**MUSA 620: Data Wrangling and Data Visualization**

SCHEDULING Class: Wednesdays from 9am to 12pm. Office hours: Mondays from 1pm to 5pm, by appointment.

OBJECTIVE The purpose of this course is to familiarize students with the “pipeline” approach to data science. This involves the process of gathering data, storing the data, analyzing the data, and visualizing the data such that non-technical decision makers can make sense of it. The course is broken down accordingly into four sections.

1. **Data collection:** Students will learn how to gather data by way of web scraping, APIs, and other unstructured sources.
2. **Databases:** This part of the course teaches students how to store this data for efficient retrieval and analysis.
3. **Analytics:** Students will learn a range of machine-driven techniques for analyzing structured and unstructured data.
4. **Data visualization:** The last part of the course teaches students how to present the results of their analysis visually using R and the web application framework Shiny.

FORMAT The course will be conducted in weekly sessions devoted to lectures, demonstrations and discussions.

ASSIGNMENTS There is one required final project at the end of the semester. Homework will be assigned before the close of each class and will be due at the end of the following week’s class. Four of the homework assignments will be explicitly required. The remainder are optional, but will count toward the participation component of your final grade.

For the final project, students will replicate the pipeline approach on a dataset (or datasets) of their choosing. The final deliverable will be a web-based data visualization and accompanying description including a summary of the results and the methods used in each step of the process (collection, storage, analysis and visualization).

GRADING The grading breakdown is as follows: 50% for homework; 40% for final project, 10% for participation

SOFTWARE This course relies on use of the R Statistical Package in conjunction with Shiny and other associated extensions.

SCHEDULEWeek 1 Jan 18 Introduction

Week 2 Jan 25 Web scraping with R

Week 3 Feb 1 Census data: American Community Survey API

Week 4 Feb 8 Unstructured data: Twitter API

Week 5 Feb 15 Large datasets: NYC Taxi & Uber trip data with Google BigQuery

Week 6 Feb 22 Spatial databases: PostGIS

Week 7 Mar 1 Data frames and data manipulation with R: dplyr

Spring Break

Week 8 Mar 15 Natural language processing: sentiment analysis

Week 9 Mar 22 Data visualization with R: ggplot2

Week 10 Mar 29 Geocoding / mapping with R: ggmap

Week 11 Apr 5 Shiny

Week 12 Apr 12 Shiny

Week 13 Apr 19 Shiny

Week 14 Apr 26 In-class work on final projects