Department of Decision, Operations and Information Technologies
University of Maryland

BMGT758B
Big Data and Artificial Intelligence for Business
Fall 2019

Instructors: **Guodong (Gordon) Gao**
VMH 4325
ggao@umd.edu

Tech support: BigDataHelpMe@gmail.com
Office Hours: go.umd.edu/bigdatahelpme

Office Hours: Wed 2-3. Appointment or Drop by welcome
Version: **V2.0**

Course Introduction

Big data represents unprecedented opportunities for companies to generate insights and create wealth. Internet of Things (IoT) is connecting almost all the components together in every aspect of business and our daily life. As a result, huge amount of data is being generated. 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 to tech-savvy companies on how to leverage the big data to gain competitive advantages. Challenges and opportunities coexist. To extract the great value from the data, we should be equipped with advanced techniques. Artificial Intelligence (AI) is penetrating our daily routines deeply, and shows great promise in exciting areas such as healthcare and autonomous driving cars.

This course uses a hands-on, learning-by-doing approach to understanding the concepts behind Big Data and AI, the strategic drivers of these technologies and the value propositions that they provide to industries. In addition, the course will also serve as an introduction to some of the key technologies within this ecosystem, such as Hadoop, AWS, Pig, Hive, Amazon Web Services and Spark. Examples of AI using Deep Learning will be conducted in class. The focus is on creating awareness of the technologies, allowing some level of familiarity with them through assignments, and enabling some strategic thinking around the use of these in business.

The technologies are still evolving very 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 and when 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
The course has two primary objectives:

1. To allow students to have working knowledge and exposure to key elements of a big data technology platform, and a basic AI example
2. To allow students to understand critical business and strategic issues around the use of these technologies in organizations and to help guide the successful design and implementation of complex data strategy

Though mastery of this content requires more than one course, an introductory course, such as this, is useful in allowing students to gain much-needed familiarity with these technologies and concepts. That is the objective of this course.

Prerequisites

1. Databases, specifically working knowledge of relational databases and SQL
2. Working knowledge of Linux/Unix is useful but not required.
3. A laptop with at least **8 GB memory** and **30 GB free space on the hard drive**.

Software Needed
Much of the software needed for big data applications tend to be open source. Therefore, the source programs are free and will be provided in class. We will be using materials provided by the Cloudera Academic Partnership, Amazon Web Services, as well as Google's TensorFlow.

1. Cloudera CDH VM – provided by instructors (Hadoop, Pig, Hive, Impala)
2. Amazon Web Services – provided through the instructors
3. Python + TensorFlow – open source

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. A Human's Guide to Machine Intelligence, by Kartik Hosanagar (https://www.amazon.com/Humans-Guide-Machine-Intelligence-Algorithms/dp/0525560882)
5. 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 twice each week. Mondays will be mostly lecturing, and Wednesdays will be mostly in-class exercise.

Assignments

We have case analysis, lab assignments, and individual reports throughout the semester.

Class project
There is a class project for each group. Each team will be assigned a dataset and a specific set of questions associated with that dataset – the team will be responsible for analyzing the data and providing a series of analyses that will help answer the questions. In addition, the team will provide a short report on the project outcomes, including screen shots and samples of code, as specified. Further details of the projects will be provided in class.

Grading

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

Class participation:	10%*1	= 10%
Case Analysis:	5%*2	= 10%
Lab Assignment:	5%*8	= 40%
Quiz:	5%*2	= 10%
Class project:	30%*1	= 30%

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 learn more information about the University's *Code of Academic Integrity* ***here***.

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.

Schedule (subject to change)

8/27/2019	Introduction		
8/29/2019	Business Value of Big Data		
9/3/2019	Case Discussion – Kyruus Dataset Introduction	Case Analysis	Case write-up due
9/5/2019	Case Discussion –IBM Watson	Case discussion	Case write-up due
9/10/2019	Introduction to AI - 1		
9/12/2019	Introduction to AI - 2		
9/17/2019	Overview of the Hadoop Ecosystem -1 HDFS		
9/19/2019	Overview of the Hadoop Ecosystem -2 MapReduce		
9/24/2019	Go through the virtual machine	Lab 2: Set up Cloudera Training Virtual Machine	
9/26/2019	Apache Foundation. Yarn and Hue	Lab 3: Yarn and Hue	
10/1/2019	Sqoop	Lab 4: Sqoop	Quiz 1
10/3/2019	Pig	Lab x: Pig	
10/8/2019	Hive	Lab 6: Hive	
10/10/2019	HBase		
10/15/2019	NoSQL		
10/17/2019	AWS – 1		
10/22/2019	AWS – 2	Lab on AWS (Pig and Hive on Cloud)	
10/24/2019	AWS – 3	Lab on AWS (SSH and CLI)	
10/29/2019	AWS – 4		
10/31/2019	Spark – 1	Lab 10: RDD Lab 13: Spark App	
11/5/2019	Spark – 2	Lab 16 – K-means Clustering	
11/7/2019	Spark – 3		#Spark Lab 1 assignment due
11/12/2019	Spark – 4		
11/14/2019	Spark – 5		
11/19/2019	Deep Learning -1		#Spark Lab 2 assignment due
11/21/2019	Deep Learning -2	Lab: PyTorch	
11/26/2019	Deep Learning -3	Lab: CNN and RNN	Quiz 2
11/28/2019	Thanksgiving		
12/3/2019	Deep Learning -4		
12/5/2019	Deep Learning -5		
12/10/2019	Group Project Presentations Session		Group Project Report Due