



Intro. to Data Mining

Chapter 1. Introduction

Dr. Chengcui Zhang (czhang02@uab.edu)

Dept. of Computer Science

Univ. Alabama at Birmingham

1

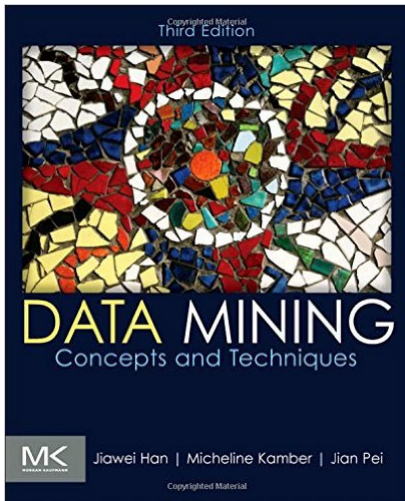
1



2

2

Course Page & Class Schedule



- Textbook
 - Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques (3rd ed)*, Morgan Kaufmann, 2011
- Class Homepage: canvas
- My office hours: Monday 1-3pm via Zoom (or by appointment)
- Class attendance is critical
- TAs and office hours: see syllabus

3

3

Course Work and Grading (see syllabus)

- Assignments, Programming Assignments, Attendance and Exams
 - Assignments (around 5): 50% for CS 463/663 students and 40% for 763 students
 - There will be additional requirements for 663/763 students.
 - Lecture attendance: 10% (**≥ 125 minutes out of 150 minutes/lecture**)
 - Midterm exam: 20%
 - Term Group Project and Presentation: 20% (bonus pts available!)
 - At least 4 students / group! No more than 5 students/ group.
 - Term Research Paper presentation: 10% for 763 students
 - **Need help and/or discussions?**
 - **Post your questions to the 'Post Your Questions here' forum on Canvas**
- Check your homework/exam scores:
 - Canvas
- Class communications: **via Inbox in Canvas!**

4

4

Assignment Submission

- ❑ Assignments are due at 11:59pm on the due date. Late submissions are subject to penalty at 20%. The assignment submission will close 2 days (48 hours) after the due date.
- ❑ **NO** late submission of the final project and the final assignment!
- ❑ **All assignments must be turned in even if they are late. Failure to submit any assignment will result in a grade of F.**
- ❑ Each student can request a **one-time only** waiver of a late penalty, but he/she must submit the assignment before the assignment is closed.
- ❑ All assignments and projects should be submitted to Canvas.
- ❑ **NO** handwritten submissions will be accepted/graded!

5

5

Class Attendance

- ❑ Attendance is mandatory for the lecture portion of this course. Students will start receiving penalty for each absence beyond the 2nd unexcused absence.
- ❑ **No makeup exams are possible.** If you cannot be present for the examinations, you should not take this course.
- ❑ UAB calendar and withdrawal deadlines:
<https://www.uab.edu/students/academics/academic-calendar>

6

6

- ❑ In case you have a medical/family emergency and need more than a usual extension (e.g., 48 hours) to complete your assignment, try your best to let the instructor know BEFORE the deadline, NOT after. Explain the situation and submit appropriate documents afterwards (e.g., doctor's note.) The one-time late penalty waiver is exactly for this purpose (please read the syllabus) but if you need a bit more time on that assignment, it MAY be granted based on the actual situation. If you expect to request more than one such extension because of the medical/family emergency, you may want to consider 'Incomplete' option (if you have completed >80% assignments and exams) or withdrawal, while the former option requires proper documents.
- ❑ What cannot be accommodated?
 - ❑ Inform the instructor after 48 hours of the due time!
 - ❑ Travel for vacation or visiting a friend/family member!
 - ❑ ...

7

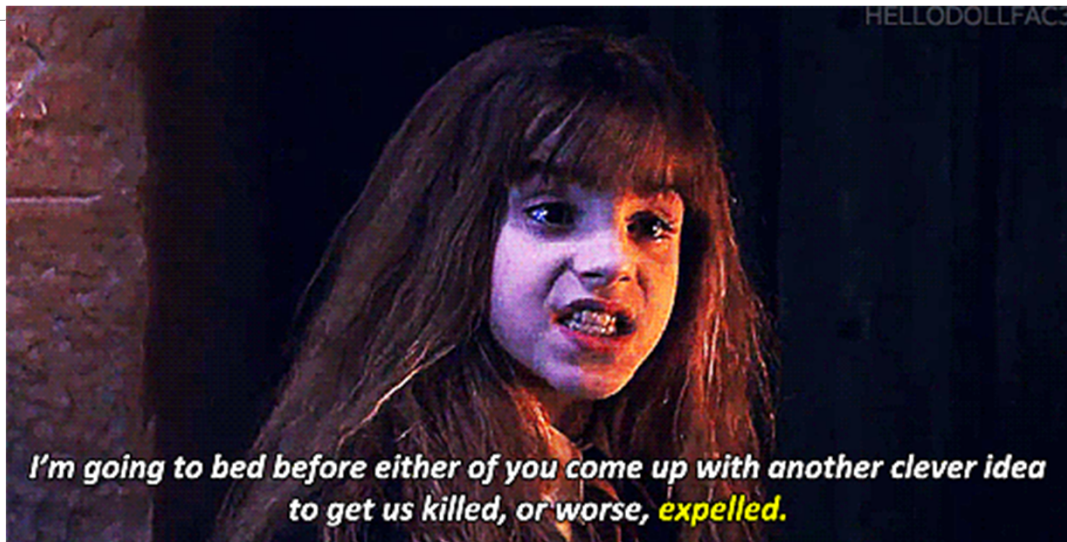
7

Academic Honesty

- ❑ Students who plagiarize a computer program (or parts of a program), get others to write a program (or parts of a program), collaborating on an assignment/exam that is supposed to be completed individually, or are found cheating on a quiz/exam, will be reported for academic dishonesty.
- ❑ Anyone who is caught cheating will receive a 0 on a given test or assignment. If a second offense occurs, the student will receive an F in the class. This includes both the **provider** of the information as well as the **receiver** of the information.
- ❑ Any student who violates the university's code of integrity will be reported for academic discipline. Further, you will no longer be eligible for any student assistantship (e.g., TA)!
- ❑ Cheggs!

8

8



9

9

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

10

10

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 (clicks and navigation), 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

11

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

12

12

If you torture the data long enough,
it will confess.
Ronald Coase



online-behavior.com

13

13

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”?
 - Simple search and query processing
 - (Deductive) expert systems

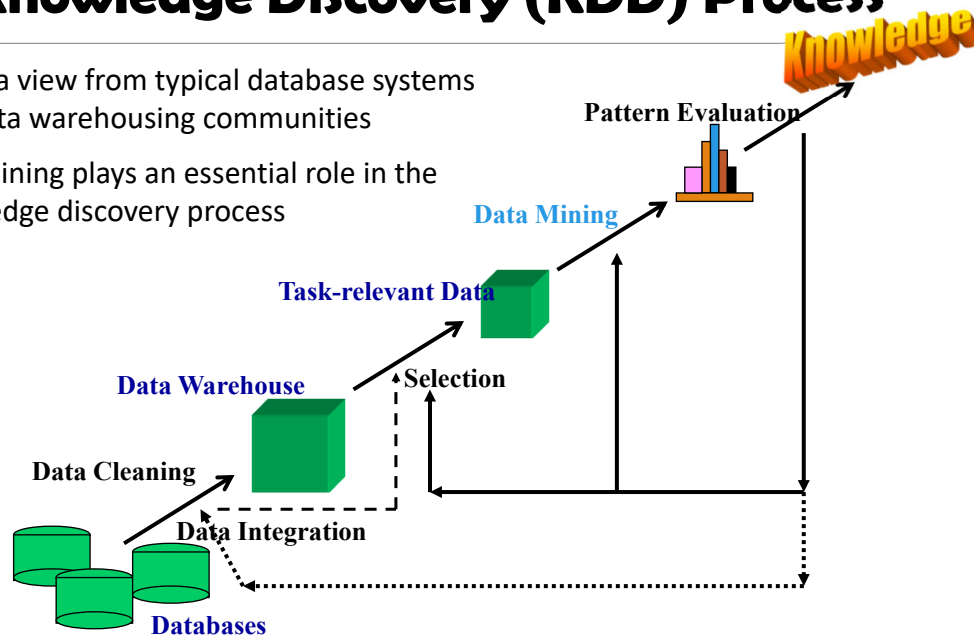


14

14

Knowledge Discovery (KDD) Process

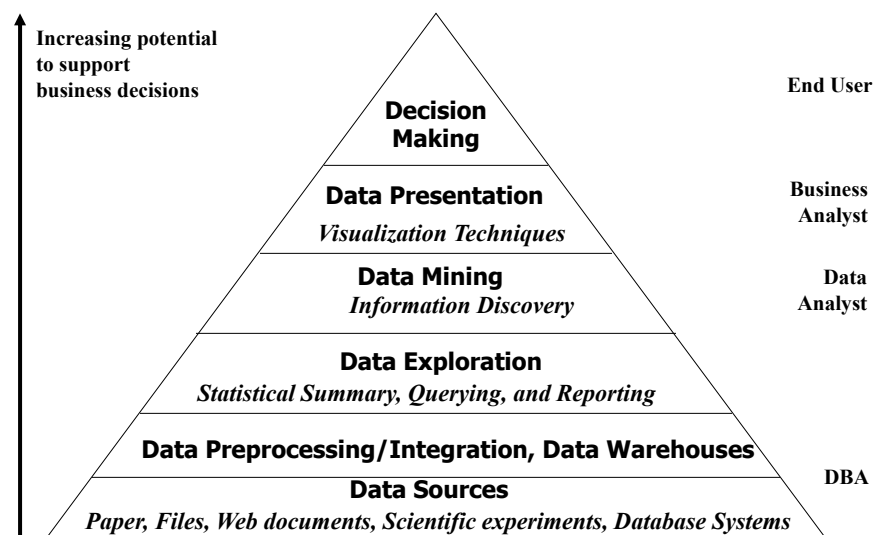
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



15

15

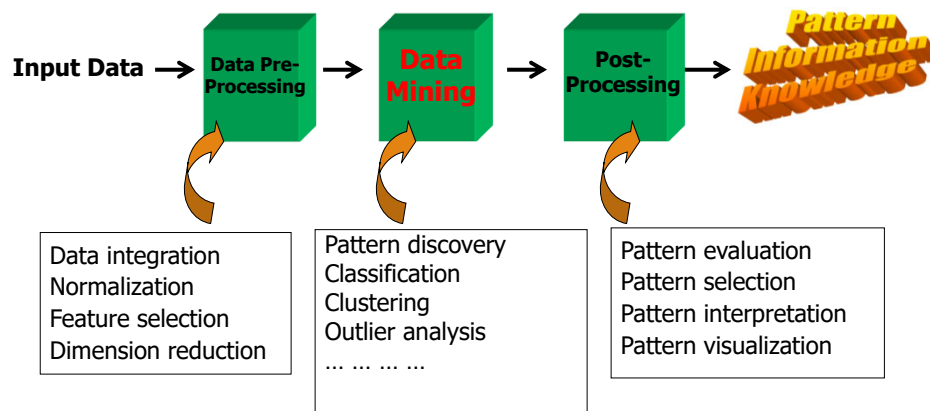
Data Mining in Business Intelligence



17

17

KDD Process: A View from ML and Statistics



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

18

18

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

19

19

Data Mining vs Statistics	
Data Mining	Statistics
Explorative – Dig out the data first, discover novel patterns and then make theories.	Confirmative – Provide theory first and then test it using various statistical tools.
Involves Data Cleaning	Statistical methods applied on Clean Data
Usually involves working with large datasets	Usually involves working with small datasets
Makes generous use of heuristics think	There is no scope for heuristics think
Inductive process	Deductive (Does not involve making any predictions)
Numeric and Non-Numeric Data	Numeric Data
Less concerned about data collection.	More concerned about data collection.
Some of the popular data mining methods include – Estimation, Classification, Neural Networks, Clustering, Association, and Visualization.	Some of the popular statistical methods include – Inferential and Descriptive Statistics.
https://www.dezyre.com/article/data-mining-vs-statistics-vs-machine-learning/349	

20

20

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

21

21

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.

22

22

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

23

23


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

24

24

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

25

25

Data Mining Functions: (1) Generalization

- Information integration and data warehouse construction
 - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
 - Scalable methods for computing (i.e., materializing) multidimensional aggregates
- Multidimensional concept description:
 - Characterization and discrimination
 - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region, honest sellers vs not so honest ones

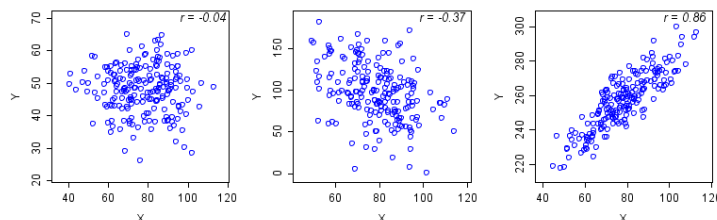


26

26

Data Mining Functions: (2) Pattern Discovery

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



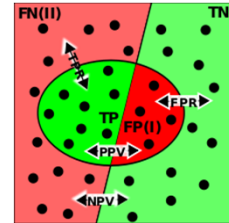
- A typical association rule
 - Diaper \rightarrow Beer [0.5%, 75%] (support, confidence)
 - Are strongly associated items also strongly correlated? **Yes!**
- How to mine such patterns and rules efficiently in large datasets? Complexity issues?
- How to use such patterns for classification, clustering,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

27

27

Data Mining Functions: (3) Classification

- Classification and label prediction
 - Describe and distinguish classes or concepts for future prediction
 - Ex. 1. Classify countries based on (climate)
 - Ex. 2. Classify cars based on (type)
 - Construct models (functions) based on some training samples
 - Predict some unknown class labels
- Typical methods
 - **Decision trees, naïve Bayesian classification, rule-based classification,** support vector machines, neural networks, pattern-based classification, logistic regression*, ...
- Typical applications:
 - Credit card fraud detection, direct marketing, object classification, diseases, web-pages, ...

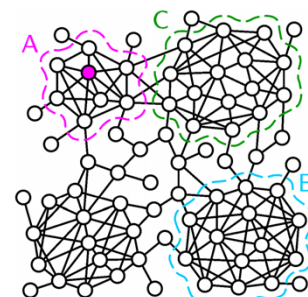
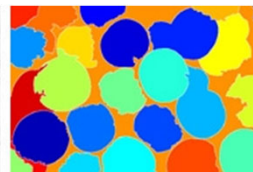
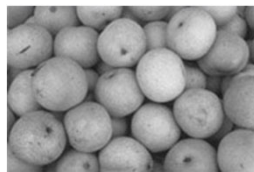
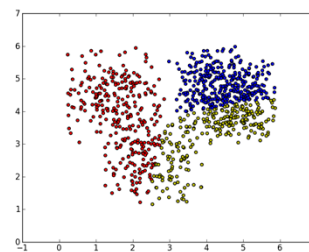


28

28

Data Mining Functions: (4) 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 spatial distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications



29

29

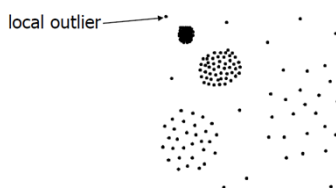
- Mid-term exam: mid March during class time (~1.5 hrs)
- Term group project formation
 - 4-5 students per group
 - You need to send me your group members by **Feb. 18th**. After that if you are still ungrouped, you will be assigned randomly to a group that still has <5 members or you will be grouped with the remaining unassigned students.

30

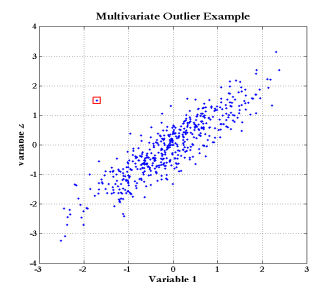
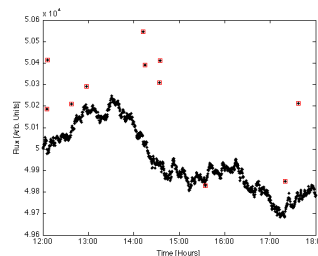
30

Data Mining Functions: (5) 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



global outliers

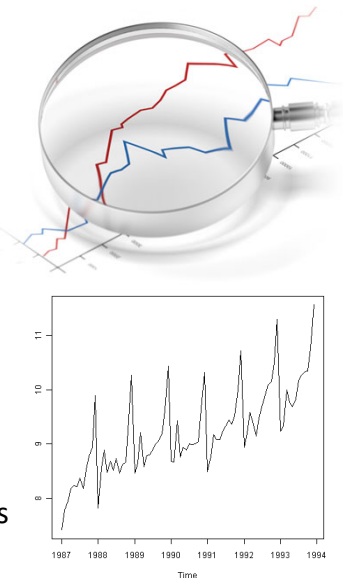


31

31

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

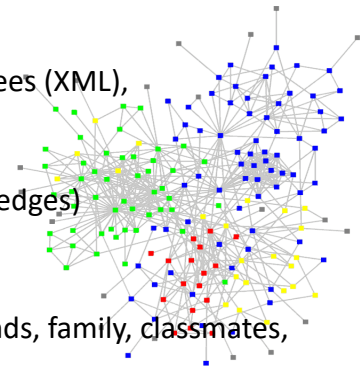


32

32

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 on 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, ...



33

33

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
 - ❑ ...



34

34

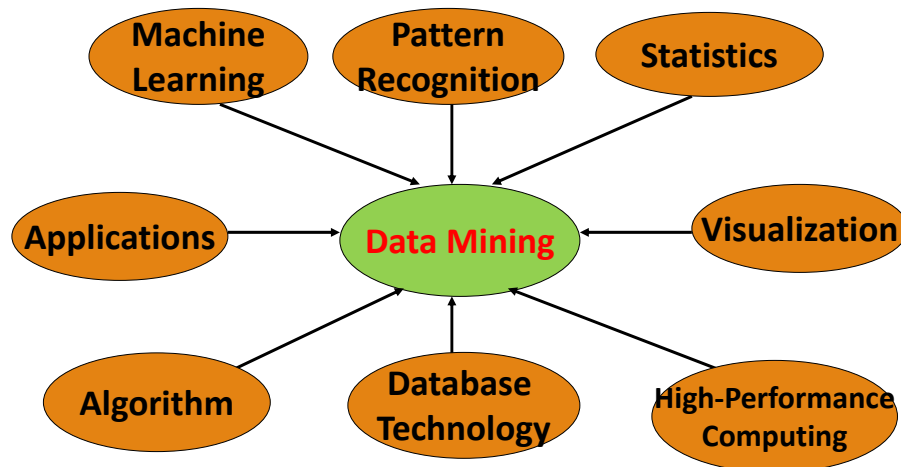
Chapter 1. Introduction

- ❑ 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

35

Data Mining: Confluence of Multiple Disciplines



36

36


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

37

37

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

38

38

Applications of Data Mining


- ❑ Web page analysis: classification, clustering, ranking
- ❑ Collaborative analysis & recommender systems
- ❑ Shopping cart data analysis for 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!, I
- ❑ **Major dedicated data mining systems/tools**
 - ❑ **Matlab, Python scikitlearn** (<https://scikit-learn.org/stable/>), SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools, R, etc.
 - ❑ UAB student software: https://uabprod.service-now.com/service_portal?id=sc_category&sys_id=3fc9e61437d34e8024a67c1643990ebd



39

39

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

40

40

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 <https://www.youtube.com/watch?v=CdQic2Xtegg>
 - ❑ Incorporation of background/domain knowledge
 - ❑ **Presentation and visualization of data mining results**

41

41

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 and ethical impacts of data mining
 - ❑ Privacy-preserving data mining
 - ❑ Invisible data mining
 - ❑ <https://www.quora.com/What-are-ethical-issues-in-data-mining>

42

42

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



43

43

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)

44

44

Conferences and Journals on Data Mining

- | | |
|---|--|
| <ul style="list-style-type: none"> ❑ KDD Conferences <ul style="list-style-type: none"> ❑ 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) | <ul style="list-style-type: none"> ■ Other related conferences <ul style="list-style-type: none"> ■ DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ... ■ Web and IR conferences: WWW, SIGIR, WSDM ■ ML conferences: ICML, NIPS ■ PR conferences: CVPR, ICCV ■ Journals <ul style="list-style-type: none"> ■ Data Mining and Knowledge Discovery (DAMI or DMKD) ■ IEEE Trans. On Knowledge and Data Eng. (TKDE) ■ KDD Explorations ■ ACM Trans. on KDD |
|---|--|

45

45

Where to Find References? DBLP, CiteSeer, Google, and UAB Stern Library (IEEE/ACM digital libraries)!!!

- ❑ 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, JIS, 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.

46

46


CS763 student research paper presentations

- ❑ April 20th (Last day of this class!)
 - ❑ ~15-20 minutes **each** followed by 5 minutes of Q&A)
 - ❑ Attendance will be taken
 - ❑ *You can get bonus points if you ask GOOD questions! (for 463 and 663 students only.)*
- ❑ CS 763 students
 - ❑ I need to receive the research paper you want to present by the end of Feb.
 - ❑ The paper cannot be more than 5 years old.
 - ❑ I need to receive your slides no later than the night of April 19th.

47

47

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 

48

48

Summary

- ❑ **Data mining:** Discovering interesting, previously unknown 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, prediction, etc.
- ❑ Data mining technologies and applications
- ❑ Major issues in data mining

49

49

Recommended Reference Books

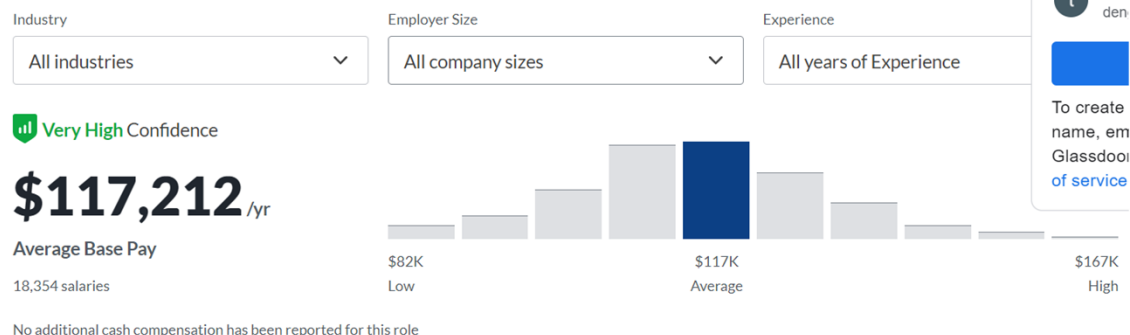
- ❑ Charu C. Aggarwal, *Data Mining: The Textbook*, Springer, 2015
- ❑ E. Alpaydin. *Introduction to Machine Learning*, 2nd ed., MIT Press, 2011
- ❑ R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2ed., Wiley-Interscience, 2000
- ❑ U. Fayyad, G. Grinstein, and A. Wierse, *Information 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

50

50

https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

How much does a Data Scientist make?



How much does a Data Scientist make? The national average salary for a Data Scientist is \$117,212 in United States. Filter by location to see Data Scientist salaries in your area. Salary estimates are based on 18,354 salaries submitted anonymously to Glassdoor by Data Scientist employees.

51

51