# CSCI 4502/5502: Data Mining
# Homework 1

Due at **12:30pm on Thursday, Jan 29, 2015**. Submit two files electronically at moodle: (1) "**LastName_FirstName_Homework1.pdf**" containing your solutions and (2) "**Last Name_FirstName_Homework1.py**" containing your python source code for Question 3. Make sure to include your name, student id, and the Honor Code Pledge (`http://honorcode.colorado.edu/student-information/honor-code-pledge`).

1. Identify two data sets that are publicly available online and answer the following questions for each of the two data sets:
   (a) URL of the webpage to access the data set.
   (b) Brief description of the data set, including number of objects, number of attributes, attribute types, and other relevant information.
   (c) What knowledge may be mined from this data set?
   (d) How would the knowledge be useful in some applications?
2. (**Required for CSCI 5502 students, optional and 10-point extra credit for CSCI 4502 students**) Check out recent KDD conferences (`http://dl.acm.org/event.cfm?id=RE329&CFID=620349577&CFTOKEN=39120347`, Publication Archive). Pick one paper that interests you and answer the following questions:
   (a) Conference venue (e.g., KDD'14), paper title, authors, and affiliations.
   (b) What problem is addressed? Why is the problem important and challenging?
   (c) A high-level, brief description of the proposed solution (no need to include the details).
   (d) How is the proposed solution evaluated? (data sets, metrics, etc.)
3. Given the Housing Data Set (`http://archive.ics.uci.edu/ml/datasets/Housing`),
   (a) Write a python program that takes two command line variables $i$ and $j$ and compute the following for the $i$-th and $j$-th attributes ($i, j \in [0, 13]$):
      i. $i$-th attribute: $N$ (number of objects), min, max, mean, standard deviation.
      ii. $j$-th attribute: $Q1$, median, $Q3$, and $IQR$.
   (b) Generate a scatter plot using the last two attributes. You can use any plotting tool, such as excel, matlab, gnuplot, R, python, etc. Include the scatter plot in the PDF file you submit.