COURSE AND CONTACT INFORMATION

Course: **DATS 6101 INTRODUCTION TO DATA SCIENCE**
Semester: **FALL 2015**

INSTRUCTOR

Name: Dr. Tatiyana (Tanya) V Apanasovich,
Campus Address: 1922 F Street, NW (Old Main) office 301 F
Phone: 202-994-9721
E-mail: apanasovich@gwu.edu
Office hours: MON 4:30-6:00

COURSE DESCRIPTION

The class covers the basic techniques of data science, algorithms for data mining, and basics of statistical modeling. The class focuses on concepts and abstractions, as well as a set of concrete, practical techniques that every data scientist should know. It is an Introduction class hence it is less focused on the details of specific software tools and statistical methodologies. After completing this class, you can expect to be prepared to dive in to deeper data science concepts and to apply the basic techniques in practice. Here is an outline of the course material.

**Introduction**
Data science examples, data science history and context

**Data Management**
Databases and the relational algebra
MapReduce, Hadoop, relationship to databases, algorithms, extensions;
key-value stores and NoSQL; tradeoffs of SQL and NoSQL
Data cleaning, entity resolution, data integration, information extraction

**Data Analytics**
Topics in statistical modeling: linear regression, parametric vs nonparametric regression, Frequentist vs. Bayesian Perspectives, Introduction to Machine Learning: supervised learning, decision trees/forests, simple nearest neighbor

**Interpreting and Communicating Results**
Visualization, visual data analytics
Ethics, privacy, unreliable methods, irreproducible results
COURSE PREREQUISITE(S)

You are expected to have a basic knowledge of statistics (STAT 2118 Regression Methods or a similar course that covers analysis of research data through simple and multiple regression and correlation).

We will use R primarily for its libraries in statistics and machine learning rather than as a general purpose programming environment.

This class is less about specific tools and more about concepts and techniques

TEXTS

The course material will be based on notes prepared by the instructor

OBJECTIVES

1. Explain the major trends in methodology and software tools in the fields of data science and big data.
2. Explain the relative strengths and weaknesses of the data analytics platforms: MapReduce-based platforms, NoSQL solutions, cloud services.
3. Describe a selection of data mining techniques and how they can be implemented
4. Identify and articulate the recurring concepts and abstractions in statistical modeling and machine learning

GRADING

Your final grade will be a weighted average of your homework average (15%), in-class exam (40%), and take home project (45%).

CLASS POLICIES

Homework: There will be 7-9 homework assignments, with greater frequency in the first half of the course. Some assignments will be more analytical, others will deal with data analysis and implementation of procedures in R. Assignments must be completed by students individually, but group discussion is permitted. The due time will be 11:59pm of the due date (usually class date). You have to use Dropbox to submit the homeworks. No late homework will be accepted, but the lowest score will be dropped

Midterm Exam: One open–book in-class examination will be given on Nov,16. Make-up midterm will be given only in exceptional circumstances (e.g. well-documented medical problems).

Take Home Project: Take Home Project is an INDIVIDUAL project with two deliverables:
A proposal for the project- one page long (Due Oct 19)
A report to be handed in (Due Dec 07)
ACADEMIC INTEGRITY
I personally support the GW Code of Academic Integrity. It states: “Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information.” For the remainder of the code, see: http://www.gwu.edu/~ntegrity/code.html

SUPPORT FOR STUDENTS OUTSIDE THE CLASSROOM
DISABILITY SUPPORT SERVICES (DSS)
Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Marvin Center, Suite 242, to establish eligibility and to coordinate reasonable accommodations. For additional information please refer to: http://gwired.gwu.edu/dss/

UNIVERSITY COUNSELING CENTER (UCC) 202-994-5300
The University Counseling Center (UCC) offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include:
- crisis and emergency mental health consultations
- confidential assessment, counseling services (individual and small group), and referrals
  http://gwired.gwu.edu/counsel/CounselingServices/AcademicSupportServices

SECURITY
In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.