the distance between each data point and the closest centroid

Cluster Analysis – illustration


Intra-cluster distances are minimized
Inter-cluster distances are maximized

Cluster Analysis (Slide 19 of 21)

Table 5.2: Average Distances Within Clusters (Intra-cluster)

	No. of Observations	Average Distance Between Observations in Cluster
Cluster 1	12	0.622
Cluster 2	8	0.739
Cluster 3	10	0.520

Cluster Analysis (Slide 20 of 21)

Table 5.3: Distances Between Cluster Centroids (Inter-cluster)

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0	2.784	1.529
Cluster 2	2.784	0	1.964
Cluster 3	1.529	1.964	0

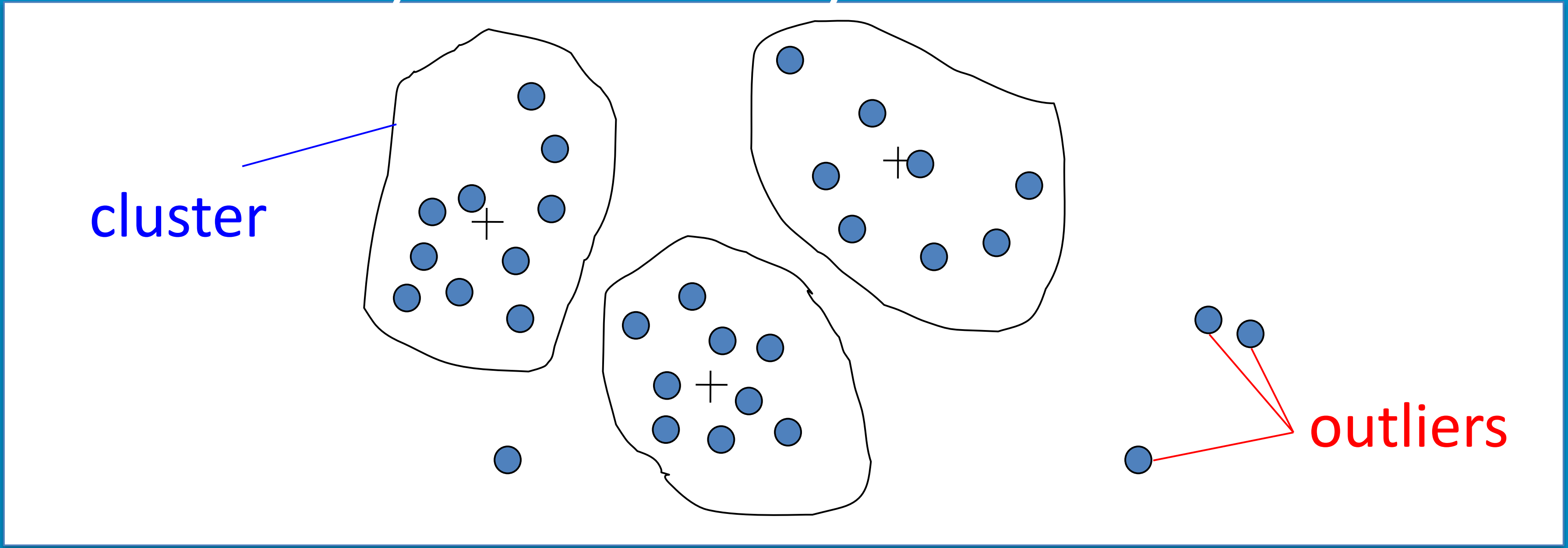
Cluster Analysis (Slide 21 of 21)

Hierarchical Clustering versus *k*-Means Clustering

Hierarchical Clustering	***k*-Means Clustering**
Suitable when we have a small data set (e.g., fewer than 500 observations) and want to easily examine solutions with increasing numbers of clusters.	Suitable when you know how many clusters you want and you have a larger data set (e.g., more than 500 observations).
Convenient method if you want to observe how clusters are nested.	Partitions the observations, which is appropriate if trying to summarize the data with *k* “average” observations that describe the data with the minimum amount of error.

Outliers

- **Outliers** are **objects that do not belong to any cluster** or form clusters of very small cardinality



- In some applications we are interested in discovering outliers, not clusters (outlier analysis)

Diagnostics – Evaluating the Model

- Do the clusters look separated in at least some of the plots when you do pair-wise plots of the clusters?
  - Pair-wise plots can be used when there are not many variables
- Do you have any clusters with few data points?
  - Try decreasing the value of K
- Are there splits on variables that you would expect, but don't see?
  - Try increasing the value of K
- Do any of the centroids seem too close to each other?
  - Try decreasing the value of K

WE ARE HUMBER

Cluster Evaluation - Summary

- **Principle:**
  - High-level similarity/low-level variation within a cluster
  - High-level dissimilarity between clusters
- **Measures:**
  - **Cohesion:** sum of squared errors (SSE), and sum of SSEs for all clusters (WC)
  - **Separation:** sum of distances between clusters (BC)
  - Combining the cohesion and separation, the ratio BC/WC is a good indicator of overall quality.

Dimensionality Reduction

HUMBER

WE ARE HUMBER

Dimensionality Reduction

Effective dimensionality reduction techniques:

- Principal Component Analysis (PCA): Reduces dimensionality by finding orthogonal axes of maximum variance.
- t-Distributed Stochastic Neighbor Embedding (t-SNE): Visualizes high-dimensional data in lower dimensions while preserving pairwise similarity.
- Autoencoders: Neural networks that learn compact representations of data.

Principal Component Analysis (PCA)

- PCA is a dimensionality reduction technique.
- It identifies and selects orthogonal axes (principal components) in the data.
- These components capture the maximum variance in the original data.
- By retaining only the most informative components, PCA reduces the dimensionality of the data while minimizing loss of information.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE is a visualization technique for high-dimensional data.
- It aims to project data into a lower-dimensional space (typically 2D or 3D) while preserving the pairwise similarity between data points.
- t-SNE is particularly effective at revealing clusters or patterns in the data, making it useful for data exploration and visualization.

Autoencoders

- Autoencoders are a class of neural networks used for dimensionality reduction and feature learning.
- They consist of an encoder and a decoder.
- Autoencoders aim to learn a compact and informative representation (encoding) of the input data.
- This compact representation can be used for tasks like denoising, anomaly detection, or even as a reduced-dimensional input for other machine learning models

Applications of Unsupervised Learning

HUMBER

WE ARE HUMBER

Applications of Unsupervised Learning

- **Text Mining:** Extracting meaningful insights and knowledge from unstructured textual data, such as documents, articles, or social media posts
- **Image Compression:** Reducing image file sizes without significant quality loss.
- **Anomaly Detection:** Identifying outliers or anomalies in data.
- **Customer Segmentation:** Grouping customers based on their behavior.
- **Natural Language Processing (NLP):** Topic modeling, word embeddings, and text clustering.
- **Recommendation Systems:** Recommending products, movies, or content based on user behavior.
- **Genomics:** Clustering genes for biological insights.

Text Mining

- Voice of the Customer at Triad Airline
- Preprocessing Text Data for Analysis
- Movie Reviews

Text Mining (1 of 12)

- Text, like numerical data, may contain information that can help solve problems and lead to better decisions.
- **Text mining** is the process of extracting useful information from text data.
- Text data is often referred to as **unstructured data** because in its raw form, it cannot be stored in a traditional structured database (rows and columns).
- Audio and video data are also examples of unstructured data.
- Data mining with text data is more challenging than data mining with traditional numerical data, because it requires more preprocessing to convert the text to a format amenable for analysis.

Text Mining (2 of 12)

Voice of the Customer at Triad Airline:

- Triad solicits feedback from its customers through a follow-up e-mail the day after the customer has completed a flight.
- Survey asks the customer to rate various aspects of the flight and asks the respondent to type comments into a dialog box in the e-mail; includes:
  - Quantitative feedback from the ratings.
  - Comments entered by the respondents which need to be analyzed.
- A collection of text documents to be analyzed is called a **corpus**.

Text Mining (3 of 12)

Table 5.6: Ten Respondents' Concerns for Triad Airlines

Concerns
The wi-fi service was horrible. It was slow and cut off several times.
My seat was uncomfortable.
My flight was delayed 2 hours for no apparent reason.
My seat would not recline.
The man at the ticket counter was rude. Service was horrible.
The flight attendant was rude. Service was bad.
My flight was delayed with no explanation.
My drink spilled when the guy in front of me reclined his seat.
My flight was canceled.
The arm rest of my seat was nasty.

Text Mining (4 of 12)

Voice of the Customer at Triad Airline:

- To be analyzed, text data needs to be converted to structured data (rows and columns of numerical data) so that the tools of descriptive statistics, data visualization and data mining can be applied.
- Think of converting a group of documents into a matrix of rows and columns where the rows correspond to a document and the columns correspond to a particular word.
- A **presence/absence or binary term-document matrix** is a matrix with the rows representing documents and the columns representing words.
  - Entries in the columns indicate either the presence or the absence of a particular word in a particular document.

Text Mining (5 of 12)

Voice of the Customer at Triad Airline (cont.):

- Creating the list of terms to use in the presence/absence matrix can be a complicated matter:
  - Too many terms results in a matrix with many columns, which may be difficult to manage and could yield meaningless results.
  - Too few terms may miss important relationships.
- Term frequency along with the problem context are often used as a guide.
- In Triad’s case, management used word frequency and the context of having a goal of satisfied customers to come up with the following list of terms they feel are relevant for categorizing the respondent’s comments: delayed, flight, horrible, recline, rude, seat, and service.

Text Mining (6 of 12)

Table 5.7: The Presence/Absence Term-Document Matrix for Triad Airlines

	Term						
Document	**Delayed**	**Flight**	**Horrible**	**Recline**	**Rude**	**Seat**	**Service**
1	0	0	1	0	0	0	1
2	0	0	0	0	0	1	0
3	1	1	0	0	0	0	0
4	0	0	0	1	0	1	0
5	0	0	1	0	1	0	1
6	0	1	0	0	1	0	1
7	1	1	0	0	0	0	0
8	0	0	0	1	0	1	0
9	0	1	0	0	0	0	0
10	0	0	0	0	0	1	0

Text Mining (7 of 12)

Preprocessing Text Data for Analysis:

- The text-mining process converts unstructured text into numerical data and applies quantitative techniques.
- Which terms become the headers of the columns of the term-document matrix can greatly impact the analysis.
- **Tokenization** is the process of dividing text into separate terms, referred to as tokens:
  - Symbols and punctuations must be removed from the document, and all letters should be converted to lowercase.
  - Different forms of the same word, such as “stacking,” “stacked,” and “stack” probably should not be considered as distinct terms.
  - **Stemming** is the process of converting a word to its stem or root word.

Text Mining (8 of 12)

Preprocessing Text Data for Analysis (cont.):

- The goal of preprocessing is to generate a list of most-relevant terms that is sufficiently small so as to lend itself to analysis:
  - Frequency can be used to eliminate words from consideration as tokens.
  - Low-frequency words probably will not be very useful as tokens.
  - Consolidating words that are synonyms can reduce the set of tokens.
  - Most text-mining software gives the user the ability to manually specify terms to include or exclude as tokens.
- The use of slang, humor, and sarcasm can cause interpretation problems and might require more sophisticated data cleansing and subjective intervention on the part of the analyst to avoid misinterpretation.
- Data preprocessing parses the original text data down to the set of tokens deemed relevant for the topic being studied.

Text Mining (9 of 12)

Preprocessing Text Data for Analysis (cont.):

- When the documents in a corpus contain many words and when the frequency of word occurrence is important to the context of the business problem, preprocessing can be used to develop a frequency term-document matrix.
- A **frequency term-document matrix** is a matrix whose rows represent documents and columns represent tokens, and the entries in the matrix are the frequency of occurrence of each token in each document.

Text Mining (10 of 12)

Movie Reviews:

- A new action film has been released, and we now have a sample of 10 reviews from movie critics.
- Using preprocessing techniques, we have reduced the number of tokens to only two: “great” and “terrible.”
- Table 4.8 displays the corresponding frequency term-document matrix.
- To demonstrate the analysis of a frequency term-document matrix with descriptive data mining, we apply *k*-means clustering with $k = 2$ to the frequency term-document matrix to obtain the two clusters in Figure 4.5.

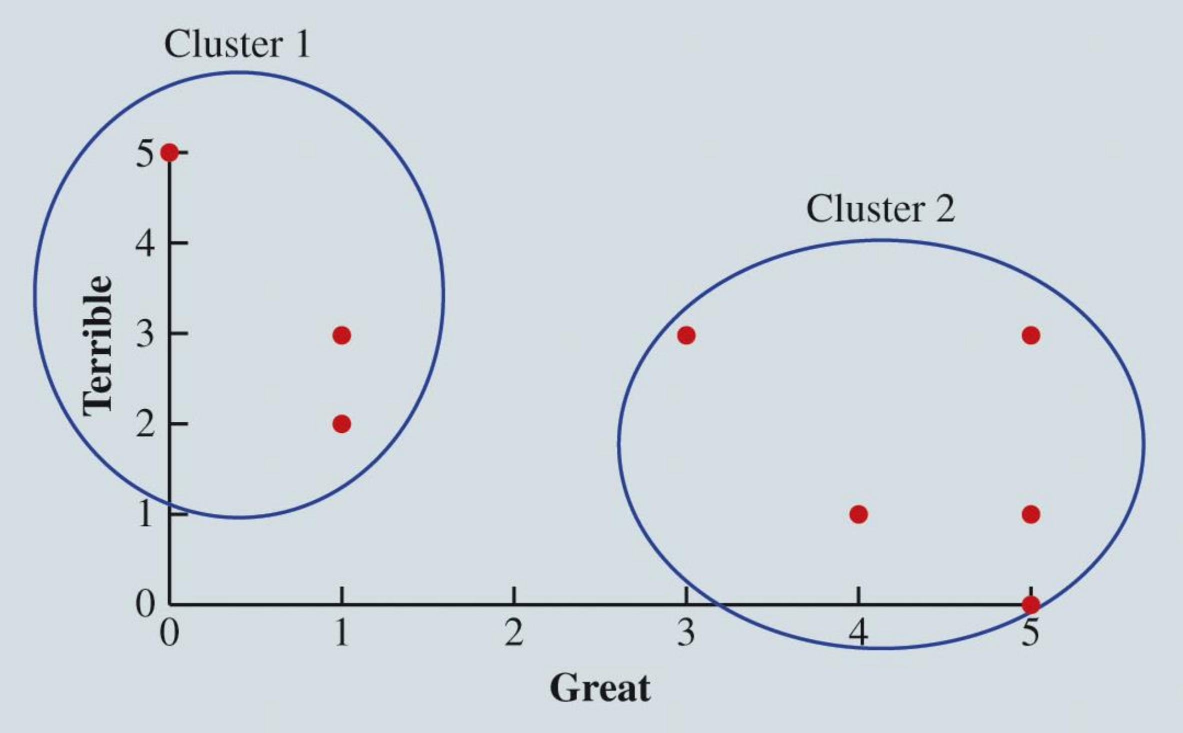
Text Mining (11 of 12)

Table 5.8: The Frequency Term-Document Matrix for Movie Reviews

	Term	
Document	**Great**	**Terrible**
1	5	0
2	5	1
3	5	1
4	3	3
5	5	1
6	0	5
7	4	1
8	5	3
9	1	3
10	1	2

Text Mining (12 of 12)

Figure 5.5: Two Clusters Using *k*-Means Clustering on Movie Reviews

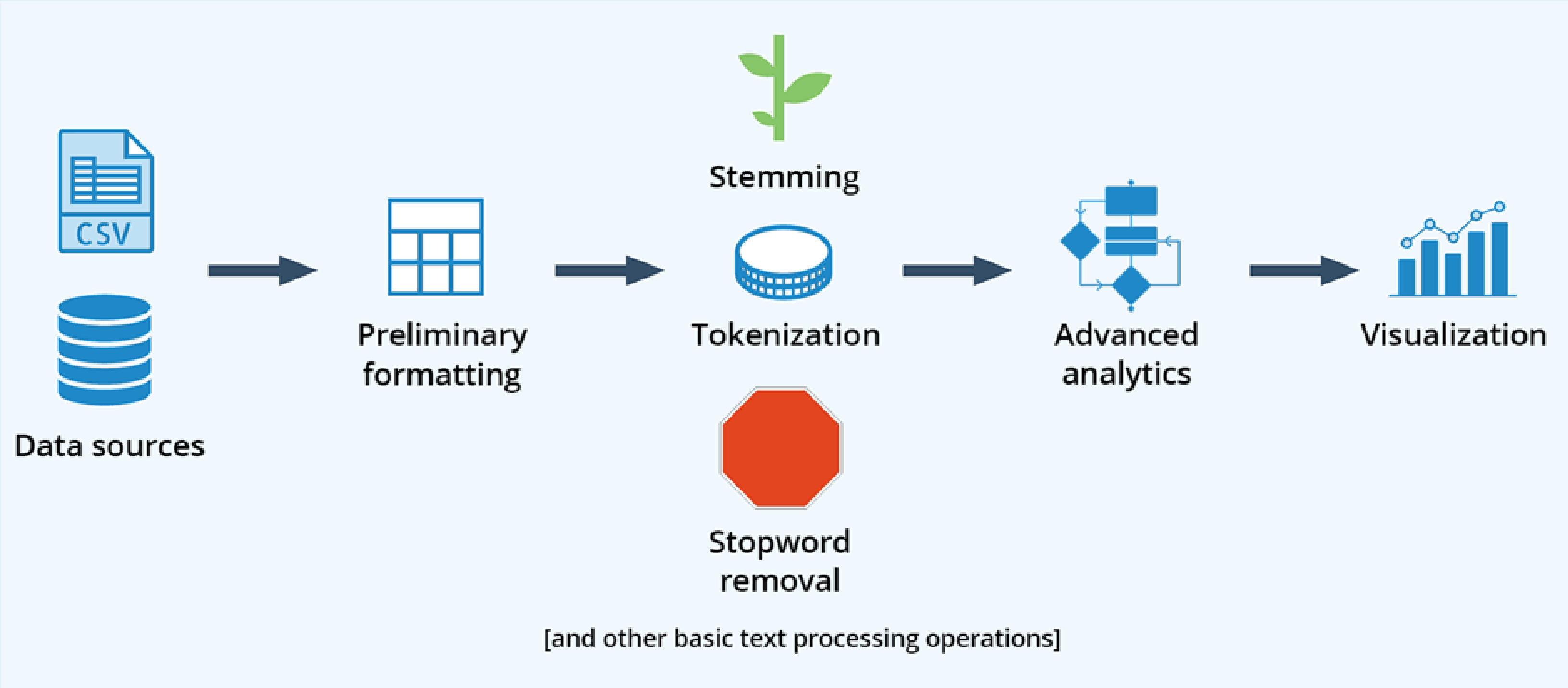


WE ARE HUMBER

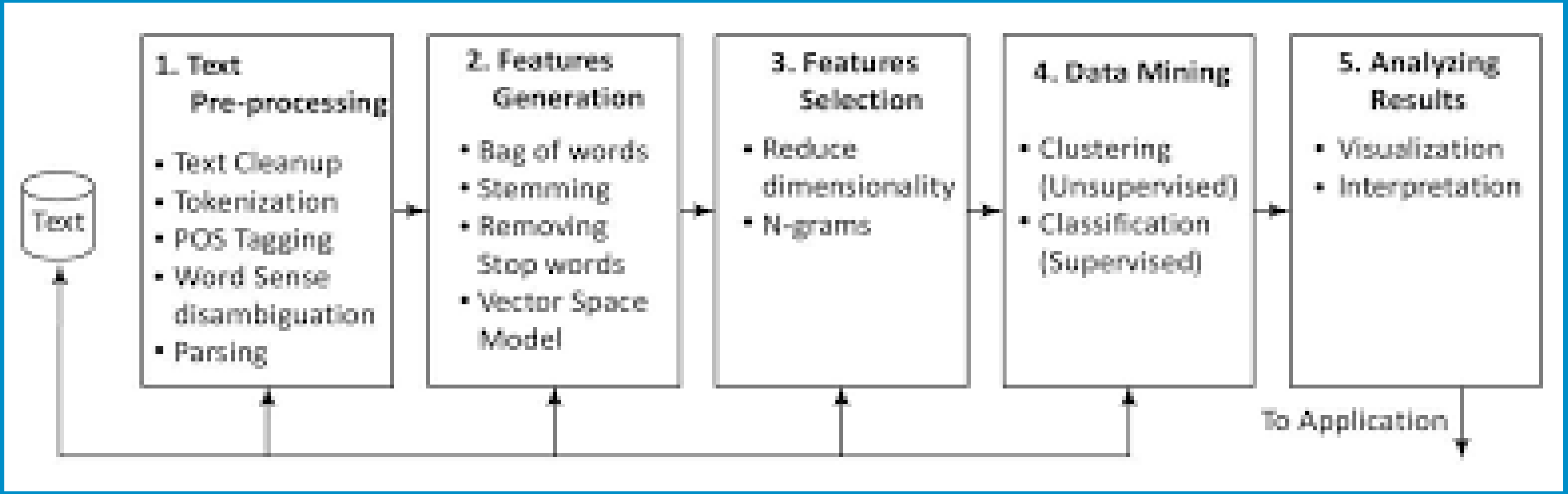
Text Mining Processes and Platforms/Solutions

- Text mining key processes
- Text mining platforms, software and tools

Text Mining Key processes (1 of 1)



Text Mining processes and underlying mining techniques



Text Mining Platforms, Software and Tools

- MonkeyLearn.
- Google Cloud NLP.
- IBM Watson.
- Amazon Comprehend.
- AYLIEN.
- Thematic.
- MeaningCloud.
- Apache OpenNLP's
- SAS Text Miner
- DiscoverText
- Levity
- RapidMiner
- …

Summary

- Unsupervised learning is a powerful tool for discovering hidden patterns and structures in data.
- It plays a crucial role in various fields, from data analysis to artificial intelligence.
- As we continue to advance in machine learning and AI, unsupervised learning techniques will become even more integral to our toolkit