Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 1 —

READ
READ
READ

Jiawei Han, Micheline Kamber, and Jian Pei

74

74

Chapter 1. Introduction

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

75

75

Why Data Mining?

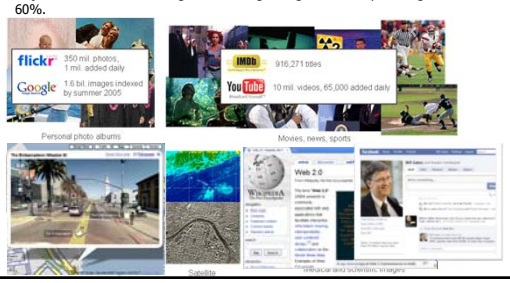
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, **Web**, computerized society, ***mobile devices, social media***
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, microblogs, 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

76

76

Understanding Visual Data on the Web

- Explosions of image/video data: digital photos, personal videos, geospatial imagery, broadcast news/sports videos, Wikipedia, social media, etc.
- IDC estimated that 1200 Exabytes of digital data will be created in 2010, and much is the visual data on the Web
- Learning and mining can unlock the wealth of information contained in the massive visual data generation is growing with a compounding annual rate of 60%.



77

77

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, ***computational*** branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics, or even social science)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations [capture]
  - The ability to economically store and manage petabytes of data online [storage]
  - The Internet and computing Grid that makes all these archives accessible [transmission]
  - HPC and GPUs provide the ability to process massive data [processing]
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!

78

78

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - *Stream* data management and mining
  - Data mining and its *applications*
  - Web technology (XML, data integration) and global information systems
  - Unstructured data!!

79

79

Chapter 1. Introduction

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

80

80

What Is Data Mining?

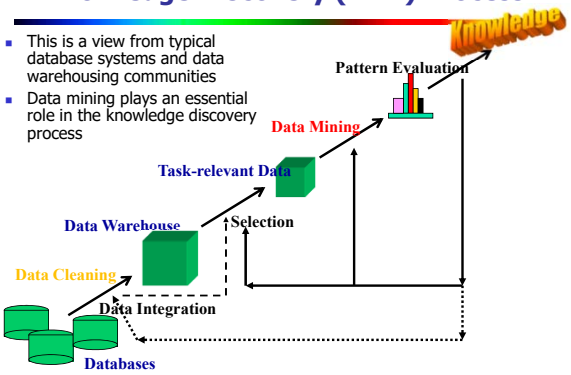
- Data mining (KDD: knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from a huge amount of data
  - Data mining: a misnomer? knowledge mining maybe?
- Alternative names
  - Knowledge discovery (mining) 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

81

81

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



82

82

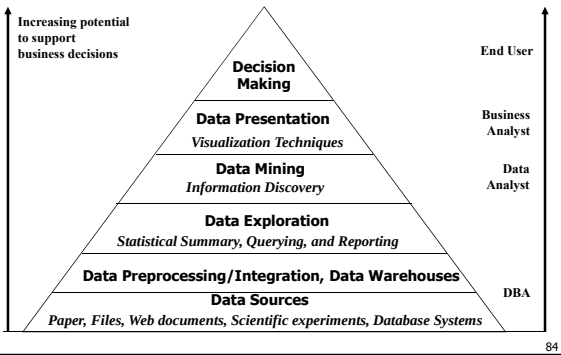
Example: A Web Mining Framework

- 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

83

83

Data Mining in Business Intelligence



84

84

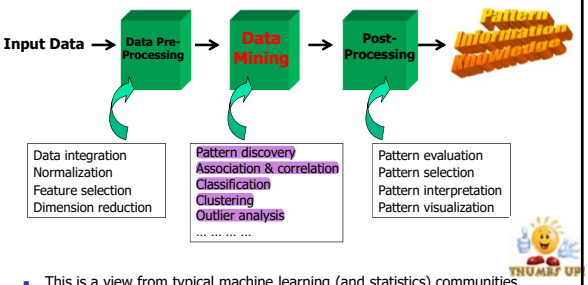
Example: Mining vs. Data Exploration

- Business intelligence view
  - Warehouse, data cube, reporting but not much mining
- Business objects vs. data mining tools
- Supply chain example: tools
- Data presentation
- Exploration

85

85

KDD Process: A View from ML and Statistics



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

86

86

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes, association (genes and diseases)
- Post-processing for presentation

87

87

Chapter 1. Introduction

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

88

88

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy, transactional data); unstructured data (data warehouse, 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, etc.
  - 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 computing, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

89

89

Chapter 1. Introduction

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

90

90

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets (mostly unstructured) and advanced applications
  - Data **streams** and **sensor** data
  - Time-series data, **temporal data, sequence data** (incl. bio-sequences)
  - Structure data**, graphs, social networks and multi-linked data**
  - Object-relational databases
  - **Heterogeneous** databases and legacy databases
  - **Spatial data and spatiotemporal data**
  - **Multimedia data (images, video, audio)**
  - Text databases
  - **The World-Wide Web**

REAL WORLD!

91

91

Chapter 1. Introduction

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

92

92

Data Mining Function: (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
  - OLAP (online analytical processing)
- Multidimensional concept description: *Characterization* and *discrimination*
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

93

93

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association vs. *correlation* vs. causality
  - A typical association rule
    - Diaper → 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?

94

94

Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some *unknown* class labels (based on attributes)
- 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, …

95

95

Data Mining Function: (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 distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

96

96

Data Mining Function: (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 (literally) could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

97

97

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., first buy digital camera, then buy large SD memory cards; tourist trajectories
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

98

98

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, movie networks, terrorist networks,
  - Multiple heterogeneous networks
    - A person could be in multiple information networks: friends, family, classmates, clubs, …
  - 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, …

99

99

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of "patterns" and knowledge
  - 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
  - …

100

100

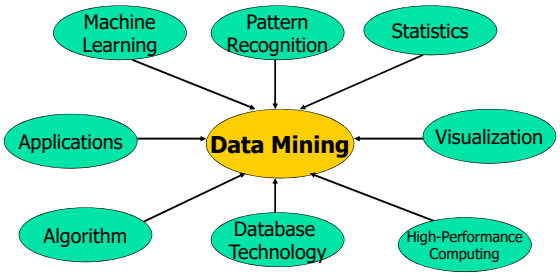
Chapter 1. Introduction

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

101

101

Data Mining: Confluence of Multiple Disciplines



102

102

Statistics, Machine Learning and Data Mining

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine Learning
  - Can be more heuristics than theory-based
  - Focused on improving performance of learning algorithms
- Data Mining and Knowledge Discovery
  - Data Mining is one step in the Knowledge Discovery process (applying the Machine Learning algorithms)
  - Knowledge Discovery, the whole process including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

103

103

Evolution of Machine Learning

- Classic: rule-based classification, MLE (Bayes), linear regression, decision trees, KNN, K-means, perceptron
- 1980s: ANN, Genetic algorithms, Fuzzy logic
- 1990s: SVM, Bayes networks
- 2000s: AdaBoost, kernel methods, random forests, sparse representation
- 2010s: Deep learning (or NN strikes back)

  "The goal of machine learning is to build computer systems that can adapt and learn from their experience." - Tom Dietterich

104

Historical Perspective

The Genesis (1950)
The First Winter (1974-1984)
IBM DeepBlue Triumphs (1997)
The Second Winter (1987-1993)
DARPA Grand Challenge (2005)
Watson wins "Jeopardy" (2011)
AlphaGo Triumphs (2016)
ADAS in Telsa (2016)
Microsoft Chatbot (2016)

105

Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of 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 networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

107

Chapter 1. Introduction

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

108

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS* algorithms
- Collaborative analysis & recommender systems
- Shopping basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

It's never too early to start thinking about your projects .....

109

Chapter 1. Introduction

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

110

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
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
  - Boosting the power of discovery in a networked environment
- User Interaction
  - Interactive mining
  - Incorporation of background/domain knowledge
  - Presentation and visualization of data mining results

federated learning
distributed learning

111

111

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 – do good, or do evil?
  - Privacy-preserving data mining
  - Invisible data mining

112

112

Chapter 1. Introduction

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

113

113

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), etc.
- ACM **Transactions** on KDD starting in 2007

114

114

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)
  - AAAI International Conference on Weblogs and Social Media (ICWSM)
- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR
  - Multimedia conferences: MM
- **Journals**
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

115

115

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - 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, MM, 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.

116

116

Chapter 1. Introduction

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

Summary

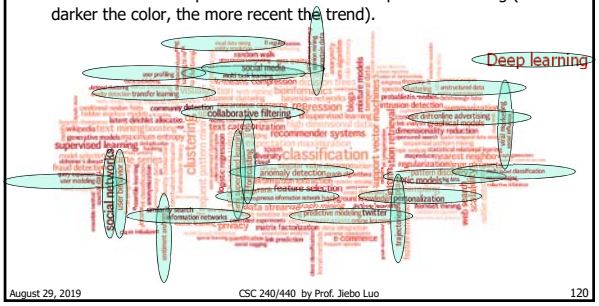
- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of database technology, in great demand, with broad applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

The Evolving Domain of KDD

- KDD community is **dynamic** in that it is quick to identify and adopt shifts in domains of interests. Check out the tag cloud of KDD research within the past decade to see how topics are evolving (The darker the color, the more recent the trend).

The Evolving Domain of KDD

- KDD community is dynamic in that it is quick to identify and adopt shifts in domains of interests. Check out the tag cloud of KDD research within the past decade to see how topics are evolving (The darker the color, the more recent the trend).



Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

Homework Assignment #1

- Part I: Textbook problems
  - 1.1, 1.2, 1.4, 1.5, 1.7, 1.9
- Part II: Playing with data
  - Decide on your programming language
  - Select a small dataset (e.g., Iris from the UCI repository)
  - Compute some statistics or plot the data, either with your own code or a toolbox (this time only)
  - Interpret the statistics or plots
- Bonus points: up to 1 pt (for a total of 6 pts)
- **report + code** Due 9/12 (one week from today)
  - Upload two files: one for the report, one for the code