CMPSC 301 – Data Analytics
Syllabus

Spring 2020

Course Instructor

Dr. Oliver BONHAM-CARTER
Classroom: Alden Hall 101
Office Location: Alden Hall 104
Office Phone: +1 814-332-2880
Email: obonhamcarter@allegheny.edu
Web Site: http://www.cs.allegheny.edu/sites/obonhamcarter/
Exam Code: A
Final deliverable due: Thursday, 30th April 2020, 9:00 pm

Instructor's Office Hours

Instructor's Office Hours

- Monday, Wednesday *and* Friday: 10:00 pm – 12:00 pm (10 minute time slots)

- By appointment, if these times do not work for you.

To schedule a meeting with me during my office hours, please visit my Web site and click the "Schedule" link in the top right-hand corner. Now, you can view my calendar or by clicking "schedule an appointment" link browse my office hours and schedule an appointment by clicking the correct link to reserve an open time slot.

Course Meeting Schedule

- **Lecture, Discussion, Presentations, and Group Work:**

    - 13th January - 29th April 2020
    - **Lecture**; Monday, Wednesday *and* Friday 9:00AM - 9:50AM, Alden Hall, Room 101
    - **Lab**; Monday, 2:30PM - 4:20PM, Alden Hall, Room 101

Academic Bulletin Description

An introduction to computational methods of data analysis with an emphasis on understanding and reflecting on the social, cultural, and political issues surrounding data and its interrogation. Participating in hands-on activities that often require teamwork, students study, design, and implement analytics software and learn how to extract knowledge from, for instance, financial, political, and scientific sources of data. Students also investigate the biases, discriminatory views, and stereotypes that may be present during the collection and analysis of data, reflecting on the ethical implications of using the resulting computational techniques. During a weekly laboratory session, students use state-of-the-art statistical software to complete projects, reporting on their findings through both written documents and oral presentations. Prerequisite: FS*102 or permission of the instructor. Distribution Requirements: QR, PD.

Course Objectives

Students successfully completing this class will have developed:

1. A "big-picture" view of data analytics.

2. An understanding of the objectives and limitations of data analytics.

3. An understanding of the main data analytics methods.

4. Practical skills using relevant software tools and programming techniques.

5. An understanding of the contemporary roles of power and difference as they relate to the knowledge derived from a data set.

6. An understanding of biases, discrimination and stereotypes that maybe present during collection, analysis, and reflection on the latent trends in real-world data sets.

The course is divided into modules, with several of the modules featuring guest visits by Allegheny professors from four distinct disciplines and consisting of investigations of real-world data in their specific field. In addition to learning specific technical and programming skills in each module students will be required to read a relevant article and prepare for a discussion related to the issues raised in the article.

Students will also enhance their ability to write and present ideas about data analytics in a clear and compelling fashion. Finally, students will gain practical experience in the design, implementation, and analysis of data for research during laboratory sessions and a final project.

An Ethical Interest: Throughout the semester students will be challenged with serious analytical questions connecting the investigated data and its analysis to arising societal issues of bias, ethical consideration and the culture of power. This step is to ensure that analytics is performed with a lens on the data, as well as its impacts (positive and negative) on culture, community, and society. We note here that there is often no clear indication of a "correct" decision as a result of an analysis of data. The so-called "right" decision ought to be made by analysis who has studied both the data, and the consequences of decision in terms of humanitarian, environmental, ecological and other factors. This class cannot give you the correct decision, however it can help to enable your critical thinking skills which will provide you with some understanding of how to navigate to worthy decisions.

Textbooks

The material for this course will be taken from the book listed below and from the additional readings that will be provided for you. It is highly recommended that you obtain a copy of the book for your study in this course.

- Wickham, Hadley, and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.*, O'Reilly Media, Inc., 2016.

    - Link to book's website: `http://r4ds.had.co.nz/`

- Silge, Julia, and David Robinson. *Text mining with R: A tidy approach.* O'Reilly Media, Inc.", 2017.

    - Link to the books website: `https://www.tidytextmining.com/`

- Along with reading the required books, you will be asked to study many additional articles from a wide variety of conference proceedings, journals, and the popular press.

Students who want to improve their technical writing skills may consult the following books.

- Crapsi, Linda. *Bugs in Writing: A Guide to Debugging Your Prose.* Technical Communication 42.4 (1995): 665-667., ISBN-10: 020137921X, ISBN-13: 978-0201379211, 704 pages, 1998.

- Zobel, Justin. *Writing for computer science.* Vol. 8. New York NY: Springer, 2004., ISBN-10: 1852338024, ISBN-13:978-1852338022, 270 pages, 2004.

Class Policies

Grading

The grade that a student receives in this class will be based on the following categories. All percentages are approximate and, if the need to do so presents itself, it is possible for the assigned percentages to change during the academic semester.

Class Participation	10%
First Examination	10%
Second Examination	10%
Laboratory Assignments	40%
Final Project (Due December 13th by 7pm)	30%

Note: final exam is code A, and thus the final project is due 30th of April at 9:00am.

Definitions of Grading Categories

- *Class Participation*: All students are required to actively participate during all of the class sessions. Your participation will take forms such as answering questions about the required reading assignments, completing in-class exercises, asking constructive questions of the other members of the class, giving presentations, and leading a discussion session in class.

- *First and Second Examinations*: The first and second examinations will cover all of the material in their associated module(s). While the second examination is not cumulative, it will assume that a student has a basic understanding of the material that was the focus of the first examination. The date for the first and second examinations will be announced at least one week in advance of the scheduled date. Unless prior arrangements are made with the course instructor, all students will be expected to take these examinations on the scheduled date and complete the tests in the stated period of time.

- *Laboratory Assignments*: These assignments invite students to explore the concepts, tools, and techniques associated with the analysis of data. All of the laboratory assignments require the use of the provided tools to study, design, implement, and evaluate systems that solve data analytics problems. To ensure that students are ready to utilize and develop software in both other classes at Allegheny College and after graduation, the instructor will assign individuals to teams for some of the laboratory assignments. Unless specified otherwise, each laboratory assignment will be due at the beginning of the laboratory session on the specified date. In addition to demonstration of the technical skills through the utilized or developed software for data analysis, most of the laboratory assignments in this course will also expect students to read a related article and to lead a discussion or to give a short presentation related to the assigned article.

- *Final Project*: This project will present you with the description of a problem and ask you to implement a full-featured solution using a wide variety of data analytics techniques. The final project in this class will require you to apply all of the knowledge and skills that you have accumulated during the course of the semester to solve a problem and, whenever possible, make your solution publicly available as a free and open-source tool. The project will invite you to draw upon both your problem solving skills and data analytics techniques.

Assignment Submission

We will be using GitHub Classroom to collect all assignments. It is expected that you are able to effectively use `git` to submit your work. If you require help, please see your peers, the Technology Leaders, or your instructor.

All assignments will have a stated due date. The electronic version of the class assignments are to be turned in at the beginning of the lab session on that due date. Submissions after the beginning of class are counted as being late. Assignments will be accepted for up to one week past the assigned due date with a 15% penalty. All late assignments must be submitted at the beginning of the session that is scheduled one week after the due date.

Extensions

Unless special arrangements are made with the course instructor, no assignments will be accepted after the late deadline. If you are requesting extensions for a lab assignment, then you are to email me with your request and also provide a *valid reason* for your extension. This request must come before the due date of the lab and not on the due date. Requests will not be granted where the reason appears to be insignificant. Extensions are 24 hours of extra time (after the original due date) and are given out at my discretion. The decision to provide you with an extension (or not) will be weighed in light of fairness to your peers who are still able to complete their labs, regardless of their own busy schedules.

The submission of homework comprises the Honor Code pledge of the student(s) completing the work. For any assignment completed in a group, students must also turn in a one-page reflection that describes each group member's contribution to the submitted deliverables.

Attendance

It is mandatory for all students to attend the class and laboratory sessions. If you will not be able to attend a session, then please see/email the course instructor at least one week in advance to describe your situation. Students who miss more than five unexcused classes, laboratory sessions, or group project meetings will have their final

grade in the course reduced by one letter grade. Students who miss more than ten of the aforementioned events will automatically fail the course.

Use of Laboratory Facilities

Throughout the semester, we will experiment with many different tools that data managers use during the phases of the data management process. The course instructor and the department's systems administrator have invested a considerable amount of time to ensure that our laboratories support the completion of both the laboratory assignments and the final project. To this end, students are required to complete all assignments and the final project while using the department's laboratory facilities. The course instructor and the systems administrator will only be able to devote a limited amount of time to the configuration of a student's personal computer.

Class Preparation

In order to minimize confusion and maximize learning, students must invest time to prepare for class discussions and lectures. During the class periods, the course instructor will often pose demanding questions that could require group discussion, the creation of a program or test suite, a vote on a thought-provoking issue, or a group presentation. Only students who have prepared for class by reading the assigned material and reviewing the current assignments will be able to effectively participate in these discussions. More importantly, only prepared students will be able to acquire the knowledge and skills that are needed to be successful in both this course and the field of data management. In order to help students remain organized and effectively prepare for classes, the course instructor will maintain a class schedule with reading assignments and presentation slides. During the class sessions students will also be required to download, use, and modify programs, diagrams, and data sets that are made available through the course GitHub repository.

Email

Using your Allegheny College email address, I will sometimes send out class announcements about matters such as assignment clarifications or changes in the schedule. It is your responsibility to check your email at least once a day and to ensure that you can reliably send and receive emails. This class policy is based on the following statement in *The Compass*, the college's student handbook.

> "The use of email is a primary method of communication on campus. ...All students are provided with a campus email account and address while enrolled at Allegheny and are expected to check the account on a regular basis."

Disability Services

The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. Students with disabilities who believe they may need accommodations in this class are encouraged to contact Disability Services at 332-2898. Disability Services is part of the Learning Commons and is located in Pelletier Library. Please do this as soon as possible to ensure that approved accommodations are implemented in a timely fashion.

Honor Code

The Academic Honor Program that governs the entire academic program at Allegheny College is described in the Allegheny Course Catalog. The Honor Program applies to all work that is submitted for academic credit or to meet non-credit requirements for graduation at Allegheny College. This includes all work assigned for this class (e.g., examinations, laboratory assignments, and the final project). All students who have enrolled in the College will work under the Honor Program. Each student who has matriculated at the College has acknowledged the following pledge:

> *I hereby recognize and pledge to fulfill my responsibilities, as defined in the Honor Code, and to maintain the integrity of both myself and the College community as a whole.*

Additionally, we expect that you will adhere to the following Department Policy:

Department of Computer Science Honor Code Policy

It is recognized that an important part of the learning process in any course, and particularly in computer science, derives from thoughtful discussions with teachers, student assistants, and fellow students. Such dialogue is encouraged. However, it is necessary to distinguish carefully between the student who discusses the principles underlying a problem with others, and the student who produces assignments that are identical to, or merely variations on, someone else's work. It will therefore be understood that all assignments submitted to faculty of the Department of Computer Science are to be the original work of the student submitting the assignment, and should be signed in accordance with the provisions of the Honor Code. Appropriate action will be taken when assignments give evidence that they were derived from the work of others.