



Chapter 4

Pre - processing





Data

- ข้อมูลแสดงในรูปแบบของตาราง
 - แถว เรียกว่า **ตัวอย่าง** (example)
 - คอลัมน์ เรียกว่า **แอตทริบิวต์** (attribute) ซึ่งมีหน้าที่ (role) ที่ใช้งานบ่อย 3 แบบ
 - **ไอดี (ID)** เป็นแอตทริบิวต์ที่แสดงหมายเลขของข้อมูล หรือ primary key ในฐานข้อมูล
 - **แอตทริบิวต์ทั่วไป (attribute)** เป็นแอตทริบิวต์ปกติที่จะใช้ในการสร้างโมเดลหรือเรียกว่าเป็น ฟีเจอร์ (feature) หรือตัวแปรต้น (independent variable)
 - **ลาเบล (label)** เป็นแอตทริบิวต์ชนิดพิเศษที่มักจะใช้แสดงคำตอบของสิ่งที่เราต้องการจะสร้างโมเดลมาทำนาย หรือ เรียกว่า คลาส (class) หรือตัวแปรตาม (dependent

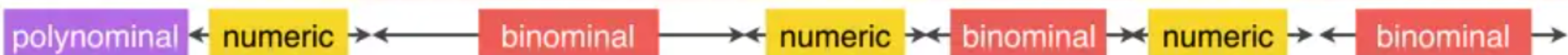
customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO



Value Type

- ค่าของข้อมูลที่เก็บในแต่ละแอตทริบิวต์
 - Polynominal ข้อมูลประเภท category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่ามากกว่า 2 ค่าขึ้นไป
 - Binominal ข้อมูลประเภท category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่าเพียง 2 ค่าเท่านั้น
 - Numeric หรือ Integer ข้อมูลประเภทตัวเลข
 - Text ข้อมูลประเภทข้อความ

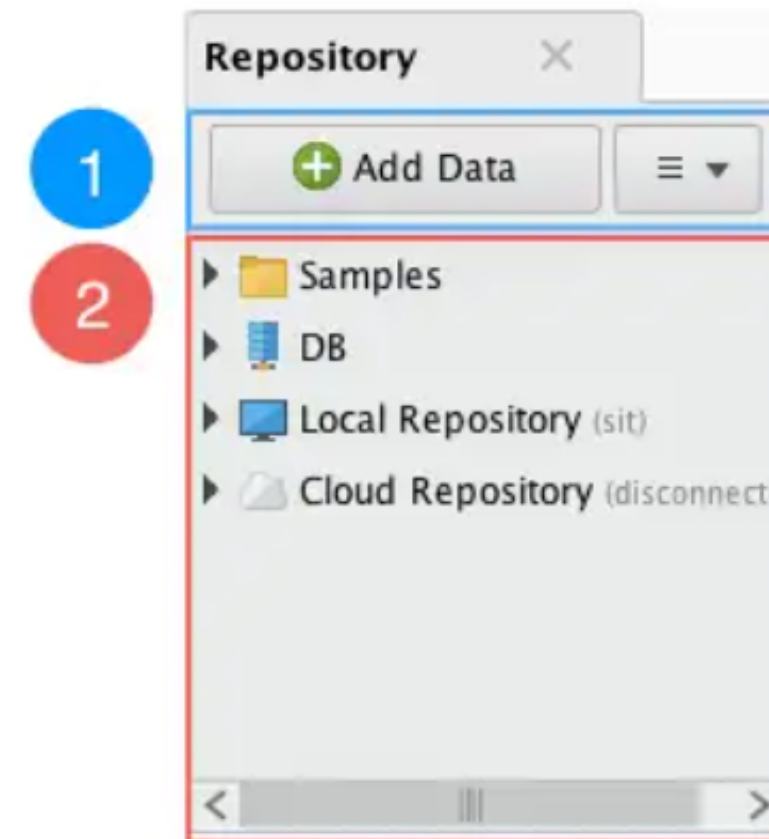
customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO





Data Management

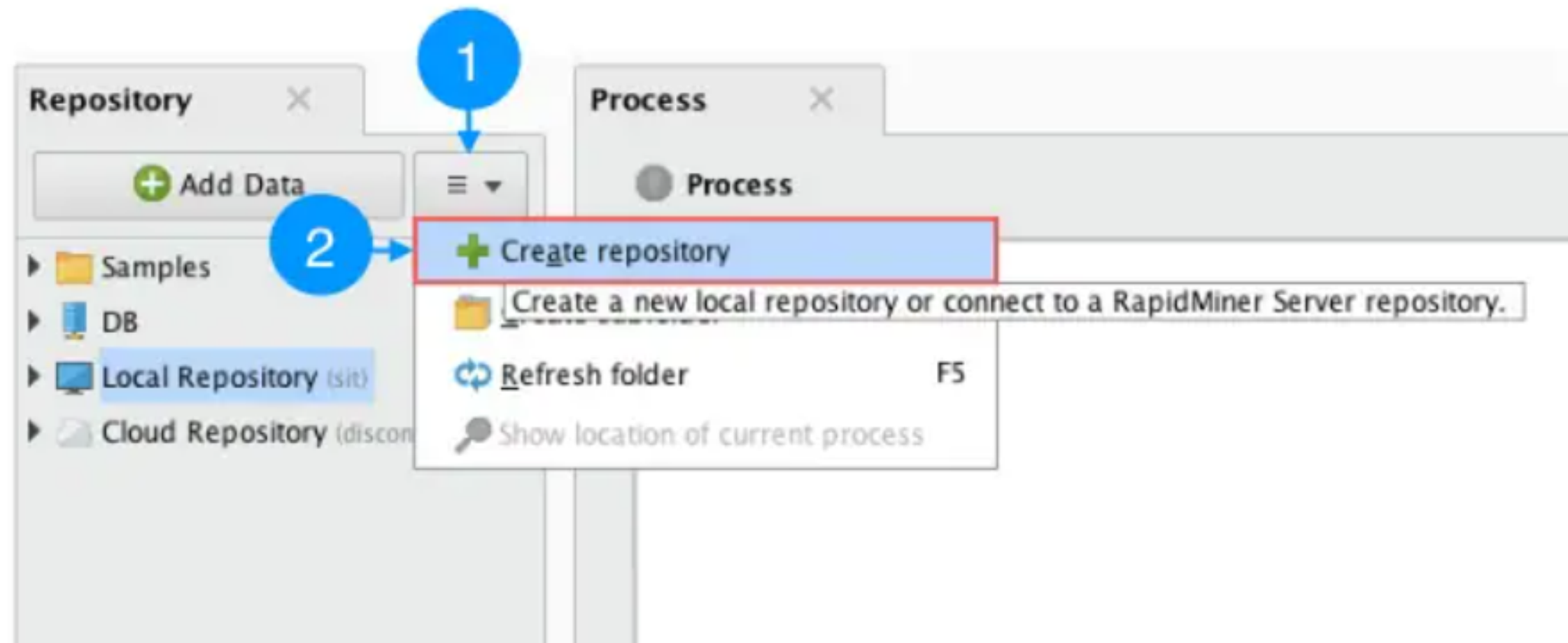
- Repository
 - เป็นที่เก็บข้อมูลและ process เพื่อใช้งานใน RapidMiner Studio 7
 - ทำให้ไม่ต้องโหลดข้อมูลจากไฟล์ใหม่ทุกครั้ง
- องค์ประกอบในส่วน Repository
 - ส่วนที่ 1
 - สำหรับสร้าง Repository ใหม่
 - โหลดไฟล์ประเภทต่างๆ เข้าไปไว้ใน Repository
 - สร้างโฟลเดอร์ใหม่
 - ส่วนที่ 2
 - ข้อมูลและ process Sample ที่ RapidMiner Studio 7 เตรียมไว้ให้
 - ข้อมูลที่เก็บอยู่ในแต่ละ Repository





Data Management

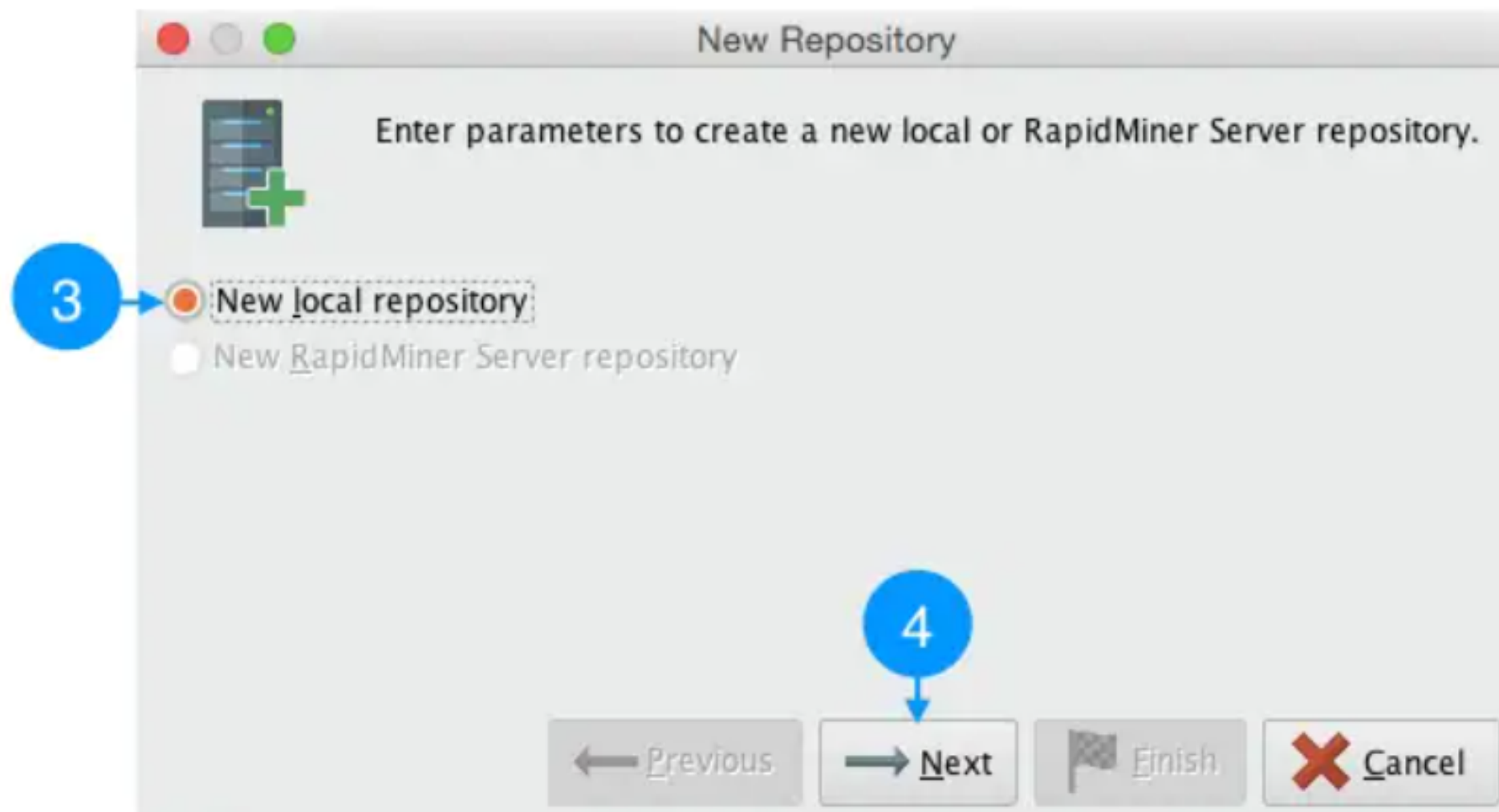
- สร้าง Repository ใหม่
 - คลิกที่ 
 - หลังจากนั้นเลือกเมนู Create repository





Data Management

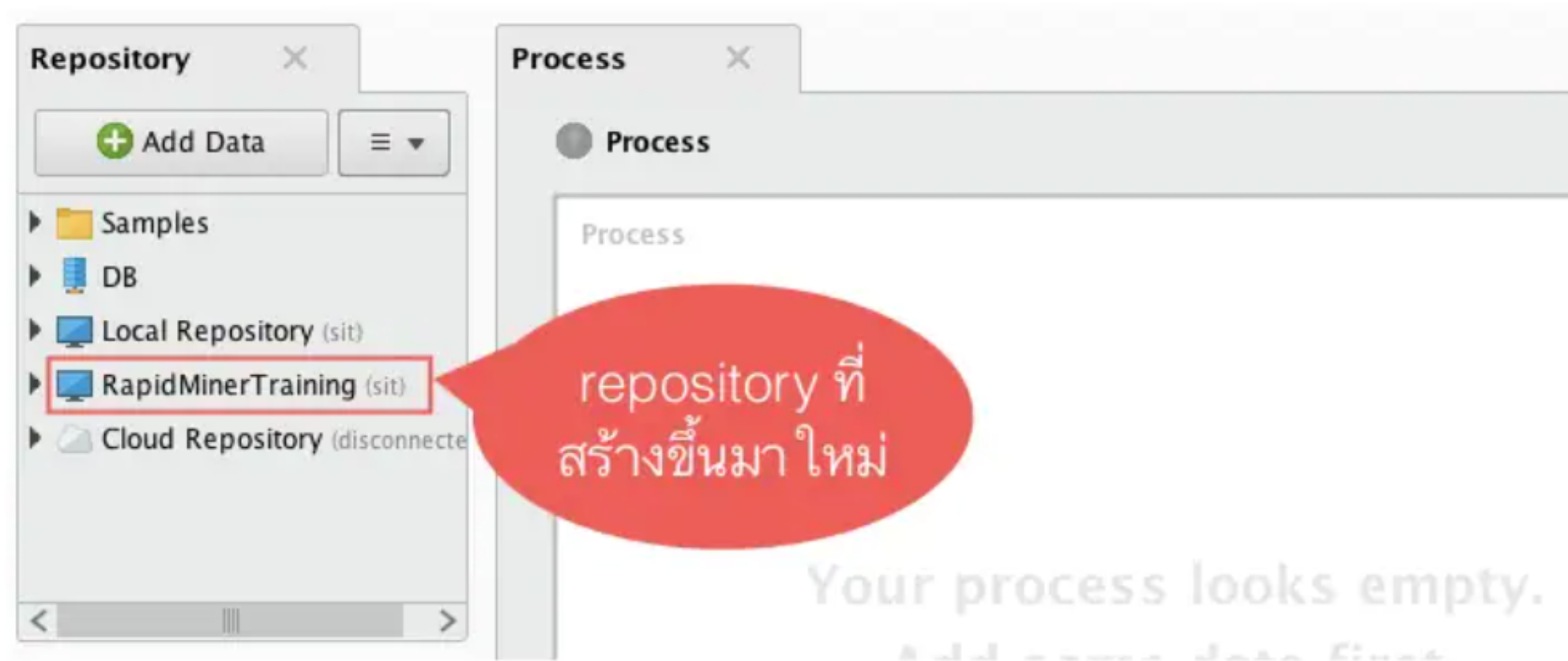
- สร้าง Repository ใหม่ (ต่อ)
 - เลือก New local repository
 - กดปุ่ม Next





Data Management

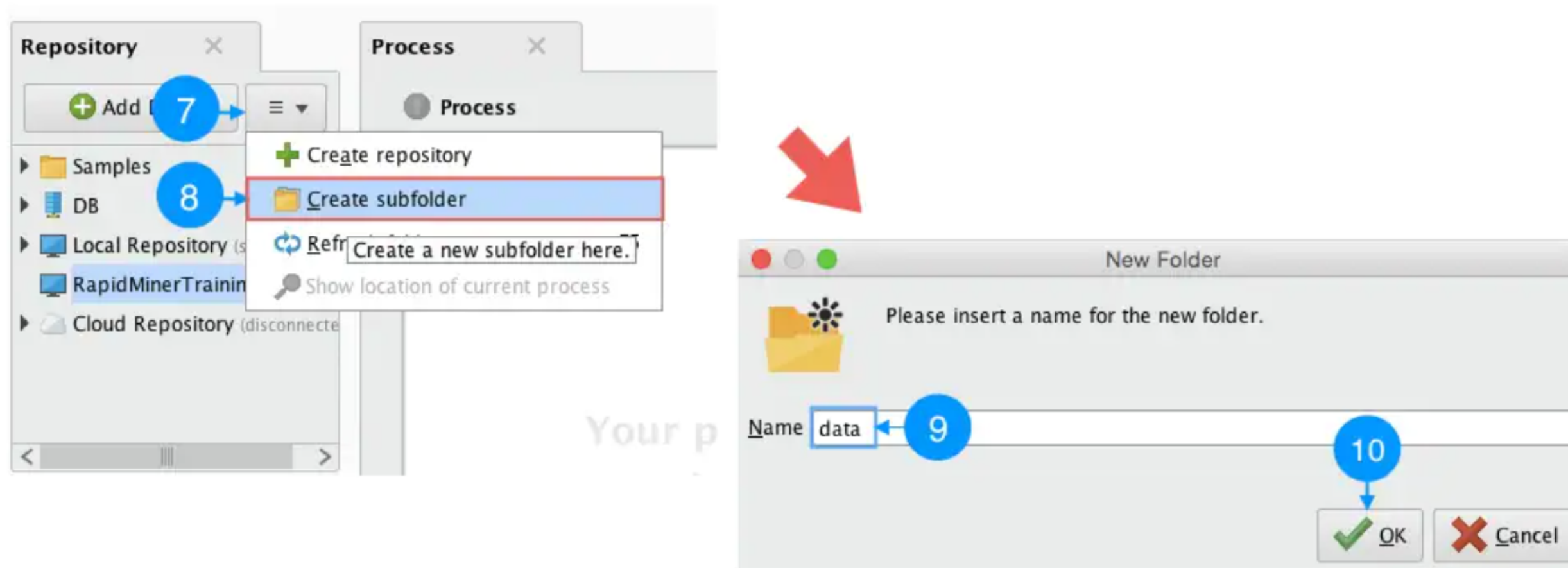
- จะปรากฏ repository RapidMinerTraining แสดงขึ้นมาในส่วนของ Repository





Data Management

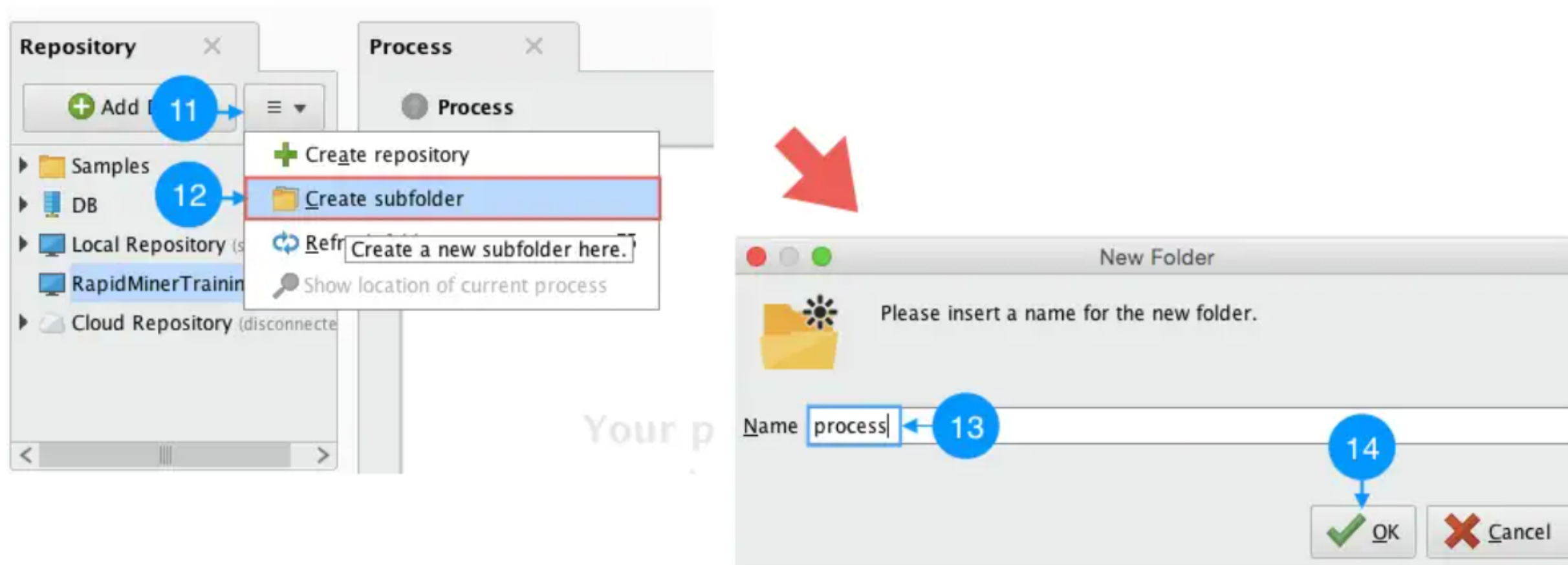
- สร้างโฟลเดอร์ใน RapidMinerTraining
 - data สำหรับเก็บข้อมูล





Data Management

- สร้างโฟลเดอร์ใน RapidMinerTraining
 - data สำหรับเก็บข้อมูล
 - process สำหรับเก็บโปรเซสที่สร้างขึ้น





CSV File

- ไฟล์ประเภท csv ย่อมาจาก Comma Separated Value
- ใช้เครื่องหมาย , (comma) คั่นระหว่างแอตทริบิวต์

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO



customer_id,age,gender,region,income,married,children,ca,response
ID12101,48,FEMALE,INNER_CITY,17546.0,NO,1,NO,NO
ID12102,40,MALE,TOWN,30085.1,YES,3,YES,NO
ID12103,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES
ID12104,23,FEMALE,TOWN,20375.4,YES,3,NO,NO

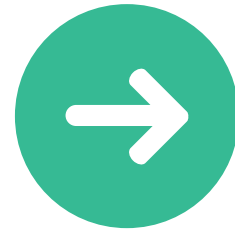
แถวแรกคือ header



สร้างไฟล์ CSV

Untitled - Notepad

File Edit Format View Help



CSV 1 - Notepad

File Edit Format View Help

```
customer_id,age,gender,region,income,married,children,car,response  
ID1,48,F,In-City,17000,No,1,No,No  
ID2,40,M,Town,30000,Yes,3,Yes,No  
ID3,51,F,In-City,16000,Yes,0,Yes,Yes  
ID4,23,F,Town,20000,Yes,3,No,NO
```



CSV 1 - Notepad

File Edit Format View Help

New	Ctrl+N	ion,income,mar
New Window	Ctrl+Shift+N	1,No,No
Open...	Ctrl+O	Yes,No
Save	Ctrl+S	,0,Yes,Yes
Save As...	Ctrl+Shift+S	No,NO
Page Setup...		
Print...	Ctrl+P	
Exit		

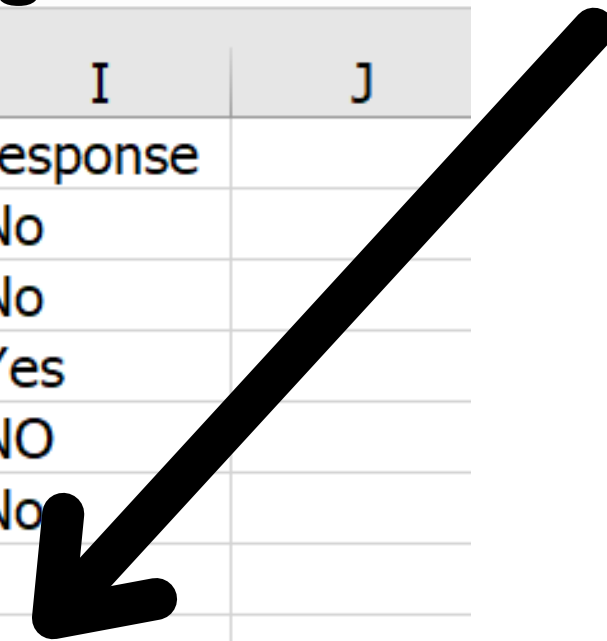


สร้างไฟล์ CSV

The screenshot shows the Microsoft Excel interface with a warning message: "POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format." The data table is as follows:

	A	B	C	D	E	F	G	H	I	J
1	customer_id	age	gender	region	income	married	children	car	response	
2	ID1	48	F	In-City	17000	No	1	No	No	
3	ID2	40	M	Town	30000	Yes	3	Yes	No	
4	ID3	51	F	In-City	16000	Yes	0	Yes	Yes	
5	ID4	23	F	Town	20000	Yes	3	No	NO	
6	ID5	20	M	Town	30000	No	0	Yes	No	
7										
8										
9										

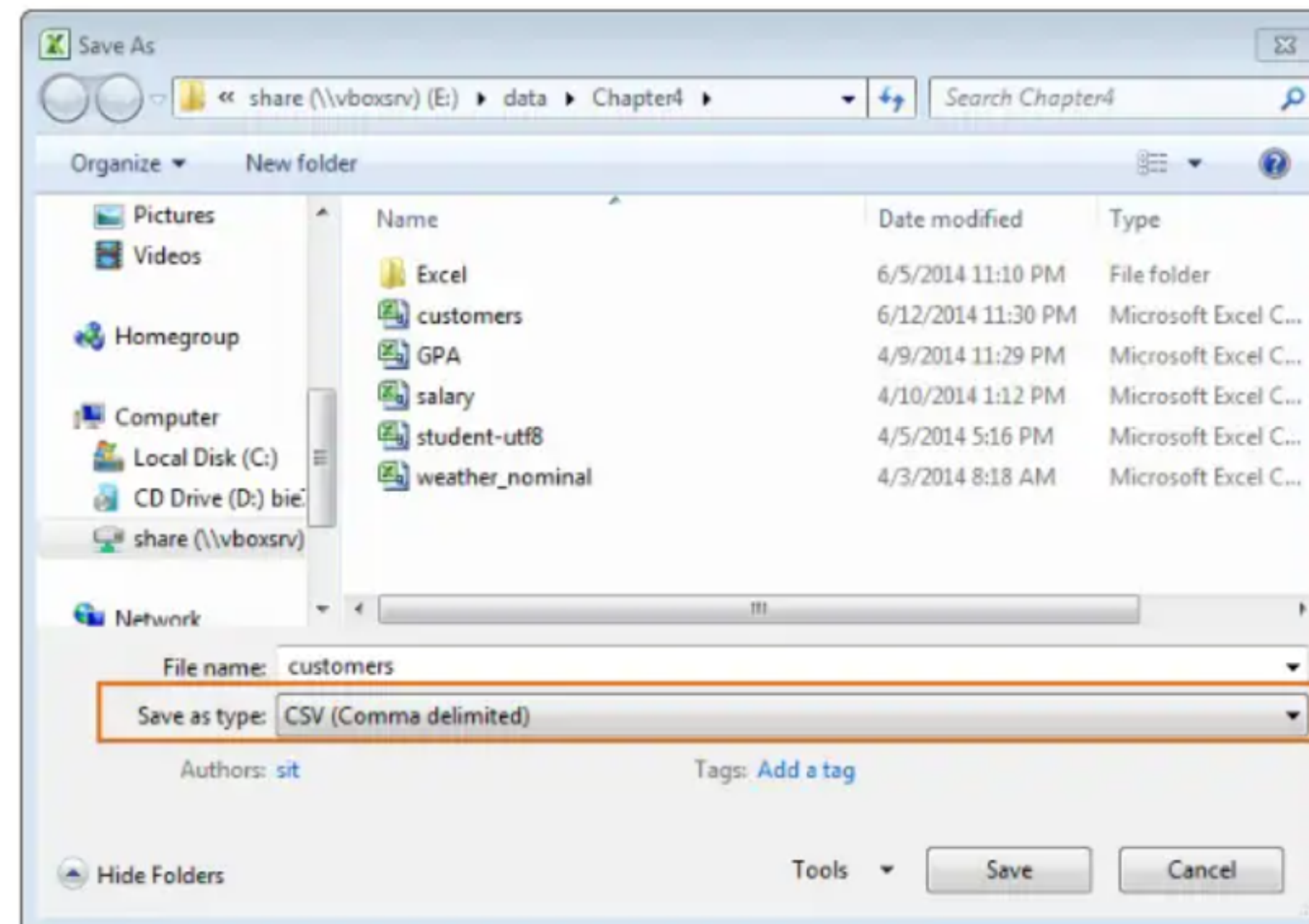
افظเตตข้อมูลได้เหมือน Excel





CSV File

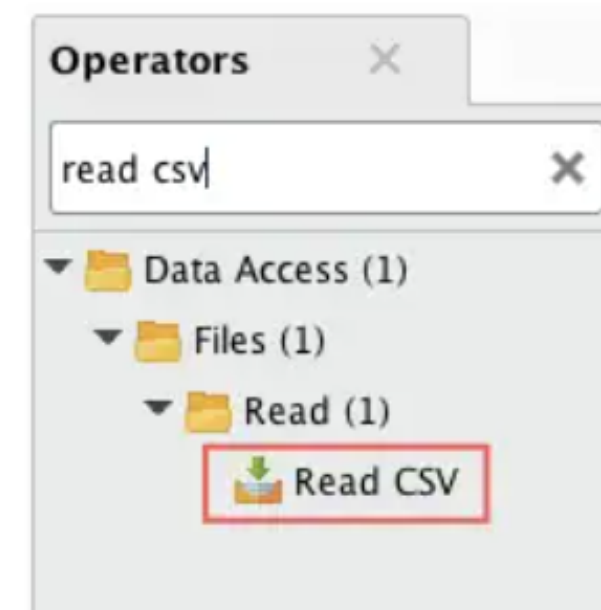
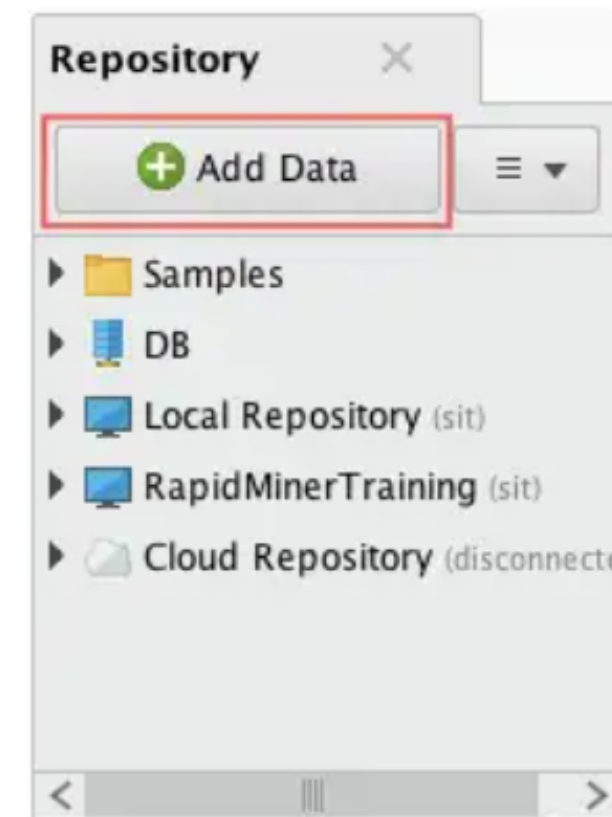
- ไฟล์ csv สามารถ export ได้จาก Excel หรือ database ต่างๆ
- Export จาก Excel
 - เลือก File > Save As > CSV (Comma delimited)





Load CSV File to RapidMiner

- การโหลดไฟล์ csv เข้าไปใช้ใน RapidMiner Studio 7 ทำได้ 2 แบบ
 - ใช้การ import ในส่วนของ Repositories
 - โหลดมาเก็บไว้ใน Repository และใช้งาน
ได้ตลอด
 - ถ้าข้อมูลในไฟล์ csv มีการเปลี่ยนแปลงจะ
ไม่ update ต้องทำการโหลดใหม่
 - ใช้โอเปอเรเตอร์ Read CSV
 - โหลดเข้ามาใช้งานโดยการอ่านจากไฟล์ csv
ทุกครั้ง เมื่อไฟล์ update ข้อมูลจะเปลี่ยนตาม





Data preparation

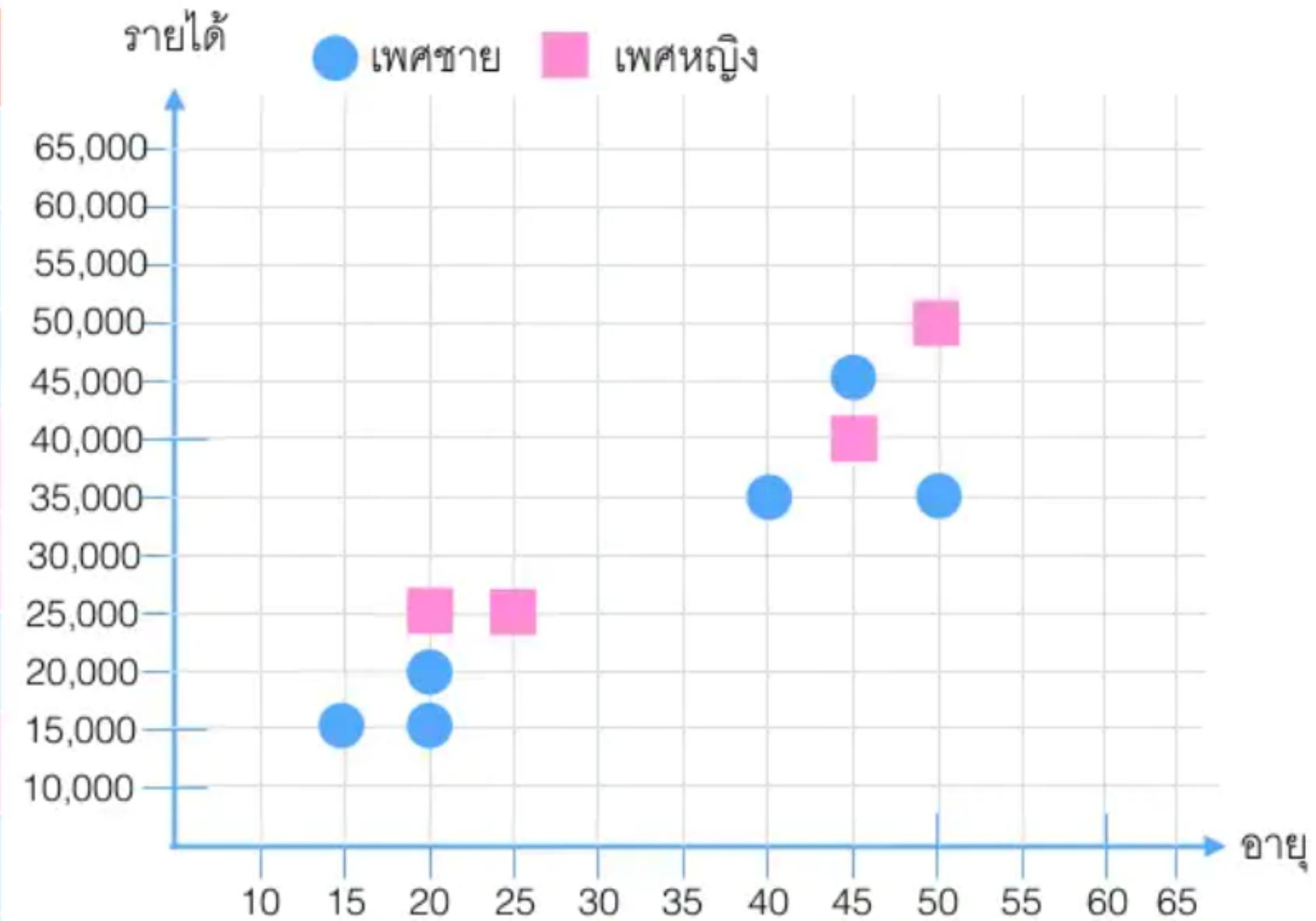
- การจัดการข้อมูล (preprocessing)
 - การเลือกแอตทริบิวต์
 - เลือกตามประเภทของแอตทริบิวต์
 - เลือกบางแอตทริบิวต์
 - การเลือกข้อมูล (example) ตามเงื่อนไข
 - การ join ข้อมูลจาก 2 ชุด
 - มีความผิดพลาดในชุดข้อมูล เช่น
 - ข้อมูลมีค่าไม่ตรงกัน
 - ข้อมูลขาดหายไป
 - ข้อมูลแปลกแยก (outlier)
- การแปลงข้อมูล
 - Discretization แปลงข้อมูล numeric ให้เป็น nominal
 - แบ่งตามเงื่อนไขที่กำหนด (user defined)
 - แบ่งตามความถี่ที่เท่ากัน (equal frequency)



Outlier

หมายเลข	อายุ	รายได้	เพศ
1	15	15,000	ชาย
2	20	15,000	ชาย
3	20	20,000	ชาย
4	20	25,000	หญิง
5	25	25,000	หญิง
6	40	35,000	ชาย
7	45	40,000	หญิง
8	45	45,000	ชาย
9	50	35,000	ชาย
10	50	50,000	หญิง

ข้อมูลลูกค้า

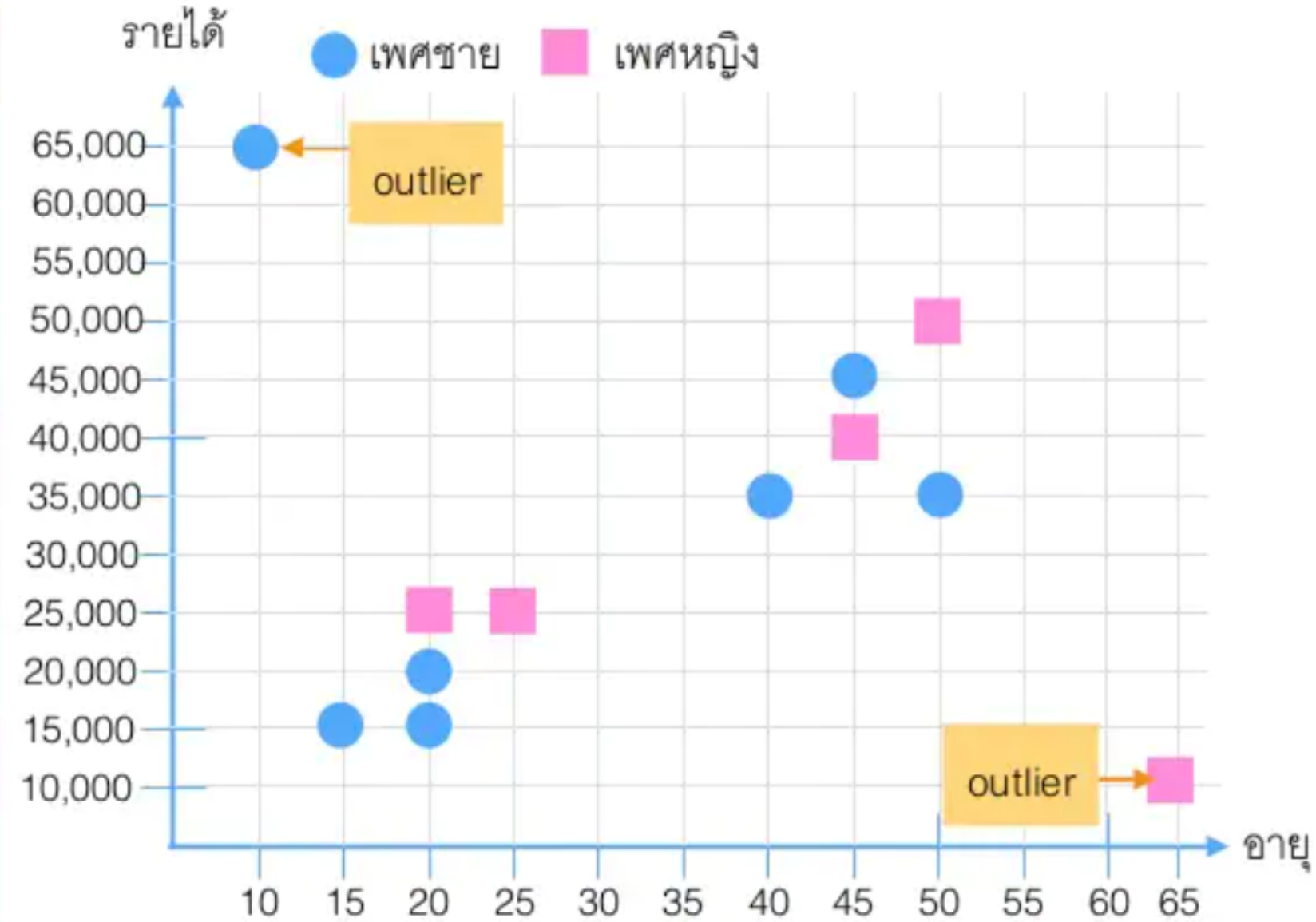


scatter plot ของข้อมูลลูกค้า



Outlier

หมายเลข	อายุ	รายได้	เพศ
1	15	15,000	ชาย
2	20	15,000	ชาย
3	20	20,000	ชาย
4	20	25,000	หญิง
5	25	25,000	หญิง
6	40	35,000	ชาย
7	45	40,000	หญิง
8	45	45,000	ชาย
9	50	35,000	ชาย
10	50	50,000	หญิง
11	10	65,000	ชาย
12	65	10,000	หญิง

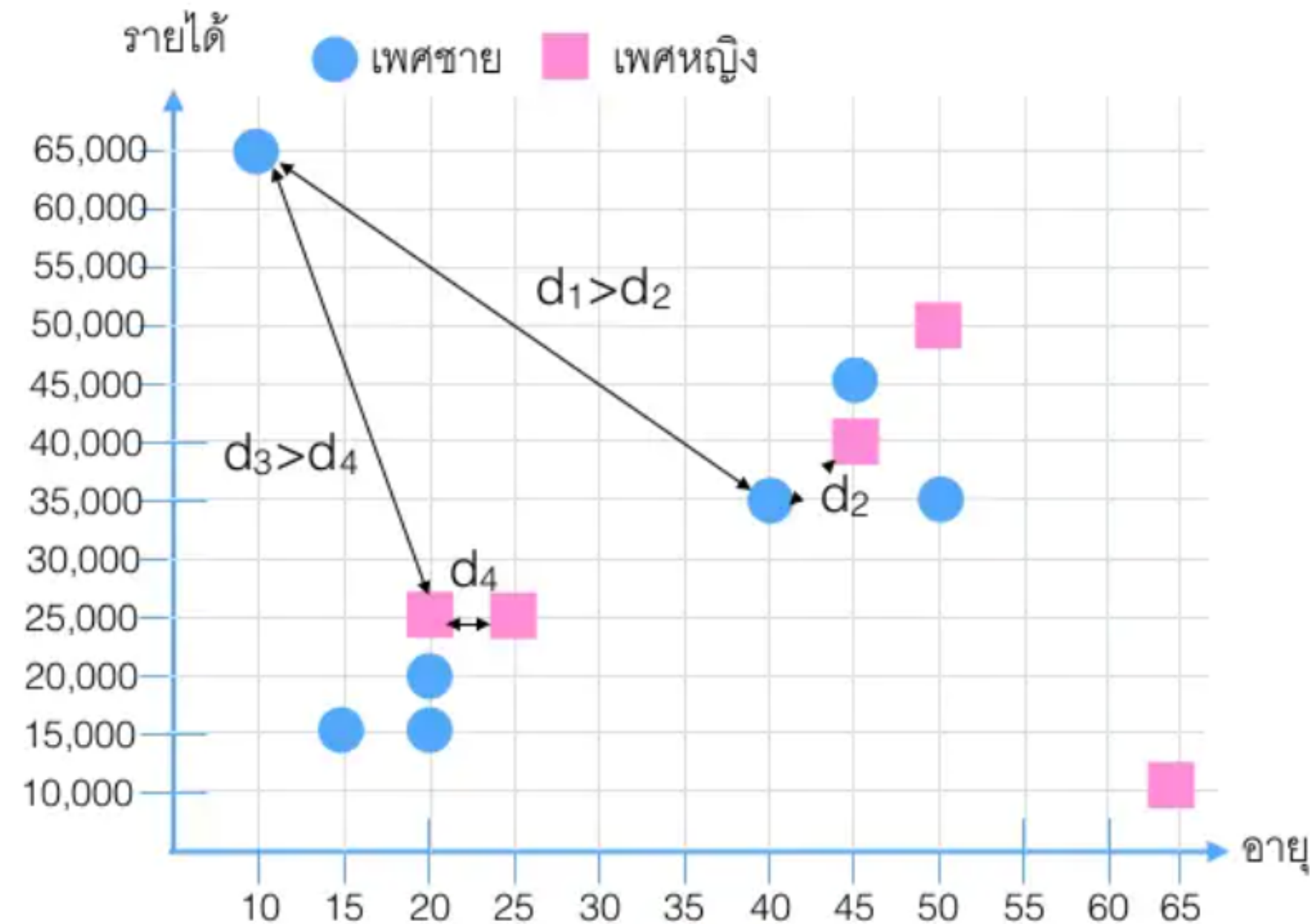


scatter plot ของข้อมูลลูกค้า



Detect outlier by distance

- วัดระยะห่างระหว่างข้อมูล (example) แต่ละตัวกับข้อมูลใกล้เคียง
- outlier คือ ข้อมูลที่มีระยะห่างกับข้อมูลอื่นๆ มากๆ





Detect outlier by distance

- ไอเปอเรเตอร์ที่เกี่ยวข้อง

ไอเปอเรเตอร์		คำอธิบาย
	Generate Data	ใช้สำหรับสร้างข้อมูล
	Detect Outlier (Distances)	ใช้ในการหา outlier ด้วยวิธีดูจากระยะห่างระหว่างข้อมูล
	Filter Examples	ลบข้อมูลตามเงื่อนไขที่กำหนด



Detect outlier by distance

- เลือกตัวอย่าง Process ที่ RapidMiner Studio 7 เตรียมไว้ให้
- เลือก Samples > Processes > 02_Preprocessing > 17_OutlierDetection

The screenshot displays the RapidMiner Studio Professional 7.0.000 interface. The main workspace shows a process flow starting with 'ExampleSetGeneral', followed by 'DistanceBasedOutlierDetection', and then 'ExampleFilter'. The 'DistanceBasedOutlierDetection' process is highlighted with a blue circle labeled '2'. The 'Parameters' panel on the right shows the following settings:

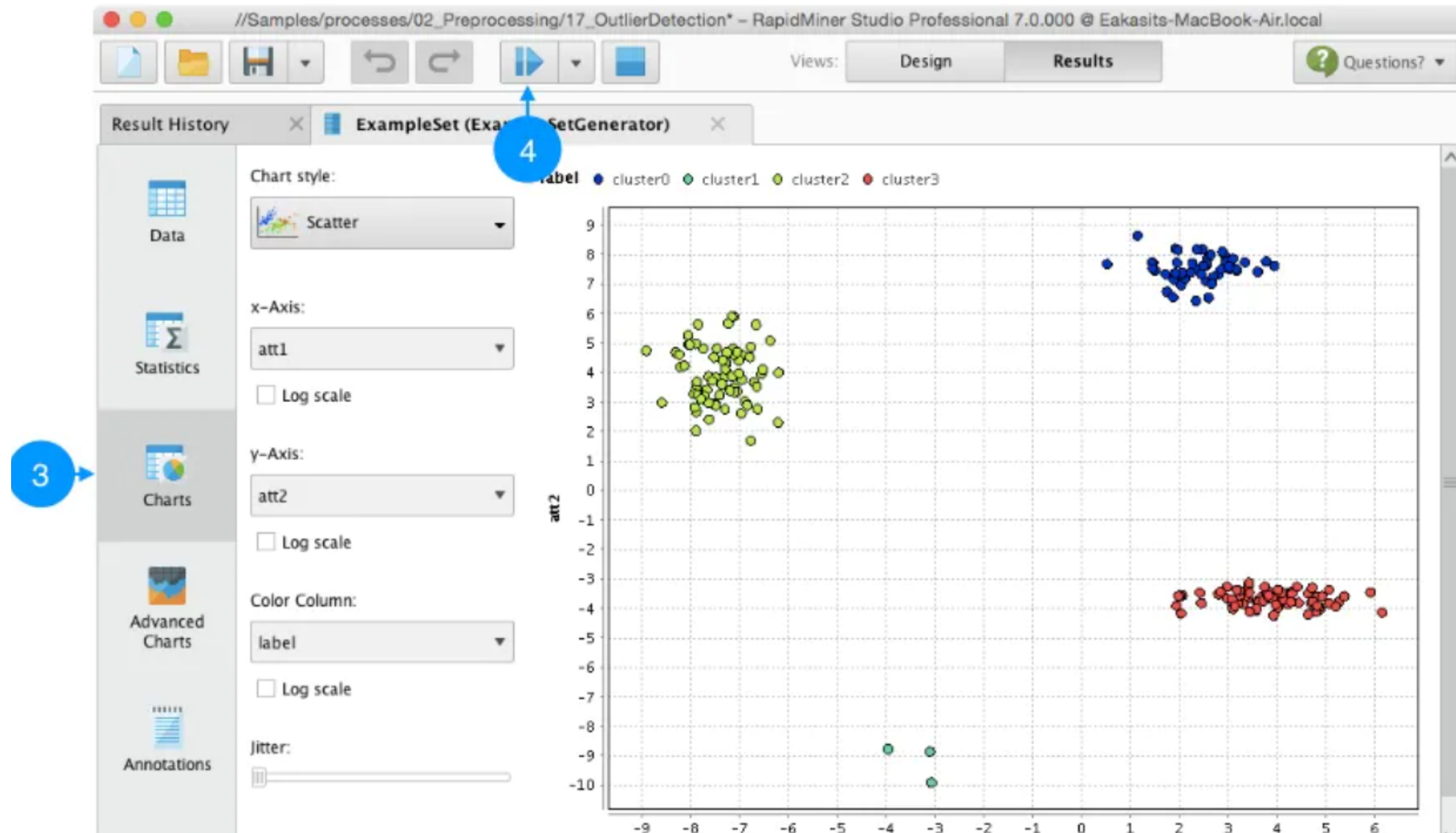
Parameter	Value
number of neighbors	4
number of outliers	12
distance function	euclidian distance

In the 'Repository' panel on the left, the '17_OutlierDetection' process is highlighted with a blue circle labeled '1'. A yellow callout box with Thai text 'double click เพื่อโหลด process' (double click to load process) points to this process. The 'Process' panel at the top shows a play button icon, which is also highlighted with a blue circle labeled '2'.



Detect outlier by distance

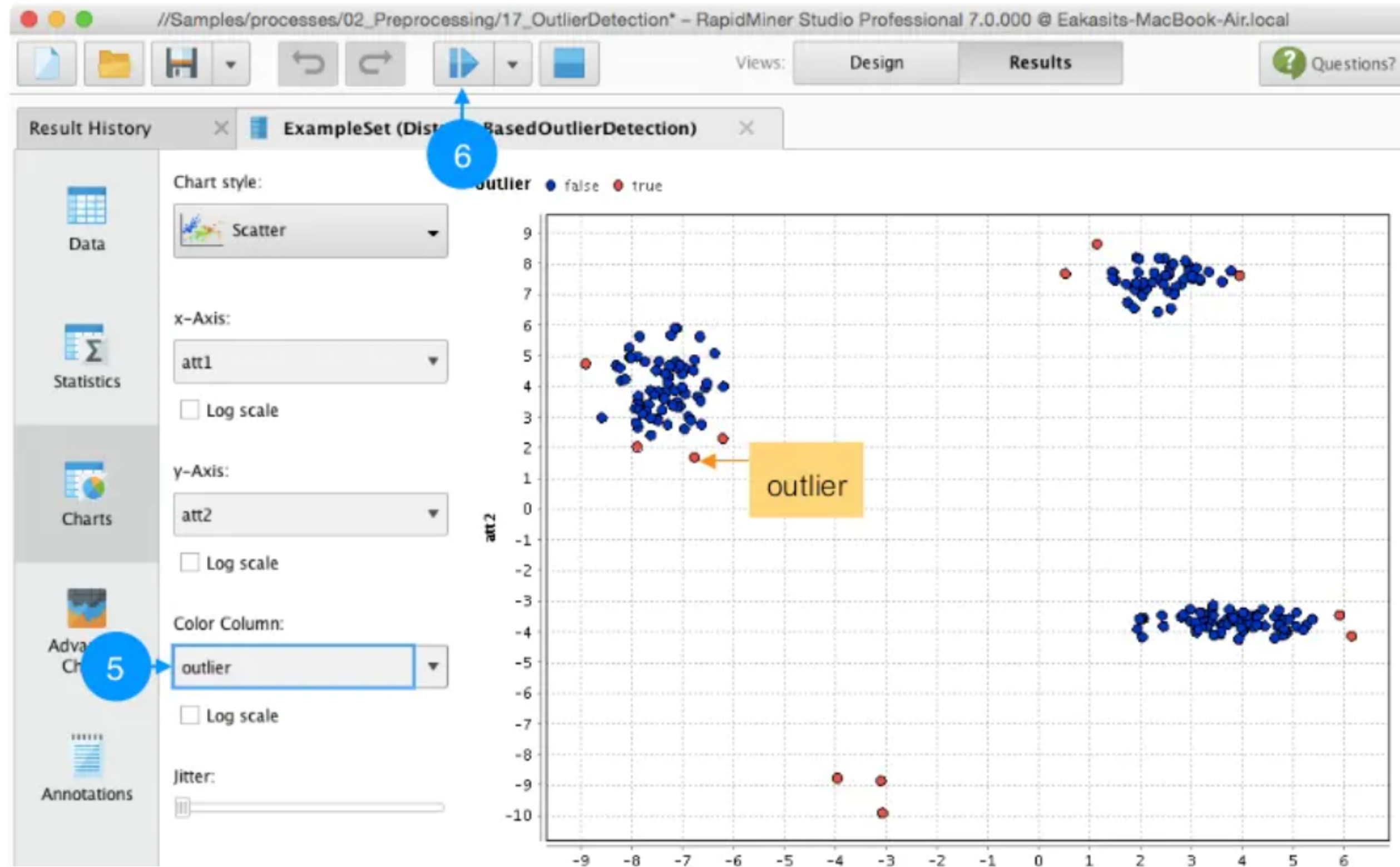
- แสดงข้อมูลแบบ scatter plot





Detect outlier by distance

- แสดงข้อมูล outlier ใน scatter plot





Detect outlier by distance

- ลบข้อมูลที่เป็น outlier ที่

Result History: ExampleSet (ExampleFilter)

Name	Type	Missing	Statistics	Filter (4 / 4 attributes):
label	Nominal	0	Least cluster1 (0) Most cluster3 (73)	Search for Attribute
outlier	Binominal	0	Least true (0) Most false (188)	
att1	Real	0	Min -8.618 Max 5.348 Average -0.622	
att2	Real	0	Min -4.197 Max 8.251 Average 1.918	

Showing attributes 1 - 4 Examples: 188 Special Attributes: 2 Regular Attributes: 2