Stony Brook University
Department of Technology and Society
College of Engineering and Applied Sciences

EMP 532 – Big Data Systems for Technology Management
Humanities 3018, Mondays 5:30pm-8:20pm

INSTRUCTOR: Dr. Thomas S. Woodson
E-MAIL: thomas.woodson@stonybrook.edu
OFFICE PHONE: (631) 632-9974
OFFICE HOURS: Mondays 2:00-5:00 or by appointment

COURSE DESCRIPTION:

Over the past two decades, the amount of data that is generated has grown exponentially and there is an increasing need to analyze all this data. This class will introduce students to the statistical software R, data analysis, text mining, and big data analyses.

Learning objectives:
- Students will learn to make plots using the GGplot package in R
- Students will learn to do linear regression and logistic regression analyses
- Students will learn how to cluster data with techniques like k-means and hierarchal clustering
- Students will learn basic text mining techniques
- Students will learn basic principles of big data analysis

Prerequisites: Must be a Master’s student in Technology and Society or receive permission from the instructor.

TEXTBOOK:
Data and Goliath: The Hidden Battles to Collect Your Data and Control your World
Bruce Schneier

HELPFUL BOOKS:
- An Introduction to Applied Multivariate Analysis with R by Brian Everitt and Torsten Hothorn
- Big Data Analytics with R, by Simon Walkowiak

ASSIGNMENTS:
Class participation and attendance: 10%
Projects: 60%
Book quizzes: 10%
R Quizzes: 20%

GRADES
A: 94-100
A-: 90-93
B+: 87-89
B: 83-86
B-: 80-82
C+: 77-79
C: 73-76
C-: 70-72
D: 60-69
F: Less than 60

OTHER INFORMATION

Late assignments:
Unless otherwise noted, assignments are due BEFORE class on the day that they are due. If the assignment is turned in late, you automatically lose 10 points on the assignment and you continue to lose another 10 points each day. After 1 week, you will receive a 0 on that assignment.

Attendance/Late Policy:
Attendance to this class is mandatory. Having more than 1 unexcused absence will impact your grade. If you miss a class, you are still expected to do all the readings and assignments for that week. Be on time to class. If you are often late to class, you will lose class participation points.

Electronics Policy:
Silence/turn off your cell phones during the class. If you have an emergency where you need to keep your cell phone on, tell the professor before class. Please NO TEXTING during class. If you use a computer to take notes, please do not surf the web. It is distracting to the other students and the professor.

Computers: Please bring your laptop to class in order to do the class examples during class. If you do not have a laptop, you may share with another classmate.

Software: We will use the software program R for this course. R is an open source data analysis software that is growing in popularity. You can download it from www.r-project.org.

Student Accessibility Support Center Statement: If you have a physical, psychological, medical or learning disability that may impact your course work, please contact Student Accessibility Support Center, ECC (Educational Communications Center) Building, Room 128, (631)632-6748. They will determine with you what accommodations, if any, are necessary and appropriate. All information and documentation is confidential.
Students who require assistance during emergency evacuation are encouraged to discuss their needs with their professors and Student Accessibility Support Center. For procedures and information go to the following website: http://www.stonybrook.edu/ehs/fire/disabilities.

**Academic Integrity Statement:** Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty is required to report any suspected instances of academic dishonesty to the Academic Judiciary. Faculty in the Health Sciences Center (School of Health Technology & Management, Nursing, Social Welfare, Dental Medicine) and School of Medicine are required to follow their school-specific procedures. For more comprehensive information on academic integrity, including categories of academic dishonesty please refer to the academic judiciary website at http://www.stonybrook.edu/commcms/academic_integrity/index.html

**Critical Incident Management:** Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of University Community Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn. Faculty in the HSC Schools and the School of Medicine are required to follow their school-specific procedures. Further information about most academic matters can be found in the Undergraduate Bulletin, the Undergraduate Class Schedule, and the Faculty-Employee Handbook.

**Extra Statistics resources:**

- Create data visualization: http://www.creativebloq.com/design-tools/data-visualization-712402
- DASL (The Data and Story Library): http://lib.stat.cmu.edu/DASL
- JASA (Journal of the American Statistical Association) Data Archive http://lib.stat.cmu.edu/jasadata/
- Statlib-Datasets Archive http://lib.stat.cmu.edu/datasets/
- University of California, Los Angeles Case Studies http://www.stat.ucla.edu/cases/
- U.S. Census Bureau http://www.census.gov
- Stats in the news, from George Mason University: http://www.stats.org/

**Online statistics textbooks and software:**

- Computing for Data Analysis
https://www.coursera.org/course/compdata
- Data Analysis
  https://www.coursera.org/course/dataanalysis
- Rice virtual lab in statistics
  http://onlinestatbook.com/rvls.html
- SISA simple interactive statistical analysis
  http://www.quantitativeskills.com/sisa/

**Online resources for R:**
- The main R project site:
  www.r-project.org
- An R online textbook
  Kickstarting R: http://cran.r-project.org/doc/contrib/Lemon-kickstart/
- Website for the Sarkar book:
  http://lmdvr.r-forge.r-project.org/figures/figures.html
- Quick R website (many helpful “how to” pages)
  http://www.statmethods.net/
- A U. of Wisconsin Stats professor’s site (B. Yandell)
  http://www.stat.wisc.edu/~yandell/software/
- Book on Regression and Anova on the R site:
  http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

**Course calendar subject to change**

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Topic</th>
<th>Assignments Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 1</td>
<td>28-Jan</td>
<td>Intro to Data and R</td>
<td></td>
</tr>
<tr>
<td>Lecture 2</td>
<td>4-Feb</td>
<td>GGPlot</td>
<td></td>
</tr>
<tr>
<td>Lecture 3</td>
<td>11-Feb</td>
<td>Maps in R/Practice</td>
<td></td>
</tr>
<tr>
<td>Lecture 4</td>
<td>18-Feb</td>
<td>R Quiz/Clustering</td>
<td>Project 1: GG Plot Due (Friday)</td>
</tr>
<tr>
<td>Lecture 5</td>
<td>25-Feb</td>
<td>Regression</td>
<td></td>
</tr>
<tr>
<td>Lecture 6</td>
<td>4-Mar</td>
<td>Logistic Regression</td>
<td></td>
</tr>
<tr>
<td>Lecture 7</td>
<td>11-Mar</td>
<td>Book Quiz 1/Practice/Work on projects</td>
<td>Project 2: Data analysis Due (Friday)</td>
</tr>
<tr>
<td></td>
<td>18-Mar</td>
<td>No class, Spring Break</td>
<td></td>
</tr>
<tr>
<td>Lecture 8</td>
<td>25-Mar</td>
<td>Text Mining</td>
<td></td>
</tr>
<tr>
<td>Lecture 9</td>
<td>1-Apr</td>
<td>Book Quiz 2/Text Mining</td>
<td></td>
</tr>
<tr>
<td>Lecture 10</td>
<td>8-Apr</td>
<td>Text Mining/Big Data/Project</td>
<td>Project 3: Text Mining Due (Friday)</td>
</tr>
<tr>
<td>Lecture 11</td>
<td>15-Apr</td>
<td>Map Reduce/Hadoop</td>
<td></td>
</tr>
<tr>
<td>Lecture 12</td>
<td>22-Apr</td>
<td>R Quiz 2/Relational databases</td>
<td></td>
</tr>
<tr>
<td>Lecture 13</td>
<td>29-Apr</td>
<td>Book Quiz 3/Big Data Lecture</td>
<td></td>
</tr>
<tr>
<td>Lecture 14</td>
<td>6-May</td>
<td>Review/Project Time</td>
<td>Project 4: Big Data in the world Due (Friday)</td>
</tr>
</tbody>
</table>