# CS670/470 Term Project Phase 2: Advanced Machine Learning in Time Series Analysis Using Recurrent Neural Networks(RNN)

Yong Zhuang, Wei Ding, UMass-Boston 04/23/2019

## 1 Educational Goal

Design and implement advanced machine learning algorithms for time series analysis; Practice how to apply the Recurrent Neural Networks(RNN) for time series analysis.

## 2 Details

**Project Goal:** Using state-of-the-art machine learning libraries to apply the basic RNN and LSTM for streamflow forecasting.
**Project Data:** Twenty-one years from 1996 to 2016, daily streamflow of the Ganges river in India.
**Due Date:** 2:00 pm, May 7, 2019
**Programming language:** Python.
**Package and framework:** scikit-learn, statsmodels, pandas, numpy, colab, Keras.
**Keras**: is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. More details at https://keras.io/

## 3 Tasks

### 3.1 Task 1

Read the online lecture notes at:
https://colab.research.google.com/drive/17I8F6unNOVJSN3xlO1GrxqtaSw0gX4Sb
to learn the workflow of time series analysis and the principle of the basic RNN and LSTM.

### 3.2 Task 2

Load the streamflow data set from http://kdl.cs.umb.edu/CS670/data/Ganges_1996_2016.csv. Create samples by splitting up the sequence into time windows as follows:

$$
\begin{array}{r|l}
\mathbf{X} & y \\
\hline
[S_{1996-01-01}, ..., S_{1996-01-20}] & S_{1996-01-21} \\
[S_{1996-01-02}, ..., S_{1996-01-21}] & S_{1996-01-22} \\
[S_{1996-01-03}, ..., S_{1996-01-22}] & S_{1996-01-23} \\
\vdots &
\end{array}
$$

Where $S_{1996-01-01}$ means the streamflow on 1996-01-01. Split all samples into 3 parts, (1996 – 2009) for training, (2010 – 2011) for validation, and (2012 – 2016) for testing, use the date of $y$ for splitting.

### 3.3 Task 3

Create your basic RNN and LSTM models using the Keras libraries and train your models on the training set. Use the validation set for unbiased evaluation of your models fit on the training dataset while tuning your models' hyperparameters. After training, evaluate your models on the test set and compare the root mean square errors (RMSE) of basic RNN and LSTM. Draw a curve to show the original data and the predictions these two models.

### 3.4 Task 4

Try to make the raw data stationary, create stationary samples as you did in Task 2, and then train a new basic RNN and a new LSTM models on them. Evaluate your models on the test set and compare their performance with the ones of the models you trained on the raw data. Draw a curve to show the original data and the predictions of your all models.

## 4 Submission Requirements

**1.** One team only need to submit one solution (colab notebook).
**2.** Share your team colab notebook with teaching.yong@gmail.com.
**3.** Submit your team colab notebook at online blackboard through the team lead's UMassOnline account.