



# Data Mining

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

- Why Data Mining?
- What is Data Mining?
- What are the typical tasks?
- What are the primitives?
- What are the typical applications?
- What are the major issues?

Why

Data Mining?

# Why Data Mining?

## “Necessity is the mother of invention”

- ❑ Explosive Growth of Data
  - ▶ Terabytes of available data
  - ▶ Data collections and data availability
  - ▶ Major sources of abundant data
  
- ❑ Pressing need for the automated analysis of massive data

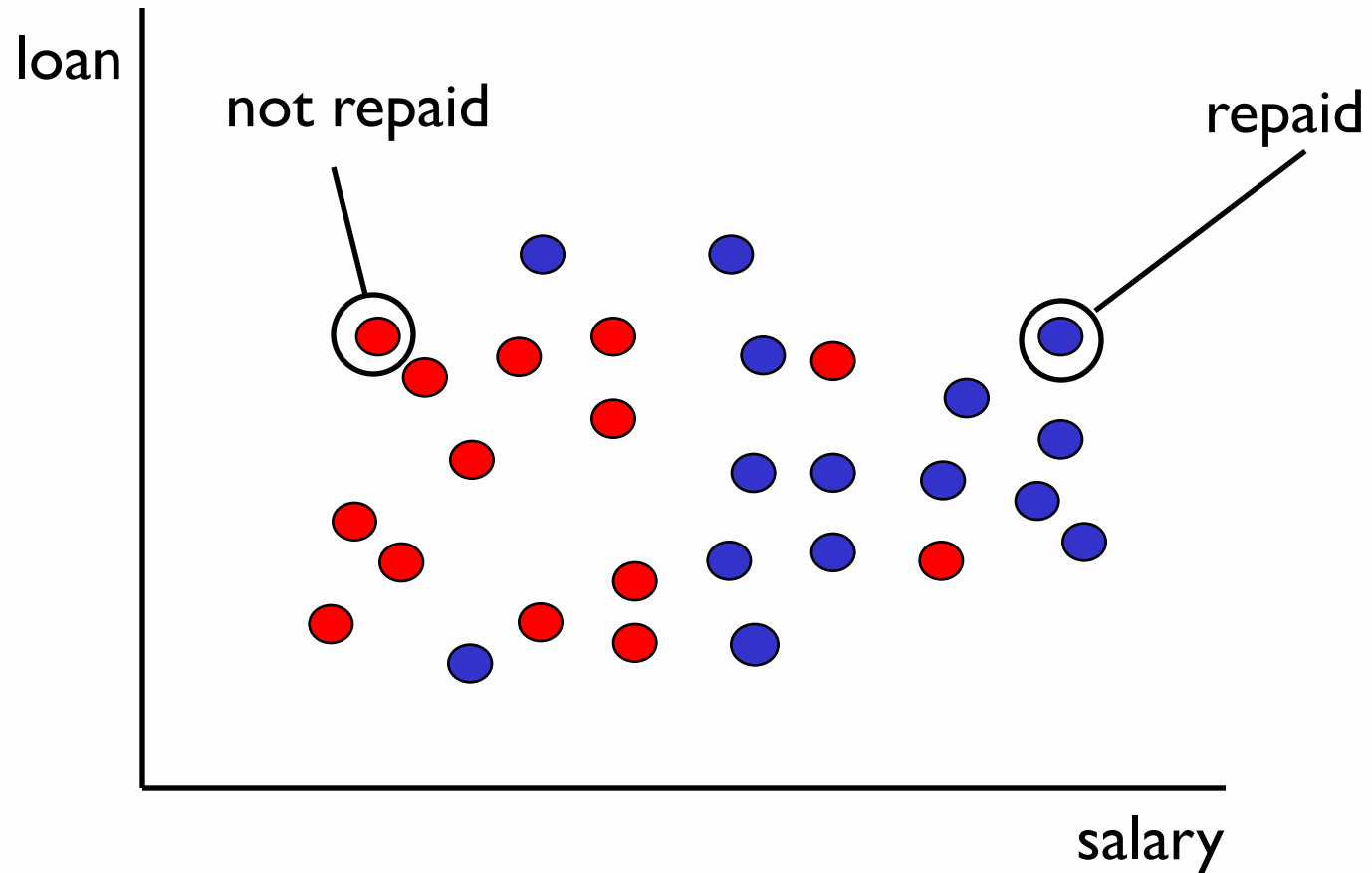
- ❑ 1960s:
  - ▶ Data collection, database creation, IMS and network DBMS
- ❑ 1970s:
  - ▶ Relational data model, relational DBMS implementation
- ❑ 1980s:
  - ▶ RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - ▶ Application-oriented DBMS (spatial, scientific, engineering, etc.)
- ❑ 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)
  - ▶ Global information systems

- ❑ In vitro fertilization
  - ▶ Given: embryos described by 60 features
  - ▶ Problem: selection of embryos that will survive
  - ▶ Data: historical records of embryos and outcome
  
- ❑ Cow culling
  - ▶ Given: cows described by 700 features
  - ▶ Problem: selection of cows that should be culled
  - ▶ Data: historical records and farmers' decisions

- ❑ Customer attrition
  - ▶ Given: customer information for the past months
  - ▶ Problem: predict who is likely to attrite next month, or estimate customer value
  - ▶ Data: historical customer records
  
- ❑ Credit assessment
  - ▶ Given: a loan application
  - ▶ Problem: predict whether the bank should approve the loan
  - ▶ Data: records from other loans

What is  
Data Mining?

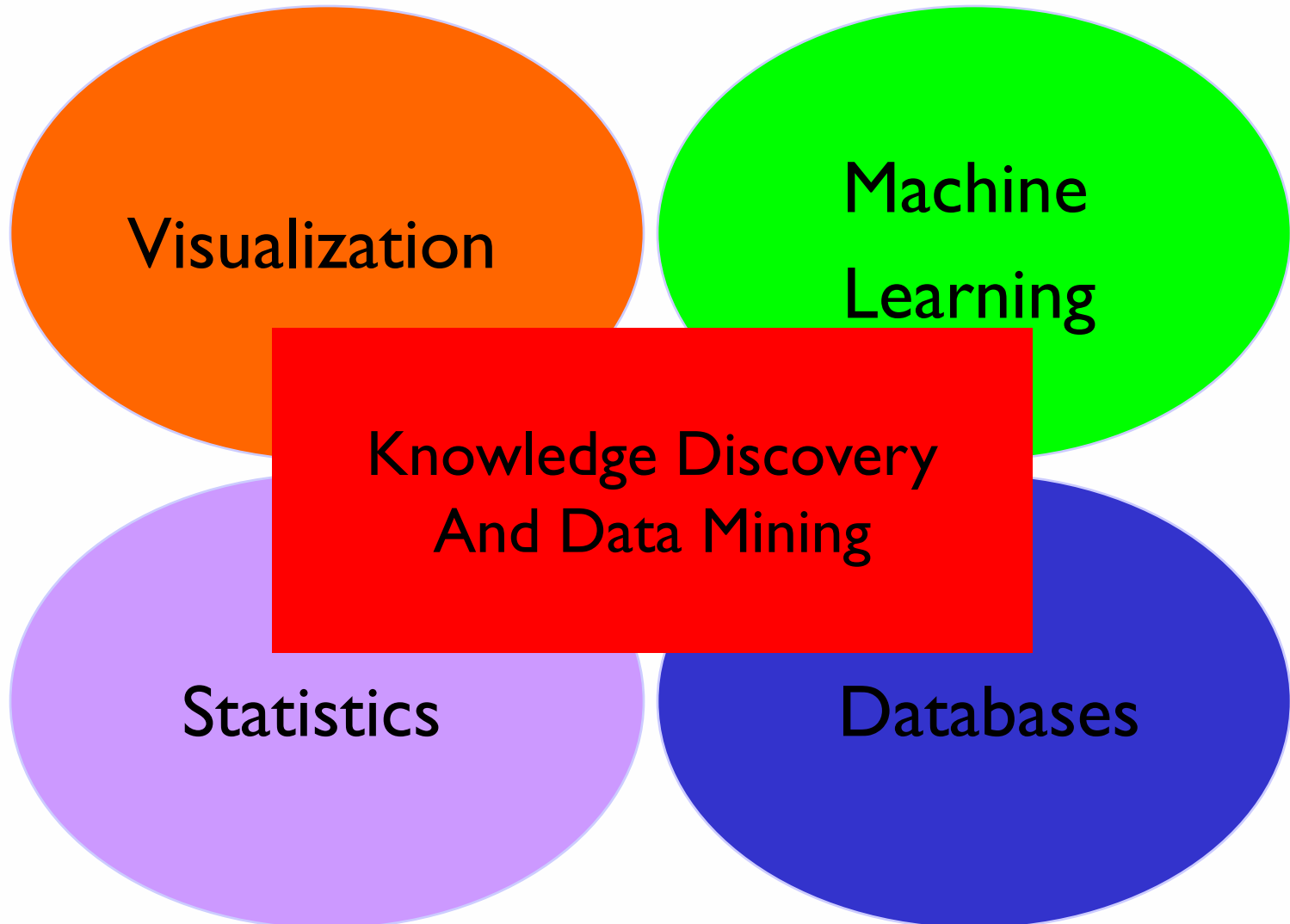
- ❑ The non-trivial process of identifying
  - ▶ valid
  - ▶ novel
  - ▶ potentially useful, and
  - ▶ ultimately understandable patterns in data.
  
- ❑ Alternative names,
  - ▶ Data Fishing, Data Dredging (1960-)
  - ▶ Data Mining (1990-), used by DB and business
  - ▶ Knowledge Discovery in Databases (1989-), used by AI
  - ▶ Business Intelligence, Information Harvesting, Information Discovery, Knowledge Extraction, ...
  
- ❑ Currently, Data Mining and Knowledge Discovery are used interchangeably





- ❑ Is it valid?
  - ▶ The pattern has to be valid with respect to a certainty level (rule true for the 86%)
  
- ❑ Is it novel?
  - ▶ The value  $k$  should be previously unknown or obvious
  
- ❑ Is it useful?
  - ▶ The pattern should provide information useful to the bank for assessing credit risk
  
- ❑ Is it understandable?

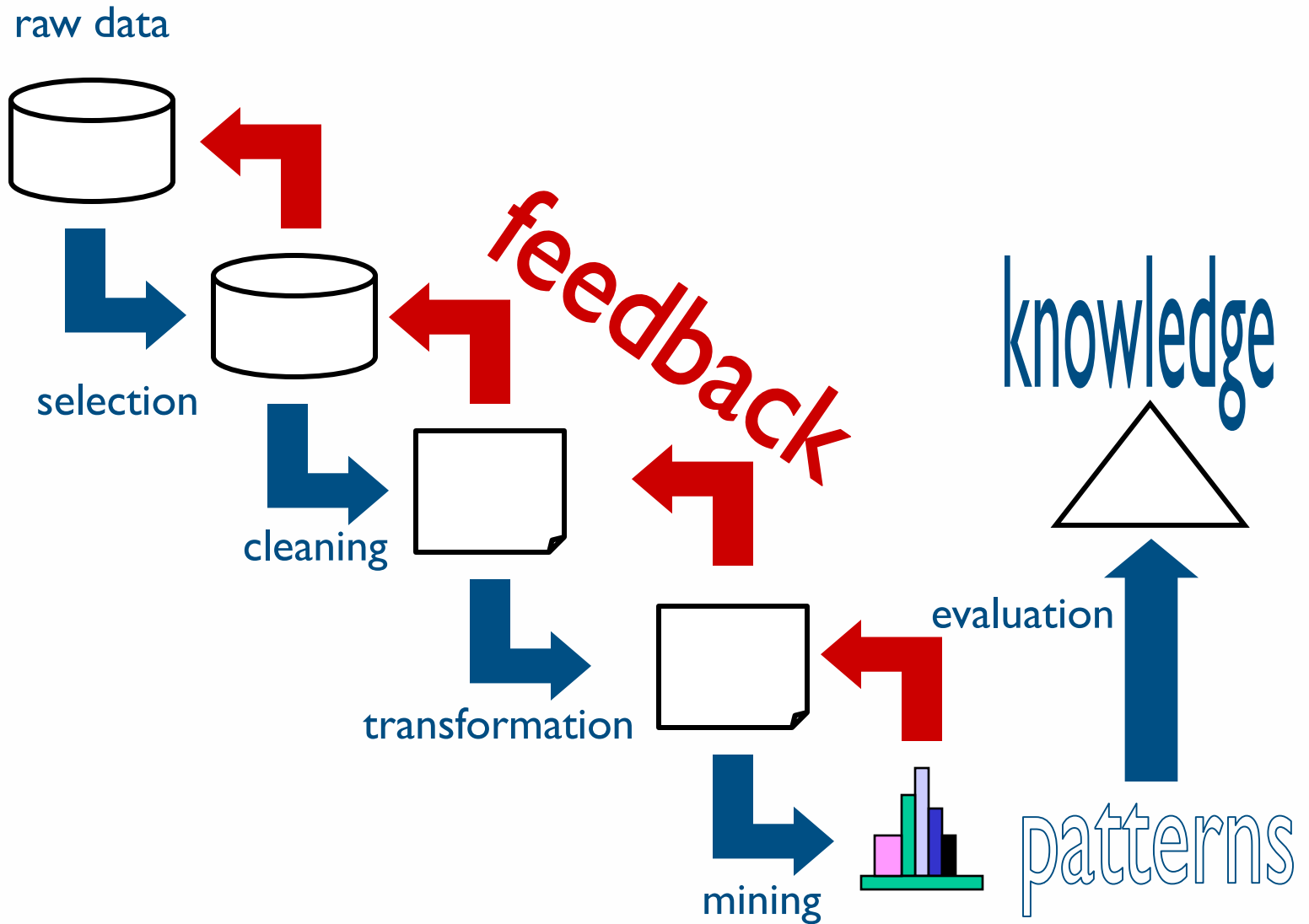
- ❑ Build computer programs that sift through databases automatically, seeking regularities or patterns
  
- ❑ There will be problems
  - ▶ Most patterns are banal and uninteresting
  - ▶ Most patterns are spurious, inexact, or contingent on accidental coincidences in the particular dataset used
  - ▶ Real data is imperfect: Some parts will be garbled, and some will be missing
  
- ❑ Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful



- ❑ Statistics:
  - ▶ more theory-based, focused on testing hypotheses
- ❑ Machine learning
  - ▶ more heuristic, focused on building program that learns, more general than Data Mining
- ❑ Knowledge Discovery
  - ▶ integrates theory and heuristics
  - ▶ focus on the entire process of discovery, including data cleaning, learning, integration and visualization
- ❑ Data Mining
  - ▶ focus on the algorithms to extract patterns from data

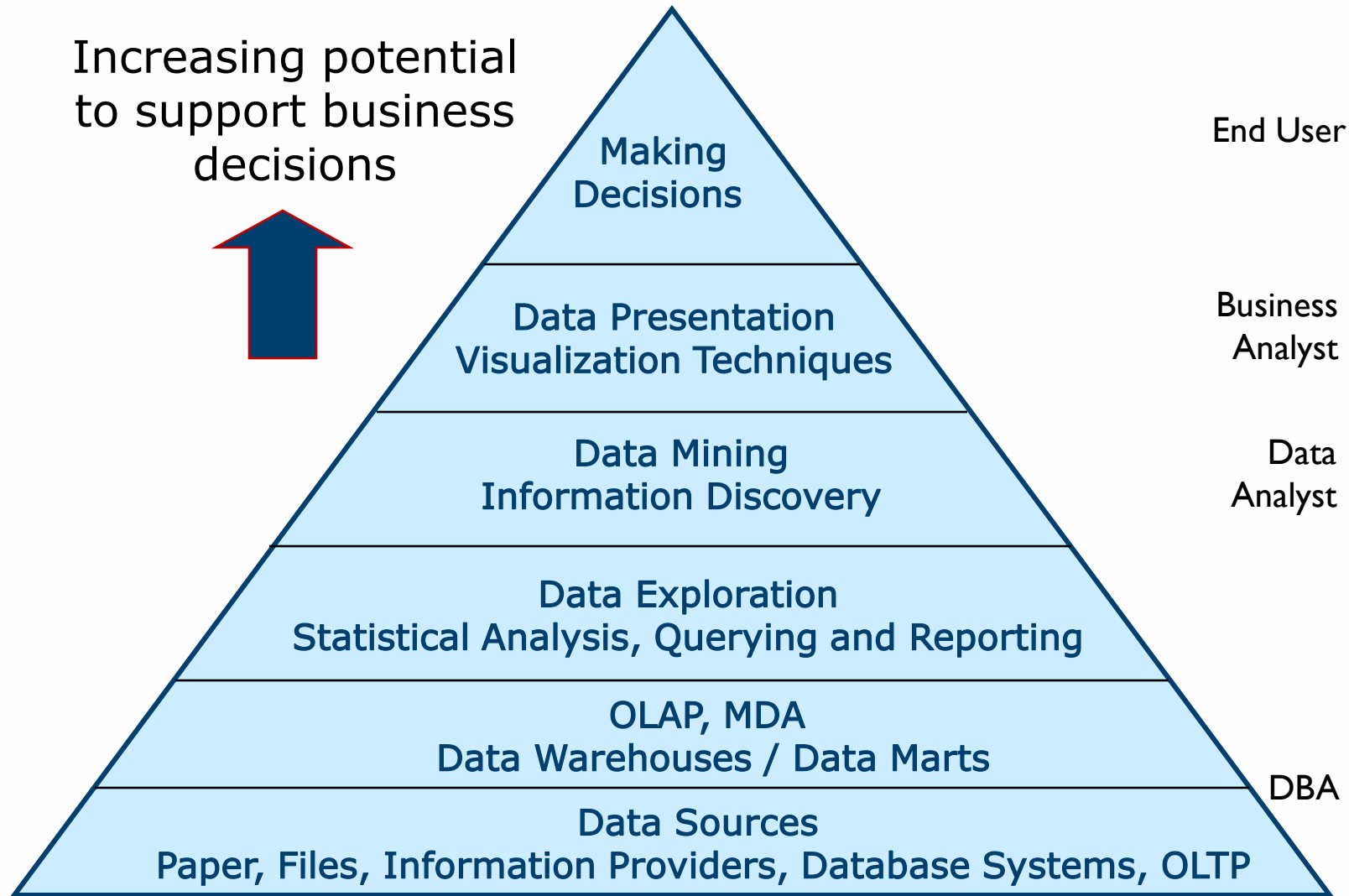
**Distinctions are blurred!**

- ❑ Tremendous amount of data
  - ▶ High scalability to handle terabytes 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



## What are the main steps?

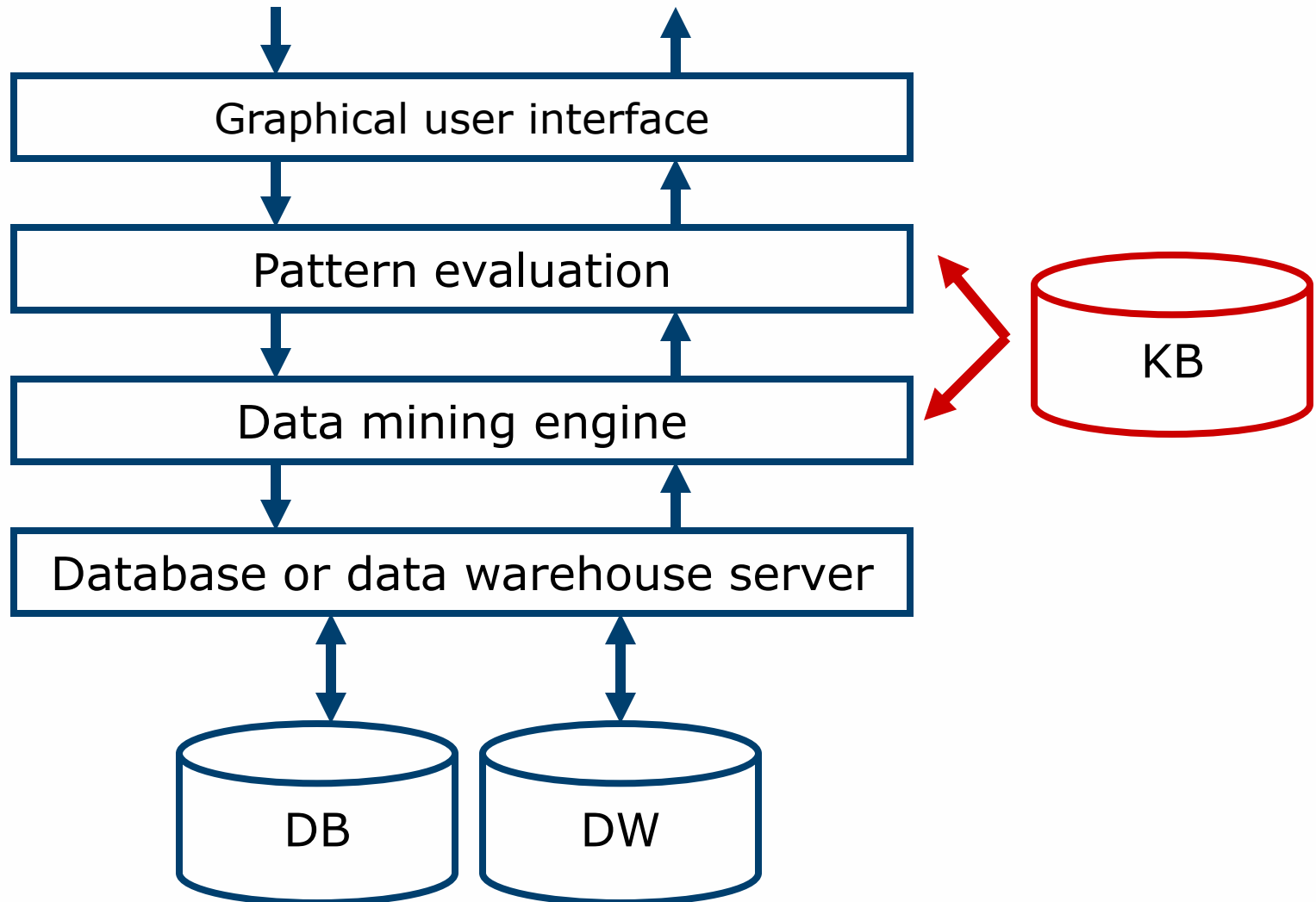
- ❑ Learning the application domain to extract relevant prior knowledge and goals
- ❑ Data selection
- ❑ Data cleaning
- ❑ Data reduction and transformation
- ❑ Mining
  - ▶ Select the mining approach: classification, regression, association, clustering, etc.
  - ▶ Choosing the mining algorithm(s)
  - ▶ Perform mining: search for patterns of interest
- ❑ Pattern evaluation and knowledge presentation
  - ▶ visualization, transformation, removing redundant patterns, etc.
- ❑ Use of discovered knowledge



- ❑ Data mining systems, DBMS, Data warehouse systems coupling
  - ▶ No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- ❑ On-line analytical mining data
  - ▶ integration of mining and OLAP technologies
- ❑ Interactive mining multi-level knowledge
  - ▶ Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- ❑ Integration of multiple mining functions
  - ▶ Characterized classification, first clustering and then association

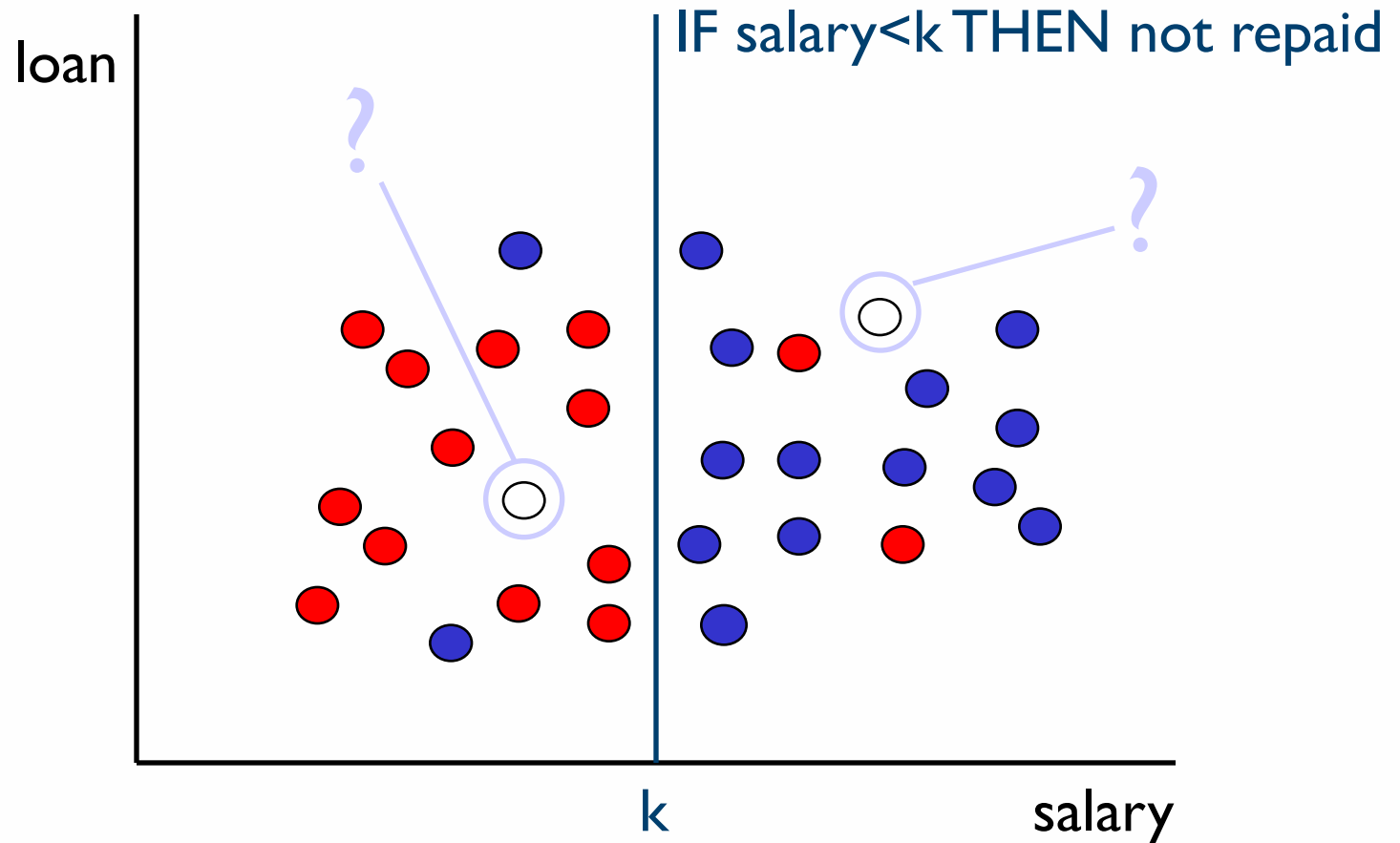
- ❑ No coupling—flat file processing, not recommended
- ❑ Loose coupling
  - ▶ Fetching data from DB/DW
- ❑ Semi-tight coupling—enhanced DM performance
  - ▶ Provide efficient implement a few data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions
- ❑ Tight coupling—A uniform information processing environment
  - ▶ DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query, indexing, query processing methods, etc.

# Architecture of a Typical Knowledge Discovery System



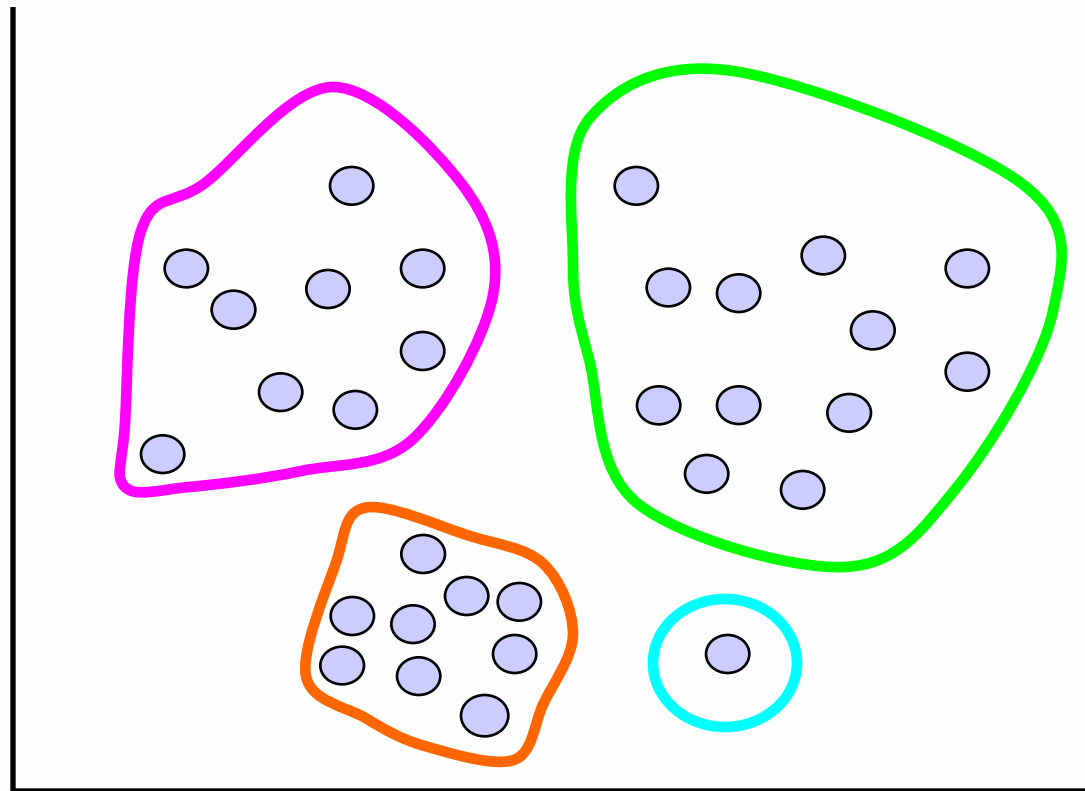
What tasks?

- ❑ Classification: predicting an item class
- ❑ Clustering: finding clusters in data
- ❑ Associations: frequent occurring events...
- ❑ Visualization: to facilitate human discovery
- ❑ Summarization: describing a group
- ❑ Deviation Detection: finding changes
- ❑ Estimation: predicting a continuous value
- ❑ Link Analysis: finding relationship



## □ Classification and Prediction

- ▶ Finding models (functions) that describe and distinguish classes or concepts
- ▶ The goal is to describe the data or to make future prediction
- ▶ E.g., classify countries based on climate, or classify cars based on gas mileage
- ▶ Presentation: decision-tree, classification rule, neural network
- ▶ Prediction: Predict some unknown numerical values



## □ Cluster analysis

- ▶ The class label is unknown
- ▶ Group data to form new classes, e.g., cluster houses to find distribution patterns
- ▶ Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

Bread  
Peanuts  
Milk  
Fruit  
Jam

Bread  
Jam  
Soda  
Chips  
Milk  
Fruit

Steak  
Jam  
Soda  
Chips  
Bread

Jam  
Soda  
Peanuts  
Milk  
Fruit

**Is there something interesting?**

Jam  
Soda  
Chips  
Milk  
Bread

Fruit  
Soda  
Chips  
Milk

Fruit  
Soda  
Peanuts  
Milk

Fruit  
Peanuts  
Cheese  
Yogurt

## □ Association Rule Mining

- ▶ Finds interesting associations and/or correlation relationships among large set of data items.
- ▶ E.g., 98% of people who purchase tires and auto accessories also get automotive services done

- ❑ Outlier analysis
  - ▶ Outlier: a data object that does not comply with the general behavior of the data
  - ▶ It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
  
- ❑ Trend and evolution analysis
  - ▶ Trend and deviation: regression analysis
  - ▶ Sequential pattern mining, periodicity analysis
  - ▶ Similarity-based analysis
  
- ❑ Text Mining, Graph Mining, Data Streams
  
- ❑ Other pattern-directed or statistical analyses

# Are all the “Discovered” Patterns Interesting?

- ❑ Data Mining may generate thousands of patterns, not all of them are interesting.
  
- ❑ Interestingness measures
  - ▶ A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
  
- ❑ Objective vs. subjective interestingness measures
  - ▶ Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - ▶ Subjective: based on user’s belief in the data, e.g., unexpectedness, novelty, etc.

# Can we find all and only interesting patterns?

- ❑ **Completeness:** Find all the interesting patterns
  - ▶ Can a data mining system find all the interesting patterns?
  - ▶ Association vs. classification vs. clustering
  
- ❑ **Optimization:** Search for only interesting patterns:
  - ▶ Can a data mining system find only the interesting patterns?
  - ▶ Approaches
    - First generate all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

- ❑ General functionality
  - ▶ Descriptive data mining
  - ▶ Predictive data mining
  
- ❑ Different views, different classifications
  - ▶ Kinds of data to be mined
  - ▶ Kinds of knowledge to be discovered
  - ▶ Kinds of techniques utilized
  - ▶ Kinds of applications adapted

What

primitives?

- ❑ Task-relevant data
- ❑ Type of knowledge to be mined
- ❑ Background knowledge
- ❑ Pattern interestingness measurements
- ❑ Visualization/presentation of discovered patterns

# Primitive 1: Task-Relevant Data

- Database or data warehouse name
- Database tables or data warehouse cubes
- Condition for data selection
- Relevant attributes or dimensions
- Data grouping criteria

## Primitive 2: Types of Knowledge to Be Mined

- Characterization
- Discrimination
- Association
- Classification/prediction
- Clustering
- Outlier analysis
- Other data mining tasks

# Primitive 3: Background Knowledge

- ❑ A typical kind of background knowledge: Concept hierarchies
- ❑ Schema hierarchy
  - ▶ E.g., Street < City < ProvinceOrState < Country
- ❑ Set-grouping hierarchy
  - ▶ E.g., {20-39} = young, {40-59} = middle\_aged
- ❑ Operation-derived hierarchy
  - ▶ email address: hagonzal@cs.uiuc.edu
  - ▶ login-name < department < university < country
- ❑ Rule-based hierarchy
  - ▶  $\text{LowProfitMargin}(X) \leq \text{Price}(X, P1) \text{ and } \text{Cost}(X, P2)$   
and  $(P1 - P2) < \$50$

# Primitive 4: Pattern Interestingness Measure

- Simplicity
- Certainty
- Utility
- Novelty

# Presentation of Discovered Patterns

- ❑ Different backgrounds/usages may require different forms of representation
  - ▶ E.g., rules, tables, crosstabs, pie/bar chart, etc.
  
- ❑ Concept hierarchy is also important
  - ▶ Discovered knowledge might be more understandable when represented at high level of abstraction
  - ▶ Interactive drill up/down, pivoting, slicing and dicing provide different perspectives to data
  
- ❑ Different kinds of knowledge require different representation: association, classification, clustering, etc.

What issues?

- ❑ 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

# Summary

- ❑ Data mining: Discovering interesting patterns from large amounts of data
- ❑ A natural evolution of database technology, in great demand, with wide applications
- ❑ A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- ❑ Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- ❑ Data mining systems and architectures
- ❑ Major issues in data mining