# Project 1: Data and Visualization

Due: see Canvas
Points: 100

Please submit your report in **PDF format.** If you want to submit code, then please submit it in an additional file containing sufficient comments to make it understandable.

For this project we will look at hospitalization data. Important topics for doctors, hospital managers and health care policy makers are the amount of time patients stay in the hospital (length of stay) and the chance that patients need to be readmitted for the same or a related problem (readmission).

Obtain the Hospital data set from the course web site: http://michael.hahsler.net/SMU/EMIS7331/data/hospital/ (username "7331" and password "datam!n!ng"). **Note that this data set is not public and cannot be shared outside this course! All copies of the data set need to be removed after the course is completed.**

The data contains a description file and more information can be obtained from http://www.heritagehealthprize.com/c/hhp (a larger set of data was used for the competition in 2012).

## *Follow the CRISP-DM framework*

**1. Business Understanding** [10]

- Why is it important to know about claims, medication, days spent in the hospital, and readmission rates? Who is interested in this information? What decisions can be informed using such data? [10 point]

**2. Data Understanding** [50]

- Describe the type of data (scale, values, etc.) for each attribute in the files Members_Y1.csv, Claims_Y1.csv and DayInHospital_Y2.csv. [10 point]
- Give simple appropriate statistics (range, mode, mean, median, variance, etc.) for the most important attributes in these files and describe what they mean or if you find something interesting. [10 points]

- Visualize the most important attributes appropriately. Provide an interpretation for each graph. Explain for each attribute type why you chose the visualization. [10 points]
- Explore relationships between attributes: Look at the attributes and then use cross-tabulation, correlation, group-wise averages, etc. as appropriate. [10 points]
- Verify data quality: Are there missing values? Duplicate Data? Outliers? Are those mistakes? How can these be fixed? [10 Points]

**3. Data Preparation** [30]

- Create additional attributes (columns) which might be of interest. For example, for each patient it might be interesting how many claims were filed. Maybe the claims need to be categorized? [20 points]
- Visualize the created attributes. [10 points]

**Exceptional Work** [10 points]