นางสาวอุมาพร คำภิชัย รหัสนักศึกษา 623020547-0

สรุป Chapter 1 Introduction

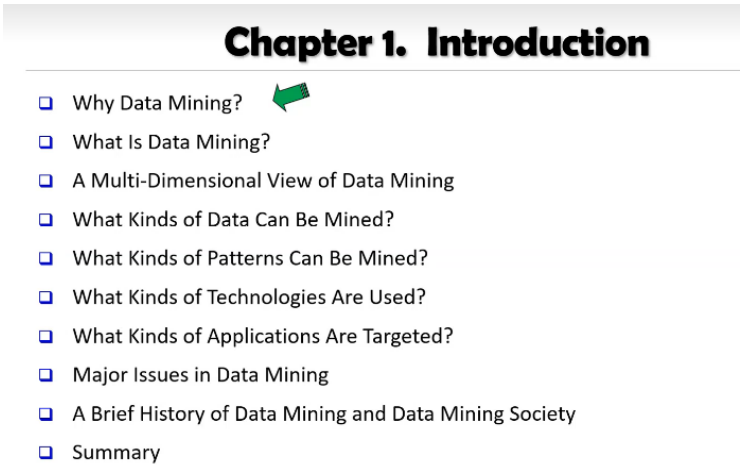
Data Warehouse & Data Mining

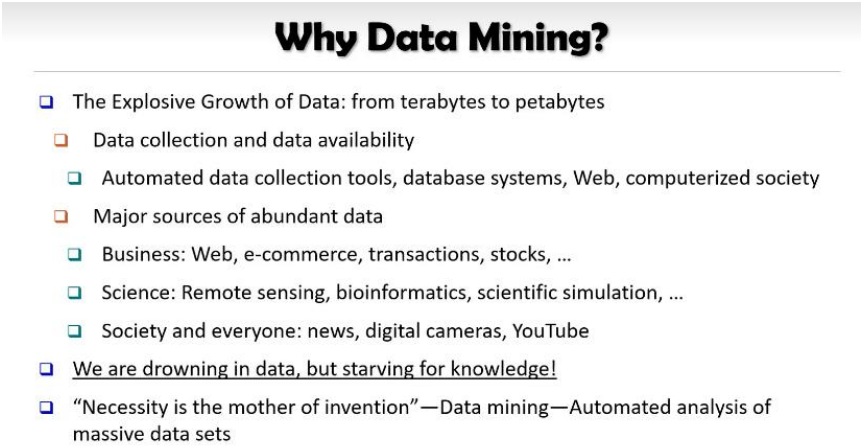
Data Warehouse คือ ข้อมูลเก็บมามีการคัดแยกและคัดเก็บ หลังจากนั้น ข้อมูลที่ถูกเก็บจะนำไปทำ Data Mining ต่อไป

Warehouse หมายถึง โรงงาน

Data Mining คือ การสิ่งที่มีค่าในข้อมูลที่มากมาย

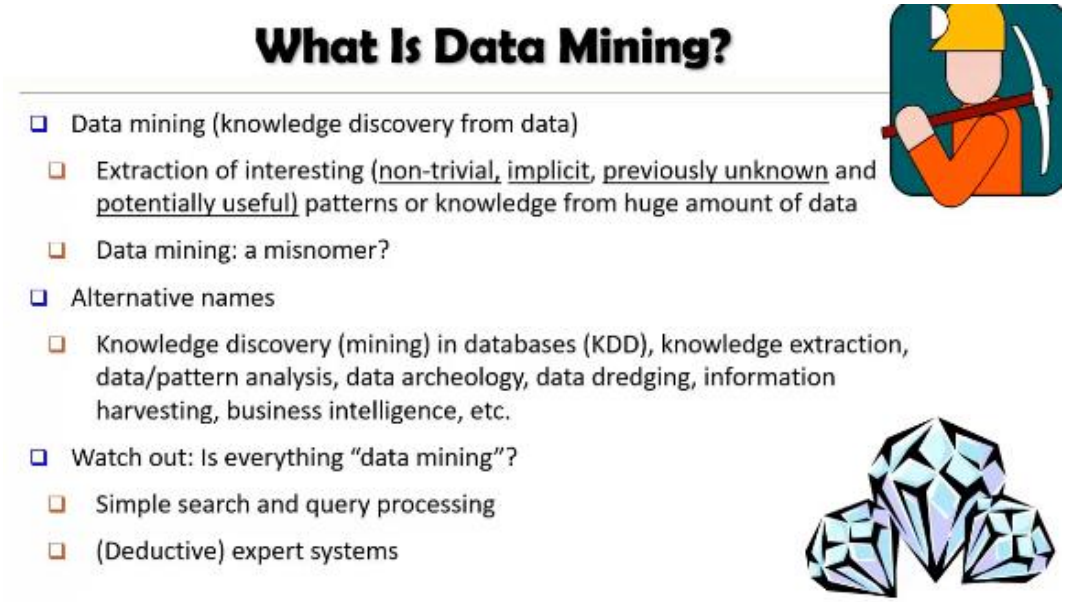
Mining หมายถึง การทำเหมือง





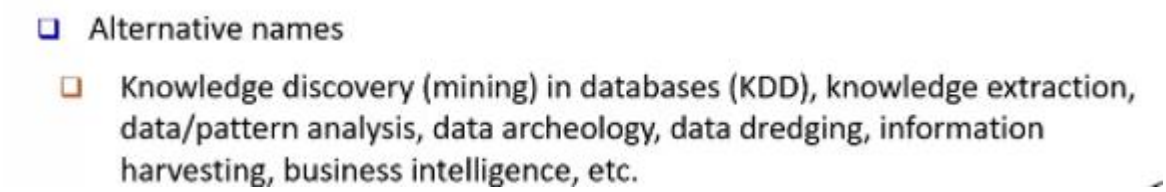
Data Mining คือการทำเหมืองข้อมูล และในปัจจุบันนี้ข้อมูลมีอยู่มากมายขึ้นเรื่อย ๆ และทุกอย่างสามารถเก็บข้อมูลได้ ยกตัวอย่างเช่น เซนเซอร์อุณหภูมิ เซนเซอร์อากาศ

เซนเซอร์ปริมาณน้ำฝน เวลาเก็บข้อมูลจะสุ่มเก็บทุกชั่วโมงทุกนาที ข้อมูลจึงเกิดการ ล้นมากขึ้นเรื่อย ๆ



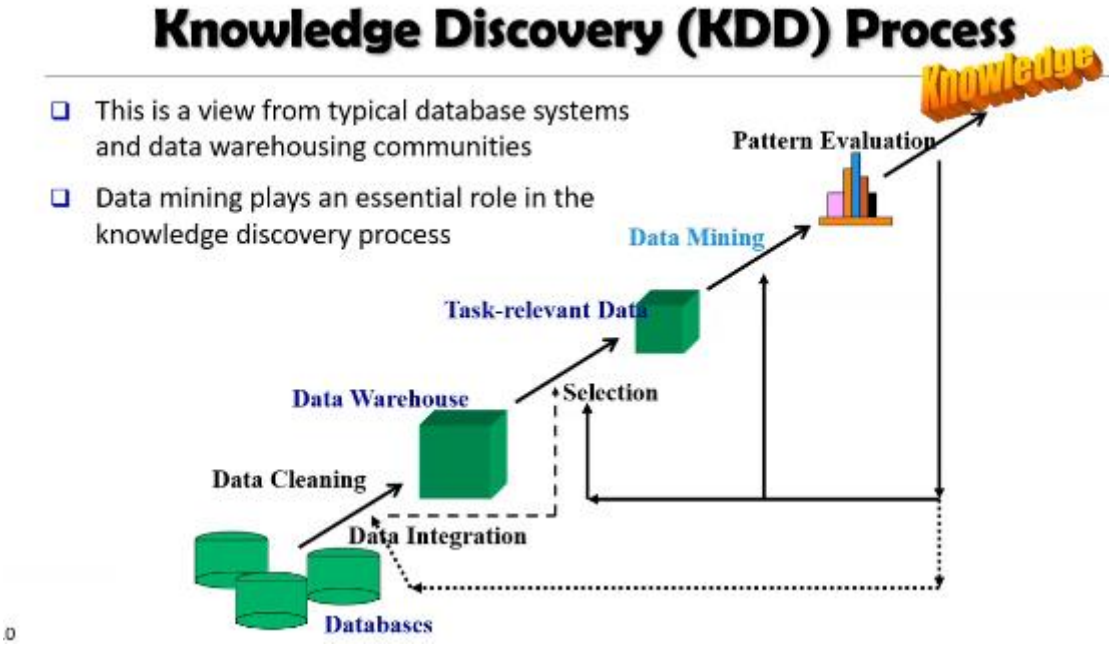
Data Mining การทำเหมืองข้อมูลคล้าย ๆ การที่ขุดเหมือง เราขุดหินดินสกัดไปเรื่อย ๆ จนเห็นเพชร ทอง Data Mining ก็เช่นเดียวกัน มีข้อมูลมากมายเต็มไปหมดแต่ไม่สามาถนำมาใช้โดยตรงได้ จึงต้องใช้ Data Mining สกัดข้อมูลนั้น ๆ

<u>การมีชื่อเรียกอื่น ๆ</u>



*แต่ละชื่อตั้งให้ตรงตามวัตถุประสงค์ของตัววิชา

ลักษณะภาพรวม



เริ่มต้น 1. Databases เราจะเก็บข้อมูลยังไงให้มีประสิทธิภาพ

2. การจัดการข้อมูล

3. เอาข้อมูลหลาย ๆ แหล่งมารวมกันเผื่อไปเก็บไว้ใน Data Warehouse ดึงข้อมูลที่เราจำเป็นที่ใช้ประโยชน์มารวมกัน สามารถดุข้อมูลแบบสรุปหรือละเอียดได้

4. เอาข้อมูลเลือกเอาเฉพาะที่เราจะเอามาสกัดเอาความรู้เอามาทำ Data Mining

5. จะได้ Pattern รูปแบบที่ซ่อนอยู่ในข้อมูล

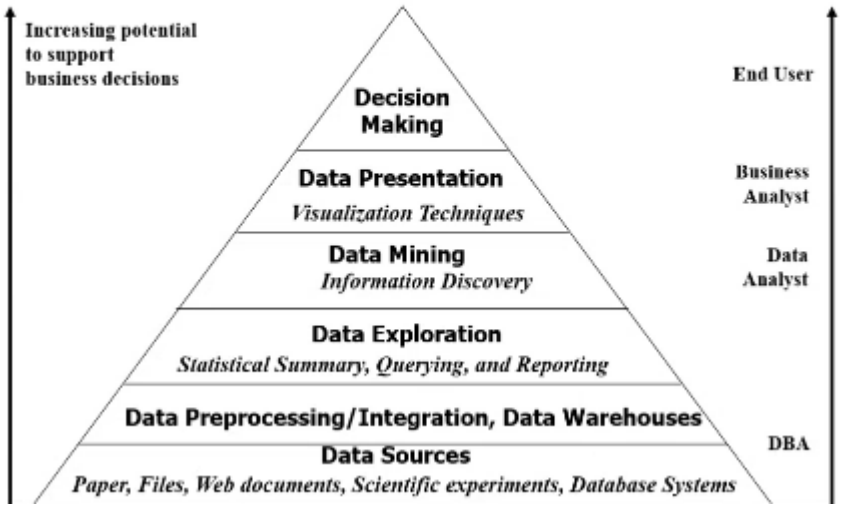
6. วัดผลดูว่าข้อมูลเราเป็นองค์ความรู้จริง ๆ จึงสรุปออกมาเป็นความรู้

Example: A Web Mining Framework

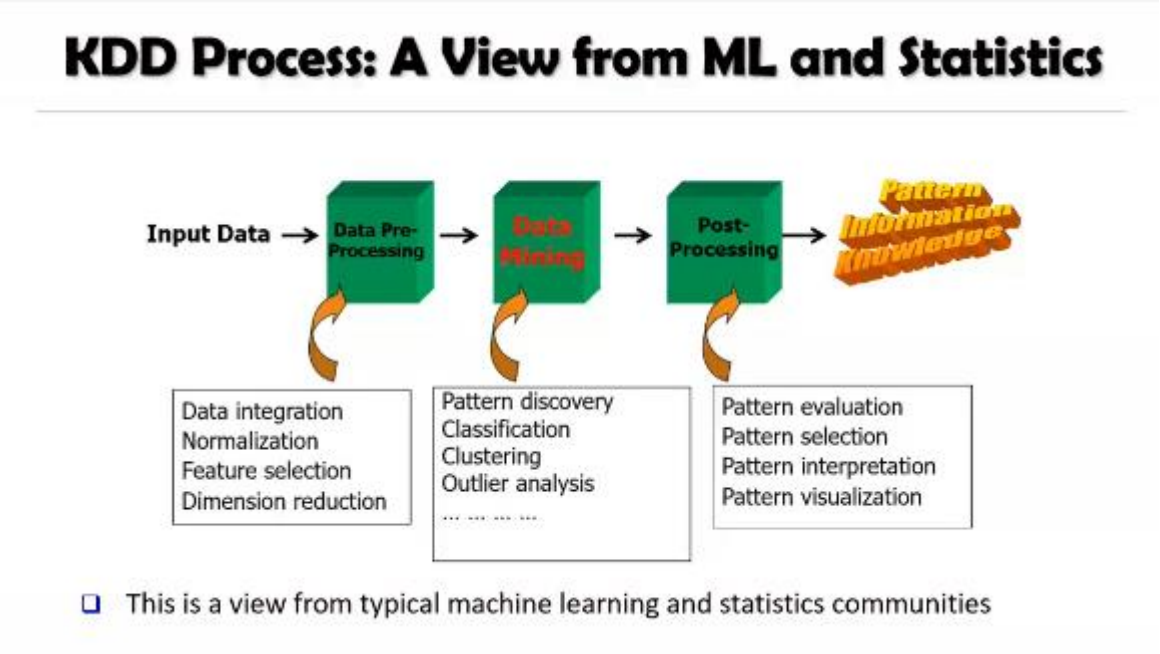
- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

ยกตัวอย่างการทำ Mining Web ว่าทำอย่างไร มีขั้นตอนอย่างไร

Data Mining in Business Intelligence



เป็นลักษณะคล้ายเดิม บอกขั้นตอน และตอนจบเป็นคนตัดสินใจว่าข้อมูลที่เราให้จะนำไปใช้อะไร

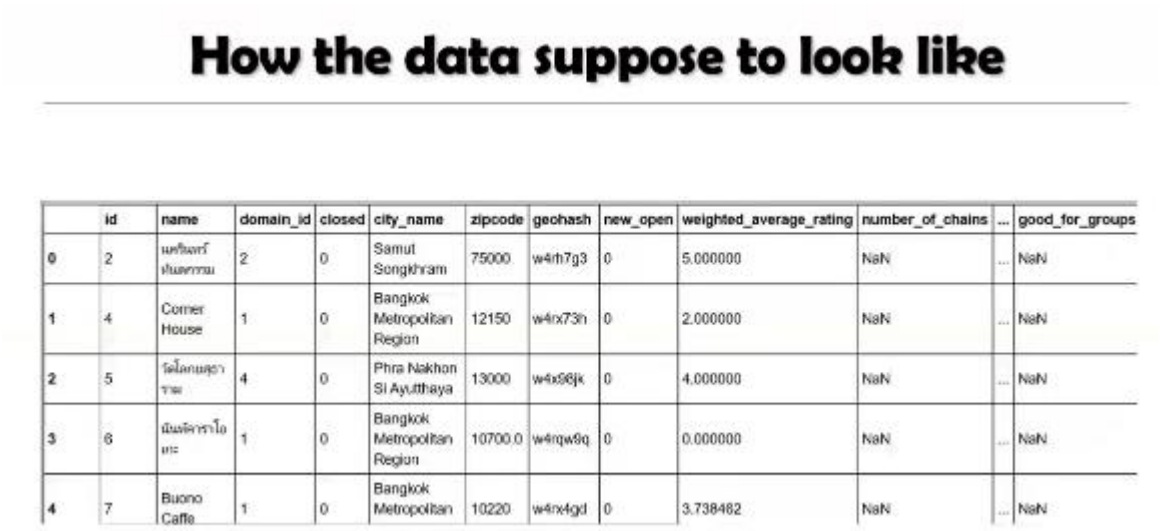


เราจะเรียนหลัก ๆ มีอยู่ 3 เรื่อง คือ

Pattern discovery (หารูปแบบที่ซ่อนอยู่ในข้อมูล)

Classification (การจำแนกข้อมูล)

Clustering (การแบ่งกลุ่มข้อมูล)



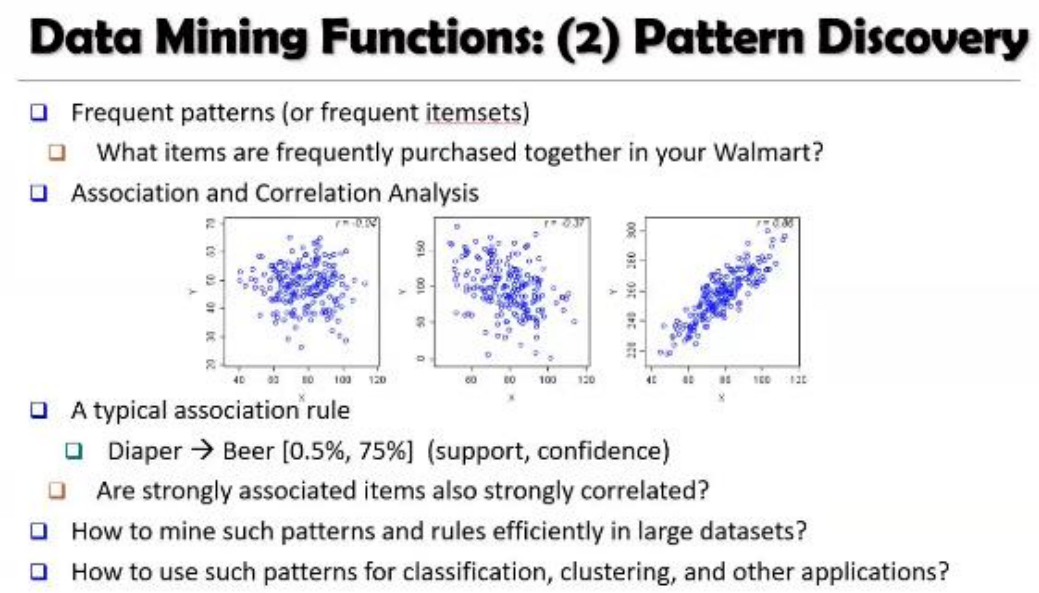
	id	name	domain_id	closed	city_name	zipcode	geohash	new_open	weighted_average_rating	number_of_chains	...	good_for_groups
0	2	[illegible]	2	0	Samut Songkhram	75000	w4rh7g3	0	5.000000	NaN	...	NaN
1	4	Corner House	1	0	Bangkok Metropolitan Region	12150	w4rx73h	0	2.000000	NaN	...	NaN
2	5	[illegible]	4	0	Phra Nakhon Si Ayutthaya	13000	w4x96jk	0	4.000000	NaN	...	NaN
3	6	[illegible]	1	0	Bangkok Metropolitan Region	10700.0	w4rqw8q	0	0.000000	NaN	...	NaN
4	7	Buono Caffe	1	0	Bangkok Metropolitan	10220	w4rx4gd	0	3.738462	NaN	...	NaN

ลักษณะข้อมูลที่เราจะเรียน มีการจัดการข้อมูลออกมาในรูปแบบนี้

แนวตั้ง (คอลลัมม์) เป็น Attributes,Field,Features คำที่ใช้อธิบายคุณสมบัติของข้อมูล

แถว : Records, Data point ข้อมูลแต่ละตัว

การสกัดหารูปแบบที่เกิดขึ้นซ้ำ ๆ ในข้อมูล

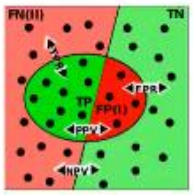


เทคนิคที่จะเรียนคือ association rule เป็นเทคนิคที่ทำให้คนรู้จัก Data Mining เพราะว่า เหมือนทำให้คนมหัศจรรย์ว่าสามารถทำอย่างนี้ด้วย

association rule ตัวอย่างที่ทำให้คนรู้จักคือ ผ้าอ้อมกับเบียร์ ใช้เทคนิคนี้ในการวิเคราะห์ใบเสร็จของคนมาช้อปปิ้ง เก็บรวมใบเสร็จกับคนมาช้อปปิ้งไว้ สิ่งที่เกิดขึ้นคือคนที่ซื้อผ้าอ้อมมักจะซื้อเบียร์ด้วย ซึ่งคนที่จะซื้อผ้าอ้อมกับเบียร์มักจะเป็นคุณพ่อที่มาซื้อ และคนที่รู้ข้อมูลนี้คือเจ้าของซุปเปอร์มาเก็ตจึงสามารถเอาข้อมูลแพทเทินนี้ไปใช้เป็นแนวทางในการจัดทางในแบบต่าง 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, …

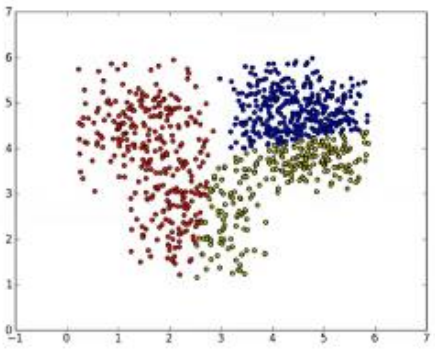


Classification เป็นการจำแนกกลุ่ม

- 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