Business Intelligence

Data Mining

Lecture Outline

- 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?

"Necessity is the mother of invention"

Explosive Growth of Data

Pressing need for the automated analysis of massive data

Emerged in the late 1980s

Major developments in the mid 1990s

Evolution of Technology

- 1960s: data collection, database creation, & 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, web technology (XML, data integration), global information systems
- 2010s: social networks

What is the Commercial Viewpoint?

- Huge amounts of data being collected and warehoused everyday
  - Web data, e-commerce
  - Purchases at department stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful

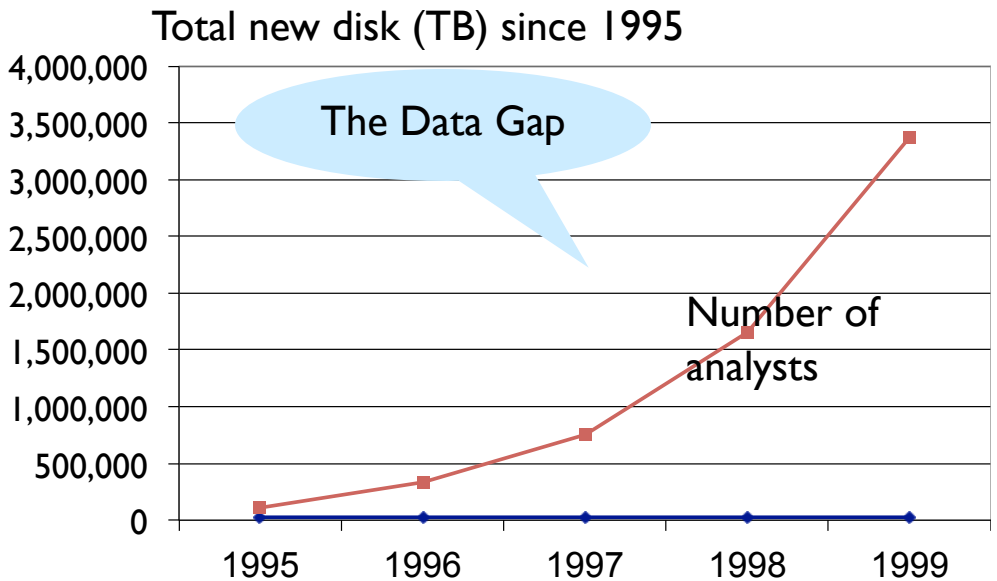
- Competitive pressure is strong to provide better, customized services (e.g., CRM or Customer Relationship Management)

What is the Scientific Viewpoint?

- Data collected and stored at enormous speeds (GB/hour or more)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

Why Mining Large Datasets?

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all!



Examples

Examples

- 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

Examples

- Customer attrition
  - Given: customer information for the past months
  - Problem: predict who is likely to leave 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?

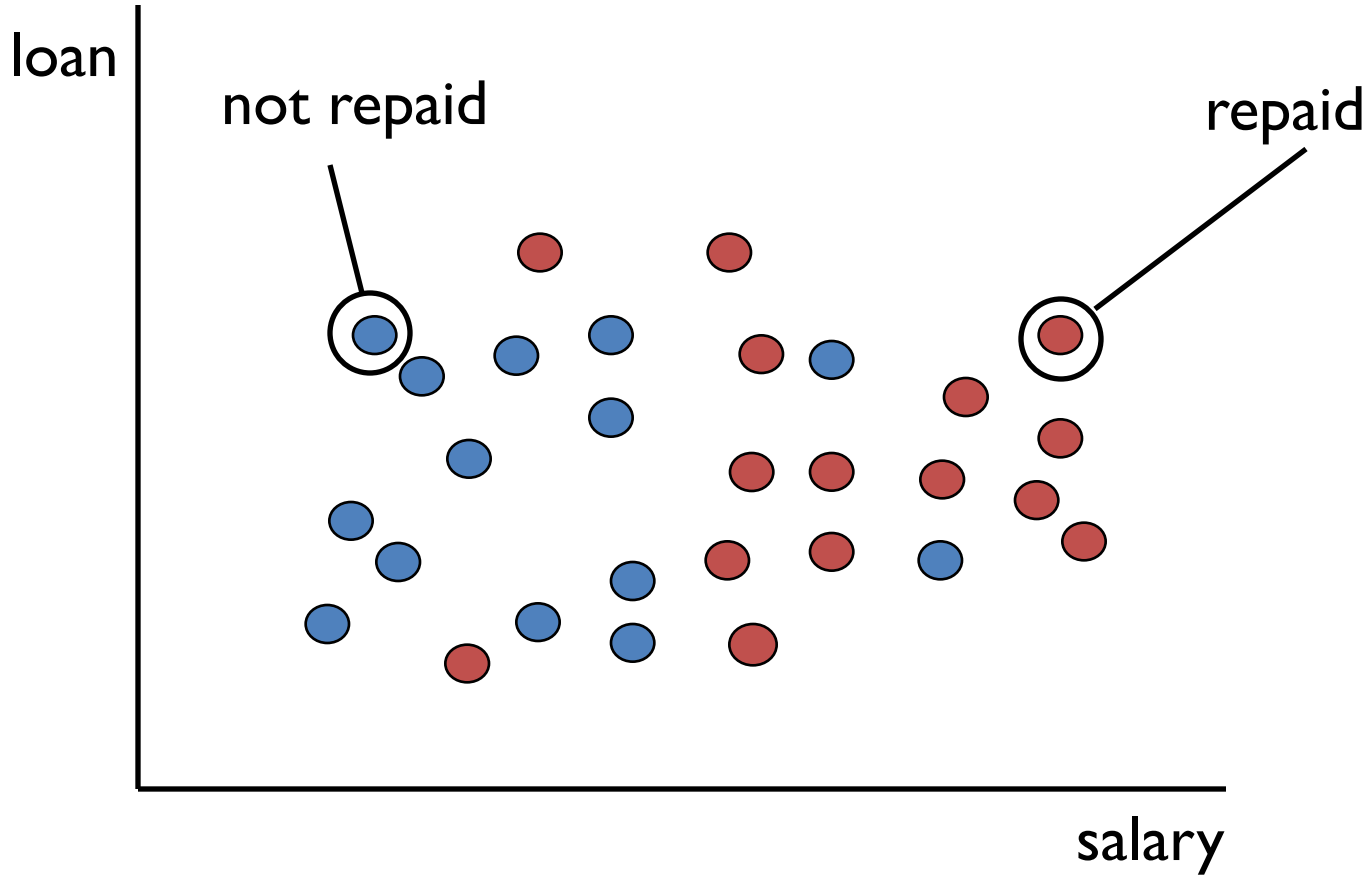
What is Data Mining?

- The non-trivial process of identifying (1) valid, (2) novel, (3) potentially useful, and (4) understandable patterns in data.

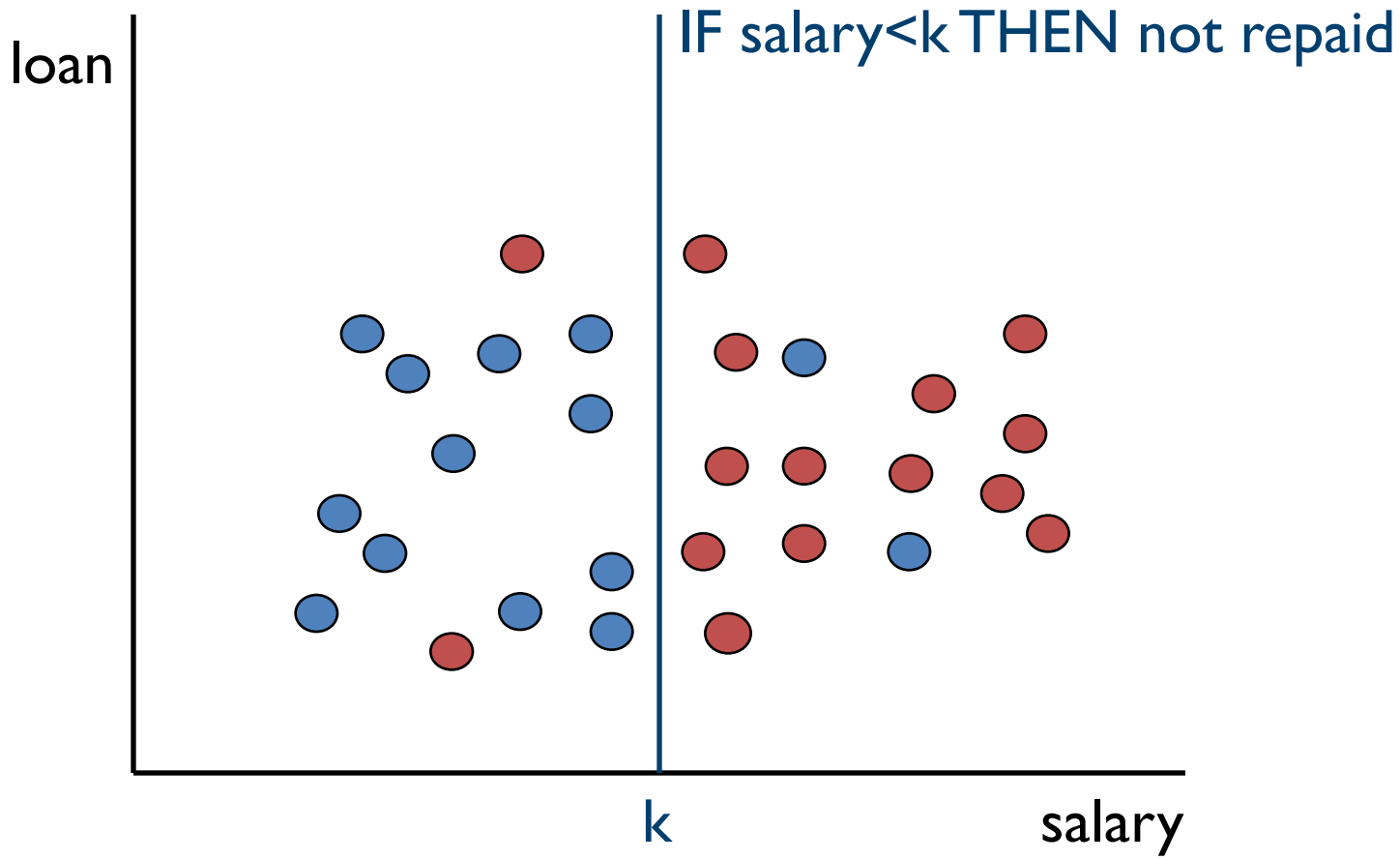
- Alternative names,
  - Data Fishing, Data Dredging (1960-)
  - Data Mining (1990-), used by DB and business
  - Knowledge Discovery in Databases (1989-), used by AI
  - Business Intelligence, Information Harvesting, Information Discovery, Knowledge Extraction, ...
  - Currently, Data Mining and Knowledge Discovery are used interchangeably

- Data Mining is not looking up in the phone directory, it is not querying a Web search engine for information about "Amazon"

Example: Credit Risk



Example: Credit Risk



Example: Credit Risk

- 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 and not obvious
- Is it useful?
  - The pattern should provide information useful to the bank for assessing credit risk
- Is it understandable?

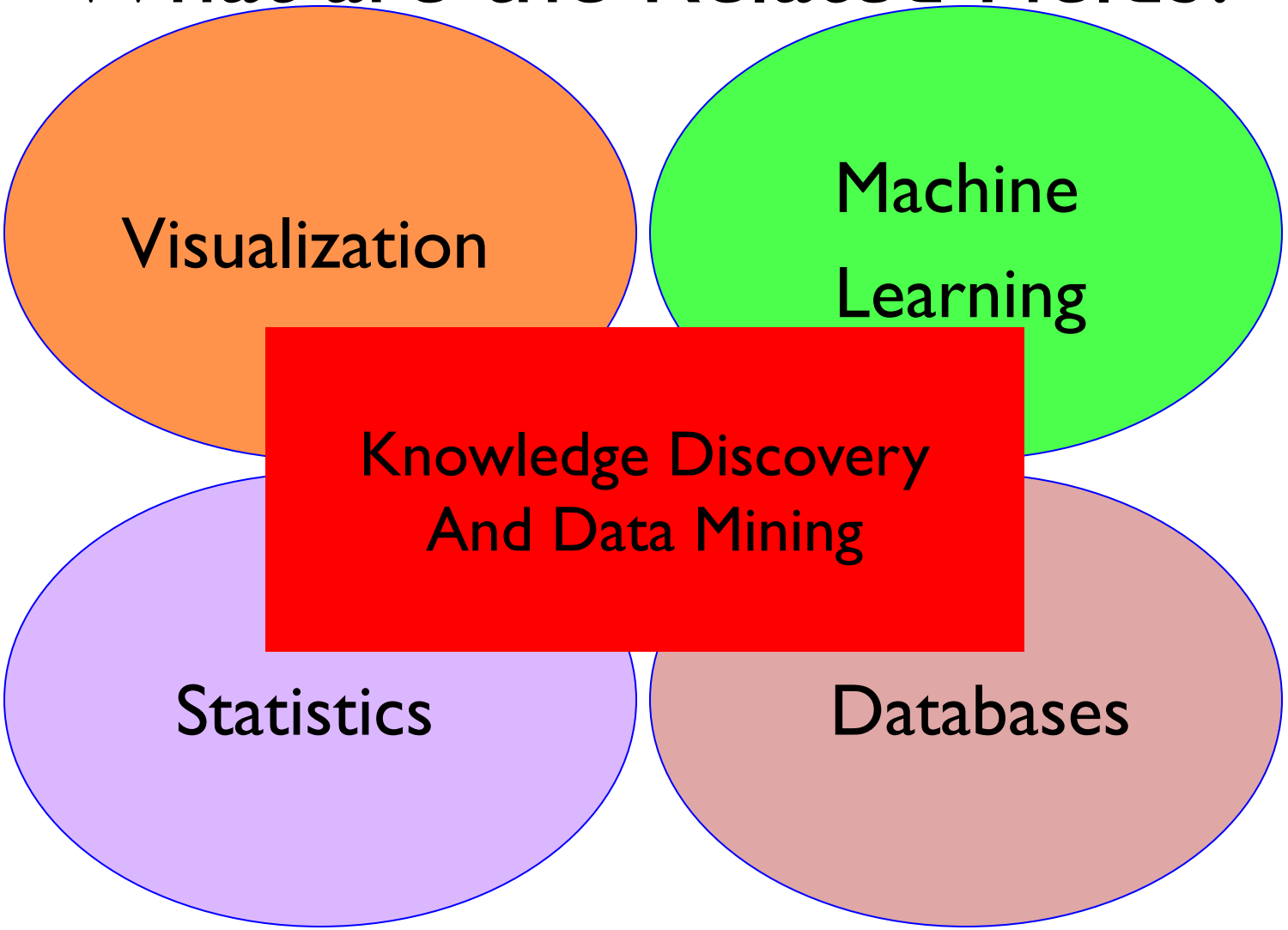
What is the General Idea?

- Build computer programs that navigate 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

What are the Related Fields?



Statistics, Machine Learning, and Data Mining

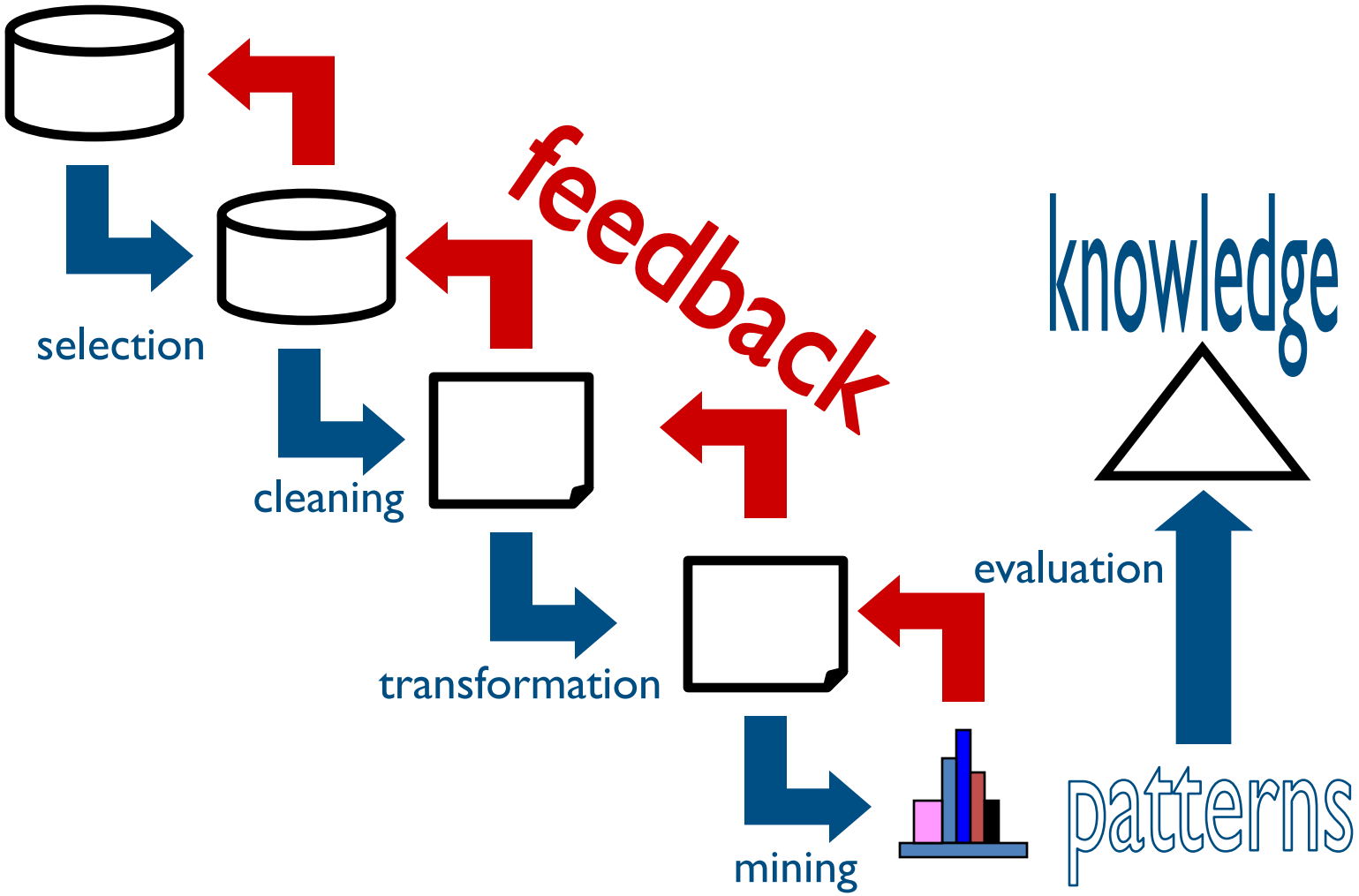
- Statistics is more theory-based, focuses on testing hypotheses
- Machine learning is more based on heuristics, focuses on building a program that learns, more general than Data Mining
- Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of discovery, including data cleaning, learning, integration and visualization

Distinctions are blurred!

Why Not Traditional Data Analysis?

- 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

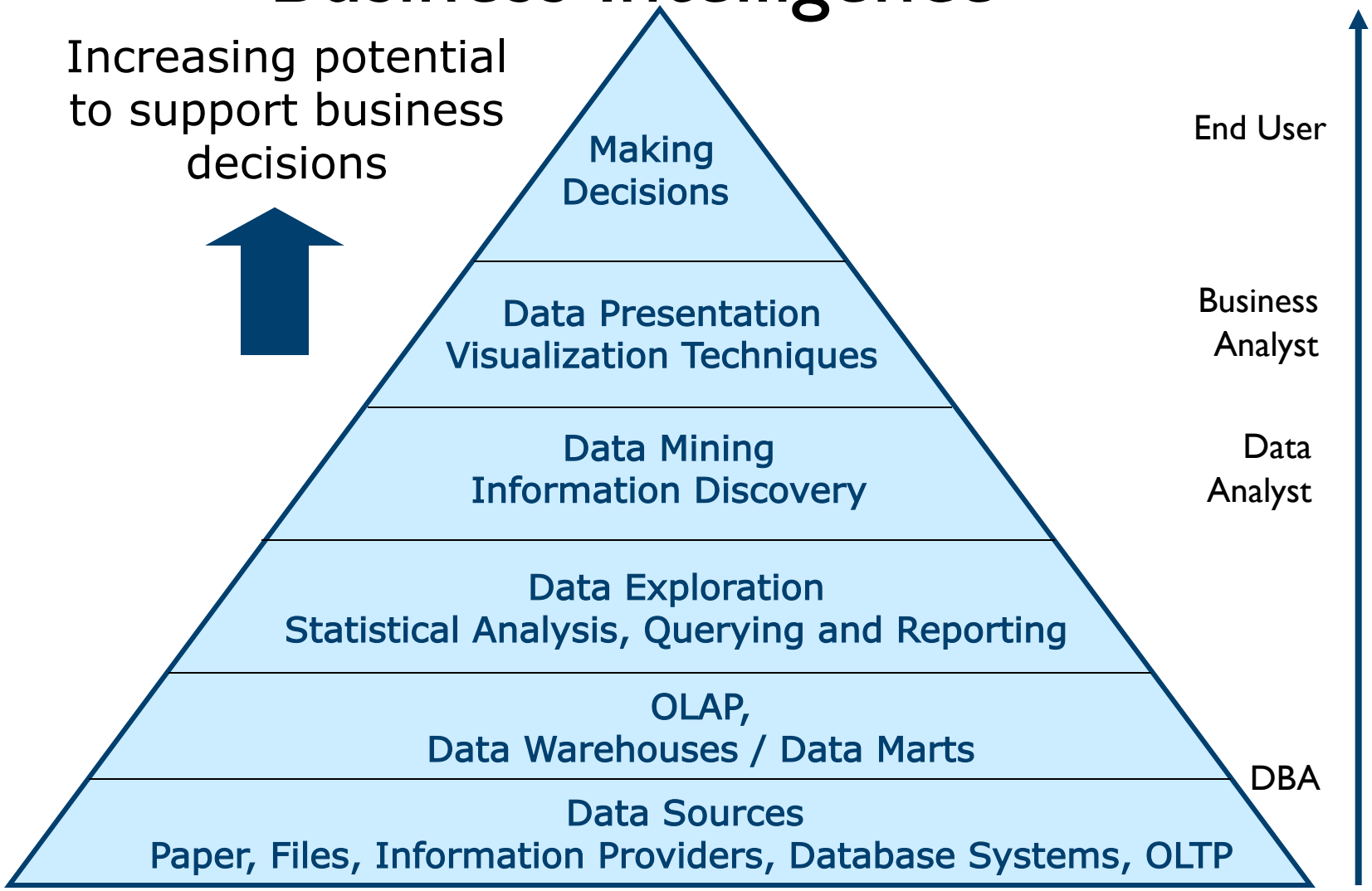
Knowledge Discovery Process



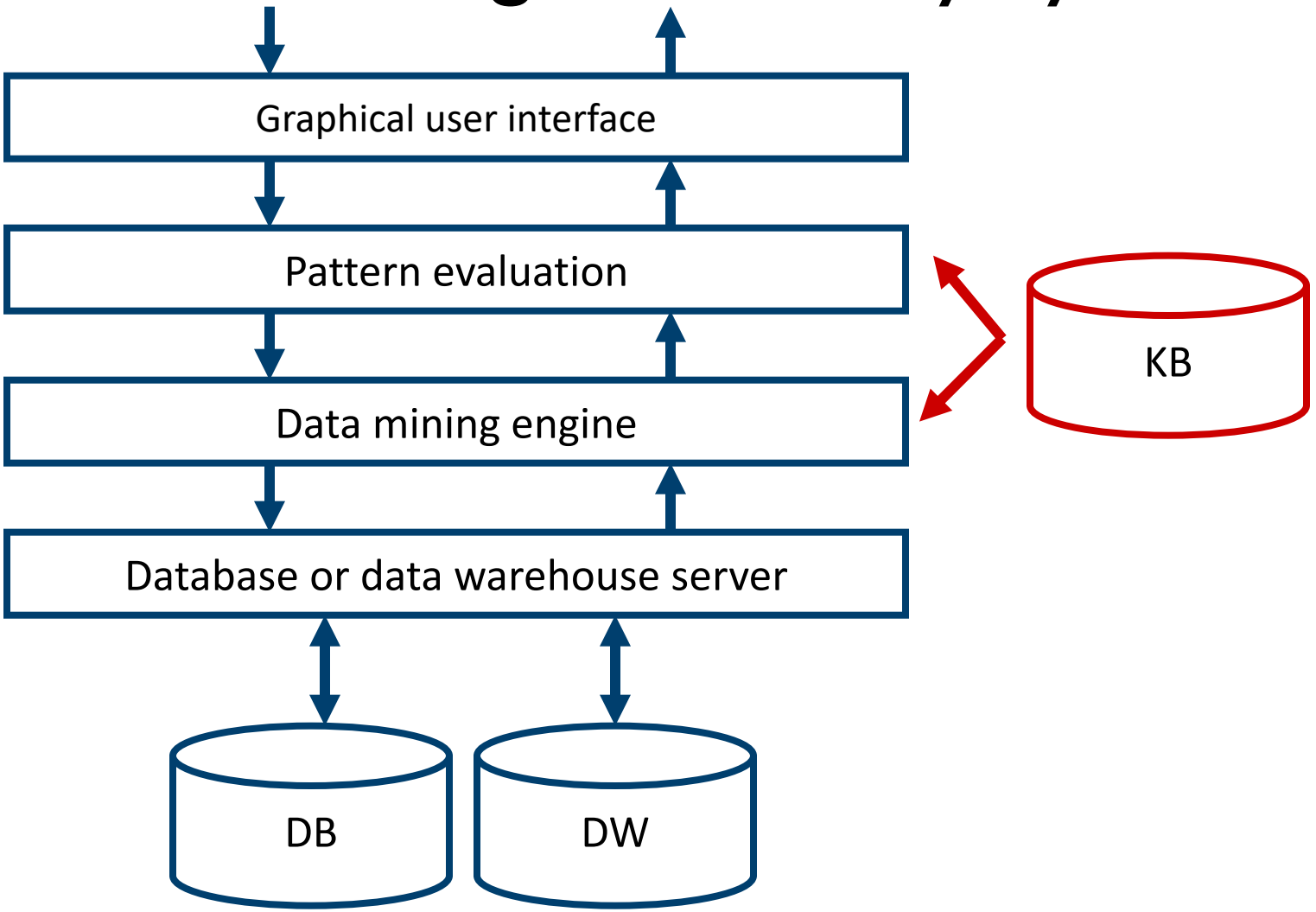
Knowledge Discovery Process What are the main steps?

- 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

Knowledge Discovery and Business Intelligence



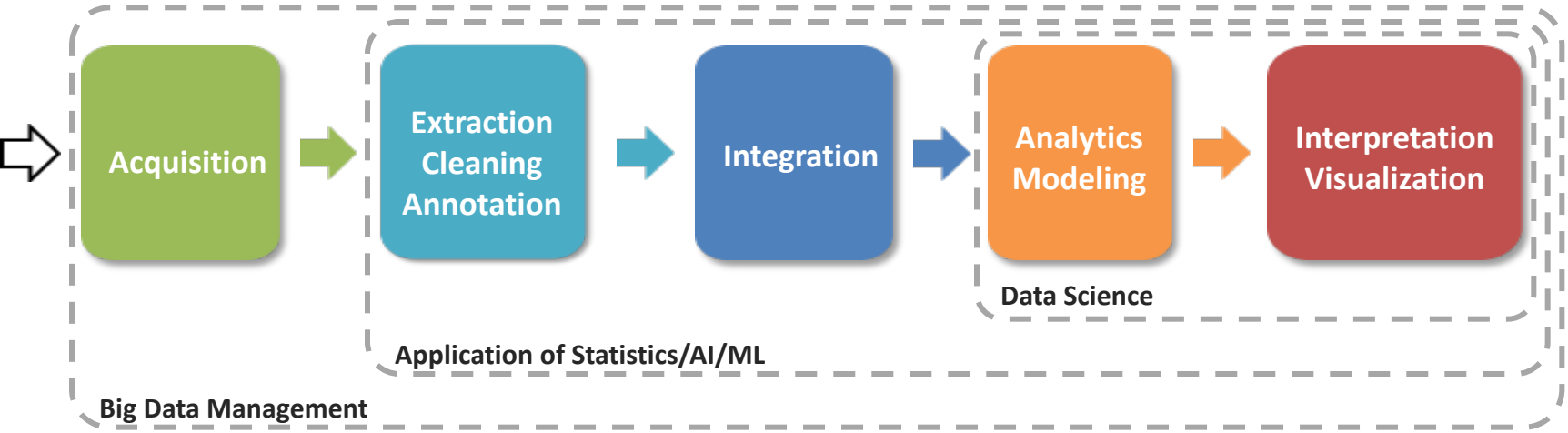
Architecture of a Typical Knowledge Discovery System



Actual efforts behind data analysis

- The actual implementation of the data mining or machine learning task is usually less than 5% lines of code in a real, non-trivial application
- The main effort (i.e. those 95% LOC) is spent on:
  - Data cleaning & annotation
  - Data extraction, transformation, loading
  - Data integration & pruning

  - Parameters tuning
  - Model training & deployment
  - …



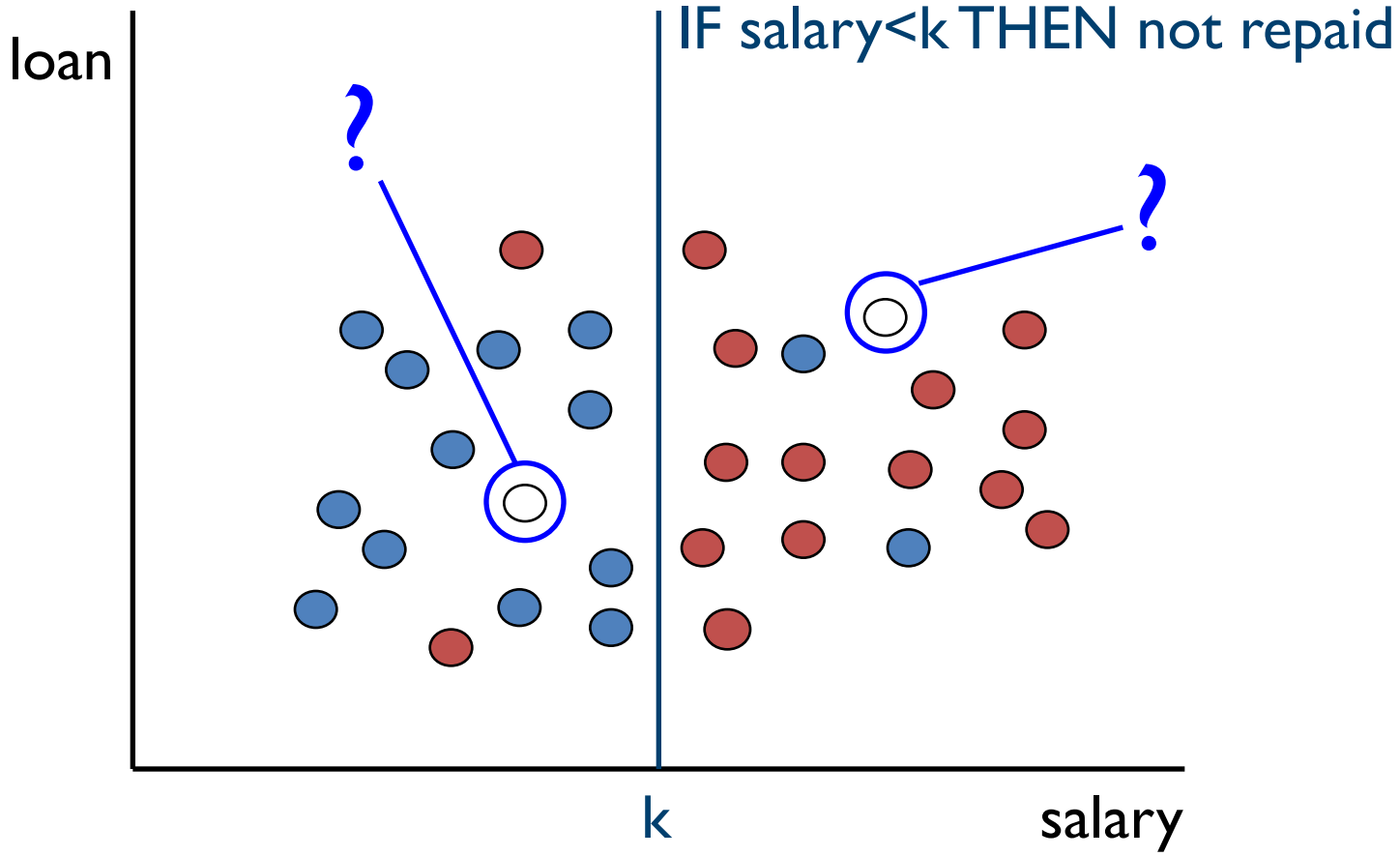
Credits: Beng Chin OOI – VLDB 2018 / A. Labrinidis, H. V. Jagadish VLDB 2012

What are the typical Data Mining tasks?

What are the Major Data Mining 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
- But many appear as time goes by, opinion mining, sentiment mining

Data Mining Tasks: classification

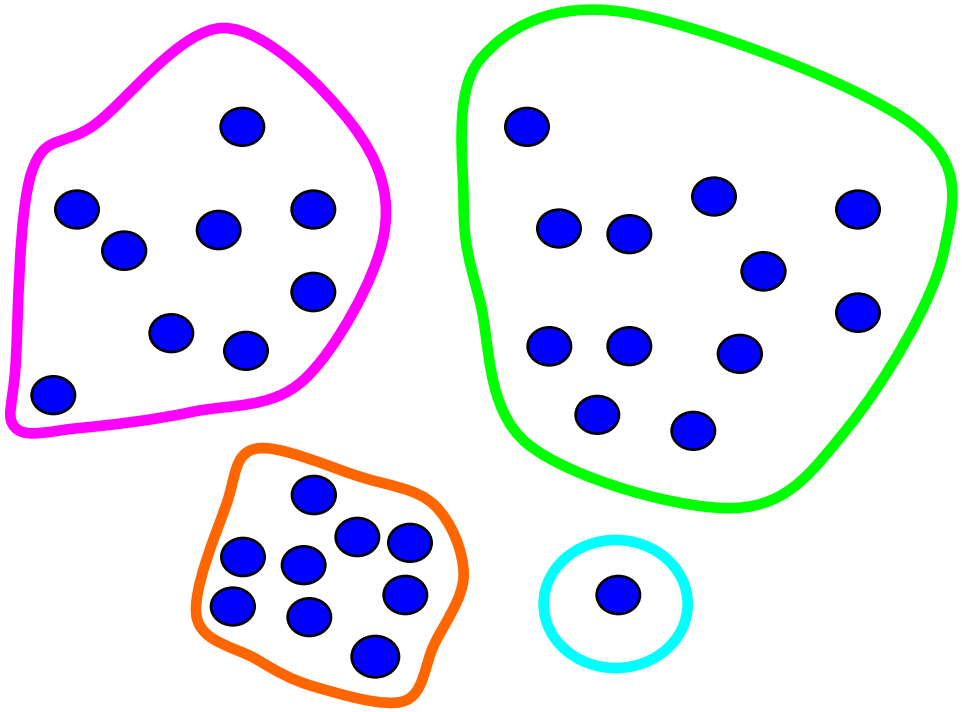


Data Mining Tasks: Classification & Prediction

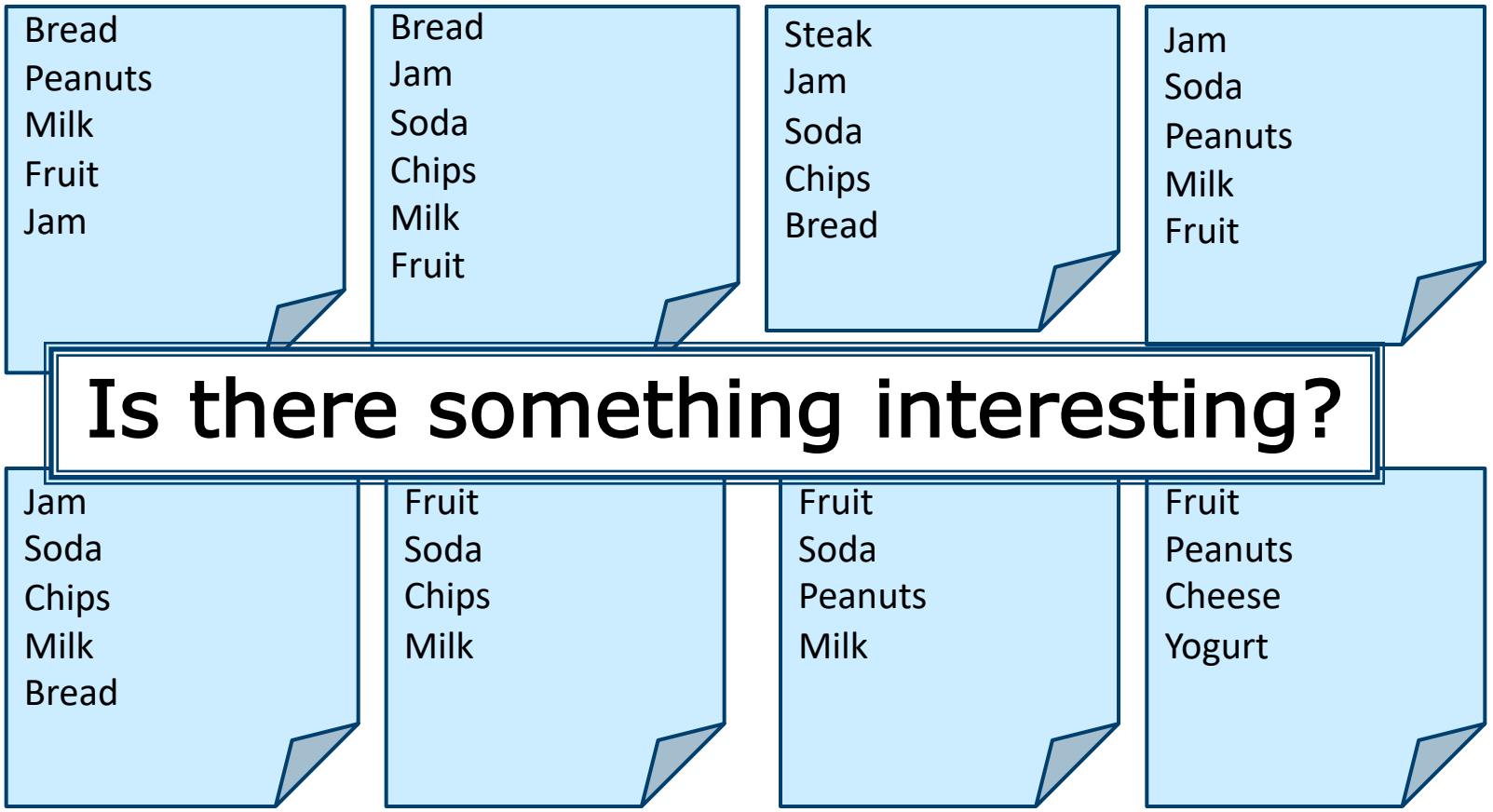
- Finding models (functions) that describe and distinguish classes or concepts
- The goal is to describe the data or to make predictions
- E.g., classify countries based on climate, or classify cars based on gas mileage
- Presentation: decision-tree, classification rule, neural network
- Prediction: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

Data Mining Tasks: Clustering

- 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



Data Mining Tasks: Associations



Data Mining Tasks: Associations

- Finds interesting associations and/or correlation relationships among large sets of data items.

- E.g., 98% of people who purchase tires and auto accessories also get automotive services done

Data Mining Tasks: Other Tasks

- 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

Relevant issues

Are all the "Discovered" Patterns Interesting?

- Data Mining may generate thousands of patterns, not all of them are interesting.
  - Suggested approach: Human-centered, query-based, focused mining

- 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?
  - Two approaches: (1) first generate all the patterns and then filter out the uninteresting ones; (2) generate only the interesting patterns—mining query optimization

What About 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
  - LowProfitMargin (X) <= Price(X, P1) and Cost (X, P2) and (P1 - P2) < $50

What the Major Issues?

- What 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

What the Major Issues?

- 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 impact
  - Domain-specific data mining & invisible data mining (embedded in other functional modules)
  - Protection of data security, integrity, and privacy

Summary

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