

## Computer Science COMP3420 (2009) – Tutorial One

**Question 1.** Among the three popular databases, the hierarchical database, the network database and the relational database, why the relational database becomes the dominant database in the commercial world?

**Question 2.** How is a data warehouse different from a database? How are they similar?

**Question 3.** What is the difference between discrimination and classification? between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?

**Question 4.** What are the major challenges of data mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)?

**Question 5.** What is a data warehouse? What is data mining?

**Question 6.** State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the *update-driven approach* which constructs and uses data warehouses, rather than the *query-driven approach* which applies wrappers and integrator. Describe a situation where the query-driven approach is preferable over the update-driven approach.

**Question 7.** Can you briefly describe the four stages of knowledge discovery (KDD)? Can you describe the multi-tiered data warehouse architecture?

**Question 8.** Briefly describe the five major tasks in data preprocessing. What are the main objectives of each of these tasks?

**Question 9.** Can you list the three measures used for measuring data central tendency? Give an example for each of the measures.

**Question 10.** Given a set of data  $\{12, 8, 45, 23, 6, 7, 8, 25, 34, 22, 19, 22, 34, 45, 41\}$ , (a) can you employ the binning method to allocate the data in the set into 3 bins by equal-frequency, smoothing by bin means and smoothing by bin boundaries respectively?

(b) Can you use the histogram method to allocate the data in the set to 4 equal intervals?

**Question 11.** What are the main metrics for measuring data dispersion? What are the main tasks of data cleaning? Briefly describe several approaches to remove data noise.

**Question 12.** Can you point out how to perform data transformation during building a

data warehouse?

**Question 13.** Outlines the major research challenges of data mining in one specific applications domain such as stream/sensor data analysis, or bioinformatics.