

# 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



ChengXiang  
Zhai

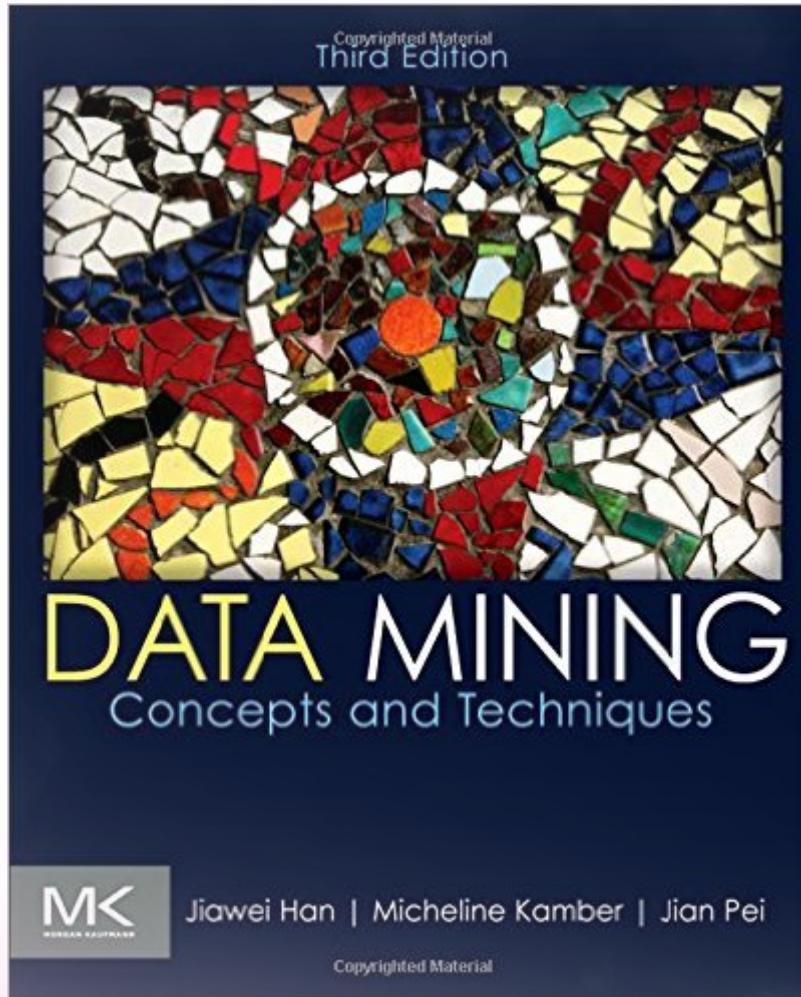
- Text Information Systems

- 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



សាស្ត្រ

- Textbook
  - Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques* (3<sup>rd</sup> 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**

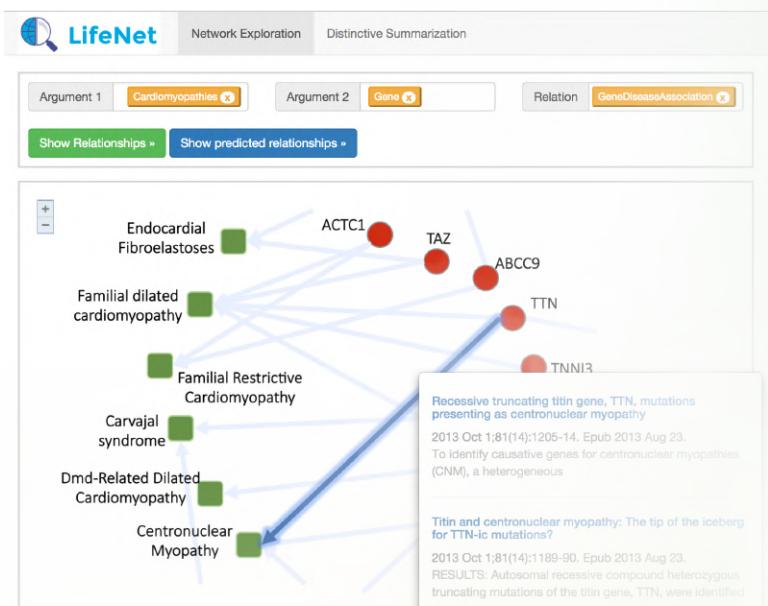
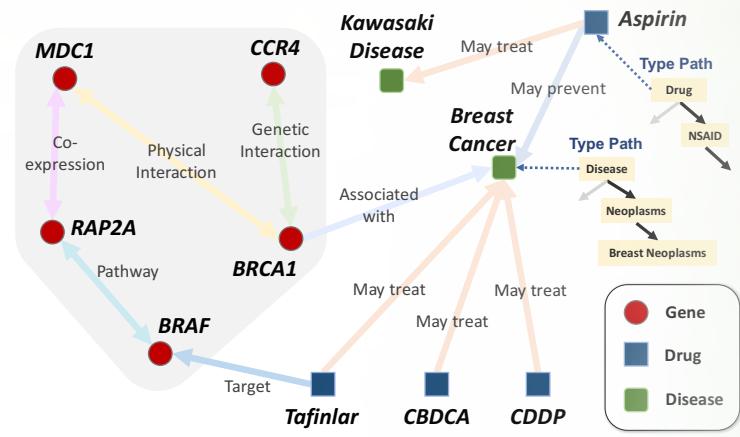
- TA office hours: **4-5pm (Mon.), 11-12pm (Wed.)@0207SC**. Additional hours before due date will be announced at Piazza
- Wait list (No wait list at this time, keep attending class, see if there is space available or there is overflow section opening)
  - If you cannot register but still desperately 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 4<sup>th</sup> credit (TA will provide concrete instructions on the 4<sup>th</sup> 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? *ផ្តល់ព័ត៌មាននៃការបោះពីរនៃបង្ហាញ*
  - A scalable system that transforms biomedical papers into a knowledge graph & supports various search/analytics functions *ការអនុវត្តន៍យកព័ត៌មានពីការបោះពីរ*
- What we already have? *មិនបានព័ត៌មាន*
  - 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](mailto:mickeysjm@gmail.com))**

# 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?

កំណត់ការអវិជ្ជមាន

ការពេញ

- 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 **លេខធម្មនៃភ័ស្តុល្អាកម្មរបស់យើង**
    - 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



គាយការពេញ ឱ្យបានការណើលួចបារប្រែបីជីថទិន្នន័យ

ឡូចធម្មន៍

ការវិទ្យាព័ត៌មានដែលមានប្រឈម

# 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

# ຈະໄດ້ຕົວແນ່ນວິທີນຸລ What Is Data Mining?

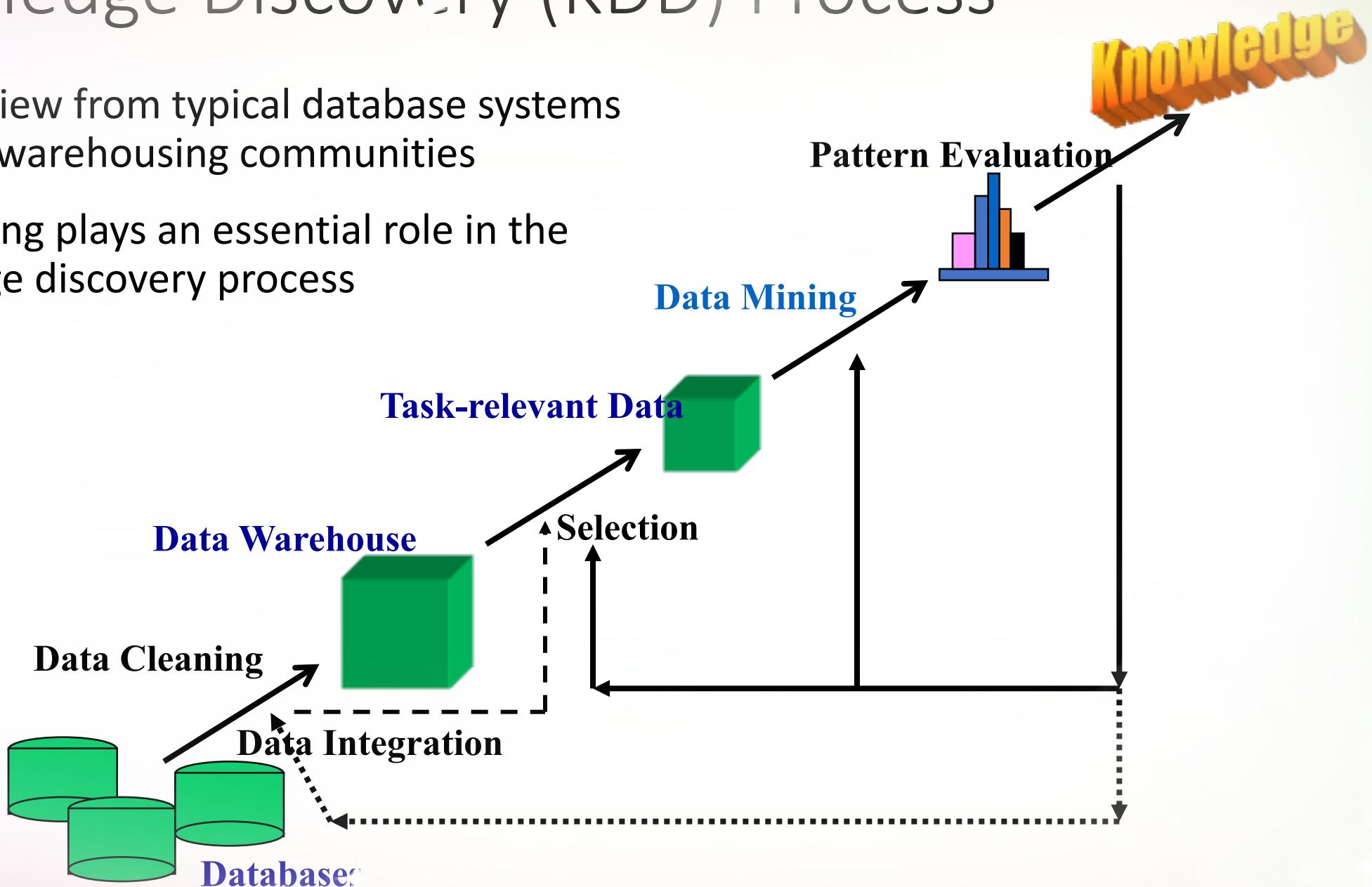


- Data mining (knowledge discovery from data)  
ການດັ່ງພບຄວາມຮູ້ຈຳດັບໂນງ  
ການຕັ້ງຽບແບບຂໍ້ອຸນຫຼວດກາກູ້ກີ່ເສັນໄວ
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?  
ໜຶບກໍສົ່ວໂລດ
- Alternative names:  
ການຄົ່ນຫາ ດາບຍຸ
  - 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”?  
ການປະໜາລັດຖະບານ ແລະ ແບບຂອບ ການອໍານໍາຍ່າຍ
  - Simple search and query processing
  - (Deductive) expert systems  
ຮັບ ກູ້ເສີ່ງຫາ ທະ



# Knowledge Discovery (KDD) Process

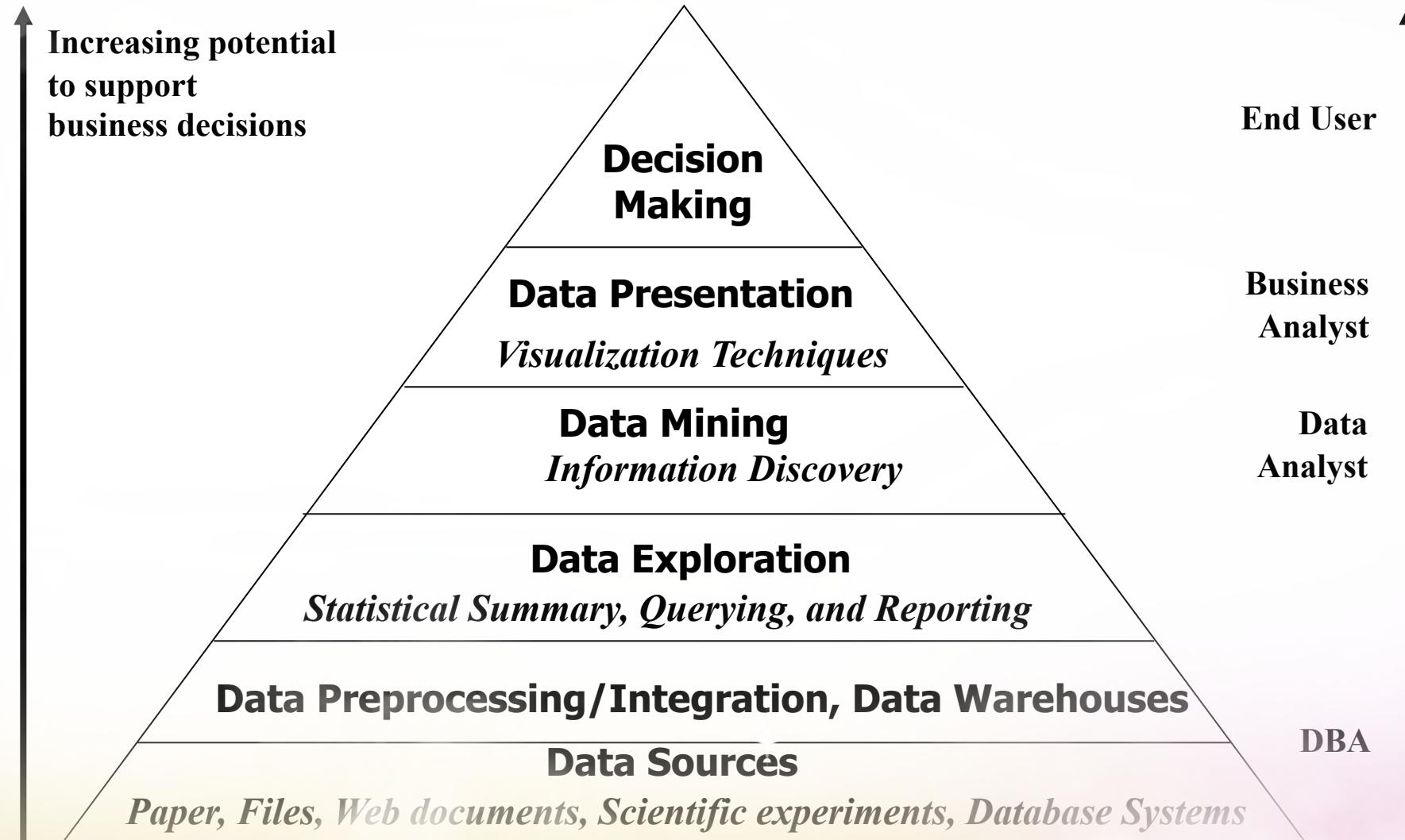
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



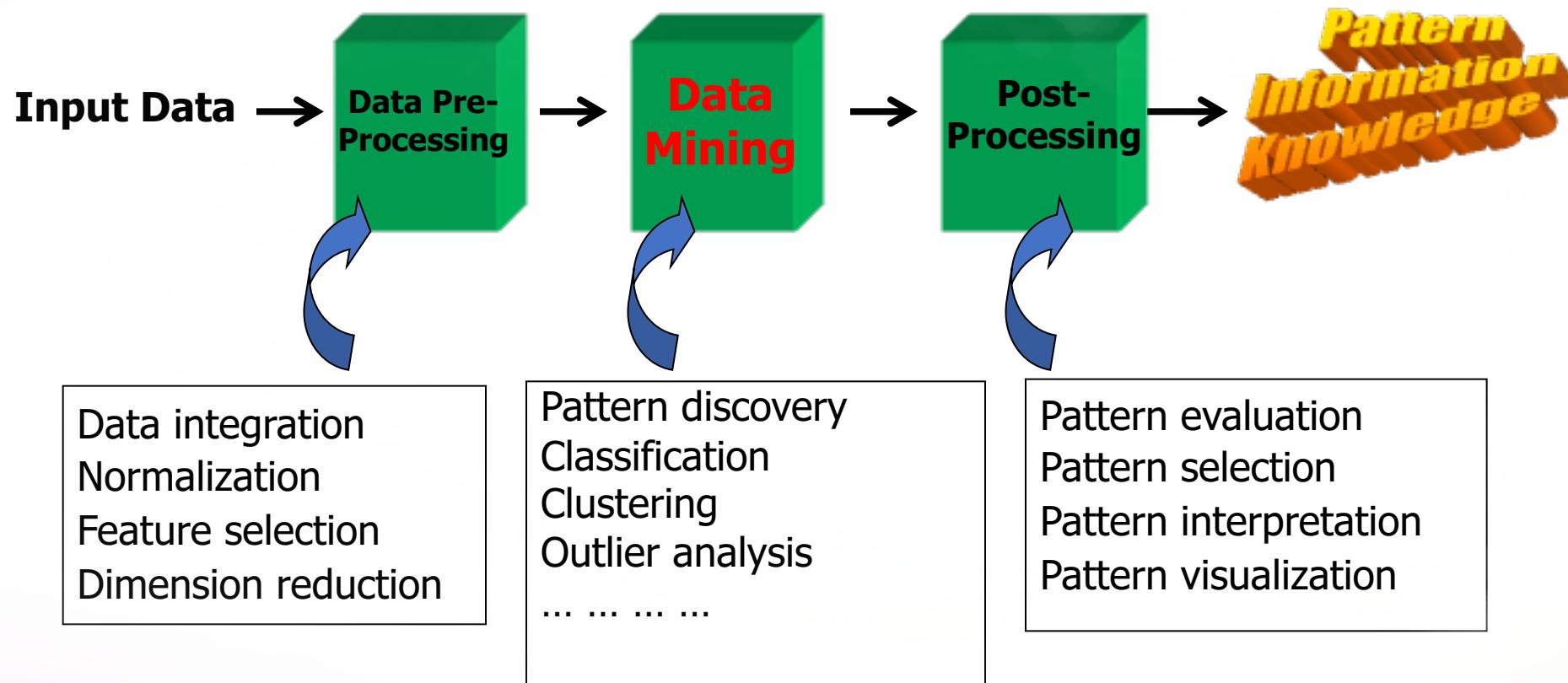
# Example: A Web Mining Framework

- Web mining usually involves
  - 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

# 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

- 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



# សំណុលលាបនិតិ

# Multi-Dimensional View of Data Mining

- Data to be mined ឯកសារក្នុងបច្ចេក  
• 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)  
ការរៀបចំឡើងឯកសារ សែនក្រឹមប្រព័ន្ធឌីជីថល ការតាំងក្រោមក្រោក ត័ត្រូវ  
• 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 ការគោលដៅ  
• 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? 
- 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? ជាមួយទៅ?

ច្បាប់

- Database-oriented data sets and applications

គ្រប់បង្កើត  
និងផ្តល់ព័ត៌មាន

ទូរគម្យ

- Relational database, data warehouse, transactional database
- Object-relational databases, Heterogeneous databases and legacy databases

ច្បាប់បង្កើត

- Advanced data sets and advanced applications

- Data streams and sensor data

ទូរគម្យ

ស៊ា

លេខ

- Time-series data, temporal data, sequence data (incl. bio-sequences)

- Structure data, graphs, social networks and information networks

សំគាល់

- 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

ការរួមទាំង នៃស្នើសុំគោលដៅ

- Information integration and data warehouse construction

- Data cleaning, transformation, integration, and multidimensional data model ផ្តល់បន្ទីត

- Data cube technology

ការបង្កើតគម្រោងការគោលដៅ

- Scalable methods for computing (i.e., materializing) multidimensional aggregates ការបង្កើតគម្រោងអំពីតិច

- OLAP (online analytical processing) ធ្វើការជាមួយ

- Multidimensional concept description: Characterization and discrimination កិណ្ឌន៍ និងសំគាល់: ឈរពារ៉ា ឬការវិភាគប្រចាំថ្ងៃ

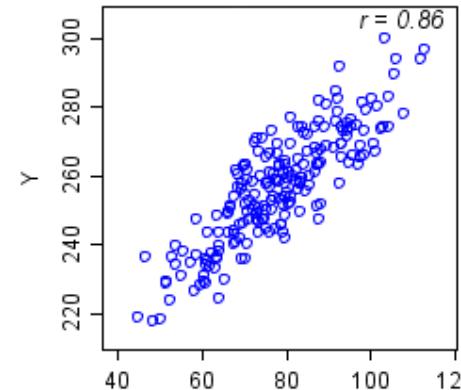
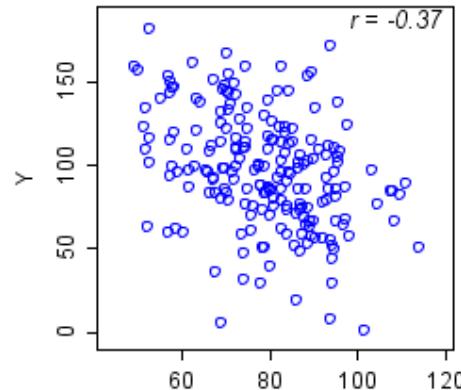
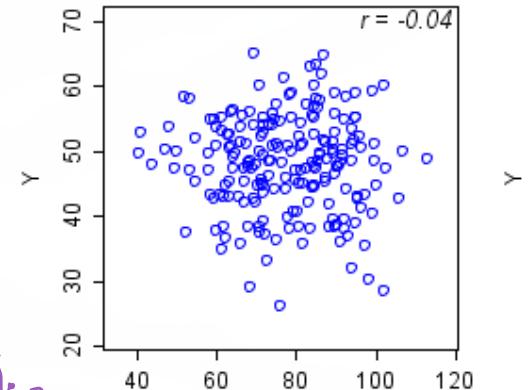
- Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region ស្រួល និងបញ្ជូនគុណភាពខ្លួន



# Data Mining Functions: (2) Pattern Discovery

ទីផ្សារ ស្តីពី ទុកដាក់ និង ការងារ

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



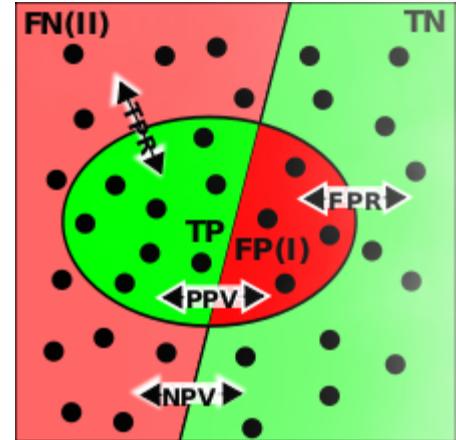
- A typical association rule  Ex. តាមឱ្យលានឯកសារអេឡិចត្រូនិក ក្នុងការងារ តាមឱ្យលានឯកសារអេឡិចត្រូនិក ក្នុងការងារ
- Diaper  $\rightarrow$  Beer [0.5%, 75%] (support, confidence)
- Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

# Data Mining Functions: (3) Classification

ការវាំងទេរស័ព្ទ

- 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, ...

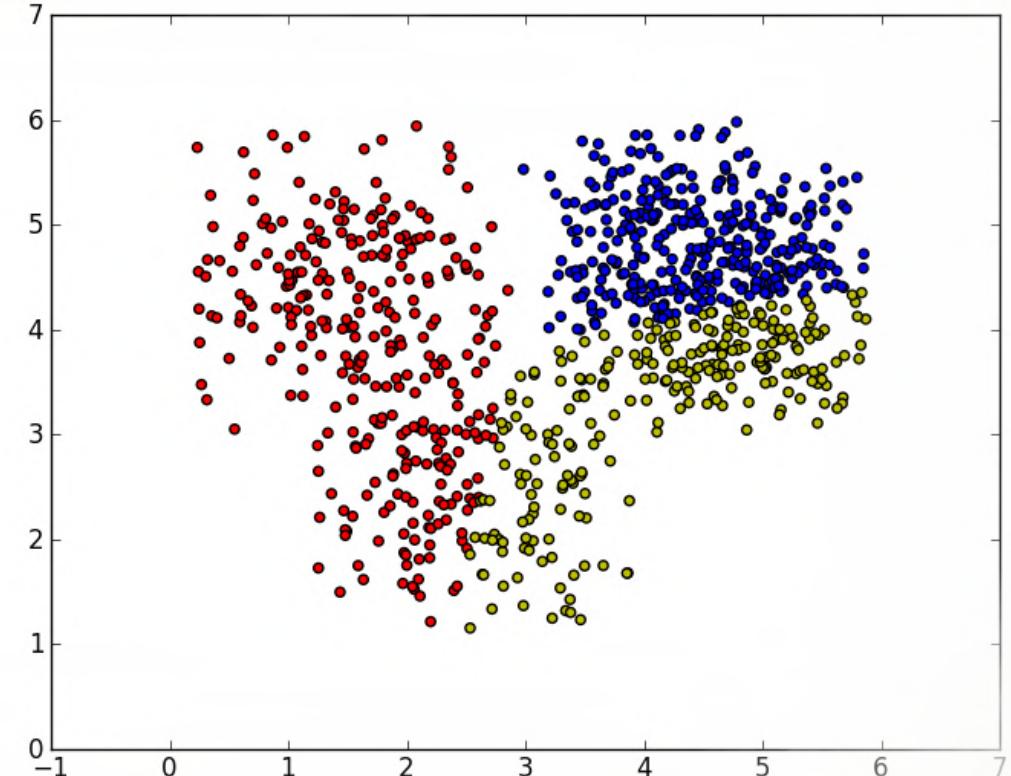
ការគ្រប់គ្រងការផ្តល់ការពិនិត្យ

ការគ្រប់គ្រងការពិនិត្យ

# Data Mining Functions: (4) Cluster Analysis

ការបង្កើរបែងចែកពួកគេ នៃ ១ប្រភេទបង្កើរបែក

- 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



# Data Mining Functions: (5) Outlier Analysis

- Outlier analysis

ໄຟສອດລົງກັບຝຳຕີດຮູບ ທີ່ຢູ່ປະນຸມວົງ

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

ຟ້ວຍການດັ່ງນີ້ຍັງ

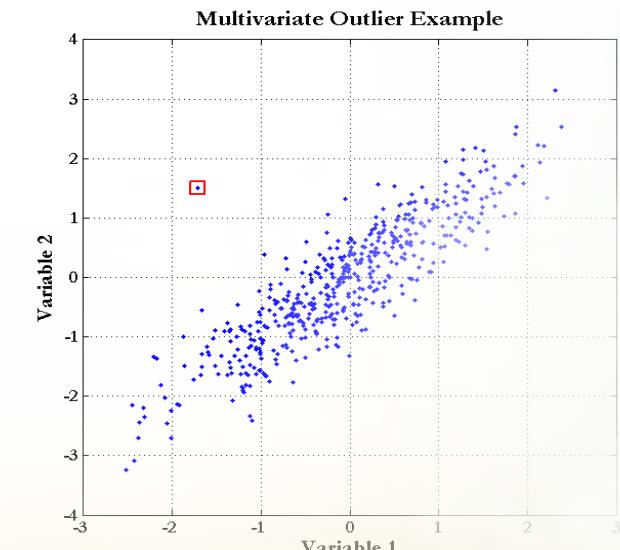
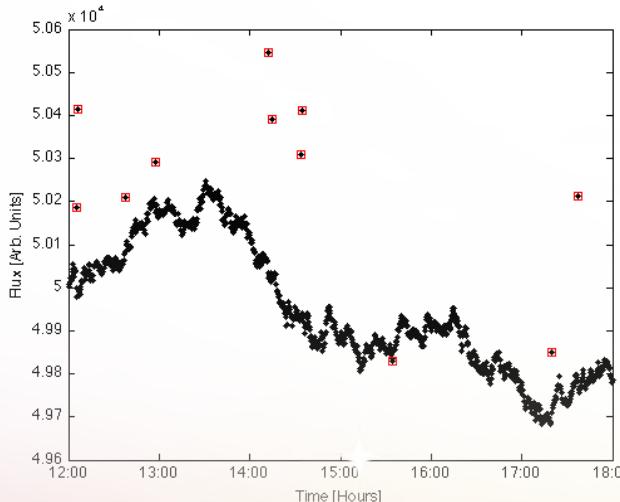
- Noise or exception?—One person's garbage could be another person's treasure

ໂຄງການຂອບດາວເຕຣະນິກຈັດກົມໆແຫ່ງດາວວ່າເຄົກະນິກດອຍ

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

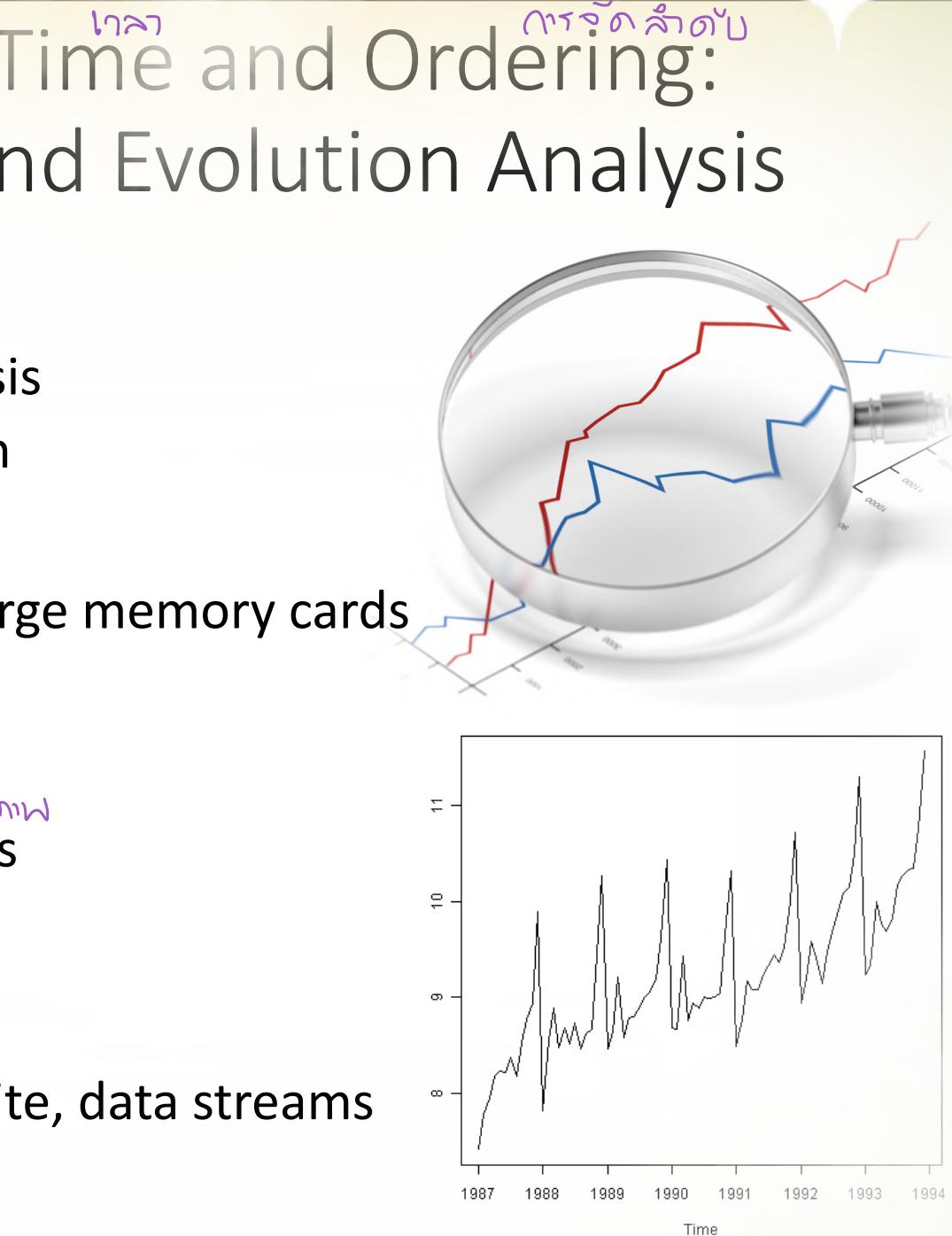
ລັບໂຍໍ້ນິກຕຽງການດ້ວຍໂລກ , ດາວເຕຣະພົມເຕັກການ

- 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**
  - Trend, time-series, and deviation analysis
    - e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., buy digital camera, then buy large memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
    - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams



# Data Mining Functions: (7) Structure and Network Analysis

ការរកទម្រង់ប្រព័ន្ធដែលមានបន្ទូល

- Graph mining ការរកបែងចែករាង

គិតថាបានបានបំផុត

- Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)

ការប្រកបណ្តុះ

- Information network analysis វិភាគប្រព័ន្ធដែលមានបន្ទូល

- Social networks: actors (objects, nodes) and relationships (edges)

គិតថាបានបំផុត

- e.g., author networks in CS, terrorist networks

- Multiple heterogeneous networks

- A person could be multiple information networks: friends, family, classmates,

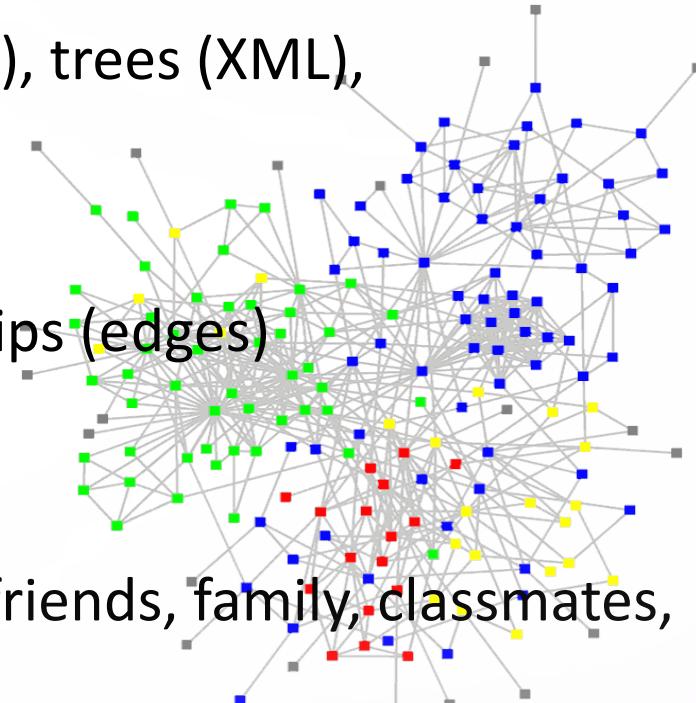
...

- 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



## ការប្រែងដែងទំនាក់ទំនង

# Evaluation of Knowledge

- Are all mined knowledge interesting? បង្ហីតីខ្ពស់ទៅការប្រែងដែងទំនាក់ទំនងទាំងអស់

- One can mine tremendous amount of “patterns”  
សរសៃក្នុងទីតាំងនៃវឌ្ឍនភាព
- Some may fit only certain dimension space (time, location, ...)  
បានត្រូវតាមលក្ខណៈកំណត់ហើយ និងពេលវេលា
- Some may not be representative, may be transient, ...  
មានតម្លៃតាមពេលវេលាដែលចំណែក និងចំណែក

- Evaluation of mined knowledge → directly mine only interesting knowledge?

- Descriptive vs. predictive  
ការពិពណ៌នា និងការពិនគរណៈ
- Coverage  
ការចាត់ការងារ
- Typicality vs. novelty  
ແណ្ឌភាពឱ្យរាយ
- Accuracy  
ការពារមិនមែន
- Timeliness  
កំណែវត្ថា, កំណែអ្នកប្រើប្រាស់
- ...

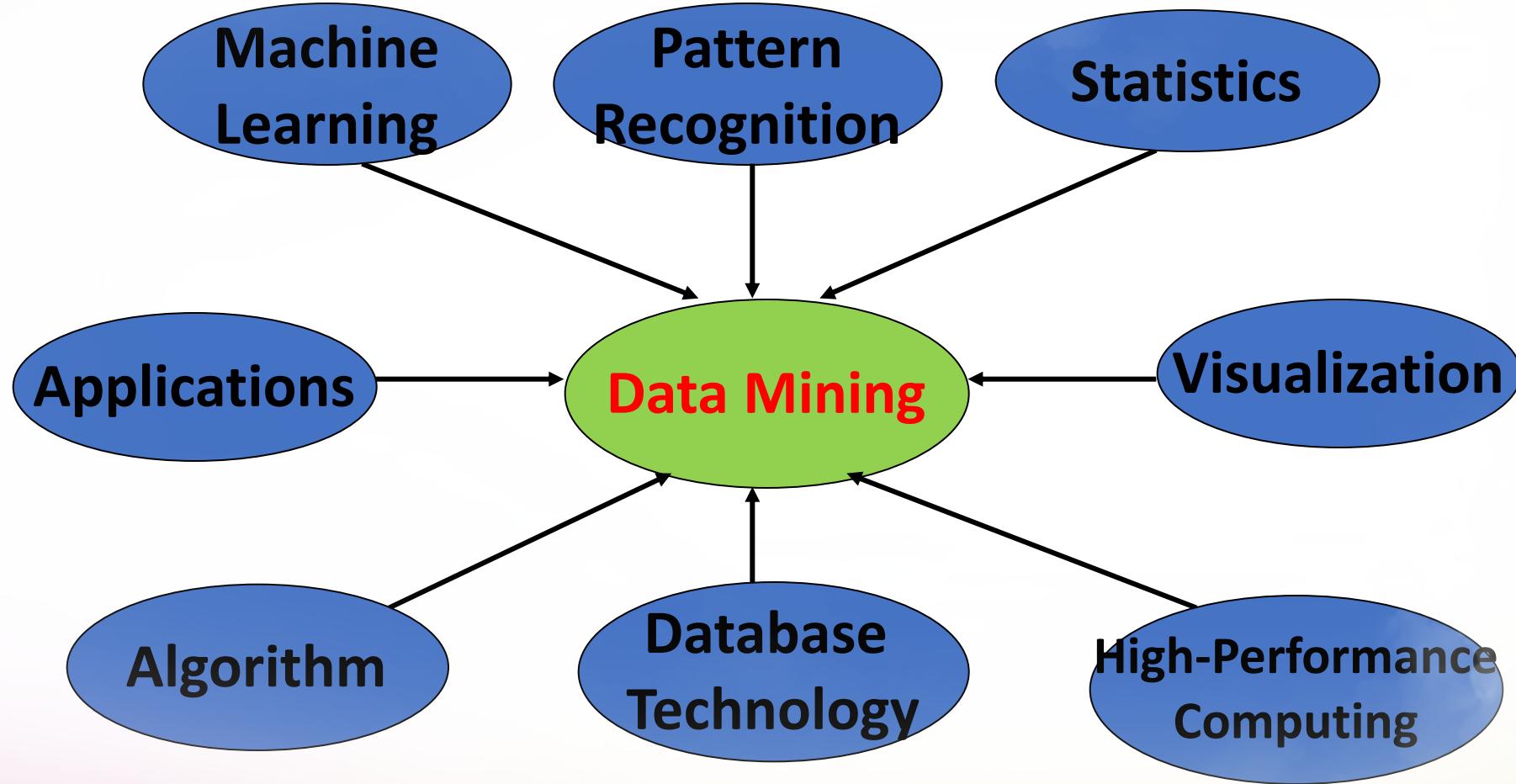


# Chapter 1. Introduction

- 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?

សំខាន់សម្រាប់បច្ចុប្បន្ន

- 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

# 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

# ການປະ: ດູກໂລກ

## Applications of Data Mining

- Web page analysis: classification, clustering, ranking  
ການວິດຄະໜົບເຫັນເວັບ ລຳແນກ ຈົດກຸ່ມ ລັດຕິນຕັບ
- Collaborative analysis & recommender systems  
ກ່ຽວຂ້ອງ: ຜົນງັນ ຜົນ: ທີ່  
ພົນດາກຄລາດເນື່ອງນັບ
- 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  
ຝຶກສົດໃຫ້ກໍາ ກຽບຂໍ້ວຸງກໍາມອນສິ້ນ
- Built-in (invisible data mining) functions in Google, MS, Yahoo!, LinkedIn, Facebook, ...
- Major dedicated data mining systems/tools  
ເຄື່ອນໄຫວທີ່ແນ່ວັນປອງຊຸກ
- 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) ប្រចែនលក្ខិនការប្រើប្រាស់

## វិធានប្រុទ

- Mining Methodology

ប្រព័ន្ធលាប នៃ ពិសេស

- Mining various and new kinds of knowledge

អភិវឌ្ឍន៍

- Mining knowledge in multi-dimensional space

រាយពហូម ឈងខ្ពស់រវាងការ

- Data mining: An interdisciplinary effort

ជំនួយការកំណត់ប្រើប្រាស់សារពាណិជ្ជកម្ម

- Boosting the power of discovery in a networked environment

- Handling noise, uncertainty, and incompleteness of data

- Pattern evaluation and pattern- or constraint-guided mining

## ការចិត្តរបស់អ្នកឱ្យ

ការចិត្តរបស់អ្នកឱ្យ

- Interactive mining

ការចិត្តរបស់អ្នកឱ្យ

- Incorporation of background knowledge

ការចិត្តរបស់អ្នកឱ្យ

- Presentation and visualization of data mining results

# Major Issues in Data Mining (2)

ក្រោមខាងក្រោម នូវ ៨ នាមីរបៀប

- Efficiency and Scalability

ស្ថិតិយវត្ថុ

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods

គ្រប់ប្រភព

- Diversity of data types

- Handling complex types of data
- Mining dynamic, networked, and global data repositories

- Data mining and society

នៅលទ្ធភាពការសង្គម

- Social impacts of data mining
- Privacy-preserving data mining
- Invisible data mining

ការអនុវត្តន៍យកដោយត្រូវបាន



# 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

ការប្រើប្រាស់ការសរសៃរបស់ខ្លួន: ការសរសៃរបស់ខ្លួនដែលមានព័ត៌មានចំណាំ

- Data mining: Discovering interesting patterns and knowledge from massive amount of data  
ការប្រើប្រាស់ការសរសៃរបស់ខ្លួនដែលមានព័ត៌មានចំណាំ
- A natural evolution of science and information technology, in great demand, with wide applications  
មែនទៀតចងក្រោម
- 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



# 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, 3<sup>rd</sup> ed. , 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2<sup>nd</sup> 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 (2<sup>nd</sup> ed. 2016)
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2<sup>nd</sup> ed. 2005
- Mohammed J. Zaki and Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms 2014

