

## สรุป Chapter 1 Introduction

### Data Warehouse & Data Mining

Data Warehouse คือ ข้อมูลเก็บมา มีการคัดแยกและจัดเก็บ หลังจากนั้น ข้อมูลที่ถูกเก็บจะนำไปทำ Data Mining ต่อไป

Warehouse หมายถึง โรงงาน

Data Mining คือ การสิ่งที่มีค่าในข้อมูลที่มีมากมาย

Mining หมายถึง การทำเหมือง

## Chapter 1. Introduction

- ☐ Why Data Mining? 
- ☐ What Is Data Mining?
- ☐ A Multi-Dimensional View of Data Mining
- ☐ What Kinds of Data Can Be Mined?
- ☐ What Kinds of Patterns Can Be Mined?
- ☐ What Kinds of Technologies Are Used?
- ☐ What Kinds of Applications Are Targeted?
- ☐ Major Issues in Data Mining
- ☐ A Brief History of Data Mining and Data Mining Society
- ☐ Summary

## Why Data Mining?

- ☐ The Explosive Growth of Data: from terabytes to petabytes
  - ☐ Data collection and data availability
    - ☐ Automated data collection tools, database systems, Web, computerized society
  - ☐ Major sources of abundant data
    - ☐ Business: Web, e-commerce, transactions, stocks, ...
    - ☐ Science: Remote sensing, bioinformatics, scientific simulation, ...
    - ☐ Society and everyone: news, digital cameras, YouTube
- ☐ We are drowning in data, but starving for knowledge!
- ☐ "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Data Mining คือการทำเหมืองข้อมูล และในปัจจุบันนี้ข้อมูลมีอยู่มากมายขึ้นเรื่อย ๆ และทุกอย่างสามารถเก็บข้อมูลได้ ยกตัวอย่างเช่น เซนเซอร์อุณหภูมิ เซนเซอร์อากาศ

เซนเซอร์ปริมาณน้ำฝน เวลาเก็บข้อมูลจะส่งเก็บทุกชั่วโมงทุกนาที ข้อมูลจึงเกิดการ ล้นมากขึ้นเรื่อย ๆ

## What Is Data Mining?



- ❑ Data mining (knowledge discovery from data)
  - ❑ Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - ❑ Data mining: a misnomer?
- ❑ Alternative names
  - ❑ Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- ❑ Watch out: Is everything “data mining”?
  - ❑ Simple search and query processing
  - ❑ (Deductive) expert systems



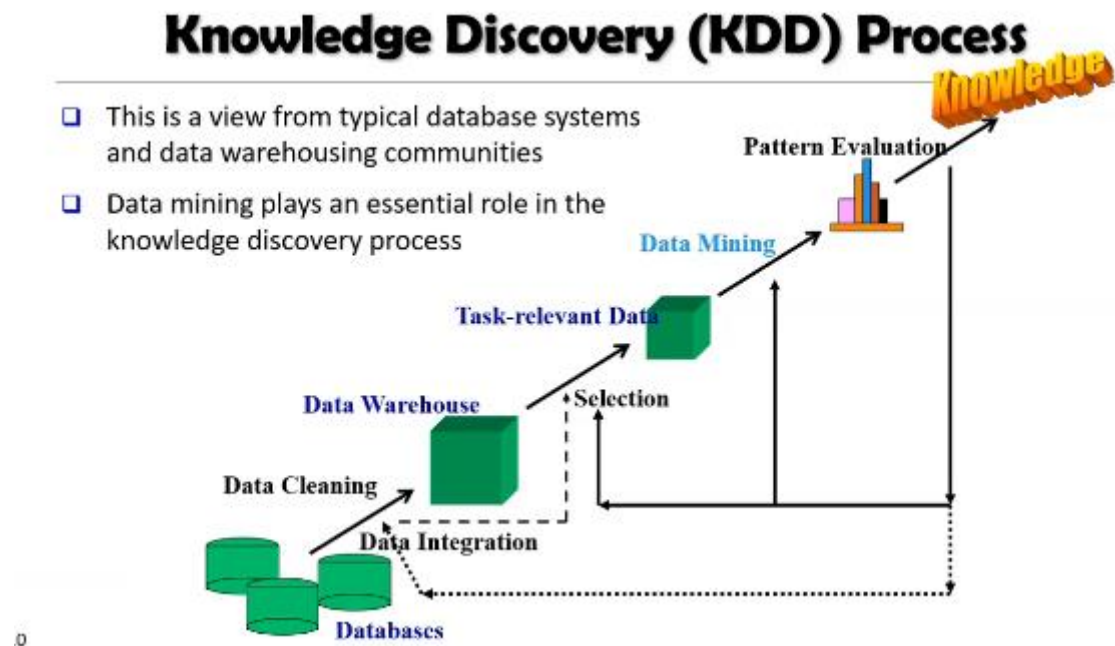
Data Mining การทำเหมืองข้อมูลคล้าย ๆ การที่ขุดเหมือง เราขุดหินดินสกัดไปเรื่อย ๆ จนเห็นเพชร ทอง Data Mining ก็เช่นเดียวกัน มีข้อมูลมากมายเต็มไปหมดแต่ไม่สามารถนำมาใช้โดยตรงได้ จึงต้องใช้ Data Mining สกัดข้อมูลนั้น ๆ

### การมีชื่อเรียกอื่น ๆ

- ❑ Alternative names
  - ❑ Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

\*แต่ละชื่อตั้งให้ตรงตามวัตถุประสงค์ของตัววิชา

## ลักษณะภาพรวม



เริ่มต้น 1. Databases เราจะเก็บข้อมูลยังไงให้มีประสิทธิภาพ

2. การจัดการข้อมูล

3. เอาข้อมูลหลาย ๆ แหล่งมารวมกันเพื่อไปเก็บไว้ใน Data Warehouse ดึงข้อมูลที่เราจำเป็นต้องใช้ประโยชน์มารวมกัน สามารถดูข้อมูลแบบสรุปหรือละเอียดได้

4. เอาข้อมูลเลือกเอาเฉพาะที่เราจะเอามาสกัดเอาความรู้เอามาทำ Data Mining

5. จะได้ Pattern รูปแบบที่ซ่อนอยู่ในข้อมูล

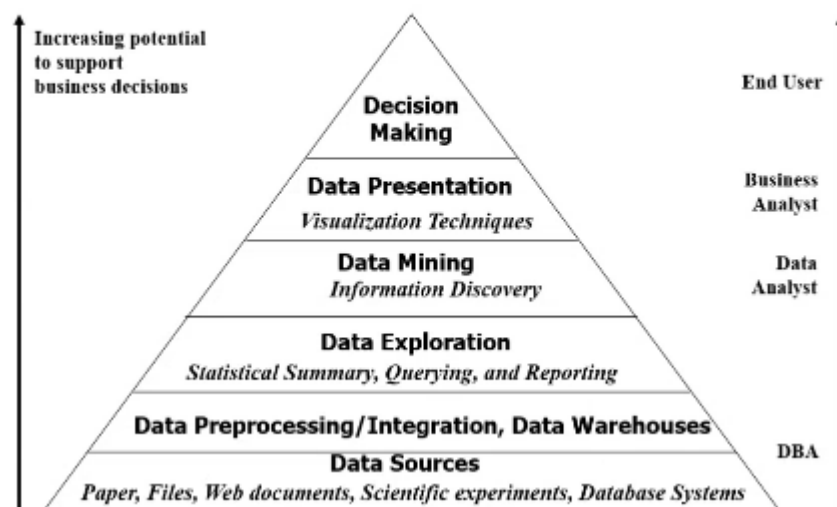
6. วัดผลดูว่าข้อมูลเราเป็นองค์ความรู้จริง ๆ จึงสรุปออกมาเป็นความรู้

## Example: A Web Mining Framework

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

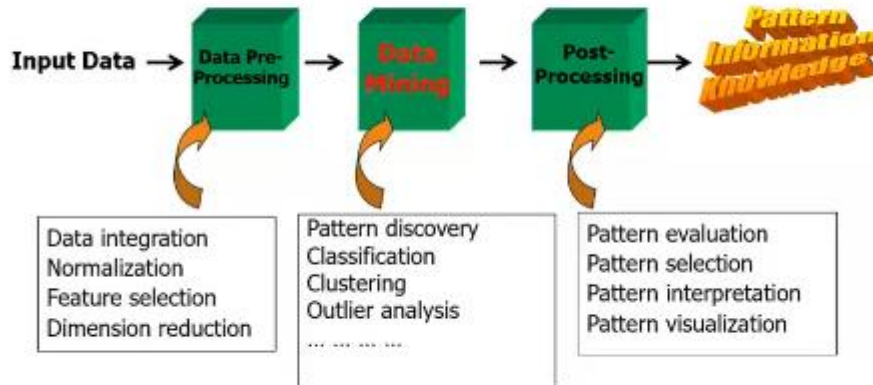
ยกตัวอย่างการทำ Mining Web ว่าทำอะไร มีขั้นตอนอย่างไร

## Data Mining in Business Intelligence



เป็นลักษณะคล้ายเดิม บอกขั้นตอน และตอนจบเป็นคนตัดสินใจว่าข้อมูลที่เรานำไปใช้อะไร

## KDD Process: A View from ML and Statistics



□ This is a view from typical machine learning and statistics communities

เราจะเรียนหลัก ๆ มีอยู่ 3 เรื่อง คือ

Pattern discovery (หารูปแบบที่ซ่อนอยู่ในข้อมูล)

Classification (การจำแนกข้อมูล)

Clustering (การแบ่งกลุ่มข้อมูล)

## How the data suppose to look like

	id	name	domain_id	closed	city_name	zipcode	geohash	new_open	weighted_average_rating	number_of_chains	...	good_for_groups
0	2	แมคโดนัลด์ สาขา...	2	0	Samut Sengkhram	75000	w4rh7g3	0	5.000000	NaN	...	NaN
1	4	Corner House	1	0	Bangkok Metropolitan Region	12150	w4rx73h	0	2.000000	NaN	...	NaN
2	5	ร้านกาแฟ...	4	0	Phra Nakhon Si Ayutthaya	13000	w4xs6jk	0	4.000000	NaN	...	NaN
3	6	ร้านกาแฟ...	1	0	Bangkok Metropolitan Region	10700.0	w4rqw8q	0	0.000000	NaN	...	NaN
4	7	Buono Caffè	1	0	Bangkok Metropolitan	10220	w4rx4gd	0	3.738482	NaN	...	NaN

ลักษณะข้อมูลที่เราจะเรียน มีการจัดการข้อมูลออกมาในรูปแบบนี้

แนวตั้ง (คอลัมน์) เป็น Attributes,Field,Features คำที่ใช้อธิบายคุณสมบัติของข้อมูล

แถว : Records, Data point ข้อมูลแต่ละตัว

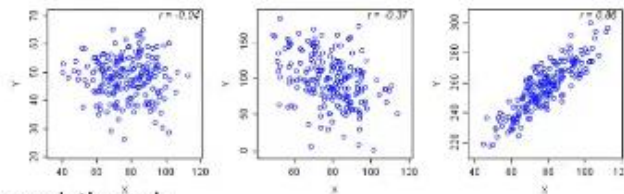
การสกัดหารูปแบบที่เกิดขึ้นซ้ำ ๆ ในข้อมูล

## Data Mining Functions: (2) Pattern Discovery

### ☐ Frequent patterns (or frequent itemsets)

- ☐ What items are frequently purchased together in your Walmart?

### ☐ Association and Correlation Analysis



### ☐ A typical association rule

- ☐ Diaper  $\rightarrow$  Beer [0.5%, 75%] (support, confidence)
- ☐ Are strongly associated items also strongly correlated?
- ☐ How to mine such patterns and rules efficiently in large datasets?
- ☐ How to use such patterns for classification, clustering, and other applications?

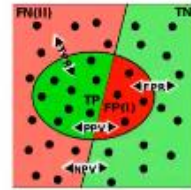
เทคนิคที่จะเรียนคือ association rule เป็นเทคนิคที่ทำให้คนรู้จัก Data Mining เพราะว่า เหมือนทำให้คนมหัศจรรย์ว่าสามารถทำอย่างนี้ด้วย

association rule ตัวอย่างที่ทำให้คนรู้จักคือ ผ้าอ้อมกับเบียร์ ใช้เทคนิคนี้ในการวิเคราะห์ใบเสร็จของคนมาซื้อป๊ิง เก็บรวมใบเสร็จกับคนมาซื้อป๊ิงไว้ สิ่งที่เกิดขึ้นคือคนที่ซื้อผ้าอ้อมมักจะซื้อเบียร์ด้วย ซึ่งคนที่ซื้อผ้าอ้อมกับเบียร์มักจะเป็นคุณพ่อที่มาซื้อ และคนที่รู้ข้อมูลนี้คือเจ้าของซูเปอร์มาเก็ตจึงสามารถเอาข้อมูลแพทเทินนี้ไปใช้เป็นแนวทางในการจัดทางในแบบต่าง ๆ



## Data Mining Functions: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - Ex. 1. Classify countries based on (climate)
    - Ex. 2. Classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...

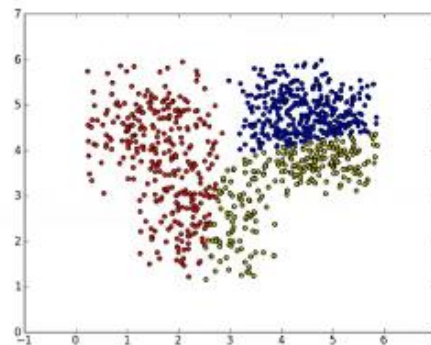


Classification เป็นการจำแนกกลุ่ม

- มีการเก็บข้อมูลทุก ๆ วัน
- มีค่า Y ในการทำนาย

## Data Mining Functions: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications



Cluster Analysis เป็นการจัดกลุ่ม

- จัดกลุ่มข้อมูล ไม่มีการทำนาย
- ข้อมูลคล้าย ๆ กันมาอยู่ด้วยกัน