# CSE5243 INTRO. TO DATA MINING

## Chapter 1. Introduction

Yu Su, CSE@The Ohio State University

Slides adapted from UIUC CS412 by Prof. Jiawei Han and OSU CSE5243 by Prof. Huan Sun

# CSE 5243. Course Page & Schedule

- **Class Homepage:**

  https://ysu1989.github.io/courses/sp20/cse5243/

- **Class Schedule:**

  **9:35-10:55 AM, Wed/Fri, Caldwell Lab 171**

- **Office hours:**
  - Instructor: Yu Su @ DL783, Fri 11:00am-12:15pm (right after class)

    **First week: No office hours**

  - TA: Jiaqi Xu (xu.1629), Wed 03:00pm-04:00pm, Baker 406

# CSE 5243. Textbook

- **Recommended but not required**
- (Primary) Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques (3rd ed)*, 2011
  - More resources: https://wiki.illinois.edu/wiki/display/cs412/2.+Course+Syllabus+and+Schedule
- (Primary) Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*, *2006*
- (Supplementary) Mohammed J. Zaki and Wagner Meira, Jr., *Data Mining Analysis and Concepts*, *2014*
- (Supplementary) Jure Leskovec, Anand Rajaraman, Jeff Ullman, Mining of Massive Datasets
  - *More resources: http://www.mmds.org/*

# CSE 5243. Course Work and Grading

- Homework, Course Projects, and Exams
  - Participation: 10% (**Online discussion and/or class participation**)
  - Homework: 50% (**No Late Submissions!**)
  - Midterm exam: 20%
  - Final exam: 20%

- Need help and/or discussions?
  - Carmen: https://osu.instructure.com/courses/76423/discussion_topics
    - Receive credits: answer questions on Carmen and engage in class discussion.

- Check your homework/exam scores
  - Carmen: https://osu.instructure.com/courses/76423/gradebook

# Videos

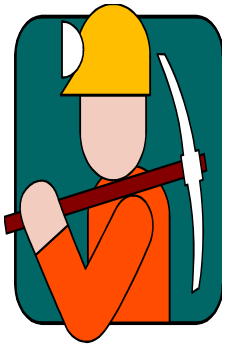
- 10 TED talks on Big Data and Analytics
  - https://www.promptcloud.com/blog/top-ted-talks-on-big-data/
  - Shyan Sanker (Director at Palantir Technologies):
    - https://www.youtube.com/watch?time_continue=19&v=ltelQ3iKybU
- 5 TED talks on Data analytics for business leaders
  - https://bigdata-madesimple.com/5-best-ted-talks-on-data-analytics-for-business-leaders/
- Data analytics for beginners
  - https://www.youtube.com/watch?v=66ko_cWSHBU (If you love sports, this TED Talk on data analytics is going to be an interesting watch)

# Chapter 1. Introduction

- What is Data Mining?
- Why 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

# What is Data Mining?

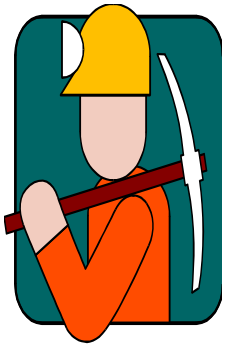
- Data mining (knowledge discovery from data, KDD)
  - Extraction of interesting (non-trivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amount of data



- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

# What is Data Mining?

- Data mining (knowledge discovery from data, KDD)
  - Extraction of interesting (non-trivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amount of data

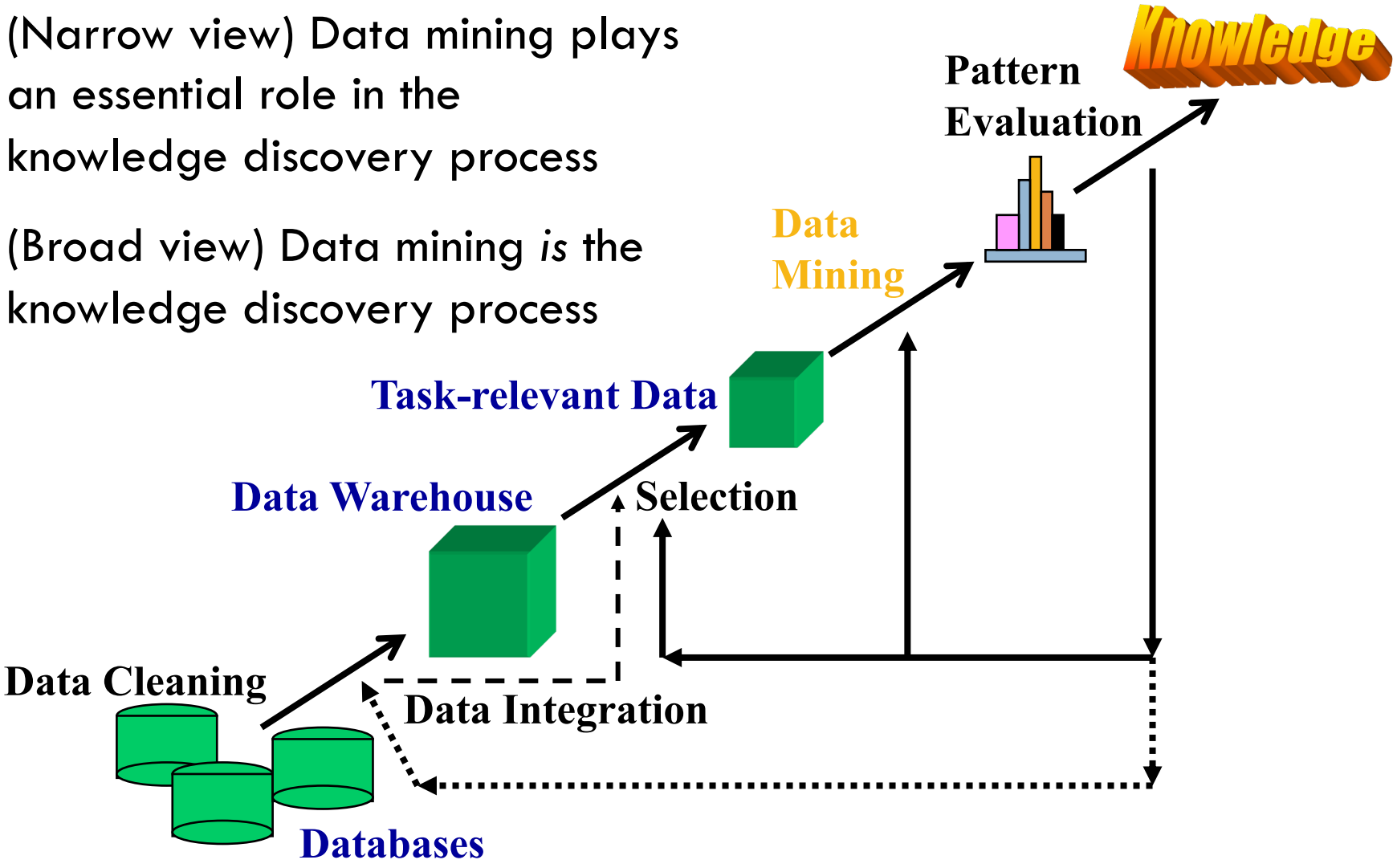


- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

One of the best conferences to publish your research work: SIGKDD (check resources)

# Knowledge Discovery (KDD) Process

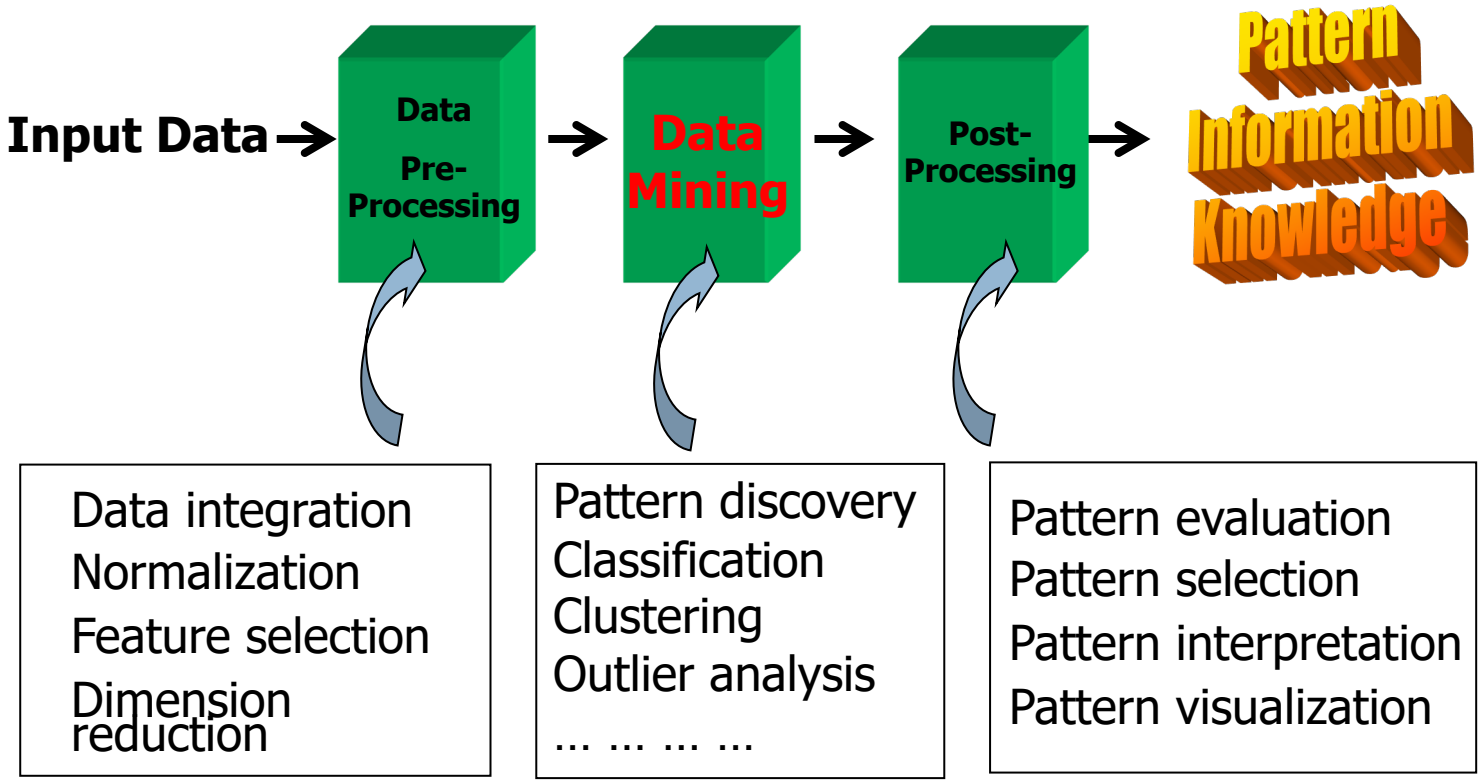
- (Narrow view) Data mining plays an essential role in the knowledge discovery process
- (Broad view) Data mining *is* the knowledge discovery process



# Example: A Web Mining Framework

- Web mining usually involves
  - Data crawling and cleaning
  - Data integration from multiple sources
  - (Optional) Warehousing the data
  - (Optional) 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

# KDD Process: A View from ML and Statistics




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

# Data Science

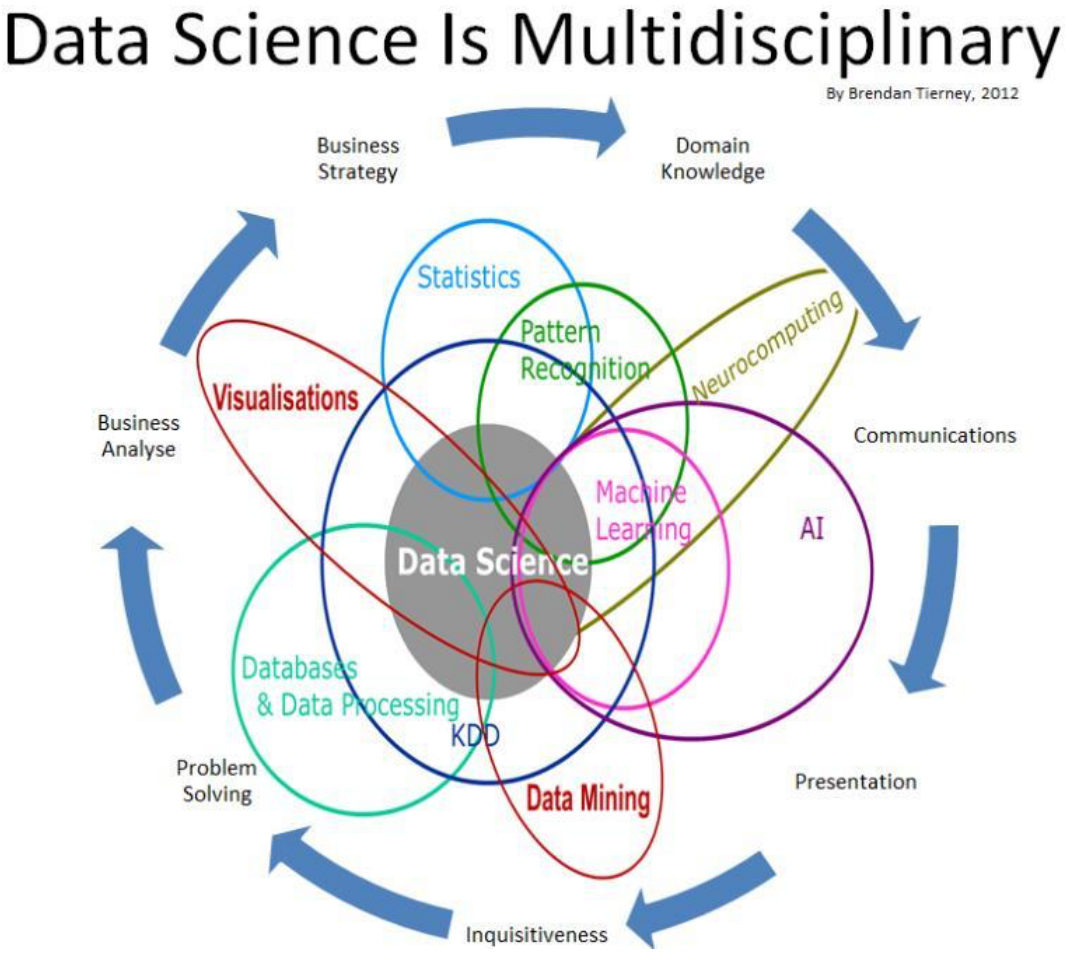



Figure from: https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining

# Chapter 1. Introduction

- What Is Data Mining?
- Why 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?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society

# 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 data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube

# A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

**500m**
tweets are sent every day
Twitter

**4PB**
of data created by Facebook, including
**350m** photos
**100m** hours of video watch time
Facebook Research

**320bn**
emails to be sent each day by 2021

**306bn**
emails to be sent each day by 2020

**294bn**
billion emails are sent
Radicati Group

**3.9bn**
people use emails

**4TB**
of data produced by a connected car
Intel

**DEMYSTIFIYING DATA UNITS**

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

| Unit | | Value | Size |
|---|---|---|---|
| b | bit | 0 or 1 | 1/8 of a byte |
| B | byte | 8 bits | 1 byte |
| KB | kilobyte | 1,000 bytes | 1,000 bytes |
| MB | megabyte | $1{,}000^2$ bytes | 1,000,000 bytes |
| GB | gigabyte | $1{,}000^3$ bytes | 1,000,000,000 bytes |
| TB | terabyte | $1{,}000^4$ bytes | 1,000,000,000,000 bytes |
| PB | petabyte | $1{,}000^5$ bytes | 1,000,000,000,000,000 bytes |
| EB | exabyte | $1{,}000^6$ bytes | 1,000,000,000,000,000,000 bytes |
| ZB | zettabyte | $1{,}000^7$ bytes | 1,000,000,000,000,000,000,000 bytes |
| YB | yottabyte | $1{,}000^8$ bytes | 1,000,000,000,000,000,000,000,000 bytes |

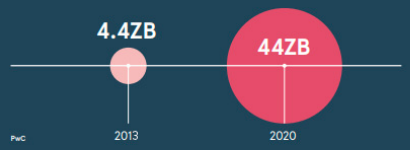
*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

**65bn**
messages sent over WhatsApp and two billion minutes of voice and video calls made
Facebook

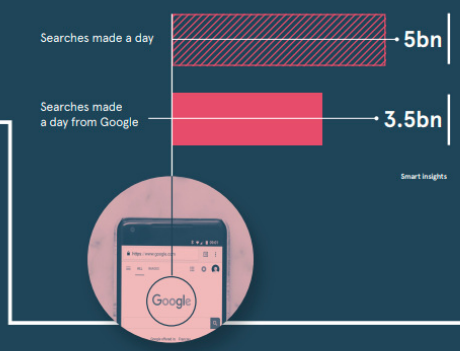
**463EB**
of data will be created every day by 2025
IDC

**95m**
photos and videos are shared on Instagram
Instagram Business

**ACCUMULATED DIGITAL UNIVERSE OF DATA**




PwC




Smart Insights

**28PB**
to be generated from wearable devices by 2020
Statista

RACONTEUR



"How much data is generated each day?" – World Economic Forum

# 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 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

# 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

- **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

# 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

# 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 warehousing (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance computing, etc.

# 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 warehousing (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance computing, etc.
- **Applications adapted**
  - 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?

- 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

# Survey

Your Name, ID, Major

Question 1: What do you think Data Mining is?

Question 2: What project have you done so far that you think is most relevant to Data Mining?

- Not necessarily research project; can be your course project or any hackathon event you participated in.

Question 3: What do you expect to learn from this course?

Briefly answer each question with a few sentences.

# Chapter 1. Introduction

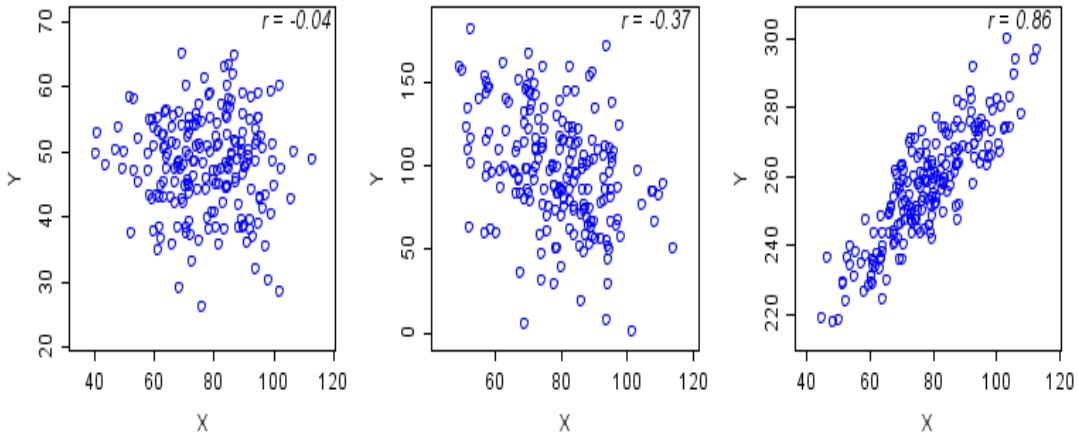
- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined? 
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

# Data Mining Functions: Pattern Discovery

- Frequent patterns
  - What items do you frequently purchase together on Amazon?

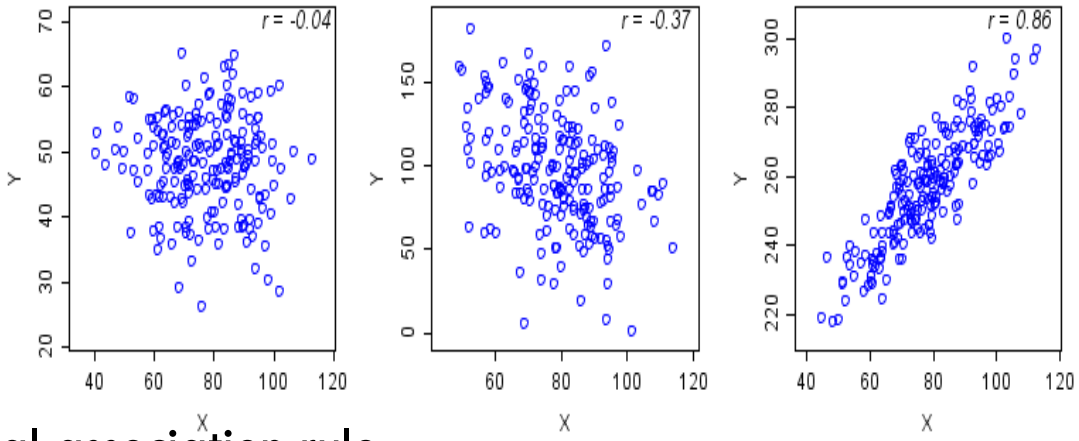
# Data Mining Functions: Pattern Discovery

- Frequent patterns
  - What items do you frequently purchase together on Amazon?
- Association and Correlation Analysis



# Data Mining Functions: Pattern Discovery

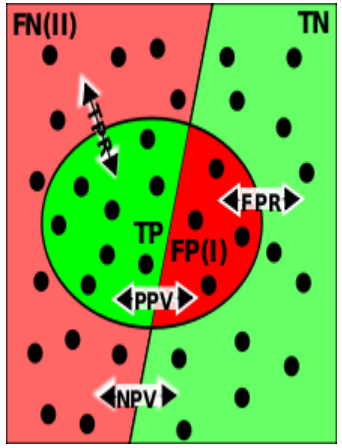
- Frequent patterns
  - What items do you frequently purchase together on Amazon?
- Association and Correlation Analysis




- A typical association rule
  - Diaper → Beer [0.5%, 75%] (support, confidence)
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?
- More: friend recommendation, motif discovery, malware detection, fraud detection, etc.

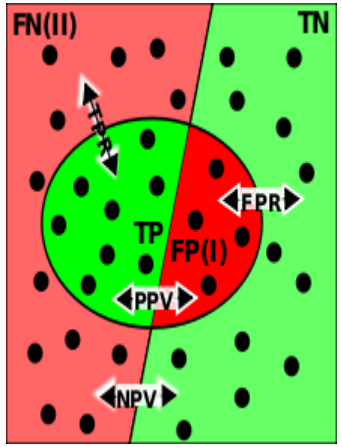
# Data Mining Functions: Classification

- 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



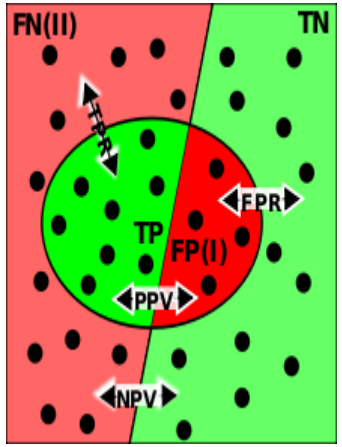
# Data Mining Functions: Classification

- 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, …



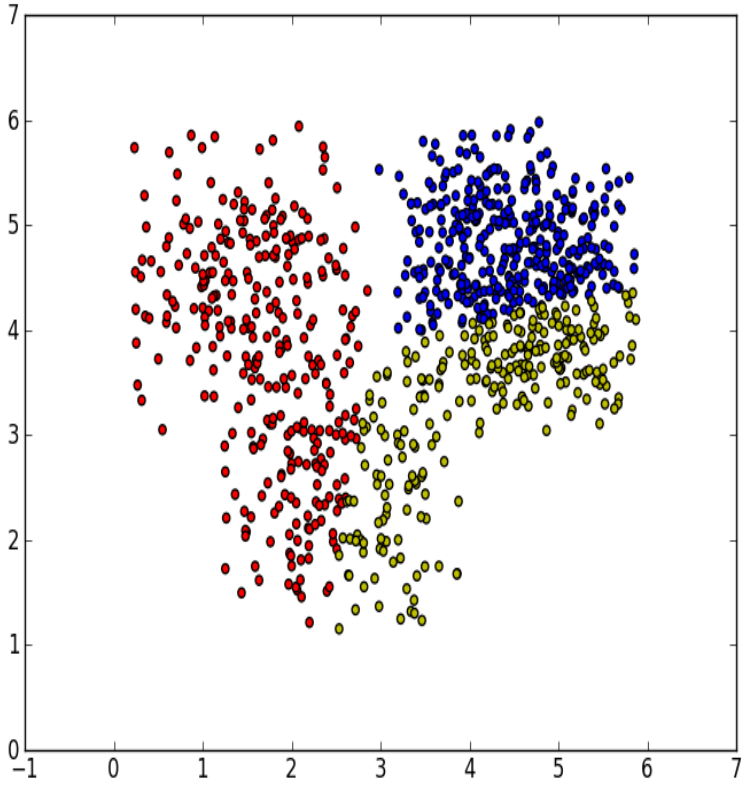
# Data Mining Functions: Classification

- 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, …



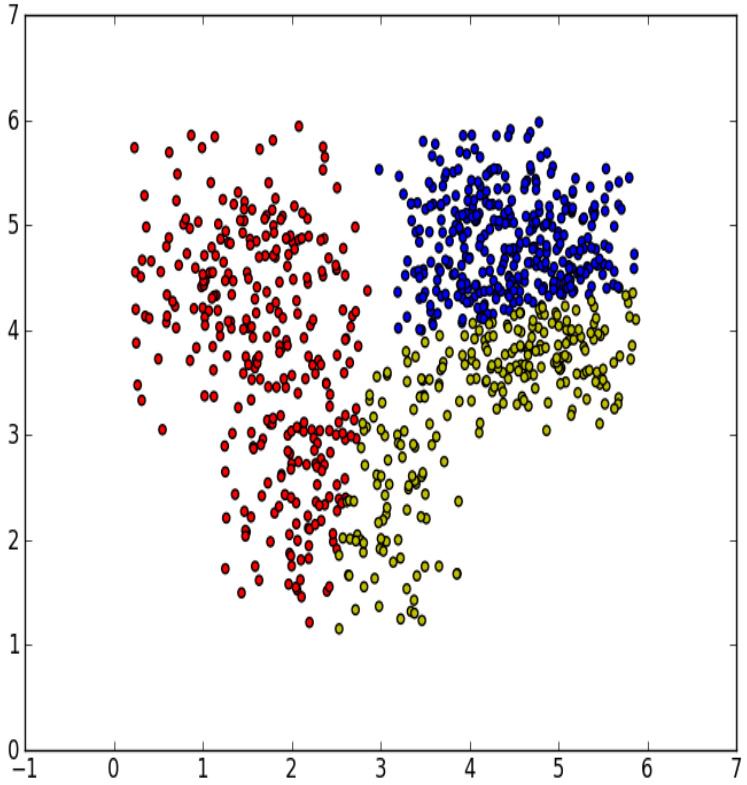
# Data Mining Functions: Cluster Analysis

- 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



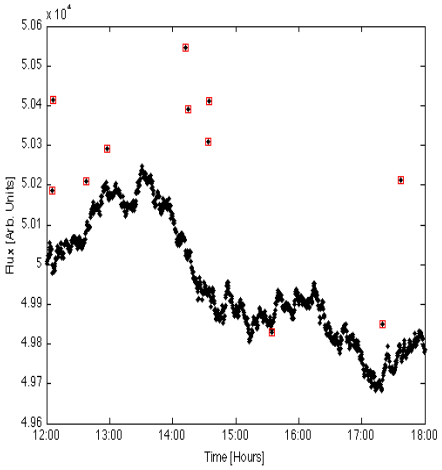
# Data Mining Functions: Cluster Analysis

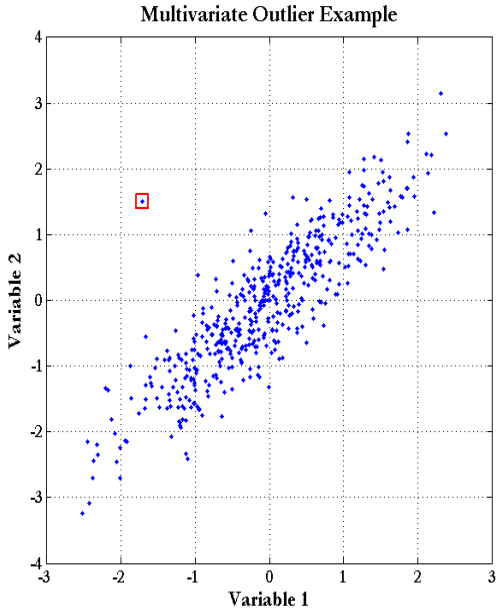
- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications



# Data Mining Functions: Outlier Analysis

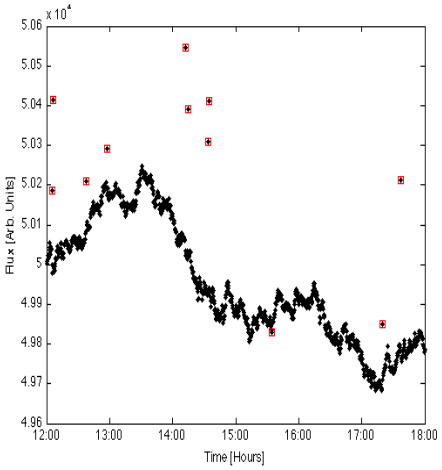
- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception?—One person's garbage could be another person's treasure

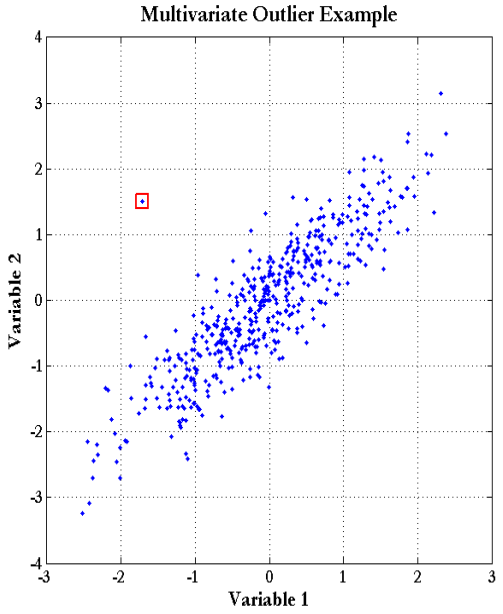


# Data Mining Functions: Outlier Analysis

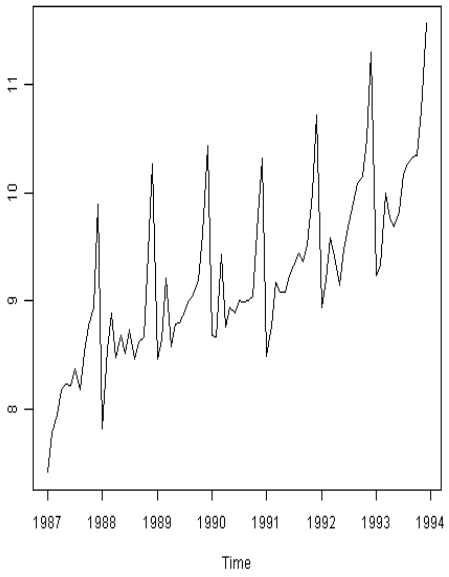
- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception?—One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis



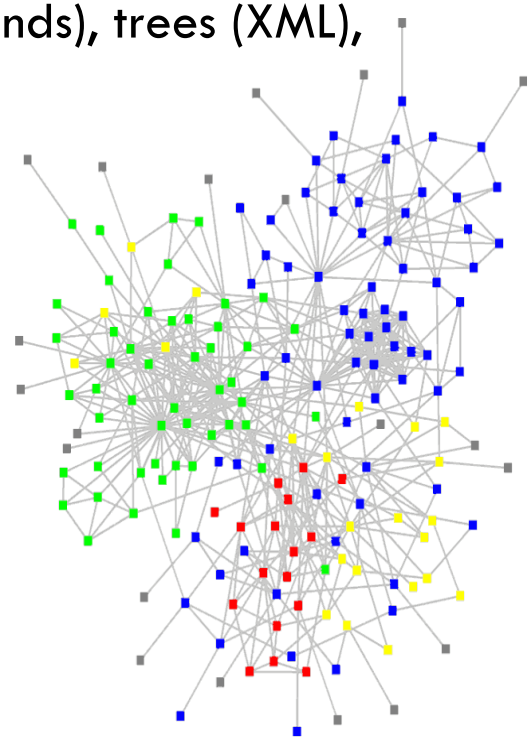
# Data Mining Functions: Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- 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



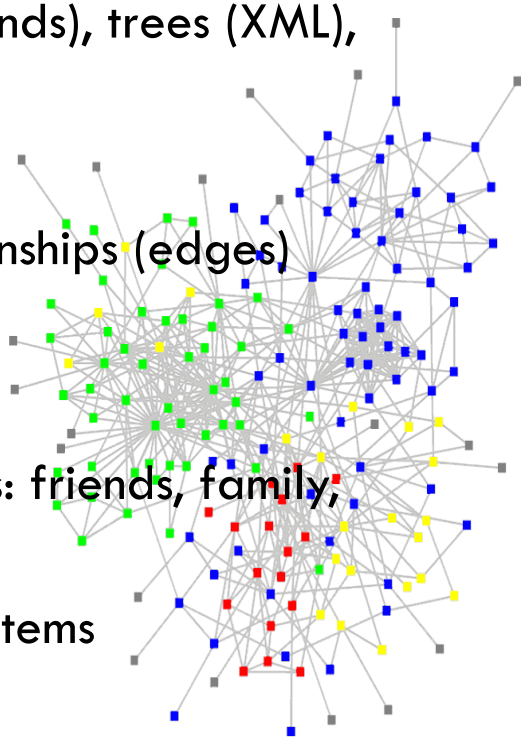
# Data Mining Functions: Structure and Network Analysis

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



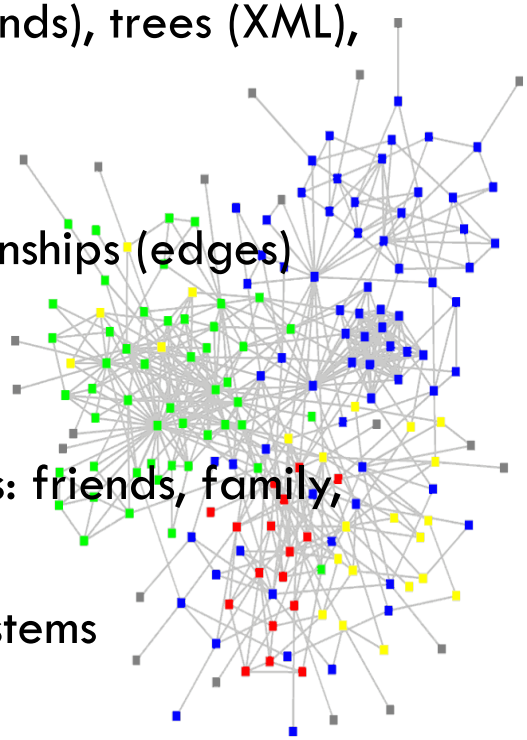
# Data Mining Functions: Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Knowledge graphs: knowledge backbone of AI systems



# Data Mining Functions: Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Knowledge graphs: knowledge backbone of AI systems
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …



# Evaluation of Knowledge

- **Are all mined knowledge interesting?**
  - One can mine tremendous amounts of "patterns"
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …

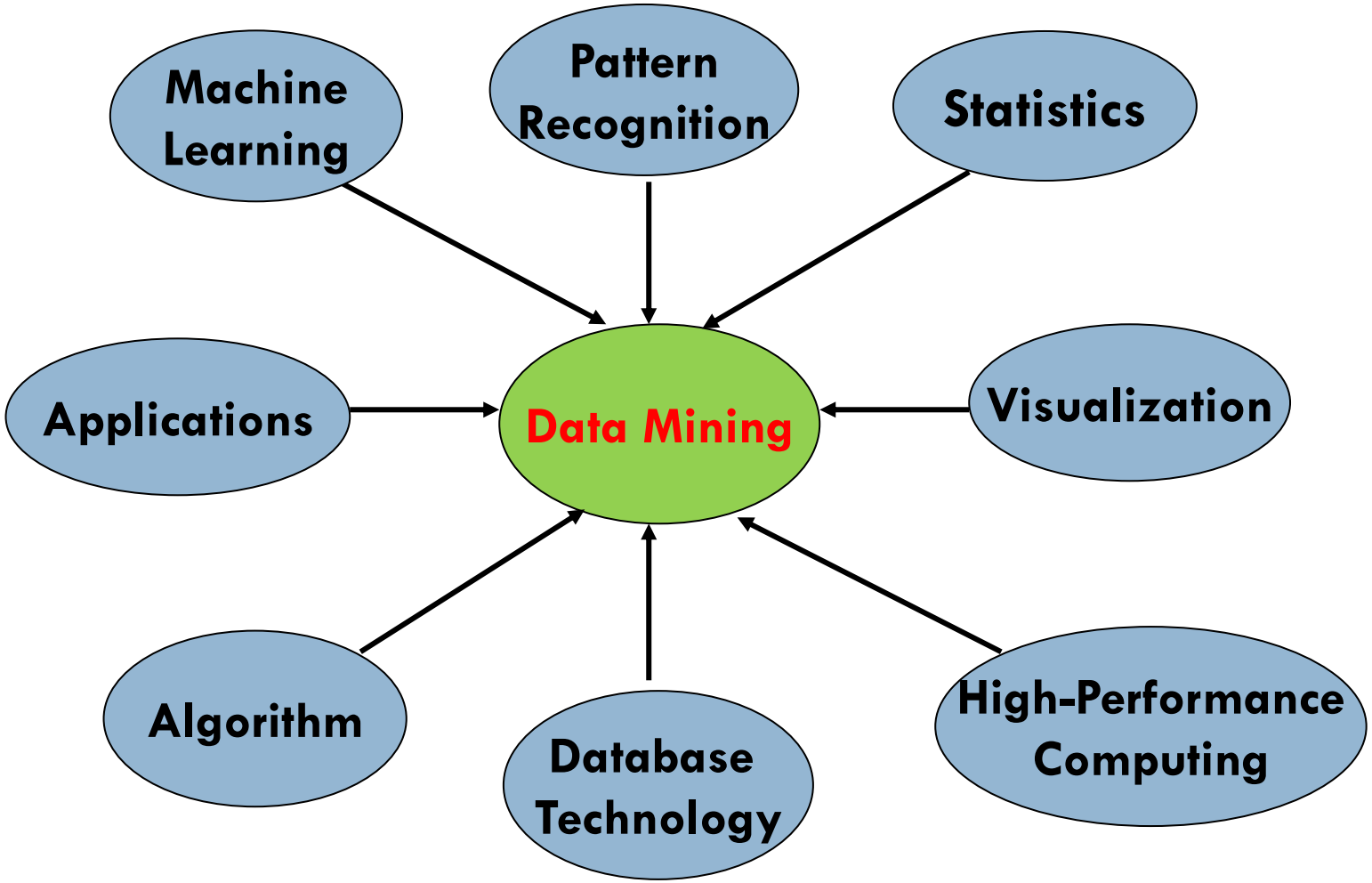
# Evaluation of Knowledge

- **Are all mined knowledge interesting?**
  - One can mine tremendous amount of "patterns"
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness
  - …

# 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?

- 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

# Applications of Data Mining

- Web page analysis: classification, clustering, ranking
- Collaborative analysis & recommender systems
- 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, …

# 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)

- 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

# 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 & Human-Machine Collaboration
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

# 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

# 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

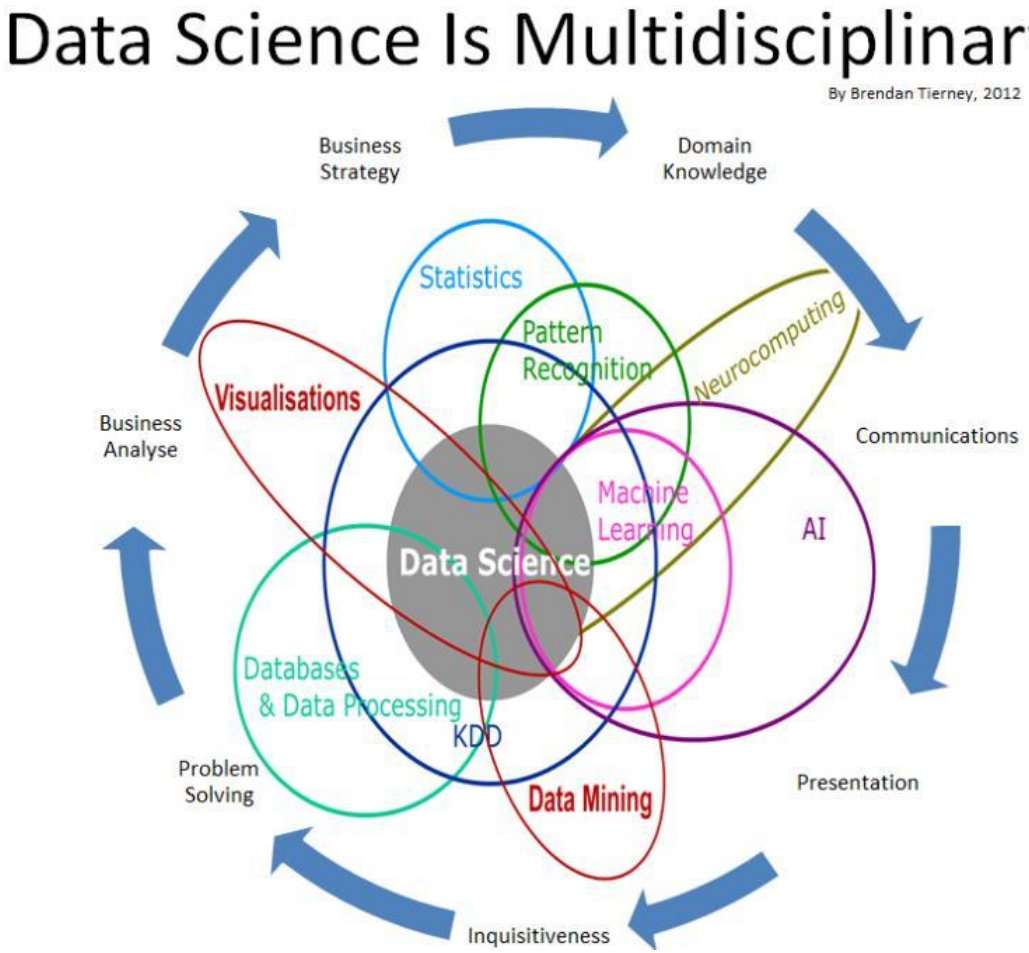
# Where to Find References? DBLP, CiteSeer, Google

- 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.

# Where to Find References? DBLP, CiteSeer, Google

- 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.

# Future of Data Science




https://www.youtube.com/watch?v=hxXIJnjC_HI (Future of Data Science @ Stanford)

Related events in OSU:

DataFest
Hackathon
Conduct research in labs

Figure from: https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining

# 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 

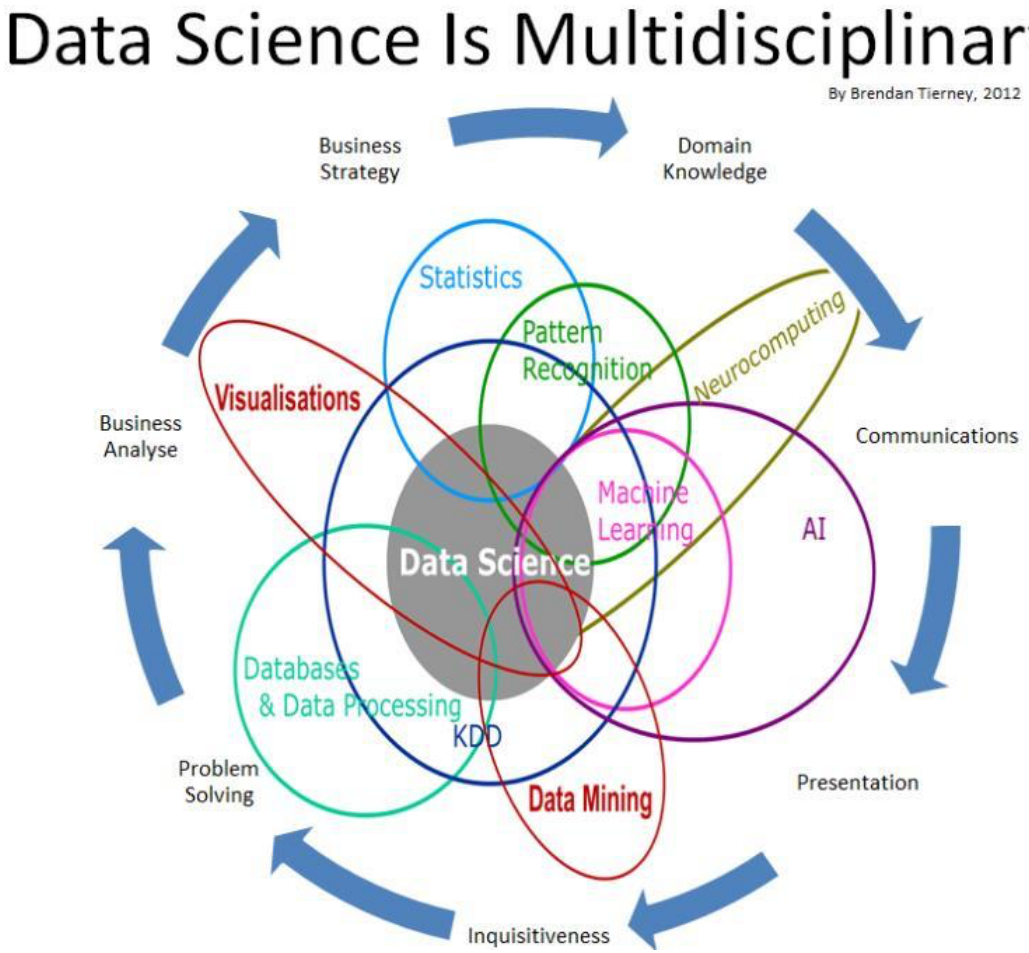
# Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- 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

# 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

# Future of Data Science




https://www.youtube.com/watch?v=hxXIJnjC_HI

Related events in OSU:

DataFest
Hackathon
Conduct research in labs

Figure from: https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining