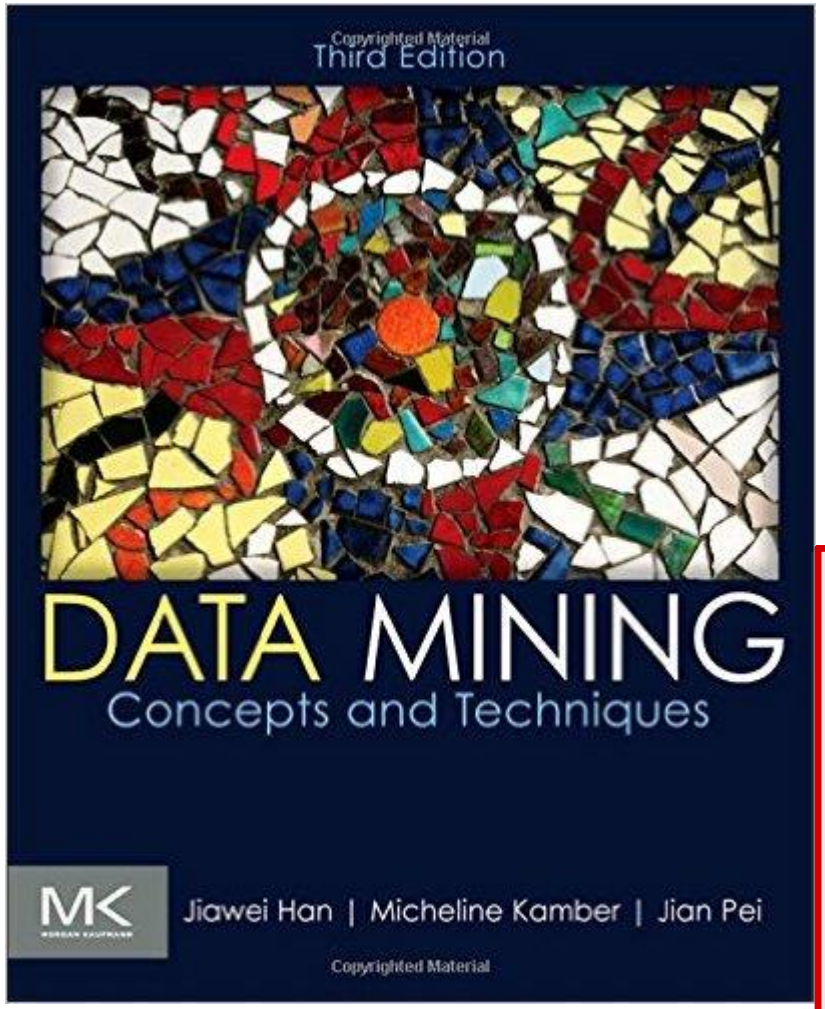
CS 412 Intro. to Data Mining

Chapter 1. Introduction

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

CS 412. Course Page & Class Schedule



- Textbook
  - Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques (3rd ed)*, Morgan Kaufmann, 2011
- Class Homepage: https://wiki.engr.illinois.edu/display/cs412
-
-
-
-

Jiawei Han

Michael Aiken Chair Professor
Department of Computer Science
Univ. of Illinois at Urbana-Champaign
Rm 2132, Siebel Center for Computer Science
201 N. Goodwin Avenue
Urbana, IL 61801, USA
E-mail: hanj[at]illinois.edu

Ph.D. (1985), Computer Science, Univ. Wisconsin-Madison

@1404

2132

ce is

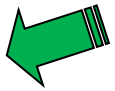
Course Work and Grading

- Project, Programming Assignments, and Exams
  - Project Presentation: ~~40%~~ 30%
  - Programming assignments: ~~25%~~ 40%
  - Midterm exam: 0%
  - Final exam: ~~35%~~ 30%
- Attendance (A)-> final score = A*score?
- Group project : 5 person จับกลุ่ม 5 คน
- Time table (Bring notebook to the class)
  - ?

Project

- Intro (What you are going to do)
- What data (Sources and How to get them) **at least 3 sources**
  - First presentation
- Data preparation & Warehousing
  - Second presentation
- Data mining
- Conclusion
  - Third presentation + Video Presentation

Chapter 1. Introduction



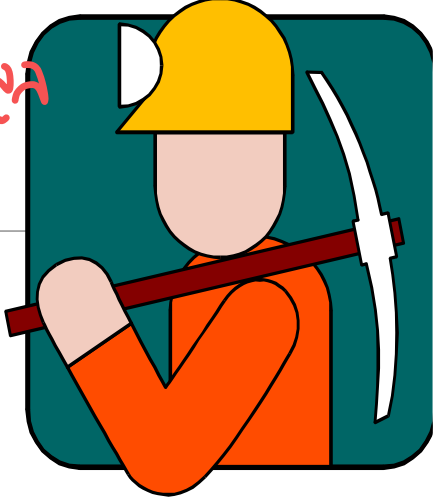
- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Why Data Mining?

เมื่องขังมูล

- 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!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

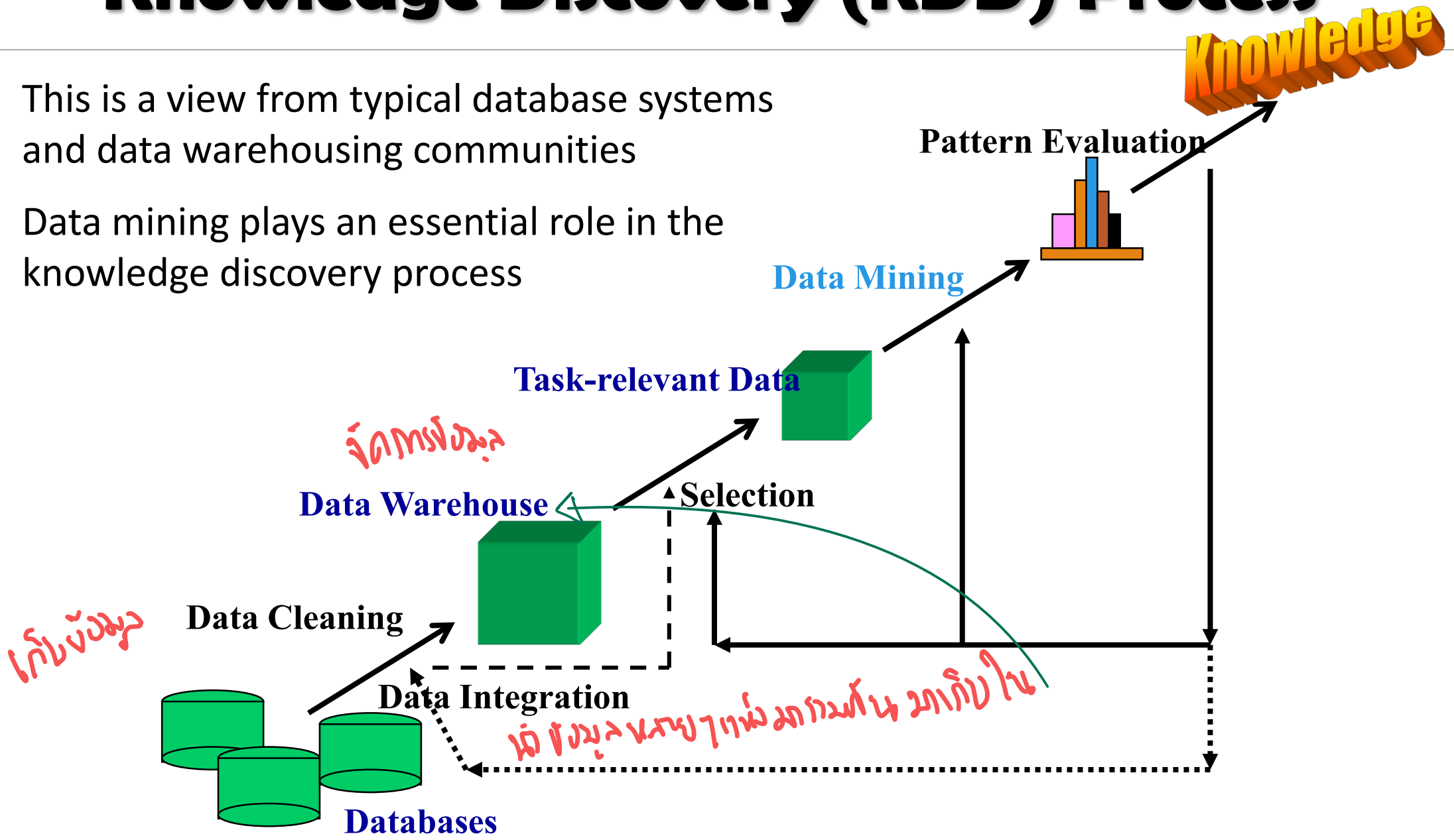
What Is Data Mining?



- 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, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Knowledge Discovery (KDD) Process

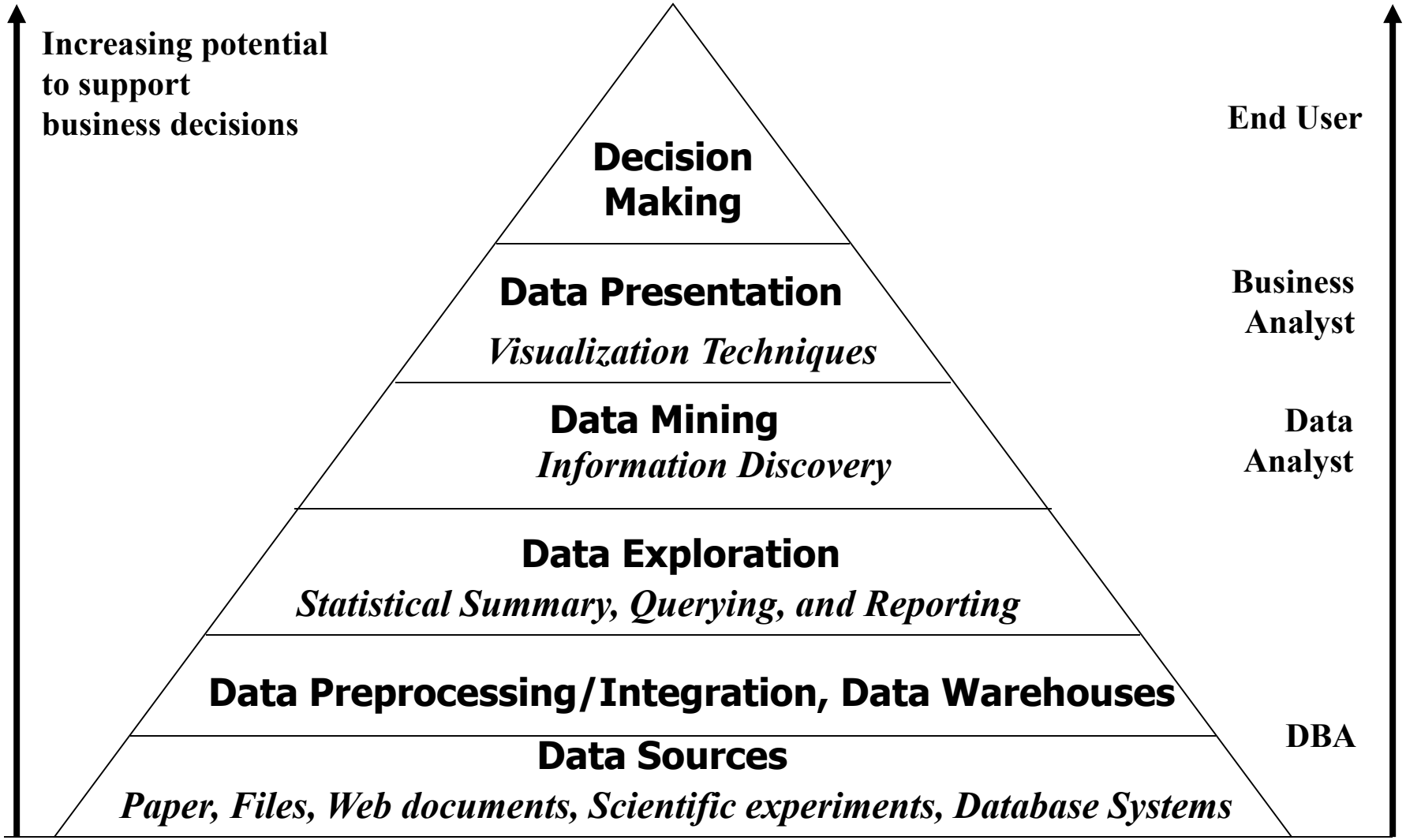
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



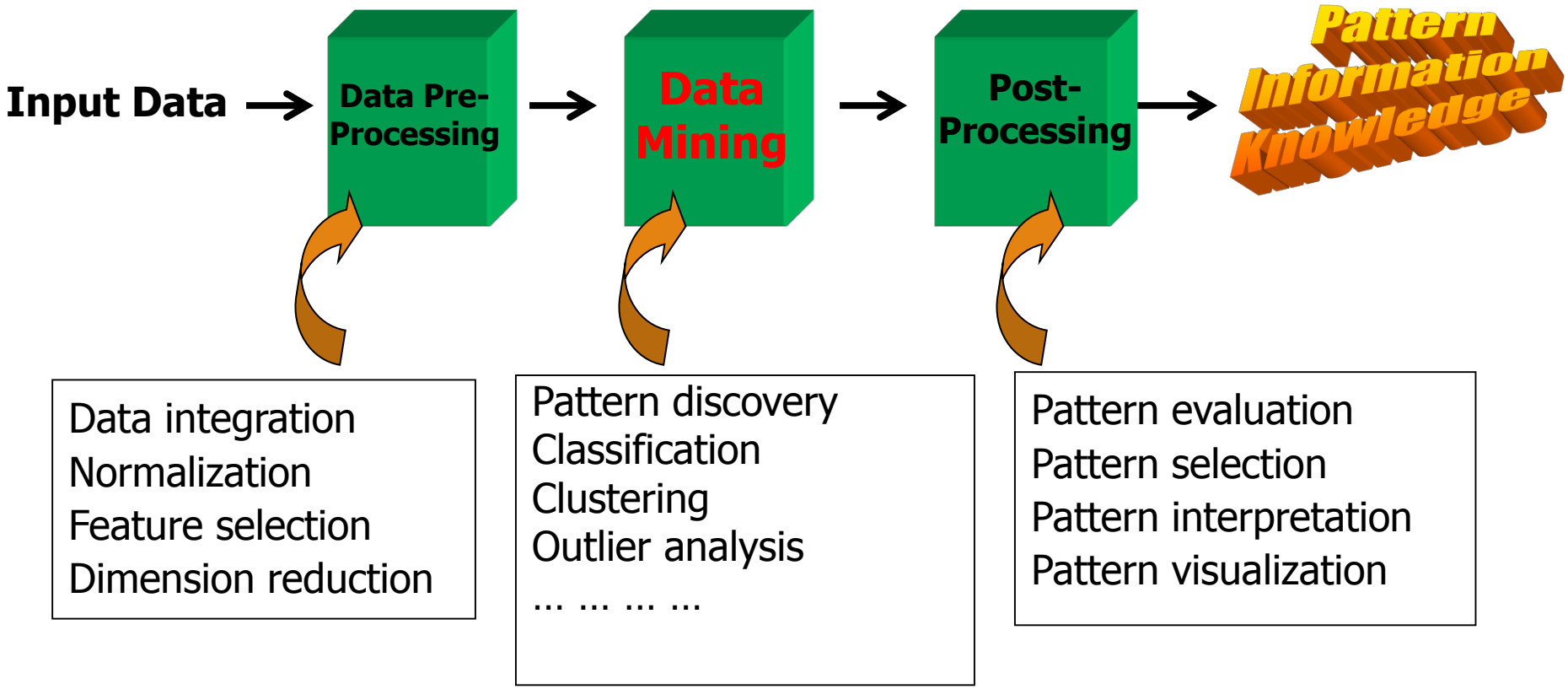
Example: A Web Mining Framework

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



KDD Process: A View from ML and Statistics



- This is a view from typical machine learning and statistics communities