**Task 1: review data mining concepts and tasks**

Answer the exercise questions 1-3 in Textbook 1.7. For Question 2, feel free to change the question scenario from “an Internet search engine company” to any organization that you would like to think of. It can be a company, government office, NGO, etc.

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

Data Mining Tasks:

1. Dividing the customers of a company according to their gender.
2. Dividing the customers of a company according to their profitability.
3. Predicting the future stock price of a company using historical records.
4. Monitoring the heart rate of a patient for abnormalities.
5. Monitoring seismic waves for earthquake activities.
6. Sorting student database based on student identification number.

Non – Data Mining Tasks:

1. Computing the total sales of a company.
2. Predicting the outcomes of tossing a (fair) pair of dice.
3. Extracting the frequencies of a sound wave.

2. Data Mining for – Amazon (e-commerce website)

Data Mining can help Amazon to boost their marketing and product development in order to raise profits.

Data Mining can be used in Amazon to obtain a lot of data about their customers as well as prospective customers. Information like the visitor’s journey through the site, what they searched for, where they went after their search, how long they are on the page, personal attributes like age, region, sex can help to analyze the behavior of customer and predict their next moves.

Data Mining Techniques like classification can help Amazon to classify customer based on their previous purchases. Also, data like Out of 100 customers 70 were male and 55/100 live in New York, and 80% of them are interested in Basketball can help data mining techniques to build profile as per type of person that visits the site.

Data Mining techniques like clustering can help Amazon to upsell by promoting other items with banners like,

1. Recommended for you. 2. customers who bought these items also bought

Clustering technique along with association rule ( people who buy protein supplements will also by protein shakers) can enable recommendation system and increase cross selling by increasing the average order size.

Data Mining technique can be used in Amazon for fraud prevention. Abnormalities like fake credit card, fake user can be identified by checking the history of the credit card and profile of the user.

3.

A. Data Privacy is an important issue:

1. IP addresses and visits time of web users who visit your website.
2. Names and addresses of people the telephone book
3. Names and email addresses collected from the web.

B. Data Privacy is not an important issue:

1. Census data collected from 1900-1950.
2. Images from Earth-orbiting satellites.

**Task 2: practice your critical thinking and writing**

Read the following two news articles. One criticized Google Flu Trend, and the other defended it. Write one paragraph to summarize the criticism, and another paragraph for the defense. Write the third paragraph to offer your own thought, e.g. is the criticism valid? Does the defense make sense? What other problems or benefit do you see in Google Flu Trend or similar big data applications?

<http>[://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/](http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/)

<http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688>[/](http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/)

1.Criticism:

Google flu trends which once the poster child for power of Big Data Analytics was under criticism for widely over estimating the number of flu cases in the United States of America. It consistently over-estimated the flu cases for two consecutive years by a large percentage of 50. The google estimates were high in 100 out of 108 weeks which clearly indicates how poor the accuracy of this trends were.

After this setback, people were skeptical of Google’s Flu trend algorithm which was used for prediction. The comparative value of the algorithm as a stand-alone flu monitor was questionable. Also, after the update of the algorithm, the google flu trend over estimated the cases by 30%. The recent trend of CDC reports from Doctors on influenza-like illness which lag by two weeks was a more accurate predictor than Google Flu trends.

2.Defense:

Google Flu trend considered 40-flu related queries to predict the prevalence of the flu. It really worked well in 2008and could predict flu trends for 2009. There lot of criticism after it over estimated cases, but the main goal was to compliment traditional practices. For the defense, it was observed that when Google Flu Trends when combined with CDC’s standard monitoring, a better result than either could provide were obtained. Greater value can be obtained GFT with other real time health data. Google flu trend was a system to assist traditional surveillance system and not replace them. Later, when a team examined how to build a better influenza model, the google flu trend model was helpful. In fact, it was the only source of external information to provide statistically significant forecast improvements over the base model. Though, Google Flu Trend did not live up to its expectations, but it has become a base model to develop anything that has predictive ability. Researchers have found GFT and it’s methods to be useful and relevant.

* Criticism that GFT over estimated the cases and it was a failure are valid. But looking at the broader perspective, GFT has laid a foundation for Big Data Analytics. The defense makes sense. Tracking 45 flu related terms over billions of searches, monitoring trends and, making correlations was a huge win for a big data approach. GFT has set an example for further big data analysis. The system served its purpose of complimenting traditional surveillance system. The problems like over-fitting, finding the right signal in the noise, data integration, speed and scalability are found in big data applications. Whereas, big data applications also set base for future generation, help identifying business levers and measuring priorities.