

# INFS4203/INFS7203

## Data Mining



---

Dr Helen Huang

DKE, ITEE, UQ

<http://www.itee.uq.edu.au/~dke>

huang@itee.uq.edu.au

# Instructors

---

- Course Coordinator: Dr Helen Huang

Phone: 3365 3239

Email: [huang@itee.uq.edu.au](mailto:huang@itee.uq.edu.au)

Room: 78-643

Consultation: by appointment

- Tutor: Mr Yang Yang

Phone: 3346 9575

Email: [yang.yang@itee.uq.edu.au](mailto:yang.yang@itee.uq.edu.au)

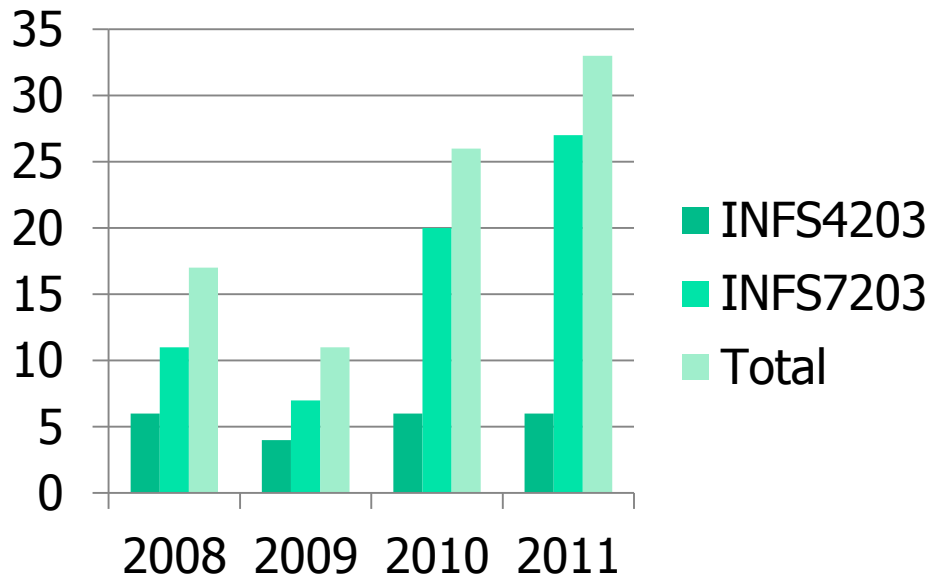
Room: 78-625

Consultation: by appointment

Tutorial: from week 2

# Students

- Course Enrolment History



- Interested in doing a PhD @ DKE?
  - Data and Knowledge Engineering

# Text Book and NewsGroup

---

## ■ Text Book:

- Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining. 1st Edition. 2006.

## ■ Reference Book:

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001

## ■ Newsgroup of INFS4203/INFS7203:

- On My-UQ Website
- Use it for the intra-class discussions for the course-related matters.
- <http://itee.uq.edu.au/~infs4203/>

# Assessment

| Assessment Task  | Due Date  | Weighting                   | Notes   |
|--|---|-----------------------------|---|
| <i>Exam - during Exam Period (School)</i><br>Final Examination | Examination Period<br>5 Nov – 19 Nov                      | 60%                         | Lecture notes;<br>closed book   |
| <i>Work-based Assessment</i><br>Individual Assignments         | 15 Aug - 23 Oct<br>Assignments on<br>Weeks 3, 5, 8, and 9 | 20%<br>(5% x 4 assignments) | No coversheet<br>required   |
| <i>Project</i>   | 6 Oct & 13 Oct<br>Demonstration Day                       | 20%                         | 1. Brief proposal<br>2. Demo system<br>3. Project report<br>4. Presentation<br>5. Group work<br><i>Up to 4 people</i> |


# Teaching Schedule

|          |   |
|----------|---|
| Week 1   | Introduction to Data Mining and Data Issues (Lecture)   |
| Week 2   | Association Rules Mining (Lecture)  |
| Week 3-4 | Classification (Lecture)  |
| Week 5-6 | Clustering (Lecture)  |
| Week 7   | Advanced Topic I -- Text Mining (Lecture)   |
| Week 8   | Advanced Topic II -- Web Mining (Guest Lecture)   |
| Week 9   | Project Consultation  |
|          | Term Break  |
| Week 10  | Project Presentation ( <b>Group Work</b> ) <i>15mins for each group</i>   |
| Week 11  | Project Presentation ( <b>Group Work</b> ) <i>15mins for each group</i>   |
| Week 12  | Revision of Previous Topics ( <b>Self Directed Learning</b> ):<br>Read the materials that are related to the final examination. |
| Week 13  | Course Revision (Lecture)   |



# **Lecture Note 1:**

# **Introduction to Data Mining**

- 
- INFS4203/7203 **Data Mining** is offered by DKE group
    - Data **Knowledge Engineering** group
    - Knowledge discovery from data
  - In this course,
    - Introduce basic data mining concepts and techniques for discovering interesting data **patterns** hidden in **large data sets**
    - Discuss algorithms for supporting a **scalable and efficient** data mining

# What Is Data?

- Lots of data is collected: web data, e-commerce, purchases at supermarkets, bank/credit card transactions, sensors on a satellite, gene data, etc.



facebook

50+ billions photos in July 2010

You Tube

35+ hours of content uploaded every minute

# How Much Data We Have?



- In a Second:
  - NASA's Space Shuttle operation will have 20,000 sensors telemetered once per second to Mission Control at Johnson Space Centre, Huston.
  - In United States there are about 50,000 security trading and up to 100,000 quotes and trades (ticks) are generated every second.
- In 24 hours:
  - AT&T records 275 million phone calls.
  - Google handles 100 million searches.
  - Wal-Mart records 20 million sales transactions.
- In a Week:
  - In Australia there are more than 80 Million SMS messages sent a week.
- In all time:
  - In scientific data collections, such as astronomical observatories, satellites imaging, and earth sensing, data can be routinely collected in gigabytes every day.

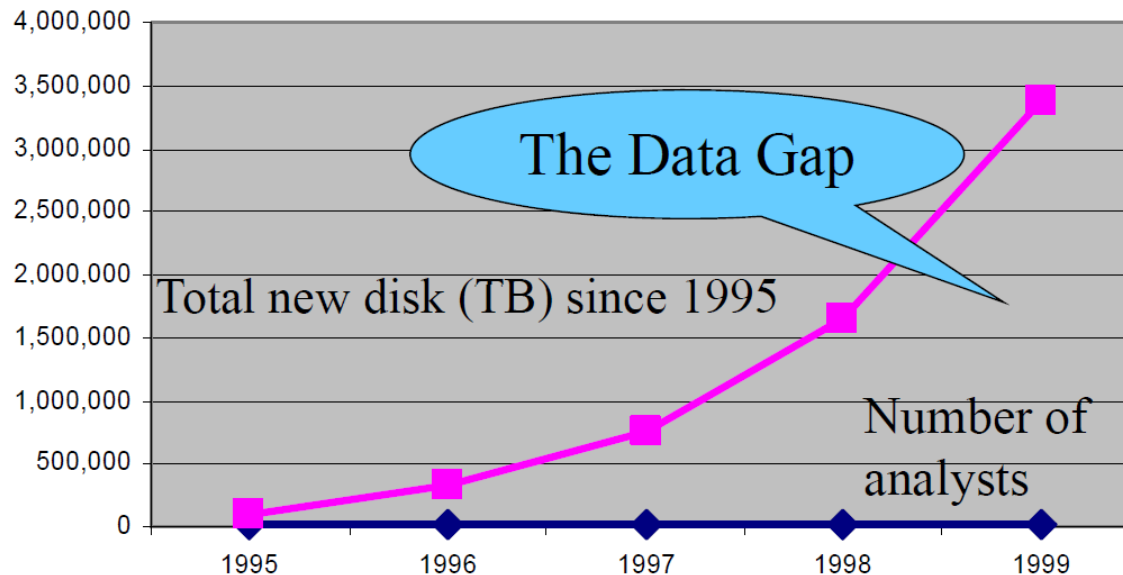
# Why Data Mining?

- We are drowning in data, but starving for knowledge!
  - Industry: Competitive pressure is strong: provide better, customized services by analysing customers' preference
  - Science: Traditional techniques are infeasible for raw data
- Solution: Data mining!
  - Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases



# Why Data Mining? (cont.)

Much of data is never analysed at all!



[Ref: Jure Leskovec, Stanford CS345a: Data Mining]

# Why Data Mining? (cont.)

## ■ Evolution of Database Technology

- 1960s: Data collection, database creation, IMS and network DBMS
  - *"How many students have been enrolled in Data Mining this year?"*
- 1970s: Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
  - *"How many students are enrolled in both Spatial Database and Data Mining?"*
- 1990s: Data mining, multimedia databases, and Web databases
- 2000s:
  - Stream data management and mining
  - Data mining with a variety of applications
  - Web technology and global information systems
  - *"Sam selected Spatial Database last year. How likely will he select Data Mining this year? Why?"*

# What Is Data Mining?



- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- 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
- Origins of Data Mining
  - Machine Learning: more heuristic, more general than Data Mining
  - Statistics: more theory-based, focused on testing hypotheses
  - Database system: information retrieval, deductive query answering
  - **Data Mining: focuses on the algorithms to extract patterns from data**

# What Is Data Mining? (cont.)

---

- Data mining applications are often structured around the specific needs of an industry sector or even tailored and built for a single organization.
- The patterns within data may be very specific.
  - Banking data mining applications may need to track client spending habits in order to detect unusual transactions that might be fraudulent.
  - In another example, a data mining application might be used by a government body to detect associations between individuals who may be involved in terrorist activities

# An Example: Market Basket Analysis



Anything interesting?

Bread → Milk (100%)  
Diapers → Beer (66%)  
Diapers → Milk (100%)

Customers who buy diapers also tend to buy beer



Identify **potential cross-selling opportunities** among related items

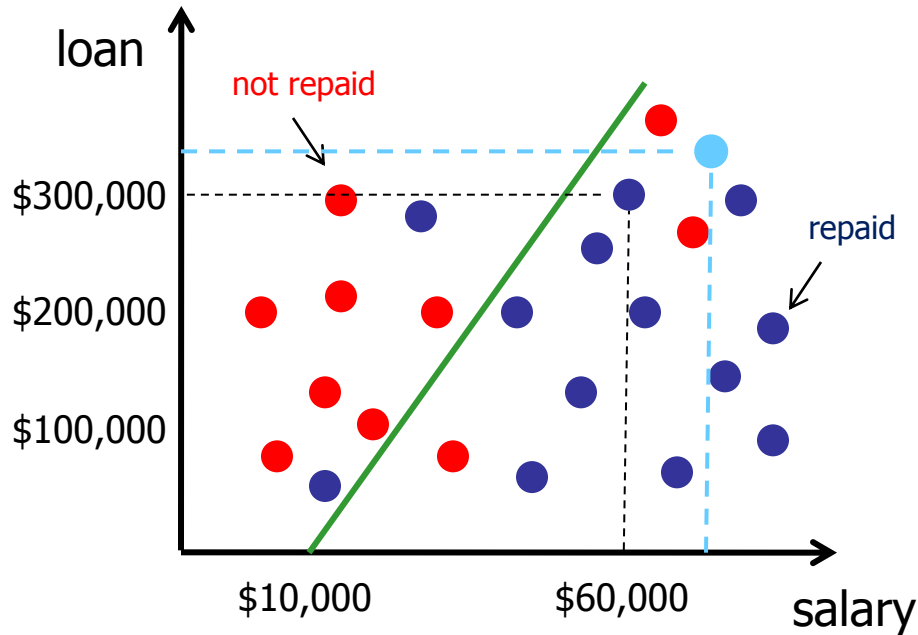
# "beer and diapers"

---

Some time ago, Wal-Mart decided to combine the data from its loyalty card system with that from its point of sale systems. The former provided Wal-Mart with demographic data about its customers, the latter told it where, when and what those customers bought. Once combined, the data was **mined extensively** and **many correlations appeared**. Some of these were obvious; people who buy gin are also likely to buy tonic. They often also buy lemons. However, one correlation stood out like a sore thumb because it was so **unexpected**.

On Friday afternoons, young American males who buy diapers (nappies) also have a predisposition to buy beer. No one had predicted that result, so no one would ever have even asked the question in the first place. Hence, this is an excellent example of data mining.

# An Example: Credit Risk

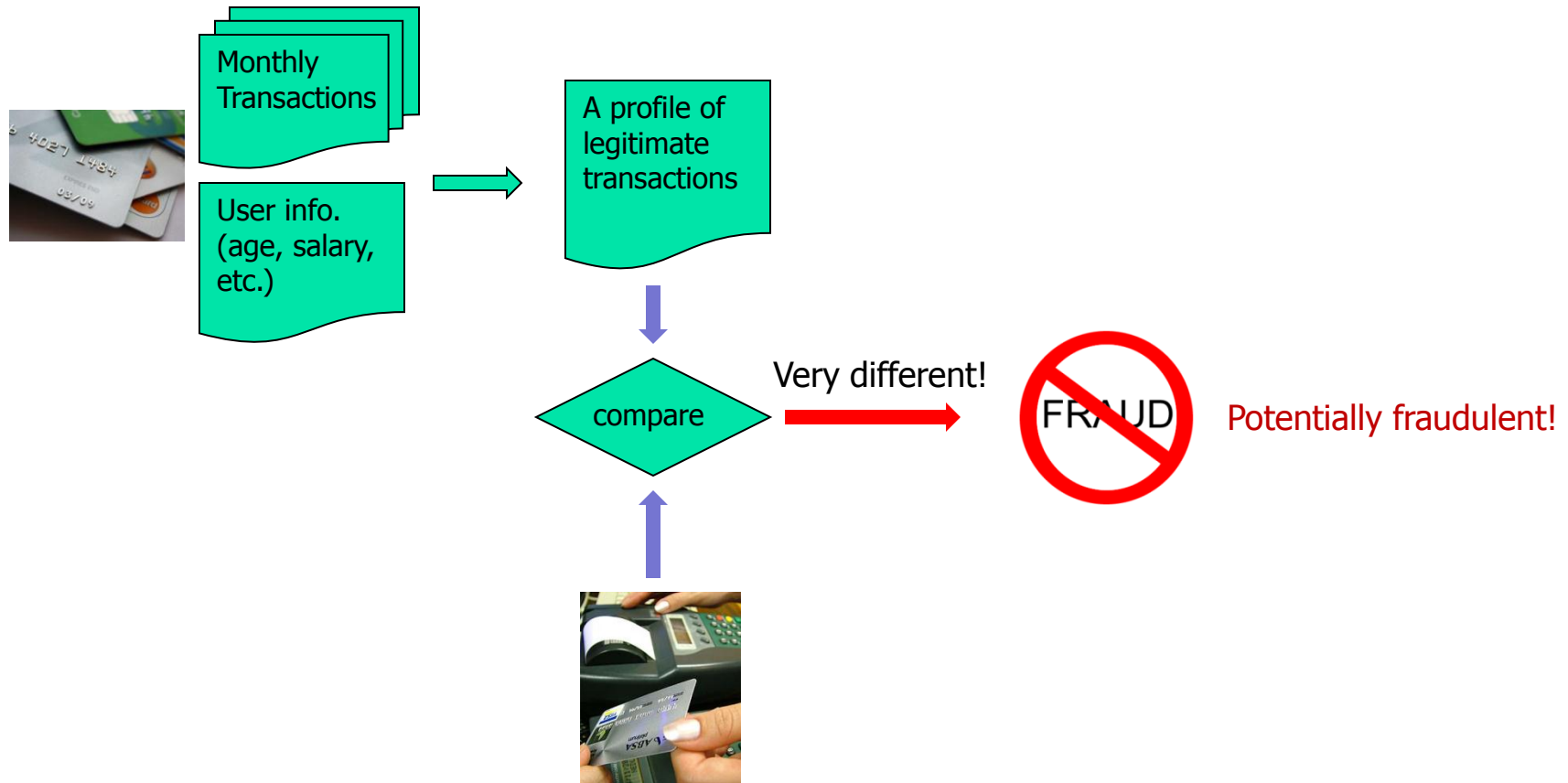


Given: a loan application

Problem: predict whether the bank should approve the loan

Data: records from other loans

# An Example: Credit Card Fraud Detection



# Potential Applications

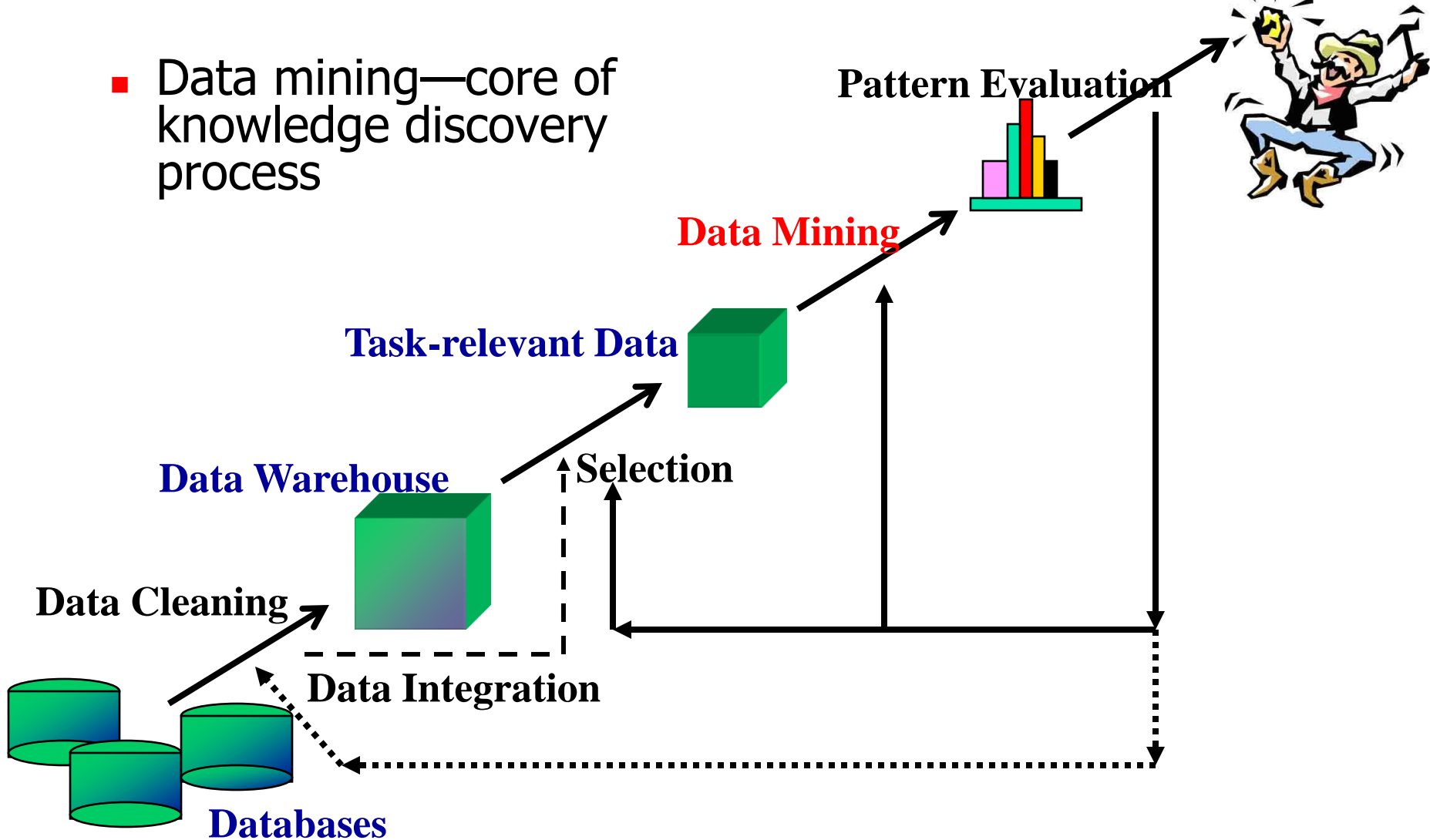


- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - DNA and bio-data analysis

# Data Mining: A KDD Process

**Knowledge**

- Data mining—core of knowledge discovery process



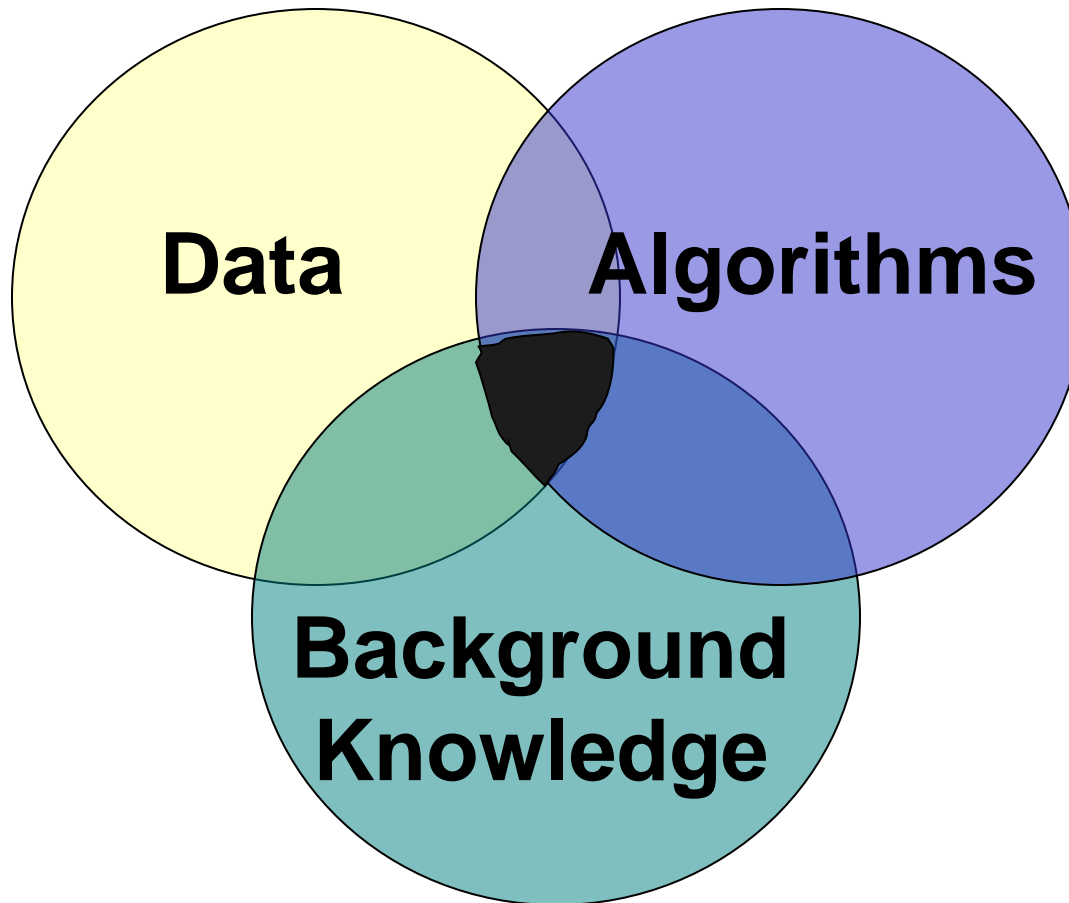
# Steps of a KDD Process

---

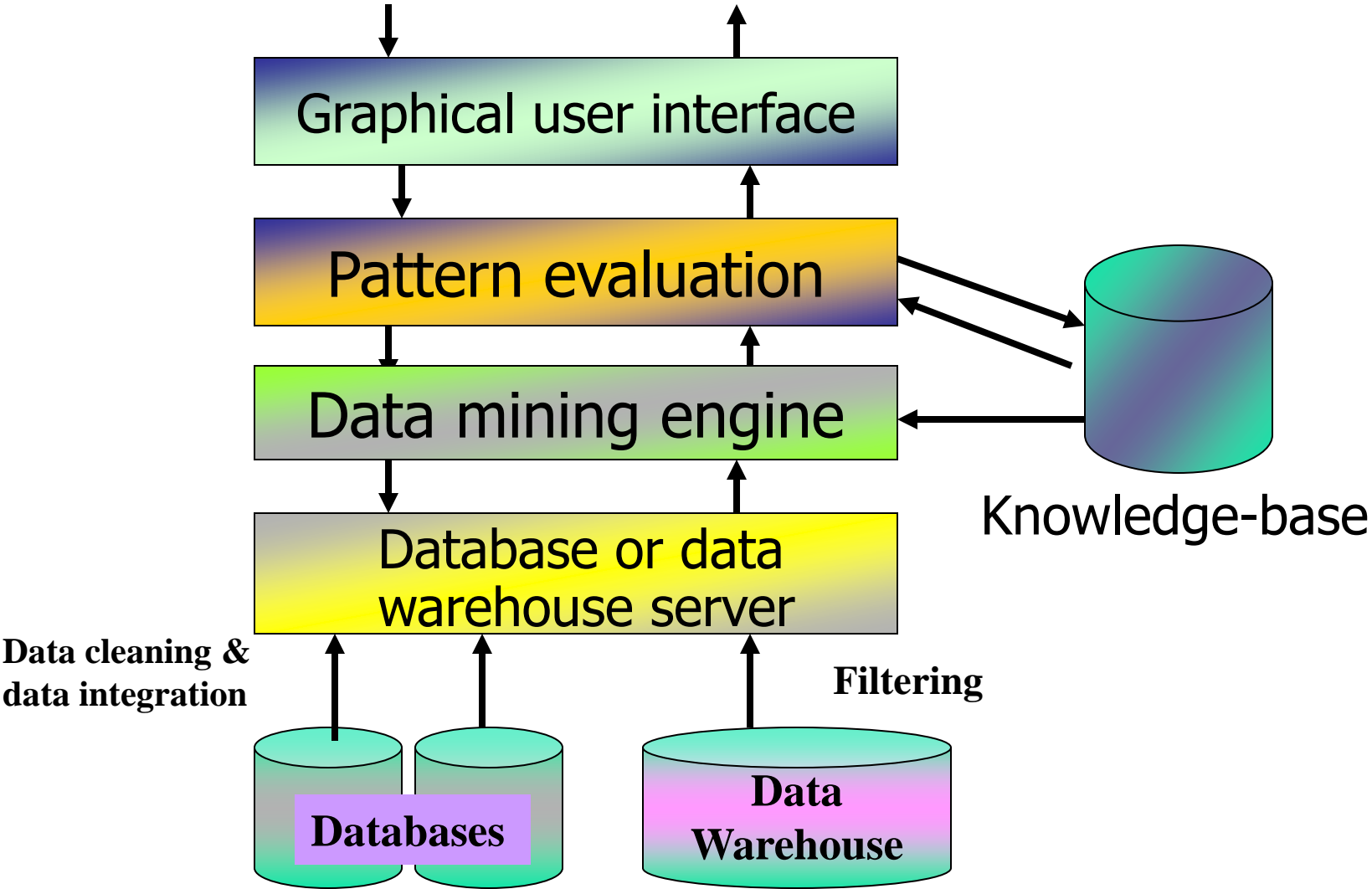
- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant 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

# Data Mining Perspectives

---



# Architecture: Typical Data Mining System



# Data Mining: On What Kinds of Data?

---

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

# First of All: What is Data?

---

- A data item has two levels meaning: the **domain** and its **value**.
  - A data domain gives data structure and prescribes its possible (legal) values.
  - A data domain is associated with its domain-specific operations. E.g.,
    - Integer: arithmetic operations
    - text string: concatenation, sub-string, etc.
  - A data value is a measurement of a real-world object or a concept.
- A data item can be either simple or complex.
  - A data item is associated to an ontology hierarchy.
  - A data item is associated to a multidimensional structure.

# First of All: What is Data? (cont.)

---

- **Associated Patterns:** dependency, associations, correlations, dimensionality, etc.
- **Associated Dynamics (changes):** monotonous changes, state transitions, etc.

# An Example: Relational database, transactional database, and multimedia database

| ID  | Name | Age | Salary(\$) |
|-----|------|-----|------------|
| 1   | Sam  | 21  | 48k        |
| 2   | Alex | 38  | 72k        |
| 3   | Mary | 27  | 52k        |
| ... |      |     |            |

Relational Database

Transactional Database

| Trans. ID | List of items             |
|-----------|---------------------------|
| T100      | Bread, Coke, Milk         |
| T101      | Beer, Bread               |
| T102      | Beer, Coke, Diaper, Milk  |
| T103      | Beer, Bread, Diaper, Milk |
| T104      | Coke, Diaper, Milk        |
| ...       |                           |

Multimedia Database

| Image ID | Hyperlink  | Colour feature         |
|----------|--|------------------------|
| I100     | <a href="http://itee.uq.edu.au/~image/1.jpg">itee.uq.edu.au/~image/1.jpg</a> | <0.1,0.4,0.15,0.8,0.2> |
| I101     | <a href="http://itee.uq.edu.au/~image/2.jpg">itee.uq.edu.au/~image/2.jpg</a> | <0.4,0.3,0.25,0.7,0.7> |
| I102     | <a href="http://itee.uq.edu.au/~image/3.jpg">itee.uq.edu.au/~image/3.jpg</a> | <0.5,0.12,0.1,0.2,0.3> |
| ...      |  |                        |

# Data Mining Functionalities

- Concept description: Characterization and discrimination: describing a group
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
  - Find human-interpretable patterns that describe the data
- Classification and Prediction: predicting an item class
  - 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
  - Presentation: decision-tree, classification rule, neural network
  - Predict some unknown, missing or future values
- Association: finding frequent occurring events
  - Diaper → Beer [0.5%, 75%]

# Data Mining Functionalities (cont.)

---

- Cluster analysis: finding clusters in data
  - 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: finding changes
  - Outlier: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! useful in fraud detection, rare events analysis
- Trend and evolution analysis: predicting a continuous value
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analysis

# 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, actionability, etc.

# Can We Find All and Only Interesting Patterns?



- Find all the interesting patterns: **Completeness**
  - Can a data mining system find **all** the interesting patterns?
  - Heuristic vs. exhaustive search
- Search for only interesting patterns: An optimization problem
  - 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

# 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
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures

# A Brief History of Data Mining Society

---

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.

Next Week:



# Mining Association Rules