**University of Waterloo**
**ECE 657A: Data and Knowledge Modeling and Analysis**
**Winter 2019**
**Homework 2:** Data Normalization and ...
**Due:** January 16th, 2019 11:59pm

# Overview

**Collaboration:** Do your work and report individually. You can collaborate on the right tools to use and setting up your programming environment.

**Hand in:** One report per person, via the LEARN dropbox. Also submit the code / scripts needed to reproduce your work. Report as a PDF or a python notebook.

**General Objective:** To study how to apply some of the methods discussed in class on two datasets. The emphasis is on analysis and presentation of results not on code implemented or used.

**Specific Objectives:**
- Establish your software stack to carry out data analysis homework, assignments and the project for the rest of the course.
- Load a simple dataset and perform some basic data preprocessing.

**Tools:** You can use libraries available in python, R or any other programs available to you. You need to mention which libraries you are using, any blogs or papers you used to figure out how to set carry out your calculations.

# Data sets

For this homework you will use the Wine Quality Data Set:

- `http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/`
- Download from Data Folder link, read data set description.

# Tasks

In the class we talked about how to normalize data and introduced basic distance metrics. In the wine quality dataset select the first 10 rows and perform the following computations and report them in a table:

- min-max normalized values, z-score normalized values, mean subtracted normalized values
- for each of the first 10 data points report the nearest and farthest out of the other first 10 points using the following distance metrics (so in the end you will have 10 rows and two columns, nearest and fasthest):
  - manhatten distance

  - euclidean distance
  - cosine distance

- **For fun (on your own):** Plot all the datapoints along each pair of two dimensions using each type of normalization.

## Resources

Blog on running doing some of these calculations in python using SciKitLearn for a similar dataset :
`http://sebastianraschka.com/Articles/2014_about_feature_scaling.html#about-standardization`