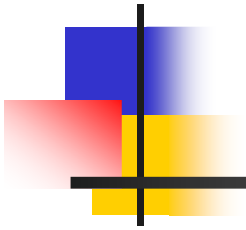


Data Mining: Concepts and Techniques



Dr. Saurabh Pal
Associate Professor & Head
Department of Computer Applications
VBS Purvanchal University, Jaunpur

Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Major issues in data mining

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 society
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, scientific simulation, ...
 - Society and everyone: news, digital cameras,
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

Evolution of Database Technology

- 1960s:
 - Data collection, database creation, IMS and network DBMS
- 1970s:
 - Relational data model, relational DBMS implementation
- 1980s:
 - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
 - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
 - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
 - Stream data management and mining
 - Data mining and its applications
 - Web technology (XML, data integration) and global information systems

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 name
 - Knowledge discovery in databases (KDD)
- Watch out: Is everything “data mining”?
 - Query processing
 - Expert systems or statistical programs

Why Data Mining?—Potential Applications

- Data analysis and decision support
 - Market analysis and management
 - Target marketing, customer relationship management (CRM), market basket analysis, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, quality control, competitive analysis
 - Fraud detection and detection of unusual patterns (outliers)

Why Data Mining?—Potential Applications

- Other Applications
 - Text mining (news group, email, documents) and Web mining
 - Stream data mining
 - Bioinformatics and bio-data analysis

Market Analysis and Management

- Where does the data come from?
 - Credit card transactions, discount coupons, customer complaint calls
- Target marketing
 - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
 - Determine customer purchasing patterns over time

Market Analysis and Management

- Cross-market analysis
 - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
 - What types of customers buy what products
- Customer requirement analysis
 - Identifying the best products for different customers
 - Predict what factors will attract new customers

Fraud Detection & Mining Unusual Patterns

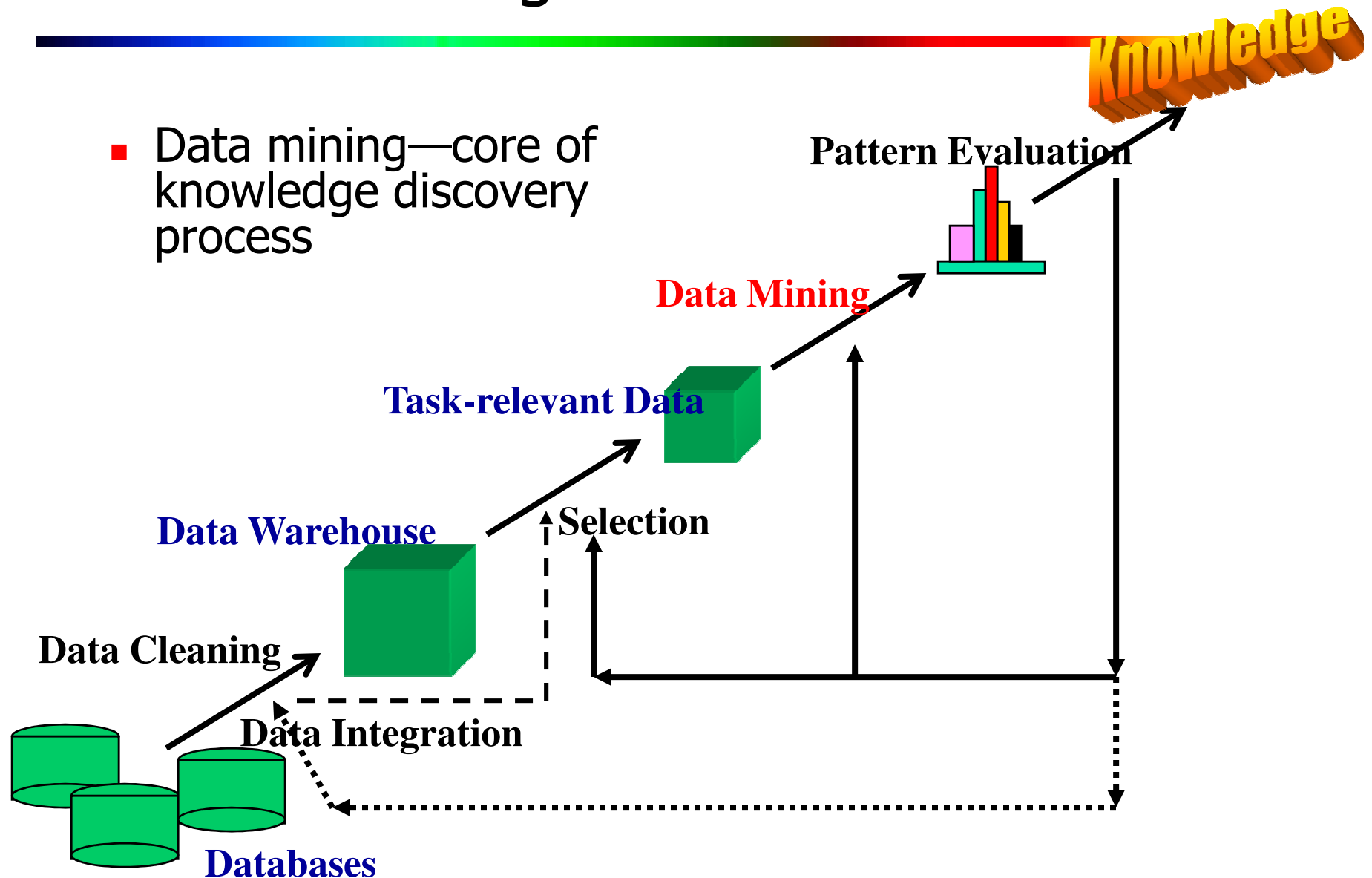
- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
 - Medical insurance
 - Professional patients, and ring of doctors
 - Unnecessary or correlated screening tests
 - Telecommunications:
 - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
 - Retail industry
 - Analysts estimate that 38% of retail shrink is due to dishonest employees

Other Applications

- Internet Web Surf-Aid
 - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Data Mining: A KDD Process

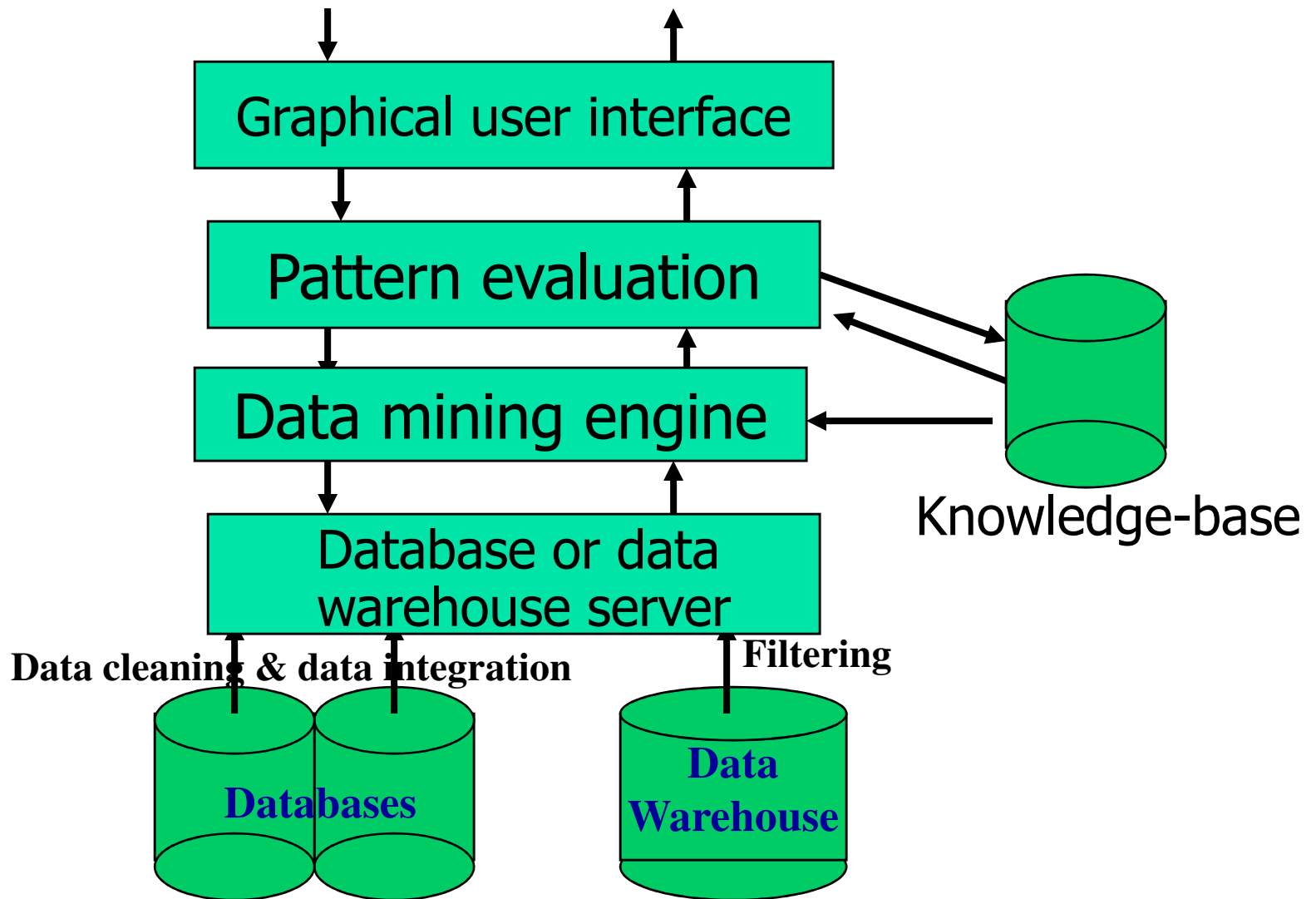
- Data mining—core of knowledge discovery process



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.
- 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

Architecture: Typical Data Mining System



Data Mining: On What Kinds of Data?

- Relational database
- Data warehouse
- Transactional database
- Advanced database and information repository
 - Spatial and temporal data
 - Time-series data
 - Stream data
 - Multimedia database
 - Text databases & WWW

Data Mining Functionalities

- Concept description: Characterization and discrimination
 - Generalize, summarize, and contrast data characteristics
- Association (correlation and causality)
 - Diaper → Beer [0.5%, 75%]
- Classification and Prediction
 - Construct models (functions) that describe and distinguish classes or concepts for future prediction
 - Presentation: decision-tree, classification rule, neural network

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: a data object that does not comply with the general behavior of the data
 - Useful in fraud detection, rare events analysis
- Trend and evolution analysis
 - Trend and deviation: regression analysis
 - Sequential pattern mining, periodicity analysis