

# Chapter 2

NO.

DATE : / /

Data Mining คือ การค้นหาความรู้จากข้อมูล

Ex

	Attribute 1	Attribute
Record 1	1	12
Record 2	2	11
Record 3	1	15

ตัวอย่าง

- ค้นหาว่า 1 record มี Data attribute 1 = 1
- ค้นหาว่า 1 record มี Data attribute 2 = 2

## ประเภทของ Data

Data มี 2 อย่าง คือ Data Object กับ Attribute

### 1. Record Data

- Data base
- Transaction data คือ ข้อมูลที่บันทึกการกระทำของ user 1 action  
ข้อมูลที่มีลักษณะ เช่น - ข้อมูลที่มีการเปลี่ยนแปลงตลอดเวลา และมีการ update ข้อมูล  
- การบันทึก 1 action
- Term-frequency ข้อมูลที่เป็น Text ที่ใช้ในการวิเคราะห์ข้อมูล

Data ที่ถูกอธิบายด้วย Attribute หรือกลุ่มข้อมูลที่เรียกว่า Data Object เรียกว่า "Data Object"

### 2. Data Graphs and Networks

Data 1 จุดที่เชื่อมกับจุดอื่น } เช่น ข้อมูลแผนที่ แผนที่ทางบก [www.thai-map.com](http://www.thai-map.com)  
หรือข้อมูลเครือข่าย

### 3. Ordered Data

- Data ที่มีการเรียงลำดับ เช่น ข้อมูล
- time series เช่น ข้อมูลย้อนหลัง 1 ปี จนถึง 99
- Video data
- Sequential Data เช่น DNA sequence ข้อมูลลำดับเบส
- Genetic sequence data

### 4. Spatial, image and multimedia Data

ข้อมูลเชิงพื้นที่ / ข้อมูลสื่อหลายมิติ เช่น ข้อมูล X, Y และภาพหรือเสียง  
ข้อมูลเชิงพื้นที่ เช่น แผนที่ 3 มิติ หรือข้อมูล 3 มิติ



## Important Characteristics of Structured Data

(ลักษณะของข้อมูลที่มีโครงสร้าง) (มีโครงสร้าง)

- Dimensionality  $\Rightarrow$  มีข้อมูลอยู่ 2 แบบ 1. ชนิดของข้อมูล  
ข้อมูลที่มี Dimension มากกว่า 1 คือ Data ที่มีหลายมิติ
- Sparsity  $\Rightarrow$  สันนิษฐานว่าค่าที่ว่างในข้อมูลมีค่าเป็น 0 (ส่วนใหญ่)
- Resolution  $\Rightarrow$  รูปแบบของข้อมูลที่มีอยู่
- Distribution  $\Rightarrow$  การกระจายของข้อมูล

## Data Objects.

เป็น Data ที่ประกอบด้วย Attribute  
เช่น Data base ใน row ก็คือ Data objects.

Attribute คือ คุณสมบัติที่จัดเป็นข้อมูล หรือเรียกอีกอย่างหนึ่งว่า - dimensions

- features

ชนิดของข้อมูลของ Attribute มีหลายแบบ

- variables

1. Nominal ข้อมูลที่เป็นชื่อหรือประเภท

2. Binary ข้อมูลที่มี 2 ค่า

3. Ordinal ข้อมูลที่เรียงลำดับได้ ไม่สามารถ  $+$   $-$   $\times$   $\div$  ได้

4. Numeric ข้อมูลที่ถูกแปลงให้เป็นตัวเลข  $+$   $-$   $\times$   $\div$  ได้ แบ่งได้เป็น

$\Rightarrow$  Interval scaled เป็นข้อมูลที่มี 0 ไม่แท้ เช่น อุณหภูมิ

$\Rightarrow$  Ratio scaled เป็นข้อมูลที่มี 0 แท้ เช่น 0 กิโลเมตร คือไม่มีการเดินทาง

ข้อมูลอาจเป็น Discrete หรือ Continuous Attribute

Discrete Attribute ข้อมูลที่ไม่ต่อเนื่อง ข้อมูลที่ไม่สามารถหาค่าที่อยู่ระหว่างกันได้  
เช่น เลขบัตรประชาชน

Continuous Attribute ข้อมูลที่อยู่ระหว่างกันได้ เช่น 1 หรือ 2 ค่าที่อยู่ระหว่าง  
1 หรือ 2 คือ 0.5