# CMPSC 301 – Data Analytics
# Syllabus

Fall 2021

## Course Instructor

Dr. Oliver BONHAM-CARTER (said and written as "Bonham-Carter," not "Carter")
Email: `obonhamcarter@allegheny.edu`
Web Site: `http://www.cs.allegheny.edu/sites/obonhamcarter/`
Class and lab meeting place: Alden 101
Final deliverable due: 9am, $13^{th}$ December 2021
Final Exam Code: D
Distribution Requirements: QR, PD

Syllabus updated on: February 14, 2022

Discord channel:

- `https://discord.gg/vsntc3R7`
  (Note: this link will expire in 7 days from $31^{st}$ August 2021!)

Google Calendar:

- `https://calendar.google.com/calendar/u/0?cid=`
  `Y18wdXI3ZHFqOHFpNmh2azd2bDAwMmJoaWlzY0Bncm91cC5jYWxlbmRhci5nb29nbGUuY29t`
  (Note: This link is all on one line!)

### The `ClassDocs/` Class Archive

- Get using ssh keys:
  - `git clone git@github.com:Allegheny-Computer-Science-301-F2021/classDocs.git`
- Get without using ssh keys:
  - `git clone`
    `https://github.com/Allegheny-Computer-Science-301-F2021/classDocs.git`

### Instructor's Office Hours

- Office hours will take place online using Zoom or in-person in my office in Alden 104.
- 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. Here, you can browse my office hours slots to schedule an appointment. If using Zoom, you will find a link with the meeting invitation. At the allotted time, I will be awaiting your meeting. If the given office hour meeting times are not convenient for your schedule, please let me know and I would be happy to work with you to find and other time which would be suitable for your schedule.
- By Zoom or in-person in Alden 104 (please let me know which it is!)
- Tuesday: 2:45 pm – 4:45 pm EST (15 minute time slots)
- Wednesdays and Thursdays : 2:00pm – 4:00 pm EST (15 minute time slots)
- By appointment, if these times do not work for you.

### Course Meeting Schedule

- **Lecture, Discussion, Presentations, Group Work, and Labs:**
  - $24^{th}$ August - $15^{th}$ December 2021
  - **Class**: Tuesday, *and* Thursday 11:10 AM - 12:25 PM EST - New York time.
  - **Lab**: Tuesday 3:00 PM - 4:50 PM EST - New York time.

## 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.

### Distribution Requirements

The following definitions were taken from the *Distribution Requirements: Learning Outcomes* website, `https://sites.allegheny.edu/registrar/academic-policies/graduation-requirements/distribution-requirement/distribution-requirements-learning-outcomes/`.

- Quantitative Reasoning (QR). Quantitative Reasoning is the ability to understand, investigate, communicate, and contextualize numerical, symbolic, and graphical information towards the exploration of natural, physical, behavioral, or social phenomena.
  - Learning Outcome: Students who successfully complete this requirement will demonstrate an understanding of how to interpret numeric data and/or their graphical or symbolic representations.
- Power, Privilege, and Difference (PD). Understanding Power, Privilege, and Difference means understanding the role of power, privilege, prejudice, discrimination, stereotypes, inequity, and oppression in human society, in both historical and contemporary contexts, and recognizing these dynamics in the learner's own life and communities.
  - Learning Outcome: Students who successfully complete this requirement will demonstrate an understanding of the historical and/or contemporary roles of power, privilege, and difference in human society.

This course meets the course distribution requirements of QR (Quantitative Reasoning) and PD (Power, Privilege, & Difference) for its use of applying concepts of computer programming to the analysis of public data concerning events where there is a detectable bias at play. In addition, the class aims to introduce an component of ethical reasoning when drawing conclusions.

## 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 consisting of investigations of real-world data in a 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. Students will receive their reported grades approximately once a week through an individual GitHub grading repository.

| | |
|---|---|
| Class Participation | 10% |
| First Examination | 10% |
| Second Examination | 10% |
| Laboratory Assignments | 40% |
| Final Project (Due $8^{th}$ December by 9 am) | 30% |

These grading categories have the definitions which are defined below.

## Definitions of Grading Categories

- *Class Activities*: All students are required to actively participate during all of the class sessions. Your participation will mainly take a form of completing class activities that are submitted via an appropriate GitHub repository. You may also be asked to answer questions about the required reading assignments, give presentations, and lead a discussion session in class.
- *First and Second Quizzes*: The first and second quizzes will cover all of the material in their associated module(s). While the second quiz is not cumulative, it will assume that a student has a basic understanding of the material that was the focus of the first quiz. Unless prior arrangements are made with the course instructor, all students will be expected to take these quizzes on the scheduled date and complete them 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. In addition

to demonstration of the technical skills through the utilized or developed software for data analysis, some of the laboratory assignments in this course may 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 or your instructor.

All assignments will have a stated due date. The electronic version of the assignments are to be turned in at the beginning of the class session on that due date. Submissions after the beginning of class are counted as being late.

### 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 an 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 assignment 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 assignments, 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.

### Bring your own computer to class

During the semester, you will be told which software to install on your machine to be prepared for class. Some of the prominent software that we may be using include;

- Git and GiHub (a software development software system): `https://github.com/`

- Atom (an editor): `https://atom.io/`
- Docker (a software container system): `https://www.docker.com/`
  - Basic tutorial from Docker: `https://www.docker.com/101-tutorial`
  - Play with Docker: `https://labs.play-with-docker.com/`
  - Please note: machines running Windows "Home" are not able to use Docker. Please verify that your machine is able to run the software by visiting the department's Approved Laptops page `https://www.cs.allegheny.edu/resources/laptops/`.
- R (a programming language): `https://www.r-project.org/`
  - Please see `https://www.r-project.org/about.html` for more information.
- RStudio (An integrated development environment for R): `https://rstudio.com/`
- Information about using Docker to run RStudio: `https://hub.docker.com/r/rocker/rstudio`

### 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, 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 analytics. In order to help students remain organized and effectively prepare for classes, the course instructor will maintain a class schedule. During the class sessions students will also be required to download, write, use, and modify programs, 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 (814) 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 and in *The Compass: Student Handbook*. 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.

omitit

## Attendance

- **Remote Attendance** If you are participating entirely remotely this semester and relying on technology to attend class meetings, occasional technology problems

that disrupt your participation will not harm your participation grade, but as with illnesses and family emergencies, chronic absences for this reason will require a more extensive discussion with me and may impact your grade.

- **Video and Microphones** Please turn off your microphone when not speaking during any meeting where you are using your computer. The microphone may allow for background sound to contribute to noise during the meeting. It is strongly encouraged that you use your video to show yourself during meeting. Enabling your video will allow the instructor to see hands to indicate questions. Showing video also helps to stimulate group discussions.

## College Messages

- **Statement of Community** Allegheny students and employees are committed to creating an inclusive, respectful and safe residential learning community that will actively confront and challenge racism, sexism, heterosexism, religious bigotry, and other forms of harassment and discrimination. We encourage individual growth by promoting a free exchange of ideas in a setting that values diversity, trust and equality. So that the right of all to participate in a shared learning experience is upheld, Allegheny affirms its commitment to the principles of freedom of speech and inquiry, while at the same time fostering responsibility and accountability in the exercise of these freedoms.

- **Learning Commons** If you are not already, you should become familiar with the Learning Commons, located in Pelletier Library (`http://sites.allegheny.edu/learningcommons/`). Among other things, the staff at the Learning Commons can assist you with study and time management skills, writing, and critical reading. You should know that if you are having trouble in this class, or if I think you can specifically benefit from their services, I will refer you to the Learning Commons. Experienced peer writing and speech consultants in the Learning Commons help writers and speakers to determine strategies for effective communication and to make academically responsible choices at any stage in the writing or speaking process and on assignments in any discipline. Both appointments and drop-in sessions are available. To view the hours of operation, and to make an appointment, visit the Learning Commons website.

- **Religious Accommodations** If you need to miss class or reschedule a final examination due to a religious observance, please speak to the professor well in advance to make arrangements. See `http://sites.allegheny.edu/religiouslife/religious-holy-days/`.