NATIONAL ENGINEERING CENTER
University of the Philippines
Diliman, Quezon City

7.0 Introduction to Unsupervised Learning and Modern Data Mining

Eugene Rex L. Jalao, Ph.D.

Associate Professor

Department Industrial Engineering and Operations Research

University of the Philippines Diliman

@thephdataminer

Module 1 of the Business Intelligence and Analytics Track of UP NEC and the UP Center of Business Intelligence

Module 1 Outline

1. Intro to Business Intelligence
   - Case Study on Selecting BI Projects
2. Data Warehousing
   - Case Study on Data Extraction and Report Generation
3. Descriptive Analytics
   - Case Study on Data Analysis
4. Visualization
   - Case Study on Dashboard Design
5. Classification Analysis
   - Case Study on Classification Analysis
6. Regression and Time Series Analysis
   - Case Study on Regression and Time Series Analysis
7. **Unsupervised Learning and Modern Data Mining**
   - **Case Study on Text Mining**
8. Optimization for BI

eljalao@up.edu.ph

Unsupervised Learning

Definition 7.1: Unsupervised Learning

- Finding **hidden patterns** within data
- No **Response/Class** variable
- No guarantee that there are **meaningful** patterns
- No easy way to **measure errors**
- Most research on **new algorithms**

Outline for This Session

- Association Rule Mining
- Clustering
  - K-Means Clustering
  - Hierarchical Clustering
- Web Mining
- Text Mining
- Case Study

eljalao@up.edu.ph

Association Rule Mining

Definition 7.2: Association Rule Mining

- Given a set of transactions, **find rules** that will predict the occurrence of an item based on the occurrences of other items in the transaction

Association Rule Mining

Example 7.1: Example Market Basket Analysis

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

$\{Diaper\} \quad \{Beer\},$
$\{Milk, Bread\} \rightarrow \{Eggs, Coke\},$
$\{Beer, Bread\} \rightarrow \{Milk\},$

Association Rule Mining

Definition 7.3: Itemset

- A collection of one or more items
  - Example: $\{Milk, Bread, Diaper\}$
- k-itemset
  - An itemset that contains k items

Definition 7.4: Support Count (σ)

- Frequency of occurrence of an itemset
  - E.g. $\sigma(\{Milk, Bread, Diaper\}) = 2$

eljalao@up.edu.ph

Association Rule Mining

Definition 7.5: Support

- Fraction of transactions that contain an itemset
  - E.g. $s(\{Milk, Bread, Diaper\}) = 2/5$

Definition 7.6: Frequent Itemset

- Frequent Itemset
  - An itemset whose support is greater than or equal to a *minsup* threshold

Association Rule Mining

Definition 7.7: Association Rule

- An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example: $\{Milk, Diaper\} \rightarrow \{Beer\}$

Definition 7.8: Confidence

- Measures how often items in Y appear in transactions that contain X

Association Rule Mining

- Given a set of transactions T, the goal of association rule mining is to find all rules having
  - $support \geq minsup$ threshold
  - $confidence \geq minconf$ threshold
- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds

  $\Rightarrow$ **Computationally prohibitive!**

Example: R Scripts

```
> library(arules)
> library(arulesViz)
> par(mar=c(2,2,2,2))
> Groceries=read.transactions("Groceries.csv",format="basket",sep=",")
> itemFrequencyPlot(Groceries,topN=20,type="absolute")
```

Association Rule Mining

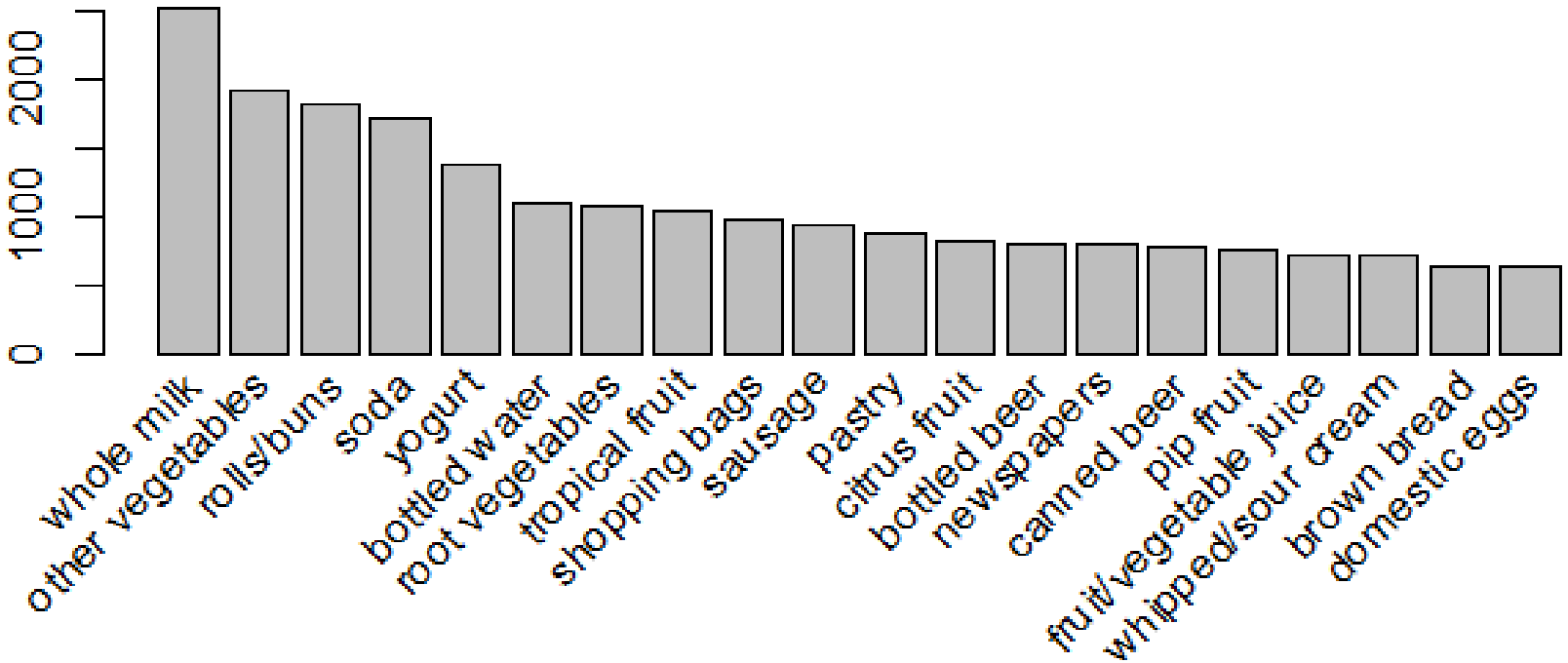


Figure 7.1: Item Frequency Plot

Example: R Scripts

```
> rules = apriori(Groceries, parameter =
  list(supp = 0.001, conf = 0.8))
> options(digits=2)
> inspect(rules[1:20])
> rules=sort(rules, by="confidence",
  decreasing=TRUE)
> inspect(rules[1:20])
```

Association Rule Mining

- Results

	lhs		rhs	support	confidence
1	{liquor, red/blush wine}	=>	{bottled beer}	0.0019	0.90
2	{cereals, curd}	=>	{whole milk}	0.0010	0.91
3	{cereals, yogurt}	=>	{whole milk}	0.0017	0.81
4	{butter, jam}	=>	{whole milk}	0.0010	0.83
5	{bottled beer, soups}	=>	{whole milk}	0.0011	0.92
6	{house keeping products, napkins}	=>	{whole milk}	0.0013	0.81

Example: R Scripts

```
› plot(rules[1:20],method="graph",interac
  tive=TRUE,shading=T)
```

Association Rule Mining

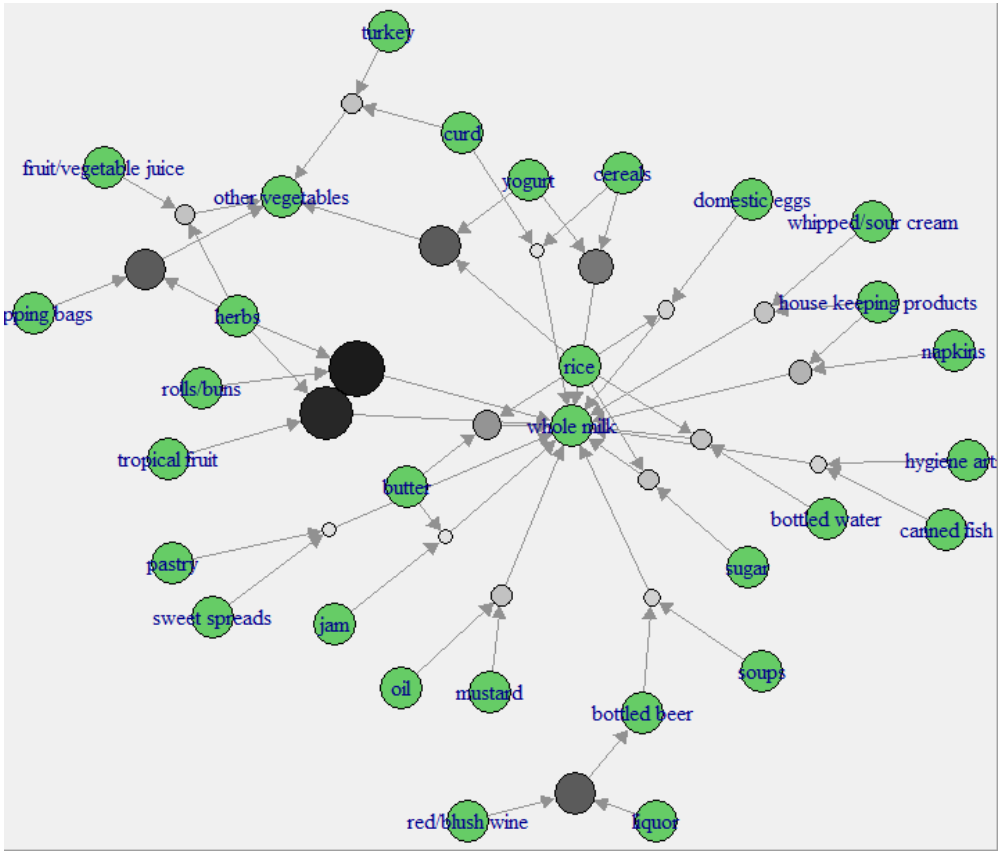


Figure 7.2: Association Graph

Outline for This Session

- Association Rule Mining
- **Clustering**
  - K-Means Clustering
  - Hierarchical Clustering
- Web Mining
- Text Mining
- Case Study

Clustering

Definition 7.9: Clustering

- Finding groups of objects such that the objects in a group will be **similar** (or related) to one another and **different** from (or unrelated to) the objects in other groups

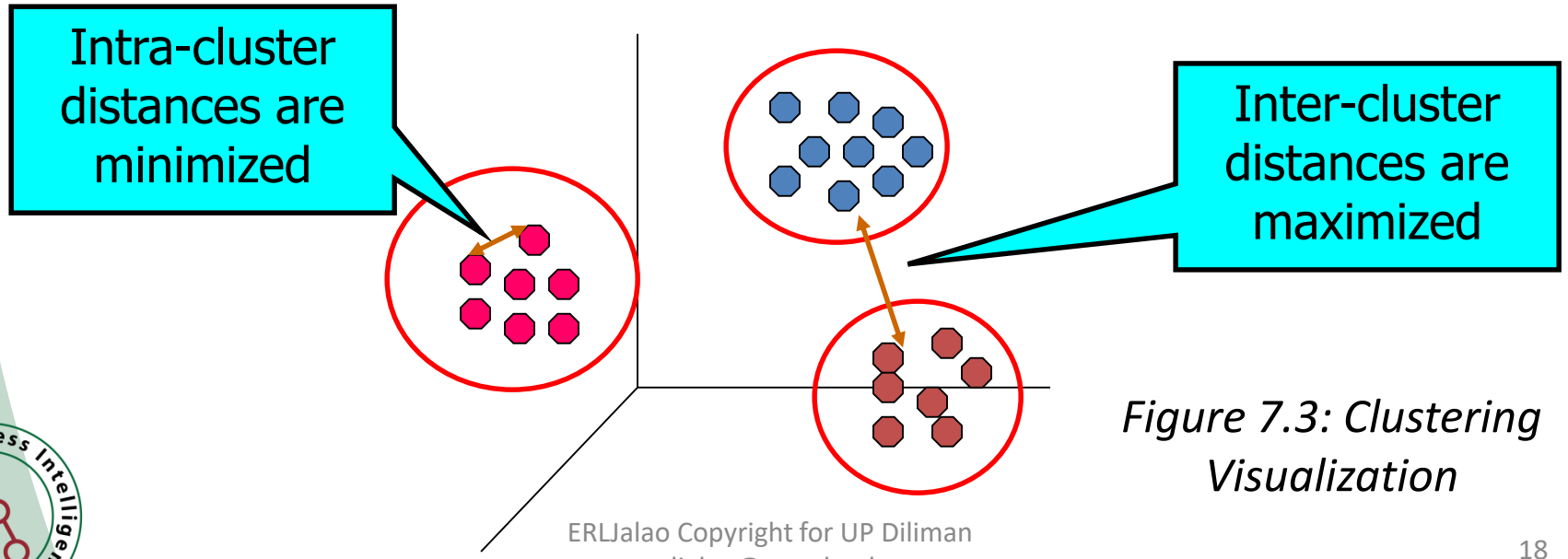


Figure 7.3: Clustering Visualization

Clustering

- Applications of Cluster Analysis
  - Understanding
    - Group **related documents** for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations
  - Summarization
    - **Reduce the size** of large data sets

Clustering

- What is not Cluster Analysis?
  - **Supervised classification**
    - Have class label information
  - **Simple segmentation**
    - Dividing students into different registration groups alphabetically, by last name
  - **Results of a query**
    - Groupings are a result of an external specification
  - **Graph partitioning**
    - Some mutual relevance and synergy, but areas are not identical

Clustering

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Figure 7.4: Notion of Cluster Could Be Ambiguous

Clustering

- Important distinction between hierarchical and partitional sets of clusters

Definition 7.10: Partitional Clustering

- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

Definition 7.11: Hierarchical Clustering

- A set of nested clusters organized as a hierarchical tree

Clustering

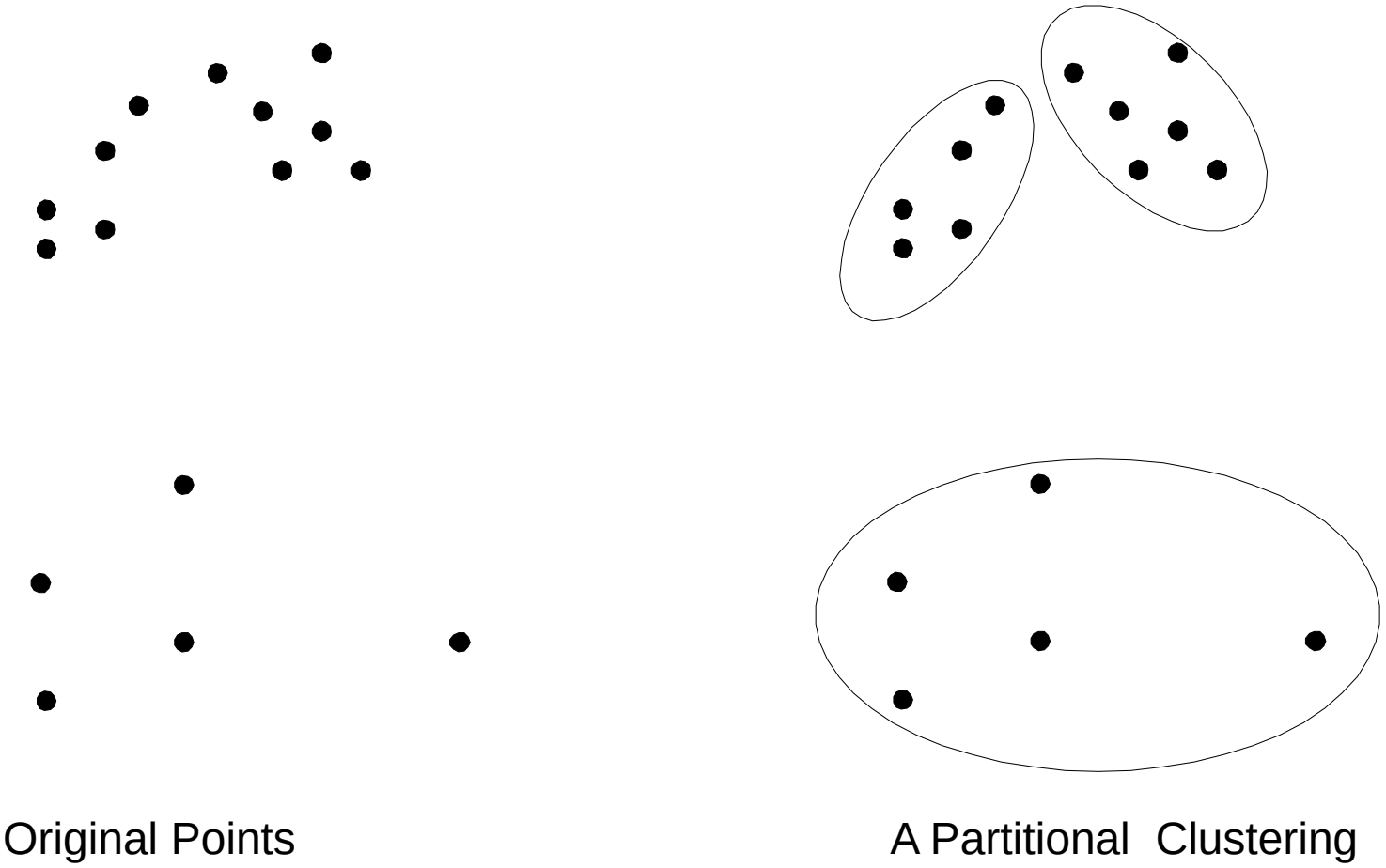
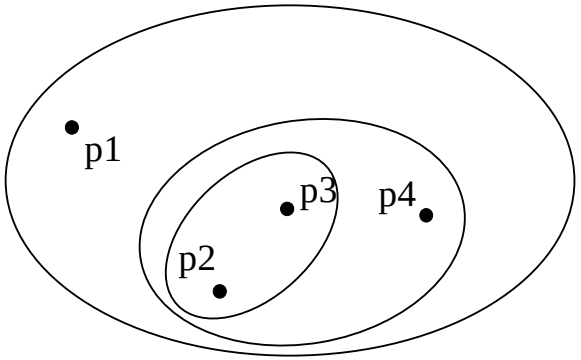
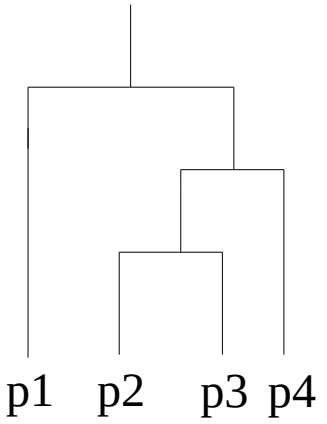


Figure 7.5: A Partitional Clustering

Clustering



Traditional Hierarchical Clustering



Traditional Dendrogram

Figure 7.6: A Hierarchical Clustering

Outline for This Session

- Association Rule Mining
- Clustering
  - **K-Means Clustering**
  - Hierarchical Clustering
- Web Mining
- Text Mining
- Case Study

K-means Clustering

Definition 7.12: K-means Clustering

- It's a partitional clustering approach where each cluster is associated with a centroid **(center point)**. Then each point is **assigned to the cluster** with the closest centroid. This process is repeated until points do not change centroids.
- Number of clusters, K, must be specified
- The basic algorithm is very simple:

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means Clustering

- Initial centroids are often chosen randomly.
  - **Clusters produced vary** from one run to another.
- The centroid is the mean of the points in the cluster.
- **Closeness** is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for **common similarity** measures mentioned above.
- Convergence happens in the **first few iterations.**
  - Often the stopping condition is changed to 'Until relatively few points change clusters'

K-means Clustering

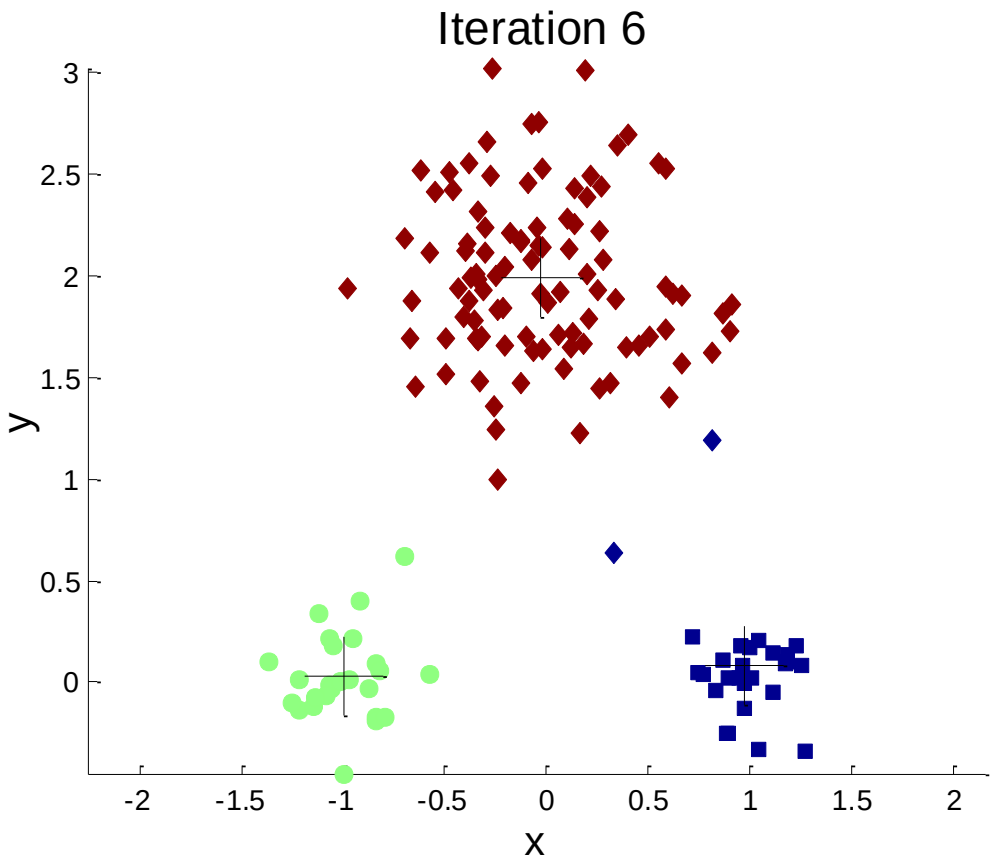


Figure 7.7: K-means Algorithm

K-means Clustering

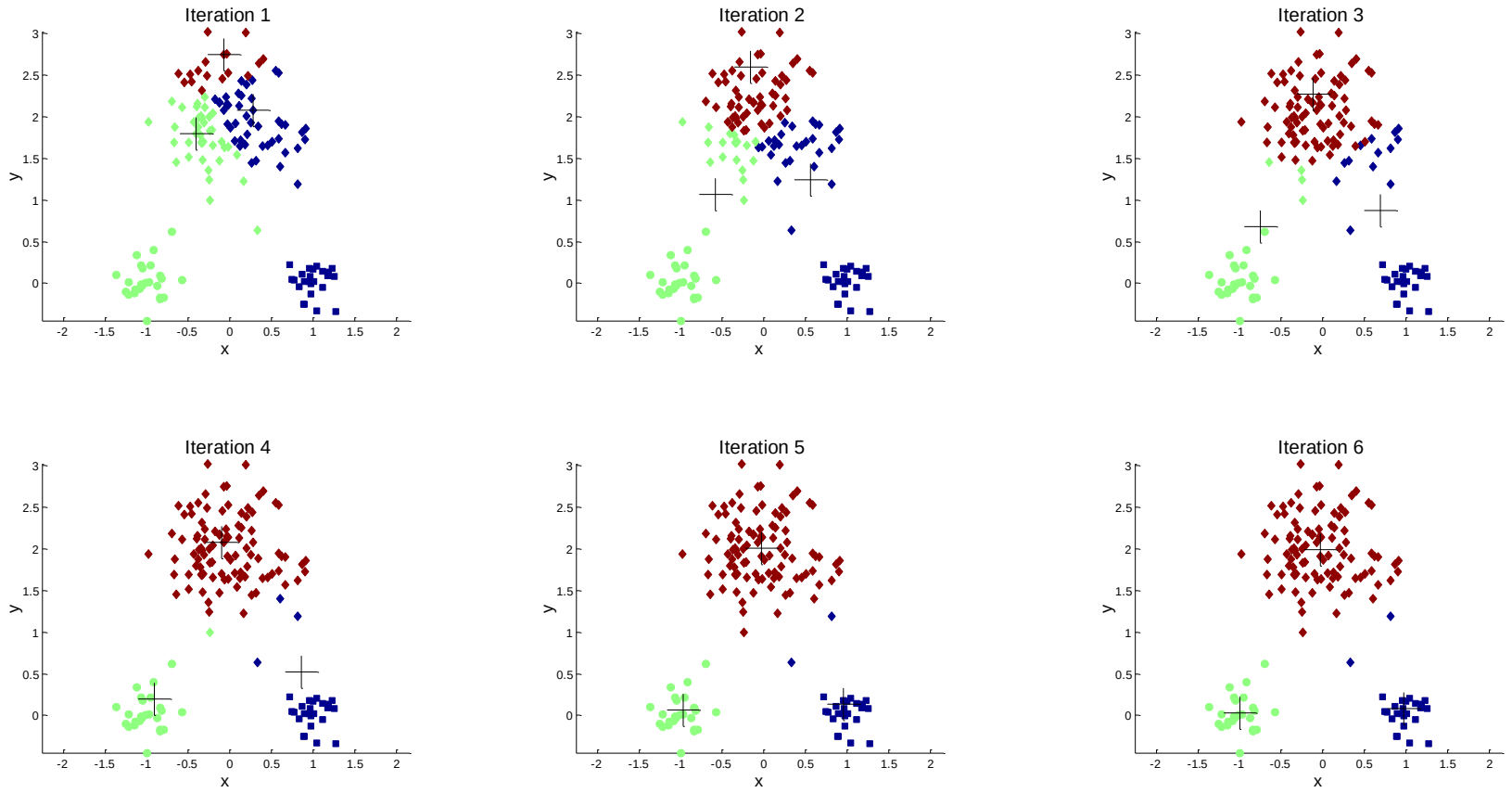


Figure 7.8: K-means Algorithm Summary

Example: Cars DataSet

```
> cars=read.csv("cars.csv")
> rownames(cars) = cars[,1]
> cars = cars[,c(2:12)]
> fit = kmeans(cars, 5)
> aggregate(cars,by=list(fit$cluster),FUN
  =mean)
> carswithclusters = data.frame(cars,
  fit$cluster)
> carswithclusters
```

Example: Cars DataSet

	Group.1	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	1	15	8.0	388	232	3.3	4.2	16	0.00	0.22	3.4	4.0
2	2	19	6.0	171	124	3.7	3.1	18	0.50	0.50	4.0	3.8
3	3	31	4.0	76	62	4.3	1.9	19	1.00	1.00	4.0	1.2
4	4	24	4.0	122	94	3.9	2.5	19	0.86	0.57	4.1	1.7
5	5	17	7.7	285	158	3.0	3.6	18	0.17	0.00	3.0	2.3

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	fit.cluster
Mazda RX4	21	6	160	110	3.9	2.6	16	0	1	4	4	2
Mazda RX4 Wag	21	6	160	110	3.9	2.9	17	0	1	4	4	2
Datsun 710	23	4	108	93	3.8	2.3	19	1	1	4	1	4
Hornet 4 Drive	21	6	258	110	3.1	3.2	19	1	0	3	1	5
Hornet Sportabout	19	8	360	175	3.1	3.4	17	0	0	3	2	1
Valiant	18	6	225	105	2.8	3.5	20	1	0	3	1	2

Example: Cars DataSet

```
> library(cluster)
> clusplot(cars, fit$cluster, color=TRUE,
  shade=TRUE, labels=2, lines=0)
```

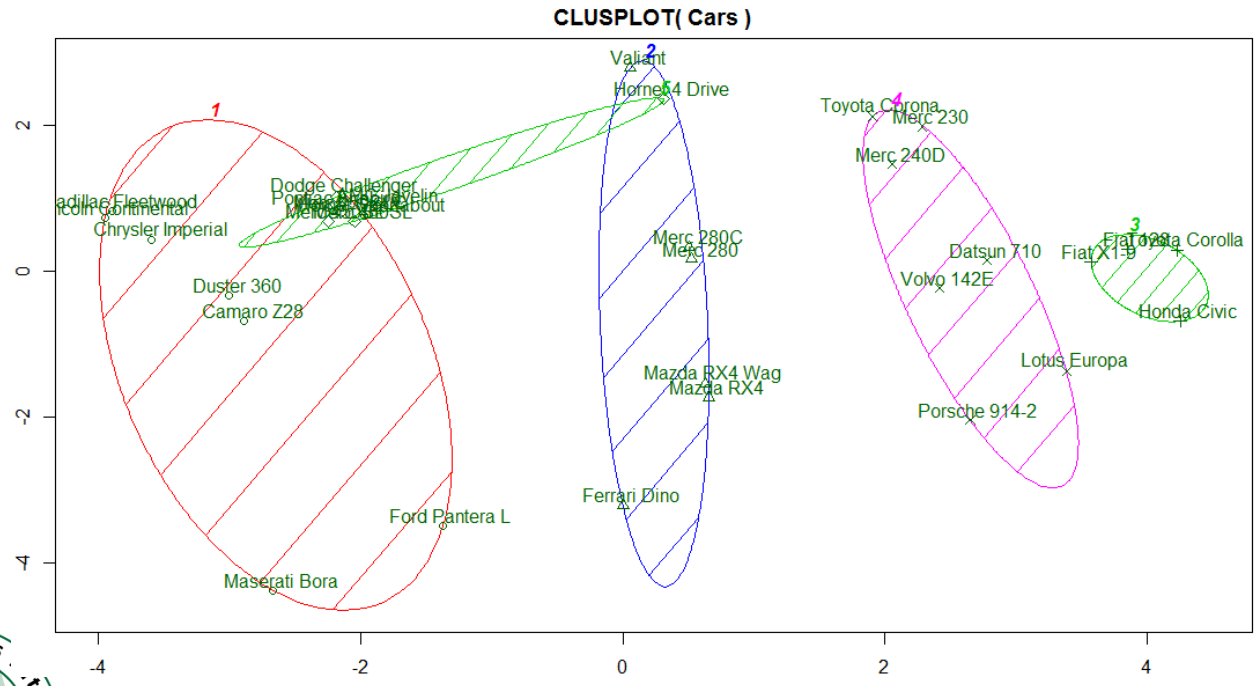


Figure 7.9: Cars Dataset in 2D

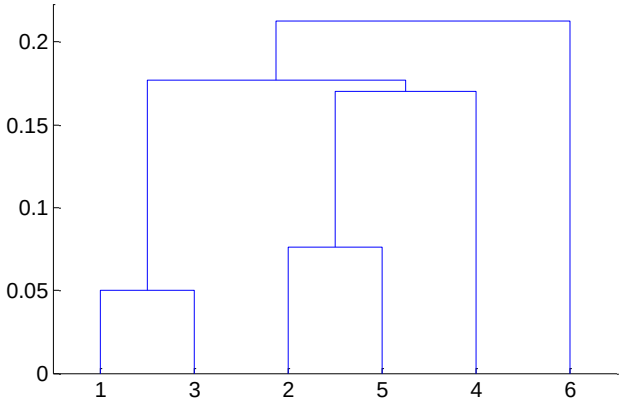
Outline for This Session

- Association Rule Mining
- Clustering
  - K-Means Clustering
  - **Hierarchical Clustering**
- Web Mining
- Text Mining
- Case Study

Hierarchical Clustering

Definition 7.12: Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a **dendrogram**
  - A tree like diagram that records the sequences of merges or splits



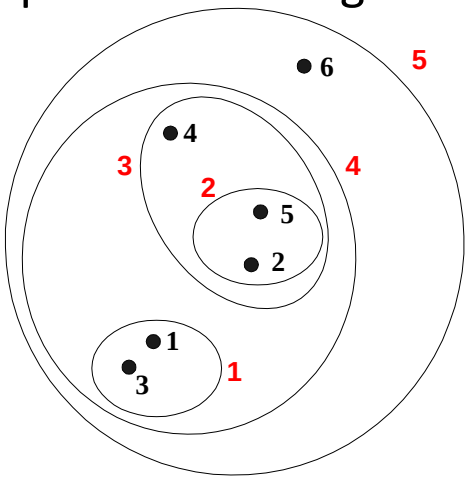


Figure 7.10: Dendogram

Hierarchical Clustering

- Strengths of Hierarchical Clustering
  - Do not have to assume any **particular number** of clusters
    - Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level
  - They may correspond to **meaningful taxonomies**
    - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

Example

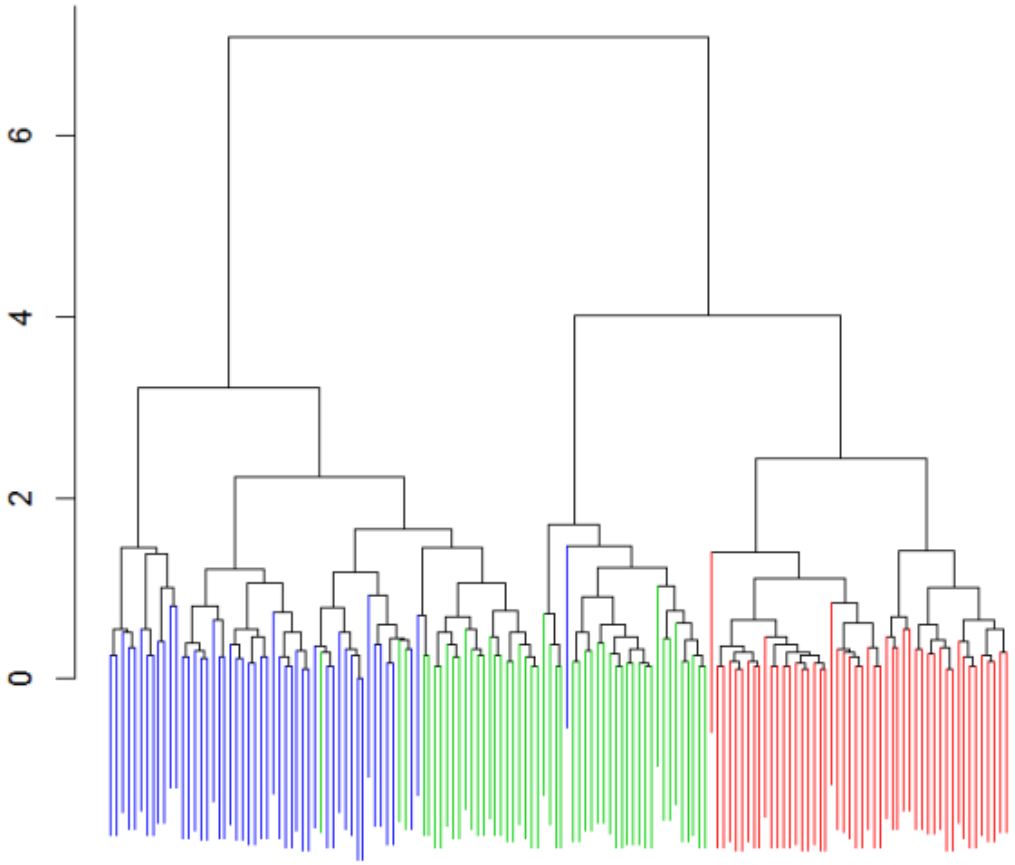


Figure 7.11: Dendogram for the Iris Dataset

Hierarchical Clustering

Definition 7.13: Agglomerative Clustering Algorithm

- Most **popular** hierarchical clustering technique
- Basic algorithm is straightforward
  - Compute the proximity matrix
  - Let each data point be a cluster
  - Repeat
  - Merge the two closest clusters
  - Update the proximity matrix
  - Until only a single cluster remains

R Code

```
cars=read.csv("cars.csv")
rownames(cars) = Cars[,1]
cars = cars[,c(2:12)]
d = dist(cars, method = "euclidean")
fit = hclust(d, method="ward")
plot(fit, main="hierarchical clustering for mtcars dataset")
groups = cutree(fit, k=5)
rect.hclust(fit, k=5, border="red")
```

Example of Clustering

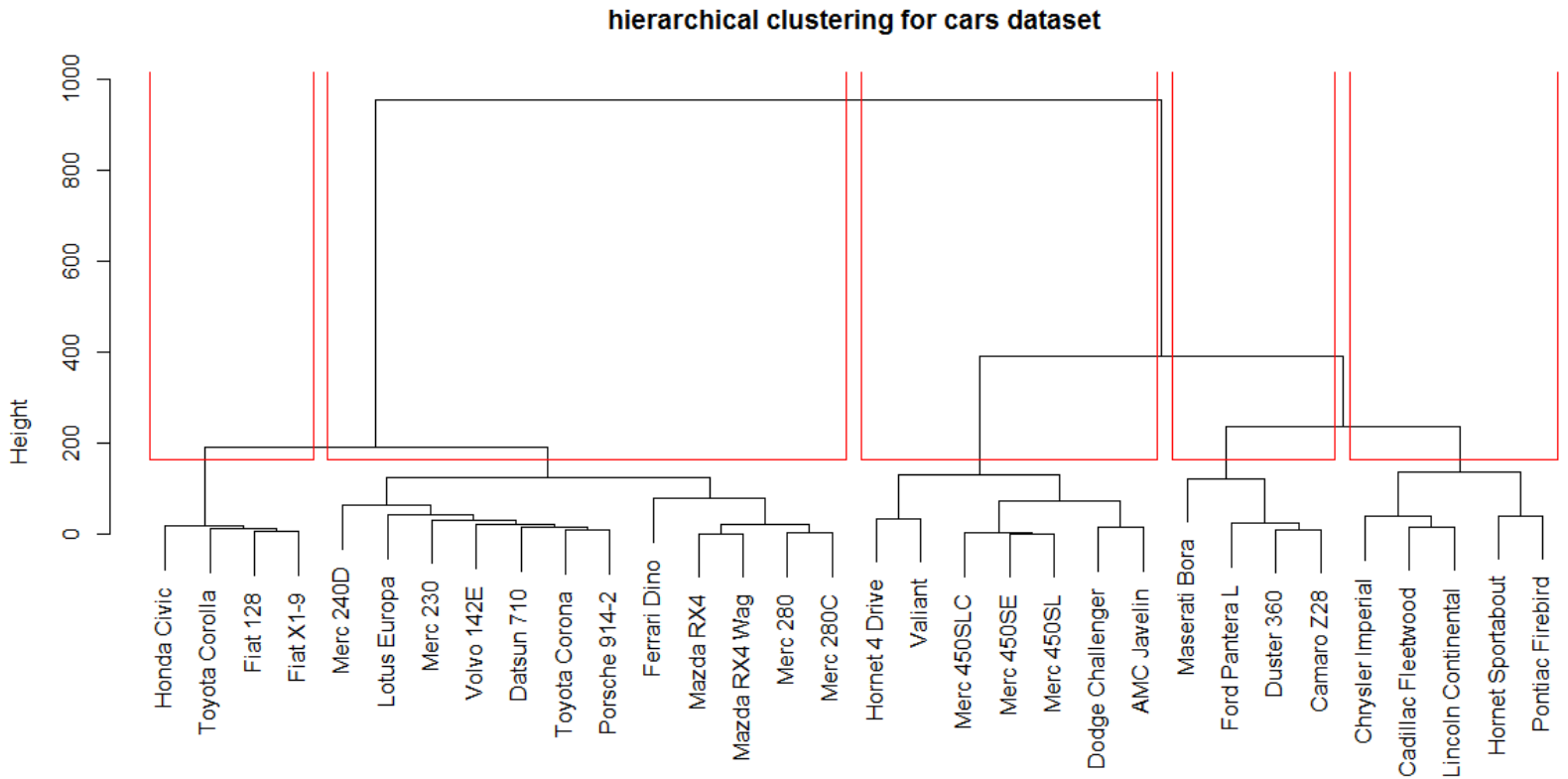


Figure 7.12: Dendogram for the Cars Dataset

Outline for This Session

- Association Rule Mining
- Clustering
  - K-Means Clustering
  - Hierarchical Clustering
- **Web Mining**
- Text Mining
- Case Study

eljalao@up.edu.ph

Web Mining

- The Web is the **largest repository** of data
- Data is in **HTML, XML, text** format
- **Challenges** (of processing Web data)
  - The Web is too big for effective data mining
  - The Web is too complex
  - The Web is too dynamic
  - The Web is not specific to a domain
  - The Web has everything
- Opportunities and challenges are great!

eljalao@up.edu.ph

Web Mining

Definition 7.14: Web Mining

- Web mining (or Web data mining) is the process of discovering **intrinsic relationships** from Web data (textual, linkage, or usage)

Web Mining

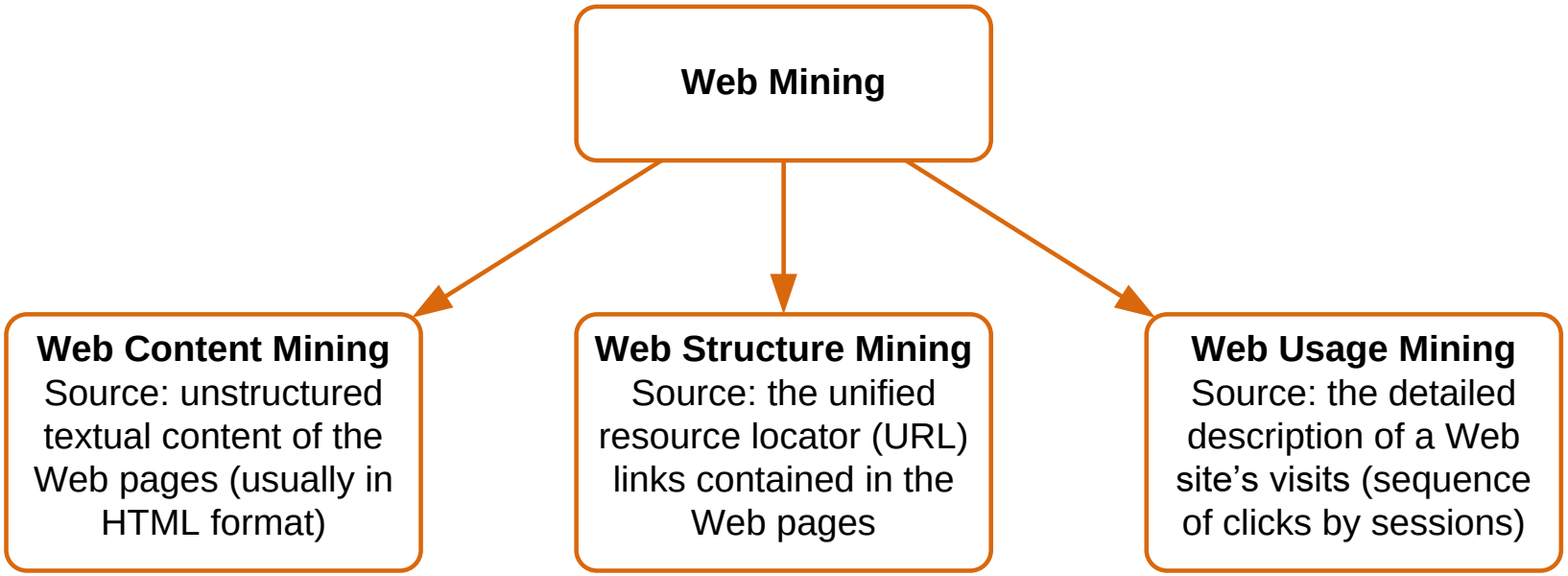


Figure 7.13: Web Mining Components

Web Mining

Definition 7.15: Web Content/Structure Mining

- Mining of the **textual content** on the Web
- **Data collection** via Web crawlers
- Web pages include **hyperlinks**
  - Authoritative pages
  - Hubs

Web Mining

- Web Structure Mining
  - Generate **structural summary** about the Web site and Web page
  - Depending upon the hyperlink, **Categorizing** the Web pages and the related Information @ inter domain level
  - Discovering the **Web Page Structure.**
  - Discovering the nature of the **hierarchy of hyperlinks** in the website and its structure.

Web Mining

- Web Structure Mining
  - **Finding Information about web pages**
    - Retrieving information about the relevance and the quality of the web page.
    - Finding the authoritative on the topic and content.
  - **Inference on Hyperlink**
    - The web page contains not only information but also hyperlinks, which contains huge amount of annotation.
    - Hyperlink identifies author's endorsement of the other web page.

Web Mining

- Web Content Mining
  - Pre-processing data before web content mining: **feature selection**
  - Post-processing data can **reduce ambiguous searching results**
  - Web Page Content Mining
    - Mines the contents of **documents directly**
  - Search Engine Mining
    - **Improves on the** content search of other tools like search engines.

Web Mining

- Web content mining is related to data mining and text mining. [Bing Liu. 2005]
  - It is related to **data mining** because many data mining techniques can be applied in Web content mining.
  - It is related to **text mining** because much of the web contents are texts.
  - Web data are mainly **semi-structured and/or unstructured**, while data mining is structured and text is unstructured.

http://www.cs.uic.edu/~liub

Web Mining

Definition 7.16: Web Usage Mining

- Extraction of information from data generated through Web page visits and transactions…
  - **data stored** in server access logs, referrer logs, agent logs, and client-side cookies
  - **user characteristics** and usage profiles
  - metadata, such as page attributes, content attributes, and usage data
- Clickstream **data**
- Clickstream **analysis**

Web Mining

- Web usage mining applications
  - Determine the **lifetime value of clients**
  - Design **cross-marketing strategies** across products.
  - Evaluate **promotional campaigns**
  - Target **electronic ads and coupons** at user groups based on user access patterns
  - Predict **user behavior** based on previously learned rules and users' profiles
  - Present **dynamic information** to users based on their interests and profiles…

Web Mining

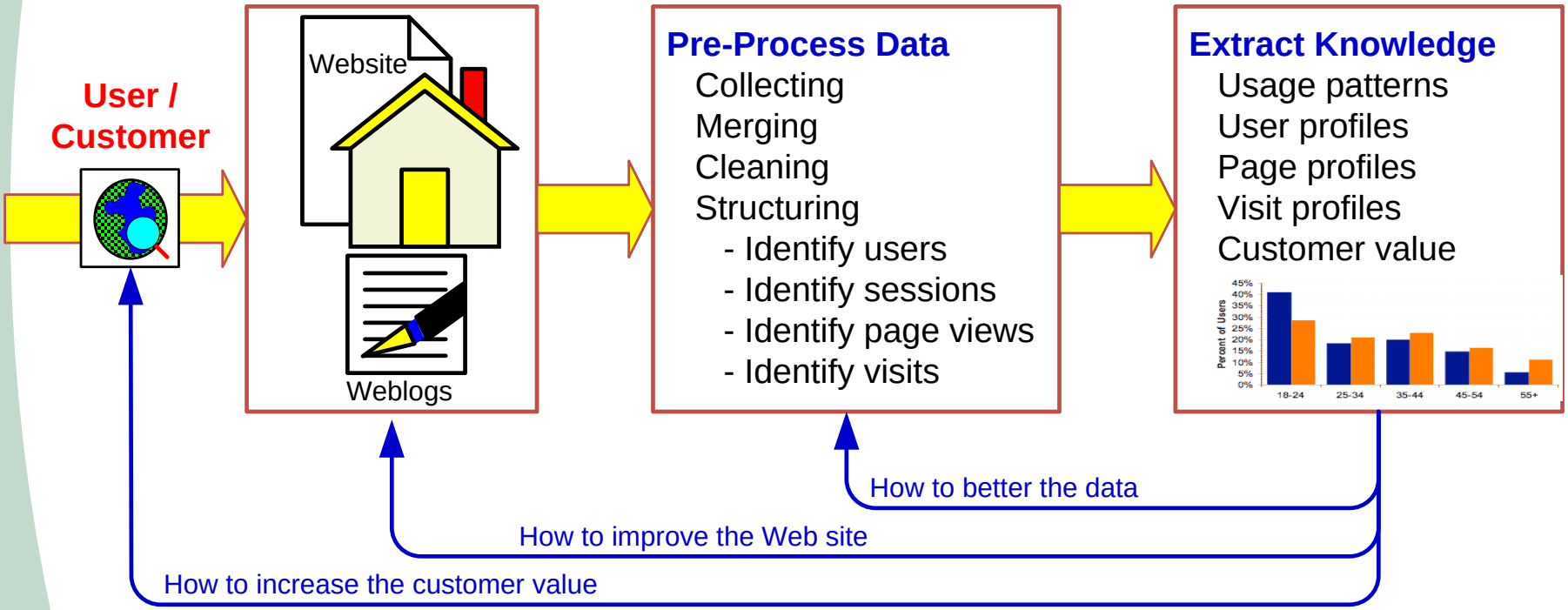


Figure 7.14: Clickstream Analysis

Web Mining

- Web Mining Success Stories
  - Amazon.com, Ask.com, Scholastic.com, …
  - Website Optimization Ecosystem

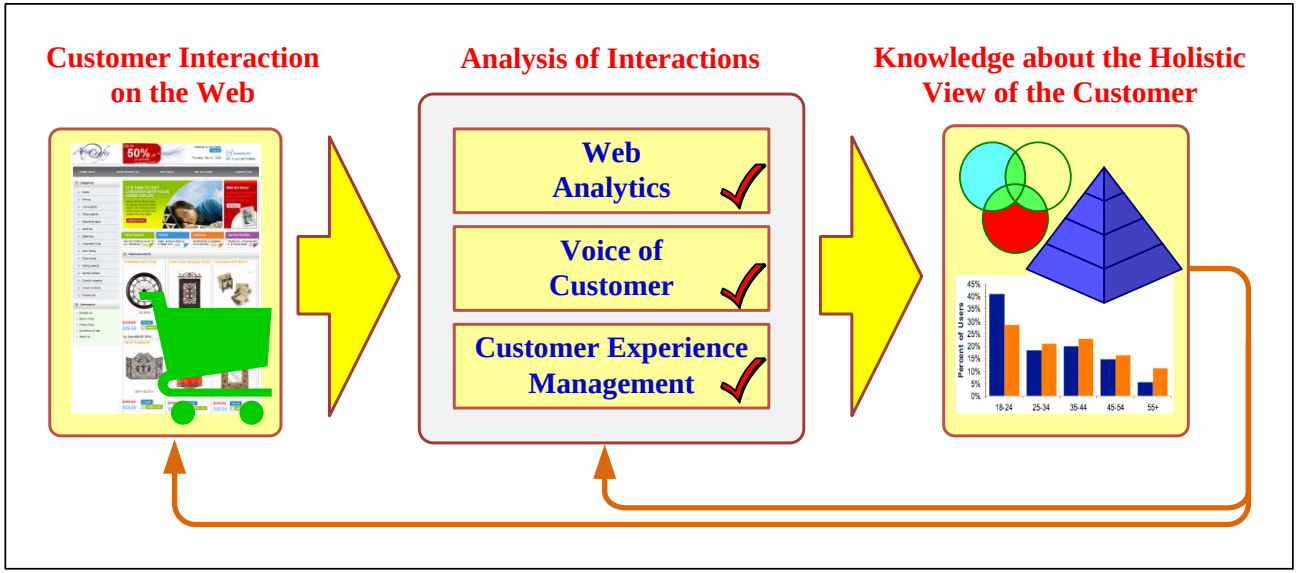


Figure 7.15: Clickstream Analysis

Web Mining

- Data Mining Techniques – **Sequential Patterns**

Example: Supermarket Cont…

Customer	Transaction Time	Purchased Items
John	6/21/05 5:30 pm	Beer
John	6/22/05 10:20 pm	Brandy
Frank	6/20/05 10:15 am	Juice, Coke
Frank	6/20/05 11:50 am	Beer
Frank	6/20/05 12:50 am	Wine, Cider
Mary	6/20/05 2:30 pm	Beer
Mary	6/21/05 6:17 pm	Wine, Cider
Mary	6/22/05 5:05 pm	Brandy

Web Mining

Product Name	URL
Angoss Knowledge WebMiner	angoss.com
ClickTracks	clicktracks.com
LiveStats from DeepMetrix	deepmetrix.com
Megaputer WebAnalyst	megaputer.com
MicroStrategy Web Traffic Analysis	microstrategy.com
SAS Web Analytics	sas.com
SPSS Web Mining for Clementine	spss.com
WebTrends	webtrends.com
XML Miner	scientio.com

Outline for This Session

- Association Rule Mining
- Clustering
  - K-Means Clustering
  - **Hierarchical Clustering**
- Web Mining
- **Text Mining**
- Case Study

Text Mining

- 85-90 percent of all corporate data is in some kind of **unstructured form** (e.g., text)
- Unstructured corporate data is **doubling in size** every 18 months
- Tapping into these information sources is not an option, **but a need** to stay competitive

Text Mining

Definition 7.17: Text Mining

- A **semi-automated** process of extracting knowledge from unstructured data sources
- a.k.a. text data mining or knowledge discovery in textual databases

Text Mining

- Data Mining versus Text Mining
  - Both seek for **novel and useful** patterns
  - Both are **semi-automated** processes
  - Difference is the nature of the data:
    - Structured versus unstructured data
    - Structured data: in databases
    - Unstructured data: Word documents, PDF files, text excerpts, XML files, and so on
  - Text mining – first, **impose structure to the data**, **then mine** the structured data

eljalao@up.edu.ph

Text Mining

- Benefits of text mining are obvious especially in **text-rich data** environments
  - e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.
- Electronic **communization records** (e.g., Email)
  - Spam filtering
  - Email prioritization and categorization
  - Automatic response generation

Text Mining

Definition 7.19: Unstructured Data

- Data does not have a **predetermined format** and stored in documents

Definition 7.20: Corpha

- Large collection of **structured texts** for knowledge discovery

Definition 7.21: Stemming

- The process of reducing **inflected words** to their stem. Stemmer, stemming, stemmed are all based on the root stem.

Text Mining

Definition 7.22: Stop Words

- Words that **are filtered out** prior to or after processing of natural language data (a, am, the, of…)

Definition 7.23: Term

- A **single word or phrase** extracted from the corpus

Definition 7.24: Tokenizing

- A **token is a categorized block of text** in a sentence. The assignment of meanings to blocks of text is called tokenizing

Text Mining

Definition 7.25: Term by Document Dataset

- Dataset the where value of each attribute is the number of times the corresponding term occurs in the object.
  - Each term is a component (attribute) of the vector,
  - Extension of a data matrix

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Figure 7.16: Example of a Document Dataset

Bag-of-Tokens Approaches

- Bag-of-Token Approach: Loses all order-specific information and severely limits context

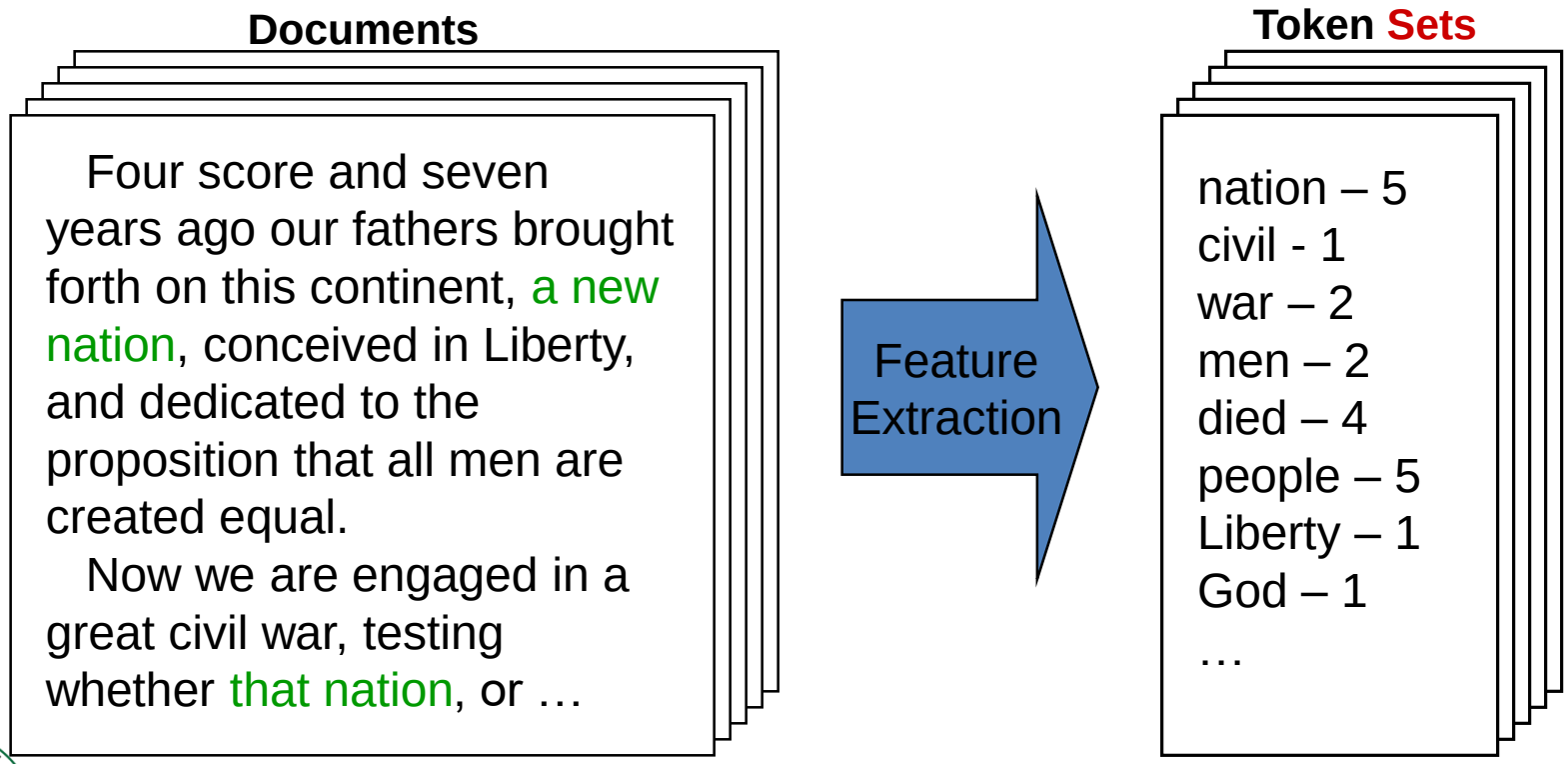


Figure 7.17: Bag-of-Tokens Approach

Text Mining

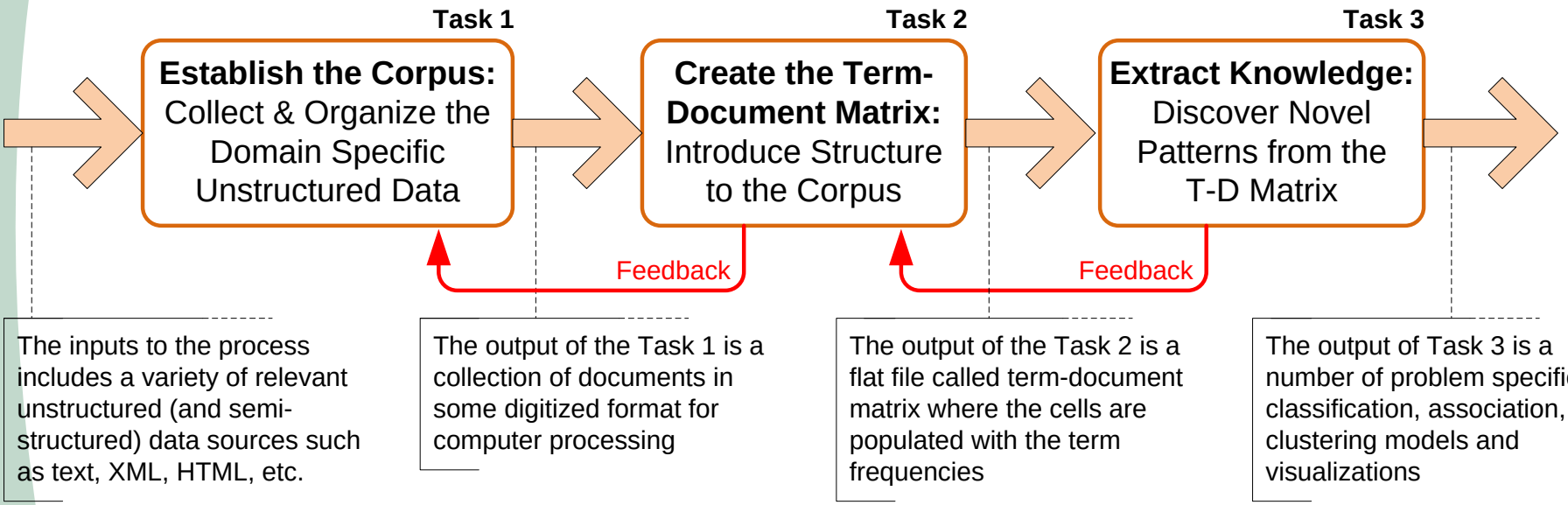


Figure 7.17: The three-step text mining process

Text Mining

- **Step 1:** Establish the corpus
  - Collect all relevant unstructured data
    - (e.g., textual documents, XML files, emails, Web pages, short notes, voice recordings…)
  - Digitize, standardize the collection
    - (e.g., all in ASCII text files)
  - Place the collection in a common place
    - (e.g., in a flat file, or in a directory as separate files)

Text Mining Process

- **Step 2:** Create the Term–by–Document Matrix

Terms / Documents	investment risk	project management	software engineering	development	SAP	...	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
...							

Figure 7.18: Example of a Document Dataset

Text Mining

- **Step 2**: Create the Term–by–Document Matrix (TDM), cont.
  - Should all terms be included?
    - Stop words, include words
    - Synonyms, homonyms
    - Stemming
  - What is the best representation of the indices (values in cells)?
    - Row counts; binary frequencies; log frequencies;
    - Inverse document frequency

Text Mining

- **Step 2:** Create the Term–by–Document Matrix (TDM), cont.
  - TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
    - Manual - a domain expert goes through it
    - Eliminate terms with very few occurrences in very few documents (?)
    - Transform the matrix using singular value decomposition (SVD)
    - SVD is similar to principle component analysis

eljalao@up.edu.ph

Text Mining

- **Step 3:** Extract patterns/knowledge
  - Classification (text categorization)
  - Clustering (natural groupings of text)
    - Improve search recall
    - Improve search precision
    - Scatter/gather
    - Query-specific clustering
  - Association
  - Trend Analysis (…)

Outline for This Session

- Association Rule Mining
- Clustering
  - K-Means Clustering
  - Hierarchical Clustering
- Web Mining
- Text Mining
- **Case Study**

eljalao@up.edu.ph

Outline for This Session

- Association Rule Mining
- Clustering
  - K-Means Clustering
  - Hierarchical Clustering
- Web Mining
- Text Mining
- Case Study

eljalao@up.edu.ph

References

- Turban et. al. 2011, Decision Support and Business Intelligence Systems
(9th Ed., Prentice Hall)
- Tan et al. Intro to Data Mining Notes
- Notes and Datasets from Montgomery, Peck and Vining, Introduction to Linear Regression Analysis 4th Ed. Wiley
- Notes from G. Runger, ASU IEE 578
- Trevor Hastie, Rob Tibshirani, Friedman: Elements of Statistical Learning (2nd Ed.) 2009

eljalao@up.edu.ph