



CS 412 Intro. to Data Mining

Chapter 1. Introduction

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





Data and Information Systems (DAIS)

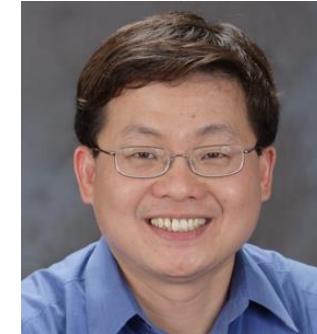
- Database Systems



Jiawei Han



Aditya
Parameswaran



Kevin Chang

- Data Mining



Hari
Sundaram

- Text Information Systems



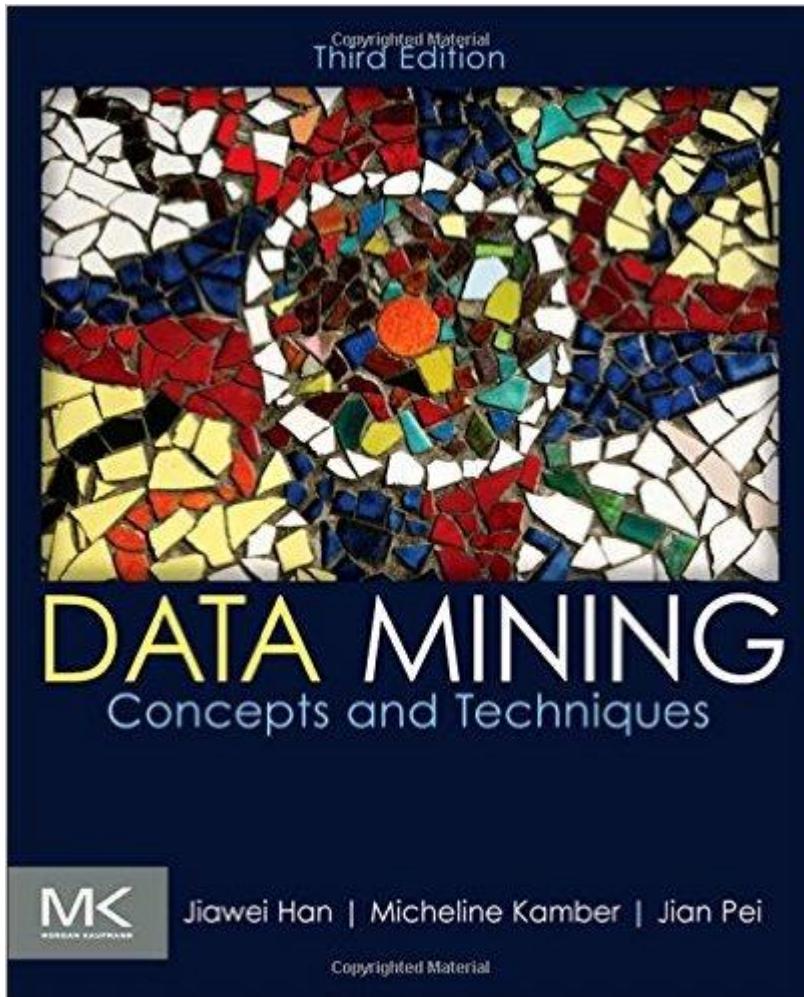
ChengXiang
Zhai

- Networks

Data and Information Systems (DAIS:) Course Structures at CS/UIUC

- Coverage: Database, data mining, text information systems, Web and bioinformatics
- Data mining
 - Intro. to data warehousing and mining (CS412)
 - Data mining: Principles and algorithms (CS512)
- Database Systems:
 - Intro. to database systems (CS411)
 - Advanced database systems (CS511)
- Text information systems
 - Text information system (CS410)
 - Advanced text information systems (CS510)

CS 412. Course Page & Class Schedule

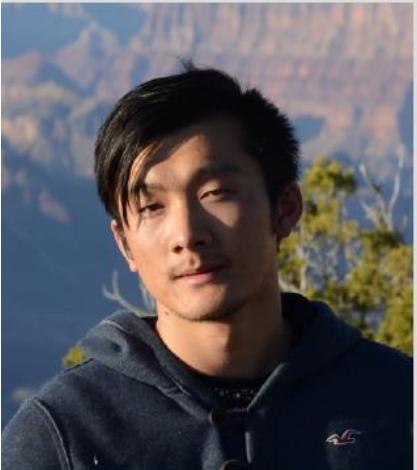


- Textbook
 - **Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques* (3rd ed), Morgan Kaufmann, 2011**
- Class Homepage:
<https://wiki.engr.illinois.edu/display/cs412>
- Bookmark on course schedule page
- **Class Schedule: 9:30-10:45 am Tues./Thurs. @1404 SC**
- Office hours: 10:45-11:30am Tues./Thurs. @2132 SC
- Lecture media: recorded; but class attendance is critical

CS 412. Fall 2017. Teach Assistants

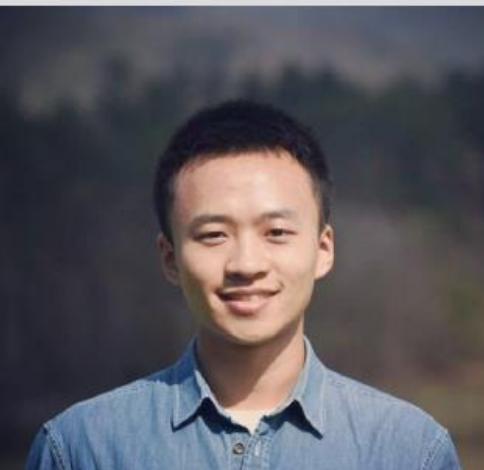


Dongming Lei



Carl Yang

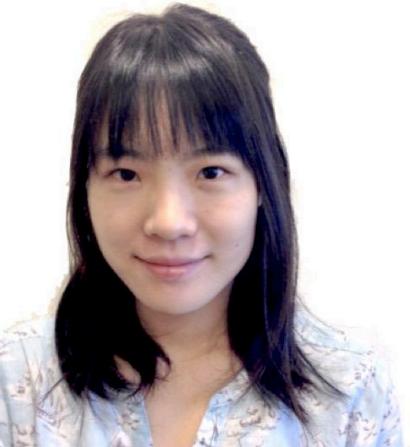
(Online Session)



Yu Shi



Chao Zhang



Shi Zhi

- TA office hours: **4-5pm (Mon.), 11-12pm (Wed.)@0207SC**. Additional hours before due date will be announced at Piazza
- Wait list (No wait list at this time, keep attending class, see if there is space available or there is overflow section opening)
- If you cannot register but still desperately want to get in, please sign on when there is “potential opening”: Explain why you have to take the course This Fall!

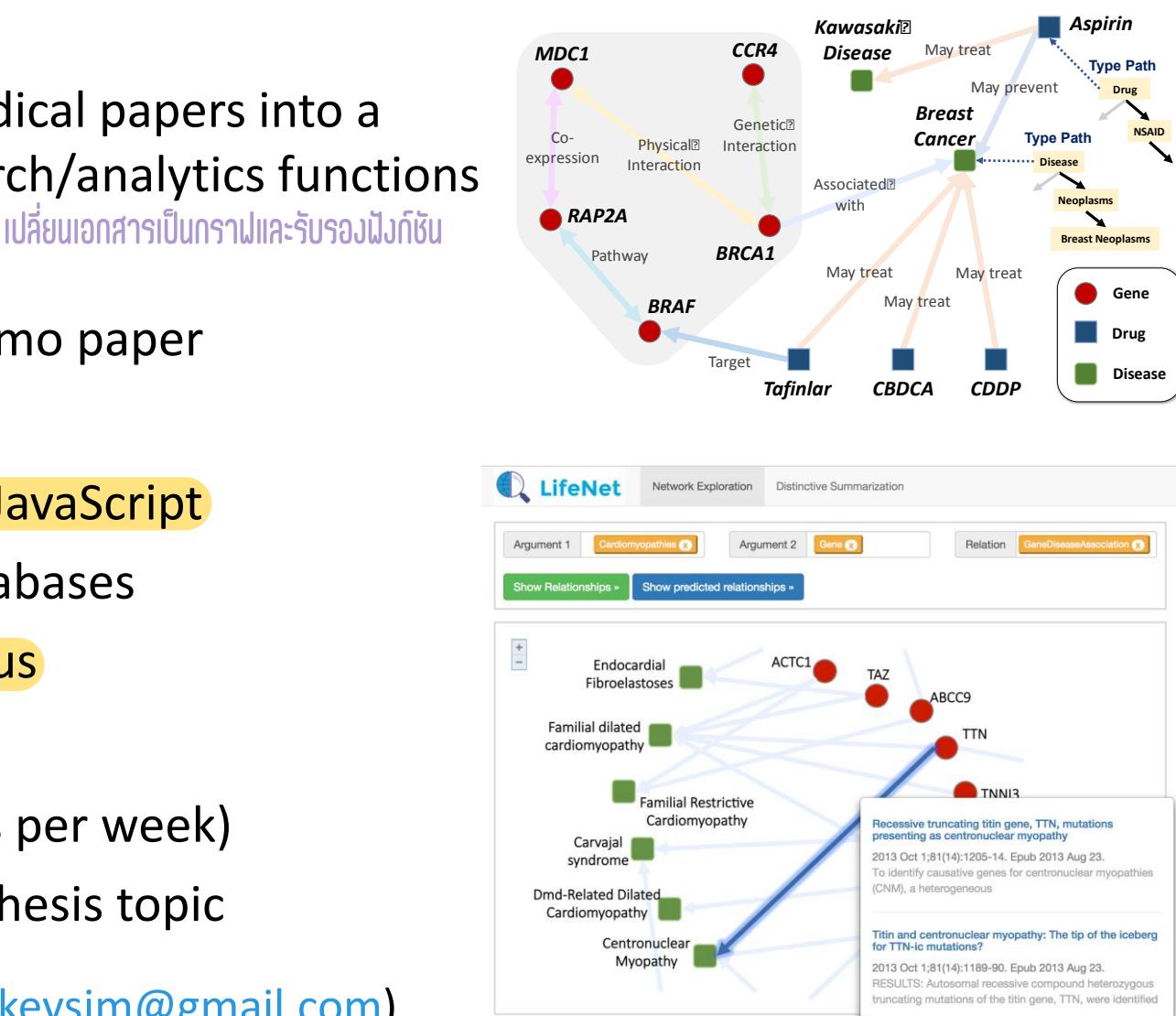
CS 412. Course Work and Grading

- ❑ Assignments, Programming Assignments, and Exams
 - ❑ Written Assignments: 15% (three homework assignments expected)
 - ❑ Programming assignments: 20% (two programming assignments expected)
 - ❑ Midterm exam: 30%
 - ❑ Final exam: 35%
- ❑ For students taking 4th credit (TA will provide concrete instructions on the 4th credit project)
 - ❑ For students registering 4 credits: 25%. The overall scores will be scaled proportionally
- ❑ Need help and/or discussions?
 - ❑ Sign on: Piazza (<https://piazza.com/illinois/cs412>)
- ❑ Check your homework/exam scores:
 - ❑ Compass

Help Needed: LifeNet—A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences

- What we are doing?
 - A scalable system that transforms biomedical papers into a knowledge graph & supports various search/analytics functions
- What we already have? สิ่งที่มีอยู่
 - A working prototype system & an ACL demo paper
- What we are looking for?
 - Students with expertise on **HTML/CSS & JavaScript**
 - Experiences on **web frameworks** and databases
 - System design experience will be a **big plus**
- What you will gain?
 - Hourly pay (\$12-\$15 per hour, 6-20 hours per week)
 - Possible research publications & a good thesis topic

Send us your resume if interested: Jiaming Shen (mickeysjm@gmail.com)



Chapter 1. Introduction

- Why Data Mining?  ทำไมต้องทำเหมืองข้อมูล
- What Is Data Mining? Data Mining คืออะไร
- A Multi-Dimensional View of Data Mining มุมมองหลายมิติของการทำเหมืองข้อมูล
- What Kinds of Data Can Be Mined? ข้อมูลประเภทใดที่สามารถขุดได้
- What Kinds of Patterns Can Be Mined? รูปแบบใดที่สามารถขุดได้
- What Kinds of Technologies Are Used? ใช้เทคโนโลยีประเภทใดบ้าง?
- What Kinds of Applications Are Targeted? แอปพลิเคชันประเภทใดที่กำหนดเป้าหมาย
- Major Issues in Data Mining ประเด็นสำคัญในการขุดข้อมูล
 ประวัติอย่างย่อ
- A Brief History of Data Mining and Data Mining Society
- Summary สรุป

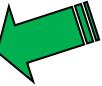
Why Data Mining?

Database เก็บข้อมูลอย่างไรให้มีประสิทธิภาพ Data Mining คือการต่อยอด

การเติบโตอย่างรวดเร็วของข้อมูล

- ❑ The Explosive Growth of Data: from terabytes to petabytes
 - ❑ Data collection and data availability การเก็บรวบรวมข้อมูลและความพร้อมใช้งานของข้อมูล
ระบบข้อมูลอัตโนมัติ
 - ❑ Automated data collection tools, database systems, Web, computerized society
 - ❑ Major sources of abundant data แหล่งข้อมูลที่สำคัญมากmany
 - ❑ Business: Web, e-commerce, transactions, stocks, ...
ธุรกิจ หุ้น
 - ❑ Science: Remote sensing, bioinformatics, scientific simulation, ...
การสำรวจทางอากาศ การจำลองวิทยาศาสตร์
 - ❑ Society and everyone: news, digital cameras, YouTube
กล้องดิจิตอล
 - ❑ We are drowning in data, but starving for knowledge! เรากำลังจมอยู่กับข้อมูล แต่หิวขาดความรู้
 - ❑ “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets
ความจำเป็นเป็นต้นกำเนิดของการประดิษฐ์ บุกข้อมูล การวิเคราะห์ข้อมูลขนาดใหญ่โดยอัตโนมัติ

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?  ឧប្បរគិតអេមីនុងខាងមុក
- 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

What Is Data Mining?

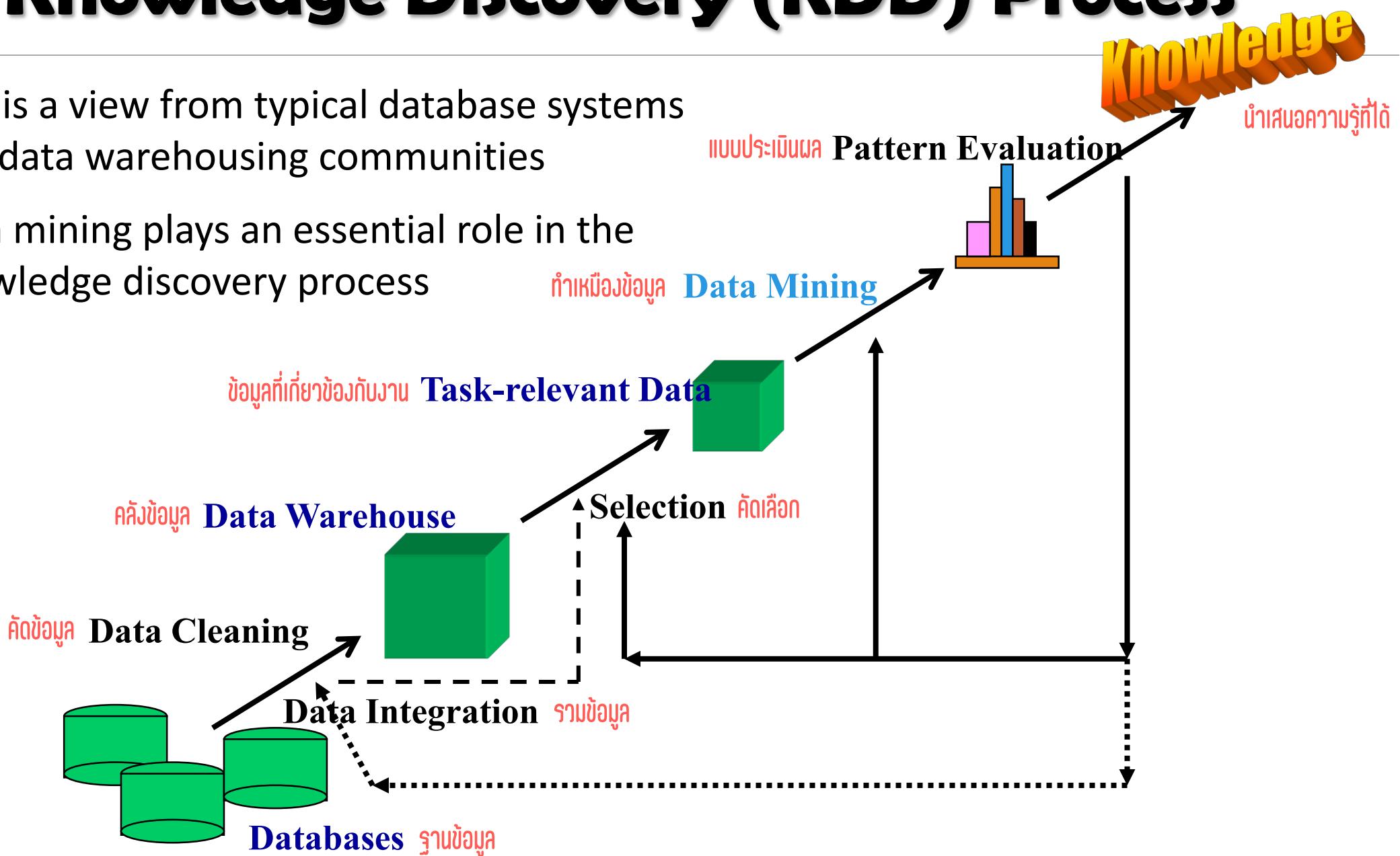


- Data mining (knowledge discovery from data)
การค้นพบความรู้จากข้อมูล
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Data mining: a misnomer?
- Alternative names
ชื่อแรกก่อนที่จะเป็น Data Mining
 - Knowledge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



Knowledge Discovery (KDD) Process

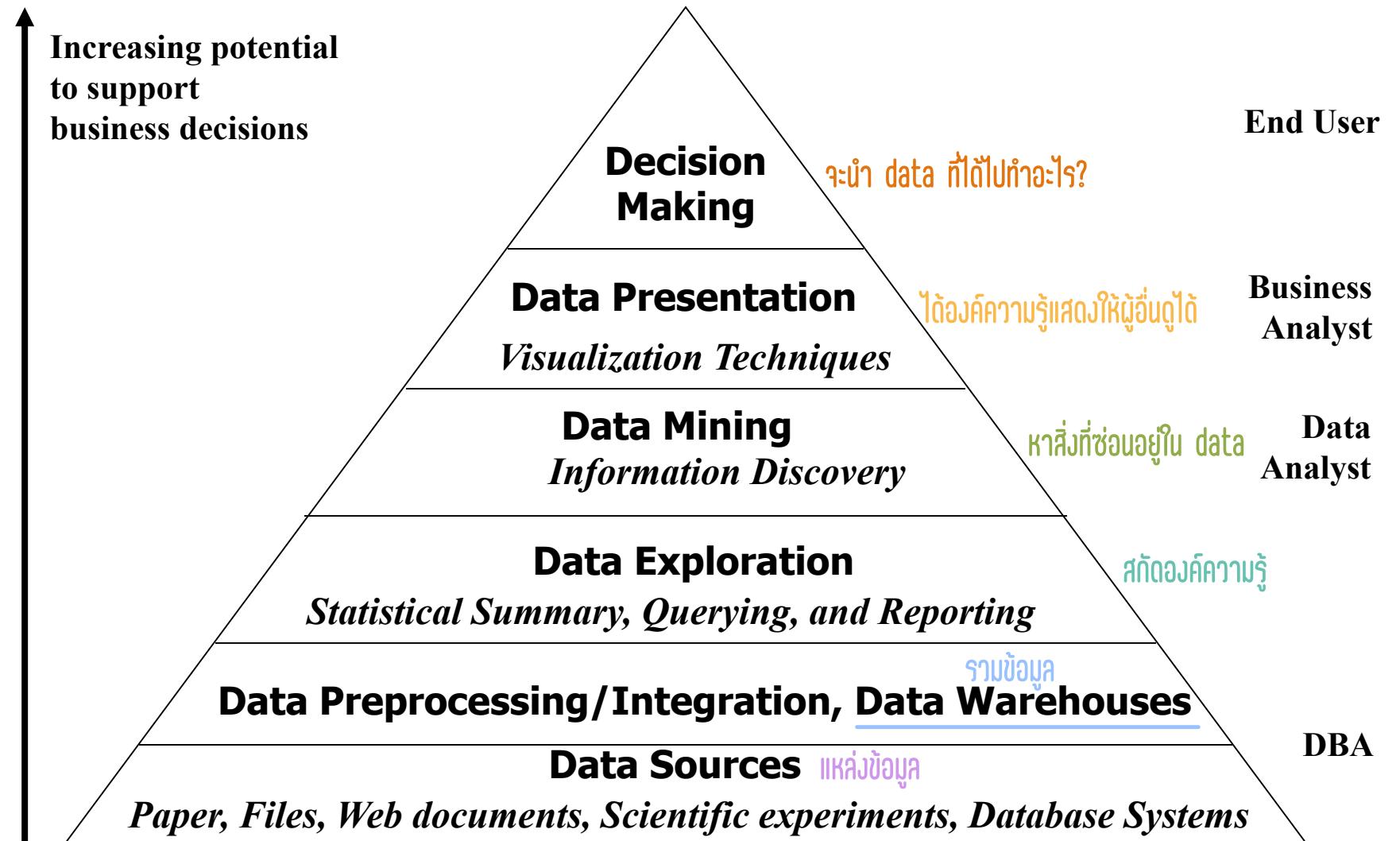
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



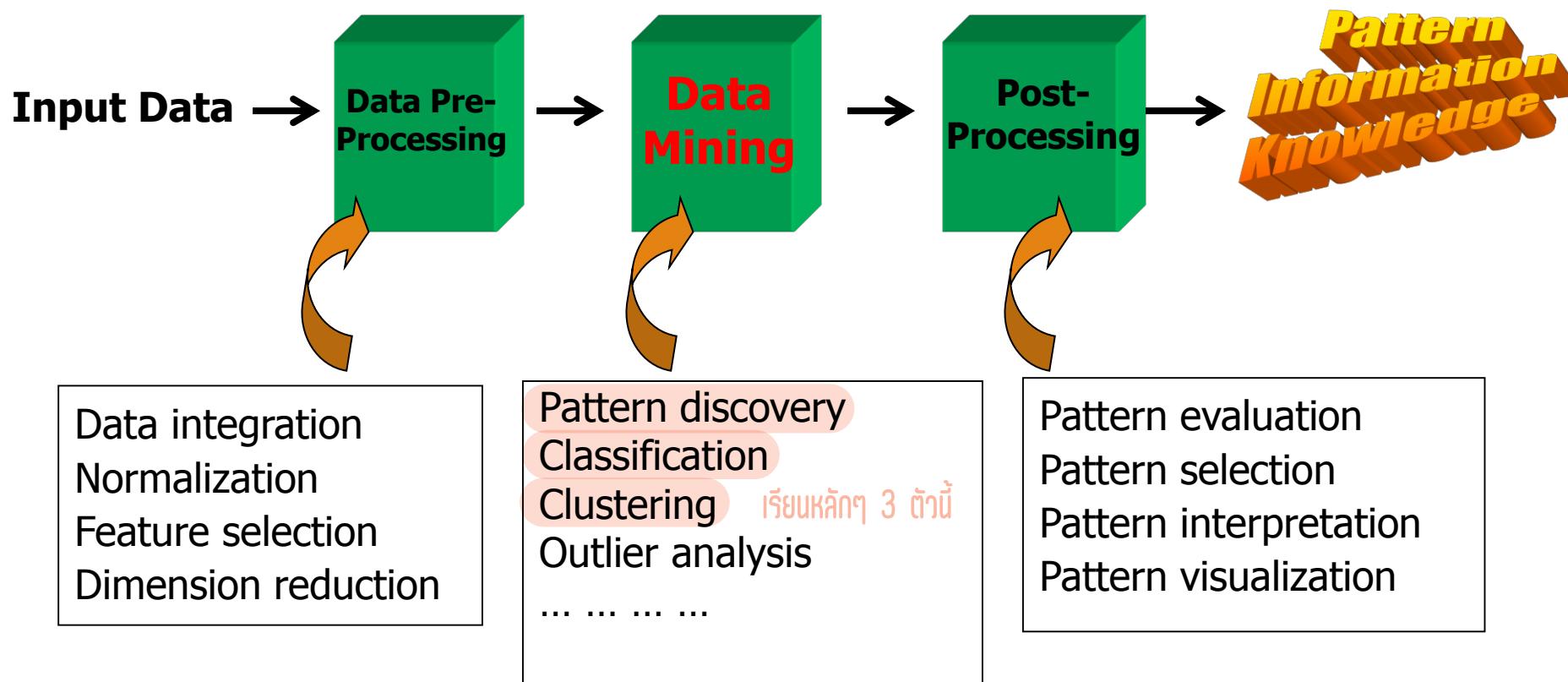
Example: A Web Mining Framework

- Web mining usually involves เริ่มจากการเก็บข้อมูล แล้วต่อด้วยกระบวนการด้านล่าง
- Data cleaning คัดข้อมูลที่ไม่เกี่ยวข้องออก
- Data integration from multiple sources รวมข้อมูลจากหลายแหล่ง
- Warehousing the data จัดเก็บข้อมูล
- Data cube construction
- Data selection for data mining วิเคราะห์ข้อมูลสำหรับ data mining
- Data mining การบุกเข้ามองข้อมูล / ค้นหาข้อมูลที่เป็นประโยชน์
- **Presentation of the mining results** ต้องสามารถแสดงองค์ความรู้ให้ผู้อื่นเข้าใจได้
- Patterns and knowledge to be used or stored into knowledge-base รูปแบบและความรู้ที่จะใช้หรือเก็บไว้ในฐานความรู้

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 vs. Data Exploration

การสำรวจข้อมูล

- ❑ Which view do you prefer?
 - ❑ KDD vs. ML/Stat. vs. Business Intelligence
 - ❑ Depending on the data, applications, and your focus

- ❑ Data Mining vs. Data Exploration
 - ❑ Business intelligence view
 - ❑ Warehouse, data cube, reporting but not much mining
 - ❑ Business objects vs. data mining tools
 - ❑ Supply chain example: mining vs. OLAP vs. presentation tools
 - ❑ Data presentation vs. data exploration

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



Multi-Dimensional View of Data Mining

ມຸມມອງຫຄາຍມີຕີ

- ❑ **Data to be mined** ข้อมูลที่ขุด
 - ❑ Database data (extended-relational, object-oriented, heterogeneous), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
 - ❑ **Knowledge to be mined (or: Data mining functions)** ความรู้ที่จะขุด
 - ❑ การกำหนดลักษณะ เสือกปฏิบัติ เชื่อมโยง การจำแนกประเภท จัดกลุ่ม แนวความคิด / การเบี่ยงเบน การวิเคราะห์ค่าผิดปกติ Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, ...
 - ❑ Descriptive vs. predictive data mining
 - ❑ Multiple/integrated functions and mining at multiple levels
 - ❑ **Techniques utilized**
 - ❑ Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
 - ❑ **Applications adapted** การดัดแปลง
 - ❑ Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?  ឯកសារមូលដ្ឋាន
- 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: On What Kinds of Data?

- Database-oriented data sets and applications ชุดข้อมูลและแอปพลิเคชันเชิงฐานข้อมูล
 - Relational database, data warehouse, transactional database
 - Object-relational databases, Heterogeneous databases and legacy databases

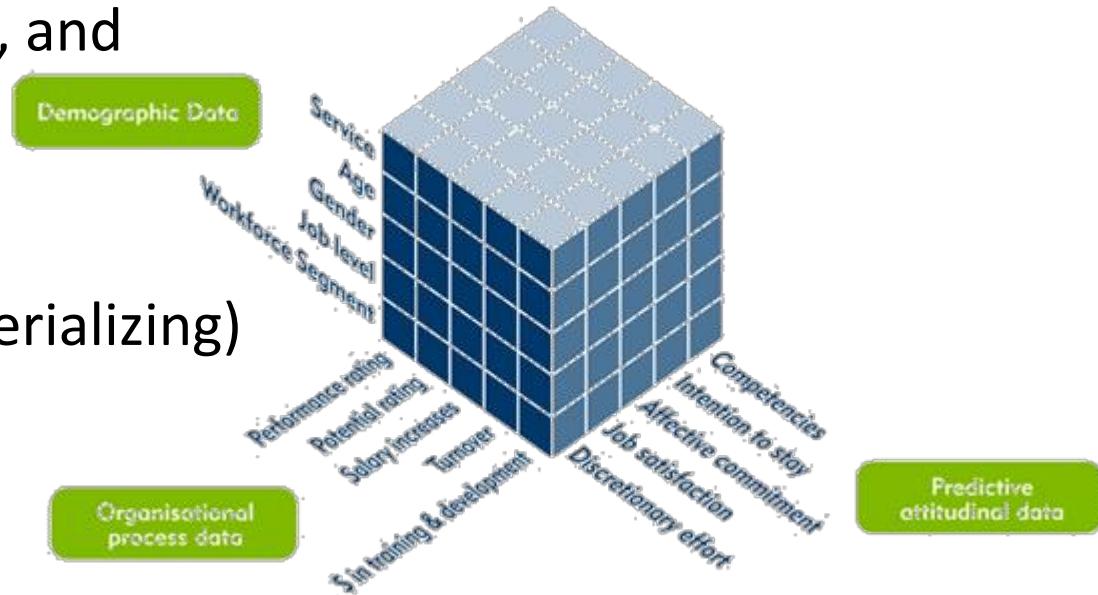
- Advanced data sets and advanced applications ชุดข้อมูลขั้นสูงและแอปพลิเคชันขั้นสูง
 - Data streams and sensor data
 - Time-series data, temporal data, sequence data (incl. bio-sequences)
 - Structure data, graphs, social networks and information networks
 - Spatial data and spatiotemporal data
 - Multimedia database
 - Text databases
 - The World-Wide Web

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 Functions: (1) Generalization

- ❑ Information integration and data warehouse construction การรวมและสร้างคลังข้อมูล
 - ❑ Data cleaning, transformation, integration, and multidimensional data model แปลง รวม
- ❑ Data cube technology ปรับขนาดได้ตามการคำนวณ
 - ❑ Scalable methods for computing (i.e., materializing) multidimensional aggregates
 - ❑ OLAP (online analytical processing)
- ❑ Multidimensional concept description: Characterization and discrimination กำหนด特徵และเฉพาะและการเลือกปฏิบัติ
 - ❑ Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region ประยุบเทียบคุณภาพข้อมูล



How the data suppose to look like

Columns: Attributes, Fields, Features: ค่าที่ใช้อธิบายคุณสมบัติของข้อมูล
ผักดอง แบบโลหะ (ก่อกร)

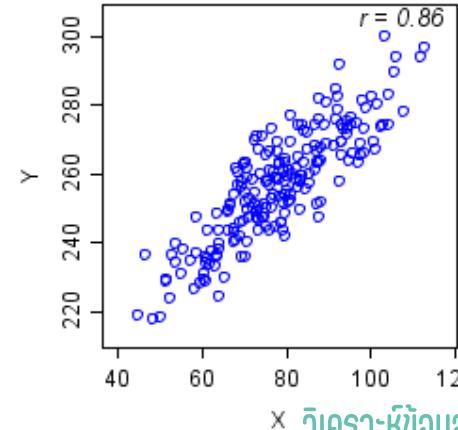
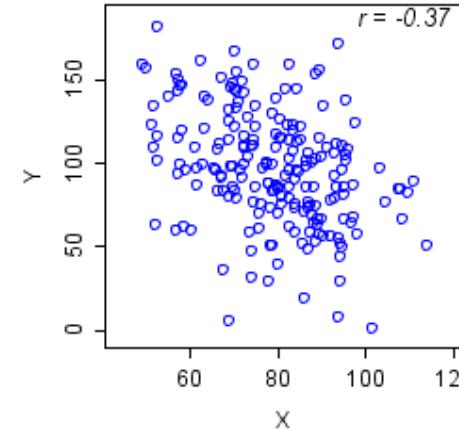
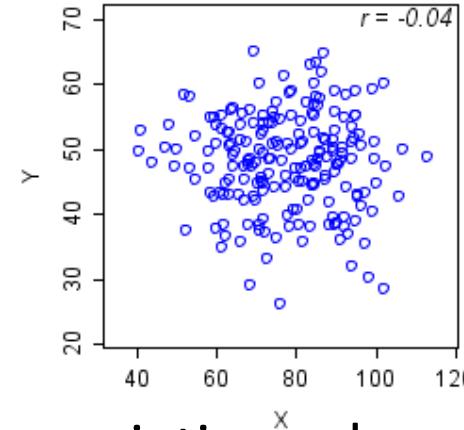
	id	name	domain_id	closed	city_name	zipcode	geohash	new_open	weighted_average_rating	number_of_chains	...	good_for_groups
0	2	นครินทร์ท่าน恭敬ราม	2	0	Samut Songkhram	75000	w4rh7g3	0	5.000000	NaN	...	NaN
1	4	Corner House	1	0	Bangkok Metropolitan Region	12150	w4rx73h	0	2.000000	NaN	...	NaN
2	5	วัดโคกน้ำสุราษฎร์	4	0	Phra Nakhon Si Ayutthaya	13000	w4x98jk	0	4.000000	NaN	...	NaN
3	6	นันท์คาرافอร์เกช	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

Data Mining Functions: (2) Pattern Discovery

รูปแบบที่ใช้บ่อย

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



- กฎการเชื่อมโยง
- 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?

วิเคราะห์ข้อมูลจากใบเสร็จ

- คนที่ซื้อผ้าอ้อมมักจะซื้อยีนส์ด้วย

แนวทางที่ร้านจะปรับใช้

1. ตั้งเปย์ร์กับผ้าอ้อมห่างกัน เพราะผู้ซื้อจะได้เลือกสินค้าอื่นด้วย
2. ตั้งเปย์ร์กับผ้าอ้อมติดกัน เพื่ออำนวยความสะดวกให้ผู้ซื้อ

Data Mining Functions: (3) Classification

□ Classification and label prediction

สร้างแบบจำลอง

- Construct models (functions) based on some training examples

อธิบาย

แยกแยะ

แนวคิด

ทำนายอนาคต

- Describe and distinguish classes or concepts for future prediction

- Ex. 1. Classify countries based on (climate) จำแนกประเทศตามภูมิอากาศ

- Ex. 2. Classify cars based on (gas mileage)

- Predict some unknown class labels

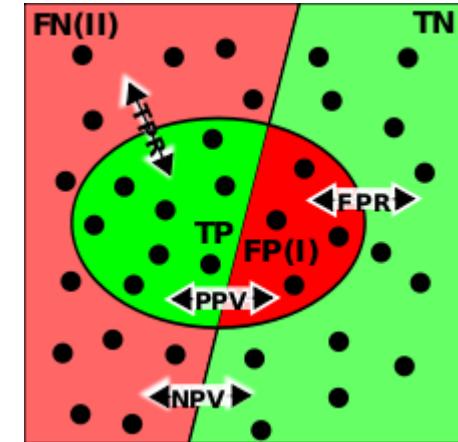
□ Typical methods

- Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...

□ Typical applications:

การตรวจจับการฉ้อโกงบัตรเครดิต

- Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...



Data Mining Functions: (4) Cluster Analysis

การวิเคราะห์แบบกลุ่ม

การเรียนรู้แบบไม่มีผู้ดูแล (เช่น ไม่รู้จักป้ายกำกับชั้นเรียน)

- Unsupervised learning (i.e., Class label is unknown)

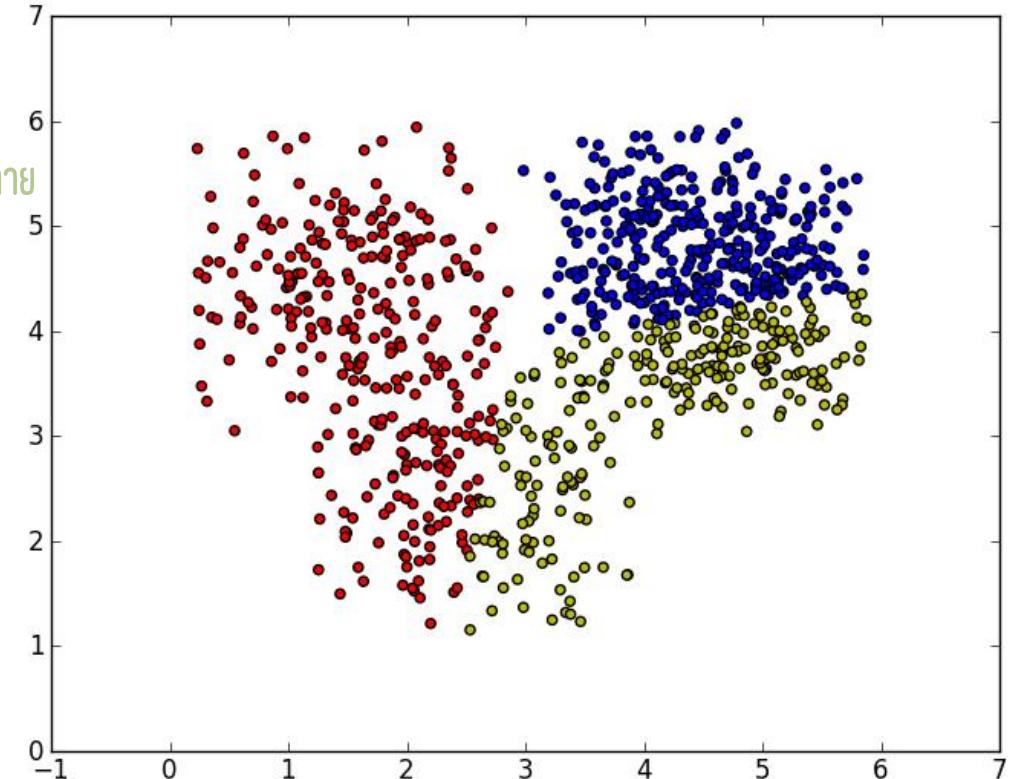
จัดกลุ่มข้อมูลเพื่อสร้างหมวดหมู่ใหม่ (เช่น คลัสเตอร์) เช่น บ้านคลัสเตอร์เพื่อค้นหารูปแบบการกระจาย

- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns

เพิ่มความคล้ายคลึงกันภายในคลาสให้มากที่สุด & ลดความคล้ายคลึงระหว่างคลาสให้น้อยที่สุด

- Principle: Maximizing intra-class similarity & minimizing interclass similarity

- Many methods and applications



Data Mining Functions: (5) Outlier Analysis

การวิเคราะห์ค่าผิดปกติ

□ Outlier analysis

วัตถุข้อมูลที่ไม่สอดคล้องกับพฤติกรรมทั่วไปของข้อมูล

- Outlier: A data object that does not comply with the general behavior of the data

เสียงหรือข้อยกเว้น? ขยะของคนหนึ่งอาจเป็นสมบัติของอีกคนหนึ่งได้

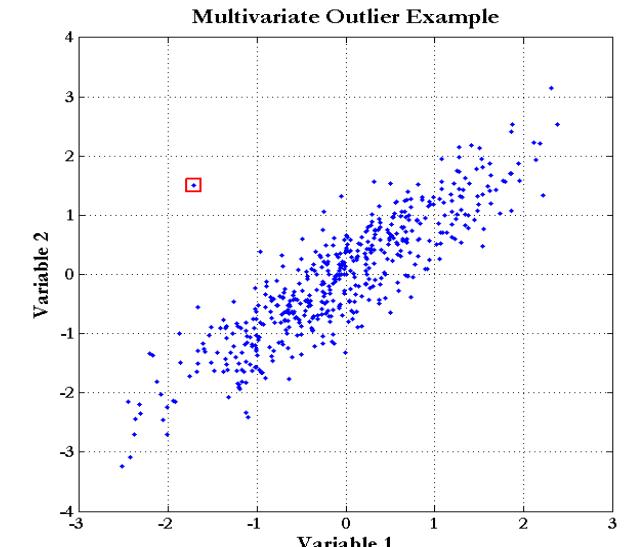
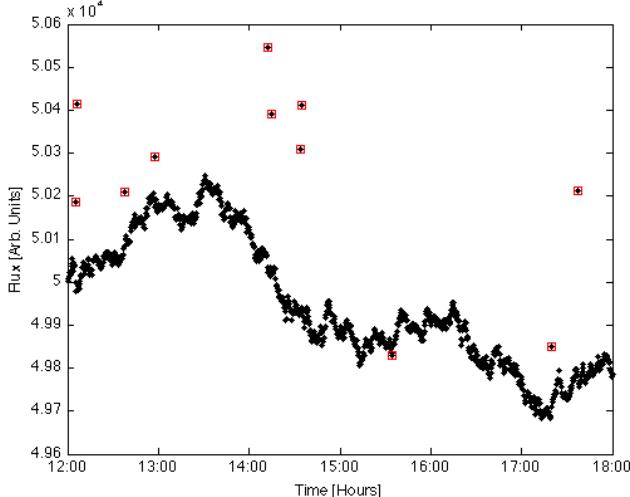
- Noise or exception?—One person's garbage could be another person's treasure

วิธีการ: โดยผลคุณของการจัดกลุ่มหรือการวิเคราะห์การถดถอย

- Methods: by product of clustering or regression analysis, ...

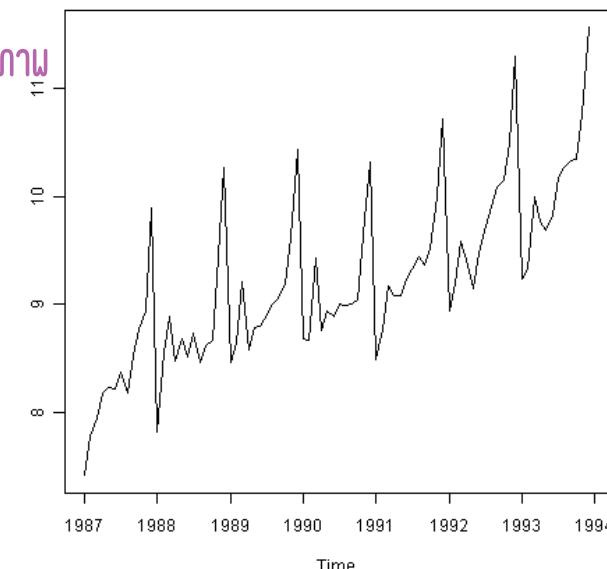
มีประโยชน์ในการตรวจสอบการจ้อโกง การวิเคราะห์เหตุการณ์หายาก

- Useful in fraud detection, rare events analysis



Data Mining Functions: (6) Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

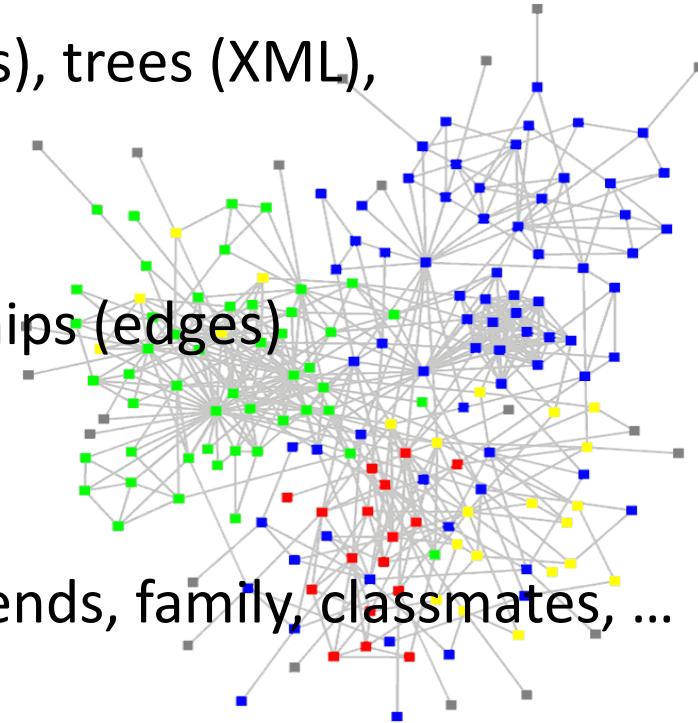
- Sequence, trend and evolution analysis การวิเคราะห์ลำดับ แนวโน้ม และวิวัฒนาการ
 - Trend, time-series, and deviation analysis
 - e.g., regression and value prediction
 - Sequential pattern mining การขุดรูปแบบตามลำดับ
 - e.g., buy digital camera, then buy large memory cards
 - Periodicity analysis การวิเคราะห์เป็นระยะ
 - Motifs and biological sequence analysis คาดคะเนและการวิเคราะห์ลำดับทางชีวภาพ
 - Approximate and consecutive motifs
 - Similarity-based analysis การวิเคราะห์ตามความคล้ายคลึงกัน
- Mining data streams ประพันตามเวลา
 - Ordered, time-varying, potentially infinite, data streams



Data Mining Functions: (7) Structure and Network Analysis

การวิเคราะห์โครงสร้างและเครือข่าย

- Graph mining
 - การค้นหากราฟย่อยที่ใช้บ่อย
- Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis การวิเคราะห์เครือข่ายข้อมูล
 - Social networks: actors (objects, nodes) and relationships (edges)
 - e.g., author networks in CS, terrorist networks
 - Multiple heterogeneous networks เครือข่ายที่ต่างกันมากmany
 - A person could be multiple information networks: friends, family, classmates, ...
 - Links carry a lot of semantic information: Link mining
- Web mining
 - Web is a big information network: from PageRank to Google
 - Analysis of Web information networks การวิเคราะห์เครือข่ายข้อมูลเว็บ
 - การค้นพบชุมชนบนเว็บ
 - Web community discovery, opinion mining, usage mining, ...



Evaluation of Knowledge

การประเมินความรู้

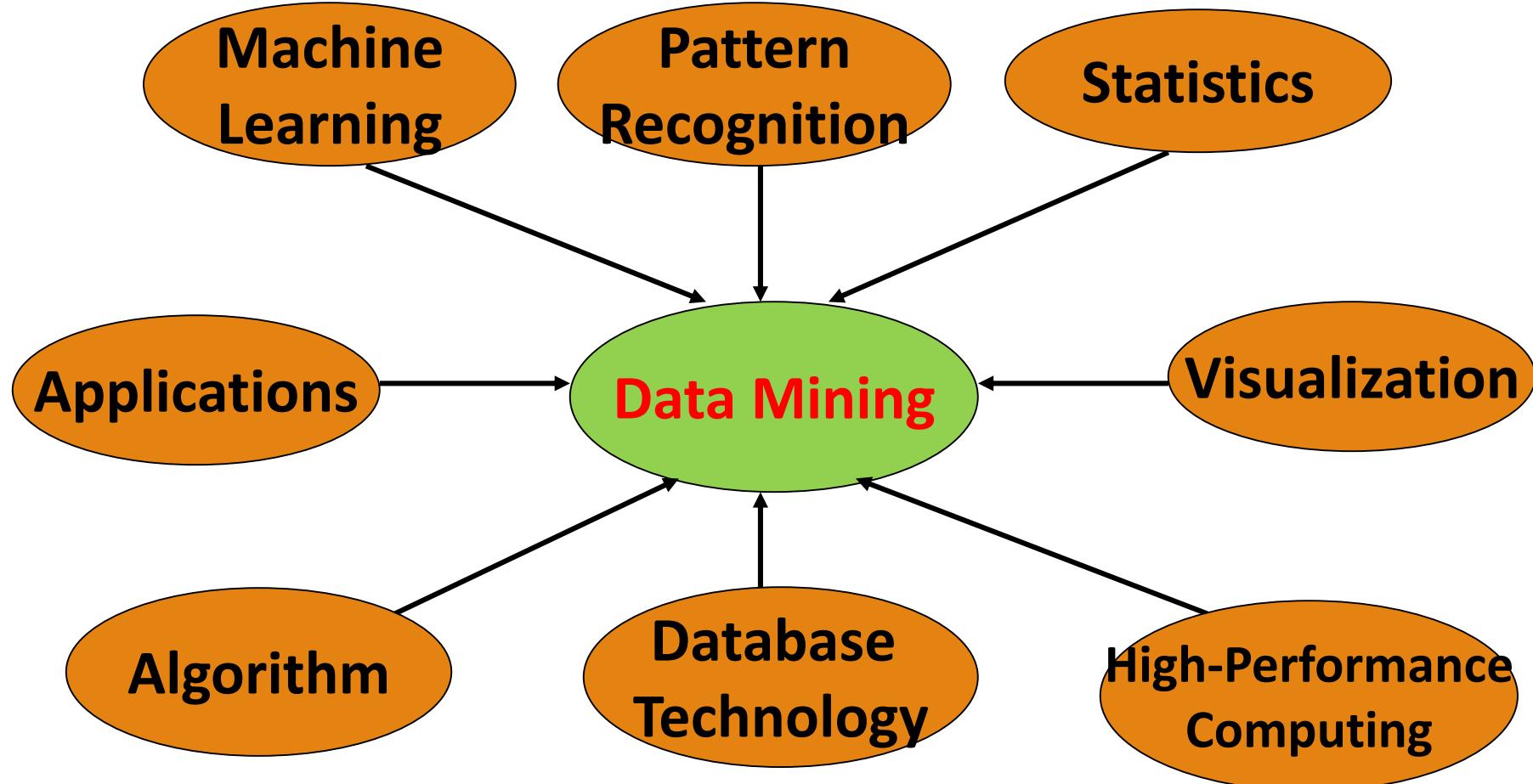
- Are all mined knowledge interesting? ความรู้ที่ขุดทั้งหมดน่าสนใจหรือไม่?
 - One can mine tremendous amount of “patterns” เราสามารถขุดรูปแบบได้จำนวนมหาศาล
 - Some may fit only certain dimension space (time, location, ...) บางส่วนอาจพอดีกับพื้นที่มิติบางส่วนเท่านั้น
 - Some may not be representative, may be transient, ...
ตัวแทน
บุคคลพิเศษที่โดดเด่น
 - Evaluation of mined knowledge → directly mine only interesting knowledge?
 - Descriptive vs. predictive
คำอธิบาย
คำทำนาย
 - Coverage ความคุ้มครอง
ความครอบคลุม
 - Typicality vs. novelty
ความนิยม
ความแปลกใหม่
 - Accuracy ความแม่นยำ
 - Timeliness ทันเวลา
 - ...



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: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

บรรจบ

- Tremendous amount of data ปริมาณข้อมูลมหาศาล
 - Algorithms must be scalable to handle big data อัลกอริทึมต้องปรับขนาดได้เพื่อจัดการกับข้อมูลขนาดใหญ่
- High-dimensionality of data
 - Micro-array may have tens of thousands of dimensions ชิบซ่อน
- High complexity of data
 - Data streams and sensor data
 - Time-series data, temporal data, sequence data
 - Structure data, graphs, social and information networks ข้อมูลเชิงผืนผ้า
 - Spatial, spatiotemporal, multimedia, text and Web data
 - Software programs, scientific simulations
 - New and sophisticated applications ชิบซ่อน

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

Applications of Data Mining

- Web page analysis: classification, clustering, ranking
การวิเคราะห์ร่วมกัน
- Collaborative analysis & recommender systems
เป้าหมาย
- Basket data analysis to targeted marketing
- Biological and medical data analysis
- Data mining and software engineering
- Data mining and text analysis
- Data mining and social and information network analysis
- Built-in (invisible data mining) functions in Google, MS, Yahoo!, Linked, Facebook, ...
- Major dedicated data mining systems/tools ระบบ/เครื่องมือการทำเหมืองข้อมูลเฉพาะที่สำคัญ
- SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)



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



Major Issues in Data Mining (1)

ประเด็นหลักในการขุดข้อมูล

วิธีการ

- Mining Methodology
 - Mining various and new kinds of knowledge ขุดหาความรู้ซึ่งนิดใหม่ๆ
 - Mining knowledge in multi-dimensional space ขุดความรู้ในมิติที่หลายมิติ
 - Data mining: An interdisciplinary effort
 - Boosting the power of discovery in a networked environment เพิ่มพลังแห่งการค้นพบในสภาพแวดล้อมแบบเครือข่าย
 - การจัดการเสียง
 - ความไม่แน่นอน
 - ความไม่สมบูรณ์ของข้อมูล
 - Handling noise, uncertainty, and incompleteness of data
 - Pattern evaluation and pattern- or constraint-guided mining การประเมินรูปแบบและการขุดตามรูปแบบหรือข้อจำกัด
 - User Interaction การตั้งตอบกับผู้ใช้
 - Interactive mining เชิงโต้ตอบ
 - Incorporation of background knowledge การรวมความรู้พื้นฐาน
 - Presentation and visualization of data mining results การนำเสนอและการแสดงผลการขุดข้อมูล

Major Issues in Data Mining (2)

- Efficiency and Scalability ประสิทธิภาพและความสามารถ
 - Efficiency and scalability of data mining algorithms
ขบวน กระฉ�ย พัฒนา
 - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types ความหลากหลายของประเภทข้อมูล
 - Handling complex types of data การจัดการข้อมูลที่ซับซ้อน
 - Mining dynamic, networked, and global data repositories
- Data mining and society การทำเหมืองข้อมูลและสังคม
 - Social impacts of data mining
ผลกระทบทางสังคม
 - Privacy-preserving data mining
การรักษาความเป็นส่วนตัว
 - Invisible data mining
ไม่ปรากฏ

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



A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
 - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
 - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
 - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
 - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

Conferences and Journals on Data Mining

- ❑ KDD Conferences
 - ❑ ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining ([KDD](#))
 - ❑ SIAM Data Mining Conf. ([SDM](#))
 - ❑ (IEEE) Int. Conf. on Data Mining ([ICDM](#))
 - ❑ European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining ([ECML-PKDD](#))
 - ❑ Pacific-Asia Conf. on Knowledge Discovery and Data Mining ([PAKDD](#))
 - ❑ Int. Conf. on Web Search and Data Mining ([WSDM](#))
- Other related conferences
 - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
 - Web and IR conferences: WWW, SIGIR, WSDM
 - ML conferences: ICML, NIPS
 - PR conferences: CVPR,
- Journals
 - Data Mining and Knowledge Discovery (DAMI or DMKD)
 - IEEE Trans. On Knowledge and Data Eng. (TKDE)
 - KDD Explorations
 - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD)
 - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD)
 - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
 - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
 - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
 - Conferences: SIGIR, WWW, CIKM, etc.
 - Journals: WWW: Internet and Web Information Systems,
- Statistics
 - Conferences: Joint Stat. Meeting, etc.
 - Journals: Annals of statistics, etc.
- Visualization
 - Conference proceedings: CHI, ACM-SIGGraph, etc.
 - Journals: IEEE Trans. visualization and computer graphics, etc.

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



Summary

- ❑ Data mining: Discovering interesting patterns and knowledge from massive amount of data การขุดข้อมูล: ค้นพบรูปแบบและความรู้ที่น่าสนใจจากข้อมูลจำนวนมาก
 - ❑ A natural evolution of science and information technology, in great demand, with wide applications วิวัฒนาการตามธรรมชาติของวิทยาศาสตร์และเทคโนโลยีสารสนเทศที่มีความต้องการสูงพร้อมการใช้งานที่หลากหลาย
 - ❑ A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
 - ❑ Mining can be performed in a variety of data การขุดสามารถทำได้ด้วยข้อมูลที่หลากหลาย
 - ❑ Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc. คัดแยก แยกแยะ
การจำแนกประเภท การจัดกลุ่ม
 - ❑ Data mining technologies and applications เทคโนโลยีและแอปพลิเคชันการทำเหมืองข้อมูล
 - ❑ Major issues in data mining ประเด็นสำคัญในการการทำเหมืองข้อมูล

Recommended Reference Books

- Charu C. Aggarwal, Data Mining: The Textbook, Springer, 2015
- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005 (2nd ed. 2016)
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005
- Mohammed J. Zaki and Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms 2014

