Department of Decision, Operations and Information Technologies
University of Maryland

BUDT 758C
Big Data: Strategy and Analytics
Fall 2017

Instructors: **Guodong (Gordon) Gao** and **Anand Gopal**
4325 / 4307 Van Munching Hall
301-405-2218/ 301-405-9681, {ggao, agopal}@rhsmith.umd.edu

Lab Session Lead: **Zheng Gong** <zheng.gong@rhsmith.umd.edu>
Yage Wang <yage.wang@rhsmith.umd.edu>
Yan Yang <yan.yang2@rhsmith.umd.edu>

Class Meets: DC Campus, Tu 6:25pm - 10:00pm

Office Hours: On request
Version: **V3**

Course Introduction

Digitization is occurring in every aspect of business and our daily life. As a result, huge amount of data is being generated. Big data represents unprecedented opportunities for companies to generate insights and create wealth. At the same time, much of the big data is unstructured, in real time and only loosely connected. It defies the traditional ways of managing databases. This creates challenges even for tech-savvy companies on how to leverage the big data to gain competitive advantages. In recent years, the move into Big Data has also helped kickstart developments in AI and deep learning, with great potential for the disruption of several technology-based industries.

This course has three objectives. First, it will provide an overview, and some hands-on experience with, the basic elements of the big data technical ecosystem, as well as an introduction into the concepts that underlie the big data architecture. Second, we will discuss the potential and observed disruption of existing industry segments where opportunities from big data have emerged, through a series of case studies. These case discussions will highlight opportunities as well as threats and limitations that are faced by firms as they navigate through elements of the big data ecosystem. Finally, through a group project, you will study the impact of big data and its related functionalities in specific industry verticals – the objective here to is to understand, from a managerial perspective, where value can be created using big data technologies in a specific industry setting, and what factors may influence this process of creating value.

In summary, we will use a hands-on, learning-by-doing approach to understanding the concepts behind Big Data and AI, the essential technologies used within this ecosystem, the strategic drivers of technology and the value propositions provided to industries. The focus is on creating awareness of the technologies,

allowing some level of familiarity with them, enabling strategic thinking around the use of these in business, and implementation and management of Big Data and AI.
The technology of Big Data and AI continues to evolve rapidly. Therefore, there is a level of experimentation with new material that will take place during the semester. Students are required to be flexible as topics or material in class are revised or modified. We will do our best to ensure that no undue burden is placed on students.

Learning Objectives
At the end of the course, students should be able to:

1. Have working knowledge on key elements of the big data technology platform, such as Hadoop, NoSQL, and Spark;
2. Understand major types of big data, methods to capture and store big data, and analytical tools on big data
3. Gain understanding of the Deep Learning based Artificial Intelligence platforms.
4. Learn how to identify strategic values of big data, how to manage a big data project, and avoid pitfalls in big data.

Prerequisites

1. Experience with databases, specifically working knowledge of relational databases and SQL, is highly desirable.
2. Working knowledge of Linux/Unix is useful but not required.
3. A willingness to experiment and play with software in a relatively unstructured environment – not all the technologies in this space are fully tested, and so some frustration is inevitable.
4. **A laptop with at least 8 GB memory and 30 GB free disk space**.

Software Needed
Much of the software needed for big data applications tend to be open source. Therefore, the source programs are free. However, for the purposes of the course, we will be using versions provided by the Cloudera Academic Partnership, Amazon Web Services, and Google Cloud.

All the access will be provided by the instructors.

1. Cloudera CDH Virtual Machine
2. Amazon Web Services
3. Google TensorFlow

Required Reading Material
A significant proportion of the reading material for this course is available online and is free. When necessary, additional reading material will be posted on Canvas/ELMS.

Optional useful sources are listed below; these are not required but are good reference material.

1. Hadoop: The Definitive Guide, by Tom White (http://it-ebooks.info/book/5629/)
2. Big Data: A Revolution That Will Transform How We Live, Work, and Think, by Viktor Mayer-Schonberger and Kenneth Cukier (http://www.big-data-book.com/)
3. Mining of Massive Datasets. Hardcopy: Amazon.com E-version: Free available here
4. Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (https://github.com/HFTrader/DeepLearningBook/blob/master/DeepLearningBook.pdf)

Course Format and Grading

Classes

We meet once each week. The class format will be lectures and case discussion, hands-on exercise, and labs. Given the diversity of the topic, there will be guest lectures in class, with visitors from industry presenting their own perspectives on the value of big data.

Assignments
We have multiple lab sessions and homework assignments. These assignments are mainly from the lectures. These assignments will help you understand concepts and ideas you've learned from the class. Screenshots showing successful completion of the labs will need to be turned in on Canvas – more details will be provided in class.

Class project
The class project will be carried out in groups of no more than 5 students. The objective of the project is to collect information, understand, and evaluate the potential disruptive nature of big data and AI technologies in specific industry sectors – think of this as a consulting engagement you may be asked to perform for your own CEO in order to answer the broad question: "What are the threats and opportunities offered by big data and AI in our industry segment? Where should we focus our attention?"

Rather than focus on a firm, you should use the industry segment as the unit of analysis. Therefore, candidate industry segments (verticals) include

- Mobile retail
- Entertainment (movies, video, streaming services)
- Hospitality
- Education (higher ed, public school systems, online and MOOCs)
- Transportation
- Economic development and philanthropy
- Insurance and financial services
- Electricity generation, storage, and distribution – retail and wholesale

Each of these sectors are facing tremendous opportunities from the use of big data/AI. You may also choose an alternative segment, but please discuss this first with the instructors. Further details on specific project deliverables will be provided in class.

Case Write-ups
There are four cases in the course, of which you are required to turn in at least two for the purposes of grading. You may choose which ones to turn in. If you turn in more than 2, the best two will count towards your grade. The case write-ups should address the relevant questions on the value of big data highlighted in the case, and should be professionally prepared. Each case write-up should be no longer than 2 pages in length, Times New Roman 11 point-font, single-line spacing, 1" margins. Please ensure that you make the best use of this space in addressing the points in the case. The actual cases to be used will be provided in class.

Quizzes
There will be two short quizzes conducted in class – each quiz will be for 15 minutes duration and will address the relevant details of the course material covered prior to that point in the semester.

Lab Assignments
Part of the value of this course comes from being able to gain some hands-on experience with big data technologies, namely elements of the Hadoop and Spark ecosystems. In each session, we will spend some time walking you through various operations on the Hadoop system through the use of a virtual machine provided by Cloudera. You are required to follow along with the lab on your own machines, but are not

required to turn in your work on these labs for credit. Should you choose to do so, there will be instructions provided on how to submit your completed labs for extra credit, upto a maximum of 10%.

Class Participation

An interactive approach to learning is vital in this class and class participation is an important part of the learning experience. We will try to provide an environment conducive to discussion in the class and will expect you to contribute to the class through your thoughts and ideas. Since 15% of your grade depends on class participation, poor attendance will affect the class participation grade.

Grading

Your final grade for the course will be composed from the following items:

Class participation (individual):	15%*1 = 15%
Case write-ups (individual):	10%*2 = 20%
Quizzes (individual):	15%*2 = 30%
Final project (group):	35%*1 = 35%
(bonus) Lab Assignments (individual):	2%*5 = 10%

Academic Integrity

The Robert H. Smith School of Business recognizes honesty and integrity as necessary cornerstones to the pursuit of excellence in academic and professional business activities. The University's *Code of Academic Integrity* is designed to ensure that the principles of academic honesty and integrity are upheld. All students are expected to adhere to this Code. The Smith School does not tolerate academic dishonesty. All acts of academic dishonesty will be dealt with in accordance with the provisions of this code. Please visit the following website for more information on the University's *Code of Academic Integrity:* http://www.inform.umd.edu/CampusInfo/Departments/JPO/AcInteg/code_acinteg2a.html

Plagiarism Policy: Inevitably in a programming course, it seems that a few people will turn in work that is not their own. You should understand that it is usually easy to detect copying of programs -- even when a program is modified to try to disguise its source. Copying a program, or letting someone else copy your program, is a form of academic dishonesty and the penalties can be found ***here***.

Schedule (subject to change)

Session	Topics	Lab	Assignment Due
8/29/2017	Introduction. Big Data – Why business should care		
	Business Value of Big Data – Frameworks	Lab 2: Set up Cloudera Training Virtual Machine	
9/5/2017	Overview of the Hadoop Ecosystem, HDFS, MapReduce, Yarn, and Hue	Lab 3: Yarn and Hue	Team formation due
	Case #1		Case write-up due
9/12/2017	Essential Data Tools in Big Data: Sqoop	Lab 4: Sqoop	Case write-up due Quiz 1 (take home)
	Case #2		
9/19/2017	Essential Data Tools in Big Data: Pig and Hive	Lab 6: Hive Lab x: Pig Set up AWS	
9/26/2017	Big Data on the Cloud: AWS NoSQL, Databases on Clusters	Run Pig / Hive on AWS	Case write-up due
	Case #3		
10/3/2017	Spark – Introduction		
	Spark Application Development (Clustering – quick explanation)	Lab 10: RDD Lab 16: K-means clustering	Quiz 2 (take home)
10/10/2017	AI and Deep Learning		Project Discussion
	Case #4		
10/17/2017	Group project presentation		Final Group Project Report Due