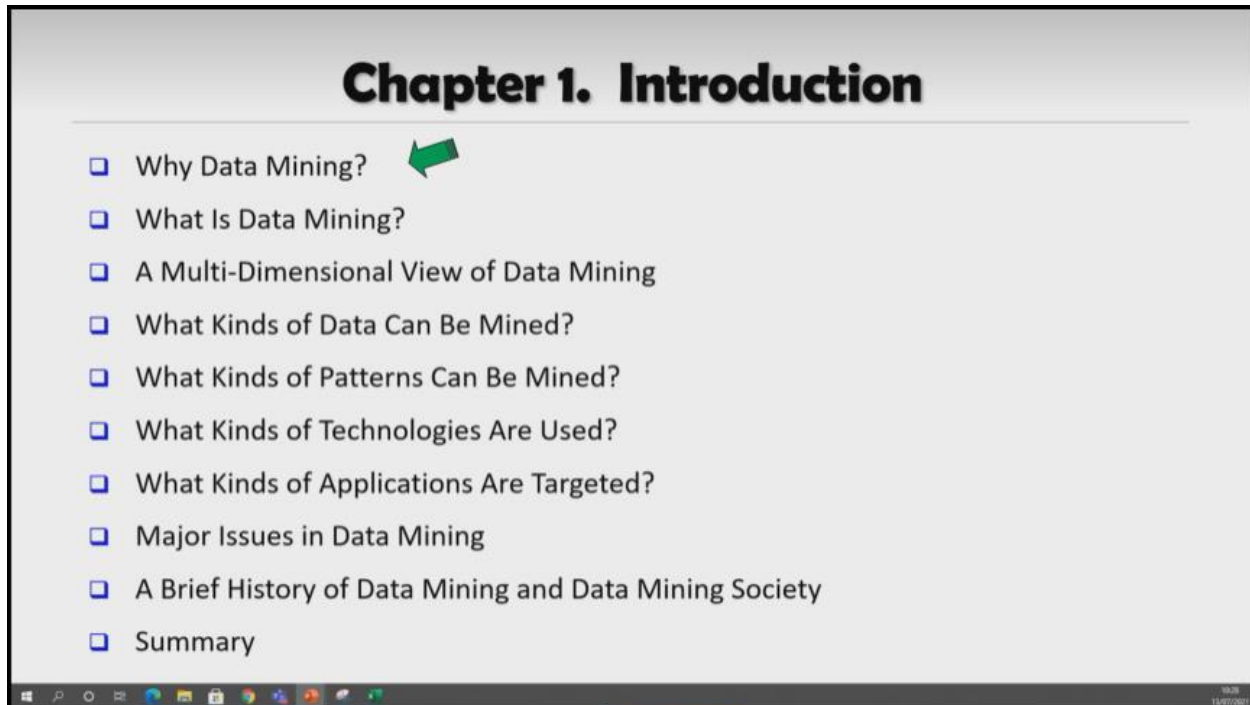



Data Warehouse & Data Mining

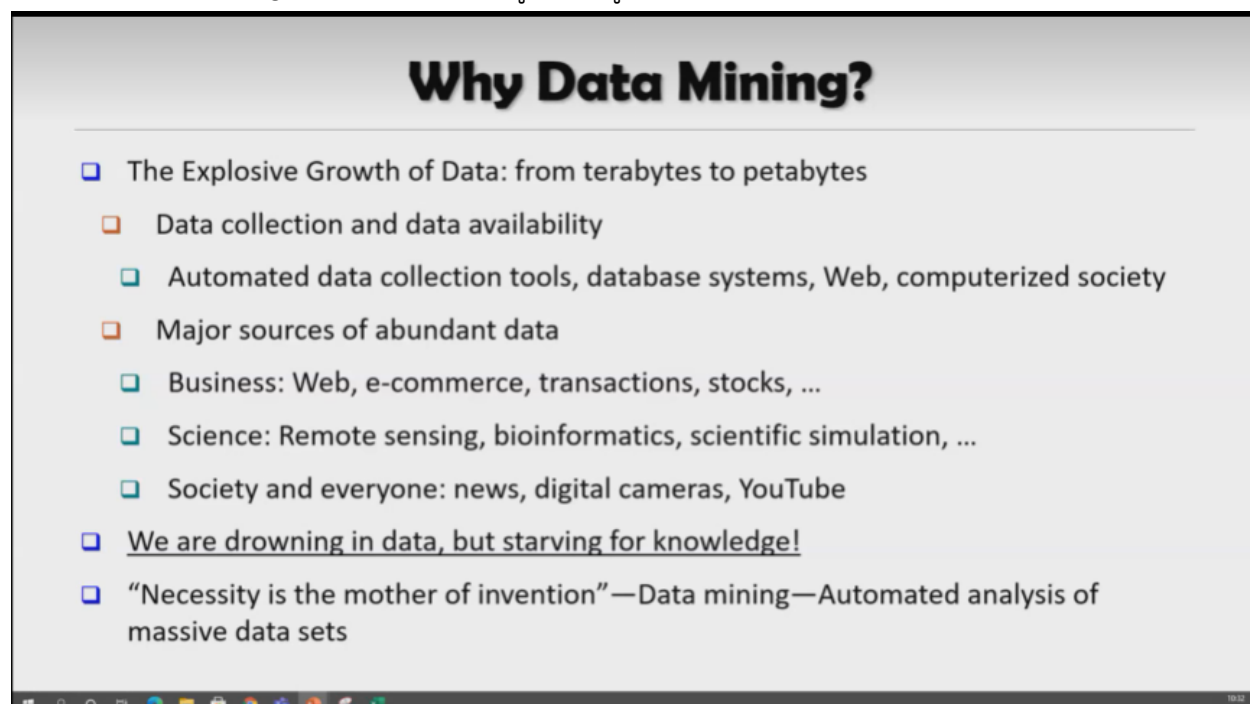
Chapter 1 Introduction



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

Data Mining คือ การค้นพบความรู้จากข้อมูล (เหมือนการทำเหมือง)



Why Data Mining?

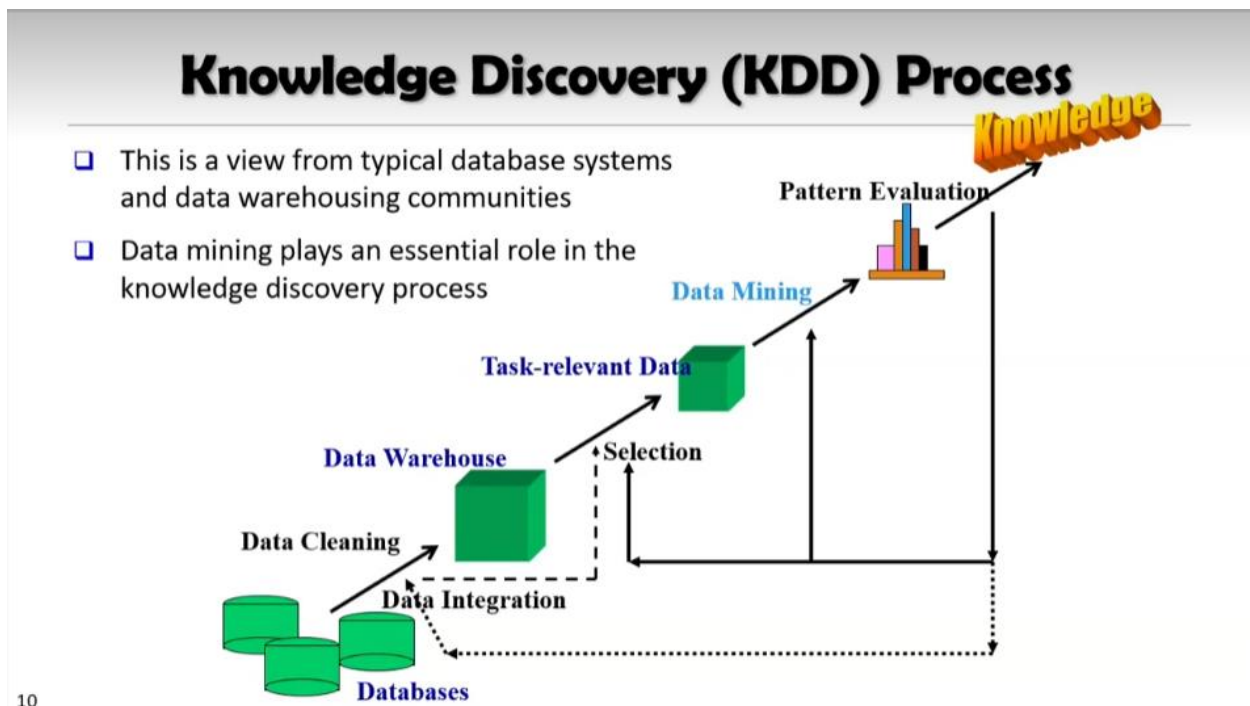
- ❑ The Explosive Growth of Data: from terabytes to petabytes
 - ❑ Data collection and data availability
 - ❑ Automated data collection tools, database systems, Web, computerized society
 - ❑ Major sources of abundant data
 - ❑ Business: Web, e-commerce, transactions, stocks, ...
 - ❑ Science: Remote 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 มาจาก *Knowledge Discovery (Mining) in databases (KDD)*

= การสกัดองค์ความรู้จากแหล่งข้อมูลหลายๆแหล่ง ที่กล่่าวมาคล้ายกับการทำเหมือง
(แหล่งข้อมูลสาธารณะ :: <https://data.go.th/>)

เราจะเก็บข้อมูลอย่างไรให้มีประสิทธิภาพ?

- เมื่อเราจัดการข้อมูลเสร็จแล้ว เราจะนำข้อมูลที่ได้มาจากหลายๆแหล่งรวมกันเพื่อเก็บไว้ใน Data Warehouse



** Data Warehouse เป็นขั้นตอนเบื้องต้น เช่น การนำเอาข้อมูลมารวมกัน/ดูข้อมูลอย่างละเอียด
หลังจาก Data warehouse คลีนข้อมูลแล้ว จะนำเอา data ที่จะสกัดองค์ความรู้มาทำ Data Mining
เพื่อหารูปแบบที่ซ่อนในข้อมูล วิดผล ตรวจสอบ และสรุป

ตัวอย่าง

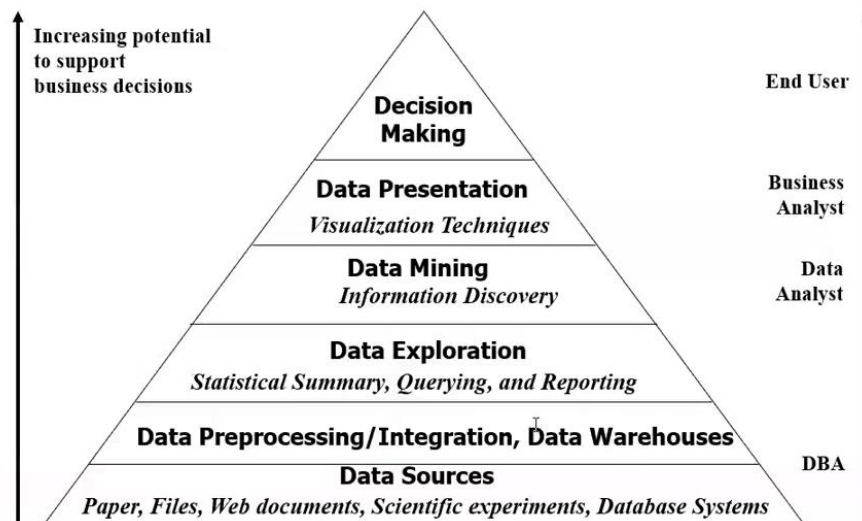
Example: A Web Mining Framework

- ❑ 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

11

สำคัญ ในการทำ Data Mining ต้องนำองค์ความรู้ที่สกัดมาได้ มาแสดงให้เห็นเพื่อนเข้าใจ ถึงจะถือว่าบรรลุ

Data Mining in Business Intelligence

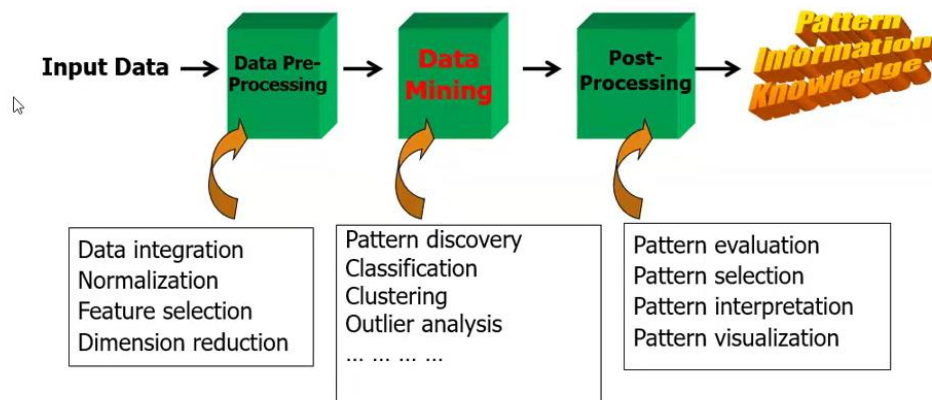


12

ที่เราจะเรียนหลักๆ (ที่เป็น Data Mining) คือ

- Pattern Discovery (หารูปแบบที่ซ่อนอยู่ในข้อมูล)
- Classification (การจำแนกข้อมูล)
- Clustering (การแบ่งกลุ่ม)

KDD Process: A View from ML and Statistics



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

13

How the data suppose to look like

Columns: Attributes, Fields, Features: ค่าที่ใช้อธิบายคุณสมบัติของข้อมูล

	id	name	domain_id	closed	city_name	zipcode	geohash	new_open	weighted_average_rating	number_of_chains	...	good_for_groups
0	2	เคอรี่ ทันตกรรม	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	วัดโลกยสุธา ราม	4	0	Phra Nakhon Si Ayutthaya	13000	w4x98jk	0	4.000000	NaN	...	NaN
3	6	นันทคาราโอ เกะ	1	0	Bangkok Metropolitan Region	10700.0	w4rqw9q	0	0.000000	NaN	...	NaN
4	7	Buono Caffe	1	0	Bangkok Metropolitan	10220	w4rx4gd	0	3.738462	NaN	...	NaN

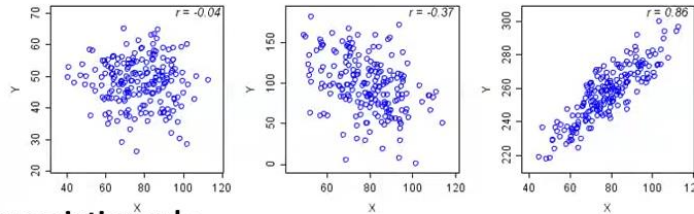
→ Rows: Records, Data point: ข้อมูลแต่ละตัว

17

Slide ที่ 18 คือ ตัวอย่างข้อมูล (Pattern Discovery) เทคนิค Association rule

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?

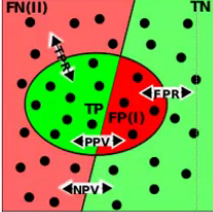
18

- จากใบเสร็จ คนที่ซื้อผ้าอ้อมมักจะซื้อเปียร์ด้วย วิเคราะห์ได้ว่า คนประเภทนี้คือ คุณพ่อที่ออกมา Shopping มักจะซื้อเปียร์กลับไปด้วย จึงทำให้เกิดการจัดร้าน 2 แนวทาง คือ
 1. จัดร้านให้ผ้าอ้อมกับเปียร์อยู่คนละที่กัน ทำให้เดินผ่านสินค้าอื่นๆด้วย
 2. จัดเปียร์กับผ้าอ้อมให้อยู่ติดกัน เพราะรู้ว่าคนที่มาซื้อผ้าอ้อมต้องซื้อเปียร์กลับไปแน่นอน(เป็นการอำนวยความสะดวกให้ลูกค้า) ทำให้เพิ่มยอดขายได้ด้วย

Slide ที่ 19 คือ ตัวอย่างข้อมูล (Classification)

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, ...



- ทำได้โดยเลือกข้อมูล 1 column มาทำนาย ต่างจาก Multivariate ตรงที่ค่าในตารางไม่ใช่ตัวเลข

Slide ที่ 20 คือ ตัวอย่างข้อมูล (Cluster Analysis)

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

