

การทำเหมืองข้อมูล

1101031 พื้นฐานวิทยาการข้อมูล 1 (DATA SCIENCE FOUNDATION I)

ผู้สอน: อ.ดร.พิชญ์สินี กิจวัฒนาการ

เอกสารประกอบการเรียน:
รศ.ดร.นิตยา เกิดประสม และ อ.ศรีญญา กาญจนวัฒนา

Introduction

2

- Knowledge discovery in databases = KDD
- KDD refers to the broad process of finding knowledge in data.
- KDD = data mining
 - knowledge extraction
 - information discovery
 - data archeology
 - etc.

Introduction

3

- “Data mining” used by statisticians, data analysts and MIS community.
- “KDD” used by AI and machine learning researchers.
- Related fields to KDD: machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and data warehousing.

Introduction

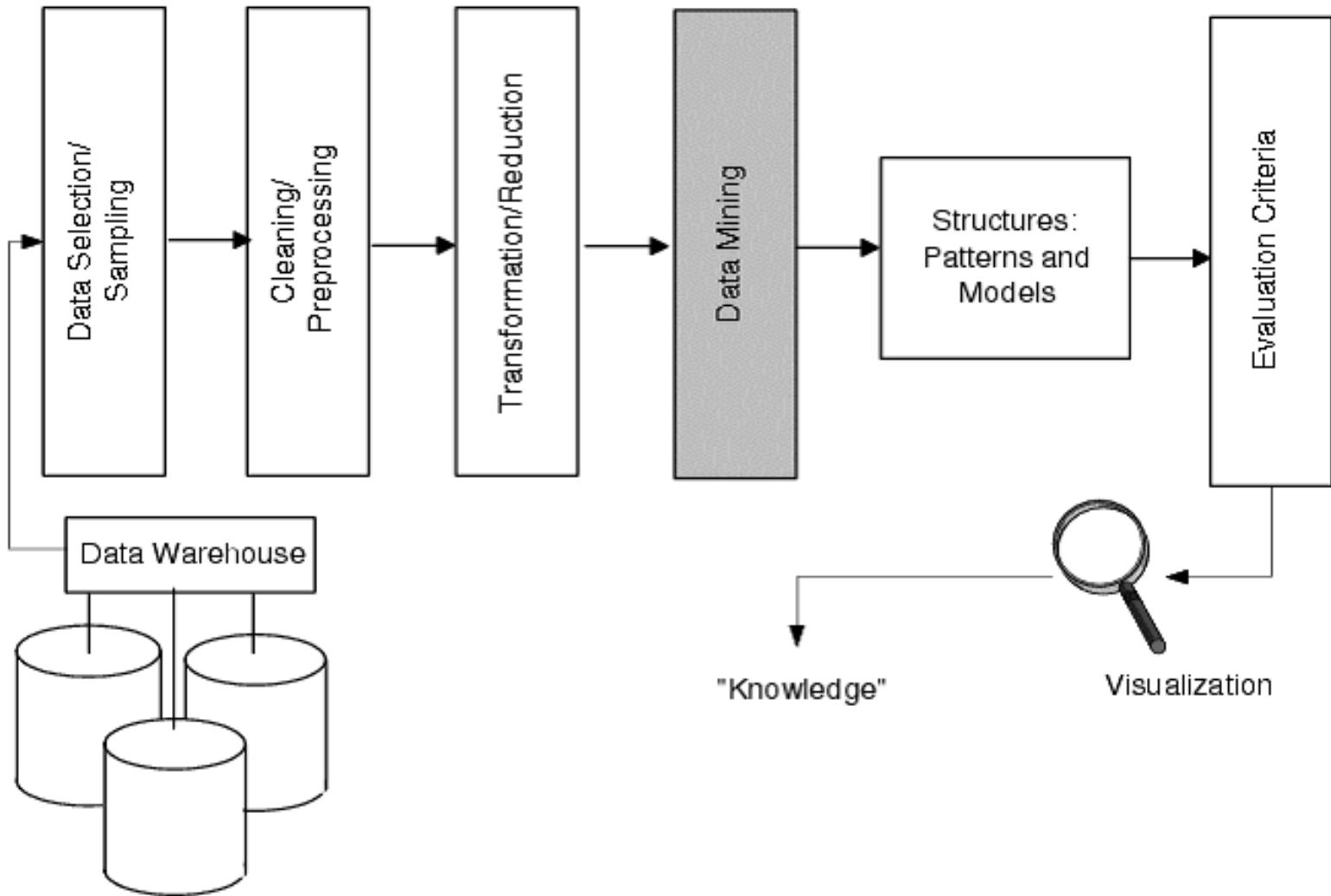
4

- **KDD** = the process of identifying valid, novel, potentially useful, and ultimately understandable structure in data.

- **Data Mining** = a step in the KDD process that, under acceptable computational efficiency limitations, enumerates structures (patterns or models) over the data.

KDD Process

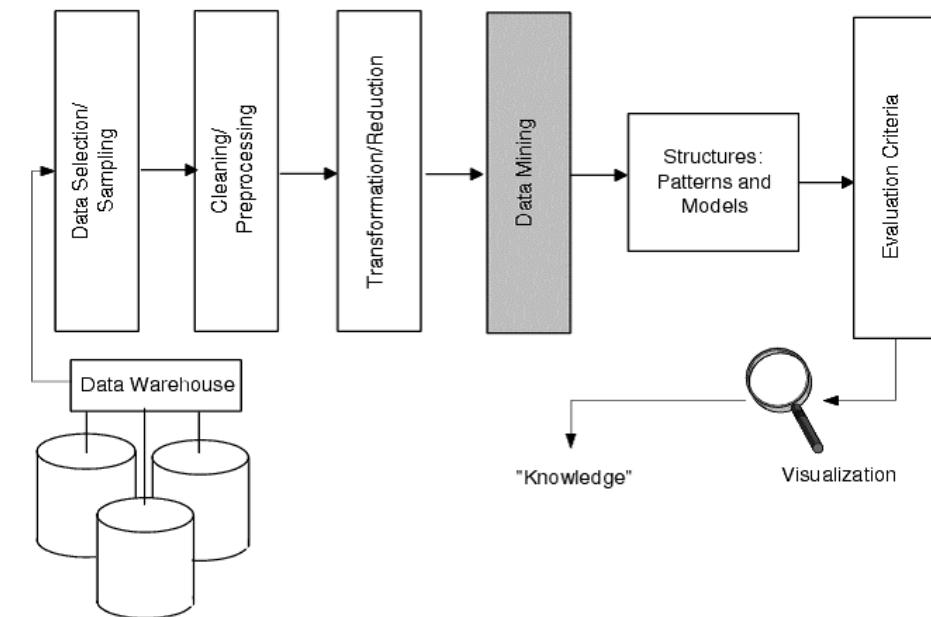
5



The task of data mining

6

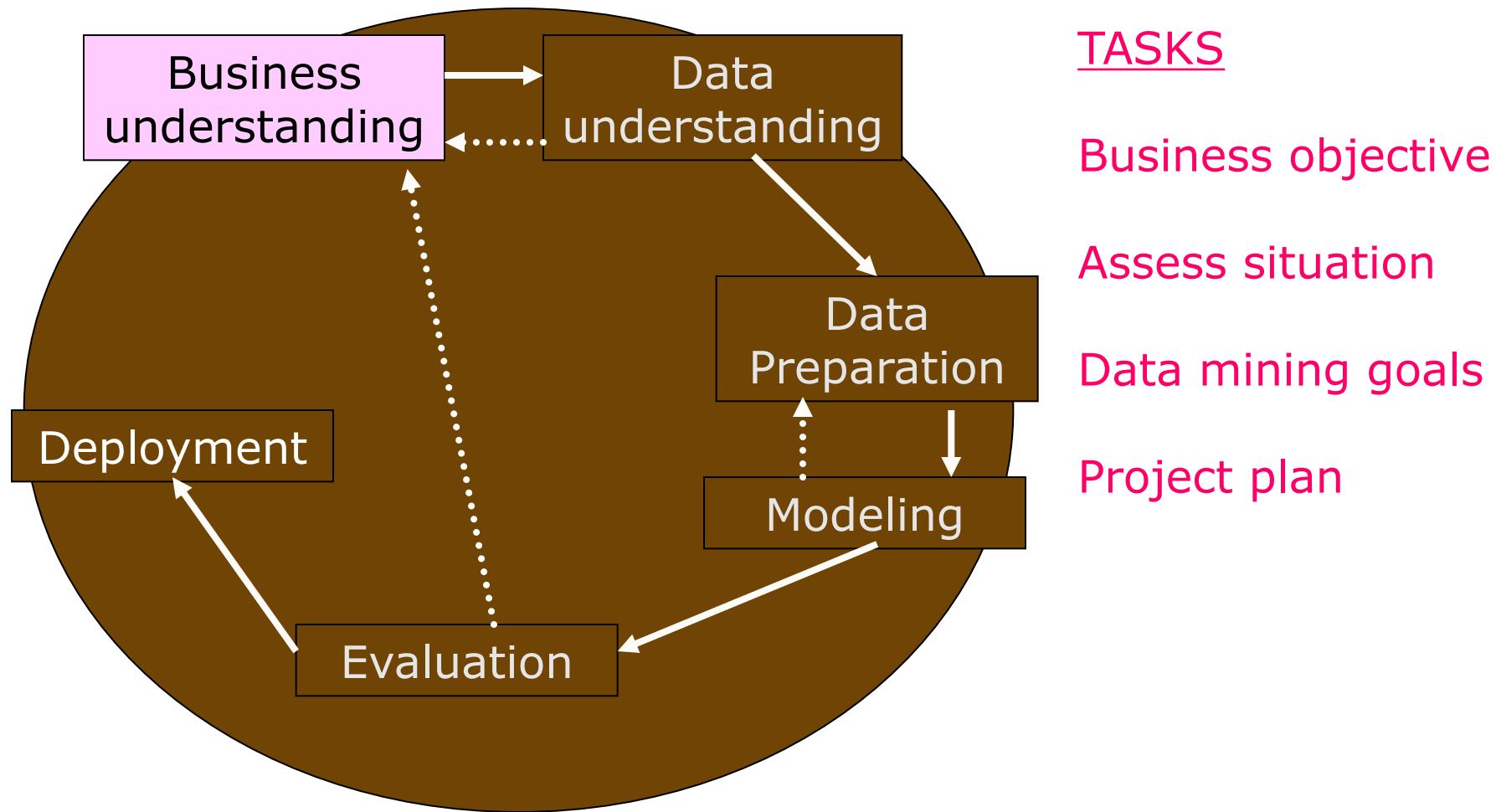
- Phase 1: data preparation
- Phase 2: data reduction
- Phase 3: data modeling/discovery
- Phase 4: solution analysis



CRISP-DM Methodology

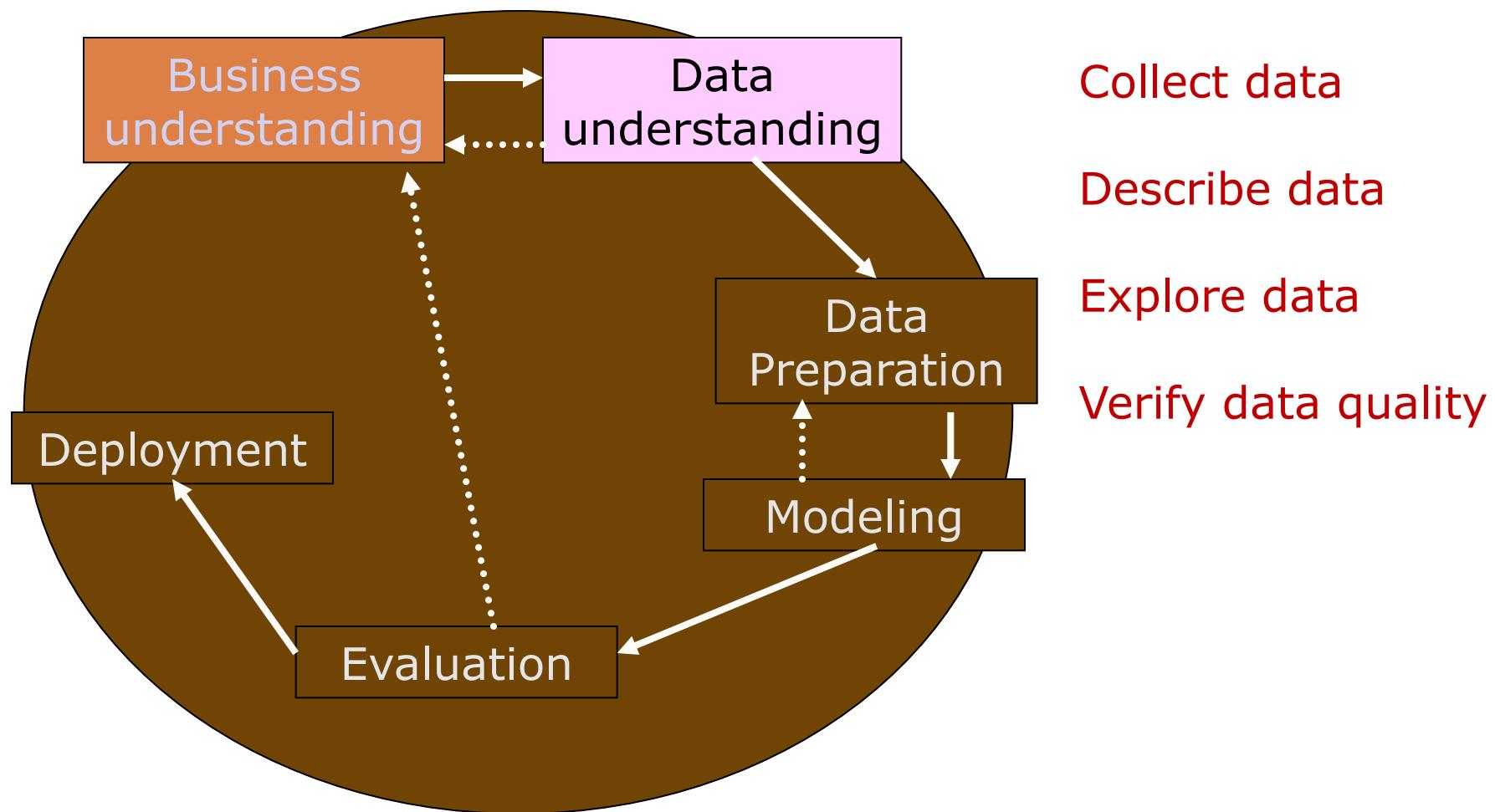
Cross Industry Standard Process for Data Mining

7



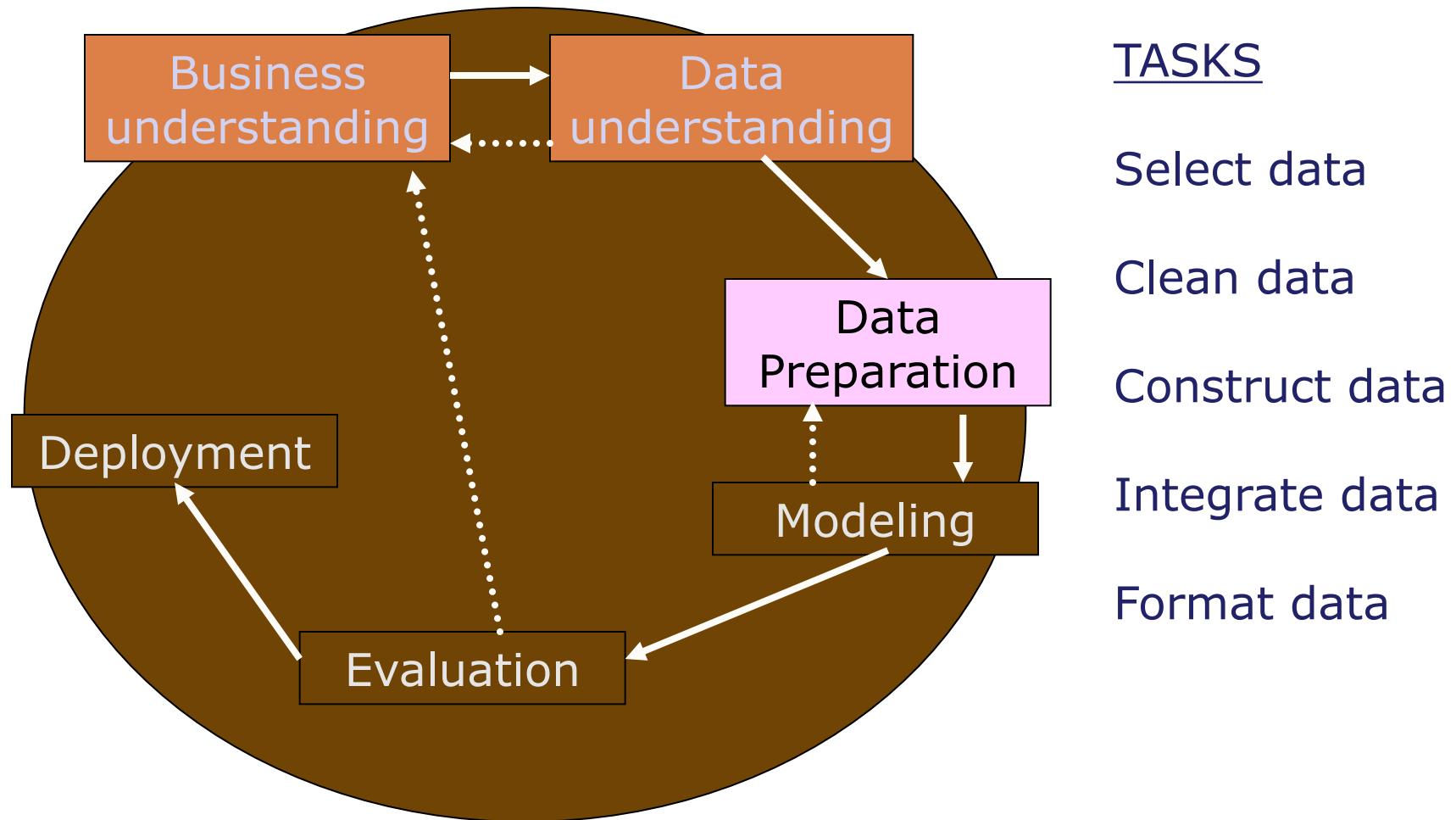
CRISP-DM Methodology

8



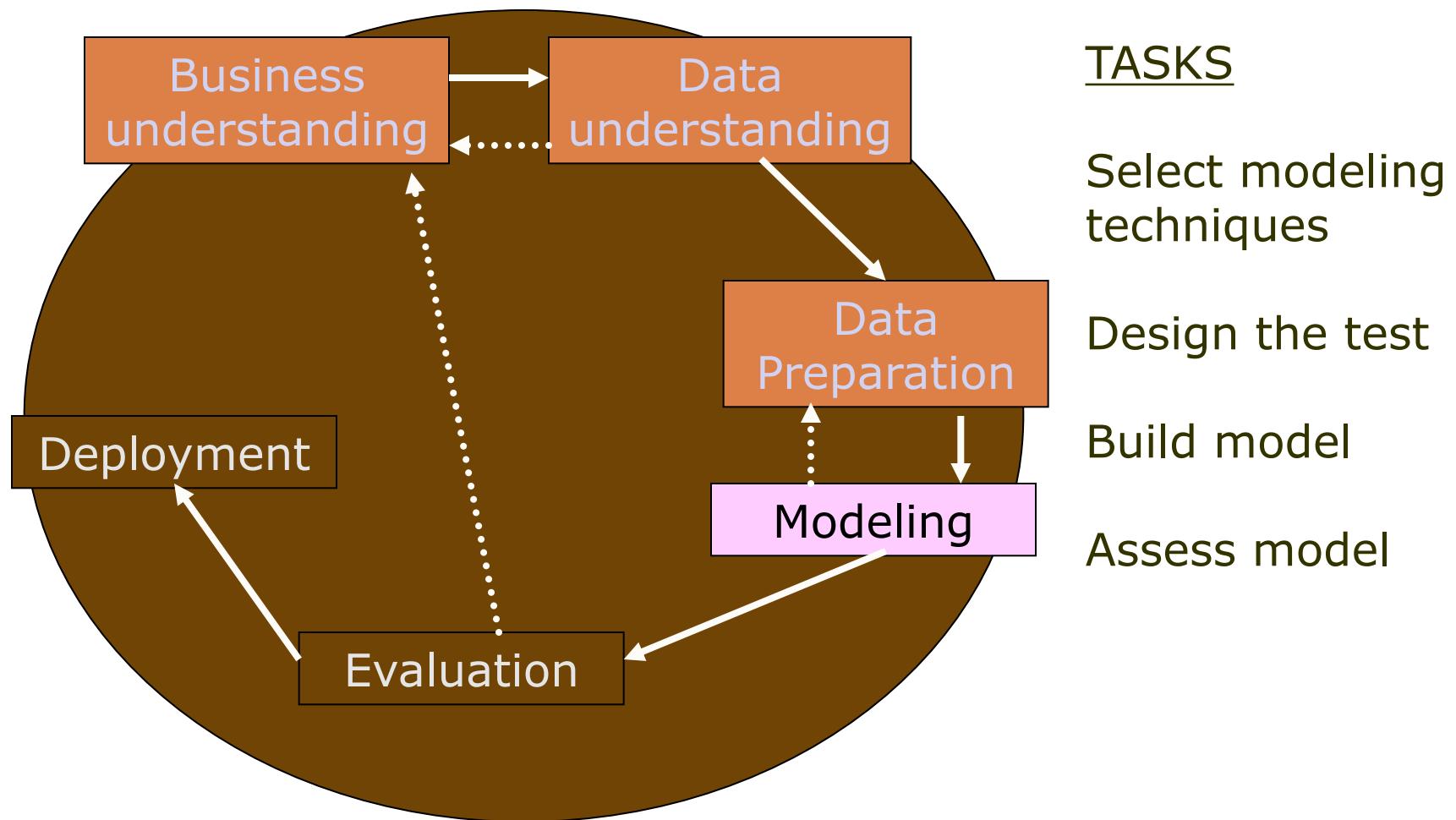
CRISP-DM Methodology

9



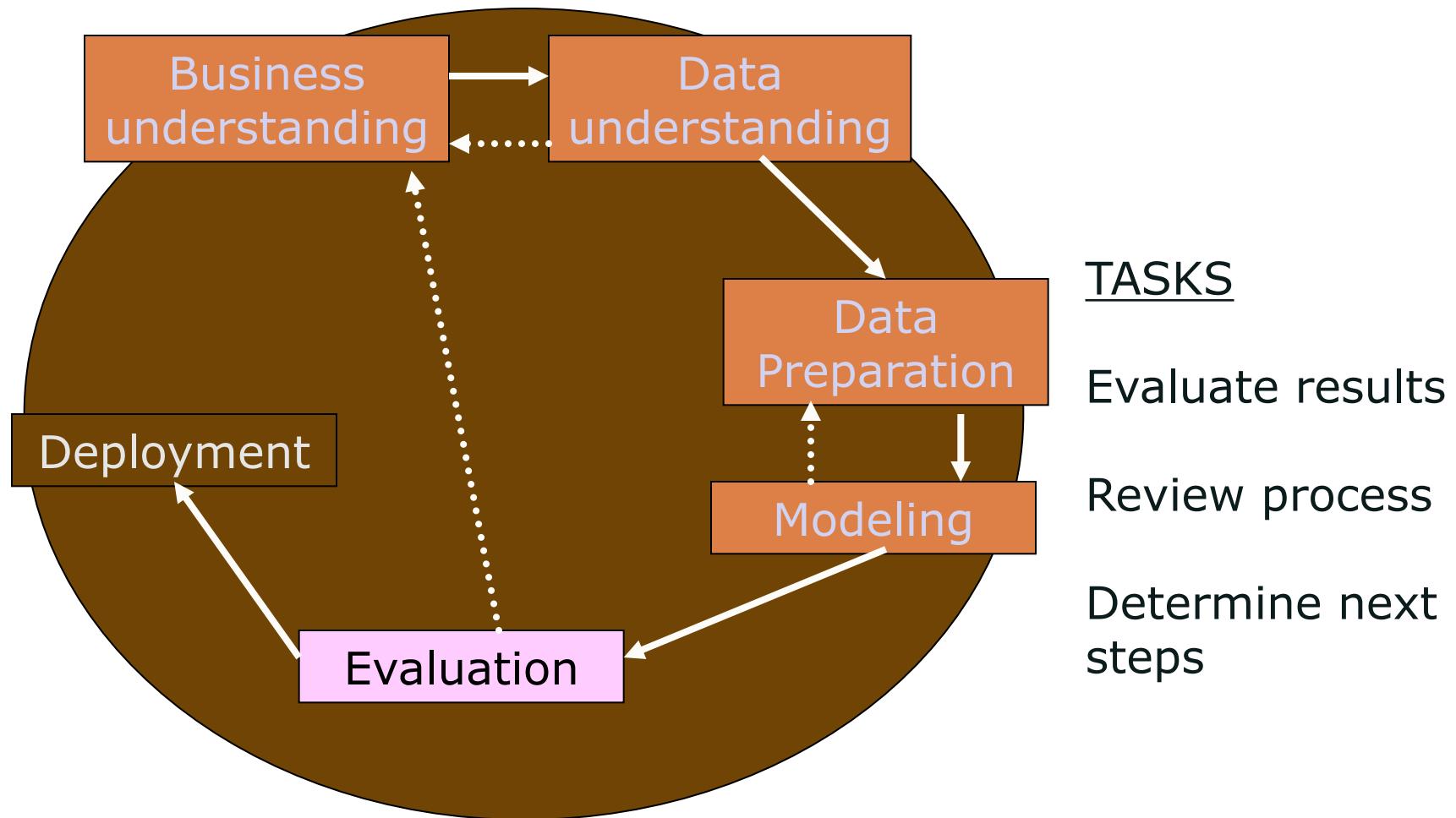
CRISP-DM Methodology

10



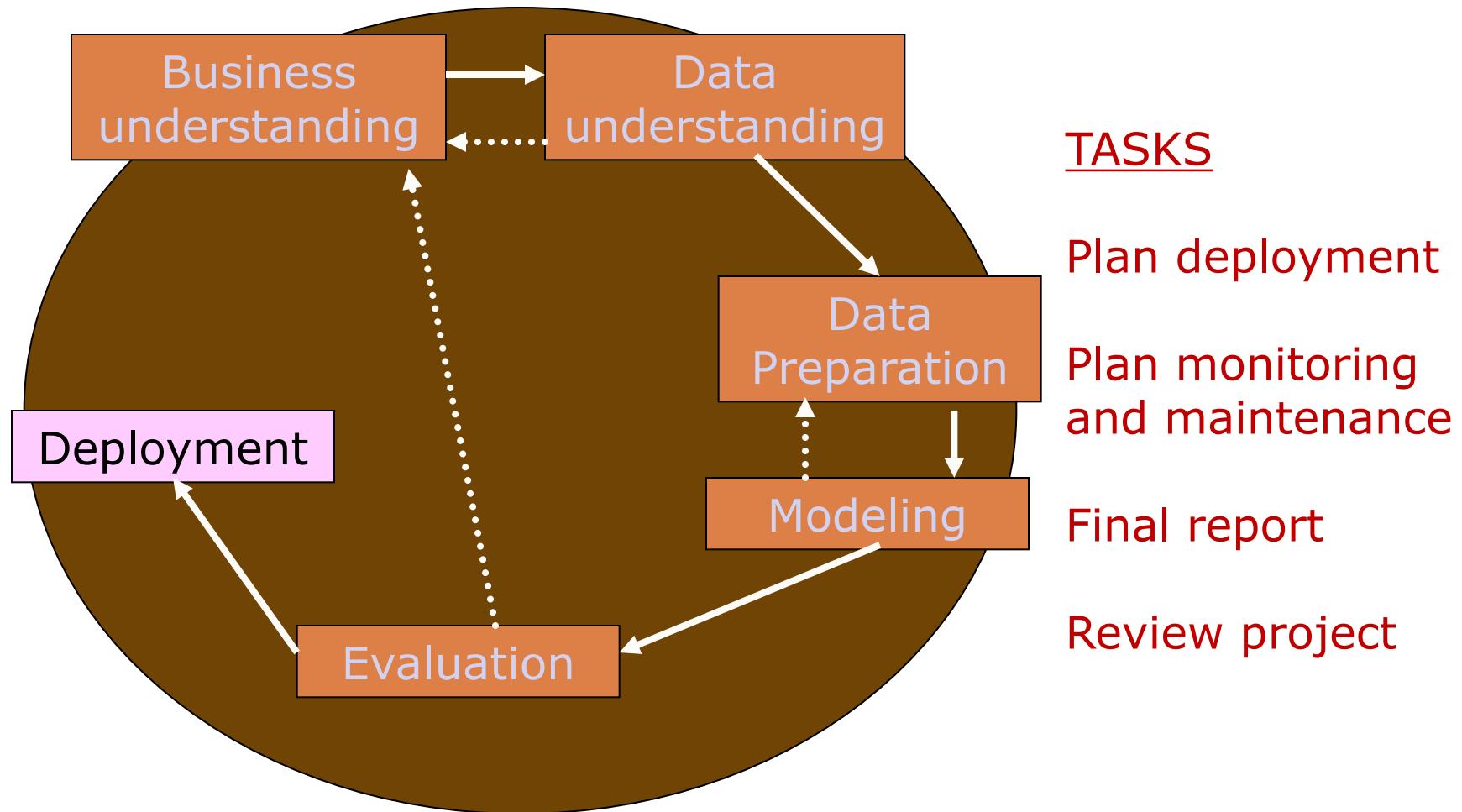
CRISP-DM Methodology

11



CRISP-DM Methodology

12



What is data mining?

13

The image shows a screenshot of a data mining application. At the top, there is a menu bar with options: Table, Edit, Generate, and Help. Below the menu is a table with four rows of data. The table has four columns with headers: id, age, gender, and responder. The data rows are as follows:

id	age	gender	responder
M12106	57	FEMALE	YES
M12108	58	MALE	NO
M12120	31	MALE	NO
M12138	36	FEMALE	YES

At the bottom of the table, there are navigation buttons: a left arrow, a right arrow, and a double-right arrow.

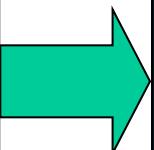
Discovering meaningful **patterns** in your data

What is data mining?

14

As the data grows...

Table Edit Generate Help			
id	age	gender	responder
ID12106	57	FEMALE	YES
ID12108	58	MALE	NO
ID12120	31	MALE	NO
ID12138	36	FEMALE	YES



id	age	gender	responder
ID12106	57	FEMALE	YES
ID12108	58	MALE	NO
ID12120	31	MALE	NO
ID12121	61	MALE	YES
ID12138	36	FEMALE	YES

The relationships become
more complicated

What is data mining?

Table Edit Generate Help

id	age	gender	region	income	married	children	car	save_act	current_act	mortgage	responder
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO	NO	NO	YES
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES	YES	NO	NO
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO
ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO
ID12106	57	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YES
ID12107	22	MALE	RURAL	8877.07	NO	0	NO	NO	YES	NO	YES
ID12108	58	MALE	TOWN	24946.6	YES	0	YES	YES	YES	NO	NO
ID12109	37	FEMALE	SUBURBAN	25304.3	YES	2	YES	NO	NO	NO	NO
ID12110	54	MALE	TOWN	24212.1	YES	2	YES	YES	YES	NO	NO



What data mining can do ?

Table 1.1 • Hypothetical Training Data for Disease Diagnosis

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

What data mining can do ?

Table 1.1 • Hypothetical Training Data for Disease Diagnosis

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

It can build
model !!

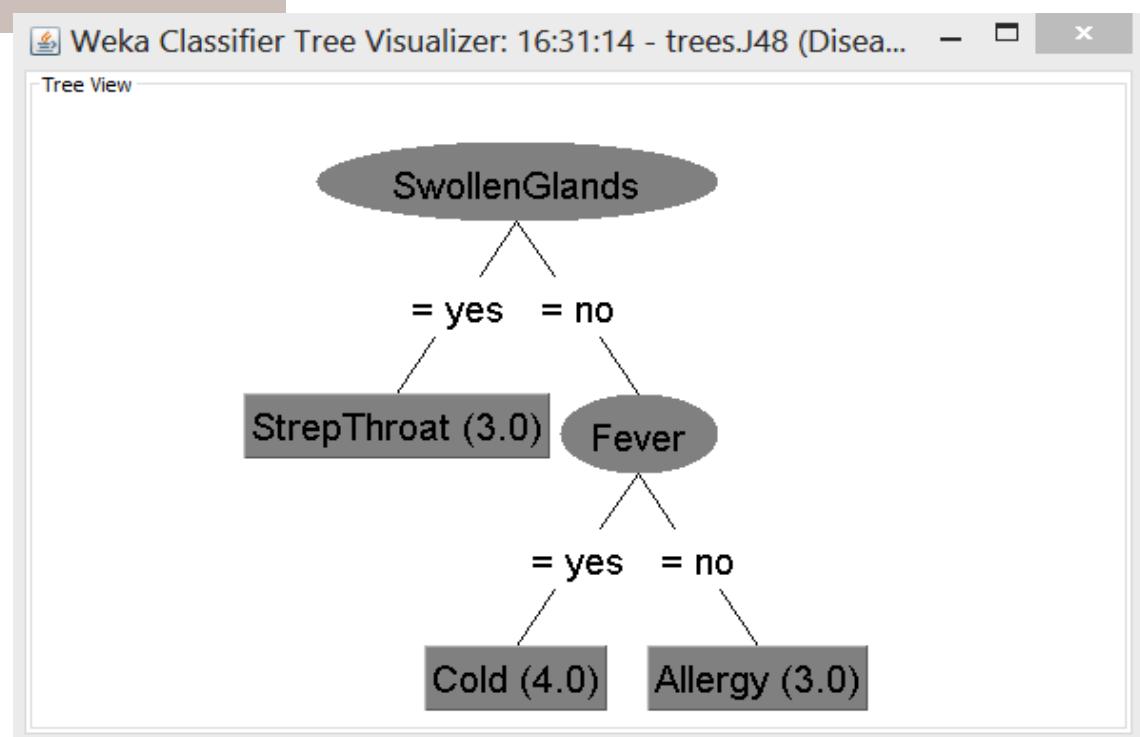
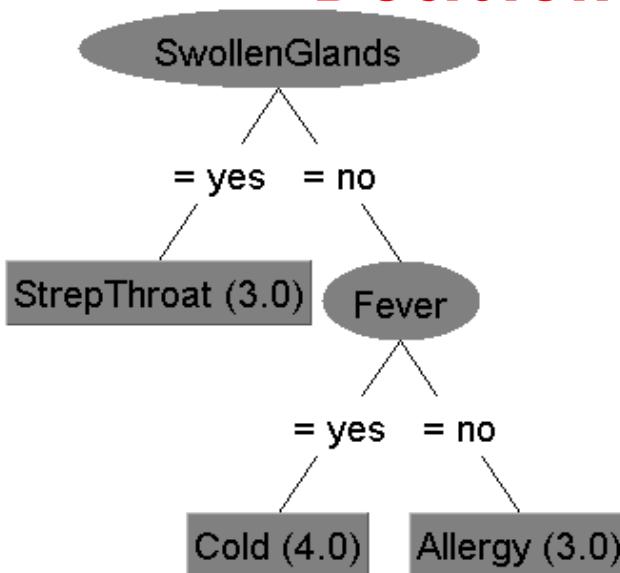


Table 1.1 • Hypothetical Training Data for Disease Diagnosis

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Decision Tree



1. IF *SwollenGlands* = Yes
THEN *Diagnosis* = *StrepThroat*
2. IF *SwollenGlands* = No & *Fever* = Yes
THEN *Diagnosis* = *Cold*
3. IF *SwollenGlands* = No & *Fever* = No
THEN *Diagnosis* = *Allergy*

Model as Rules

Use of the Decision Tree for Prediction

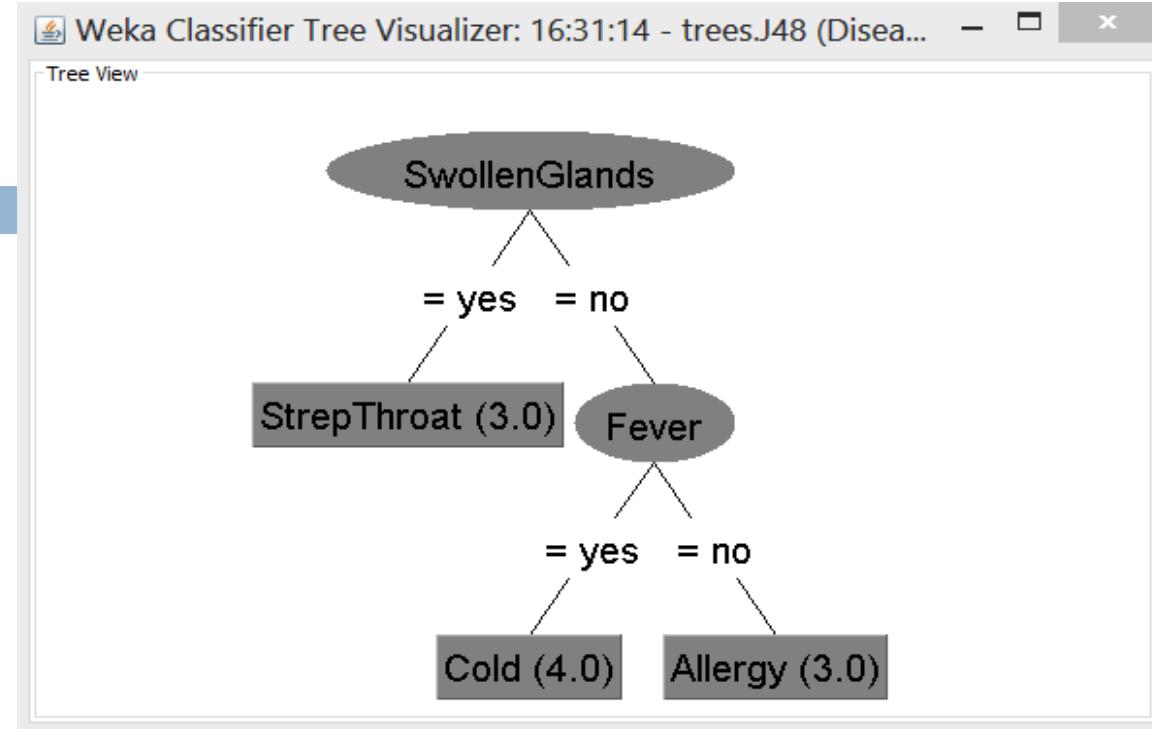


Table 1.2 • Data Instances with an Unknown Classification

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
11	No	No	Yes	Yes	Yes	?
12	Yes	Yes	No	No	Yes	?
13	No	No	No	No	Yes	?

What Data Mining Can Do?

20

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Discovering association

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper}, \text{Milk}\} \rightarrow \{\text{Beer}\}$

The Sad Truth About Diapers and Beer

21



Applications of KDD

22

- Medicine: drug side effects, genetic sequence analysis
- Finance: credit approval, bankruptcy prediction, stock market prediction
- Agricultural: soybean and tomato disease classification
- Social: demographic data, voting trend, election results
- Marketing and sales: retail shopping patterns, product analysis, sales prediction

Applications of KDD

23

- Insurance: detection of fraudulent and excessive claims
- Engineering: automotive diagnostic expert system
- Physics and chemistry: superconductivity research
- Law enforcement: tax and welfare fraud, fingerprint matching
- Space science: astronomy, space data analysis

Types of data mining problems

25

□ Prediction

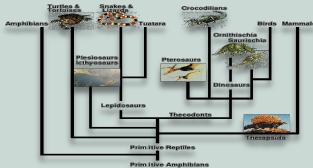
- classification
- regression
- time series

● Knowledge discovery

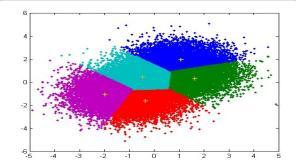
- deviation detection
- database segmentation
- clustering
- association rules
- summarization
- visualization
- text mining
- web mining

Data Mining Techniques

26



Supervised Learning: Classification and Prediction



Unsupervised Learning: Clustering



Association Rule Discovery



Other Techniques

- Deviation Detection
- Regression
- Sequential Pattern Discovery

Data Mining Software

27

□ Commercial Software

- SAS Enterprise Miner



- DB2 Intelligent Miner



- Microsoft SQL Server 2008



□ Open Source Software

- Weka



- RapidMiner



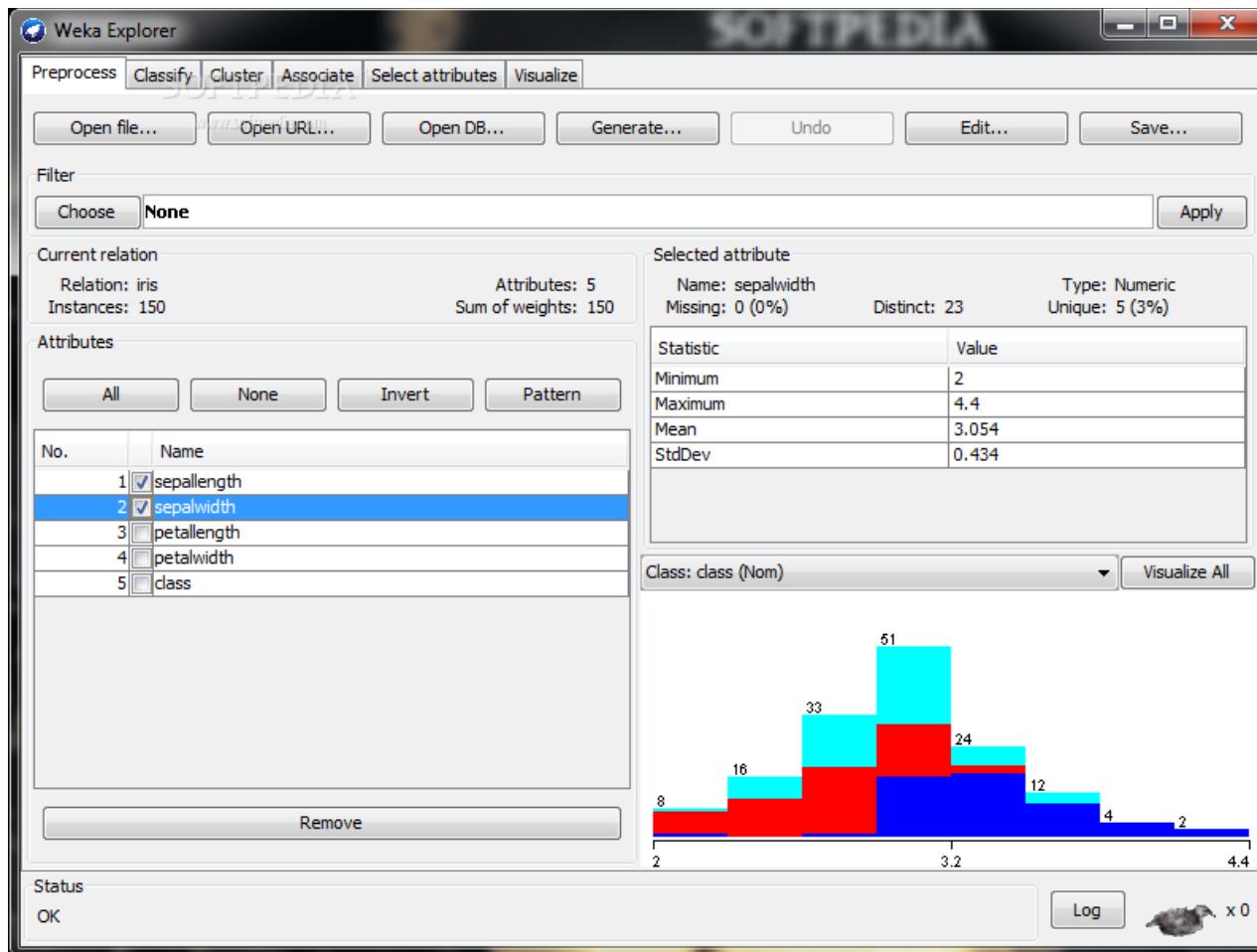
- KNIME (Konstanz Information Miner)



Data Mining Software (Cont.)

28

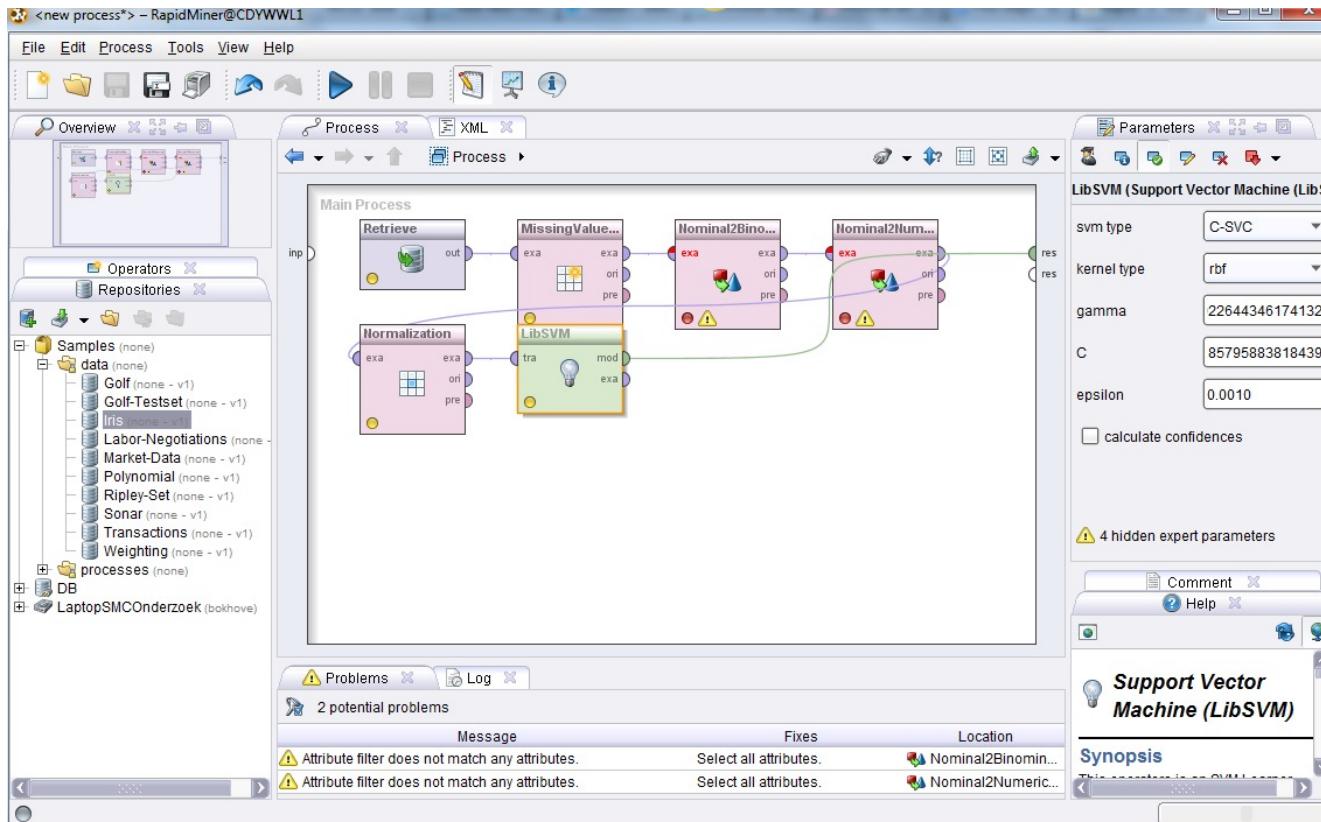
□ WEKA



Data Mining Software (Cont.)

29

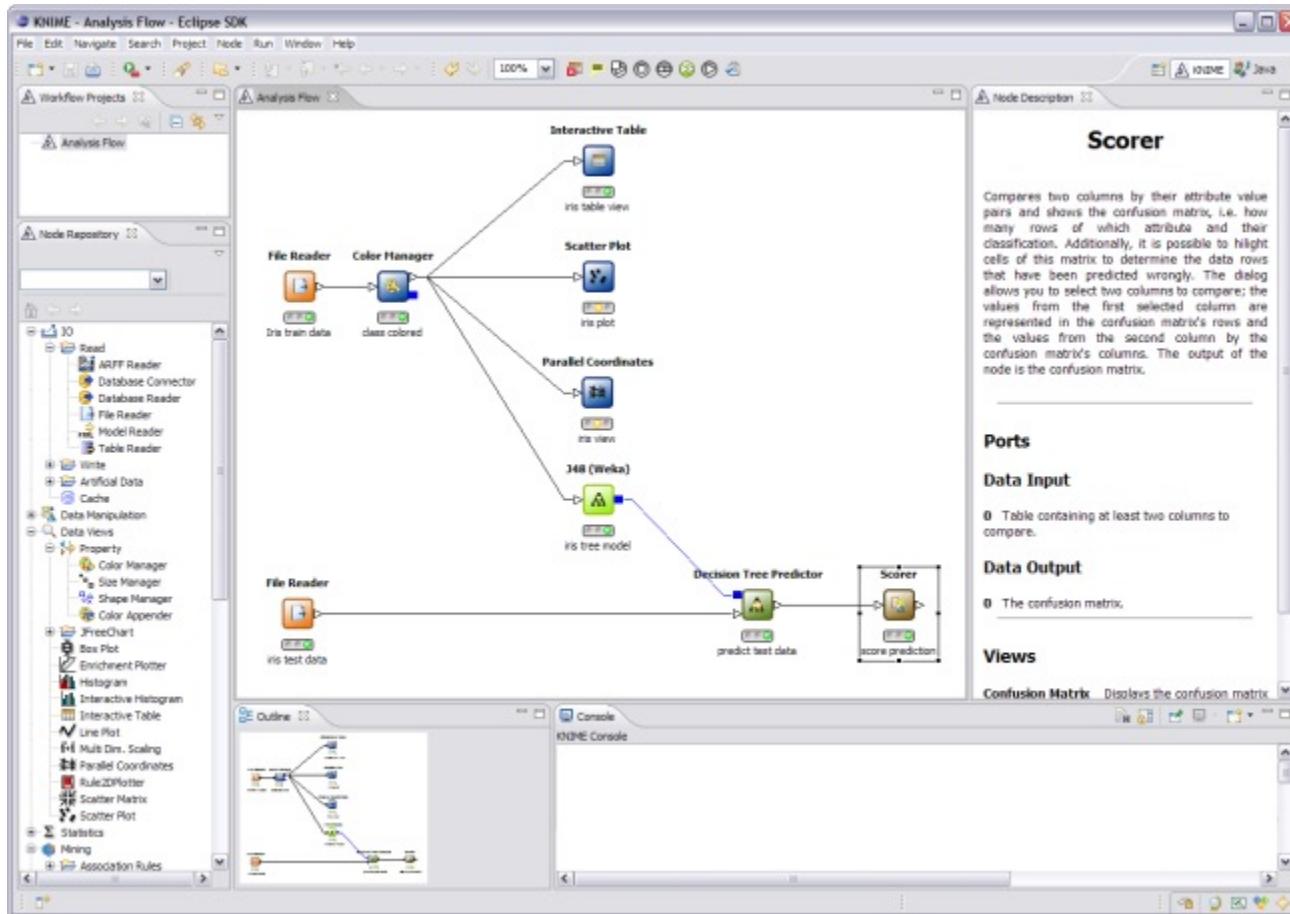
□ RapidMiner



Data Mining Software (Cont.)

30

□ KNIME



What is WEKA?

31

- **Waikato Environment for Knowledge Analysis**
 - It's a data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand.
 - Weka is also a bird found only on the islands of New Zealand.



Download and Install WEKA

32

- Website: <http://www.cs.waikato.ac.nz/ml/weka/>
- Support multiple platforms (written in java):
 - Windows, Mac OS X and Linux

The image shows two screenshots of the WEKA website. The left screenshot displays the main homepage with a large blue circular logo featuring a white bird, the title 'WEKA', and the subtitle 'The workbench for machine learning'. Below this is a detailed description of the software and links to 'Download', 'Docs', 'Courses', and 'Book'. The right screenshot shows a sub-page titled 'Downloading and Installing Weka' under the 'Weka Wiki' header. This page includes a sidebar with links for 'Snapshots', 'Stable version', 'Windows', 'Mac OS', 'Linux', 'Other platforms', 'Developer version', 'Requirements', 'Documentation', 'Getting help', 'Citing Weka', 'Literature', and 'Development'. The main content area contains sections for 'Snapshots' (describing nightly builds), 'Stable version' (describing the latest stable release), 'Windows' (with a link to download a self-extracting executable), and 'Mac OS' (with a link to download a disk image). Both pages have a dark header bar with navigation links: 'Weka', 'Book', 'Courses', 'Blog', and 'Wiki'.

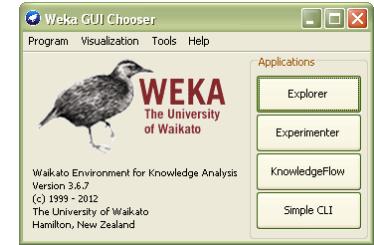
Main GUI

33

□ Three graphical user interfaces

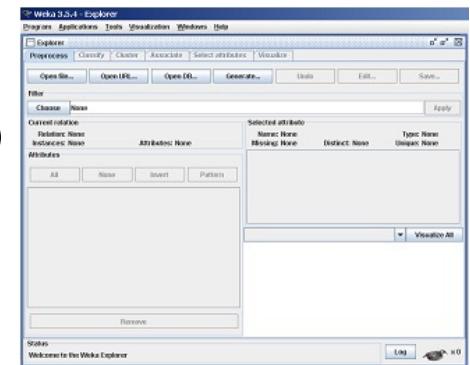
□ “The Explorer” (exploratory data analysis)

- ใช้งานโดยการคลิกผ่านหน้าจอ interface



□ “The Experimenter” (experimental environment)

- ใช้เพื่อหาค่า parameter ที่เหมาะสมในการทำงานของแต่ละโปรแกรม

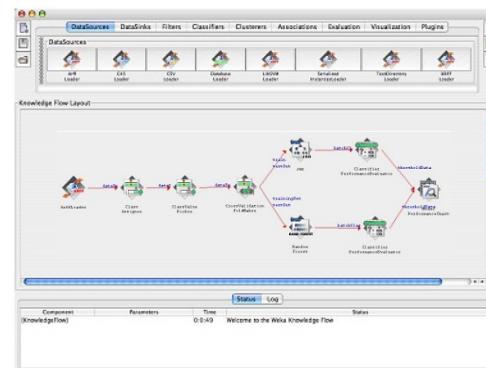
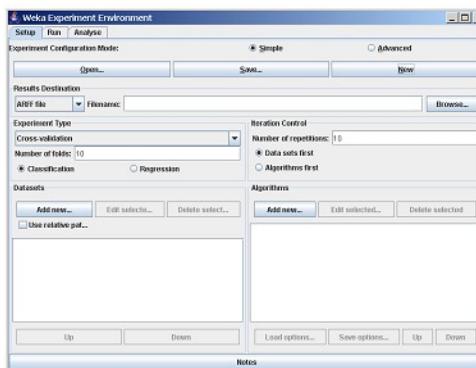


□ “The KnowledgeFlow” (new process model inspired interface)

- นำส่วนต่าง ๆ ของ weka มาเชื่อมต่อกัน

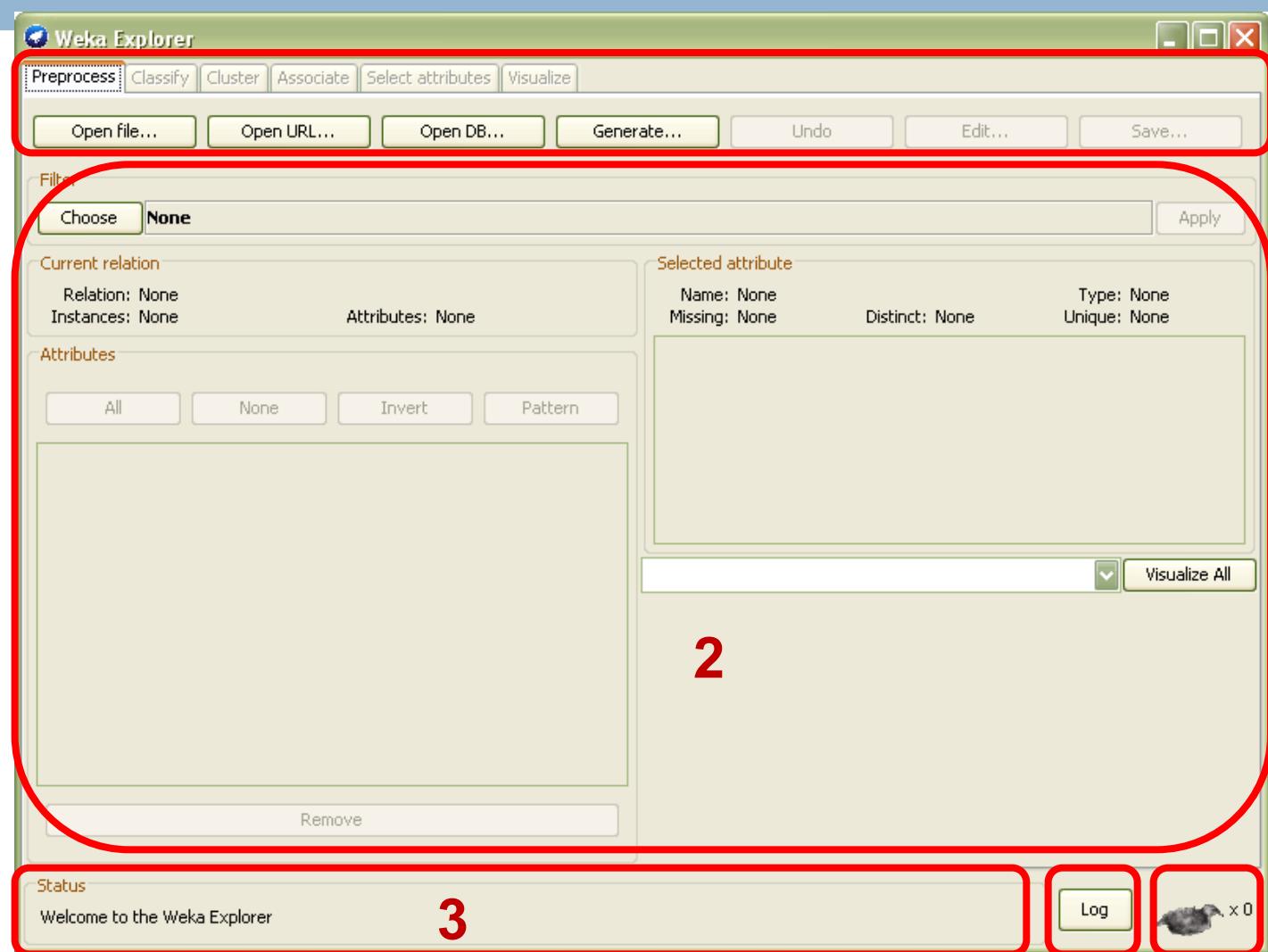
□ “Simple CLI”

- ใช้งานโดยการพิมพ์คำสั่งต่าง ๆ

A screenshot of the SimpleCLI window. It has a text area at the top with the text "Welcome to the WEKA SimpleCLI". Below that, it says "Enter commands in the textfield at the bottom of the window. Use the up and down arrows to move through previous commands." It then provides instructions for command completion: "Command completion for classnames and files is initiated with <Tab>. In order to distinguish between files and classnames, file names must be either absolute or start with '.'." It also shows how to use the delete key: "<Alt>+<BackSpace> is used for deleting the text in the commandline in chunks." At the bottom, there is a text input field and a status bar at the bottom right.

WEKA Explorer

34



Load data into Weka

35

□ ข้อมูลที่ใช้เป็น input สำหรับ Weka

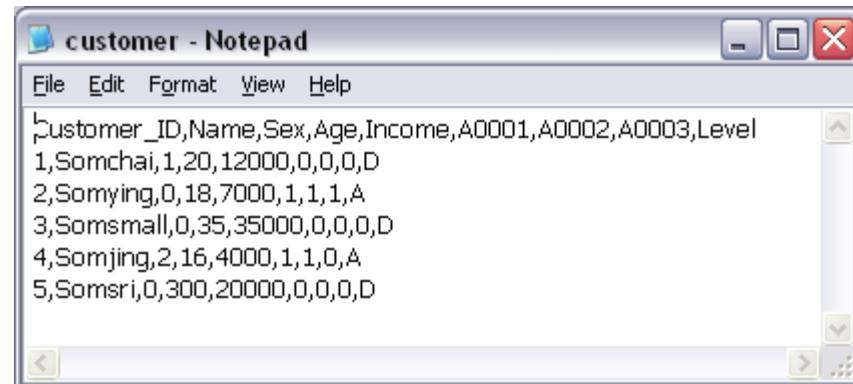


Comma-Separated Value (CSV)

36

- ใช้เครื่องหมาย comma (,) แบ่งระหว่างแต่ละบิวต์
- เป็นไฟล์ที่ใช้กันอย่างแพร่หลาย
- Export ได้จากโปรแกรม Excel หรือจากฐานข้อมูล MySQL, SQL Server

	A	B	C	D	E	F	G	H	I
1	Customer_ID	Name	Sex	Age	Income	A0001	A0002	A0003	Level
2	1	Somchai	1	20	12000	0	0	0	D
3	2	Somying	0	18	7000	1	1	1	A
4	3	Somsmall	0	35	35000	0	0	0	D
5	4	Somjing	2	16	4000	1	1	0	A
6	5	Somsri	0	300	20000	0	0	0	D



Attribute-Relation File Format (ARFF)

```
@relation customer  
  
@attribute Customer_ID numeric  
@attribute Name string  
@attribute Sex {0,1,2}  
@attribute Age numeric  
@attribute Income numeric  
@attribute A0001 numeric  
@attribute A0002 numeric  
@attribute A0003 numeric  
@attribute Level {D,A}
```

```
@data  
1,Somchai,1,20,12000,0,0,0,D  
2,Somying,0,18,7000,1,1,1,A  
3,Somsmall,0,35,35000,0,0,0,D  
4,Somjing,2,16,4000,1,1,0,A  
5,Somsri,0,300,20000,0,0,0,D
```

- Tag พิเศษที่มีในส่วน header
 - ▢ @relation <relation-name>
 - ▢ ใช้ในการบอกชื่อเรียกชุดข้อมูล
 - ▢ @attribute <attribute-name> <datatype>
 - ▢ ชื่อแอตทริบิวต์
 - ▢ ประเภทของข้อมูลในแอตทริบิวต์นั้นๆ
- Tag พิเศษที่มีในส่วน data
 - ▢ @data
 - ▢ บรรทัดถัดไปจะนี้จะเป็นส่วนของข้อมูล
 - ▢ ใช้เครื่องหมาย comma (,) แบ่งแต่ละแอตทริบิวต์
 - ▢ % แทน คอมเมนต์ (comment) หรือคำอธิบาย

Attribute in ARFF

38

```
@relation customer
```

```
@attribute Customer_ID numeric  
@attribute Name string  
@attribute Sex {0,1,2}  
@attribute Age numeric  
@attribute Income numeric  
@attribute A0001 numeric  
@attribute A0002 numeric  
@attribute A0003 numeric  
@attribute Level {D,A}
```

```
@data
```

```
1,Somchai,1,20,1,2000,0,0,0,D
```

```
2,Somying,0,18,7,000,1,1,1,A
```

```
...
```

- ประเกทข้อมูลในแต่ละแอตทริบิวต์
- ข้อมูลที่เป็นตัวเลข (Numeric) ที่ต่อเนื่อง
 - จำนวนเต็ม (Integer), เลขทศนิยม (real)
 - ใช้ keyword “numeric”
 - Customer_ID
 - Age
 - Income
- ข้อมูลที่ไม่ใช่ตัวเลข (Nominal) ซึ่งไม่ต่อเนื่อง
 - ไม่ใช่ข้อมูลที่เป็นตัวเลขที่เรียงต่อกัน
 - ชื่อคน, สิงของ, สถานที่, เพศ
 - ระบุค่าที่เป็นไปได้ทั้งหมด ในรูปแบบของ Set
 - Sex {0,1,2}

Data in ARFF

39

@relation customer

@attribute Customer_ID numeric

@attribute Name string

@attribute Sex {0,1,2}

@attribute Age numeric

@attribute Income numeric

@attribute A0001 numeric

@attribute A0002 numeric

@attribute A0003 numeric

@attribute Level {D,A}

@data

1,Somchai,1,20,1,2000,0,0,0,D

2,Somying,0,18,7000,1,1,1,A

...

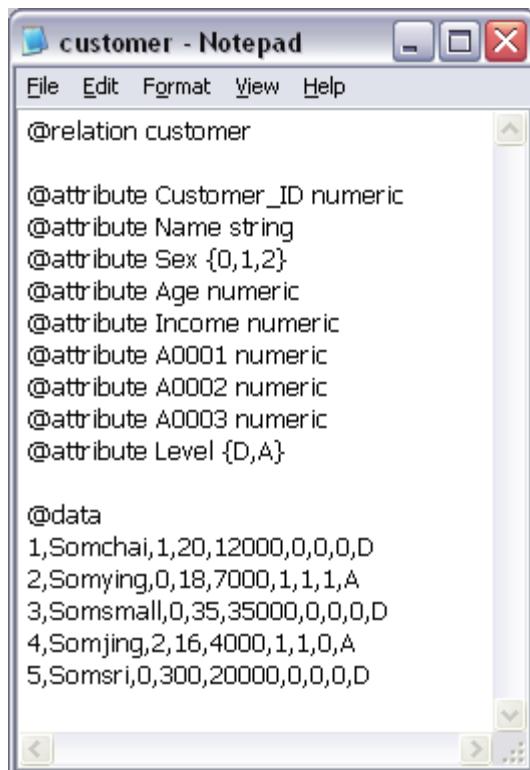
- ข้อมูลในแอ็ตทริบิวต์ต่าง ๆ ที่ใช้ไว้เคราะห์
- อุปบรรทัดถัดจาก Tag พิเศษ @data
- บรรทัดละ 1 อินสแตนซ์
- ข้อมูลที่ขาดหายจะถูกแทนด้วยเครื่องหมาย ?
- ข้อมูลในแอ็ตทริบิวต์ใดที่ระบบประเภทเป็น ไม่ใช่ตัวเลข (Nominal) จะต้องมีค่าตามที่ระบุไว้ในส่วนแอ็ตทริบิวต์เท่านั้น
(อักษรตัวเล็ก-ตัวใหญ่ถือว่าเป็นคนละค่ากัน)
- หากข้อมูลที่ไม่ใช่ตัวเลขนั้นมีช่องว่างสมญู่ จะต้องอุปภัยในเครื่องหมาย single quoted ("")

Difference between ARFF & CSV

40

ARFF file

- มีรายละเอียดของแอ็ตทริบิวต์



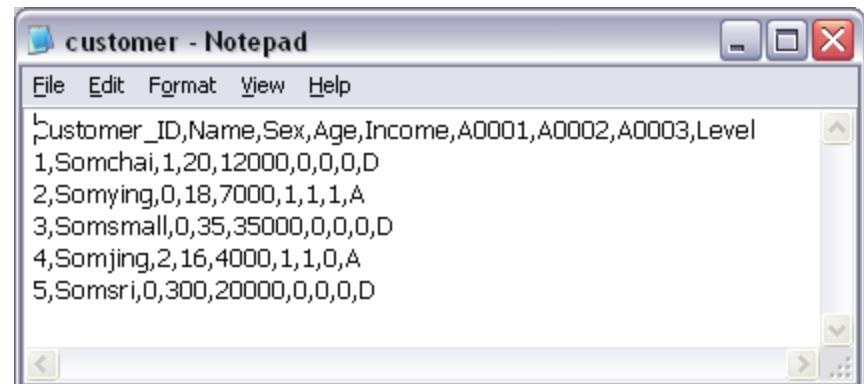
```
customer - Notepad
File Edit Format View Help
@relation customer

@attribute Customer_ID numeric
@attribute Name string
@attribute Sex {0,1,2}
@attribute Age numeric
@attribute Income numeric
@attribute A0001 numeric
@attribute A0002 numeric
@attribute A0003 numeric
@attribute Level {D,A}

@data
1,Somchai,1,20,12000,0,0,0,D
2,Somying,0,18,7000,1,1,1,A
3,Somsmall,0,35,35000,0,0,0,D
4,Somjing,2,16,4000,1,1,0,A
5,Somsri,0,300,20000,0,0,0,D
```

CSV file

- ไม่มีรายละเอียดของแอ็ตทริบิวต์

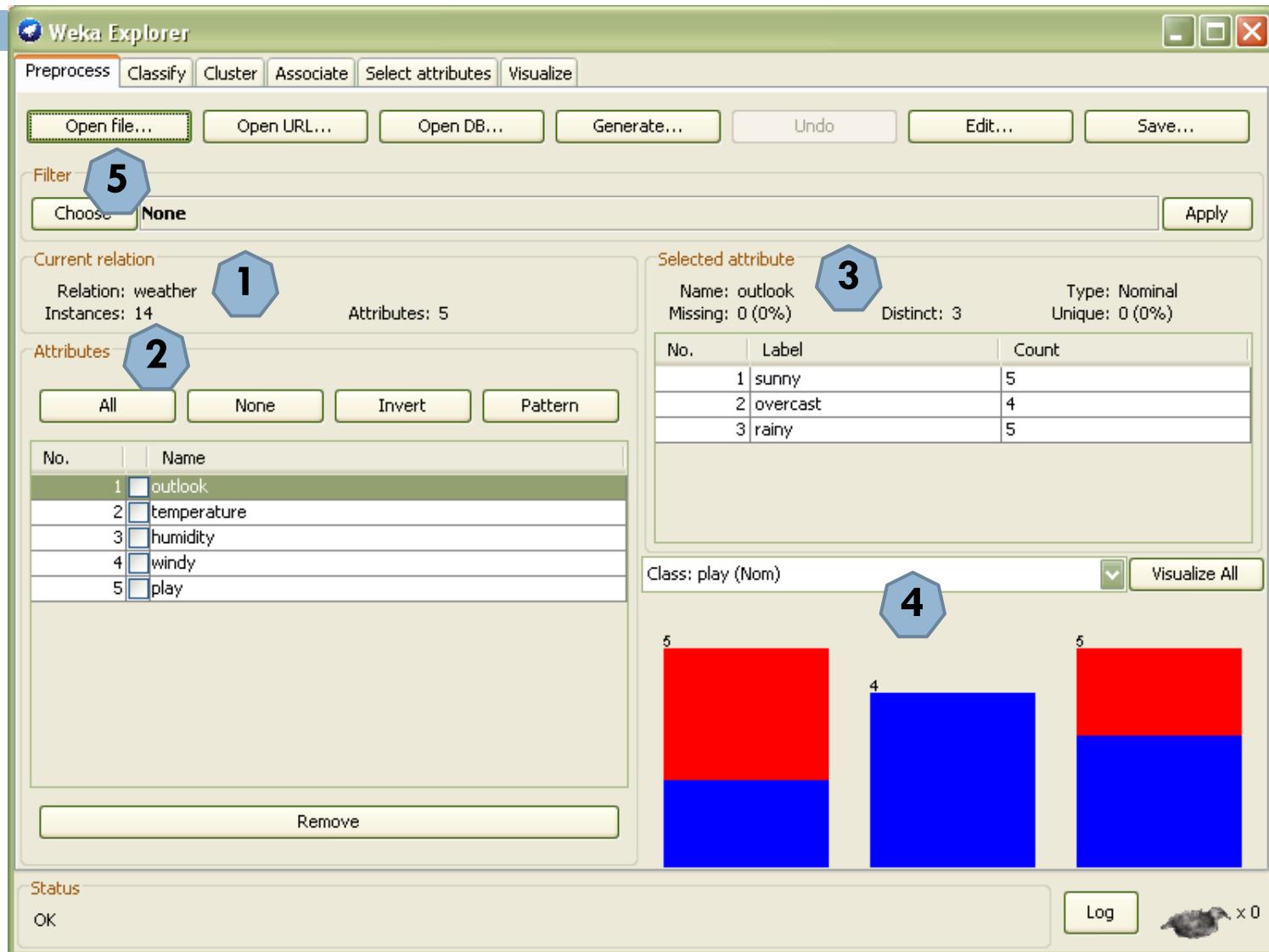


```
customer - Notepad
File Edit Format View Help
Customer_ID,Name,Sex,Age,Income,A0001,A0002,A0003,Level
1,Somchai,1,20,12000,0,0,0,D
2,Somying,0,18,7000,1,1,1,A
3,Somsmall,0,35,35000,0,0,0,D
4,Somjing,2,16,4000,1,1,0,A
5,Somsri,0,300,20000,0,0,0,D
```

Load data to Weka

ที่ tab Preprocess: กดปุ่ม Open file ... > เลือกไฟล์ C:\Program Files\Weka-3-8-4\data\weather.numeric.arff

41



Example data: Weather

42

- สภาพภูมิอากาศมีผลต่อการแข่งขันเบสบอล (baseball) ในอเมริกา
 - Outlook - ทัศนียภาพ
 - Temperature - อุณหภูมิ
 - Humidity - ความชื้น
 - Wind - ลม
- เก็บข้อมูลสภาพภูมิอากาศย้อนหลัง 14 วัน
- ไฟล์ชื่อ weather.nominal.arff อยู่ใน C:\Program Files\Weka-3-8-4\data

Load data to Weka (Cont.)

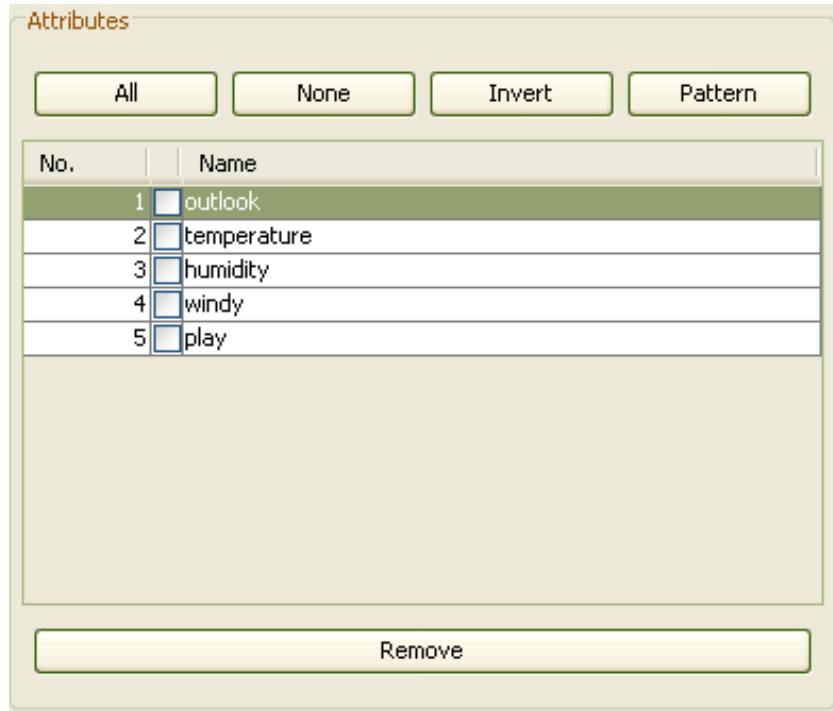
□ 1: Current Relation



- เป็นการนับรายการและอีดของชุดข้อมูลที่โหลดเข้ามา
 - Relation: ชื่อของชุดข้อมูลที่เลือก
 - Instances: จำนวนแถวในไฟล์ข้อมูล
 - Attributes: จำนวนคอลัมน์ในไฟล์ข้อมูล

Load data to Weka (Cont.)

□ 2: Attributes



- เป็นส่วนที่ช่วยในการจัดการและแทรกตัวอย่างในชุดข้อมูล
- **ปุ่ม All** ใช้ในการเลือกทุกแอตทริบิวต์
- **ปุ่ม None** ใช้ในการเคลียร์แอตทริบิวต์ที่ไม่เลือกอยู่ก่อนแล้ว
- **ปุ่ม Invert** ใช้ในการสลับสถานะของแอตทริบิวต์ ระหว่างถูกเลือกและไม่ถูกเลือก
- **ปุ่ม Pattern** ใช้ในการเลือกแอตทริบิวต์ที่มีชื่อตามเงื่อนไขที่กำหนด
 - ใช้ Regular Expression
- **ปุ่ม Remove** จะเป็นการลบแอตทริบิวต์ที่เลือกออก

Load data to Weka (Cont.)

□ 3: Selected Attribute

Selected attribute			
Name:	outlook	Type:	Nominal
Missing:	0 (0%)	Distinct:	3
No.	Label	Count	
1	sunny	5	
2	overcast	4	
3	rainy	5	

แอตทริบิวต์ที่มีข้อมูลเป็นประเภท

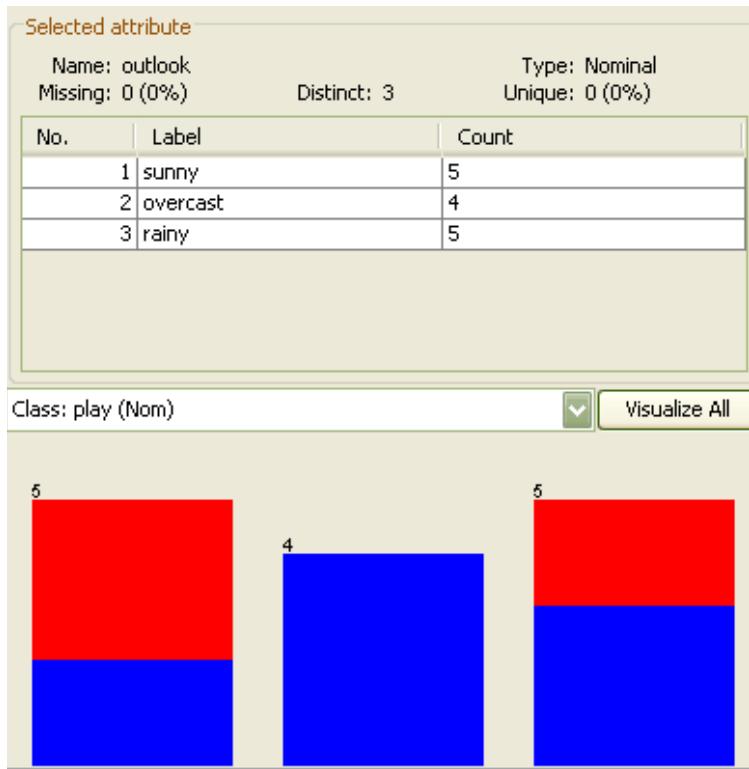
Selected attribute			
Name:	temperature	Type:	Numeric
Missing:	0 (0%)	Distinct:	12
Unique:	10 (71%)		
Statistic	Value		
Minimum	64		
Maximum	85		
Mean	73.571		
StdDev	6.572		

แอตทริบิวต์ที่มีข้อมูลเป็นตัวเลข

- แสดงรายละเอียดของแอตทริบิวต์ที่เลือกอยู่
 - Name: ชื่อของแอตทริบิวต์
 - Type: ประเภทข้อมูลในแอตทริบิวต์
 - Numeric ข้อมูลที่มีลักษณะเป็นตัวเลข หรือ เป็นเชิงปริมาณ
 - Nominal ข้อมูลที่มีลักษณะเป็นประเภท หรือ ไม่ใช่ตัวเลข
 - Missing: จำนวนข้อมูลในแอตทริบิวต์ที่ขาดหายไป
 - Distinct: จำนวนของข้อมูลที่เป็นไปได้ทั้งหมด เช่น sunny, overcast, rainy
 - Unique: จำนวนข้อมูลที่มีการปรากฏขึ้นแค่ครั้งเดียวในแอตทริบิวต์

Load data to Weka (Cont.)

□ 4: Visualization



ในตัวอย่างเลือกแอดทริบิวต์ outlook เทียบกับแอดทริบิวต์ play

- สีแดง คือ ค่า No ในแอดทริบิวต์ play
- สีน้ำเงิน คือ ค่า Yes ในแอดทริบิวต์ play

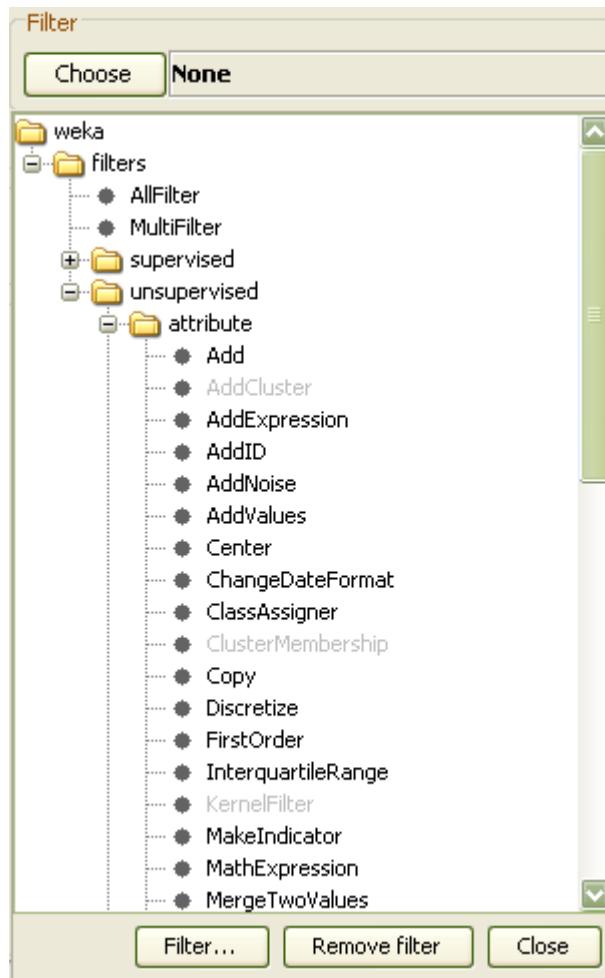
- กราฟแท่งแบ่งแยกตามค่าที่ group ได้ในแอดทริบิวต์
- แสดงสัดส่วนของจำนวน อินสแตนซ์ (ແລວ) เทียบกับอีก แอดทริบิวต์ที่เลือก
 - ปกติแล้วจะเทียบกับคลาส (class) หรือข้อมูลที่ต้องการคาดเดาหรือทำนาย
- ปุ่ม Visualize All ใช้แสดงกราฟของทุกแอดทริบิวต์

Load data to Weka (Cont.)

47

□ 5: Filter

□ ส่วนสำคัญของ Preprocess



The screenshot shows the Weka Filter dialog box. The 'filters' category is expanded, revealing various filter options under 'attribute'. The 'ReplaceMissingValue' option is highlighted with a blue square icon. The 'Apply' button is visible at the top right.

- แก้ไขข้อมูลที่ผิดพลาด
 - **ReplaceMissingValue:** เพิ่มข้อมูลที่ขาดหายไป (missing value)
- การค้นหา Outlier
 - **InterquartileRange:** พิจารณาจากการกระจายตัวของข้อมูล
- แปลงข้อมูล
 - **Discretize:** แปลงข้อมูลตัวเลข (numeric) ให้เป็นข้อมูลประเภท (nominal)
 - **StringToNominal:** แปลงข้อมูลตัวอักษร (string) ให้เป็นข้อมูลประเภท (nominal)

DATA PREPROCESSING



Data - Record Data

49

Customer_ID	Name	Sex	Age	Income
1	Somchai	1	20	12000
2	Somying	0	18	7000
3	Somsmall	0	35	35000
4	Somjing	2	16	4000
5	Somsri	0	300	20000

แอตทริบิวต์
(Attribute)

ใช้แสดงลักษณะต่าง ๆ ของวัตถุ (คน)

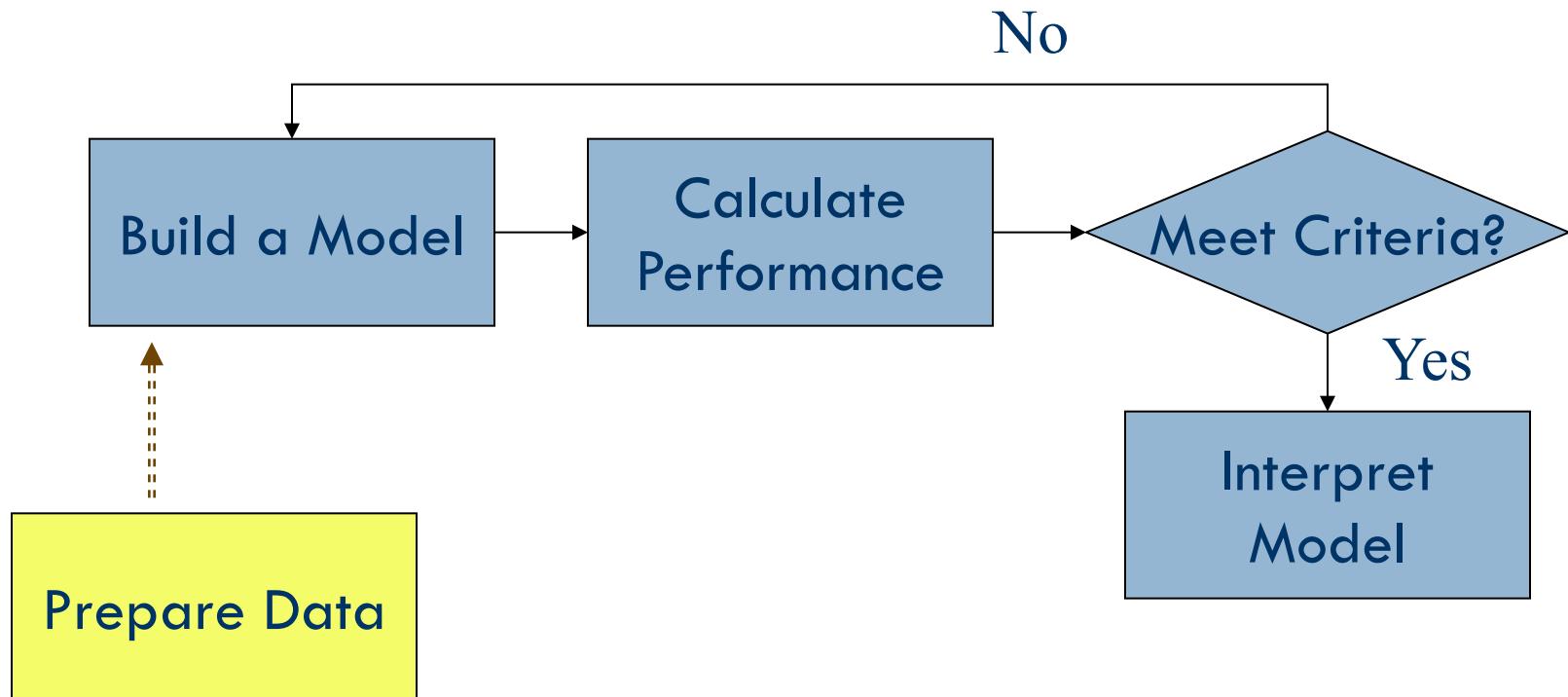
วัตถุ
(Object)

แทนแต่ละวัตถุ (คน)

- **Attribute** is also known as **variable**, **field**, or **feature**
- **Object** is also known as **record**, **case**, **sample**, or **instance**

Data Mining Process

50

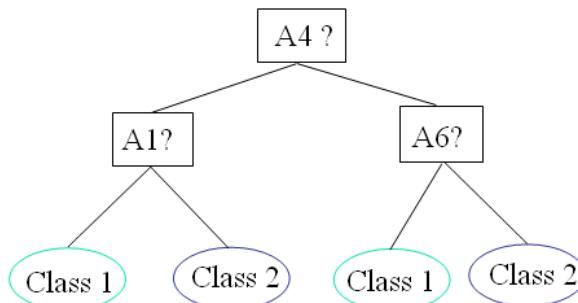


Prepare Data (Cont.)

51

- ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุด
- เนื่องจากโมเดลที่ได้จากการทำด้วยมือจะให้ผลลัพธ์ที่ถูกต้องหรือไม่นั่น
 ขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ แบ่งออกได้เป็น 3 ขั้นตอนย่อยคือ
 - คัดเลือกข้อมูล (**Data Selection**) - วิเคราะห์อะไร, เลือกเฉพาะข้อมูลที่เกี่ยวข้อง
 - กลั่นกรองข้อมูล (**Data Cleaning**) - จัดการข้อมูลซ้ำซ้อน, ขาดหาย, ผิดพลาด
 - แปลงรูปข้อมูล (**Data Transformation**) - พร้อมนำไปใช้วิเคราะห์

Initial attribute set:
 $\{A_1, A_2, A_3, A_4, A_5, A_6\}$



-----> Reduced attribute set: $\{A_1, A_4, A_6\}$

Choosing the sample size

52

- จำนวนของอินสแตนซ์ (records)
 - ▣ $\geq 5,000$
 - ▣ ถ้ามีข้อมูลที่น้อย โมเดลที่ได้จะมีความน่าเชื่อถือน้อยตามไปด้วย



8000 points



2000 Points



500 Points

Why Do We Preprocess Data?

53

- Raw data often incomplete, noisy
- May contain:
 - Obsolete fields
 - Missing values*
 - Outliers*
 - Data in form not suitable for data mining
 - Erroneous values

Handling Missing Data

54

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
	1	14.00	8	350.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	4	400.00	150.00
5	30.50	4	200.65	106.53
6	23.00	4	350.00	125.00
7	13.00	4	351.00	158.00
8	14.00	8	200.65	215.00
9	25.40	5	200.65	77.00
10	37.70	4	89.00	62.00

Replace Missing Values with Mode or Mean

- Mode of categorical field cylinders = 4
- Missing values replaced with this value
- Mean for non-missing values in numeric field cubicinches = 200.65
- Missing values replaced with 200.65

Missing Values

55

- Replace Missing Value กำหนดค่าของข้อมูลที่ขาดหายไป
 - ใช้เครื่องหมาย ? แทนข้อมูลที่หายไป หรือที่ต้องการให้มีการ replace
 - Numeric หรือ ข้อมูลตัวเลข
 - แทนค่าด้วยค่าเฉลี่ย (mean) ของค่าในแอตทริบิวต์
 - Nominal หรือ ข้อมูลประเภท
 - แทนค่าด้วยข้อมูลที่ปรากฏบ่อยที่สุด (mode) ในแอตทริบิวต์

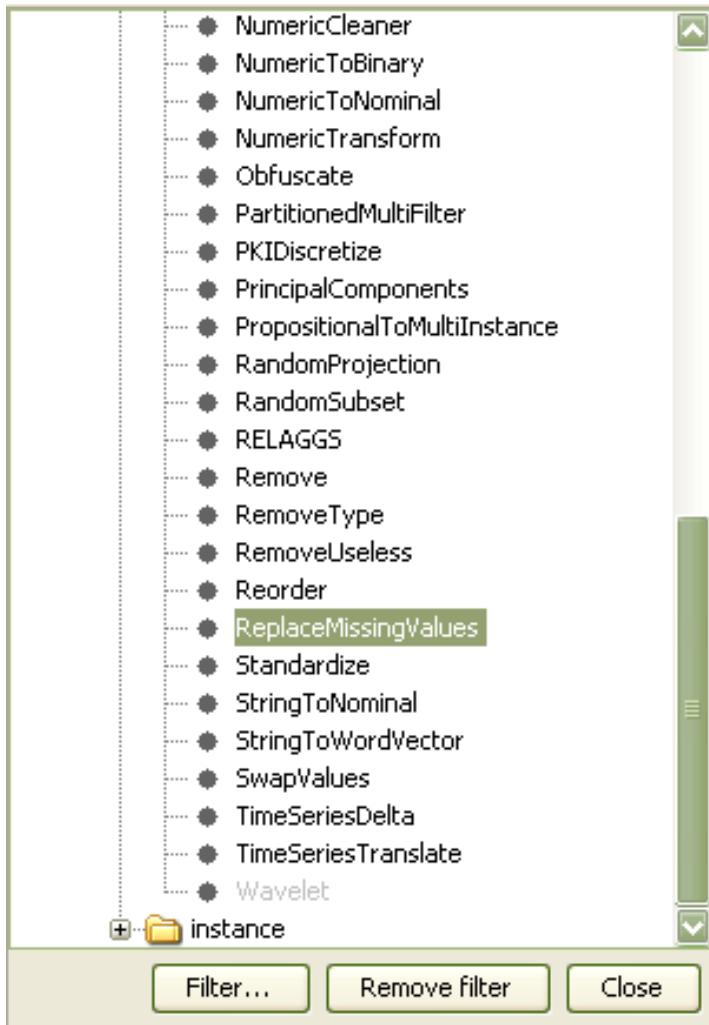
Customer_ID	Name	Sex	Age	Income
1	Somchai	ชาย	20	12000
2	Somying	หญิง	18	7000
3	Somsmall	หญิง	35	35000
4	Somjing	เด็ก	16	4000
5	Somsri	หญิง	300	20000



Customer_ID	Name	Sex	Age	Income
1	Somchai	ชาย	20	12000
2	Somying	หญิง	18	7000
3	Somsmall	หญิง	35	35000
4	Somjing	หญิง	16	4000
5	Somsri	หญิง	22.25	20000

Replace missing values in Weka

56



- กดปุ่ม Choose
 - ▢ เลือก filters
 - ▢ เลือก unsupervised
 - ▢ เลือก attribute
 - ▢ เลือก ReplaceMissingValues
- กดปุ่ม Apply

Discretization

57

- Discretization เป็นการแปลงข้อมูลลักษณะตัวเลข (numeric) ให้เป็นข้อมูลลักษณะประเภท (nominal)
- บางเทคนิคทำงานได้เฉพาะกับข้อมูลที่เป็น nominal เท่านั้น
- แบ่งกลุ่มของข้อมูลเป็นระดับ ตามเงื่อนไข
 - เด็ก อายุในช่วง 0-12 ปี, วัยรุ่น อายุในช่วง 13-20 ปี, ผู้ใหญ่ อายุมากกว่า 20 ปี
 - รายได้น้อย มีรายได้ในช่วง 0-15,000 บาท, รายได้มาก มีรายได้มากกว่า 15,000 บาท

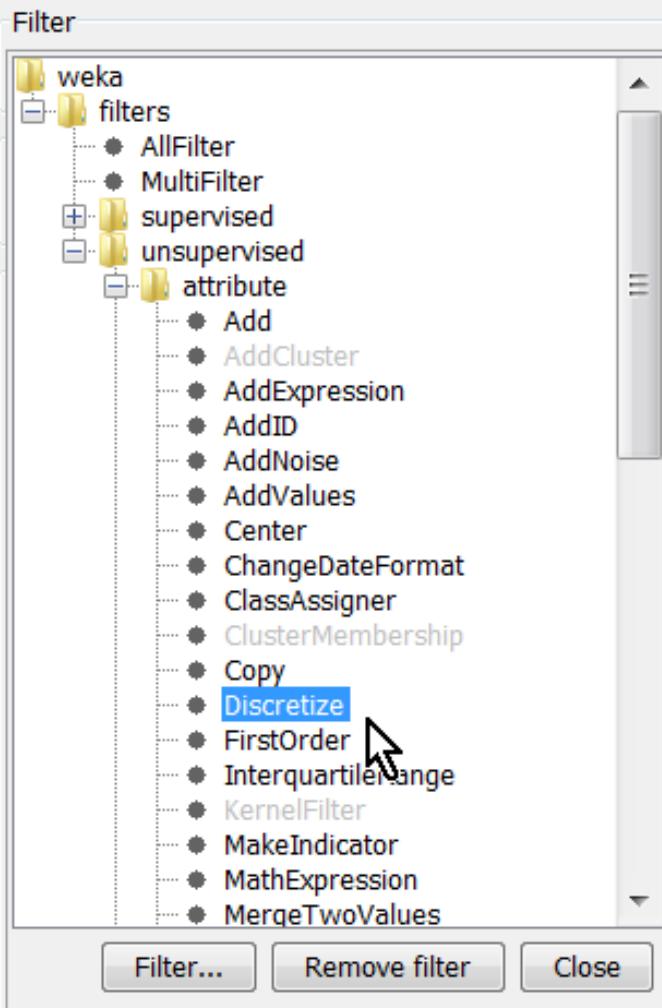
Customer_ID	Age	Income
1	20	12000
2	18	7000
3	35	35000
4	16	4000
5	-	20000



Customer_ID	Age	Income
1	วัยรุ่น	น้อย
2	วัยรุ่น	น้อย
3	ผู้ใหญ่	มาก
4	วัยรุ่น	น้อย
5	-	มาก

Discretization in Weka

58

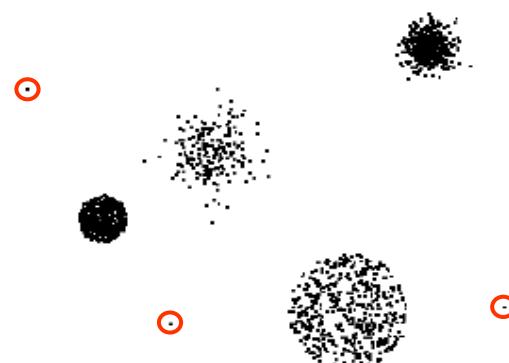


- กดปุ่ม Choose
 - ▢ เลือก filters
 - ▢ เลือก unsupervised
 - ▢ เลือก attribute
 - ▢ เลือก Discretize
- กำหนดเงื่อนไขการแบ่งกลุ่มข้อมูล เช่น
จำนวนกลุ่ม (bin) จากนั้นกดปุ่ม OK
- กดปุ่ม Apply

Graphical Methods for Identifying Outliers

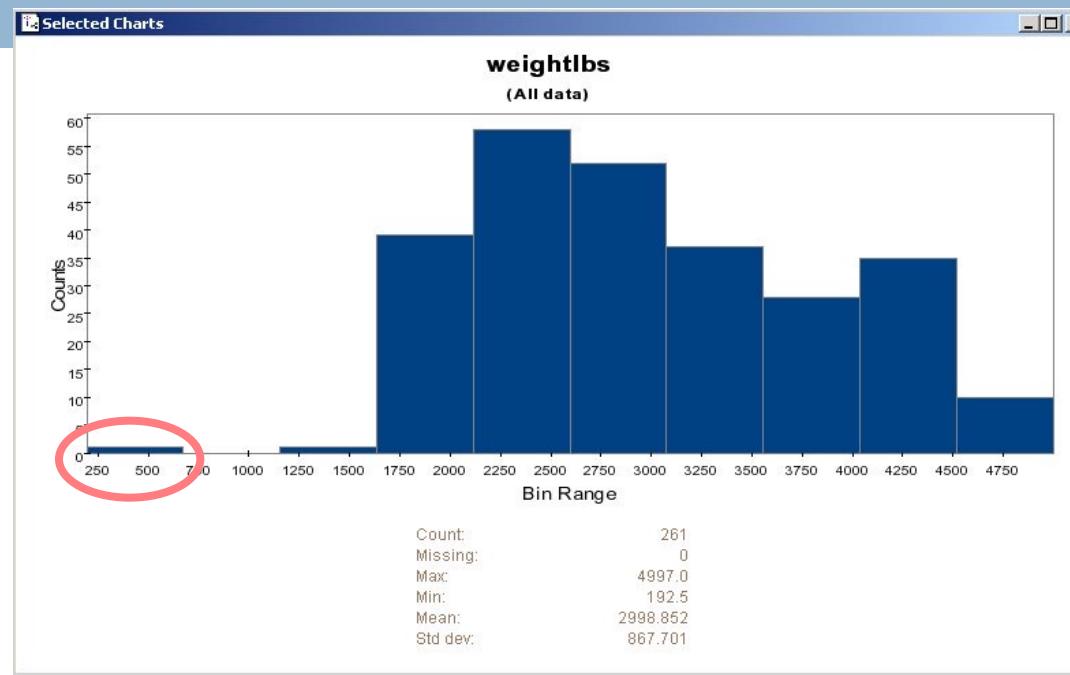
59

- Outliers are values that lie near extreme limits of data range
- Outliers may represent errors in data entry
- Certain statistical methods very sensitive to outliers and may produce unstable results
- Neural Networks and k -Means benefit from normalized data



Graphical Methods for Identifying Outliers

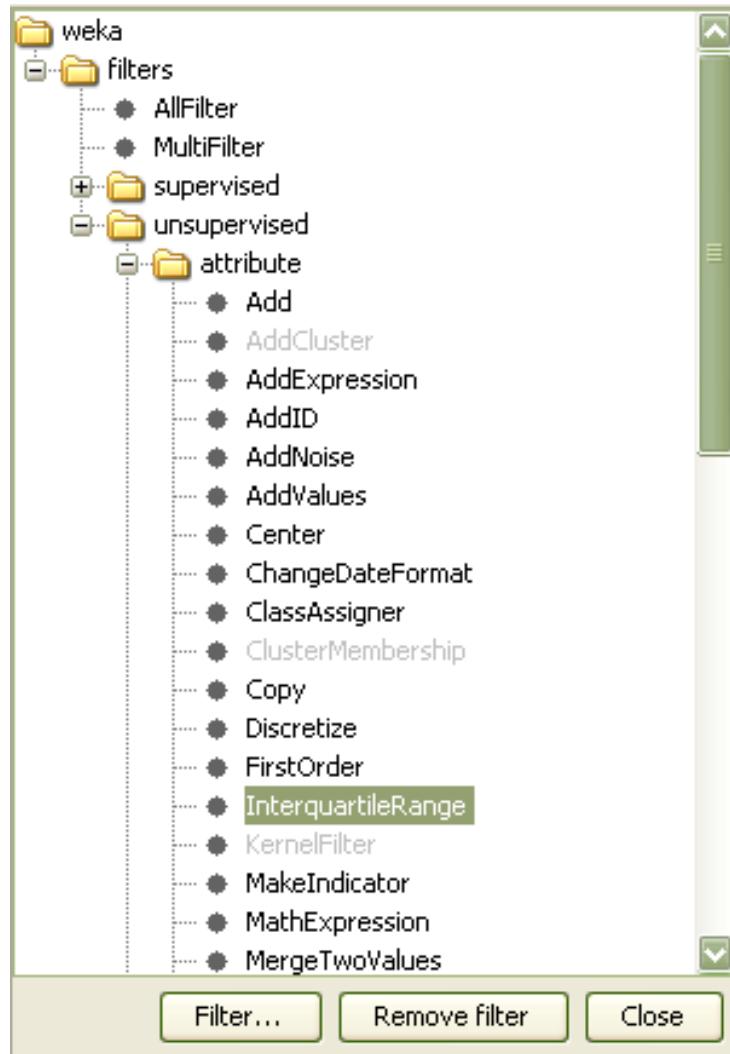
60



- This histogram shows vehicle weights for cars data set
- The extreme left-tail contains one outlier weighing several hundred pounds (192.5)
- Should we doubt validity of this value?
 - Discuss the meaning of the value with someone familiar with database content

Detect outlier in Weka

61



- กดปุ่ม Choose
 - เลือก filters
 - เลือก unsupervised
 - เลือก attribute
 - เลือก InterquartileRange
- กดปุ่ม Apply

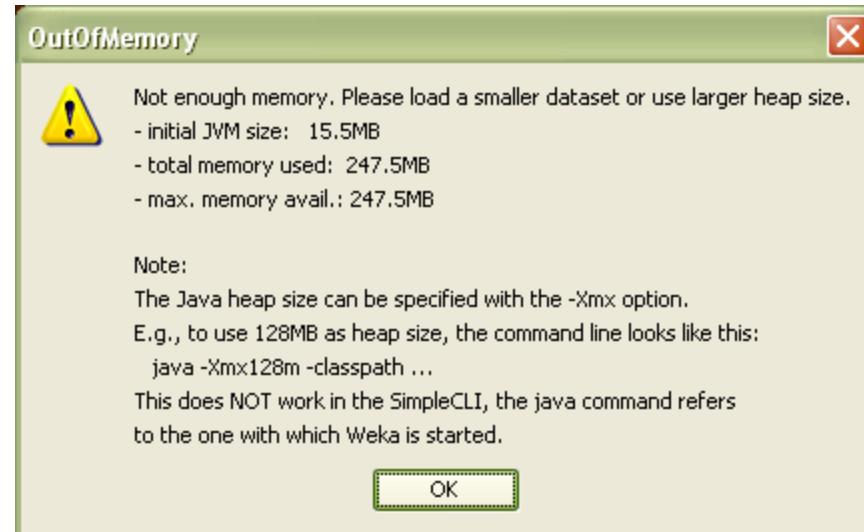
Memory Error !!

62

- วัตถุประสงค์: เพื่อแก้ไขปัญหาหน่วยจำ (memory) ไม่พอที่จะทำให้ Weka ทำงาน

- วิธีการ:

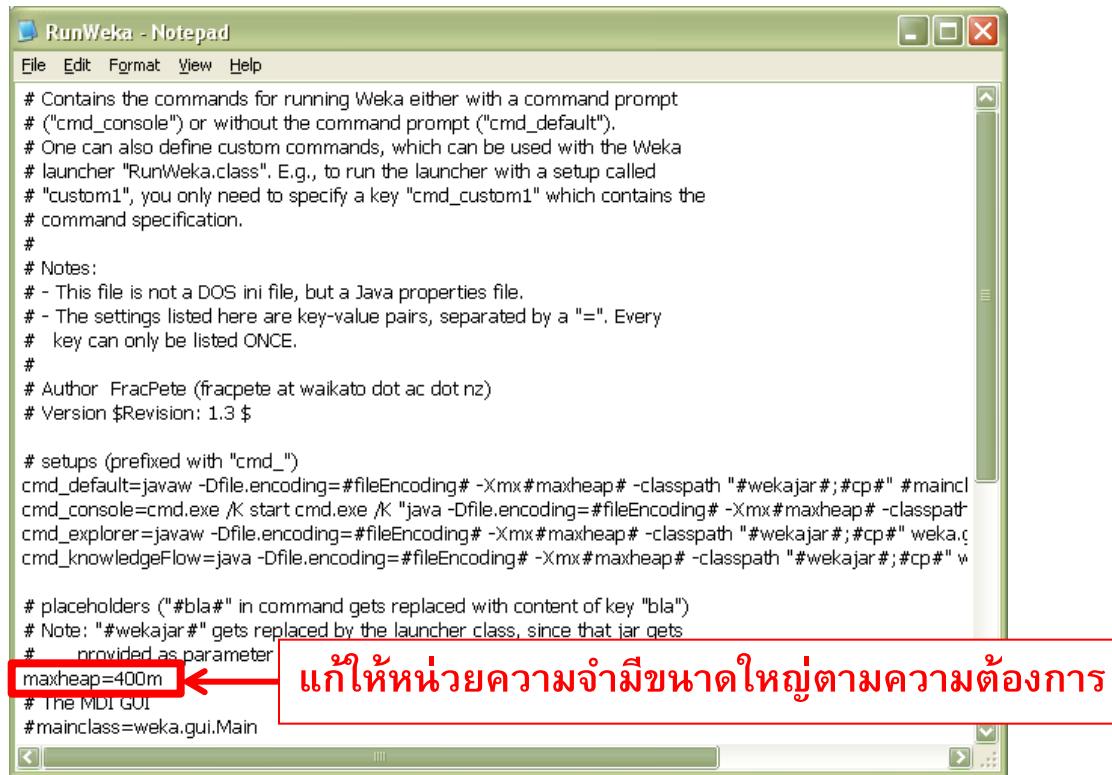
- กดปุ่ม Generate...
 - ตั้งค่า numExamples เป็น 1,000,000
 - กดปุ่ม Generate



Memory Error (Cont.)

63

- เกิด Error ขึ้น เพราะหน่วยความจำ (memory) ที่ Weka ใช้ไม่พอ
 - ต้องเพิ่มขนาดของ (memory) โดยแก้ไขในไฟล์ C:\Program Files\Weka-3-8-4\RunWeka.ini
 - เปลี่ยนค่าของ maxheap ให้มากขึ้น (แต่ไม่เกินขนาดของหน่วยความจำภายในเครื่อง)



```
# Contains the commands for running Weka either with a command prompt
# ("cmd_console") or without the command prompt ("cmd_default").
# One can also define custom commands, which can be used with the Weka
# launcher "RunWeka.class". E.g., to run the launcher with a setup called
# "custom1", you only need to specify a key "cmd_custom1" which contains the
# command specification.

#
# Notes:
# - This file is not a DOS ini file, but a Java properties file.
# - The settings listed here are key-value pairs, separated by a "=".
#   Every
#   key can only be listed ONCE.
#
# Author: FracPete (fracpete at waikato dot ac dot nz)
# Version $Revision: 1.3 $

# setups (prefixed with "cmd_")
cmd_default=javaw -Dfile.encoding=#fileEncoding# -Xmx#maxheap# -classpath "#wekajar#;#cp#" #maincl
cmd_console=cmd.exe /K start cmd.exe /K "java -Dfile.encoding=#fileEncoding# -Xmx#maxheap# -classpath "#wekajar#;#cp#" weka.c
cmd_explorer=javaw -Dfile.encoding=#fileEncoding# -Xmx#maxheap# -classpath "#wekajar#;#cp#" weka.e
cmd_knowledgeFlow=java -Dfile.encoding=#fileEncoding# -Xmx#maxheap# -classpath "#wekajar#;#cp#" weka.k

# placeholders ("#bla#" in command gets replaced with content of key "bla")
# Note: "#wekajar#" gets replaced by the launcher class, since that jar gets
#       provided as parameter
maxheap=400m
# The MDI GUI
#mainclass=weka.gui.Main
```

แก้ให้หน่วยความจำมีขนาดใหญ่ตามความต้องการ

Approximate Memory

64

- ขนาดของหน่วยความจำที่ใช้ในการเก็บข้อมูลนั้นสามารถประมาณค่าได้จาก

$$\text{approximate_memory} = \text{number of attribute} * \text{number of instances} * 8$$

โดยที่

- number of attribute คือ จำนวนแอตทริบิวต์
- number of instances คือ จำนวนอินสแตนซ์
- หมายเลข 8 คือ จำนวน byte ที่ใช้เก็บข้อมูลตัวเลข 1 ตัว

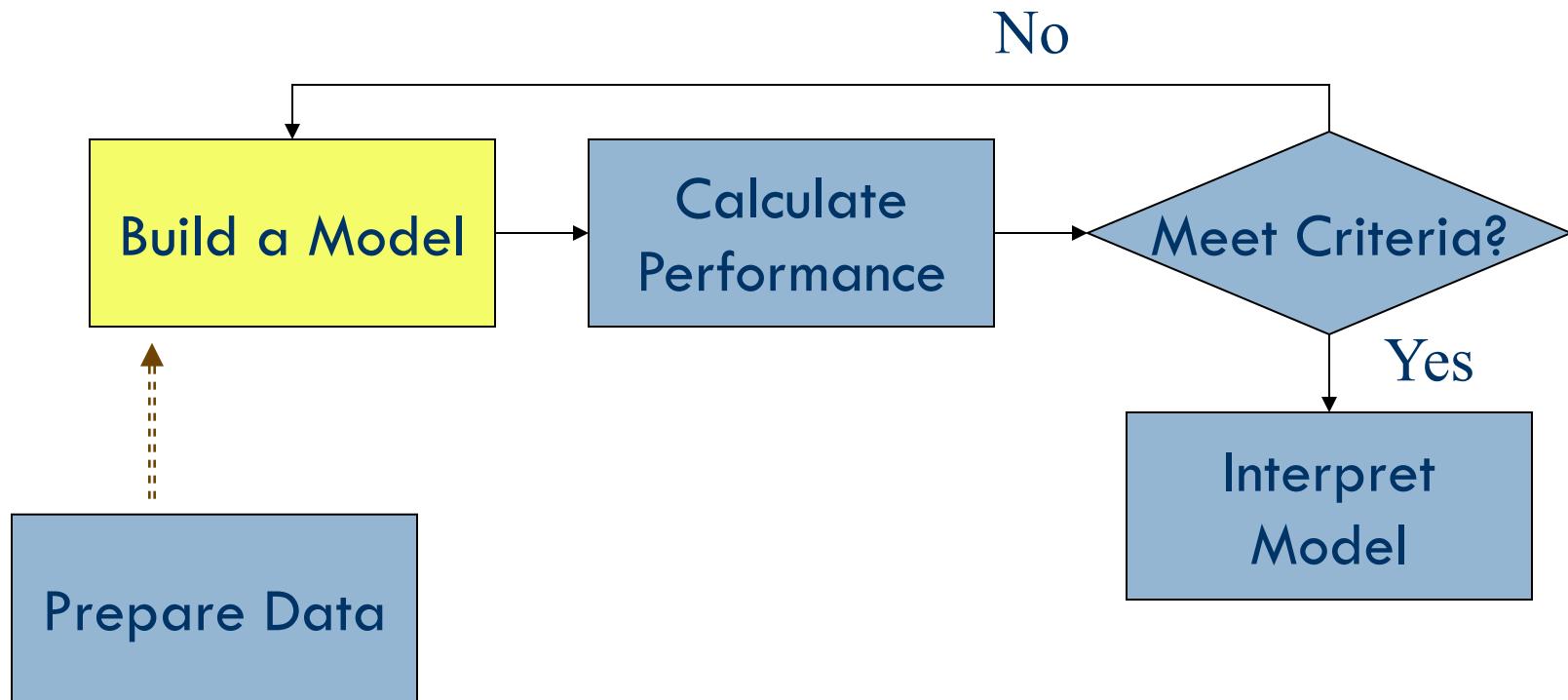
ตัวอย่าง ข้อมูล 10,000,000 instances และมี 10 attributes จะต้องใช้หน่วยความจำอย่างน้อย 800 MB

CLASSIFICATION



Data Mining Process

66



Build a Model

67

- เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคด้วยตัวมันเอง เช่น

- Classification

- Decision Tree
 - Bayesian Network
 - Neural Network
 - SVM

- Clustering

- K-Means
 - Fuzzy C Mean

- Association Rule

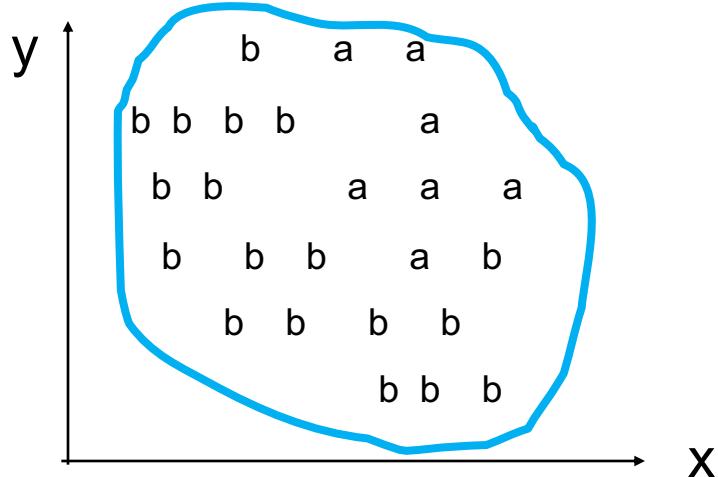
- Apriori

Classification Idea

69

Given: data in two dimensions,

Objective: to classify a class of “a”



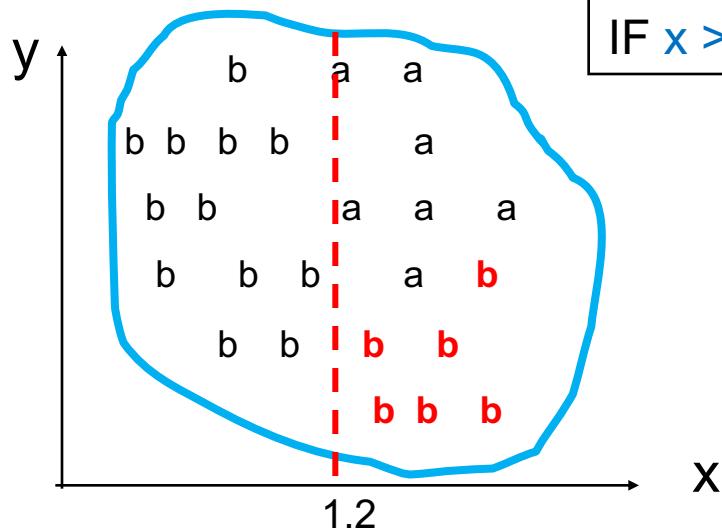
a = positive examples/instances

b = negative examples/instances

Classification Idea

70

Problem: to classify class of “a”



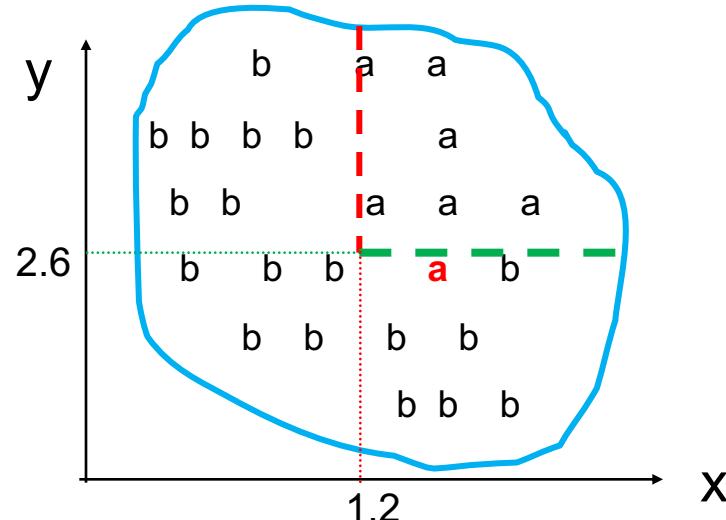
IF $x > 1.2$ THEN class = a

Coverage = 100%
Accuracy = <100%

Classification Idea

71

Problem: to classify class of “a”



Coverage = <100%

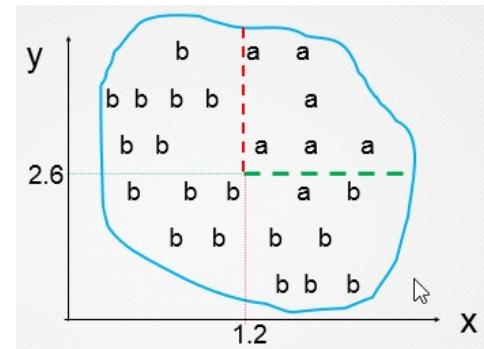
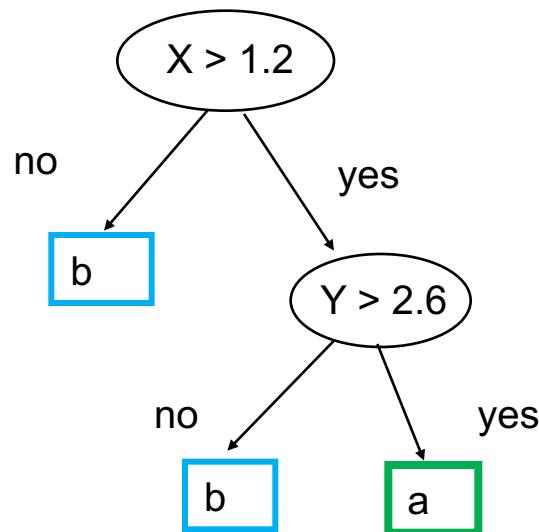
Accuracy = 100%

IF $(x > 1.2)$ AND $(y > 2.6)$ THEN class = a

Classification Idea

72

Decision tree for the same problem



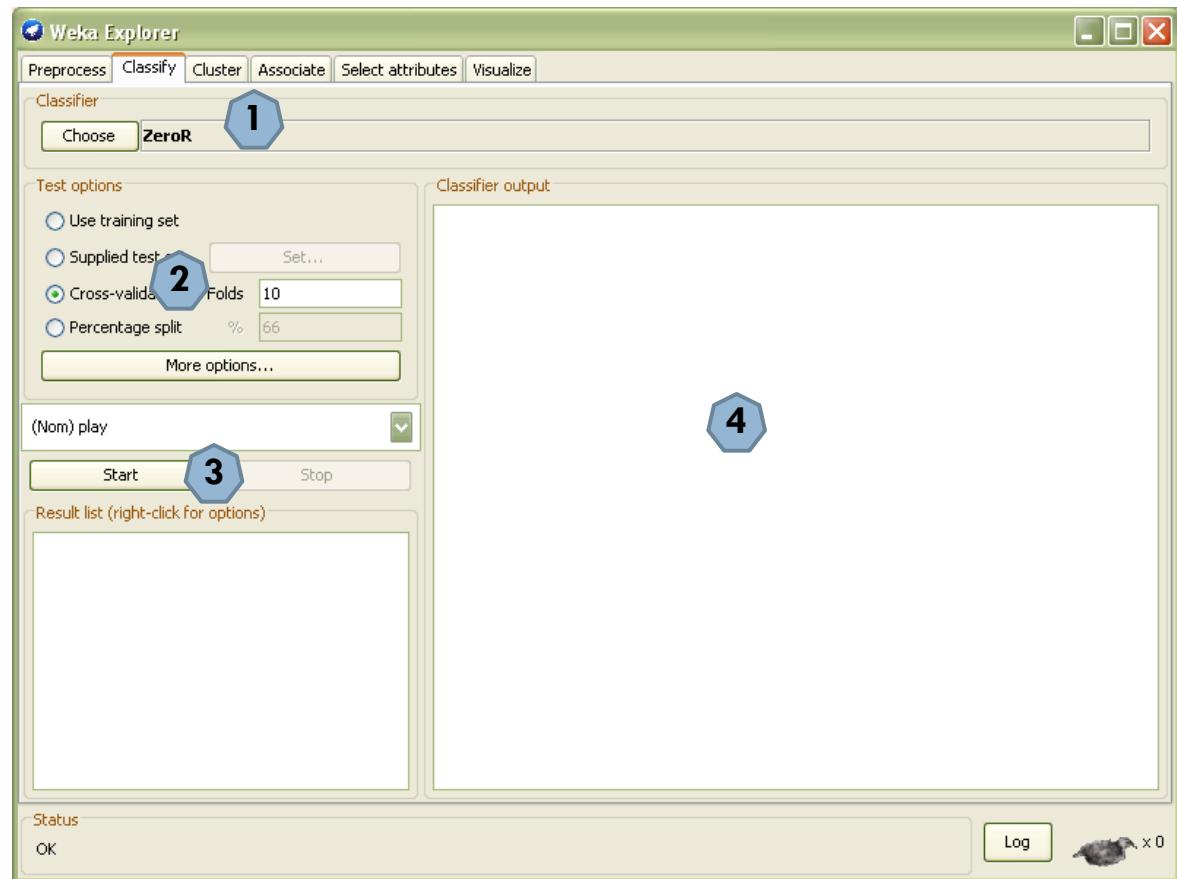
IF $(x > 1.2)$ AND $(y > 2.6)$
THEN class = a

IF $(X \leq 1.2)$ THEN class = b
IF $(X > 1.2)$ AND $(Y \leq 2.6)$ THEN class = b
IF $(X > 1.2)$ AND $(Y > 2.6)$ THEN class = a

Classification in Weka

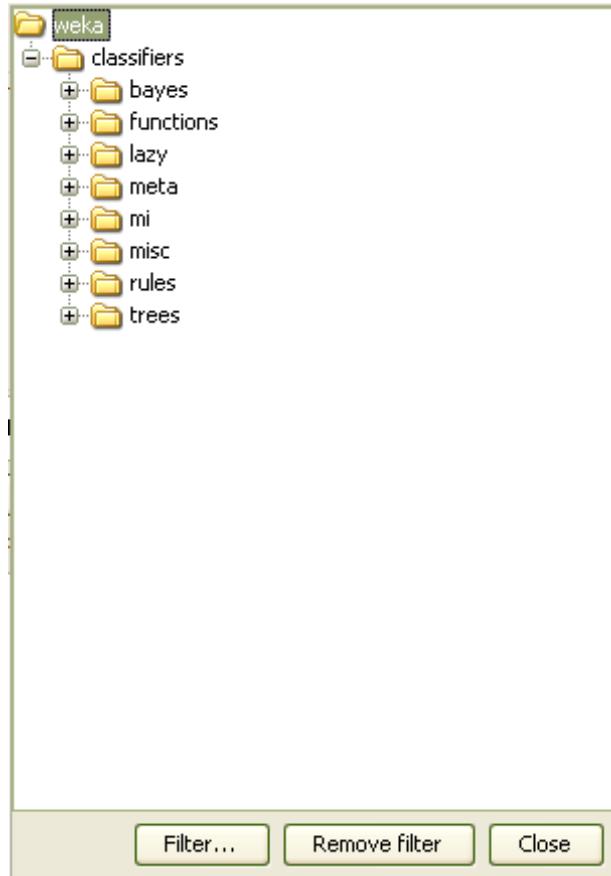
73

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file ... > เลือกไฟล์ที่ต้องการเปิด
- คลิกที่ tab Classify



1: Classifier

74



□ เทคนิคในการทำ Classification แบบต่าง ๆ

□ bayes

- สร้างโมเดลโดยอาศัยการคำนวณความน่าจะเป็น (probability) ของข้อมูลต่าง ๆ

□ functions

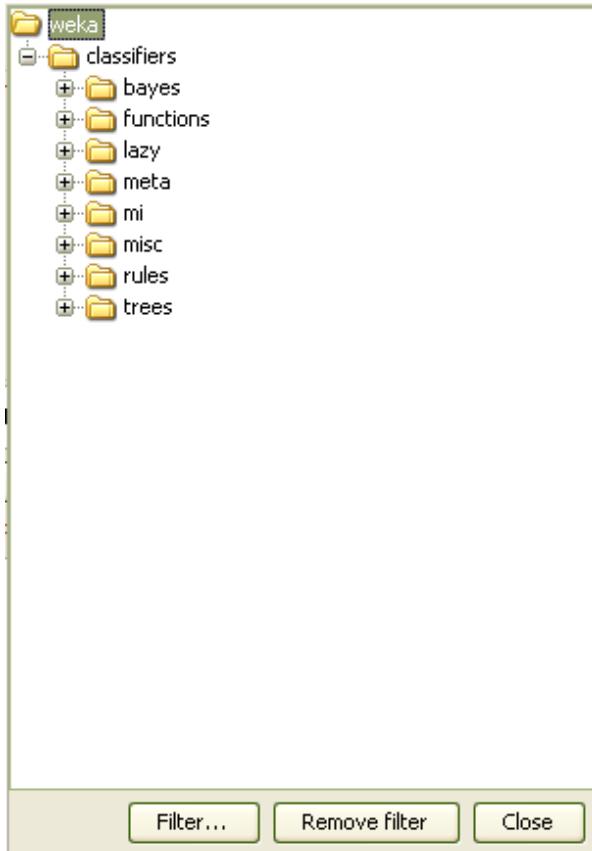
- สร้างโมเดลโดยอาศัยการคำนวณทางคณิตศาสตร์
- โมเดลเป็นรูปแบบของสมการ

□ lazy

- ต่างจากเทคนิค Classification แบบอื่น ๆ
- ไม่มีการสร้างโมเดลไว้ก่อน
- ใช้ข้อมูลเรียนรู้เพื่อจำแนกประเภทข้อมูลของข้อมูลใหม่ได้เลย

1: Classifier (Cont.)

75



□ เทคนิคในการทำ Classification แบบต่าง ๆ

□ meta

- เป็นการสร้างโมเดลโดยอาศัยเทคนิคของ Classification หลาย ๆ เทคนิค
- เพื่อเพิ่มความถูกต้องในการจำแนกประเภทข้อมูล (classification) หรือคาดเดา (prediction)

□ trees

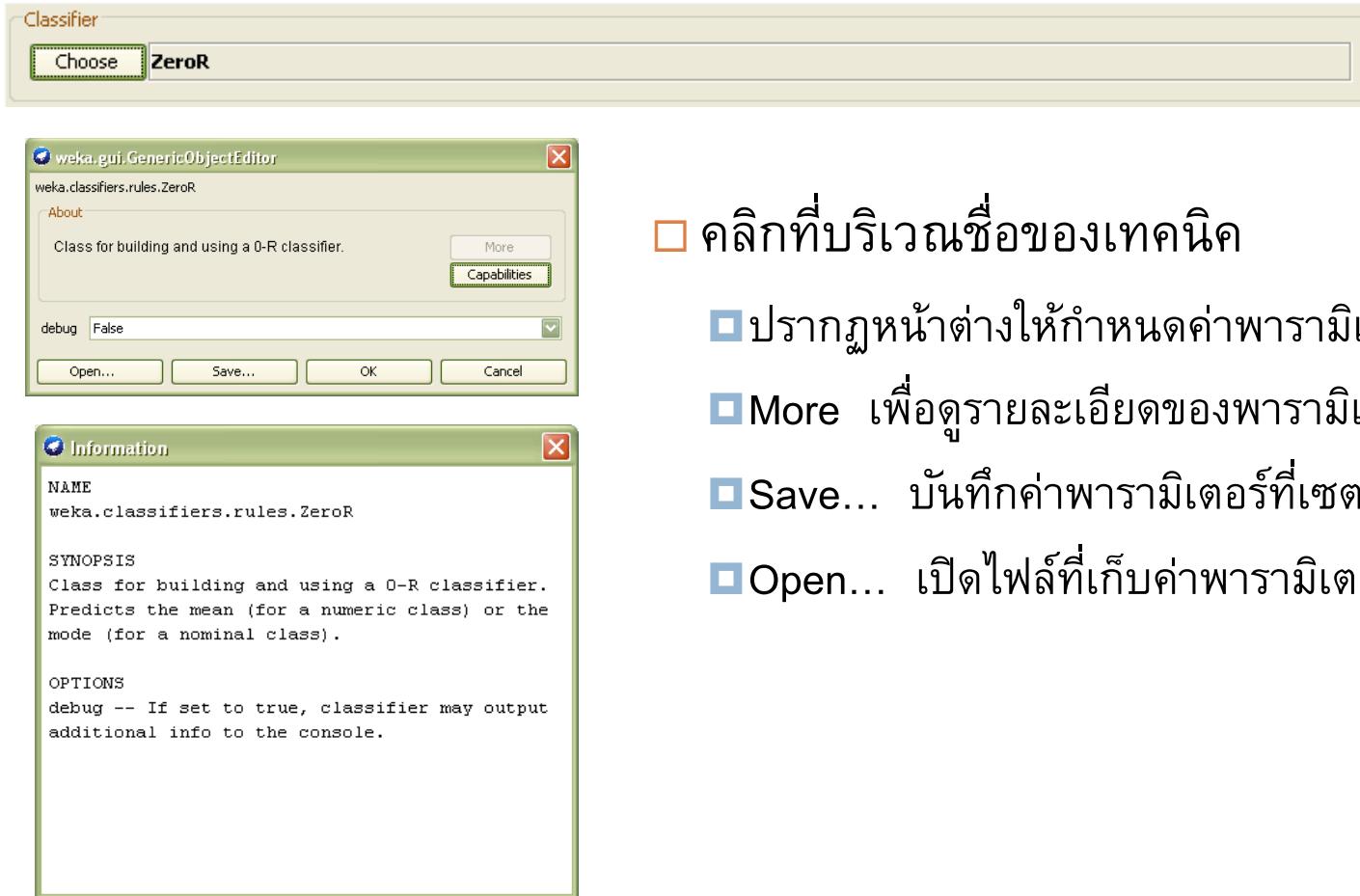
- สร้างโมเดลแบบต้นไม้ (tree) ในการจำแนกข้อมูล (classify)

□ rules

- สร้างโมเดลในรูปแบบของกฎ

1: Classifier (Cont.)

76



□ คลิกที่บริเวณชื่อของเทคนิค

- pragakृuhnातांगให้กำหนดค่าพารามि�เตอร์ต่าง ๆ
- More เพื่อดูรายละเอียดของพารามิเตอร์ต่าง ๆ
- Save... บันทึกค่าพารามิเตอร์ที่เซตไว้
- Open... เปิดไฟล์ที่เก็บค่าพารามิเตอร์ที่เซตไว้

2: Test options

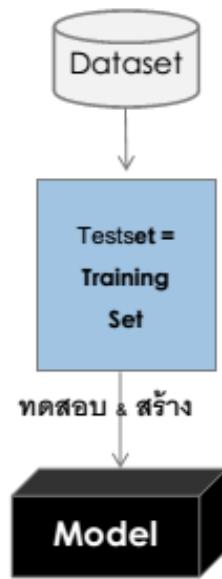
77



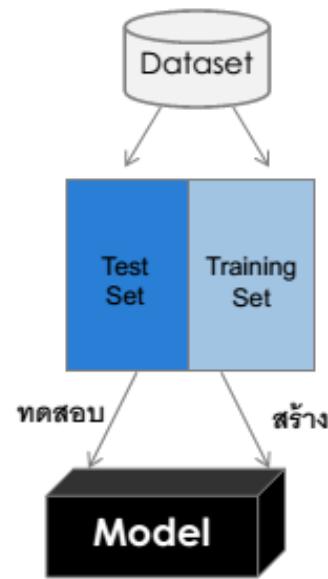
แต่ทริบิวต์ที่จะใช้เป็นคลาส
(class) ในการจำแนกประเภทข้อมูล
(ปกติจะเป็นแต่ทริบิวต์สุดท้าย)

- เลือกข้อมูลเพื่อใช้ทดสอบ
 - Use training set
 - ใช้ข้อมูลเรียนรู้ทั้งหมดเพื่อเป็นตัวทดสอบ
 - Supplied test set <-- Hold out method
 - ใช้ข้อมูลใหม่ (unseen data) เพื่อทำการทดสอบ
โดยเดลที่สร้างขึ้น
 - Cross-validation <-- Leave-one-out method
 - แบ่งข้อมูลทดสอบออกเป็น k ส่วนเท่า ๆ กัน
(folds) เพื่อใช้ในการทดสอบ
 - ระบุจำนวน k ในช่อง folds (แต่ห้ามเกินจำนวน
instance)
 - Percentage split <-- Hold out method
 - แบ่งข้อมูลเรียนรู้ออกเป็น $x\%$ เพื่อใช้ในการสร้าง
โดยเดล ส่วนที่เหลือใช้เป็นข้อมูลทดสอบ
 - กำหนดในช่อง %

Validation Techniques

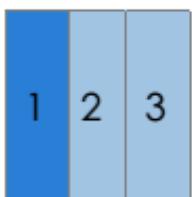


Holdout Method



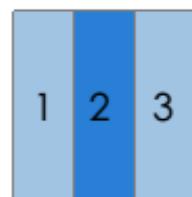
k-fold Cross Validation

3-fold cross validation



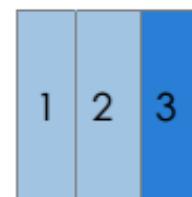
ทดสอบครั้งที่ 1

80%



ทดสอบครั้งที่ 2

75%



ทดสอบครั้งที่ 3

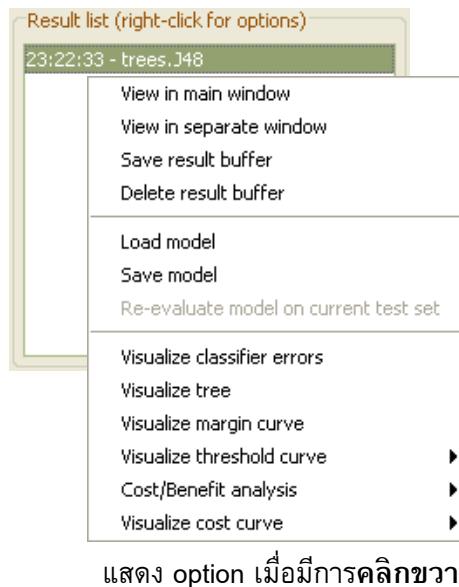
90%

$k-1$ = Training Set
 1 = Test Set
Validate = k times

เฉลี่ย 3 ครั้ง = 81.67%

3: Result list

79



□ แสดงผลการทำงานครั้งก่อนๆ

□ แสดงเวลา

□ เทคนิคที่ใช้

□ คลิกขวาที่ผลจะแสดง option

เพิ่มเติม

■ Load model: เปิดโมเดลที่ได้เคยเก็บไว้

■ Save model: บันทึกโมเดลไว้ใช้ในคราวต่อๆไป

■ Visualize tree (เมื่อพำนเทศน์ J48) ใช้แสดงโมเดล (decision tree) ในรูปแบบต้นไม้

4: Classifier output

80

The screenshot shows the 'Classifier output' window from the Weka interface. It displays the following sections:

- Run information:** Shows the scheme (weka.classifiers.trees.J48) and parameters used (C 0.25, M 2).
- Classifier model (full training set):** Displays a pruned J48 tree for the 'weather' dataset. The tree structure is:

```
outlook = sunny
| humidity <= 75:
| humidity > 75: no
outlook = overcast: yes
```
- Summary:** Provides a summary of the classification results:

	Correctly Classified Instances	64.2857 %
Incorrectly Classified Instances	5	35.7143 %
Kappa statistic	0.186	
Mean absolute error	0.2857	
Root mean squared error	0.4818	
Relative absolute error	60	%
Root relative squared error	97.6586	%
Total Number of Instances	14	

- แสดงผลการจำแนกประเภทข้อมูล (classify)

- Run information**

- แสดงรายละเอียดของข้อมูลที่ใช้
- เทคนิคและพารามิเตอร์ที่เลือก
- การทดสอบประสิทธิภาพ

- Classifier model (full training set)**

- แสดงโมเดล เช่น tree ที่สร้างได้จากข้อมูล เรียนรู้ทั้งหมด

- Summary**

- ค่าความถูกต้อง (accuracy)
 - กรณีที่คลาสเป็นข้อมูลแบบประเภท (nominal)
- ค่าความคลาดเคลื่อน (error)
 - กรณีที่คลาสเป็นข้อมูลแบบตัวเลข (numeric)

Predictor Error Measures

- Test error (generalization error): the average loss over the test set

- Mean absolute error:

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

- Mean squared error:

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

- Relative absolute error:

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

- Relative squared error:

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

Predictor Error Measures: Mean absolute error

Classifier

Choose **LinearRegression -S 0 -R 1.0E-8**

Test options

Use training set
 Supplied test set **Set...**
 Cross-validation Folds 10
 Percentage split % 66
More options...

(Num) Buy

Start Stop

Result list (right-click for options)

13:52:57 - functions.LinearRegression
13:54:20 - functions.LinearRegression
14:01:26 - functions.LinearRegression

$$\frac{\sum_{i=1}^d |y_i - y'_i|}{d}$$

Classifier output

Linear Regression Model

Buy =

41.6667 * Country=France,England +
25 * Age=Young +
53.3333

Time taken to build model: 0 seconds

==== Predictions on test set ===

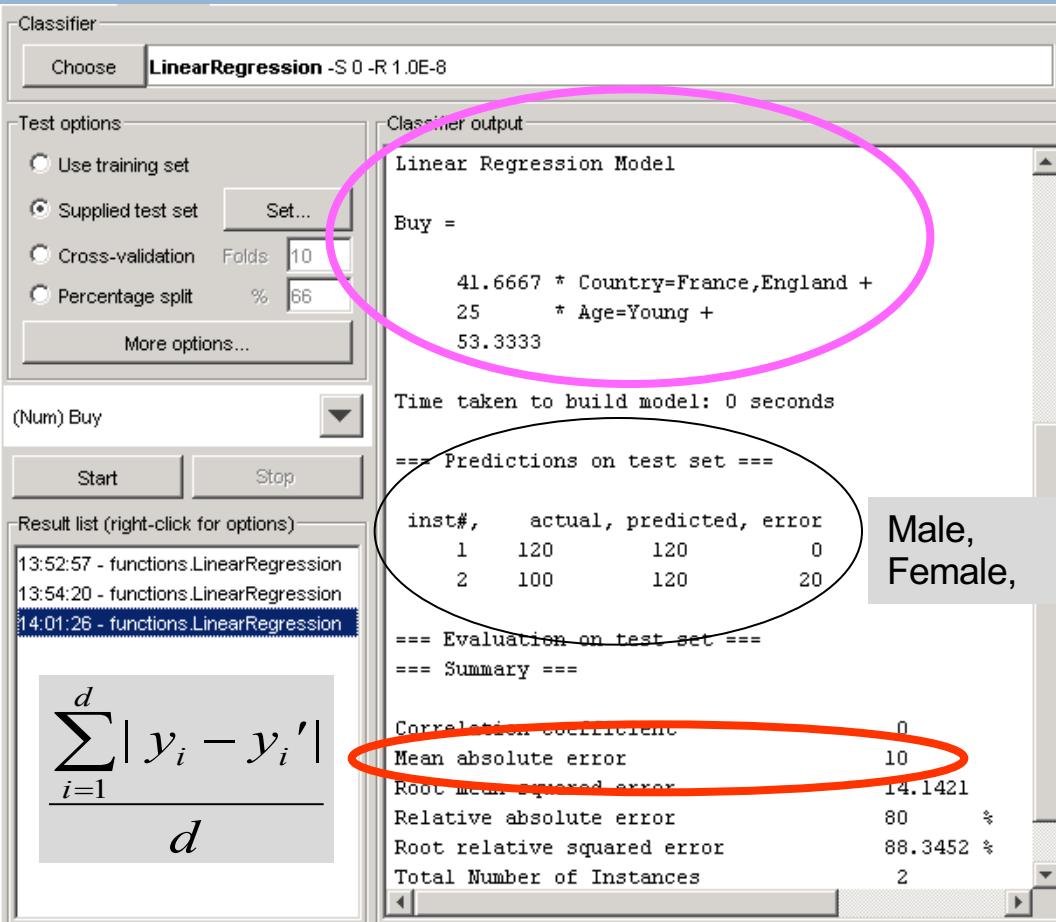
inst#	actual	predicted	error
1	120	120	0
2	100	120	20

==== Evaluation on test set ===

==== Summary ===

Correlation coefficient	0
Mean absolute error	10
Root mean squared error	14.1421
Relative absolute error	80 %
Root relative squared error	88.3452 %
Total Number of Instances	2

Male, France, Young, 120
Female, England, Young, 100



4: Classifier output (Cont.)

83

The screenshot shows the Weka interface with two windows open:

- Classifier output**:
 - Run information:
 - Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
 - Relation: weather
 - Instances: 14
 - Attributes: 5
 - outlook
 - temperature
 - humidity
 - windy
 - play
 - Test mode: 10-fold cross-validation
 - Classifier model (full training set):
 - J48 pruned tree
 -
 - outlook = sunny
| humidity <= 75: yes
| humidity > 75: no
outlook = overcast: yes
- Detailed Accuracy By Class**:
 - Stratified cross-validation Summary:

	Correctly Classified Instances	64.2857 %
TP Rate	0.778	0.6
FP Rate	0.4	0.222
Precision	0.7	0.5
Recall	0.778	0.4
F-Measure	0.737	0.444
ROC Area	0.789	0.789
 - Weighted Avg.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Weighted Avg.	0.643	0.465	0.629	0.643	0.632	0.789

□ แสดงผลการจำแนกประเภทข้อมูล
(classify)

□ Detailed Accuracy By Class

■ ค่าทางสถิติเมื่อแยกตามคลาส

■ TP Rate: ค่าที่ทายถูก

■ FP Rate: ค่าที่ทายผิด

□ Confusion Matrix

■ คอลัมน์: ค่าที่ทำนายได้

■ 行: ค่าจริง

□ ข้อมูล 2 ส่วนท้ายนี้จะไม่มีเมื่อคลาสเป็นตัวเลข !!!

4: Classifier output (Cont.)

84

```
Classifier output

==== Stratified cross-validation ====
==== Summary ====

Correctly Classified Instances          9           64.2857 %
Incorrectly Classified Instances        5           35.7143 %
Kappa statistic                         0.186
Mean absolute error                     0.2857
Root mean squared error                 0.4818
Relative absolute error                  60         %
Root relative squared error             97.6586 %
Total Number of Instances                14

==== Detailed Accuracy By Class ====

      TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area
      0.778     0.6       0.7       0.778     0.737     0.789
      0.4       0.222    0.5       0.4       0.444     0.789
Weighted Avg.    0.643     0.465    0.629     0.643     0.632     0.789

==== Confusion Matrix ====
a b  <-- classified as
7 2 | a = yes
3 2 | b = no


```

		PREDICTED CLASS	
		Class=Yes	Class>No
ACTUAL CLASS	Class=Yes	True positive (TP)	False negative (FN)
	Class>No	False positive (FP)	True negative (TN)

$$TP \text{ rate} = \frac{TP}{TP + FN}$$

$$FP \text{ rate} = \frac{FP}{FP + TN}$$

$$Precision (p) = \frac{TP}{TP + FP}$$

$$Recall (r) = \frac{TP}{TP + FN}$$

$$F - measure (F) = \frac{2rp}{r + p} = \frac{2TP}{2TP + FP + FN}$$

Example 1: Decision Tree

85

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file ... > เลือกไฟล์ data\weather.numeric.arff
- คลิกที่ tab classify
- กดปุ่ม Choose
 - เลือก classifier
 - เลือก tree
 - เลือก J48
- กดปุ่ม Start
- สังเกตผลลัพธ์
- คลิกขวาที่โมเดลในช่อง Result list
- เลือก Visualize tree และ Save model

Example 1: Decision Tree (Cont.)

86

□ Testing file

- คลิกขวาที่โมเดลในช่อง Result list
- เลือก Load model
- เลือก Supplied test set
- กดปุ่ม Set... เลือกไฟล์ weather_test.arff
- กดปุ่ม More options...
 - เลือก Output predictions
- คลิกขวาที่โมเดล เลือก Re-evaluate model on current test set
- สังเกตผลลัพธ์

Example 2: K-Nearest Neighbors

87

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file ... > เลือกไฟล์ data\weather.numeric.arff
- คลิกที่ tab classify
- กดปุ่ม Choose
 - เลือก classifier
 - เลือก lazy
 - เลือก ibk
- กดปุ่ม Start

Example 3: Neural Networks

88

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file ... > เลือกไฟล์ data\weather.numeric.arff
- คลิกที่ tab classify
- กดปุ่ม Choose
 - เลือก classifier
 - เลือก functions
 - เลือก MultilayerPerceptron
- กดปุ่ม Start

Goal Identification

149

การจัดโปรโมชั่นส่งเสริมการขายได้ตรงกับกลุ่มเป้าหมาย

การจัดวางของบนชั้นขายของใกล้กัน เพื่อลดเวลาในการซื้อสินค้า

การทำนายยอดขายในปีถัดไป เพื่อใช้วางแผนการผลิต

การแบ่งกลุ่มลูกค้าที่น่าจะซื้อ ช่วยลดค่าโฆษณา

การแบ่งกลุ่มเอกสาร ช่วยลดเวลาในการดำเนินการ

การตรวจสอบการจ้อโงบบัตรเครดิต