CIS410/510 Data Mining
Assignment 1 Sample Solutions:

1. (warm-up) Based on your understanding, please define what is data mining. Please also give at least three examples for applications.

   **Answer:** Data mining refers the process or method that extracts or “mines” interesting knowledge or patterns from large amounts of data.

   It is applied in many fields. For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures. WalMart is pioneering massive data mining to transform its supplier relationships. WalMart captures point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits this data to its massive 7.5 terabyte Teradata data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, WalMart computers processed over 1 million complex data queries.

2. A data warehouse has three dimensions (time, doctor and patient) and two measures (count and charge). Here charge is the fee that a doctor charges for a visit from a patient.

   a) List three classes of schemas that are popularly used for modeling data warehouses.

   **Answer:** Three classes of schemas popularly used for modeling data warehouses are the star schema, the snowflake schema, and the fact constellations schema.

   b) Draw a star schema diagram for the given data warehouse.
c) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2000.

**Answer:** The operations to be performed are:
- Roll-up on *time* from *day* to *year*.
- Slice for *time* = 2000
- Roll-up on *patient* from individual patient to all.

d) To obtain the same list, write an SQL query, if the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge). (Note: If you haven't taken CIS451/551 or don't know SQL, you can give an sample table with several data instances, which are enough to give an example answer. And calculate and report the list you get)

**Answer:**
```
select doctor, SUM(charge)
from fee
where year = 2000
group by doctor
```

3. A data warehouse for UO has four dimensions (student, course, term, and instructor), and two measures (count, and GPA). When at the lowest conceptual level (for example, for a given student, course, term and instructor combination), the GPA measure stores the actual course grade
of the student. At the higher conceptual level, GPA stores the average grade for the given combination.

a) Draw a snowflake schema diagram for the data warehouse.

b) Starting with the base cuboid [student, course, term, instructor], what specific OLAP operations (e.g., roll-up from term to year) should one perform in order to list the GPA of CIS courses for each UO student.

Answer: The specific OLAP operations to be performed are:

a. Roll-up on course from course_id to department
b. Roll-up on student from student_id to all_student
c. Dice on course, student with department = "CIS"
d. Drill-down on student from all_student to student_name

4. Discuss the motivation behind OLAP mining (OLAM).
The motivations behind OLAP mining are the following: The high quality of data (i.e., integrated, consistent, and cleaned data) in data warehouses serves as a valuable source for OLAP as well as for data mining.

The available information processing infrastructure surrounding data warehouses means that comprehensive information processing and data analysis infrastructures will not need to be constructed from scratch.

OLAP-based exploratory data analysis can be realized by coupling on-line analytical mining with data/knowledge visualization tools to allow users to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms.

On-line selection of data mining functions allows users who may not know what kinds of knowledge they would like to mine the flexibility to select desired data mining functions and dynamically swap data mining tasks.