BDA 501
Special Topics
Introduction to Business Data Analytics of Large Scale Data
The City University of New York – Macaulay Honors College – Summer 2016

Dr. Jonathan Peters
Phone: (718) 982-2958
E-mail: jonathan.peters@csi.cuny.edu

W. Michael Johnson
Phone: (302) 438-1969
E-mail: william.johnson@csi.cuny.edu

BDA 501 Course Objectives:

BDA 501 provides an introduction to the analysis techniques that can be applied to large scale data sets. The course will focus on teaching data analysis skills to interested students in a fun and engaging environment. This course is structured to engage the students in direct hands on analysis and to broaden the skills sets that are generally provided by the undergraduate programs in Business, Accounting and Economics. Students from outside of the Business School fields are welcomed to attend and should provide documentation of their level of computational preparation.

Prerequisites:

BUS 215 Information Management
4 hours; 4 credits An introduction to the use and application of Business Information Systems and Technology. It examines how information technologies support business functions, satisfy management needs, and promote collaboration. Students analyze and develop methods of information technology management for organizational effectiveness, strategic advantage, and value creation for different types of organizations. They discuss the impact of information technology on individuals, organizations, and society. Students apply information technology and management concepts and skills to solve practical business problems. Prerequisites: MGT 110

ECO 230 Introduction to Economic and Managerial Statistics
(Also MGT 230)
4 hours; 4 credits Development and application of modern statistical methods, including such elements of descriptive statistics and statistical inference as correlation and regression analysis, probability theory, sampling procedures, normal distribution and binomial distribution, estimation, and testing of hypotheses. Prerequisites: Successful completion of the CUNY Assessment Test in Writing and the CUNY Assessment Test in Reading, and (ECO 101 or ECO 111 or ECO 112) and (MTH 121 or MTH 123 or higher) and (BUS 150 or BUS 215 or BUS 250 or CSC 102 or CSC 126)

Entry Skills:
Strong working knowledge of Excel and the basic softwares of the Microsoft Office set. It is encouraged that students also have a background in Economics and Statistics – with an intermediate undergraduate level of knowledge in these areas is desirable.
Exiting Skills:
Students are expected to acquire skills in statistical analysis programing (SAS, R or STATA) and Geospatial Analysis (ESRI ARC GIS) at an introductory level. Students should be able with guidance be able to produce basic analysis, maps and results using these tools. In addition, students will have an introductory experience with the Supercomputing tools that are available at the CUNY High Performance Computing Center. This will include system operation, data handling and computational tools.

Texts:
1) The Little SAS® Book: A Primer, Fifth Edition Lora D. Delwiche, Susan J. Slaughter 978-1-61290-343-9 $45.32 MSRP.
2) Too Big to Ignore: The Business Case for Big Data, Phil Simon. Publisher: John Wiley Sons Inc. Copyright Date: March 2013. ISBN# 978-1-118-63817-0. $50.00 MSRP

Articles and Papers:
3). Duhigg, C. "What Facebook Learned from Its Quest to Building the Perfect Team" NYT February 25, 2016.

Videos:
Barabasi, L. https://www.youtube.com/watch?v=ni_A2bAkUww
Crawford, S. https://vimeo.com/search?q=susan+crawford+intellectual+property
Nisbett, RL. https://www.edge.org/conversation/richard_nisbett-the-crusade-against-multiple-regression-analysis

Overview
The class will be taught in two sections divided into eight modules. Four modules will be taught by Dr. Peters and four will be taught by Professor Johnson.

Division A- Analytical Skills Acquired By Module:

Module A1 - Data
Students will move from basic analytics in Excel on data sets that push the limits of spreadsheet type analysis – say in the range of 50,000 to 100,000 observations. Further analysis of larger scale data will be performed in a PC environment using a higher level analysis system such as SAS. Here students will utilize the basic analytical skills applied at the introductory level to large scale data sets (3-5 Million Records) that are beyond the capabilities of spreadsheet analysis.

Module A2 – Geospatial Analysis
The second module will introduce students to the tools of geospatial analysis and the use of mapping software in complex large scale data analysis. We look to explore some of the subtle patterns that exist in data systems and teach students to reveal these structures through the use of Geographic Information Systems (GIS). The geospatial component will introduce GIS software (ESRI ARC GIS is a field standard and is site licensed at CUNY) and geospatial methods to the students. Basic data handing in GIS Systems, buffer analysis and data selection will be explored. Finally, an introduction as to how one can do network and feature analysis will be discussed.

Module A3 – Digital Exhaust
We live in a world where data resources are flowing around us as digital exhaust from various information sources – from cell phones to bikeshare systems to Twitter – we find a wealth of raw data flowing out of these systems that can inform and be sourced for system analysis. This module focuses on how we can acquire and datize the vast cloud of information that flows out of these various systems. We will explore the world of Apps and phone technology as well as web site scraping as methods of data acquisition and sources of important content.

Module A4 – Large Scale Data Analysis
This module will introduce students to the method and challenges of large scale data. Discussions of machine limitations, methods and processing techniques will be explore. Students will experiment with large scale data in various formats to develop a familiarity in the challenges that large scale data presents to researchers and policy makers. Students will be introduced to the computational resources of the CUNY High Performance Computing Center.

Division B- Analytical Skills Acquired By Module:

Module B1 - Social Network Analysis
Social networks now seem ubiquitous. But while many have long understood the intrinsic value of social connections, it is only in the last 20 years that Network Science has systemically been able to analyze varied networked relationships. This module will introduce students to network analysis, network visualization, and non-linear data analysis. Discussions will focus on seminal articles that have studied the network structure of organizations and how business networks can change over time. Students will be introduced to Gephi, a network visualization software.

Module B2 - Bias
Intentional and unintentional biases often obscure and limit data analysis. This module will systematically review types of bias that influence and limit the types of information that can be inferred from very large data sets. Students will review how our analytical assumptions and social behavior are subject to systematic biases, by examining case studies in behavioral economics and game theory. Students will use Netlogo to study non-linear systems, computational models, and
network growth models.

Module B3 - Information Policy
We live in a world where data resources are flowing around us both as digital exhaust and vast troves of open source data and code that permit an historically unprecedented exchange of data, information, and knowledge. Yet, as Alphabet CEO Eric Schmidt recently noted “Information at scale wants to regulated by governments.” Students will discuss the impact of the government, private sector, and institutional rules governing big data and machine learning. Students will experiment with large-scale NYC data to develop a familiarity in the challenges that large-scale data presents to researchers and policy makers.

Module B4 – Privacy and Ethics
The final B module will introduce students to the ethical and practical challenges of analyzing and interpreting large-scale data in the context of privacy, machine learning, and personal digital devices. Discussions will examine relationships among laws, corporate governance, institutional review, intellectual property, and personal privacy will affect the future use of large-scale data. The module will focus on identity tracking across space and introduce subtler form of digital identification that have the potential to restrict “choice architectures”.

Notes: BDA 501

1. Students are expected to attend class. More than two absences (excused or unexcused) may result in penalties ranging from grade reduction to course failure.
2. Two hourly examinations will be scheduled during the semester to cover assigned materials and lectures.
3. The final examination will be comprehensive and cover the work of the entire semester, except as specified by the instructor.
4. Makeup examinations are permitted only in extreme and unusual circumstances. Such circumstances must be documented, with the instructor determining the appropriateness of the request. If permission is granted to take such a test, it will be administered after the end of the semester.
5. Cheating is a breach of trust and will be cause for course failure and referral to the appropriate college officials.
6. The instructors reserves the right to add course requirements, as he deems appropriate.
7. Avoid discussion when the instructors are talking.
8. Your final grade will be determined by the average of your exam scores and project (90 percent) with both examinations weighted equally. In addition, 10% of your final grade will be based upon class participation. Class participation grades can be both positive and negative.

Grading Summary: BDA 501

<table>
<thead>
<tr>
<th>Component</th>
<th>Title</th>
<th>Weight in</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam 1</td>
<td>30.0%</td>
</tr>
<tr>
<td>Exam 2</td>
<td>30.0%</td>
</tr>
<tr>
<td>Exam Subtotal</td>
<td>60.0%</td>
</tr>
<tr>
<td>Project</td>
<td></td>
</tr>
<tr>
<td>First Draft</td>
<td>5.0%</td>
</tr>
<tr>
<td>Final Version</td>
<td>15.0%</td>
</tr>
<tr>
<td>Presentation</td>
<td>10.0%</td>
</tr>
<tr>
<td>Project Subtotal</td>
<td>30.0%</td>
</tr>
<tr>
<td>Class Participation</td>
<td>10.0%</td>
</tr>
<tr>
<td>Total</td>
<td>100.00%</td>
</tr>
</tbody>
</table>