Chapter 1: Introduction

What is Data Mining? (Definition, Process, Research Issue)

Dong-Kyu Chae

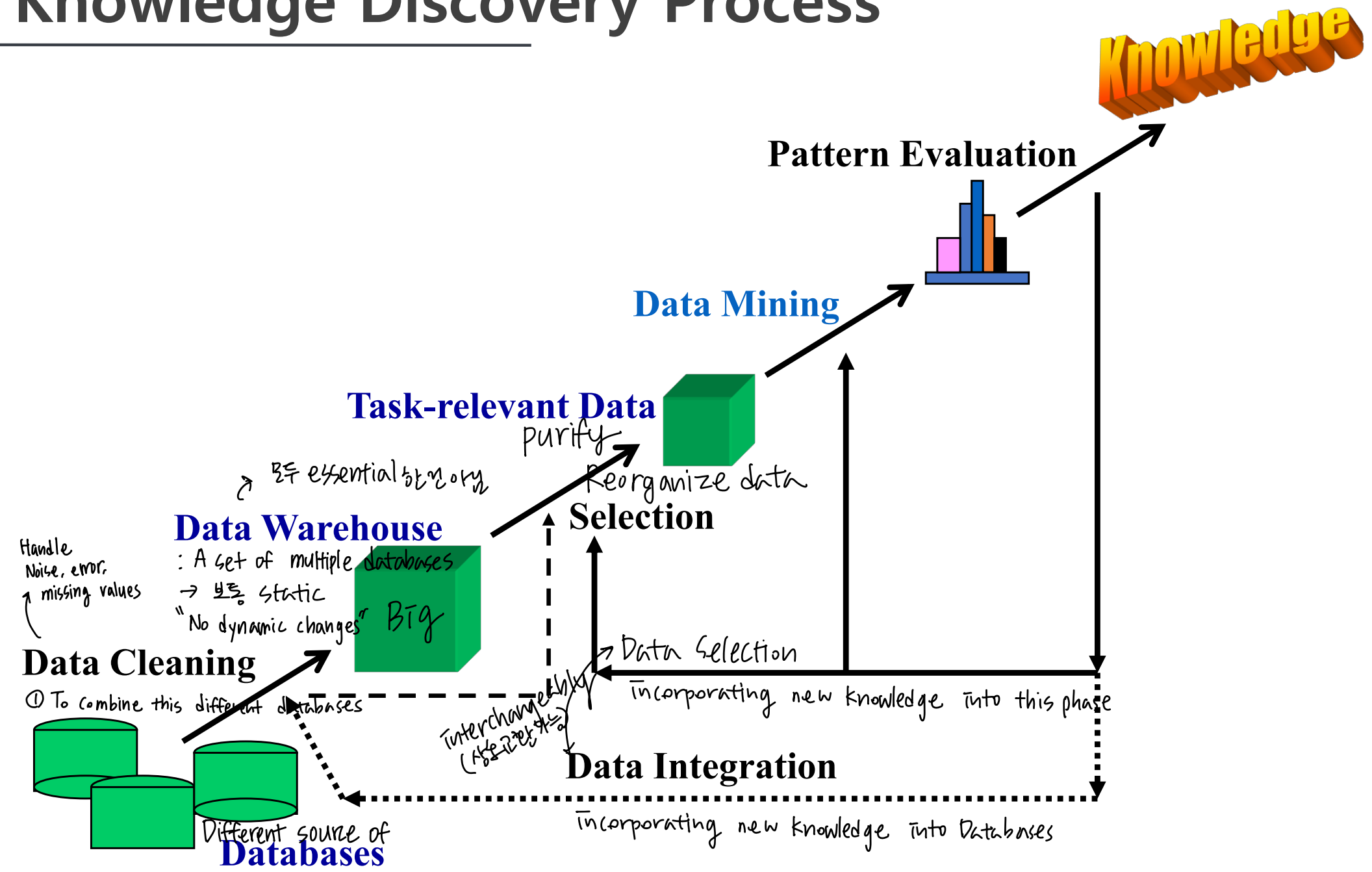
PI of the Data Intelligence Lab @HYU
Department of Computer Science & Data Science
Hanyang University

Data
Intelligence
LAB

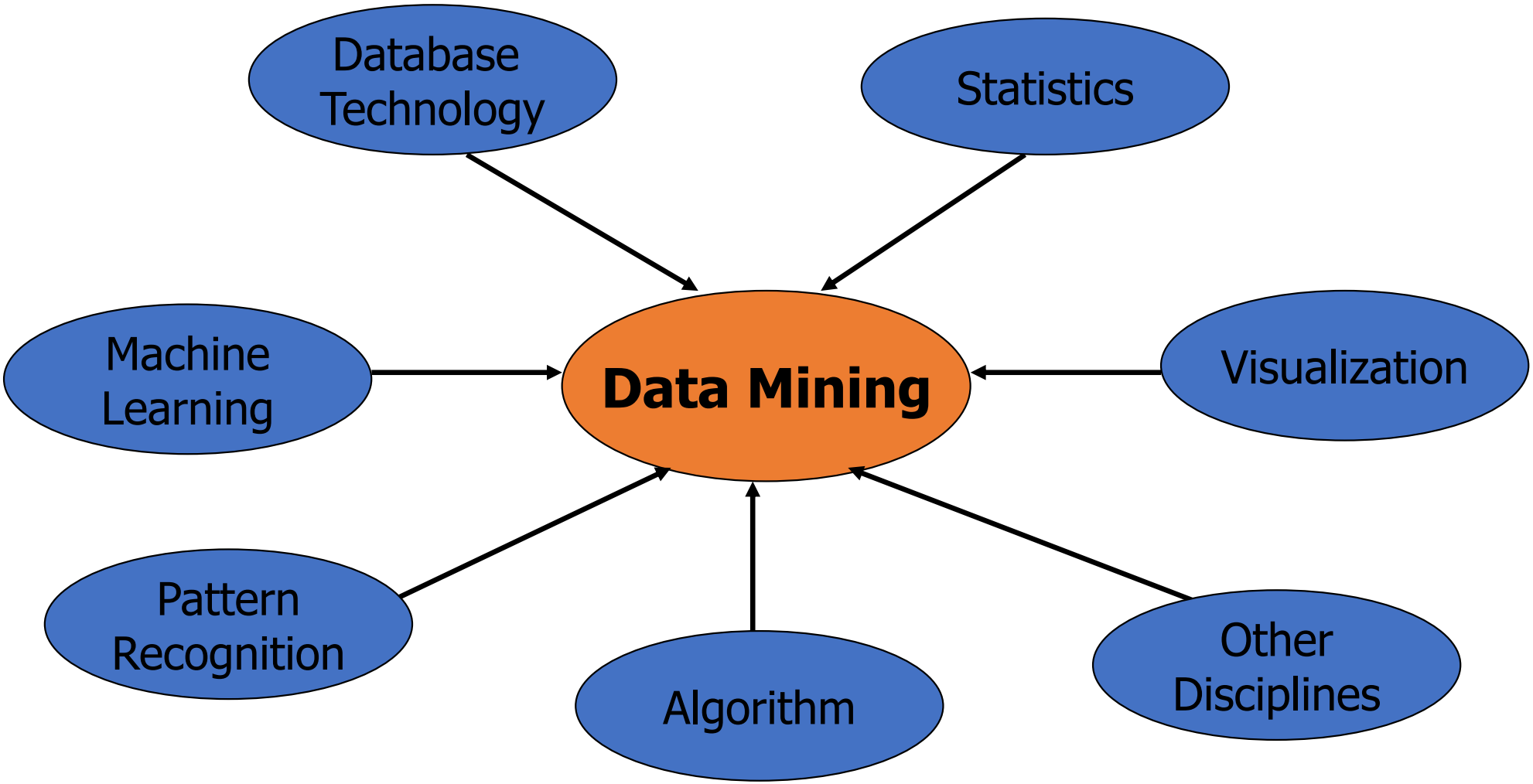
What Is Data Mining?

- **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**

Knowledge Discovery Process



Data Mining: Confluence of Multiple Disciplines



Functionalities for Data Mining

- **Frequent patterns, association rules**
  - Diaper → Beer
- **Classification and regression**

continuous value 예측
ex) 주식, 비트코인, 날씨, 기온 등

  - 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** Interesting data itself
  - 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

    Identify similar group of data
- **Outlier analysis**
  - Outlier: Data object that does not comply with the general behavior of the data — far from normal data
  - Noise or exception? Useful in fraud detection, rare-events analysis
- **Trend and evolution analysis** ← Time series data
  - Sequential pattern mining: e.g., digital camera → large SD memory

    → similar with association rules

    But diffrence in the pattern in the same time

    sequential은 time 다름 time stamp 가 different

Research Issues in Data Mining

- **Mining methodology**
  - Mining valuable 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

Integration of automatic data mining method and the background knowledge of human experts

Data cleaning

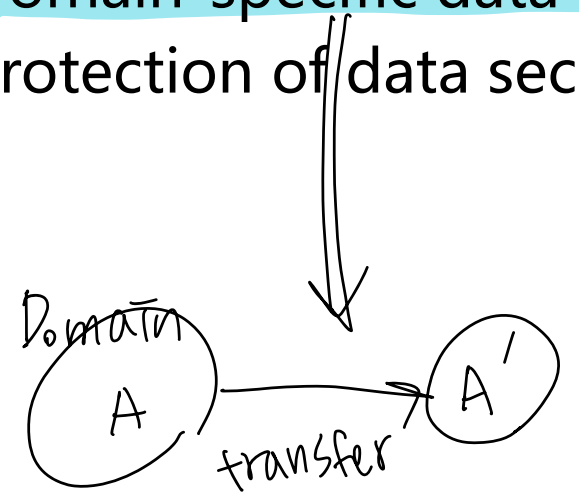
This is related with the big size data

or knowledge 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 복잡한 data를 user가 잘 이해하기 위해서 abstraction 필요

- **Applications and social impacts**
  - Domain-specific data mining ← Customize general data mining technique to specific data mining application
  - Protection of data security, integrity, and privacy



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