Practical 1

**Introductory Review Questions**

# What are we doing?

Using what you learned from lectures (Lecture 1&2) and relevant reading materials, you will answer some review questions. These questions are for your self-review on topics covered. You will need to review lecture and reading materials or seek for other resources (e.g. Googling), in order to answer questions.

**Submission:**

You are required to submit one document containing your answers via the weekly-practical submission box (available on CP1407 LearnJCU)

# Self-Review Questions

1. Briefly define the term “Data Science”, “Data Mining” and “Machine Learning”, and identify the difference between these terms.
2. Broadly, there are two types of knowledge, shallow and deep. Shallow knowledge is simply what makes up a computer’s response. If we can retrieve any answer by framing a data query (using SQL) from an existing database system (e.g. JCU student database system), the output result retrieved will constitute shallow knowledge about the data. For example, we may learn that Australian Stock Exchange generally follows the lead of Wall Street, but we wouldn’t necessarily know why. Deep knowledge is the underlying reason behind such relationships. Hidden knowledge is the top layer of this deep knowledge, which normally a data mining technique can unveil. Data mining will not give us the causes or the significance, but it can point to various associations and links.

Data query is about searching in the data when we know exactly what we are looking for. Example are:

* A list of customers who used MasterCard to buy medicine from a pharmacy.
* A list of employees who will reach retiring age next year.

These are all in the domain of shallow knowledge, which can easily be obtained by simple data queries using, for instance, SQL. By contrast, let us consider the following examples:

* Develop a profile of MasterCard holders who will take advantage of the forthcoming sale promotion at the pharmacy
* Develop a list of employees who are likely to avail themselves of the voluntary early retirement scheme when they reach retirement age.

These are examples of hidden knowledge whose answers cannot be obtained from data queries, although data mining techniques can unveil the information.

Identify whether each of the following is a data query or a data mining task(s):

1. A social worker is interested in learning about the proportion of males to females in the population of a particular region.
2. A stock market analyst has been asked by his client to predict the future prices of 10 stocks three months in advance.
3. Do single men play more golf than married men?
4. Determine the characteristics of a successful used car salesperson.
5. Determine whether a credit card transaction is valid or fraudulent.
6. Why is a fully automated data mining tool not desirable? Discuss the need for human intervention in the data mining process.
7. How can data mining help a business analyst?
8. Data mining is a powerful technology that can bring about positive benefits but it has also caused a certain degree of suspicion and concerns over ethical issues. Find suitable examples to highlight that such concerns are valid and reasonable.
9. The main objectives of data mining can be broadly categorized into *classification, estimation, prediction* and *data description*.

* Classification: Object are classified into one of a set of pre-defined classes. In order to do this, a classification model is built from a set of data examples. The accuracy of the classification of the model is then evaluated to give some degree of confidence to the result. Once a reliable classification model has been developed, it is then used to classify data records whose class outcomes are unknown.
* Estimation: Instead of classifying an object into a discrete class, this task involves building a model (based on a set of data examples) to estimate the value of a continuous outcome variable.
* Data Description: This task is about describing general or specific features of the selected data set. It includes summary statistics, clustering and characteristic rule mining.
* Prediction: It overlaps significantly with the classification and estimation, but is more concerned with a future outcome of the output variable. For instance, historical data recordings on weather conditions are used to predict tomorrow’s weather. Solutions for classification and estimation are widely used for prediction too.

Categorize each of the following data mining activities as classification, estimation or description. State clearly the reason behind your decision. Can any patterns discovered be used for prediction purposes?

1. A real-estate agency has accumulated a large number of property sale records. The properties can be studio flats, semi-detached houses, detached houses or mansion houses. The agency wants to investigate from the data set what kinds of customer are likely to purchase which types of property.
2. It is interesting for the same real-estate agency to make significant links between descriptors of the properties sold and the characteristics of their customers. For instance, customers who are married with young children may be more likely to purchase a three-bedroom, detached house with a single garage.
3. In recent years, we have seen increasing amounts of toxic waste dumped into our environment. Waste water from manufacturing processes, farming land run-off and sewage water from treatment plants have broken the chemical balance of the water in our rivers. The organic matter in the water has resulted in excessive growth of algae, which in turn leads to a reduction of the oxygen level in the water. Causing the deaths of fish and other wild life. Therefore, environment agencies want to monitor closely the growth of algae in the rivers and lakes. One agency has collected water samples from a number of different sites and analysed them for various chemical substances. They have also collected algae samples at the same locations to determine the population distributions of different algae. The agency wants to use the sample data to build a model that can approximate the distribution of algae population based on amounts of the chemical substances.

# Laboratory Questions

1. Visit [KDnuggets](http://www.kdnuggets.com/) website and try to explore the site freely to get useful news and information in the field of Business Analytics, Data Mining, and Data Science. Try to find and list some practical applications of data mining tools. (Hint: you can refer various [polls](http://www.kdnuggets.com/polls/index.html) results provided by this site)
2. [*UCI Repository*](http://archive.ics.uci.edu/ml/index.php) is one of popular web sites where provide a repository of databases, domain theories and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

Visit the site and retrieve one data set. With this example dataset, imagine you are a data miner who uses this data set to process a data mining project to solve a real-world problem. Define the following:

* What problem do you target to solve using this data set?
* What part of the data (which attributes of the data) would be used as input for your data mining model?
* What data mining methods (e.g. classification, clustering, association rule mining etc.) can be applied?
* What would be the output of this data mining process?