Chapter 1. Introduction

- Motivation: Why data mining?

- What is data mining?

- Data Mining: On what kind of data?

- Data mining functionality

- Classification of data mining systems

- Top-10 most popular data mining algorithms

- Major issues in data mining

- Overview of the course

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

- <u>We are drowning in data, but starving for knowledge!</u>

- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

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.)
  - 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
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

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

What Is Data Mining?
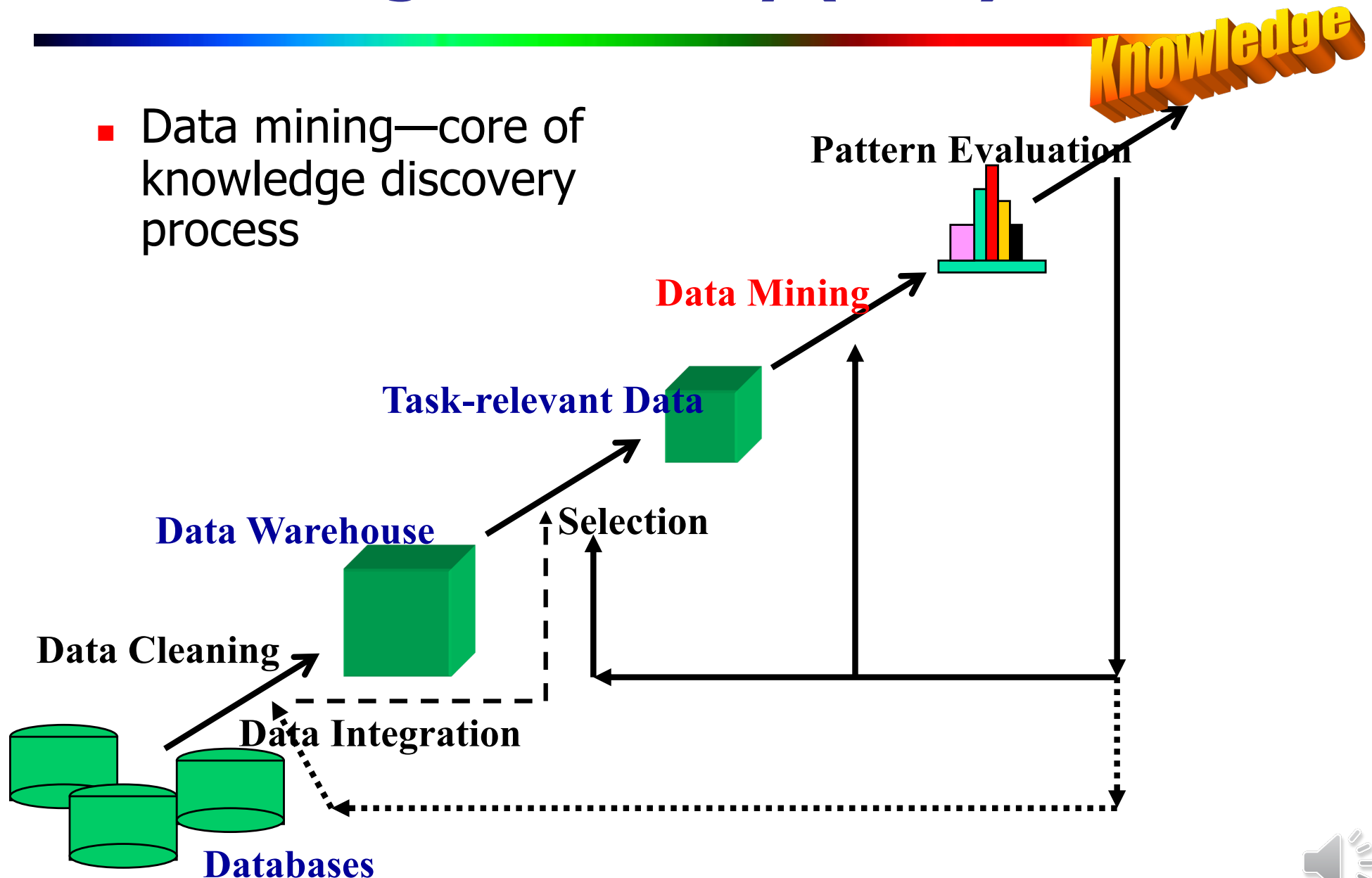
- Data mining (knowledge discovery from data)
  - Extraction of interesting (<u>non-trivial,</u> <u>implicit,</u> <u>previously</u> <u>unknown</u> and <u>potentially useful)</u> patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) 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
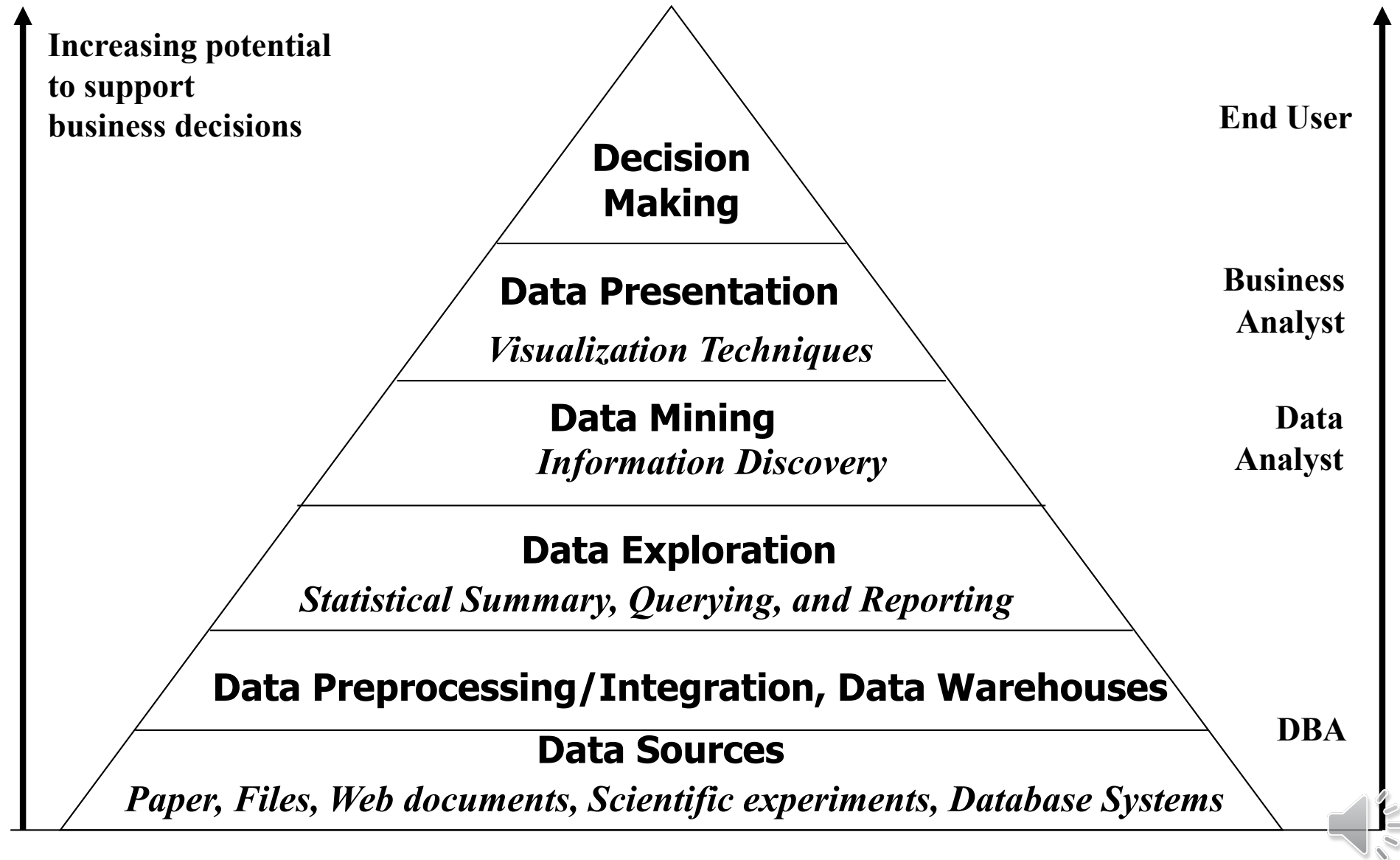 - (Deductive) expert systems

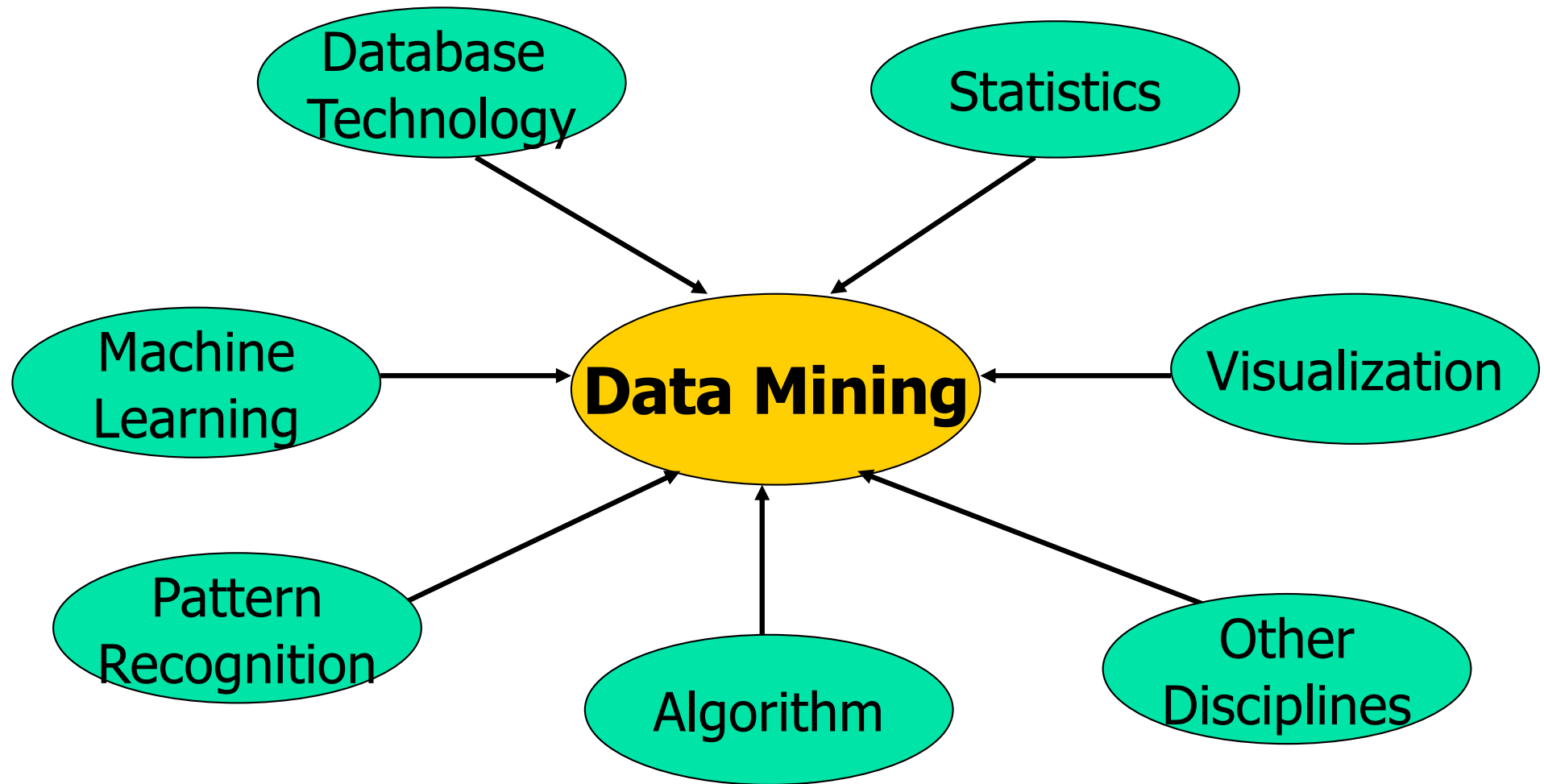# Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Data Warehouse

Selection

Data Cleaning

Data Integration

Databases

Data Mining and Business Intelligence

Increasing potential to support business decisions

Decision Making

End User

Data Presentation

Visualization Techniques

Business Analyst

Data Mining

Information Discovery

Data Analyst

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

Paper, Files, Web documents, Scientific experiments, Database Systems

DBA

Data Mining: Confluence of Multiple Disciplines

Data Mining: Concepts and Techniques

Why Not Traditional Data Analysis?

- 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

Data Mining: Concepts and Techniques

Data Mining: Classification Schemes

- General classification

  - Descriptive data mining

  - Predictive data mining

- Different views lead to different classifications

  - Data view: Kinds of data to be mined

  - Knowledge view: Kinds of knowledge to be discovered

  - Method view: Kinds of techniques utilized

  - Application view: Kinds of applications adapted

Multi-Dimensional View of Data Mining

- **Data to be mined**
    - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW

- **Knowledge to be mined**
    - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
    - Multiple/integrated functions and mining at multiple levels

- **Techniques utilized**
    - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.

- **Applications adapted**
    - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data for Mining

- Database-oriented data sets and applications

    - Relational database, data warehouse, transactional database

- 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 multi-linked data

    - Object-relational databases

    - Heterogeneous databases and legacy databases

    - Spatial data and spatiotemporal data

    - Multimedia database

    - Text databases

    - The World-Wide Web

Functionalities for Data Mining

- Multidimensional concept description: Characterization and discrimination
    - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation vs. causality
    - Diaper → Beer [0.5%, 75%] (Correlation or causality?)
- Classification and prediction
    - Construct models (functions) that 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 or missing numerical values

Data Mining: Concepts and Techniques

Functionalities for Data Mining (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis

Top-10 Most Popular DM Algorithms: 18 Identified Candidates (I)

- Classification
  - #1. C4.5: Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann., 1993.
  - #2. CART: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984.
  - #3. K Nearest Neighbours (kNN): Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. TPAMI. 18(6)
  - #4. Naive Bayes Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- Statistical Learning
  - #5. SVM: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
  - #6. EM: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York. Association Analysis
  - #7. Apriori: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
  - #8. FP-Tree: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.

The 18 Identified Candidates (II)

- Link Mining
  - #9. PageRank: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
  - #10. HITS: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. SODA, 1998.
- Clustering
  - #11. K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
  - #12. BIRCH: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
- Bagging and Boosting
  - #13. AdaBoost: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.

The 18 Identified Candidates (III)

- Sequential Patterns
  - #14. GSP: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
  - #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- Integrated Mining
  - #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- Rough Sets
  - #17. Finding reduct: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- Graph Mining
  - #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

Top-10 Algorithm Finally Selected at ICDM'06

- **#1: C4.5 (61 votes)**
- **#2: K-Means (60 votes)**
- **#3: SVM (58 votes)**
- **#4: Apriori (52 votes)**
- **#5: EM (48 votes)**
- **#6: PageRank (46 votes)**
- **#7: AdaBoost (45 votes)**
- **#7: kNN (45 votes)**
- **#7: Naive Bayes (45 votes)**
- **#10: CART (34 votes)**

Major Issues in Data Mining

- <u>Mining methodology</u>
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- <u>User interaction</u>
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- <u>Applications and social impacts</u>
  - Domain-specific data mining
  - Protection of data security, integrity, and privacy

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

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)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)

- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Data Mining: Concepts and Techniques

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, 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.

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, 2nd ed., 2006

- 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, Springer-Verlag, 2001

- 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

Summary

- Data mining: Discovering interesting patterns from large amounts of data

- A natural evolution of database 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 information repositories

- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.

- Data mining systems and architectures

- Major issues in data mining