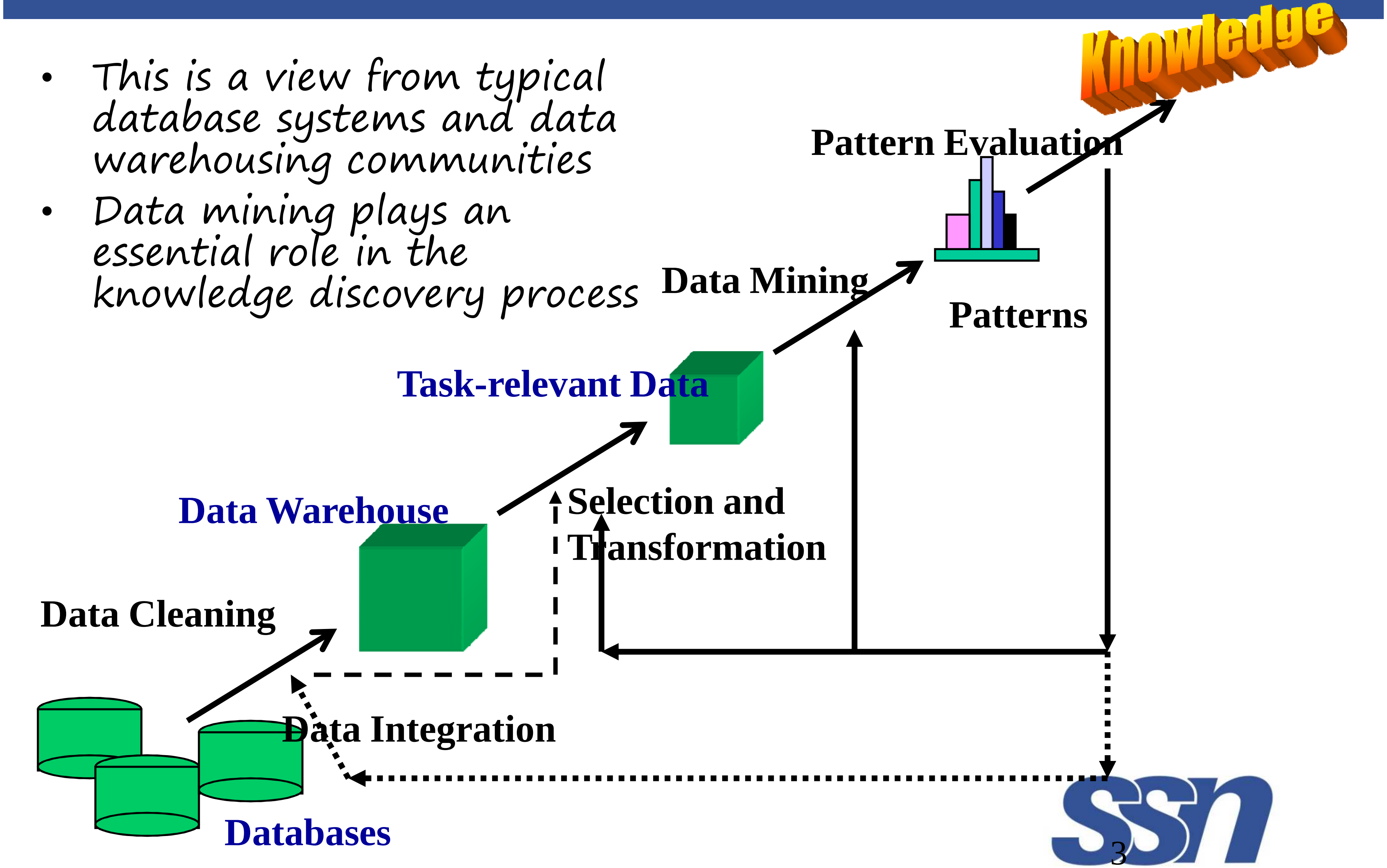
Data Mining – Introduction II

Overview

- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

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



Knowledge

Pattern Evaluation

Data Mining

Patterns

Task-relevant Data

Selection and Transformation

Data Warehouse

Data Cleaning

Data Integration

Databases

Patterns

- Are all the “Discovered” Patterns are interesting?
- What makes the pattern interesting?
- Can a data mining system generate only the interesting patterns?

Are all patterns are intersting?

- **A pattern is interesting if**
  - Easily understood by humans.
  - Valid on new or test data with some degree of certainty.
  - Potentially useful
  - Novel
  - Validates a hypothesis that the user sought to confirm
- An interesting pattern represents **knowledge**

Objective measures of pattern Interestingness

- **Objective measures** based on structure and statistics of discovered pattern. e.g., **support, confidence**, etc.
- For an Association rule x=>y **support** represents the percentage of transactions from a transaction database that satisfies the rule. **$P(X \cup Y)$**
- **Confidence:** Measures the degree of certainty of the detected association **$[P(Y|X)]$**
- **Accuracy** : Tells the percentage of the data that are correctly classified by the rule
- **Coverage:** Percentage of data to which a rule applies.

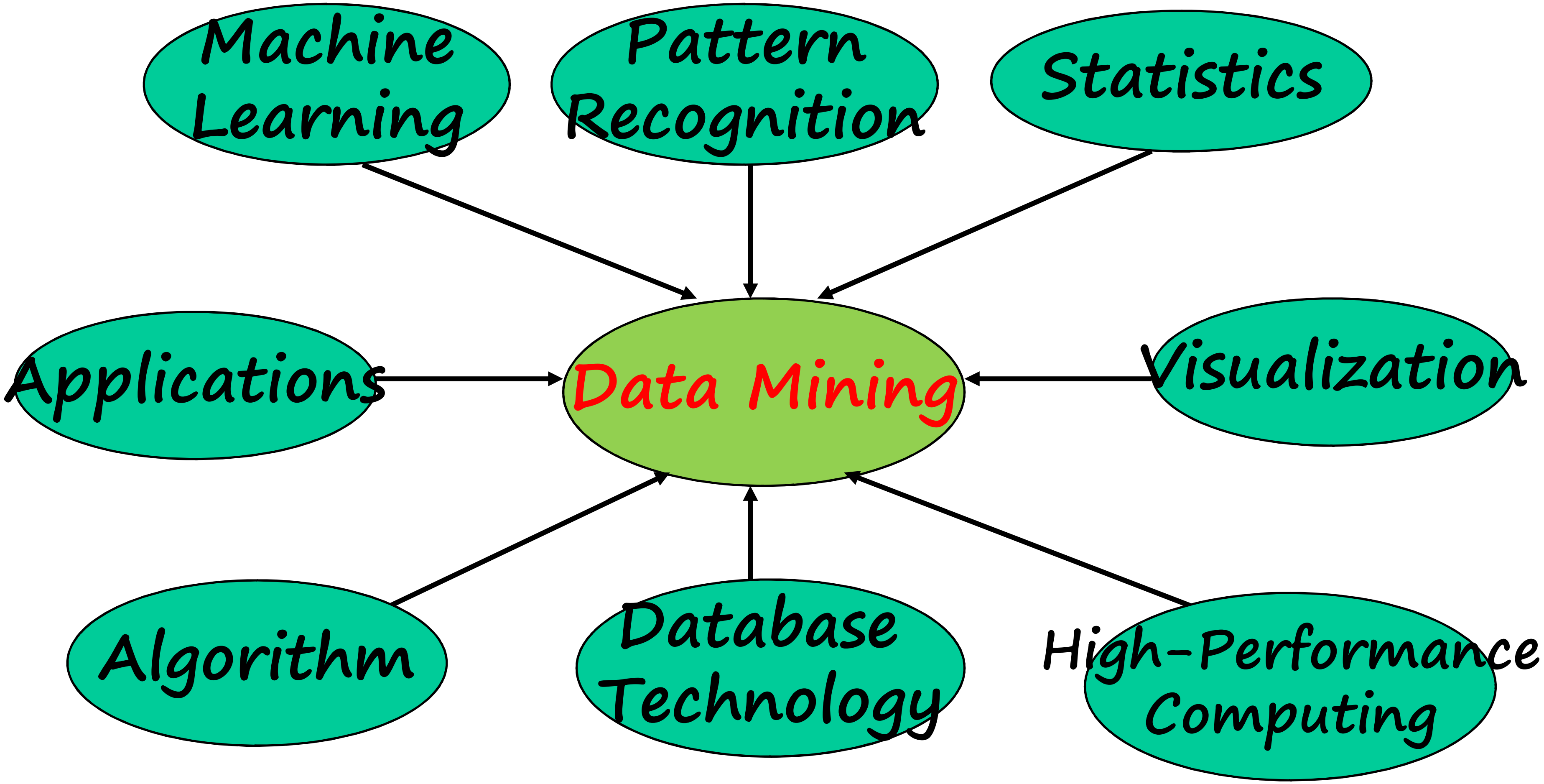
Subjective measures of pattern Interestingness

- **Subjective measures:** Based on user beliefs in the data , e.g., **unexpectedness, novelty, actionability, etc.**
- Measures find pattern interesting if the patterns are **unexpected patterns** providing contradicting information of the user beliefs
- **Actionable:** Offer strategic information on the which user act.
- **Find all the interesting patterns: Completeness**
  - Can a data mining system find all the interesting patterns?
  - Constraints and interestingness measures focus the search
- Eg: Association rule Mining can ensure the completeness with the help of constraints and interestingness measures

Can a data mining system find only the interesting patterns?

- Search for only interesting patterns: An optimization problem
  - Approaches
    - Measures of pattern interestingness are essential for efficient discovery of patterns by target users
    - Generalize and rank all the interestigness patterns and then filter out the uninteresting ones.
    - Access the methods for pattern interestingness and should use to improve data mining efficiency

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- **Tremendous amount of data**
  - Algorithms must be scalable to handle big data
- **High-dimensionality of data**
  - Data can have tens or thousands of features (e.g DNA, micro array)
- **High complexity of data**
  - Data can be highly complex, can be of different types and can include different descriptors.
  - Images can be described using text and visual features such as color, texture and contour etc.,
  - Videos can be described using text images and their descriptors
  - Social networks have complex structures.
- **New and sophisticated applications**
  - Applications can be difficult (eg.Medical applications)

Statistics

- Data mining (DM) has inherent connection with statistics.
- Statistics studies the collection, analysis, interpretation or explanation and presentation of data.
- Set of mathematical functions describe the behavior of the objects in the target class in terms of random variable and their associated probability distributions.
- USES:
  - Model target data and data classes (outcome of DM)
  - DM can be built on the top of statistical models
  - Helps to develop tools for prediction and forecasting
  - Used to summarize or describe collection of data and to draw inferences about the process.
  - To verify DM results (Statistical Hypothesis test)

Machine Learning

- Machine learning is computer program that automatically learn to recognize complex patterns and make intelligent decisions based on data.
  - Supervised learning (classification)
  - Unsupervised Learning(Clustering)
  - Semi-supervised learning(combination of unsupervised and supervised)
  - Active learning

Information Retrieval(IR)

- IR is science of searching for documents or information in documents.
- IR assumes data under search are
  - Unstructured
  - Queries are formed by keywords
- IR adopt probabilistic models for generating bag of words by means of documents language model .
- Integration of IR models with DM techniques helps to find major topics in the collection of document and for each document

Applications of Data Mining

- Web page analysis: classification, clustering, ranking
- Collaborative analysis & recommender systei
- 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
- Built-in (invisible data mining) functions in Google, MS, Yahoo!, Linked, Facebook, …
- Major dedicated data mining systems/tools
  - SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)

Major Issues in Data Mining

- Major issues in data mining are partition into five groups
  - Mining methodology
  - User interaction
  - Efficiency and scalability
  - Diversity of data types
  - Data mining and society

Major Issues in Data Mining-Mining Methodology

- **Mining various and new kinds of knowledge**
  - Different mining tasks use the same DB
  - New mining tasks continue to emerge due to the diversity of applications making DM dynamic and fast growing field
  - Eg: Integrated clustering and ranking led to discovery of high quality clusters in n/w mining.
- **Mining knowledge in multi-dimensional space**
  - Searching for interesting patterns among combinations of dimensions at varying levels of abstraction. (exploratory)
- **Data mining: An interdisciplinary effort**
  - Power of DM can be enhanced by integrating new methods from multiple discipline.
  - Eg: Text mining fuses DM with NPL and IR

Major Issues in Data Mining (1)

- **Boosting the power of discovery in a networked environment**
  - Semantic links across multiple data objects and knowledge derived can be used to boost the discovery of knowledge in a related or semantically linked set.
- **Handling noise, uncertainty, and incompleteness of data**
  - Errors and other uncertainty leads to erroneous patterns
  - Data cleaning, preprocessing, outlier detection and removal are integrated with DM process
- **Pattern evaluation and pattern- or constraint-guided mining:** Techniques are needed to assess the interestingness patterns based on subjective measure

Major Issues in Data Mining –User Interaction

- **Interactive mining:**
  - Build flexible UI and an exploratory mining environment.
  - Interactive mining is needed to dynamically change the focus of a search.
- **Incorporation of background knowledge**
  - Background knowledge, constraints, rules and other information should be incorporated.
  - Knowledge can be used for pattern evaluation and to guide search toward interesting pattern.
- **Presentation and visualization of data mining results**
  - Expressive knowledge representation, user friendly interfaces and visualization techniques

Major Issues in Data Mining –User Interaction

- **Ahoc data mining and data mining query language:**
  - Query languages plays important role in searching that allow users to pose ad hoc queries.
  - High-level DM languages or high-level flexible user interfaces defines adhoc mining tasks
  - Facilitates specification of relevant sets of data for analysis, domain knowledge, the kinds of knowledge to be mined and conditions and constraints to be enforced on the discovered patterns.

Major Issues in Data Mining –Efficiency and Scalability

- **Efficiency and scalability of data mining algorithms –**
  - DM algorithm must be efficient and scalable in order to effectively extract the information from huge amount of data in many data repositories or in dynamic data streams
  - Key criteria: Efficiency, scalability, performance, optimization and ability to execute in real time for development of new DM algorithms
- **Parallel, distributed, and incremental mining algorithms**

  – The factors such as huge size of databases, wide distribution of data, and computational complexity of data mining methods motivate the development **of parallel and distributed data mining algorithms.**

Major Issues in Data Mining –Efficiency and Scalability

- These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged.
- **Cloud and cluster computing** use computers in a distributed fashion to tackle large scale computational tasks.
- The incremental data mining algorithms incorporates new data updates without having to mine the entire data again from scratch.

Major Issues in Data Mining –Diversity of Data Types

- **Handling of relational and complex types of data –**
  - Diverse applications generate wide spectrum of new data types.
  - It is not possible for one system to mine all these kind of data, given the diversity of data types and different goals
  - Domain and application oriented DM are constructed for indepth mining of specific kinds of data.
  - The construction of efficient and affective DM tools for diverse applications remains challenging

Major Issues in Data Mining –Diversity of Data Types

- **Mining information from heterogeneous databases and global information systems –**
- The data is available at different data sources on LAN or WAN.
- These data source may be structured, semi structured or unstructured with diverse semantics poses great challenge to DM.
- Therefore mining the knowledge from them adds challenges to data mining.

Major Issues in Data Mining -Data mining and society

- **Social impacts of data mining**
  - How can we use DM technology to benefit society?
  - How can we guard against its misuse?
  - Proper disclosure or use of data and potential violation of individual privacy and protection rights are need to be addressed.
- **Privacy-preserving data mining**
  - DM helps in scientific discovery, business management, economy recovery and security protection.
  - Risks of disclosing the personal information
  - Studies on privacy preserving data publishing and data mining

Major Issues in Data Mining -Data mining and society

- **Invisible Data mining:**
  - More and more systems should have DM functions to built within so that people perform DM by mouse clicking without knowledge algorithm
  - Internet search engines and Internet based stores perform such invisible DM by incorporating DM into their components to improve the functionality and performance.

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
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications
- Major issues in data mining