INTRODUCTION TO DATA MINING

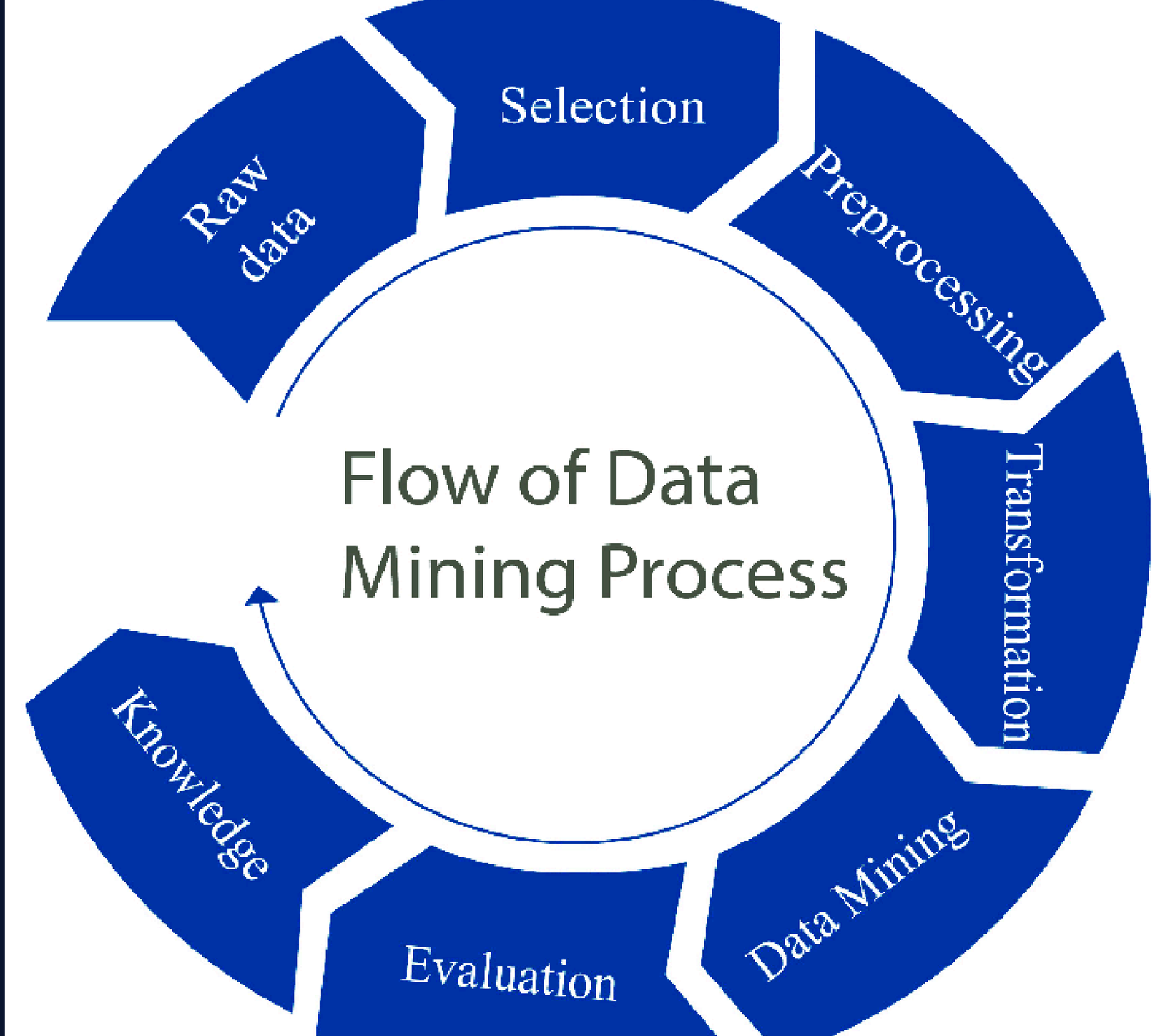
- KDD Process
- Data Preprocessing
- Data Warehousing
- Association Rule Mining
- Classification
- Clustering and Anomaly Detection
- Sequence Analysis

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized
  - 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 mas

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
- 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

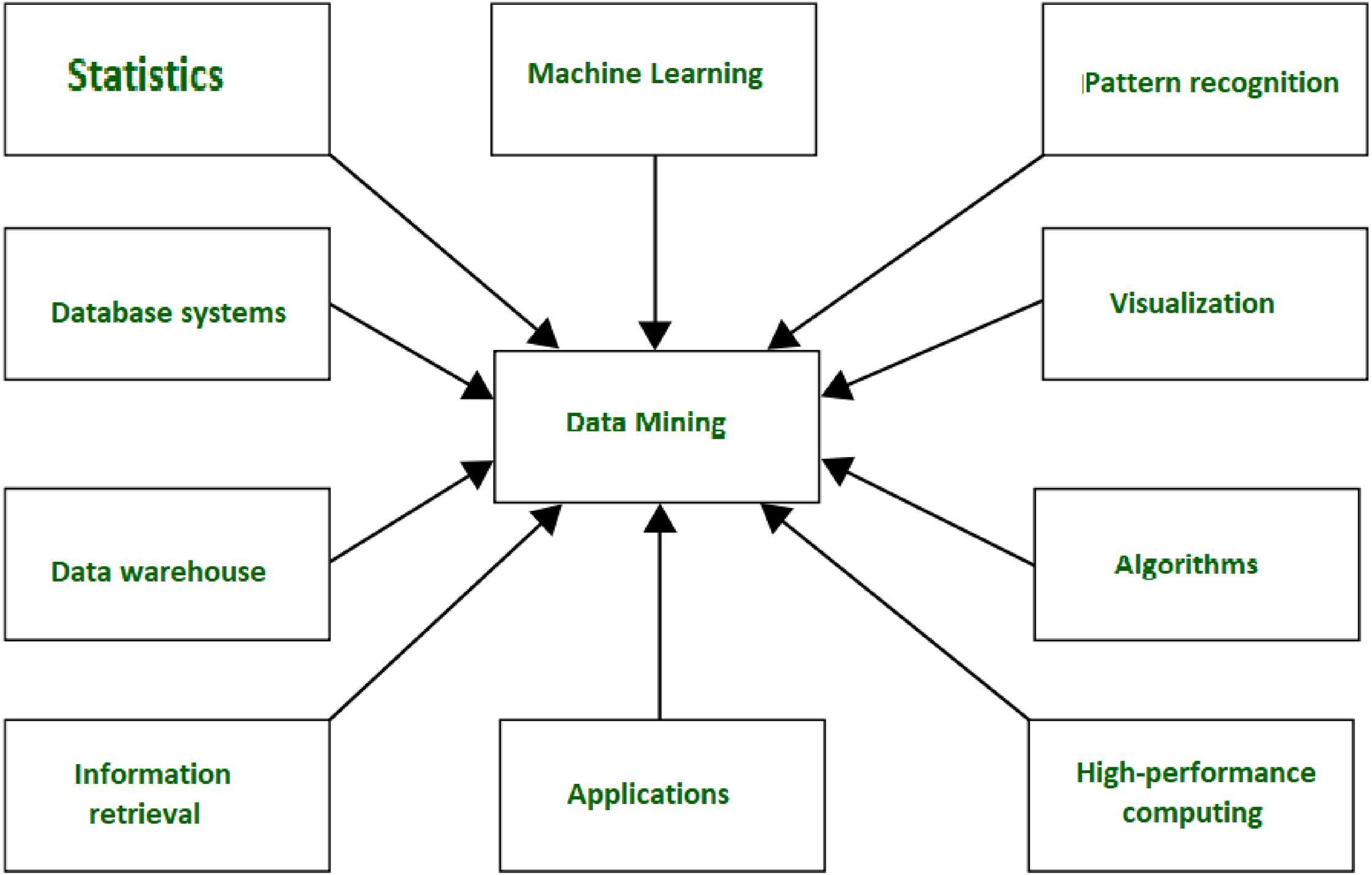

Flow of Data Mining Process
Raw data
Selection
Preprocessing
Transformation
Data Mining
Evaluation
Knowledge

DATA MINING

FIVE ESSENTIAL STAGES

1. PROBLEM DEFINITION
2. DATA COLLECTION
3. DATA ANALYSIS
4. EVALUATION
5. DEPLOYMENT

dataaspirant.com


Statistics
Machine Learning
Pattern recognition
Database systems
Visualization
Data Mining
Data warehouse
Algorithms
Information retrieval
Applications
High-performance computing

Why Not Traditional Data Analysis?

Tremendous amount of data

- Algorithms must be highly scalable to handle such as tera-bytes of data

High-dimensionality of data

- Micro-array may have tens of thousands of dimensions

High complexity of data

- Data streams and sensor data
- Time-series data, temporal data, sequence data
- Structure data, graphs, social networks and multi-linked data
- Heterogeneous databases and legacy databases
- Spatial, spatiotemporal, multimedia, text and Web data
- Software programs, scientific simulations

New and sophisticated applications

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Data Mining Functionalities

Multidimensional concept description: Characterization and discrimination

- Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions

Frequent patterns, association, correlation vs. causality

- Tea → Sugar [0.5%, 75%] (Correlation or causality?)

Classification and prediction

- 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

Data Mining Functionalities

Cluster analysis

- Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
- Maximizing intra-class similarity & minimizing interclass similarity

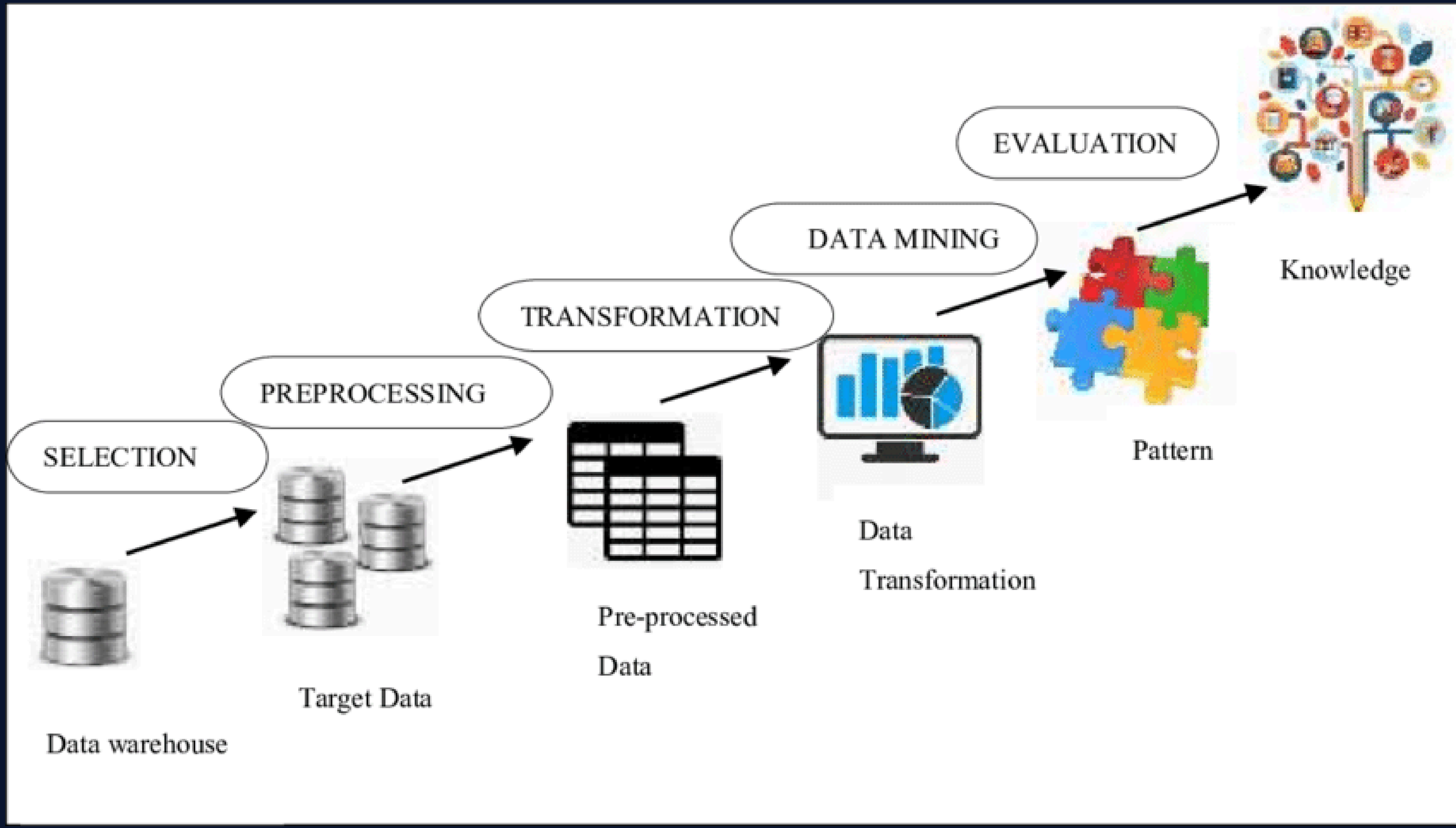
Outlier analysis

- 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

- Trend and deviation: e.g., regression analysis
- Sequential pattern mining: e.g., digital camera → large SD memory
- Periodicity analysis
- Similarity-based analysis

Other pattern-directed or statistical analyses


EVALUATION
DATA MINING
TRANSFORMATION
PREPROCESSING
SELECTION
Knowledge
Pattern
Data
Transformation
Pre-processed
Data
Target Data
Data warehouse

Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of 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
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy