**Lecture 1: In class examples**

1. Data mining definition: “Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.” – Wikipedia. This definition differs from the one in the lecture because it doesn’t mention that the discovery process is automatic.
2. Watched videos of many different categories on YouTube. Google could mine this data to check if I watched any videos pertaining to products. Google’s other web pages or Chrome could give me advertisements on these and similar products.
3. Classification: {binary, discrete, or continuous}, {qualitative (nominal or ordinal) or quantitative (interval or ratio)}
   1. Number of telephones in your house – Discrete, Ordinal
   2. Size of French Fries (Medium or Large or X-Large) – Discrete (since there are 3 possible values), Ordinal (Since there is a specific order)
   3. Ownership of a cell phone – Binary (Either have a phone or don’t have one)
   4. Number of local phone calls you made in a month – Discrete, Interval (the number of calls is always between 0 and some positive number)
   5. Length of longest phone call - Continuous
   6. Length of your foot - Continuous
   7. Price of your textbook - Continuous
   8. Zip code - Discrete
   9. Temperature in degrees Fahrenheit - Continuous
   10. Temperature in degrees Celsius - Continuous
   11. Temperature in Kelvins – Continuous

**Assignment 1:**

I study in a postgraduate program that requires a lot of lookups on the Internet. One possible data mining problem could be to mine the Internet usage data of a student to find out what academic concepts the student mostly looks up.

Data – The data is the Internet usage logs that the students leave on the default gateway of the network when they browse the Web.

Benefit – By looking for usage patterns in the data, we could understand the topics that students look up most often. These topics could be:

1. Academic – The student is spending a lot of time on the network searching for information about certain subjects on the Internet. This can be further divided into the following categories:
   1. The student looks up topics that are taught in class – This means that there is some sort of divide between the student’s competency level and the level expected from the class.
   2. The student looks up extensions to the topics taught – This means that the student is seeking to study the topics further, either for personal enrichment or for some project(s).
   3. The student looks up topics unrelated to those taught – This could be chalked up to either personal curiosity or personal project(s).
2. Non-academic – This situation also needs careful analysis. One main consideration to take is the amount of time that the student spends on the Internet searching for these topics. Another consideration is the scores that the student achieves in tests. Any conclusions drawn from the data and any actions resulting from them must take the above two points into consideration.

Based on the information retrieved above, the University could attempt to tweak the academic programs. For instance, the academic topics that most students look up most often could be made part of the curriculum, if they are not already. Additionally, the University could use the data to focus on students who are either struggling or not paying attention by intervening at the right time.

Relevant type of data mining – Classification and Clustering seem like they would be relevant to this problem. Since the idea revolves around checking whether the visited pages are relevant or not to the topics taught, clustering and classification would help.

Irrelevant type of data mining – Anomaly detection would be irrelevant here, since we are not looking for data that conforms to a specific pattern.