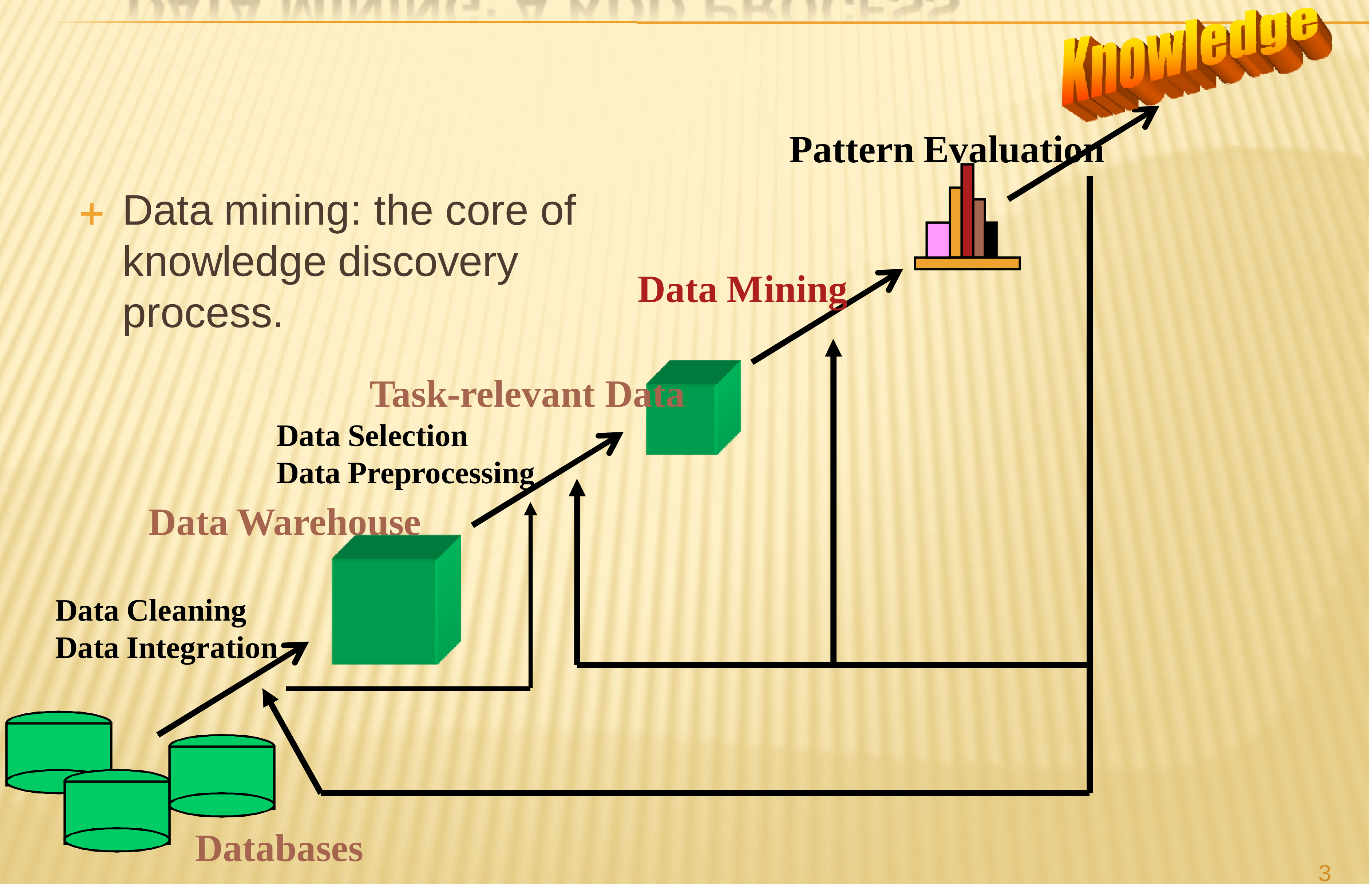
COURSE NAME:
DATA WAREHOUSING & DATA MINING

LECTURE 12
TOPICS TO BE COVERED:

- KDD versus data mining
- Data mining techniques
- Tools and applications.

DATA MINING: A KDD PROCESS

- Data mining: the core of knowledge discovery process.



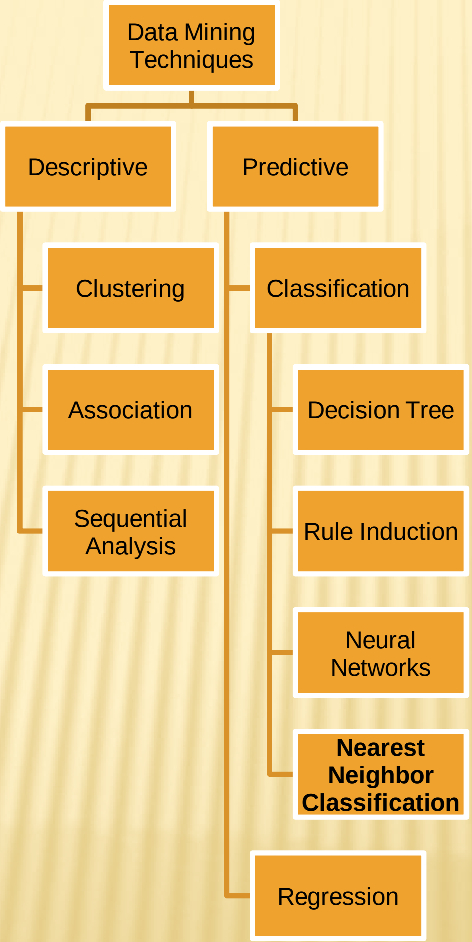
STEPS OF A KDD PROCESS

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- **Data cleaning** and preprocessing: (may take 60% of effort!)
- **Data reduction and transformation**:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- **Data mining**: search for patterns of interest
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

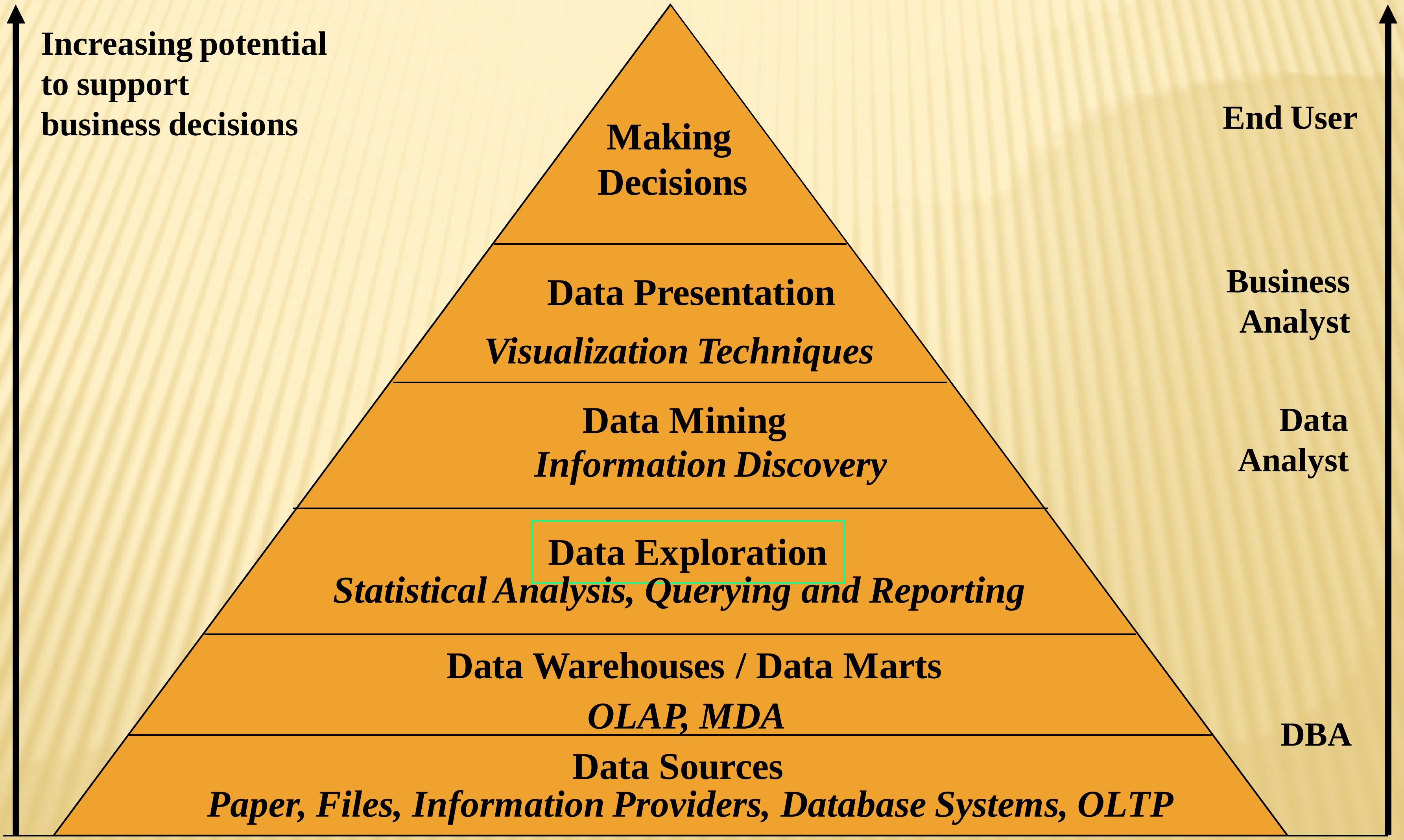
STEPS OF A KDD PROCESS

- **Data cleaning** (to remove noise and inconsistent data)
- **Data integration** (where multiple data sources may be combined)
- **Data selection** (where data relevant to the analysis task are retrieved from the database)
- **Data transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
- **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns)
- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

DATA MINING TECHNIQUES



DATA MINING AND BUSINESS INTELLIGENCE



Increasing potential to support business decisions

Making Decisions

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Analysis, Querying and Reporting

Data Warehouses / Data Marts

OLAP, MDA

Data Sources

Paper, Files, Information Providers, Database Systems, OLTP

End User

Business Analyst

Data Analyst

DBA

DBMS, OLAP, AND DATA MINING

	DBMS	OLAP	Data Mining
Task	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
Type of result	Information	Analysis	Insight and Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

MAJOR ISSUES IN DATA WAREHOUSING AND MINING

- Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
- Performance and scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods

MAJOR ISSUES IN DATA WAREHOUSING AND MINING

- Issues relating to the diversity of data types
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)
- Issues related to applications and social impacts
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy

DATA MINING APPLICATIONS: OTHER APPLICATIONS

- Customer segmentation
  - All industries can take advantage of DM to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis.
- Manufacturing
  - Through choice boards, manufacturers are beginning to customize products for customers; therefore they must be able to predict which features should be bundled to meet customer demand.
- Warranties
  - Manufacturers need to predict the number of customers who will submit warranty claims and the average cost of those claims.
- Frequent flier incentives
  - Airlines can identify groups of customers that can be given incentives to fly more.