**Question 1.1: Tan Chapter 1, Questions 1 and 3**

Question 1: Discuss whether or not each of the following activities is a data mining task.

1. Dividing the customers of a company according to their gender.
   1. No because this is a simple database query task not a data mining task
2. Dividing the customers of a company according to their profitability.
   1. No, this is an accounting calculation, followed by the application of the threshold. However, predicting the profitability of the new customer would be data mining.
3. Computing the total sales of a company
   1. No this is just a accounting calculation not a data mining task
4. Sorting a student database based on student identification numbers
   1. Similarly to part 1, this is a database query not a data mining task
5. Predicting the outcomes of tossing a (fair) pair of dice.
   1. No, since the dice are fair dice then it is more of a probability calculation.
6. Predicting the future stock price of a company using historical data
   1. Yes, we would attempt to create a model that can predict the continuous value of the stock price. This is an example of the area of data mining known as predictive modeling.
7. Monitoring the heart rate of a patient for abnormalities
   1. Yes. We would build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection. This could also be considered as a classification problem if we had examples of both normal and abnormal heart behavior.
8. Monitoring seismic waves for earthquake activities
   1. Yes, in this case, we would build a model of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. This is an example of the area of data mining known as classification.
9. Extracting the frequencies of sound waves
   1. No this is just a signal processing task

Question 3: For each of the following data sets, explain whether or not data privacy is an important issue.

1. Census data collected from 1900-1950
   1. No since it is data from over 60 years ago it is more outdated then collecting data from today.
2. IP Addresses and visit times of web users who visit your website
   1. Yes because IP Addresses can be used for malicious purposes against people.
3. Images from Earth-orbiting satellites
   1. No because it isn’t exposing anything person, due to the fact that the images aren’t precise enough to be exposing
4. Names and addresses of people from the telephone book
   1. No because this data is already public so the risk was already taken when the user decided to expose this information
5. Names and email addresses collected from the Web.
   1. No because this data is more public then for instance IP addresses and website actions.

**Question 1.2: Tan Chapter 2, Questions 2,3,7 and 12**

Question 2: Classify the following attributes as binary, discrete or continuous. Also classify them as qualitative(nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there might be some ambiguity.

1. Times in terms of AM or PM.
   1. Binary, Qualitative, nominal
2. Brightness measured by a light sensor
   1. Continuous, quantitative, tatio
3. Brightness measured by people's judgements
   1. Discrete, qualitative, ordinal
4. Angles as measured by degrees of 0 to 360
   1. Continuous, quantitative, ratio
5. Bronze, Silver and Gold medals awarded at the Olympics
   1. Discrete, qualitative, ordinal
6. Height above sea level
   1. Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)
7. Number of patients in a hospital
   1. Discrete, quantitative, ratio
8. ISBN numbers of books
   1. Discrete, qualitative, nominal (but ISBN numbers do have some order information so it could be ordinal if you use that information)
9. Ability to pass light in terms of the following values: opaque, translucent, transparent
   1. Discrete, qualitative, ordinal
10. Military Rank
    1. Discrete, qualitative, ordinal
11. Distance from center of campus
    1. Continuous, quantitative, interval/ratio(depends)
12. Density of a substance in grams per cubic centimeter
    1. Continuous, quantitative, ratio
13. Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)
    1. Discrete, qualitative, nominal (or ordinal if you are using the order information)

Question 3: What is aggregation? What are the motivations for aggregation? How are the values of attributes handled when aggregating data?

1. Aggregation is the process of finding, collecting, and presenting the data in a summarized format to perform statistical analysis of business schemes or analysis of human patterns. Data Aggregation is a need when a dataset as a whole is useless information and cannot be used for analysis. So, the datasets are summarized into useful aggregates to acquire desirable results and also to enhance the user experience or the application itself. The values of the attributes which are affected by the aggregation usually consist of operations performed on one or more attributes for instance the sum of attributes which then are used to be filtered by a desired attribute to aggregate the data.

Question 7: What is a document term matrix?

1. A document term matrix is a matrix which is used to represent a collection of documents. Which is a type of sparse data matrix which are special cases in which data attributes are of the same type and are asymmetric: i.e. only non-zero values are important. Transaction data is an example of a document term matrix.

Question 12: Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

1. Since the data in a document-term matrix is compiled of several data sets from documents making where there are data sets with discrete and continuous data, we can focus on asymmetric data for key data sets in which we want to know about from the compiled information from the document data vectors which are in the matrix.