CS 412 Summer 2017: Introduction to Data Mining

Chapter 1. Introduction

Meng Jiang

CS412 Summer 2017:

Introduction to Data Mining

Data and Information Systems (DAIS)

- Database Systems (Aditya Parameswaran)
- Data Mining (**Jiawei Han**, **Hari Sundaram**)
- Text Information Systems (**Kevin Chang**, ChengXiang Zhai)

DAIS Course Structures

- Data mining
  - Introduction to data warehousing and mining (CS412)
  - Data mining: Principles and algorithms (CS512)
- Database Systems
  - Introduction to database systems (CS411)
  - Advanced database systems (CS511)
- Text information systems
  - Text information system (CS410)
  - Advanced text information systems (CS510)

Official Description

- Concepts, techniques, and systems of data warehousing and data mining. Design and implementation of data warehouse and on-line analytical processing (OLAP) systems; data mining concepts, methods, systems, implementations, and applications.
- Course Information: 3 undergraduate hours. 3 or 4 graduate hours.
- Prerequisite: CS 225 (Data Structures).

About Me

- Dr. Meng Jiang (www.meng-jiang.com)

Carnegie Mellon University — Visiting Ph.D.

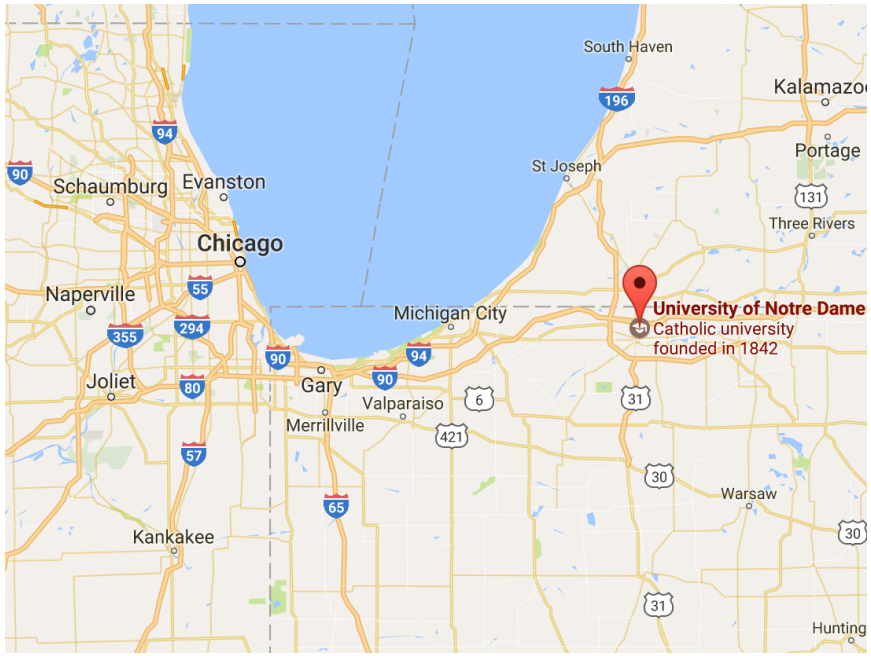
B.S. and Ph.D. — Tsinghua University 清華大學

Postdoc Researcher

Assistant Professor

Visiting Researcher

University of Notre Dame



R.A. Recruiting

- I have competitive research assistantship offers (fully funded) for self-motivated, hard-working, and mature Ph.D. students interested in data science research.
- Being talented is a great plus, but I do not require that. I only expect basic Computer Science backgrounds matching your final degree. You will graduate with lots of talents anyway. :-)

Course Page and Class Schedule

- Syllabus and schedule: https://wiki.illinois.edu/wiki/display/CS412S17/Course+Syllabus+and+Schedule
  - http://www.meng-jiang.com/teaching.html (My homepage)
  - https://wiki.illinois.edu/wiki/pages/viewpage.action?spaceKey=cs412&title=2.+Course+Syllabus+and+Schedule (Fall 2016 by Prof. Jiawei Han)
- Time: 11:00 am – 12:15 pm M/T/R @ 0216 SC
- Office hours: 12:15 pm – 1:00 pm M/R @ 2130 SC
- ONL (online session): http://engineering.illinois.edu/online/current-students/engineering-online-student-portal.html

Teaching Assistants

Sheng Wang
swang141@illinois.edu

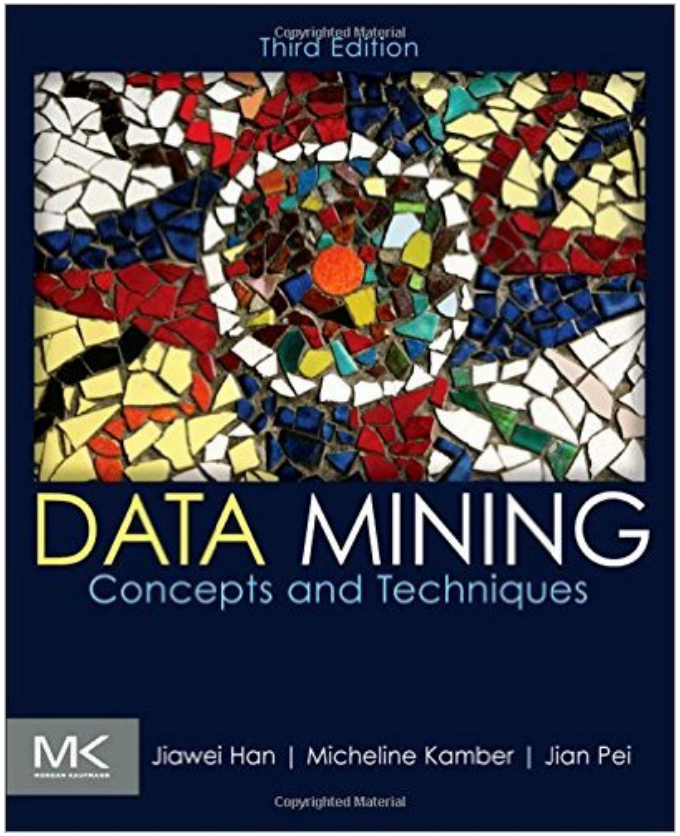
Xuan Wang
xwang174@illinois.edu

- TA office hours: 3 - 4 pm Monday @2113 SC
- Piazza: https://piazza.com/class#summer2017/cs412

https://piazza.com/class/j2pkefesw6u67z (same as above)

Textbook

- Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques (3rd ed), Morgan Kaufmann, 2011



Course Work and Grading

- Assignments and Exams
  - Written assign.: 15% (three expected)
  - Programming assign.: 20% (two expected)
  - Midterm exam: 30%
  - Final exam: 35%
- For students taking 4th credit
  - Concrete instructions on the project next week
- Piazza: https://piazza.com/class/j2pkefesw6u67z
- Check your homework/exam scores: Compass

Why Data Mining?

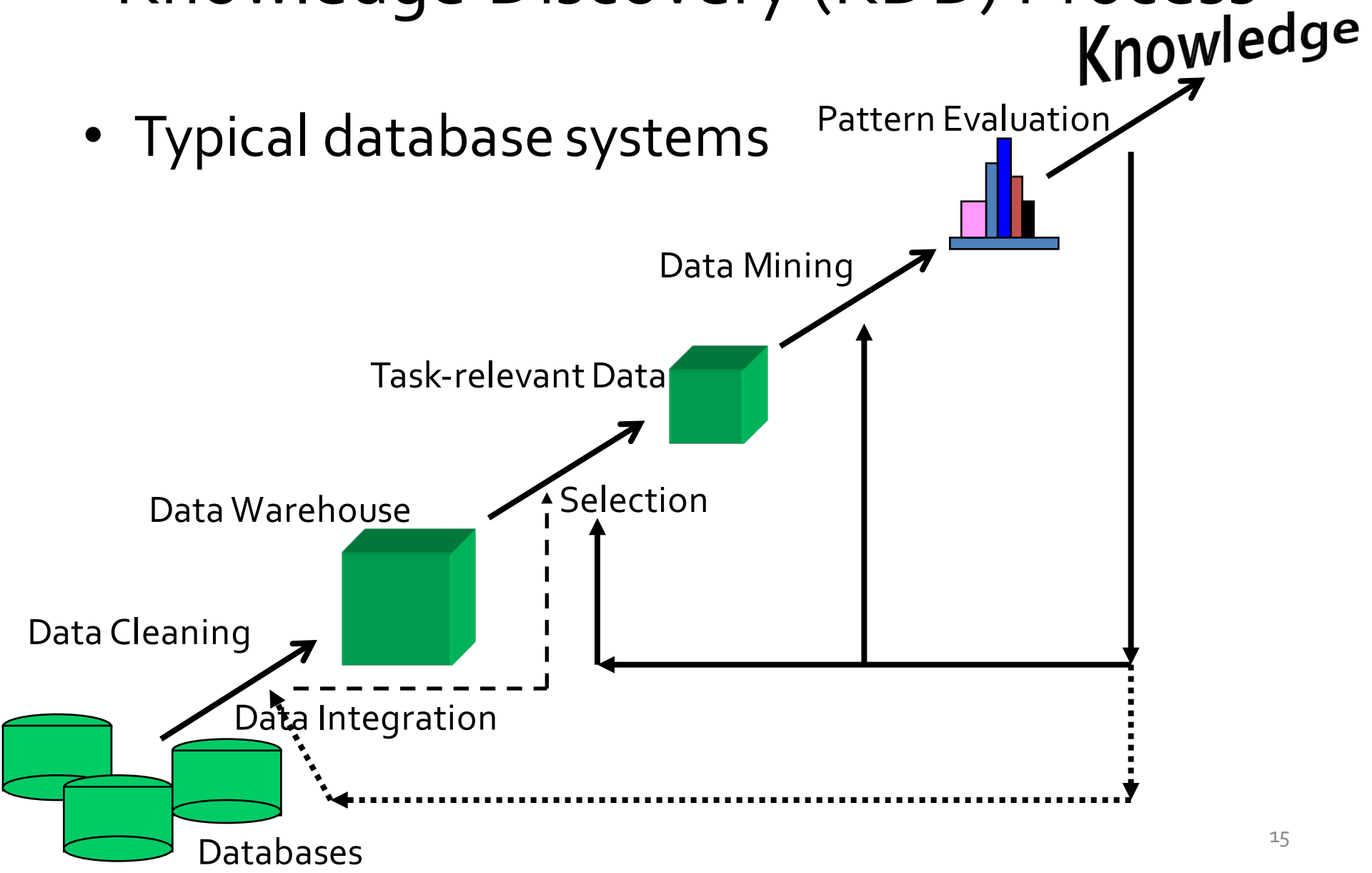
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!

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
- 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"?

Knowledge Discovery (KDD) Process

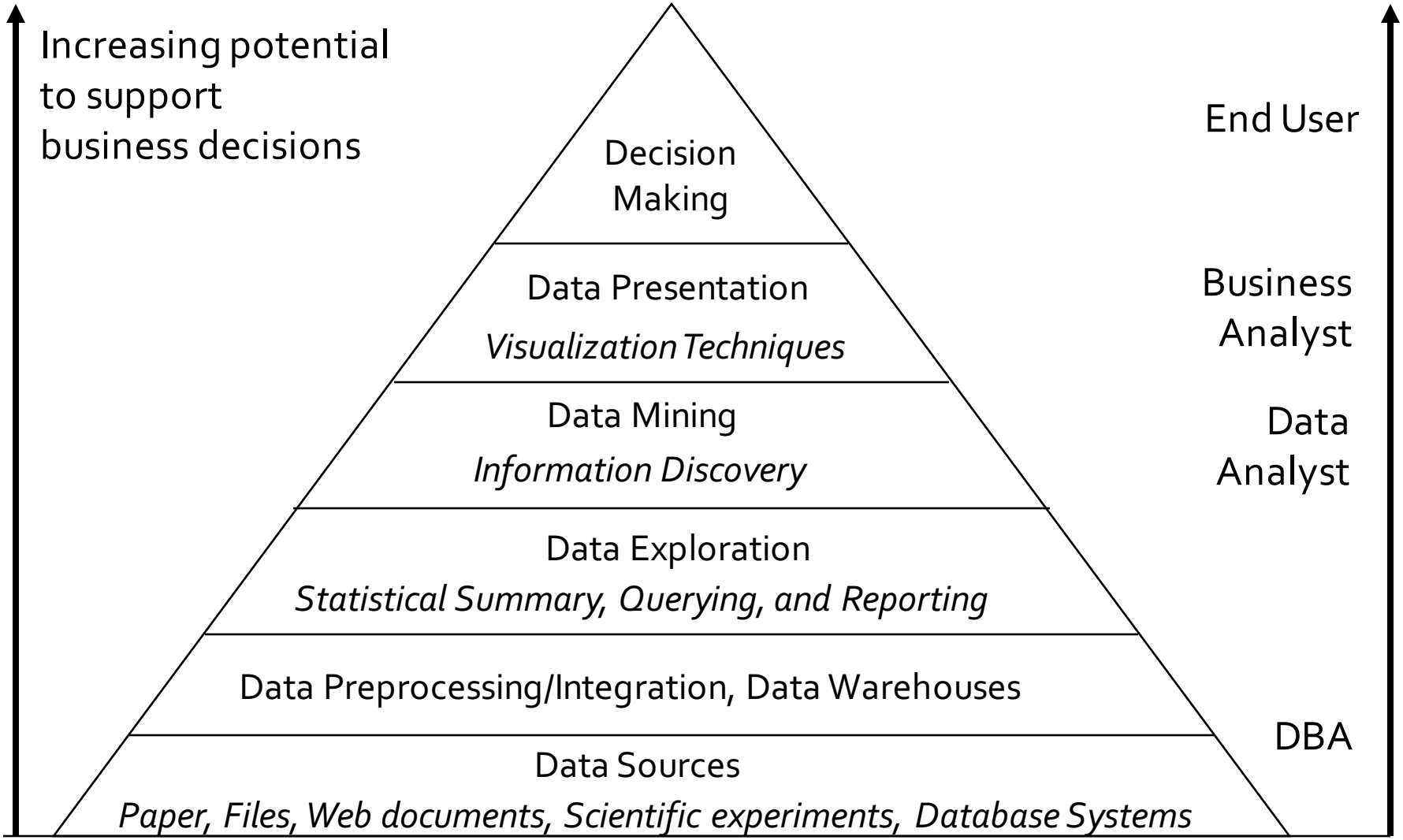
- Typical database systems



Example: Web Mining

- 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

Example: Business Intelligence

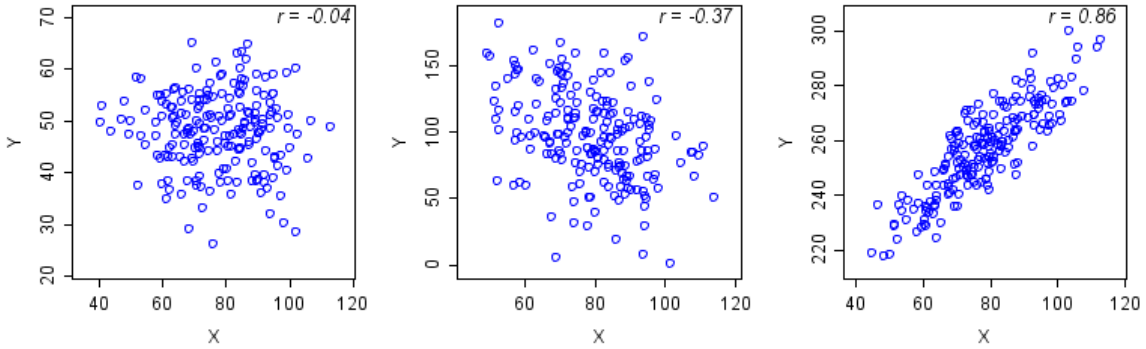


Data Mining Functions: (1) Generalization

- Information integration and data warehouse construction
- Data cube technology
- Multidimensional concept description: Characterization and discrimination

Data Mining Functions: (2) Pattern Discovery

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- A typical association rule
  - Diaper → Beer [0.5%, 75%] (support, confidence)
  - **Support:** the proportion of transactions in the dataset which contains the itemset (Diaper, Beer).
  - **Confidence:** the proportion of the transactions that contains Diaper which also contains Beer.
- **Association** and **Correlation** Analysis

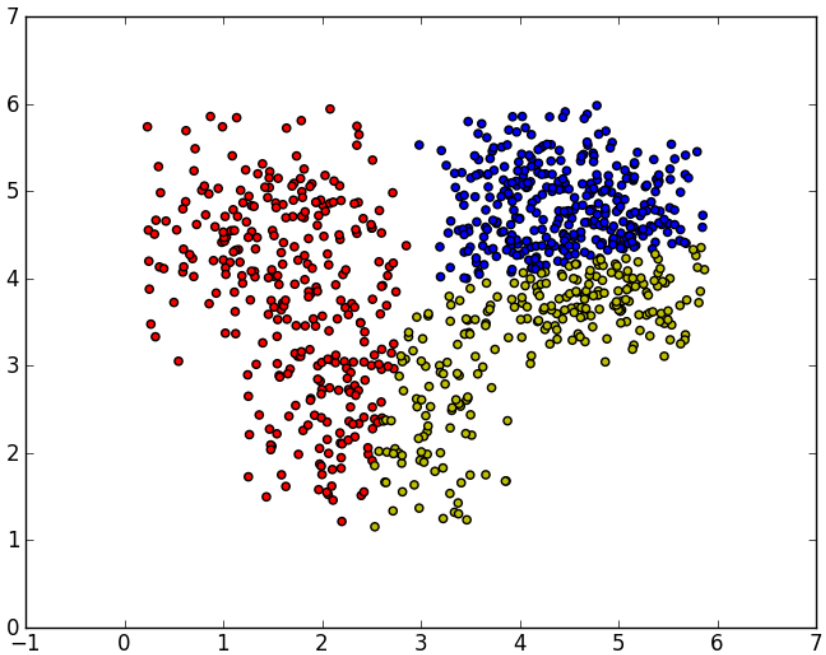


Data Mining Functions: (3) Classification

- 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, …

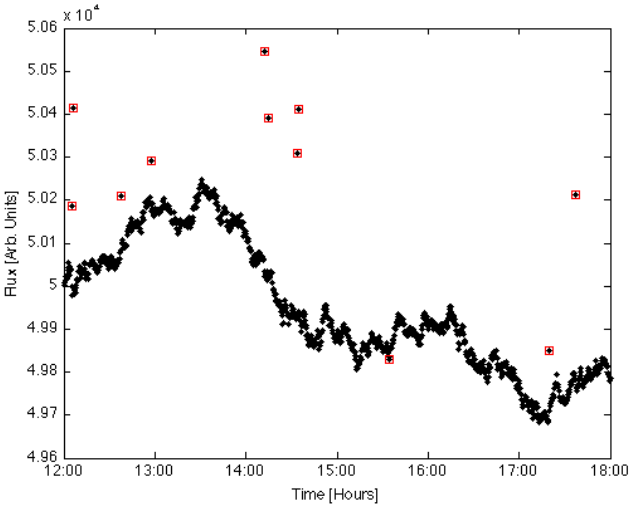
Data Mining Functions: (4) Clustering

- 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



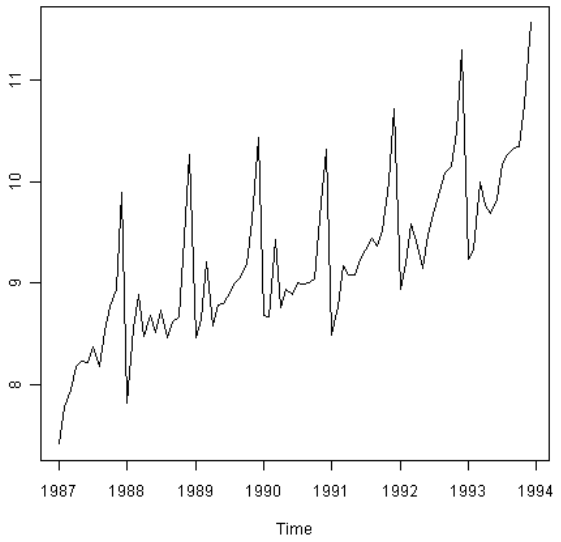
Data Mining Functions: (5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Methods: by product of clustering or regression analysis…
  - Useful in fraud detection, rare events analysis



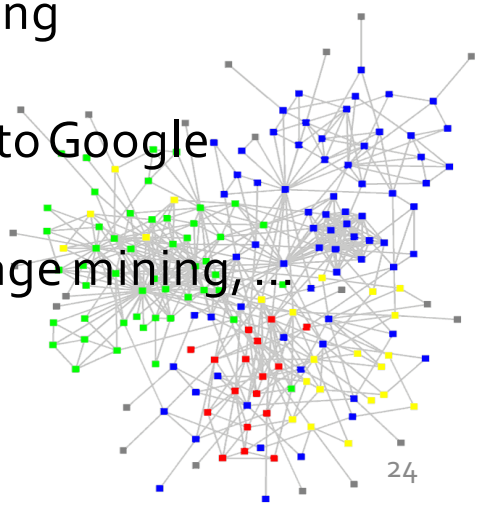
Data Mining Functions: (6) Sequential Pattern, Evolution Analysis

- 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



Data Mining Functions: (7) 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, …
  - 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
    - Web community discovery, opinion mining, usage mining, …



Syllabus

- Intro (1)
- Know Your Data (2): Assign. 1
- **Data Preprocessing** (2.5)
- **Data Warehousing** & OLAP (1.5): Assign. 2
- **Data Cube** Tech (2)
- Mining **Frequent Patterns and Associations** (3+5): Assign. 3 and Midterm Exam
- **Classification** (3.5+3.5): Assign. 4 and Assign. 5
- **Cluster** Analysis (4): Final Exam

Discussion

Choose one company:
Amazon, Yahoo!, Walmart, Google, Facebook, Twitter, Snapchat, Bloomberg, CNN…

Answer:
Given what data,

- What to find?
- Why?
- How?

Applications

- 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

Multidimensional 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.

Amazon, Yahoo!, Walmart, Google, Facebook, Twitter, Snapchat, Bloomberg, CNN…

Research Challenges

- 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

Research Challenges (cont.)

- 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

History

- 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)

Venues

- Data mining and KDD (SIGKDD)
  - Conferences: ACM SIGKDD, IEEE ICDM, SIAM DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, IEEE TKDE, ACM TKDD
- Database systems (SIGMOD)
  - Conferences: ACM SIGMOD, ACM-PODS, VLDB, IEEE ICDE, EDBT, ICDT, DASFAA
  - Journals: ACM TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: ICML, AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, IEEE PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: 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.

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
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, …
- Data mining technologies and applications
- Research challenges in data mining

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

What you can expect from this course

Fundamental data warehousing and data mining theories

Basic concepts and methods for mining datasets

Not included:

State-of-the-art machine learning/artificial intelligence algorithms

Full coverage of specific skills that your start-up ideas require

References

- 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

Discussion

- Given a **task** – design a machine that can find the only outlier from a set of concepts, …
  - Q: apple, banana, eggplant, strawberry
  - A: eggplant
  - How?

Discussion

- Given self-introductory one-minute one-page text **data** of your classmates, …
  - Q:?
  - A:?
  - How?

Write down your self-intro and what you want to do