Slides related to:

Data Mining: Concepts and Techniques

— Chapter 1 and 2 —

— Introduction and Data preprocessing —

Jiawei Han and Micheline Kamber

Department of Computer Science

University of Illinois at Urbana-Champaign

www.cs.uiuc.edu/~hanj

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!
- → Data mining—Automated analysis of massive data sets

Ex. 1: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, lifestyle studies, …
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, …
  - Determine customer purchasing patterns over time
- Customer profiling
  - What types of customers buy what products
- Cross-market analysis
  - Find associations/co-relations between product sales
  - Predict based on such associations
- Customer requirement analysis
  - Identify the best products for different groups of customers
  - Predict what factors will attract new customers

Ex. 2: Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications:
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests

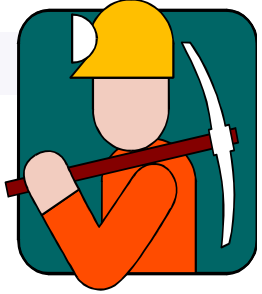
Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - Advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, temporal, multimedia, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases

Evolution of Database Technology

- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems
- 2010s
  - Big data (Volume, Velocity, Veracity, Variety, Variability, …)
  - NoSQL databases, graph databases
  - Knowledge graphs

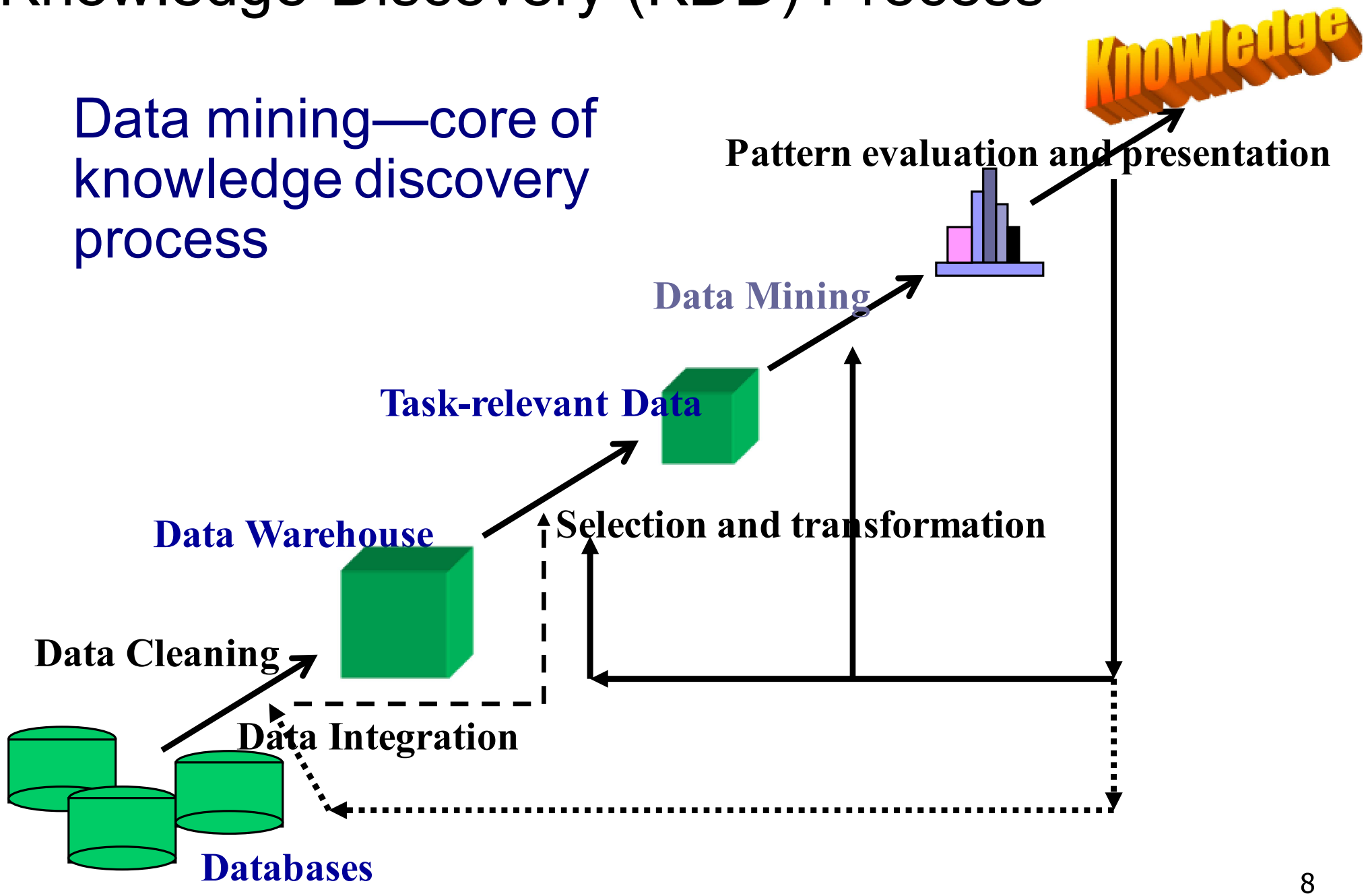
What Is Data Mining?



- 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
  - 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, ...
- Watch out: Is everything "data mining"?
  - Not: Simple search and query processing
  - Not: (Deductive) expert systems

Knowledge Discovery (KDD) Process

Data mining—core of knowledge discovery process



Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthdate=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

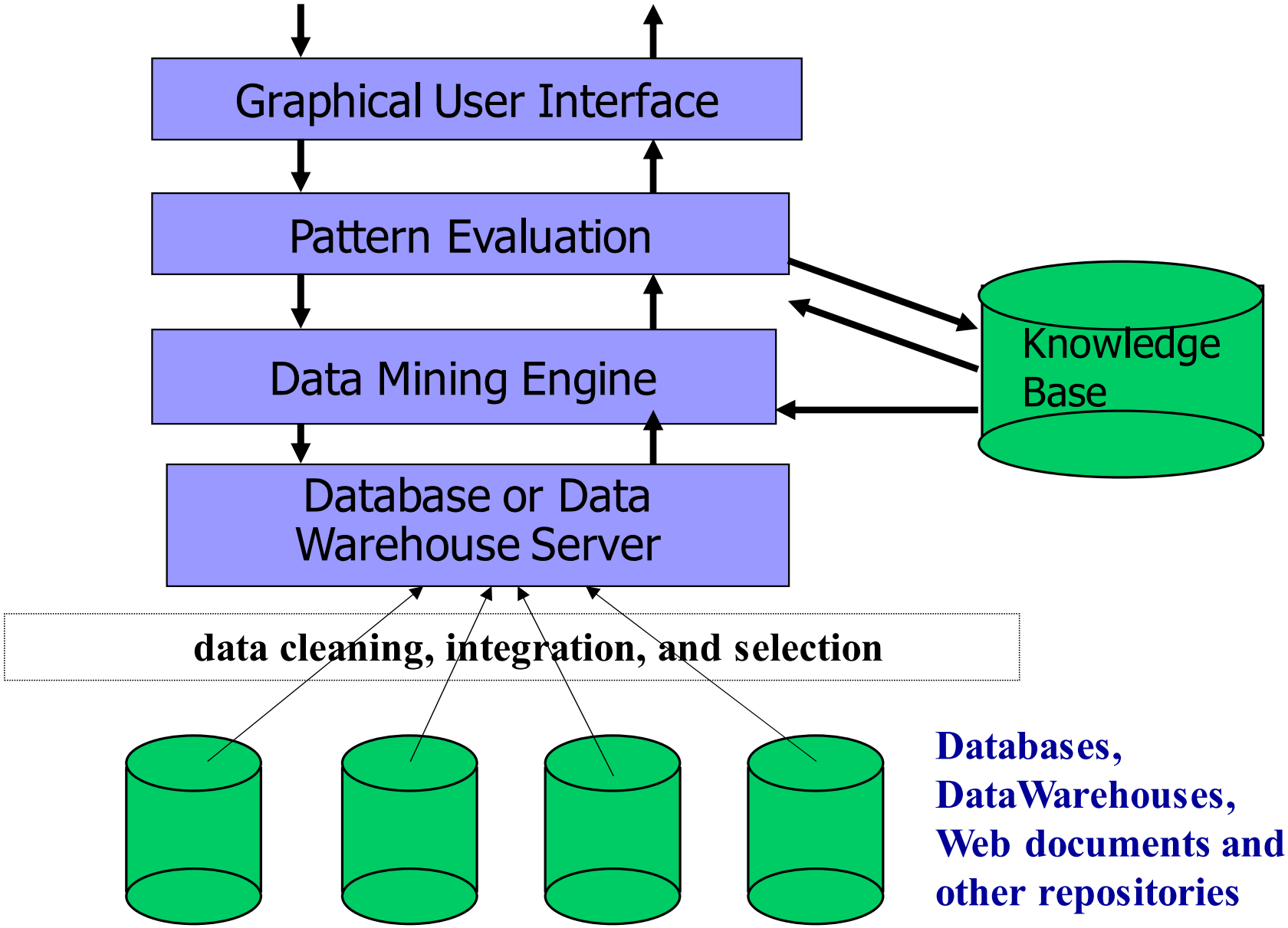
Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Architecture: Typical Data Mining System



Traditional Data Analysis

- Summaries
- Aggregations
- Views

Why New Kinds of Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle large amounts 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

Data Mining: Classification Schemes

- General functionality
  - 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

Data Mining: on what kinds of data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Object-relational databases
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Spatial data and spatiotemporal data
  - Text databases and Multimedia databases
  - Data streams and sensor data
  - The World-Wide Web
  - Heterogeneous databases and legacy databases

Data Mining – what kinds of patterns?

- Concept/class description:
  - Characterization: summarizing the data of the class under study in general terms
    - E.g. Characteristics of customers spending more than 10000 sek per year
  - Discrimination: comparing target class with other (contrasting) classes
    - E.g. Compare the characteristics of products that had a sales increase (target class) to products that had a sales decrease last year (contrasting class)

Data Mining – what kinds of patterns?

- Frequent patterns, association, correlations
  - Frequent itemset
  - Frequent sequential pattern
  - Frequent structured pattern

  - E.g. buy(X, “Diaper”) → buy(X, “Beer”) [support=0.5%, confidence=75%]
    confidence: if X buys a diaper, then there is 75% chance that X buys beer
    support: of all transactions under consideration 0.5% showed that diaper and beer were bought together
  - E.g. Age(X, ”20..29”) and income(X, ”20k..29k”) → buys(X, ”car”) [support=2%, confidence=60%]

Data Mining – what kinds of patterns?

- Classification and prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction.

    The derived model is based on analyzing training data – data whose class labels are known.
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown or missing numerical values

Data Mining – what kinds of patterns?

- Cluster analysis
  - Class label is unknown: Group data to form new classes,
    - E.g., cluster customers to find target groups for marketing
  - 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
  - Trends and deviation

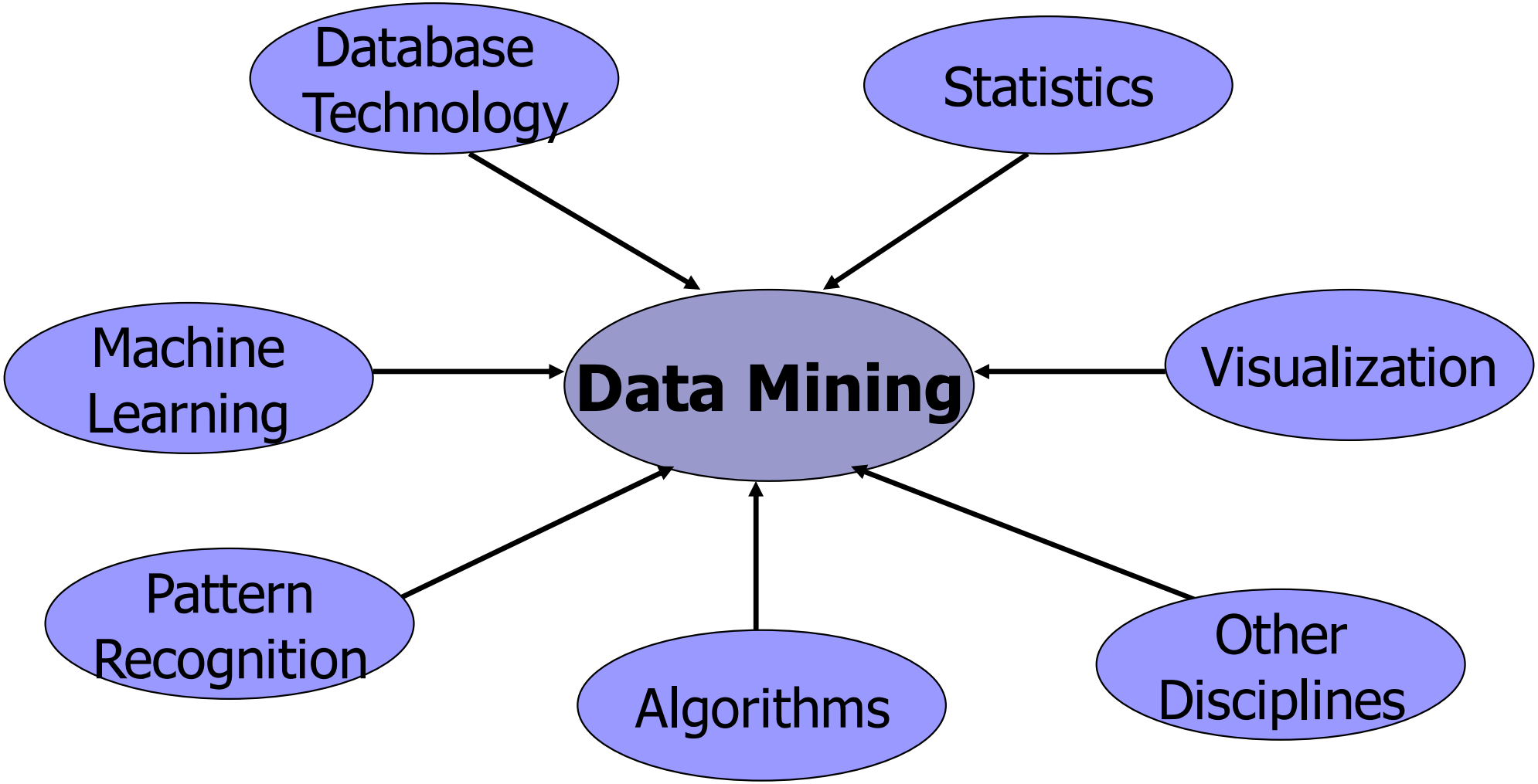
Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns
    - e.g., support, confidence, ...
  - Subjective: based on user's belief in the data
    - e.g., unexpectedness, novelty, actionability, ...

Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns? Do we need to find all of the interesting patterns?
  - Heuristic vs. exhaustive search
  - Association vs. classification vs. clustering
- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First generate all the patterns and then filter out the uninteresting ones
    - Generate only the interesting patterns—mining query optimization

Data Mining – what techniques used?



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), now ECML-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