



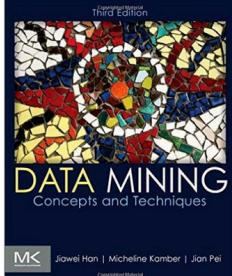
CS 412 Intro. to Data Mining

Chapter 1. Introduction

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



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



Jiawei Han

CS 412. Course Work and Grading

Score

- Midterm (data preprocessing ปฏิบัติ (เดี่ยว)) 25%
 - Final(ทฤษฎี data mining เดี่ยว) 25%
 - Project (data preprocessing + data mining (จัดกลุ่มเอง 5-6 คน)) 20%
 - Homework (แบ่งกลุ่มใหม่ทุกครั้ง) 15%
 - Quiz (เดี่ยว สามในห้อง) 10%
 - GitHub 5%
- Final Score = Score * %attendance

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

5

Why Data Mining?

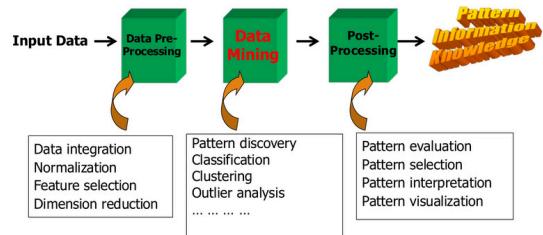
- The Explosive Growth of Data: from terabytes to petabytes
 - Data collection and data availability
 - Automated data collection tools, database systems, Web, computerized society
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, 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

6

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

KDD Process: A View from ML and Statistics



This is a view from typical machine learning and statistics communities

12

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



8

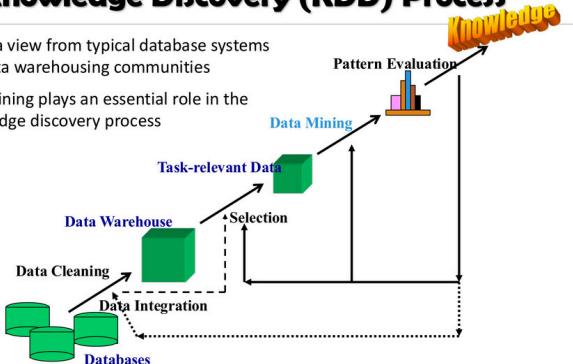
Data Mining vs. Data Exploration

- 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

13

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



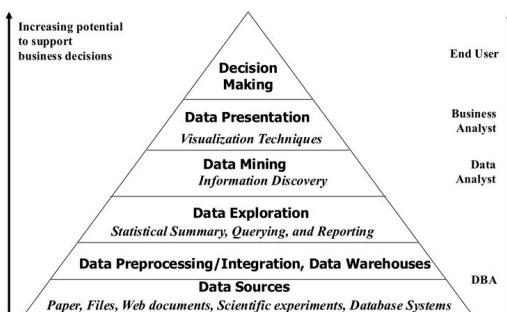
9

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

10

Data Mining in Business Intelligence



11

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

14

Multi-Dimensional View of Data Mining

- ❑ **Data to be mined**
 - ❑ 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)**
 - ❑ 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**
 - ❑ Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

15

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

16

Data Mining: On What Kinds of Data?

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

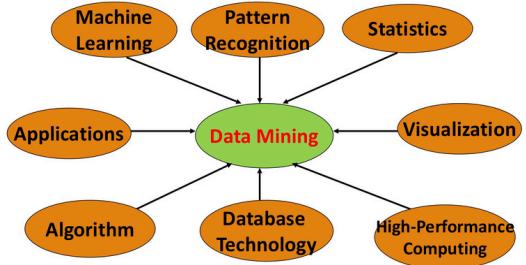
17

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

27

Data Mining: Confluence of Multiple Disciplines



28

Why Confluence of Multiple Disciplines?

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

29

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

30

Applications of Data Mining



- Web page analysis: classification, clustering, ranking
- Collaborative analysis & recommender systems
- 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
 - SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)

31

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

32

Major Issues in Data Mining (1)

- ❑ Mining Methodology
 - ❑ Mining various and new kinds of knowledge
 - ❑ Mining knowledge in multi-dimensional space
 - ❑ Data mining: An interdisciplinary effort
 - ❑ Boosting the power of discovery in a networked environment
 - ❑ Handling noise, uncertainty, and incompleteness of data
 - ❑ Pattern evaluation and pattern- or constraint-guided mining
- ❑ User Interaction
 - ❑ Interactive mining
 - ❑ Incorporation of background knowledge
 - ❑ Presentation and visualization of data mining results

33

Major Issues in Data Mining (2)

- ❑ Efficiency and Scalability
 - ❑ Efficiency and scalability of data mining algorithms
 - ❑ Parallel, distributed, stream, and incremental mining methods
- ❑ Diversity of data types
 - ❑ Handling complex types of data
 - ❑ Mining dynamic, networked, and global data repositories
- ❑ Data mining and society
 - ❑ Social impacts of data mining
 - ❑ Privacy-preserving data mining
 - ❑ Invisible data mining

34

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

35

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)

36

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 (**ECCML-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

37

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.

38

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 

39

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications
- 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
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

40

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](#)

41

สรุปบทที่ 1: บทนำสู่การทำเหมืองข้อมูล (Introduction to Data Mining)

1. Why Data Mining? (ทำไมต้องทำเหมืองข้อมูล?)

1.1 ยุคแห่งข้อมูลท่วมท้น (The Explosive Growth of Data)

ในปัจจุบัน โลกของเรามีกำลังเพิ่มขึ้นกับปริมาณข้อมูลมหาศาล (Big Data)

- ปริมาณข้อมูล (Volume): ข้อมูลเติบโตจากระดับ Terabytes ไปสู่ Petabytes และ Exabytes
- การรวบรวมข้อมูล (Data Collection): เครื่องมือจับเก็บข้อมูลมีอยู่ทุกที่และทำงานอัตโนมัติ (Automated Data Collection Tools) เช่น:
 - เว็บเพจและการคลิกของผู้ใช้งาน (Web clickstreams)
 - ธุรกรรมทางการเงินและการซื้อขาย (E-commerce transactions)
 - เซ็นเซอร์ต่างๆ (IoT, GPS, กล้องวงจรปิด)
 - ข้อมูลทางวิทยาศาสตร์ (ภาพถ่ายดาวเทียม, ข้อมูลจีโนม, กล้องโทรทรรศน์)
 - สังคมออนไลน์ (Social Media)

1.2 ปัญหาหลัก: "Drowning in Data, but Starving for Knowledge"

ประโยชน์คือห้าใจสำคัญของปัญหานี้ แปลว่า "เรากำลังจะมองข้อมูลตาย แต่กลับอดอยากความรู้"

- Rich Data but Information Poor: เรา้มีข้อมูลดีบกับมหาศาล แต่ขาดความรู้หรือสารสนเทศที่สรุปอุปกรณ์เพื่อใช้ประโยชน์ได้
- ข้อจำกัดของมนุษย์: มนุษย์ไม่สามารถอ่านหรือวิเคราะห์ข้อมูลปริมาณมหาศาลนี้ด้วยตาเปล่าได้ทัน
- ความจำเป็น (Necessity): เราจึงต้องการเครื่องมืออัตโนมัติ (Automated Analysis Tools) ที่จะมาช่วย "ขุด" (Mine) หาความรู้ที่ซ่อนอยู่อุปกรณ์

1.3 วิวัฒนาการของเทคโนโลยีฐานข้อมูล (Evolution of Database Technology)

Data Mining ไม่ได้เกิดขึ้นมาโดยๆ แต่เป็นผลพวงจากการวิวัฒนาการทางเทคโนโลยี:

- 1960s: การประมวลผลข้อมูลยุคแรก (Data Collection, Database Creation, IMS, Network DBMS)
- 1970s: ระบบจัดการฐานข้อมูลเชิงสัมพันธ์ (Relational DBMS - RDBMS) เริ่มมีการใช้ SQL
- 1980s: ระบบฐานข้อมูลขั้นสูง (Advanced RDBMS, Object-Oriented DB)
- 1990s: คลังข้อมูลและการทำเหมืองข้อมูล (Data Warehousing, Data Mining, OLAP) – เริ่มน้นการวิเคราะห์
- 2000s - ปัจจุบัน: การจัดการข้อมูลขั้นสูง (Stream, Web, Text, Cloud, Big Data, Deep Learning)

2. What Is Data Mining? (การทำเหมืองข้อมูลคืออะไร?)

2.1 นิยาม (Definition)

Data Mining คือกระบวนการ ค้นหา (Discovery) แพทเทิร์นที่น่าสนใจ (Interesting Patterns) และความรู้ (Knowledge) จากข้อมูลปริมาณมหาศาล

- ชื่ออื่นที่ใช้เรียก: Knowledge Discovery (Mining) in Databases (KDD), Knowledge Extraction, Data/Pattern Analysis, Data Archaeology, Business Intelligence
- ข้อควรระวัง: Data Mining ไม่ใช่แค่การค้นหาข้อมูล (Search/Query) แบบ Google หรือ SQL ปกติ แต่เป็นการทำ "ความสัมพันธ์ที่ซ่อนอยู่"

2.2 กระบวนการ KDD (Knowledge Discovery in Databases Process)

Data Mining เป็นเพียง "ขั้นตอนหนึ่ง" (แต่เป็นหัวใจสำคัญ) ในกระบวนการ KDD ทั้งหมด ซึ่งประกอบด้วย:

1. Data Cleaning: กำจัดข้อมูลเสียง (Noise) และข้อมูลที่ไม่สอดคล้องกัน
2. Data Integration: รวมข้อมูลจากหลายแหล่ง (เช่น Database + Web + CSV)
3. Data Selection: ดึงเฉพาะข้อมูลที่เกี่ยวข้องกับงานวิเคราะห์ออกมากจากฐานข้อมูล
4. Data Transformation: แปลงข้อมูลให้เหมาะสมกับการทำเหมือง (เช่น การทำ Normalization, สร้างฟีเจอร์ใหม่)
5. Data Mining (หัวใจสำคัญ): การใช้อัลกอริทึมอัจฉริยะเพื่อสกัดแพทเทิร์นออกมา
6. Pattern Evaluation: ประเมินว่าแพทเทิร์นที่ได้นั้น "น่าสนใจ" หรือไม่ (วัดจากความถูกต้อง, ประโยชน์, ความเปลี่ยนแปลง)
7. Knowledge Presentation: นำเสนอผลลัพธ์ให้ผู้ใช้เข้าใจ (Visualization)

2.3 Data Mining vs. Data Exploration

- Data Exploration (สำรวจข้อมูล): ใช้สถิติพื้นฐาน กราฟ เพื่อดูภาพรวม (Summary Statistics)
- Data Mining (เหมืองข้อมูล): ใช้โมเดลซับซ้อนเพื่อหาความสัมพันธ์ที่มองไม่เห็นด้วยตาเปล่า

3. A Multi-Dimensional View of Data Mining (มุมมองหลายมิติของการทำเหมืองข้อมูล)

การทำเหมืองข้อมูลสามารถจำแนกได้หลายมุมมอง (Multi-dimensional view):

1. Data to be Mined (ชนิดของข้อมูล): ฐานข้อมูล, คลังข้อมูล, ธุรกรรม, สถิติ, ความเชื่อมโยง, เว็บฯลฯ
2. Knowledge to be Mined (ชนิดของความรู้): การจัดกลุ่ม (Clustering), การจำแนก (Classification), กฎความสัมพันธ์ (Association), การวิเคราะห์แนวโน้ม (Trend)
3. Techniques Utilized (เทคนิคที่ใช้): เน้นฐานข้อมูล (Database-oriented), เน้นคลังข้อมูล (Data Warehouse - OLAP), เน้นเครื่องจักรเรียนรู้ (Machine Learning), เน้นสถิติ (Statistics), เน้นการมองเห็นภาพ (Visualization)
4. Applications Adapted (การประยุกต์ใช้): การค้าปลีก, โทรคมนาคม, การธนาคาร, การตรวจสอบการอัปโหลด, ตลาดหุ้น, การแพทย์

4. What Kinds of Data Can Be Mined? (ข้อมูลชนิดใดบ้างที่ทำเหมืองได้?)

4.1 ฐานข้อมูลทั่วไป (Database-oriented Data)

- Relational Database: ตารางข้อมูลที่มีความสัมพันธ์กัน (Table, Record, Attribute)
- Data Warehouse: คลังข้อมูลขนาดใหญ่ที่รวบรวมจากหลายแหล่ง ออกแบบมาเพื่อการวิเคราะห์โดยเฉพาะ (มักเก็บข้อมูลย้อนหลังหลายปี)
- Transactional Database: ข้อมูลการทำธุรกรรม เช่น ในเรียบสินค้า (Transaction ID + รายการสินค้าที่ซื้อ)

4.2 ข้อมูลขั้นสูง (Advanced Data Sets)

- Data Streams: ข้อมูลที่ไหลเข้ามาต่อเนื่องและรวดเร็ว (เช่น ข้อมูล Sensor, หุ่น, Network Traffic)
- Time-Series Data: ข้อมูลที่มีลำดับเวลาเข้ามาเกี่ยวข้อง (เช่น ราคาหุ้นรายวัน, อุณหภูมิรายชั่วโมง)
- Spatial Data: ข้อมูลเชิงพื้นที่ (เช่น แผนที่, GPS, ข้อมูลทางภูมิศาสตร์ GIS)
- Sequence Data: ข้อมูลลำดับ (เช่น ลำดับ DNA, ลำดับการเข้าชมเว็บไซต์)
- Text & Multimedia: ข้อความ, รูปภาพ, วิดีโอ, เสียง
- Graph & Social Networks: กราฟความสัมพันธ์, เพื่อนใน Facebook
- World Wide Web (Web Mining): ข้อมูลจากอินเทอร์เน็ต

5. What Kinds of Patterns Can Be Mined? (รูปแบบความรู้ชนิดใดที่ค้นหาได้?)

นี่คือ "ฟังก์ชันหลัก" ของ Data Mining ซึ่งแบ่งเป็น 2 กลุ่มใหญ่: Descriptive (เชิงพรรณนา - อธิบายสิ่งที่เป็นอยู่) และ Predictive (เชิงทำนาย - ทำนายอนาคต)

5.1 Class/Concept Description (การอธิบายคลาส/คอนเซปต์)

เป็นการสรุปลักษณะเด่นของข้อมูล

- Characterization (การสรุปลักษณะ): สรุปคุณสมบัติเด่นของกลุ่มข้อมูล เช่น:
 - ตัวอย่าง: ลูกค้าที่ใช้จ่ายเกิน 30,000 บาท/ปี มักจะเป็น "ผู้ชาย อายุ 40-50 ปี ทำงานบริหาร"
- Discrimination (การเปรียบเทียบความต่าง): เปรียบเทียบลักษณะของกลุ่ม เป้าหมายกับกลุ่มอื่น
 - ตัวอย่าง: เปรียบเทียบลูกค้าที่ "ซื้อคอมพิวเตอร์" vs "ไม่ซื้อคอมพิวเตอร์" ว่ามีอะไรต่างกันบ้าง (เช่น รายได้ต่างกันไหม?)

5.2 Frequent Patterns, Associations, and Correlations (รูปแบบที่พบบ่อยและความสัมพันธ์)

- Frequent Itemsets: กลุ่มของสินค้าที่มักปรากฏพร้อมกันบ่อยๆ
- Association Rules: กฎความสัมพันธ์แบบ "ถ้า...แล้ว..."
 - ตัวอย่างคลาสสิก: Beer & Diapers (คนซื้อลูกอ่อนที่มาซื้อผ้าอ้อม มักจะซื้อเบียร์กลับไปดื่มด้วย) -> Buy(Diaper) -> Buy(Beer) [Support=2%, Confidence=60%]

5.3 Classification (การทำนายประเภท) และ Prediction (การทำนาย)

- Classification: ทำนายข้อมูลให้อยู่ในกลุ่ม (Class) ที่กำหนดไว้ล่วงหน้า (Discrete Labels)
 - ตัวอย่าง: ทำนายว่าลูกค้าจะ "อนุมัติบัตรเครดิต" หรือ "ไม่อนุมัติ" (Yes/No)
 - โมเดลที่ใช้: Decision Tree, Neural Network, Naïve Bayes, Support Vector Machine (SVM)
- Regression (Prediction): ทำนายค่าตัวเลขต่อเนื่อง (Continuous Values)
 - ตัวอย่าง: ทำนายยอดขายเดือนหน้า (เป็นตัวเลขบาท), ทำนายราคาบ้าน

5.4 Cluster Analysis (การวิเคราะห์การจัดกลุ่ม)

- ต่างจาก Classification ตรงที่ "ไม่มีเล憋 (Labels) ล่วงหน้า" (Unsupervised Learning)
- หลักการ: จัดกลุ่มโดยให้ "ข้อมูลในกลุ่มเดียวกันเหมือนกันมากที่สุด (High Intracluster Similarity)" และ "ข้อมูลต่างกลุ่มกันต่างกันมากที่สุด (Low Interclass Similarity)"
 - ตัวอย่าง: การแบ่งกลุ่มลูกค้า (Segmentation) ตามพฤติกรรมการซื้อ เพื่อทำการตลาดเฉพาะเจาะจงกลุ่ม

5.5 Outlier Analysis (การวิเคราะห์ข้อมูลผิดปกติ)

- ค้นหาข้อมูลที่แปลกแยกจากกลุ่มส่วนใหญ่ (Outliers / Anomalies)
- การใช้งาน: การตรวจจับการโกรธบัตรเครดิต (Fraud Detection), การตรวจจับการบุกรุกเครือข่าย (Intrusion Detection), การวิเคราะห์ข้อมูลผิดพลาด

5.6 Trend and Evolution Analysis (การวิเคราะห์แนวโน้มและวิวัฒนาการ)

- วิเคราะห์การเปลี่ยนแปลงของข้อมูลตามเวลา
- ตัวอย่าง: รูปแบบการจราจรที่เปลี่ยนไปตามช่วงเวลา, พฤติกรรมลูกค้าที่เปลี่ยนไปตามฤดูกาล

6. What Kinds of Technologies Are Used? (ใช้เทคโนโลยีอะไรบ้าง?)

Data Mining ไม่ได้ทำงานโดยเดียว แต่เป็นศาสตร์ที่เกิดจาก การบรรจบกัน (Confluence) ของหลายสาขาวิชา:

1. Statistics (สถิติ): รากฐานของการวิเคราะห์ตัวเลข, การทดสอบสมมติฐาน
2. Database Technology (เทคโนโลยีฐานข้อมูล): การจัดการข้อมูลจำนวนมาก, Indexing, Query Optimization
3. Machine Learning (การเรียนรู้ของเครื่อง): อัลกอริทึมที่เรียนรู้จากข้อมูล (AI)
4. Pattern Recognition (การจดจำรูปแบบ): การแยกแยะแพทเทิร์นในภาพหรือเสียง
5. Visualization (การสร้างภาพนิมิต): การแสดงผลกราฟิกให้คนเข้าใจง่าย
6. High-Performance Computing: การประมวลผลประสิทธิภาพสูงเพื่อจัดการ Big Data
7. Algorithms: อัลกอริทึมที่มีประสิทธิภาพในการคำนวณ

7. What Kinds of Applications Are Targeted? (ประยุกต์ใช้กับงานด้านใด?)

Data Mining นำไปใช้ได้แบบทุกวิธี:

- Web Page Analysis: การจัดอันดับหน้าเว็บ (PageRank), การแนะนำโฆษณา
- Collaborative Analysis & Recommender Systems: ระบบแนะนำสินค้า (เช่น Netflix และนำหนัง, Amazon และนำของ)
- Basket Data Analysis: วิเคราะห์ตระกูลสินค้าเพื่อวางแผนการขาย
- Biological & Medical Data: วิเคราะห์ DNA, การจำแนกโรค, การอักเสบแบบยา
- Data Mining for Software Engineering: หา Bug ในโค้ด, ตรวจสอบความซ้ำซ้อน
- Finance: ทำนายราคาหุ้น, ตรวจสอบการฟอกเงิน

8. Major Issues in Data Mining (ประเด็นปัญหาและความท้าทายหลัก)

การทำ Data Mining ไม่ใช่เรื่องง่าย มีความท้าทายหลายด้าน:

8.1 Mining Methodology (ระบบวิธีการทำเหมือง)

- Mining various and new kinds of knowledge: ต้องหาความรู้ได้หลายแบบ
- Mining knowledge in multi-dimensional space: ต้องหาความรู้ได้ในหลายมิติ
- Handling noise, uncertainty, and incompleteness: ข้อมูลจริงมักสกปรกและไม่สมบูรณ์ ต้องจัดการได้

8.2 User Interaction (การปฏิสัมพันธ์กับผู้ใช้)

- Interactive mining: ผู้ใช้สามารถปรับเปลี่ยนเงื่อนไขการค้นหาได้ระหว่างทาง
- Incorporation of background knowledge: ควรนำความรู้เดิมของผู้เชี่ยวชาญมาช่วยในการค้นหาได้
- Presentation: ผลลัพธ์ต้องเข้าใจง่าย สวยงาม

8.3 Efficiency and Scalability (ประสิทธิภาพและการขยายตัว)

- Efficiency: ต้องทำงานเร็ว ไม่ใช่รอนานเป็นวัน
- Scalability: ถ้าข้อมูลเพิ่มจาก 1 GB เป็น 1 TB ไม่เดลต้องยังทำงานไหว ไม่ล่ม
- Parallel & Distributed Mining: รองรับการประมวลผลแบบขนานและกระจายศูนย์

8.4 Diversity of Data Types (ความหลากหลายของข้อมูล)

- ต้องรองรับข้อมูลที่ซับซ้อน (Complex Data) เช่น กราฟ, รูปภาพ, ข้อมูลเชิงพื้นที่ ไม่ใช่แค่ตารางตัวเลข

8.5 Data Mining and Society (ผลกระทบต่อสังคม)

- Privacy-preserving data mining: การทำเหมืองข้อมูลต้องไม่ละเมิดความเป็นส่วนตัว (เช่น ข้อมูลสุขภาพ, ข้อมูลส่วนตัวลูกค้า)
- Invisible data mining: การฝัง Data Mining ลงในชีวิตประจำวันโดยที่ผู้ใช้ไม่รู้ตัว (เช่น Google Search, GPS)

9. A Brief History of Data Mining and Data Mining Society (ประวัติย่อ)

- 1989: คำว่า "Knowledge Discovery in Databases (KDD)" ถูกบัญญัติขึ้นครั้งแรก (KDD Workshop)
- 1990s: เริ่มมีการใช้คำว่า "Data Mining" อย่างแพร่หลาย
- 1995: เริ่มมีการจัดประชุมวิชาการนานาชาติ KDD Conference ครั้งแรก
- 1998: ก่อตั้งกลุ่ม ACM SIGKDD (กลุ่มวิชาชีพด้าน KDD)
- 2000s เป็นต้นมา: Data Mining แทรกซึมไปในงานวิจัยสาขาอื่น (DB, AI, Statistics, Web) และภาคธุรกิจอย่างกว้างขวาง

10. Summary (บทสรุป)

- Data Mining คือการค้นหาแพทเทิร์นที่น่าสนใจและความรู้จากข้อมูลจำนวนมหาศาล
- เป็นขั้นตอนสำคัญในกระบวนการ KDD Process
- Data Warehouse และ OLAP เป็นเทคโนโลยีที่ช่วยเตรียมข้อมูลและวิเคราะห์โดยมิติ
- พัฒนาหลัก: Characterization, Discrimination, Association, Classification, Clustering, Outlier Analysis, Trend Analysis
- Data Mining เกิดจากการรวมกันของ Database, ML, Statistics และศาสตร์อื่นๆ
- ความท้าทาย: ประสิทธิภาพ, การร้องรับข้อมูลที่หลากหลาย, และเรื่องความเป็นส่วนตัว (Privacy)