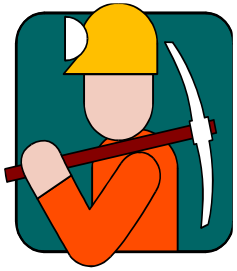
Chapter 1: Introduction

Dong-Kyu Chae

PI of the Data Intelligence Lab @HYU
Department of Computer Science & Data Science
Hanyang University

Data
Intelligence
LAB

What Is Data Mining?

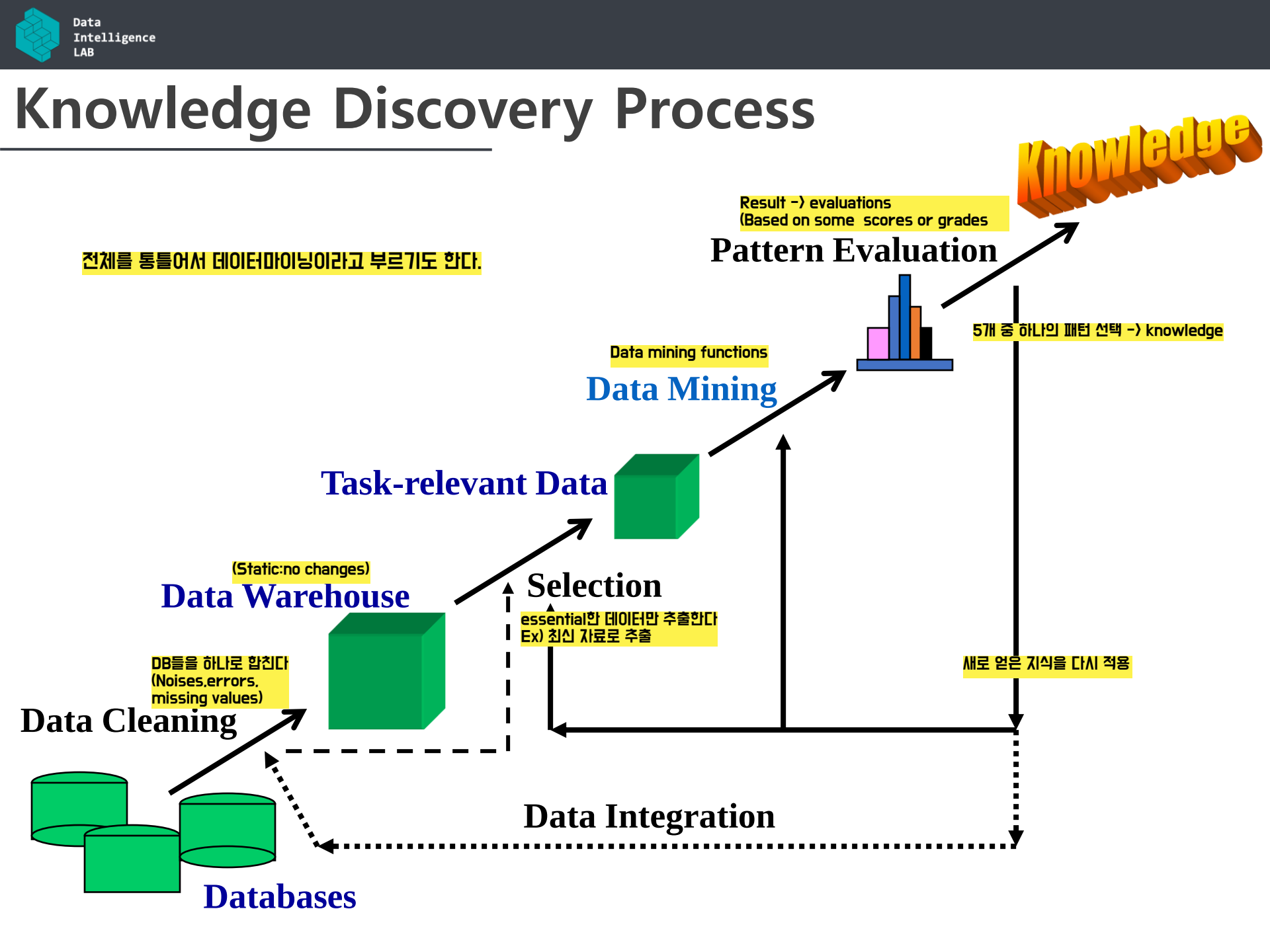


- **Data mining (knowledge discovery from data)**
  - Automatic extraction of interesting (**non-trivial**, **implicit**, **previously unknown** and **potentially useful**) patterns or knowledge from **huge amount of data**

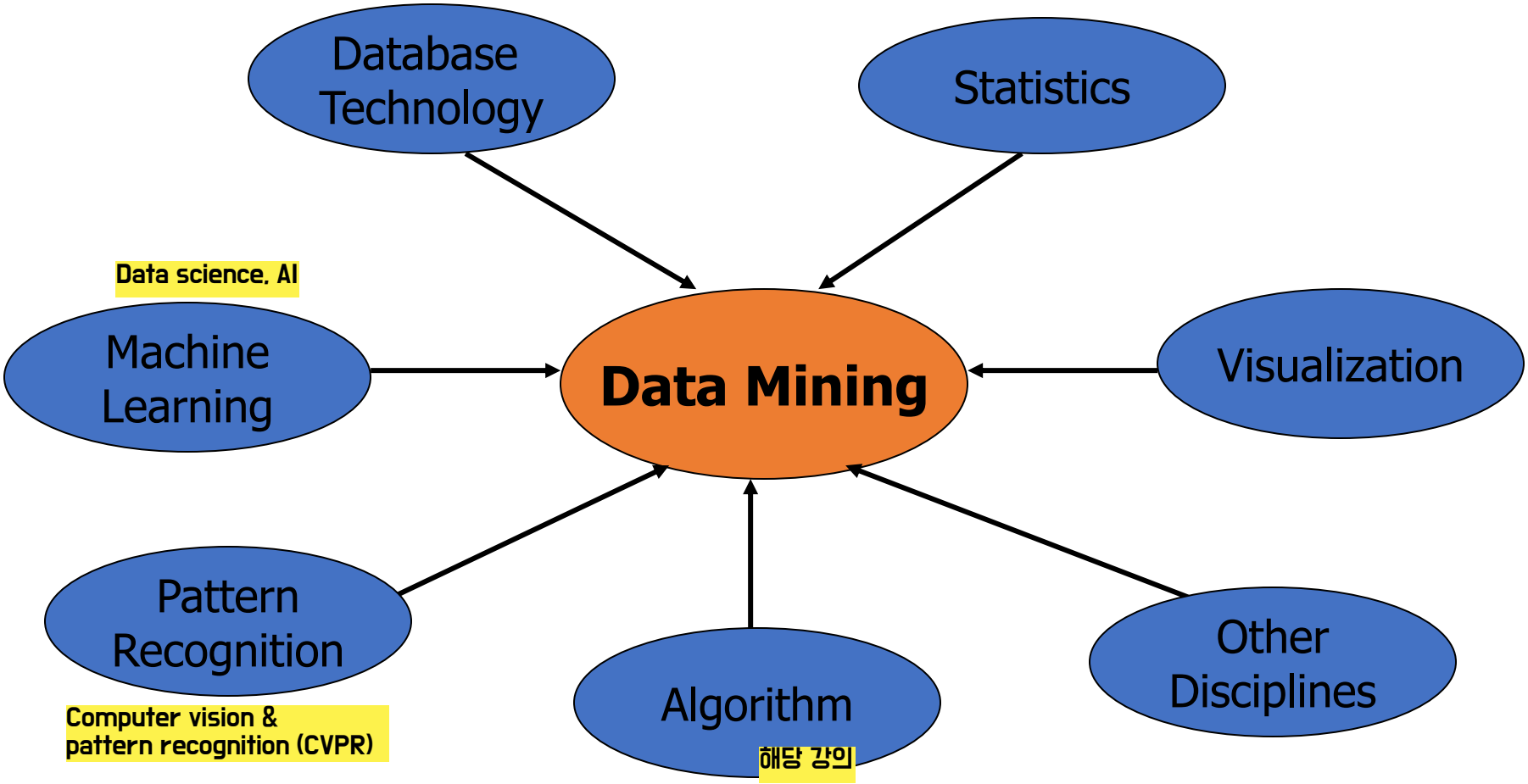
Data mining 을 하는 이유:
데이터량이 너무 크기 때문에 (테라바이트~)

-> crucial한 데이터 추출 필요

Knowledge Discovery Process



Data Mining: Confluence of Multiple Disciplines



Functionalities for Data Mining

- **Frequent patterns, association rules**
  - Diaper → Beer

기저귀를 사는 사람이 맥주를 살 확률이 높다.
-> {diaper, beer}

Ex) 고양이냐 강아지냐

연속적인 값을 예상하는 것 (날씨 등)

- **Classification and regression** (Machine learning)
  - 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

Functionalities for Data Mining

- **Cluster analysis** 분류하는 것 보다 데이터 자제에 중점을 두고 성질을 비교
Ex) 차 (아반떼K5/BMW아우디벤츠/포르쉐벤틀리)
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing inter-class similarity

- **Outlier analysis** 일반적의 데이터 분포와 다르게 모난 데이터
-) 찾아내서 지우거나 함
  - 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** (연쇄) 카메라를 샀으면 SD 메모리를 살 것
(+ 기저귀-맥주)
  - Sequential pattern mining: e.g., digital camera → large SD memory

Time stamp가 다름

Research Issues in Data Mining

- **Mining methodology**
  - Mining valuable knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability High speed and accuracy
  - Pattern evaluation: the interestingness problem 패턴을 찾으면 그 패턴이 중요한 것인지 평가
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods 데이터가 너무 큰 경우
  - Integration of the discovered knowledge with existing one: knowledge fusion

integration

Research Issues in Data Mining

- **User interaction**
  - Data mining query languages Ex) frequent pattern) 20%
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction = visualization

- **Applications and social impacts**
  - Domain-specific data mining 도메인에 따라 다름 (A ->A')
  - Protection of data security, integrity, and privacy Users' privacy issues

Summary

- Data mining: automatically 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
- Data mining functionalities: frequent patterns and associations, classification, clustering, outlier and trend analysis, etc.
- Major issues in data mining

Thank You

Data
Intelligence
Lab