CS 412 Intro. to Data Mining

Chapter 1. Introduction

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

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?

จัดหมวดหมู่ข้อมูลจำนวนมากให้ค้นหาง่ายขึ้น

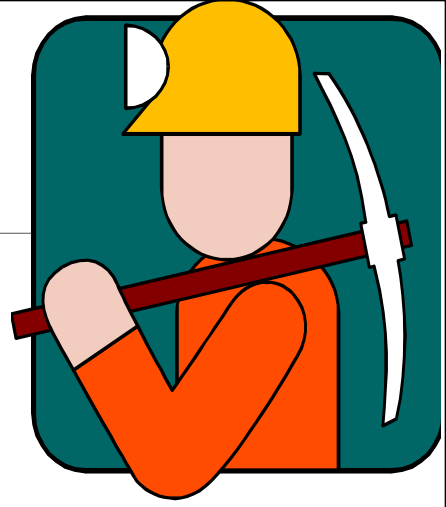
การทำเหมืองข้อมูล

- 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

ทุกอย่างรอบตัวสามารถเก็บข้อมูลได้หมด เช่น science มันสร้าง data ตลอดเวลา, Web ก็เก็บ data ของผู้ใช้งานอยู่ตลอด

Data Mining ที่คนสร้างขึ้นที่เรารู้จักและใช้บ่อย คือ Google photo, icloud

What Is Data Mining?



คือการสกัดความรู้ที่ซ่อนอยู่ภายใน

- 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

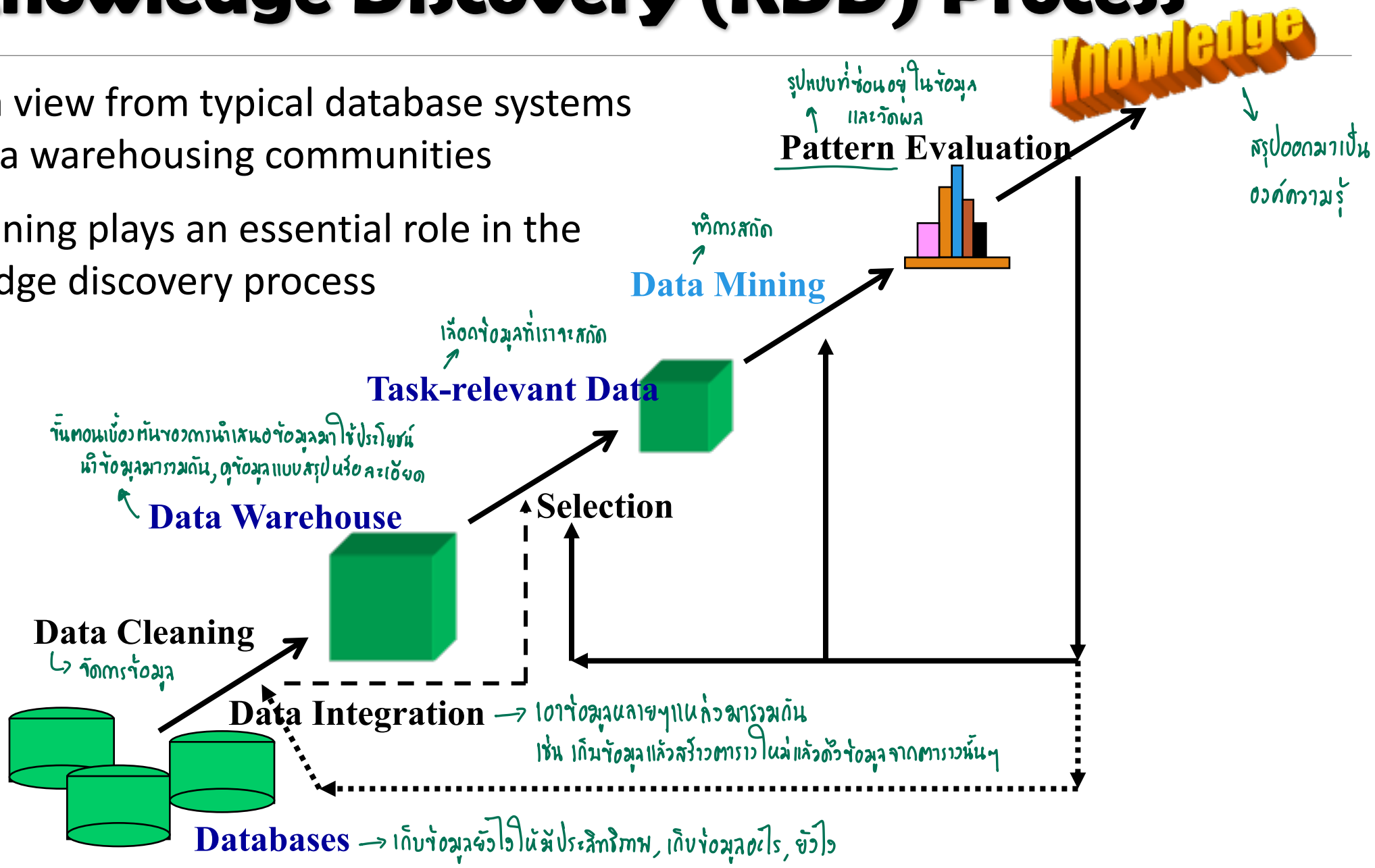
  การสกัดองค์ความรู้ของ database หลาย ๆ แหล่ง
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

ชื่อแรกก่อน จะเปลี่ยนเป็น Data mining

- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

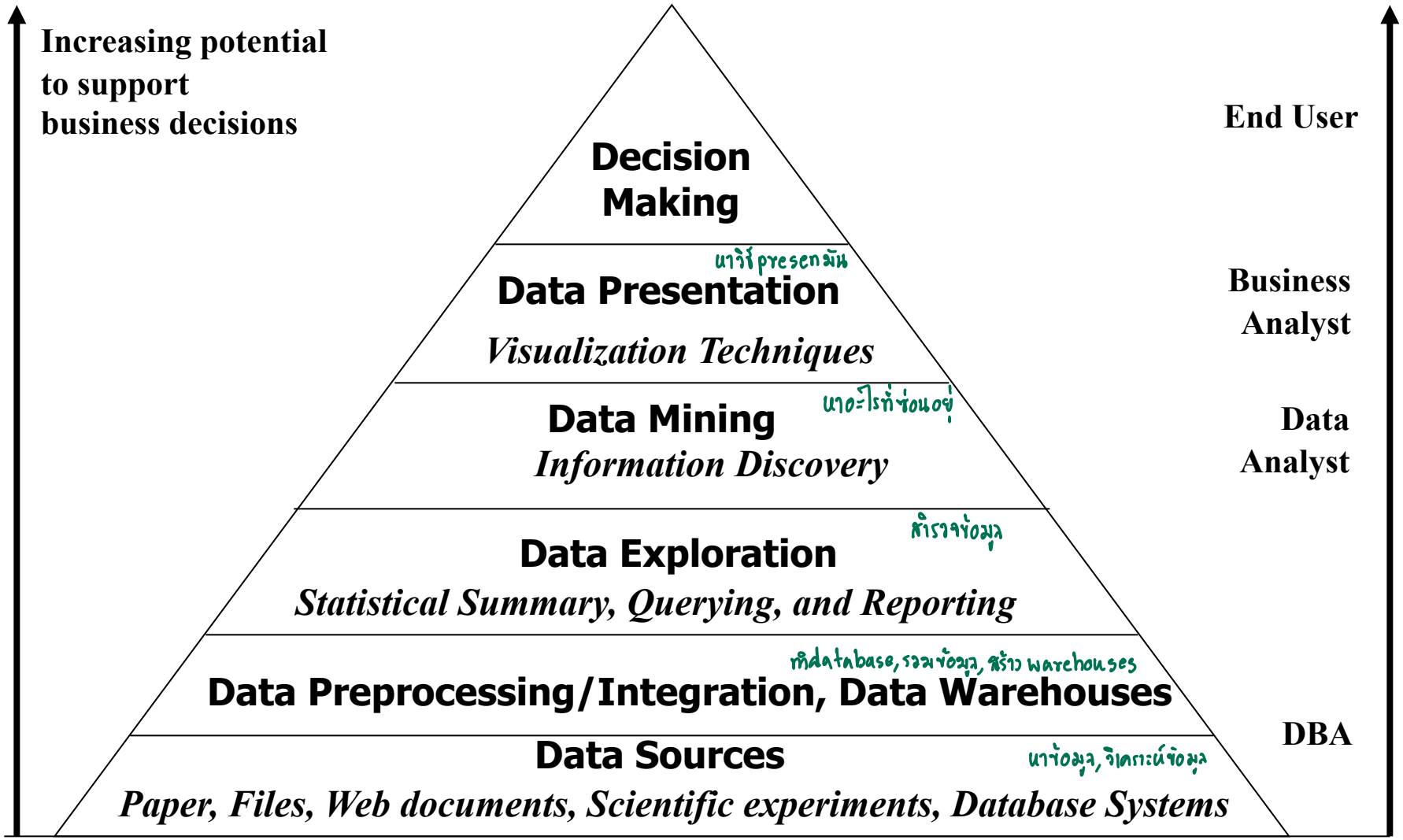


Example: A Web Mining Framework

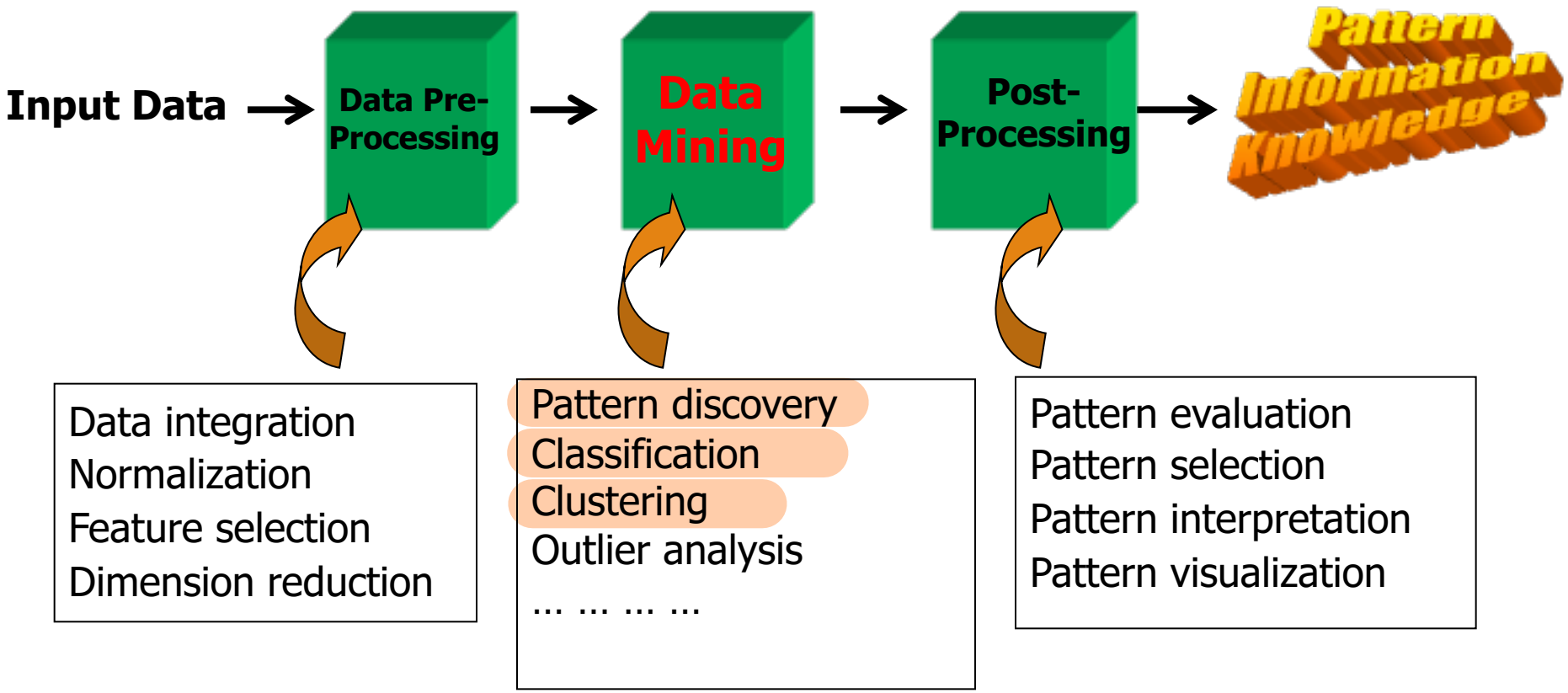
กรณี Web mining

- Web mining usually involves
  - Data cleaning → กรองข้อมูลที่ไม่เกี่ยวข้องออก
  - Data integration from multiple sources → รวมข้อมูลจากหลายแหล่งให้เป็นชุดเดียว
  - Warehousing the data → เก็บข้อมูลที่ผ่านไม่สกัด
  - Data cube construction
  - Data selection for data mining → ดึงข้อมูลสำหรับวิเคราะห์จากแหล่งข้อมูล
  - Data mining → สกัดข้อมูล
  - Presentation of the mining results → presen ข้อมูลที่เราได้มาหลังจากสกัดแล้ว.
  - Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



KDD Process: A View from ML and Statistics

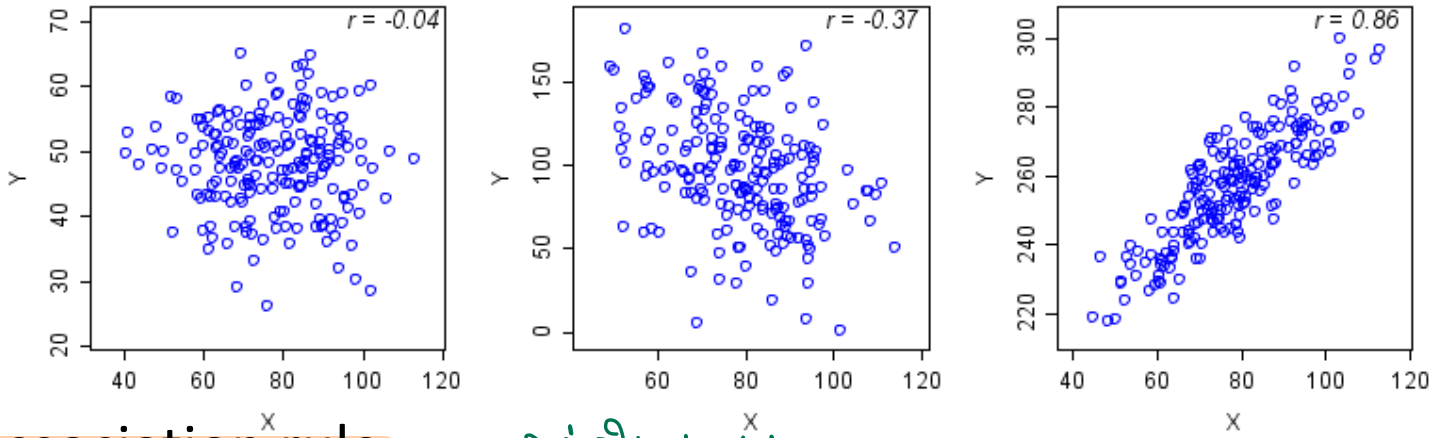


- This is a view from typical machine learning and statistics communities

Data Mining Functions: (2) Pattern Discovery

หารูปแบบที่ใช้ในข้อมูล

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis



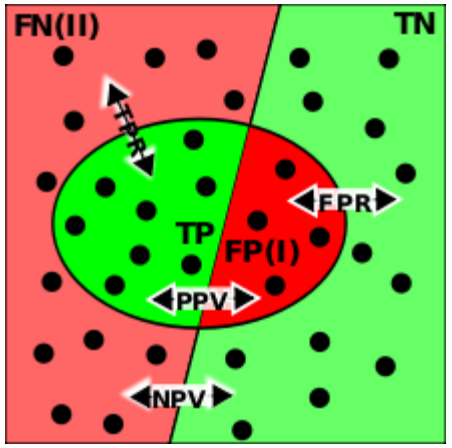
- A typical association rule → เทคนิคที่ทำให้คนรู้จัก data mining
  - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Functions: (3) Classification

การจำแนกกลุ่ม

ใช้ attributes เพื่อทำนาย attributes อื่น ใช้ทำนาย class

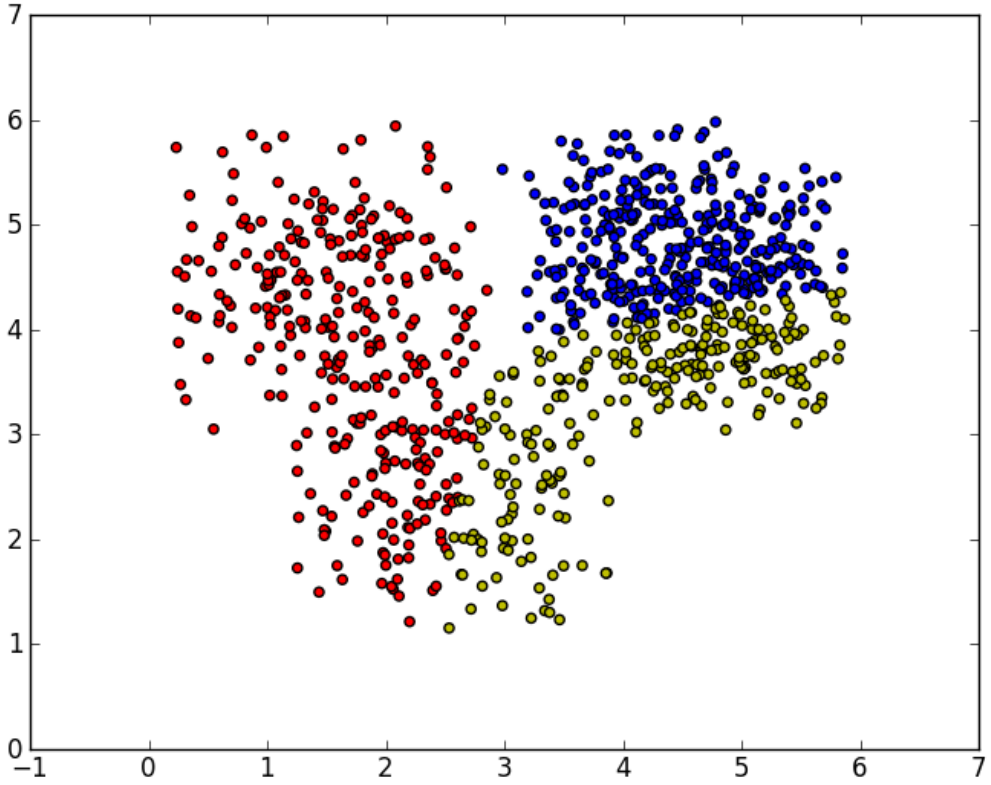
- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - Ex. 1. Classify countries based on (climate)
    - Ex. 2. Classify cars based on (gas mileage)
  - Predict some unknown class 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, …



Data Mining Functions: (4) Cluster Analysis

แบ่งกลุ่ม, จัดกลุ่ม

- 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



How the data suppose to look like

Ex. ข้อมูลที่จะใช้

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	นครินทร์ หัตถกรรม	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	วัดโลกยสุธาราม	4	0	Phra Nakhon Si Ayutthaya	13000	w4x98jk	0	4.000000	NaN	...	NaN
3	6	นันท์คาราโอเกะ	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