Introduction of Data Mining

Young-Rae Cho, Ph.D.

Associate Professor

Division of Software / Division of Digital Healthcare

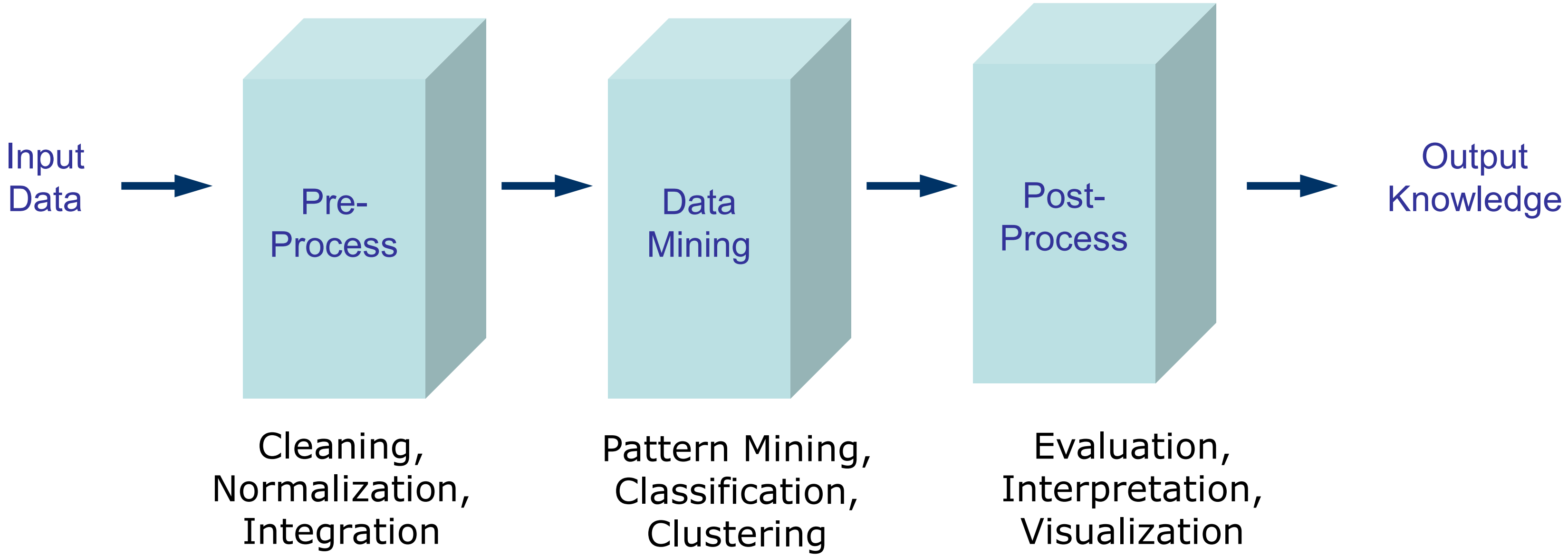
Yonsei University – Mirae Campus

What is Data Mining?

❑ Definition

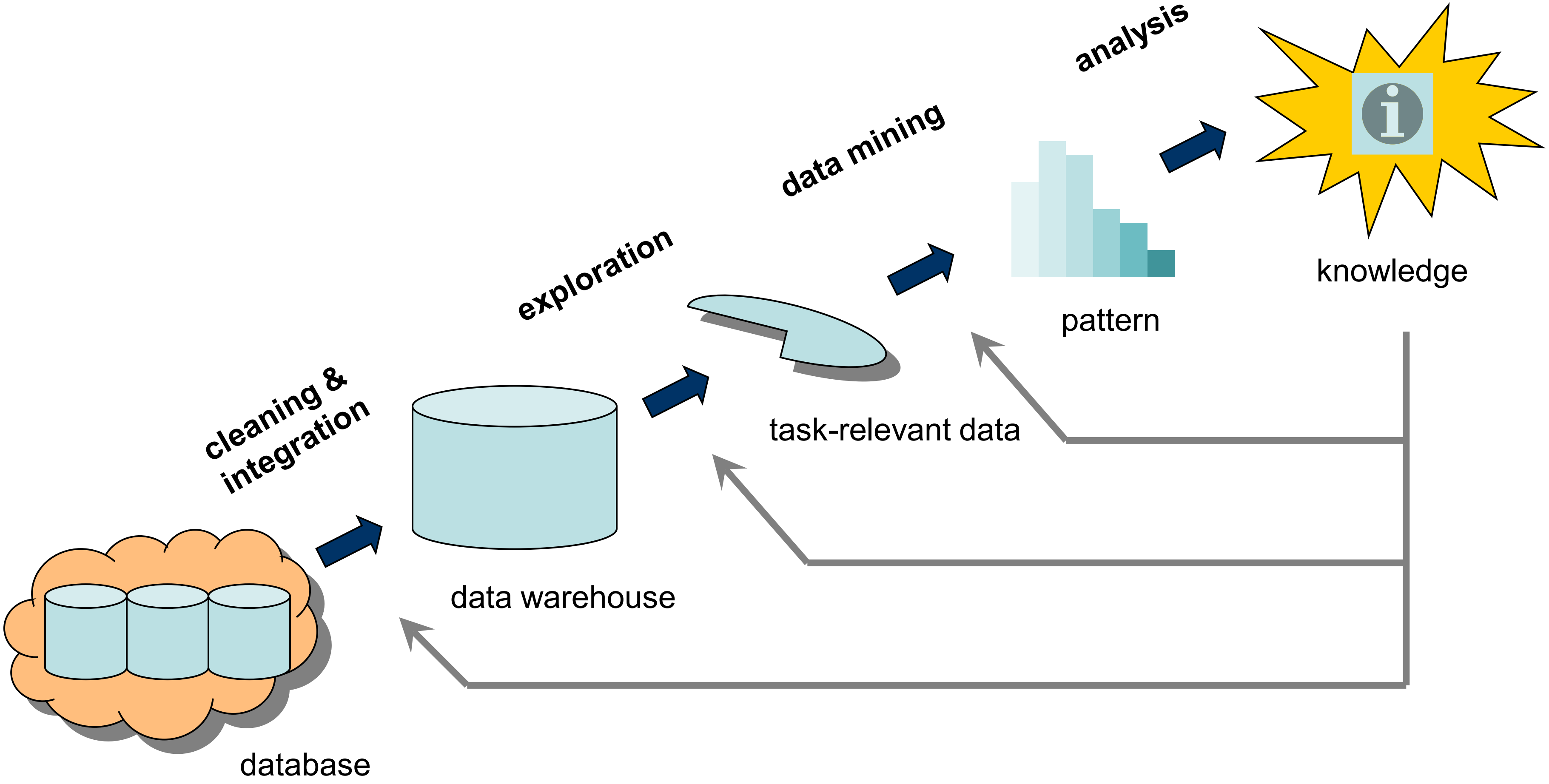
- Knowledge discovery from data (KDD)
- Informative pattern extraction from data
- Data analysis using specific algorithms or machine learning techniques

❑ General View of Data Mining Process

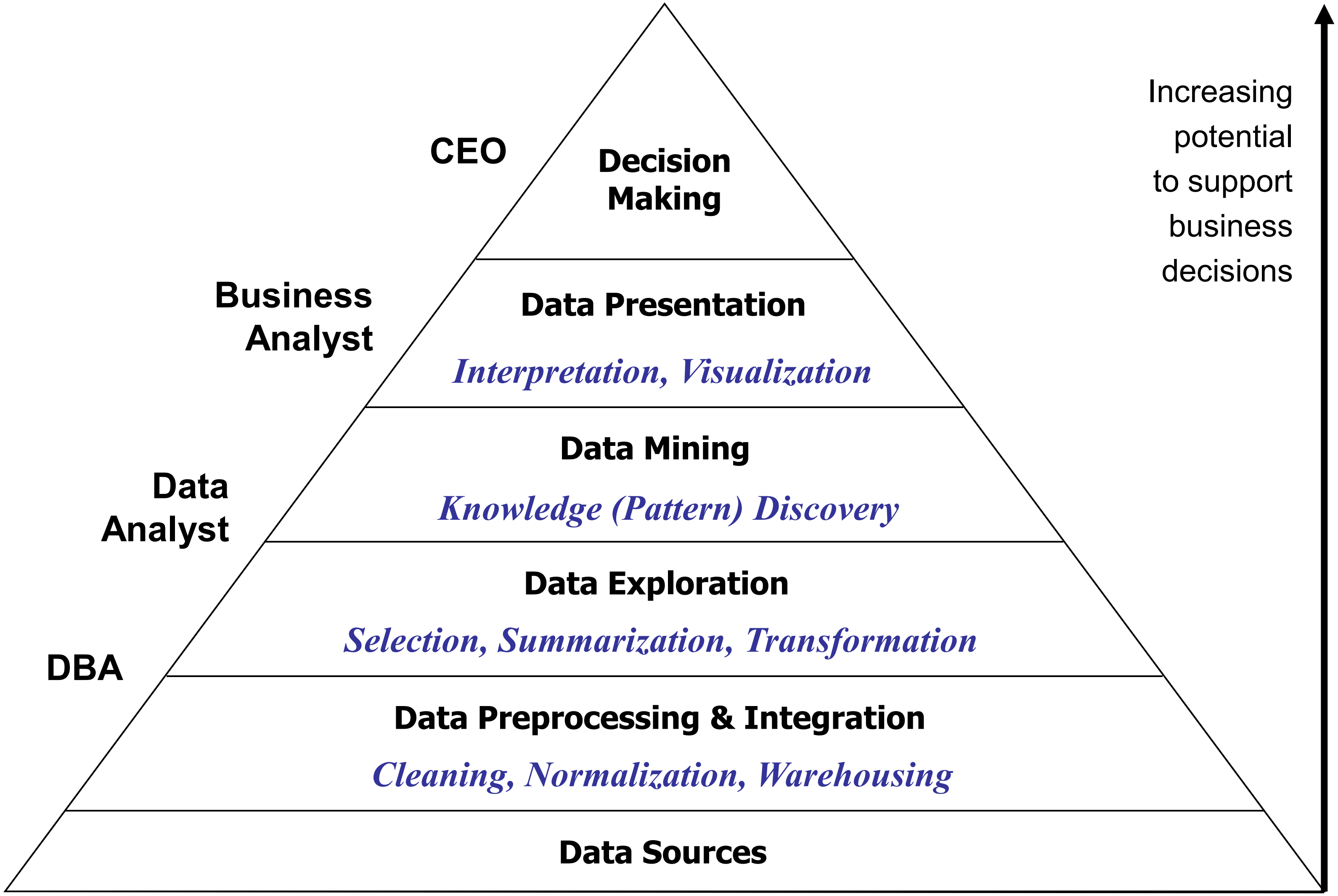

Input Data
Pre-Process
Data Mining
Post-Process
Output Knowledge
Cleaning, Normalization, Integration
Pattern Mining, Classification, Clustering
Evaluation, Interpretation, Visualization

Data Mining in Computer Science

❑ Alternative View of Data Mining Process



Data Mining in Business


CEO
Decision Making
Business Analyst
Data Presentation
Interpretation, Visualization
Data Analyst
Data Mining
Knowledge (Pattern) Discovery
DBA
Data Exploration
Selection, Summarization, Transformation
Data Preprocessing & Integration
Cleaning, Normalization, Warehousing
Data Sources
Increasing potential to support business decisions

Why Need Data Mining?

- **Business**
  - Business data analysis for decision support
    - Market analysis and management
    - Risk analysis and management
    - Fraud detection and security

- **Science and Engineering**
  - Biomedical data analysis
    - Patient treatment, disease diagnosis, and drug discovery
  - WWW data analysis
    - Information retrieval and web management
  - Geographic data analysis
    - City planning and renewal

Why Not Traditional Data Analysis?

- ❑ **Explosive Growth of Data**
  - Terabytes or petabytes of data
- ❑ **High Dimensionality of Data**
  - Hundreds or thousands of dimensions
- ❑ **High Complexity of Data**
  - Stream data, sensor data
  - Time-series data, temporal data
  - Spatial data, spatio-temporal data, multimedia data
  - Structural data, graphic data
  - Combined, heterogeneous data format

➡ **Data mining algorithms should handle these data !!**

Data Mining Functions: (1) Generalization

- **Data Cleaning and Reduction**
  - Statistical normalization methods
  - Sampling and discretizing techniques
- **Data Integration and Warehousing**
  - Multidimensional data modeling
  - Dimension reduction techniques
  - Data cube aggregation algorithms
- **Data Transformation**
  - OLAP (online analytical process) operations
  - Querying for selection and summarization

Data Mining Functions: (2) Pattern Mining

- **Frequent Pattern Mining**
  - Mining frequently occurred item-sets
  - Mining frequently occurred sequential patterns
  - Mining frequently occurred structural patterns (sub-graphs)

- **Association Rule Mining**
  - Mining one-direction relations between two sets of data

- **Correlation Mining**
  - Mining two-direction relations between two sets of data

- **Coherent Pattern Mining**
  - Mining coherent sequential patterns
  - Mining coherent structural patterns

Data Mining Functions: (3) Classification

- **Supervised Learning**
  - Training data with class labels
  - Prediction of classes of data with no class labels

- **Typical Examples**
  - Decision tree-based induction
  - Naïve bayesian classification
  - K-nearest neighbors (kNN)
  - Support vector machine (SVM)
  - Neural network
  - Logistic regression
  - Rule-based classification
  - Pattern-based classification

Data Mining Functions: (4) Clustering

- **Unsupervised Learning**
  - Grouping data with no class labels
  - Prediction of potential members with same class labels

- **Typical Examples**
  - K-means
  - Agglomerative hierarchical clustering
  - Divisive hierarchical clustering
  - Density-based clustering
  - Grid-based clustering
  - Pattern-based clustering
  - Outlier analysis

Summary of Data Mining

❑ **Inter-disciplinary Field**

- Basic disciplines: Algorithms, Databases, Statistics
- Advanced disciplines: Machine Learning, Pattern Recognition, High-Performance Computing
- Applications: Visualization, Web Applications

❑ **Origins & History**

- 1991 - 1994: Workshop on Knowledge Discovery in Databases
- 1993: Market basket problem (Agrawal et al., ACM SIGMOD Conference)
- 1994: Apriori algorithm (Agrawal and Srikant, VLDB Conference)
- 1995 - current: International Conference on Knowledge Discovery and Data Mining (KDD)

  Sponsored by ACM from 1998

❑ **Current Conferences & Journals**

- Annual Conferences: ACM KDD, IEEE ICDM, ACM CIKM, SDM, PKDD, PAKDD
- Journals: DMKD by Springer, IEEE TKDE, ACM TKDD

Questions?

❑ Lecture Slides on the Course Website, “https://ads.yonsei.ac.kr/faculty/data_mining/”