UMBC UGC New Course Request: CMSC462 – Introduction to Data Science

Date Submitted: 2/14/2020 Proposed Effective Date: 8/23/2020

	Name	Email	Phone	Dept
Dept Chair or UPD	Jeremy Dixon	jdixon@umbc.edu	5-8866	CSEE
Other Contact	Mohamed Younis	younis@umbc.edu	5-3969	CSEE

COURSE INFORMATION:

Course Number(s)	CMSC462
Formal Title	Introduction to Data Science
Transcript Title (≤30c)	Intro to Data Science
Recommended Course Preparation	
Prerequisite **NOTE:** Unless otherwise indicated, a prerequisite is assumed to be passed with a "D" or better.	CMSC 341 and (STAT 355, STAT 451, or CMPE 320) each with a grade of C or better
# of Credits Must adhere to the UMBC Credit Hour Policy	3
Repeatable for additional credit?	☐ Yes ☒ No
Max. Total Credits	3 This should be equal to the number of credits for courses that cannot be repeated for credit. For courses that may be repeated for credit, enter the maximum total number of credits a student can receive from this course. E.g., enter 6 credits for a 3 credit course that may be taken a second time for credit, but not for a third time. Please note that this does NOT refer to how many times a class may be retaken for a higher grade.
Grading Method(s)	☒ Reg (A-F) ☒ Audit ☒ Pass-Fail

PROPOSED CATALOG DESCRIPTION (Approximately 75 words in length. Please use full sentences.):

Data science is a field that involves data manipulation, analysis, and presentation, all at scale. It's typical for an organization to have a few terabytes of data maintained for different purposes by different business units stored in different formats, and for someone to have an idea about how the data might bring significant additional value. Data scientists are the bridge between the idea and the data and help extract latent value, often uncovering novel insights and novel beneficial ways to use the data in the process.

RATIONALE FOR NEW COURSE:

a) Why is there a need for this course at this time?
This course is mandatory in our data science track and has been taught many times as CMSC 491 – Special Topics in Computer Science

b) How often is the course likely to be taught?
We will most likely offer it every semester (Fall and Spring).

c) How does this course fit into your department's curriculum?
It can be used as a technical elective for all computer science majors. Additionally, it is required for all data science track – computer science majors and it can be used for the new AI/ML track.

d) What primary student population will the course serve?
This course will primarily serve Juniors, and Seniors in the CMSC department.

e) Why is the course offered at the level (ie. 100, 200, 300, or 400 level) chosen?
The course has significant analysis and design elements for an "Introductory" course. It builds on concepts introduced in CMSC 341- Data Structures, and STAT 355 - Introduction to Probability and Statistics for

Scientists and Engineers with additional considerations for databases, data manipulation, and cloud computing.

f) Explain the appropriateness of the recommended course preparation(s) and prerequisite(s).
A student is more likely to be successful in this course if they are adequately prepared with programming (CMSC 341 – Data Structures) and statistics (STAT 355 - Introduction to Probability and Statistics for Scientists and Engineers).

g) Explain the reasoning behind the P/F or regular grading method
Students are most likely to take this course using A-F but on occasion a student could audit it or taking it P-F.

h) Provide a justification for the repeatability of the course.
This course cannot be repeated for additional credit.

ATTACH COURSE SYLLABUS (mandatory):

CMSC 462: Introduction to Data Science

Prerequisites:

CMSC 341 and (STAT 355, STAT 451, or CMPE 320) each with a C or better.

Instructor:

Name: TBD
Office: TBD
Office Hours: TBD
Phone: TBD
Email: TBD

Course Description:

Data science is a field that involves data manipulation, analysis, and presentation, all at scale. It's typical for an organization to have a few terabytes of data maintained for different purposes by different business units stored in different formats, and for someone to have an idea about how the data might bring significant additional value. Data scientists are the bridge between the idea and the data and help extract latent value, often uncovering novel insights and novel beneficial ways to use the data in the process.

The goal of this class is to give students hands on experience with all phases of the data science process using real data and modern tools. Topics that will be covered include data formats, loading, and cleaning; data storage in relational and non-relational stores; data analysis using supervised and unsupervised learning, and sound evaluation methods; data visualization; and scaling up with cloud computing, MapReduce, Hadoop, and Spark.

Credits:

Three credits: not repeatable

Learning Outcomes:

At the end of the course, the student will:

- Organize and transmit data using tools such as text files, JSON, or other language specific libraries.
- Demonstrate data management processes including data loading, cleaning, summarization, and outlier detection.
- Analyze data storage options including SQL and NoSQL and where data can be stored and processed including cloud computing options.
- Use statistical methods and modeling to summarize data and identify relationships.
- Explain how to formulate new hypotheses and draw accurate conclusions from data.
- Develop effective visualizations of given data.
- Evaluate ethical and privacy considerations of data sets and apply ethical practices.

Course Format and Procedures:

The core concepts of data science are programming language independent, but Python has a powerful set of open source tools for doing data science at scale that we will leverage, as do many organizations, both large and small. Specifically, we'll use Anaconda, which bundles "over 100 of the most popular Python, R and Scala packages for data science" and provides easy access to hundreds more through the Anaconda package manager.

The elements of Anaconda that are most relevant to the tripartite structure of this course are (1) **pandas** (the Python Data Analysis Library), which provides ways to load data into a data frame for easy manipulation and analysis, (2) **scikit-learn**, which is a set of "simple and efficient tools for data mining and data analysis", and (3) **matplotlib**, which is "a python 2D plotting library [that] produces publication quality figures in a variety of hardcopy formats and interactive environments".

Two other tools that will figure prominently are Spark, an extremely powerful framework for data manipulation in cluster computing environments, and Jupyter Notebook, a "web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text". We'll use the former to explore the power of cloud computing with Amazon's EC2, and the latter to interactively explore data and present results.

Please note that you will get your hands dirty in this class! You will be required to install software, read and use online documentation, solve problems by googling for answers, read posts on Stack Overflow, and so on. Data science is a broad and rapidly changing field, so one of the most valuable skills you can cultivate is the ability to dive in and solve problems, either your own or the client's. You will by no means be on your own, with support from me, the TA, and your classmates. But the first thing I will ask when you come to me with a question is "what have you already tried?", and the list of things you've tried must have length ≥ k where k is at least 2.

Readings:

Baumer, B., Kaplan, D., & Horton, N. J. (2017). Modern Data Science with R. Boca Raton: CRC Press,Taylor & Francis Group, CRC Press.

Dietel, P. and Dietel, H. (2020). Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and The Cloud. Pearson. ISBN: 9780135404669

Dietrich, D., Heller, B., Yang, B., & EMC Education Services (Eds.). (2015). Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Indianapolis, IN: Wiley.

Larose, C. D., & Larose, D. T. (2019). Data Science Using Python and R. Hoboken, NJ: John Wiley & Sons,Inc.

Course Topics:

Students in Introduction to Data Science will participate by:

- Introduction: Introduction to data science and setting up your environment.
- Background and data acquisition: Types of data and data representations. How to acquire, process, and parse data. Data manipulation, data wrangling, and data cleaning.
- Data engineering and processing: Use SQL, NoSQL, and other storage methods.
- Visualization and basic statistics: Visualization principles and goals. How to communicate results.

Visualizing distributions and relationships.

- Statistical analysis: Evaluation, cross-validation, overfitting, clustering, dimensionality reduction, and other practical concerns. Learning predictive models from data.
- Ethical considerations: Examine how data can generate privacy and ethical considerations.

Grading:

Description	Quantity	Points	Total
Homework	6	8.5	51
Project	1	34	34
Final Exam	1	15	15

Grading is on a standard 10-point scale, so you will get an A for 90.0 or more total points, a B for 80.0 or more but less than 90.0 points, and so on.

The homework assignments will be a blend of practical exercises and questions that cement conceptual knowledge. The project can best be described as *grab some data from OpenBaltimore and do something interesting*. While that may sound glib, it's an accurate representation of a typical data science assignment in the real world. From the OpenBaltimore website:

The goal of OpenBaltimore is to provide, to the public, access to City data in an effort that supports government transparency, openness and innovative uses that will help improve the lives of Baltimore residents, visitors and businesses through use of technology

Your task is to browse the data, choose a collection of datasets, load and analyze them using the conceptual and concrete tools that we'll cover, and derive and communicate interesting insights. Imagine that you're trying to get a job as a data scientist for the city of Baltimore, and this will be your interview portfolio. You need to tell the people in city hall something they don't already know about their data and that they would be interested in knowing.

Academic Integrity

By enrolling in this course, each student assumes the responsibilities of an active participant in UMBC's scholarly community in which everyone's academic work and behavior are held to the highest standards of honesty. Cheating, fabrication, plagiarism, and helping others to commit these acts are all forms of academic dishonesty, and they are wrong. Academic misconduct could result in disciplinary action that may include, but is not limited to, suspension or dismissal. To read the full Student Academic Conduct Policy, consult the Academic Integrity Resources for Students page (https://aetp.umbc.edu/ai/resources-for-students/) or the Faculty Handbook (http://provost.umbc.edu/faculty-handbook/), specifically Sections 14.2-14.3.

If you need help with a project, see your instructor, your TA, or tutors provided by the Learning Resource Center. We also encourage you to consult textbooks and code examples provided on Blackboard. Consult Blackboard for additional Academic Integrity policies for projects.

Any act of dishonesty will be reported to the University's Academic Conduct Committee for further action, which may include, but is not limited to, academic suspension or dismissal from the University.

We will be using special software to check for cheating. The software is quite sophisticated and has surprised many students in the past. There is no difficulty in comparing every pair of assignments – even assignments submitted to other sections of this course, or from previous semesters.

This is a *non-exhaustive* list of restrictions for completing your assignments in this course.

- **If you have questions about what is acceptable, please contact a professor or TA.**

You may not look at, access, download, or obtain anyone else's work.

- You should think carefully about the assignment, and the assignment you turn in should be entirely a product of your own understanding of the material.
- You may not use any online resources to request additional help. Please contact a professor or TA for additional help.
- You may not post any part of a course document online. Posting any slides, projects, or labs will be considered a violation of this course policy and will result in an "F" for the course.
- You may not look at someone else's code "for reference," even if you put it aside before programming, and even if that person is not a CMSC student.
- You may not Google or search for the solution to an assignment, even if it's "only for reference."
- You may not copy code other than that provided in the course materials (slides, book, labs, etc.).
- You may not let someone else explain a solution to you in such detail that they are effectively dictating the code to you line by line. It does not matter if this person has never taken this course, or if they are not looking at their own code while doing so!

Student Disability Services:

UMBC is committed to eliminating discriminatory obstacles that may disadvantage students based on disability. Services for students with disabilities are provided for all students qualified under the Americans with Disabilities Act (ADA) of 1990, the ADAAA of 2009, and Section 504 of the Rehabilitation Act who request and are eligible for accommodations. The Office of Student Disability Services (SDS) is the UMBC department designated to coordinate accommodations that would allow students to have equal access and inclusion in all courses, programs, and activities at the University.

If you have a documented disability and need to request academic accommodations for access to your courses, please refer to the SDS website at sds.umbc.edu for registration information and to begin the process, or alternatively you may visit the SDS office in the Math/Psychology Building, Room 212. For questions or concerns, you may contact us through email at disAbility@umbc.edu or phone (410) 455-2459.

If you require accommodations for this class, make an appointment to meet with your instructor to discuss your SDS-approved accommodations.

Disclosures of Sexual Misconduct and Child Abuse or Neglect

As an instructor, I am considered a Responsible Employee, per UMBC's Policy on Prohibited Sexual Misconduct, Interpersonal Violence, and Other Related Misconduct (located at http://humanrelations.umbc.edu/sexual-misconduct/umbc-resource-page-for-sexual-misconduct-and-

other-related-misconduct/). While my goal is for you to be able to share information related to your life experiences through discussion and written work, I want to be transparent that as a Responsible Employee I am required to report disclosures of sexual assault, domestic violence, relationship violence, stalking, and/or gender-based harassment to the University's Title IX Coordinator. As an instructor, I also have a mandatory obligation to report disclosures of or suspected instances of child abuse or neglect (www.usmh.usmd.edu/regents/bylaws/SectionVI/VI150.pdf).

The purpose of these reporting requirements is for the University to inform you of options, supports and resources; you will not be forced to file a report with the police. Further, you can receive support and resources, even if you choose to not want any action taken. Please note that in certain situations, based on the nature of the disclosure, the University may need to act.

If you need to speak with someone in confidence about an incident, UMBC has the following Confidential Resources available to support you:
The Counseling Center: 410-455-2472
University Health Services: 410-455-2542
(After-hours counseling and care available by calling campus police at 410-455-5555)

Other on-campus supports and resources:
The Women's Center, 410-455-2714
Title IX Coordinator, 410-455-1606
Additional on and off campus supports and resources can be found at:
http://humanrelations.umbc.edu/sexual-misconduct/gender-equitytitle-ix/

Tentative Schedule:

Week	Topics	Assignments
1	Course overview, introduction to data science, setting up your environment (Anaconda and iPython Notebook)	
2	Introduction to Pandas and dataframes, CSV, json, and minimal visualization capabilities	HW1 Due
3	More visualization. Data loading, cleaning, summarization, and outlier detection	
4	SQL, NoSQL, key/value stores, connecting to a database from Python	HW2 Due
5	Building models, trees for classification, scikit-learn	
6	Trees for regression, linear regression	HW3 Due
7	Logistic regression, support vector machines	
8	Evaluation, cross-validation, overfitting, practical concerns	HW4 Due
9	Clustering, dimensionality reduction, practical concerns	
10	Data visualization	HW5 Due
11	Cloud computing, scaling up, Amazon EC2	
12	MapReduce and Hadoop	HW6 Due
13	Spark (the MapReduce killer)	
14	Spark - part 2	
15	Topics that spilled over from prior weeks (e.g., EC2, a little Spark)	Project Due
16	Final Exam (TBD)	Final Exam

This schedule is subject to change without notification from the professor.