Lecture 1:
A Brief Introduction to Data Mining

Lecture 1. A Brief Introduction to Data Mining

❑ Why Data Mining?

❑ What Is Data Mining?

❑ Data Mining: A Knowledge Discovery (KDD) Process

❑ Data Mining: A Multi-Dimensional View

   ❑ Data Mining from Data Point of View

   ❑ Data Mining from Knowledge Point of View

   ❑ Data Mining from Methodology Point of View
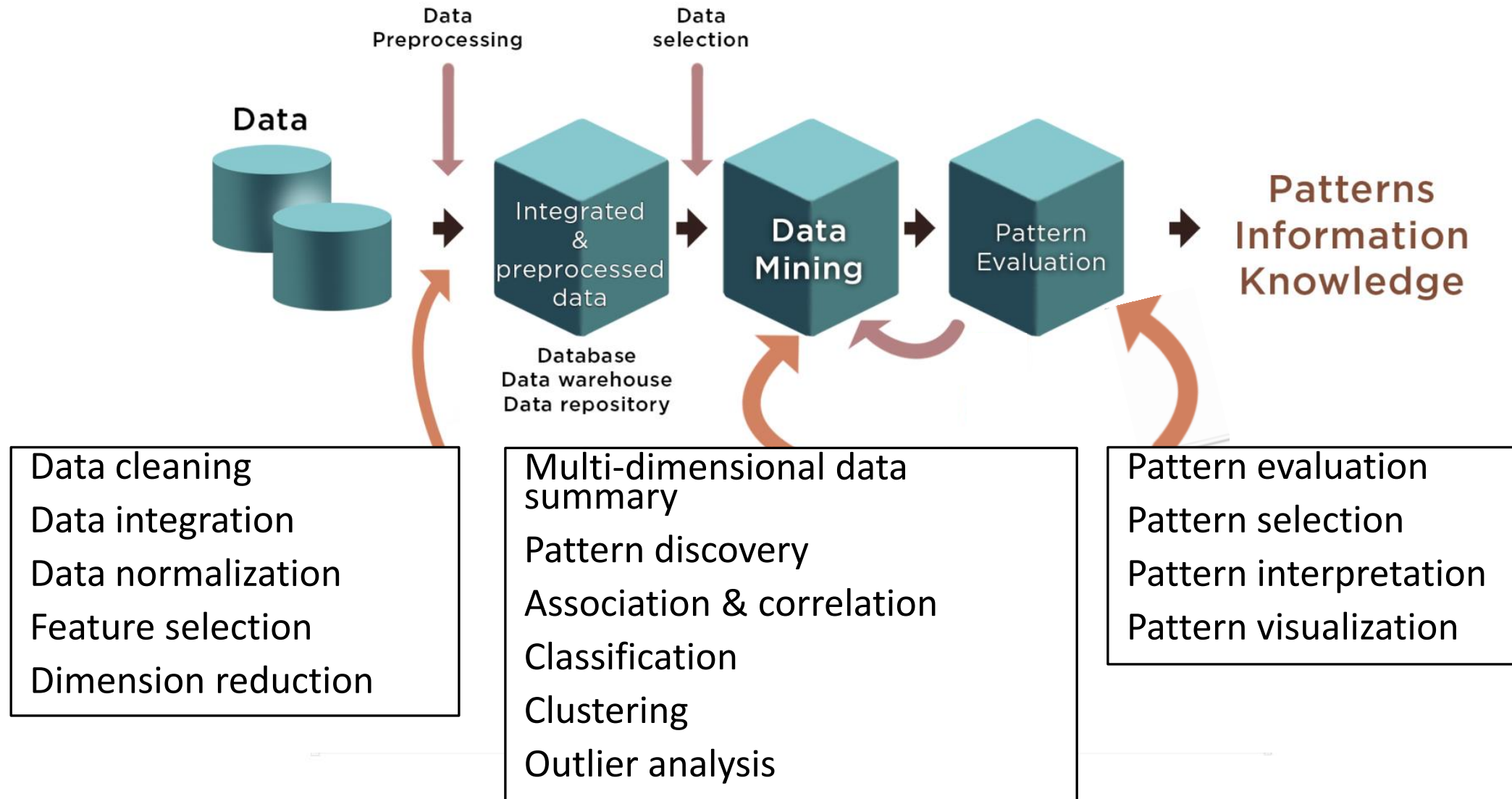
   ❑ Data Mining from Application Point of View

Why Data Mining?

❑ Explosive Growth of Data

  ❑ Data explosion: Our capability of generating, collecting, storing, and managing data has grown tremendously in the last 50 years

  ❑ Major sources of abundant data

    ❑ Web and computerized society: News, digital media, social networks, e-commerce, transactions, stocks, …

    ❑ Science and engineering: Remote sensing, bioinformatics, scientific simulation, …

❑ We are drowning in data but starving for knowledge!

❑ "Necessity is the mother of invention"—Data mining—Automated and scalable analysis of massive data sets

What Is Data Mining?

❑ Data mining: Knowledge discovery from data (KDD)

  ❑ Extraction of interesting (<u>non-trivial, implicit, previously unknown</u> and <u>potentially useful</u>) patterns or knowledge from massive data

  ❑ Data mining: A misnomer?

    ❑ Knowledge mining from data!

❑ Data mining: Its relationship with other disciplines

  ❑ Machine learning, pattern recognition, statistics, databases, business intelligence, big data, ……

Data Mining: A Knowledge Discovery (KDD) Process

Data Preprocessing

Data selection

Data

Integrated & preprocessed data

Database
Data warehouse
Data repository

Data Mining

Pattern Evaluation

Patterns
Information
Knowledge

Data cleaning

Data integration

Data normalization

Feature selection

Dimension reduction

Multi-dimensional data summary

Pattern discovery

Association & correlation

Classification

Clustering

Outlier analysis

Pattern evaluation

Pattern selection

Pattern interpretation

Pattern visualization

Data Mining: A Multi-Dimensional View

❑ Data mining can be viewed from multiple angles

    ❑ **Data** to be mined

    ❑ **Knowledge** to be mined

    ❑ **Methodologies** or techniques utilized

    ❑ **Applications** adapted

    ❑ ......

Data View: On What Kinds of Data?

❑ **Structured and semi-structured data**

  ❑ Relational data/object-relational data

  ❑ Data warehouse data

  ❑ Transactional data

❑ **Unstructured data**

  ❑ Text data and Web data

  ❑ Spatial and spatiotemporal data

  ❑ Multimedia data

  ❑ Data streams and sensor data

  ❑ Time-series data, temporal data, sequence data

  ❑ Graphs, social networks and information networks

Knowledge View: Knowledge to Be Mined

❑ **Data summary in multidimensional space**

   ❑ Data cube and OLAP (On-Line Analytical Processing)

❑ **Pattern discovery**

   ❑ Mining frequent patterns, association and correlation

   ❑ Applying pattern mining in many other tasks

❑ **Classification and predictive modeling**

   ❑ Model construction based on some training examples

   ❑ Prediction of new data based on constructed models

❑ **Cluster analysis:** How to group data to form new categories?

❑ **Outlier analysis:** Discovery of anomalies and rare events

❑ **Trend and evolution analysis**

Methodology View: Confluence of Multiple Disciplines
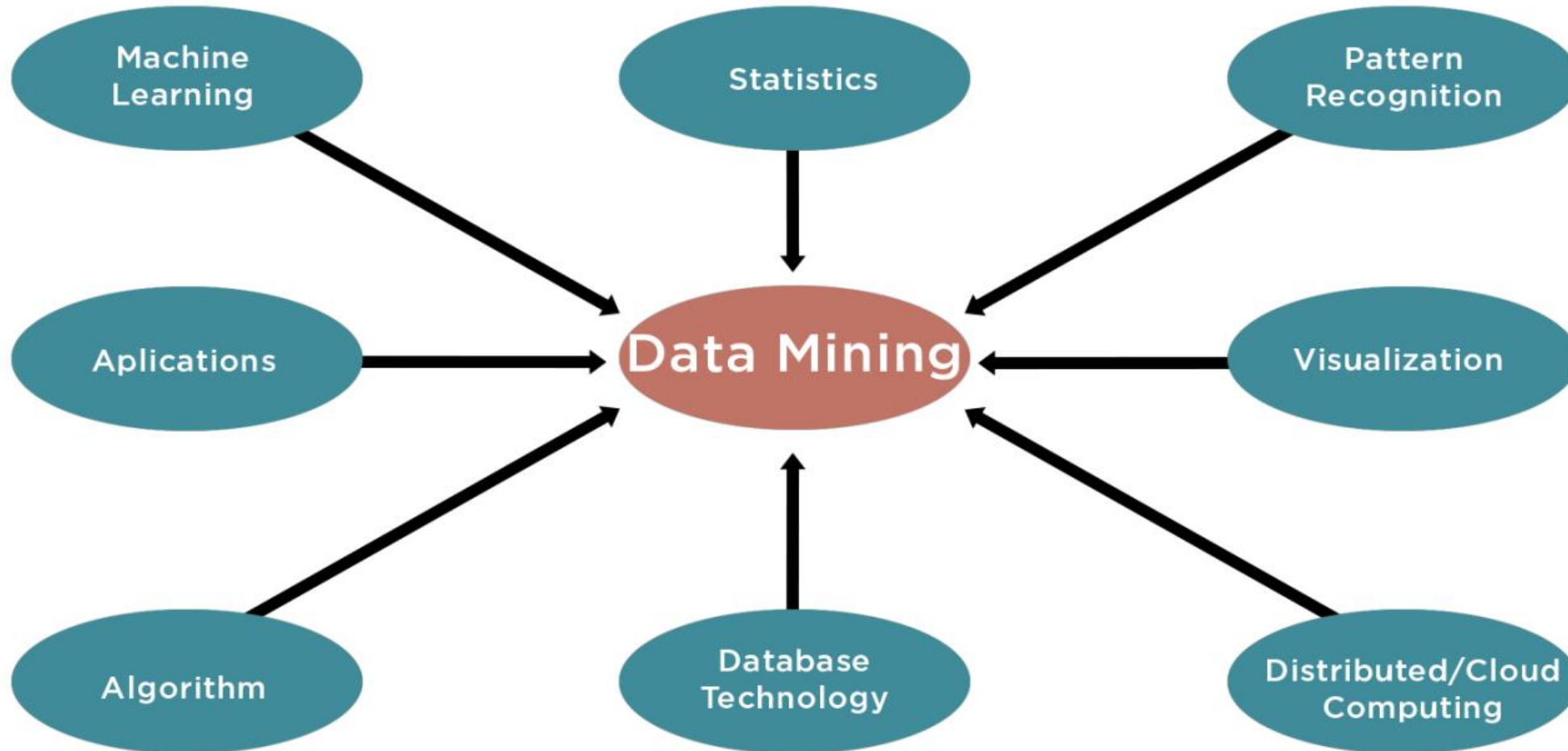
Application View: Diverse Applications

❑ Mining text data and mining the Web

 ❑ Web page classification and ranking, Weblog analysis, recommender systems, …

❑ Mining business data

 ❑ Transaction data, market basket analysis, fraud detection, …

❑ Data mining and software/system engineering, e.g., mining software bugs

❑ Mining biological and medical data

 ❑ Gene, protein, microarray data, biological networks

❑ Mining social and information networks

 ❑ Community discovery, information propagation, …

❑ Invisible data mining

Recommended Readings

❑ Text books on Data Mining:

   ❑ Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011

   ❑ Mohammed J. Zaki and Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014

   ❑ Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining, 2nd ed., Wiley, 2014

❑ Reference book on Pattern Discovery:

   ❑ Charu Aggarwal and Jiawei Han (eds.), Frequent Pattern Mining, Springer, 2014