Course Syllabus

18-788: Big Data Science
Spring 2023

Instructor: Prof. Patrick McSharry
Email Address: mcsharry@cmu.edu
Course Management Assistant: Stephanie Bolla
Email Address: sbolla@andrew.cmu.edu
Office Location: HH 1113
Waitlist Manager: Megan Oliver
Email Address: mvoliver@andrew.cmu.edu
Office Location: HH 1113B

Course Website: https://www.andrew.cmu.edu/course/18-788

Teaching Assistants:

Name:	Email:	Zoom Meeting Id:
Christian Iradukunda	Christianiradukunda2@gmail.com	https://cmu.zoom.us/my/christian.iradukunda
Seth Abayo	sabayo@andrew.cmu.edu	https://cmu.zoom.us/my/sabayo
Remy Patrick Mwizerwa	rmwizerw@andrew.cmu.edu	https://cmu.zoom.us/my/rmwizerw
Gabriel Ntwari	ntwarig@andrew.cmu.edu	https://cmu.zoom.us/my/Gabriel Ntwari
Ange Ernestine Iradukunda	aeiraduk@andrew.cmu.edu	https://cmu.zoom.us/my/aeiraduk
Neeraj Garole	ngarole@andrew.cmu.edu	https://cmu.zoom.us/j/2852423291
Innocent Mukoki	imukoki@andrew.cmu.edu	https://cmu.zoom.us/my/imukoki

Course Discipline: Electrical and Computer Engineering
Course Level: Graduate
Type of Course: Onsite (Pittsburgh) & Remote (Kigali)
Course Streamed to Pittsburgh: Yes
Course Concentration: Applied Machine Learning
Core/Elective: Core
Units: 6
Lecture/Lab/Rep Hours/Week: 3 Lecture Hours Per Week; 1 Lab/Rep Hours Per Week
Semester/Year Offered: Spring, All Years
Pre-Requisites: 18-785

Class Schedule:
Pittsburgh: *Hamerschlag Hall 1107*

Kigali: *F305*

Lectures: *1 hour and 20 minutes*

Week	A	Pittsburgh	Kigali	B	Pittsburgh	Kigali
1	Tues, Mar 14	08:00 ET	14:00 CAT	Thurs, Mar 16	08:00 ET	14:00 CAT
2	Tues, Mar 21	08:00 ET	14:00 CAT	Thurs, Mar 23	08:00 ET	14:00 CAT
3	Tues, Mar 28	08:00 ET	14:00 CAT	Thurs, Mar 30	08:00 ET	14:00 CAT
4	Tues, Apr 04	08:00 ET	14:00 CAT	Thurs, Apr 06	08:00 ET	14:00 CAT
5	Tues, Apr 11	08:00 ET	14:00 CAT	Thurs, Apr 13	08:00 ET	14:00 CAT
6	Tues, Apr 18	08:00 ET	14:00 CAT	Thurs, Apr 20	08:00 ET	14:00 CAT

Course Description:

The proliferation of mobile technology, wireless sensors and social media provides a means of monitoring socio-economic activity, consumption of resources and human mobility. Recent advances in data science are now capable of coping with the technical challenges of collecting, managing and developing actionable insights from big data. Partnerships between academia, government and the private sector are at the heart of the revolution that is currently demonstrating how data is a valuable commodity and a source of intellectual property. This course will take a practical approach to solving challenges in the public and private sectors using a collection of techniques that constitute this new multidisciplinary field known as data science. A number of different themes will be explored as case studies in order to demonstrate how big data collected from a wide range of disparate sources can be combined to provide insights, drive decisions and influence policy. The course content will be structured to provide a roadmap for deploying data science techniques using case studies, reading material and previously published models. Participants will obtain hands-on experience by working on real-world datasets during assignments.

Learning Objectives:

The objective of this course is to provide students with practical experience of the different techniques and skills that constitute the field of data science. In particular, these case studies are selected to demonstrate the technical challenges of dealing with the three V's that define big data (volume, velocity and variety). The various steps required will include: (1) exploration of data using visualization techniques; (2) construction of features; (3) evaluation of a collection of models; and (4) consideration of how a decision-maker can utilize the analysis; and (5) development of a dashboard for displaying the results of the analysis. The sources of big data will range from surveys to mobile data to satellite imagery and therefore involve both structured and unstructured data.

Outcomes

After completing this course, students should be able to:

- Identify sources of big data in response to a specific challenge
- Download and organize data for addressing the challenge
- Explore the dataset using visualization techniques
- Develop a number of features to extract information
- Construct a range of quantitative models
- Discuss the advantages and disadvantages of different models
- Select an approach that is optimal for meeting the objective
- Present conclusions and recommendations

- Communicate model output to decision-makers

Responsibility:
Students take full responsibility for reading all communications, following advice given in recitations and paying careful attention to verbal and written instructions given by the instructor and TAs. It is assumed that students taking the course have read the syllabus and agree to abide by the terms and conditions set out here.

Grading

The grades for this course will be based on students' performance on three homework assignments, Canvas quizzes, a final exam and class participation. Homework assignments will be done individually and turned in via canvas by the designated due date. Late work that is deemed complete will be accepted until 24 hours past the deadline, but it will lose 10%. The assignments will be graded based on both a written report and working code with clear comments used to achieve results presented in the report. Class participation will be evaluated based on the student's contribution to discussions both in-class and on the Piazza Discussion Board. When posting or reacting to online discussion threads, students are expected to use their own words and the post should be relevant to the topic under discussion. Please make sure to introduce, summarize and explain the article in your own words to enlighten the audience on the point the article is making.

The following is the weight distribution of the grades:

Class participation	**5%**
Canvas quizzes	**2.5%**
Piazza participation	**2.5%**
Assignment 1	**22.5%**
Assignment 2	**22.5%**
Assignment 3	**22.5%**
Final Exam (Multiple Choice)	**22.5%**

Important Dates:

Description	Release Date	Due Date	Days	Grade Date
Assignment 1	Tue, Mar 14	Mon, Mar 27	14	Fri, Mar 31
Assignment 2	Tue, Mar 28	Mon, Apr 17	21	Fri, Apr 21
Assignment 3	Tue, Apr 18	Mon, May 01	14	Fri, May 05
Final Grades				Thu, May 11

* The final exam date is to be announced but likely to be the first/second week of May.

Assignments:
Each assignment is explained in as much detail as possible in the assignment document. Piazza threads and recitations serve to provide more information and clarifications. The objective is to perform calculations, data analysis and produce key findings (summary statistics, tables, graphics and visualizations). These findings are, however, only part of the solution as students are also expected to provide step-by-step explanations, insights and recommendations in words. This ability to clearly communicate and comment on each finding is a key aspect of the course. Furthermore, submitted code should be clearly

documented with many comments in order to demonstrate knowledge of what each calculation is doing and why it is necessary. Plagiarism checks will be carried out and students sharing and copying will receive penalties. A penalty of 50% will be applied for the first offense. Repeat plagiarism offenses will result in a penalty of 100% and an AIV report being filed. Students are advised to ensure that submissions do not contain any material that could initiate a plagiarism warning.

Assignment submission format:

The assignment document provides instructions about the submission format. However, the instructor and TAs have the right to change or update the submission instructions at any time before the deadline and will communicate accordingly. Therefore it cannot be assumed that instructions for a past assignment are relevant for future assignments. The contents and format of the submissions are extremely specific and demand close attention. These submission requirements are not negotiable and form part of our approach to identify plagiarism and AIV. Only a complete submission containing all requested files in the correct format, with requested naming conventions will be accepted. Submissions deemed incomplete will be rejected. It is the responsibility of each student and not the TAs to ensure that their submissions are completed and submitted before deadlines.

Deadlines:

It is expected that deadlines are respected and taken seriously. Students have the responsibility of uploading a complete submission on time. Missing a deadline by between 0 and 24 hours will result in the deduction of 10% of the marks assigned. A submission made more than 24 hours after the deadline, will be rejected and result in zero marks. A student that submits an incomplete submission and does not complete it within 24 hours of the deadline will also obtain zero in that assignment.

Regrade requests:

Once assignment grades have been provided, students have 48 hours to submit a regrade request using the Google form that will be provided. The TA who graded the question will email and lead on all communications to ensure that appropriate steps are taken and that a decision can be provided. Please allow for up to one week for a final decision.

Exceptional circumstances:

In the case of an unavoidable exceptional circumstances which prevent a student from meeting a deadline, the student is required to follow the protocol: (1) immediately email the instructor and TAs of the course at least 48 hours in advance of the deadline; (2) copy his/her faculty advisor; and (3) provide a letter from student services or another staff member at CMU to support the claim that this exceptional circumstance was sufficiently severe to prevent the student from meeting the deadline. An example of an exceptional circumstance is the sudden onset of a serious medical illness that is outside the student's control. Internet connectivity, technical difficulties, poor time management, failure to read communications or having multiple assignments due at the same time cannot be considered as exceptional circumstances.

Participation:

The participation component is designed to reward students that communicate effectively to help others problem-solve. In order to be fair to those students that spend time being constructive and offering help to others, any excessive posting of messages on Piazza that

do not contribute any information (short posts such as Thanks, Noted, etc) may at the discretion of the TAs and instructor result in a zero for this participation component.

MATLAB:
One option is to use MATLAB software. Download MATLAB software for your computer operating system from the CMU download website. After unzipping the file, read the Matlab-Licence_Instructions.pdf file for instructions on connecting to the MATLAB server and running MATLAB.

Python
Another option is to use Python. Google Colaboratory (Colab) is an excellent platform for building notebooks which allow you to combine executable code and rich text in a single document, along with images, HTML, LaTeX.

Canvas and Piazza:
Canvas will be used for posting supplementary course materials and turning in assignments. Please familiarize yourself with navigating, uploading and downloading. Piazza will be used for questions and discussion among students, TAs and the instructor.

Topic Outline (Weeks 1-6)

Big Data Science 18-788			
Week	Activity	Lecture A	Lecture B
1	**Theme**	**Weather & agriculture**	**Climate change**
	Challenge	Agricultural productivity	Global warming
	Discussion	Historical data	Open data policy
	Case Study	Weather and tea	UK Met Office
	Analysis	Correlations	Trend analysis
	Demo	Index insurance	Climatefrontier.com
2	**Theme**	**Climate scenarios**	**Catastrophe models**
	Challenge	Economic impacts	Financial losses
	Discussion	Weather dependence	CAT model
	Case Study	Weather derivatives	Hurricanes
	Analysis	Tourism climatic index	Exceedance probability
	Demo	Tourist scenarios	EP curves
3	**Theme**	**Social trends**	**Finance**
	Challenge	Surveys versus big data	Forecasting the stock market
	Discussion	Google trends	Web searches
	Case Study	Flu forecasting	Google trends and the market
	Analysis	Unemployment	Anomalies
	Demo	CPI portal	Environmental impact
4	**Theme**	**Sentiment analysis**	**Health**
	Challenge	The meaning of words	Big data and health
	Discussion	Quantifying words	Benchmarks
	Case Study	Twitter and S&P500	Linear or nonlinear
	Analysis	Twitter sentiment	Drug trials
	Demo	Tweets and floods	Significance

5	**Theme**	**Telemedicine**	**Mobile data**
	Challenge	Personalized healthcare	Human activity
	Discussion	Weather and health	Privacy versus reward
	Case Study	Asthma	Malaria
	Analysis	Diabetes	Mobility and transport
	Demo	Activity classification	Mobility visualization
6	**Theme**	**Data4Dev**	**Socioeconomic status**
	Challenge	Monitoring and evaluation	Measuring poverty
	Discussion	Global goals	MPI
	Case Study	Social impact bonds	Give directly
	Analysis	Development impact bonds	Mobile CDRs and satellite imagery
	Demo	Ibrahim index	Predicting MPI

Grading Scale:

A : > 93% A - :]93% , 90%]	C + :]80%, 77%] C :]77%, 73 %] C - :]73%, 70%]
B + :]90% , 87%] B :]87% , 83 %] B - :]83% , 80%]	D + :]70%, 67%] D :] 67%, 60%] F : < 60 %

Education Objectives (Relationship of Course to Program Outcomes)
The ECE department is accredited by ABET to ensure the quality of your education. ABET defines 7 Educational Objectives that are fulfilled by the sum total of all the courses you take. The following list describes which objectives are fulfilled by this course and in what manner they are fulfilled. The objectives are numbered from "1" through "7" in the standard ABET parlance. Those objectives not fulfilled by this course have been omitted from the following list:

1) **an ability to identify, formulate, and solve complex engineering problems by applying principles of engineering, science, and mathematics:** The course poses many problems in homework assignments involving data from complex real-world systems that will require students to think critically, analyse, model and solve using good engineering practice.
2) **an ability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors:** The course introduces the use of analytics and machine learning to address specific challenges and facilitate improved decision-making for practitioners and managers in public and private sectors.
3) **an ability to communicate effectively with a range of audiences:** Students can practice communication skills remotely (Piazza Q&A), during class (quizzes, polls, questions and answers) and especially during recitations where challenges are discussed and volunteers with solutions are welcome to present.

4) **an ability to recognize ethical and professional responsibilities in engineering situations and make informed judgments, which must consider the impact of engineering solutions in global, economic, environmental, and societal contexts:** The ethical aspects of analytics and machine learning are discussed in relation to data sampling biases, model error, risk and uncertainty. Case studies are used to demonstrate the potential adverse consequences of AI.
5) **an ability to function effectively on a team whose members together provide leadership, create a collaborative and inclusive environment, establish goals, plan tasks, and meet objectives:** Quizzes and polls are used as a mechanism to encourage group discussion and motivate debate about the process of interpreting data and arriving at implementable solutions.
6) **an ability to develop and conduct appropriate experimentation, analyze and interpret data, and use engineering judgment to draw conclusions:** The course focuses extensively on the interpretation and analysis of data to predict outcomes and generate actionable insights based on transparent processes that can be documented and explained.
7) **an ability to acquire and apply new knowledge as needed, using appropriate learning strategies:** Students will tackle problems using real-world datasets and will select from amongst a range of techniques and models to undertake statistical learning and arrive at a robust and predictive solution.

ECE Academic Integrity Policy (http://www.ece.cmu.edu/programs-admissions/masters/academic-integrity.html**):**

The Department of Electrical and Computer Engineering adheres to the academic integrity policies set forth by Carnegie Mellon University and by the College of Engineering. ECE students should fully and carefully review Carnegie Mellon University's policies regarding Cheating and Plagiarism; Undergraduate Academic Discipline; and Graduate Academic Discipline. ECE graduate student should further review the Penalties for Graduate Student Academic Integrity Violations in CIT outlined in the CIT Policy on Graduate Student Academic Integrity Violations. In addition to the above university and college-level policies, it is ECE's policy that an ECE graduate student may not drop a course in which a disciplinary action is assessed or pending without the course instructor's explicit approval. Further, an ECE course instructor may set his/her own course-specific academic integrity policies that do not conflict with university and college-level policies; course-specific policies should be made available to the students in writing in the first week of class.

This policy applies, in all respects, to this course.

CMU Academic Integrity Policy (http://www.cmu.edu/academic-integrity/index.html**):**
In the midst of self-exploration, the high demands of a challenging academic environment can create situations where some students have difficulty exercising good judgment. Academic challenges can provide many opportunities for high standards to evolve if students actively reflect on these challenges and if the community supports discussions to aid in this process. It is the responsibility of the entire community to establish and maintain the integrity of our university.

This site is offered as a comprehensive and accessible resource compiling and organizing the multitude of information pertaining to academic integrity that is available from across the university. These pages include practical information concerning policies, protocols and best practices as well as articulations of the institutional values from which the policies and protocols grew. The Carnegie Mellon Code, while not formally an honor code, serves as the foundation of these values and frames the expectations of our community with regard to personal integrity.

This policy applies, in all respects, to this course.

The Carnegie Mellon Code
Students at Carnegie Mellon, because they are members of an academic community dedicated to the achievement of excellence, are expected to meet the highest standards of personal, ethical and moral conduct possible.
These standards require personal integrity, a commitment to honesty without compromise, as well as truth without equivocation and a willingness to place the good of the community above the good of the self. Obligations once undertaken must be met, commitments kept.
As members of the Carnegie Mellon community, individuals are expected to uphold the standards of the community in addition to holding others accountable for said standards. It is rare that the life of a student in an academic community can be so private that it will not affect the community as a whole or that the above standards do not apply.
The discovery, advancement and communication of knowledge are not possible without a commitment to these standards. Creativity cannot exist without acknowledgment of the creativity of others. New knowledge cannot be developed without credit for prior knowledge. Without the ability to trust that these principles will be observed, an academic community cannot exist.
The commitment of its faculty, staff and students to these standards contributes to the high respect in which the Carnegie Mellon degree is held. Students must not destroy that respect by their failure to meet these standards. Students who cannot meet them should voluntarily withdraw from the university.

This policy applies, in all respects, to this course.

Carnegie Mellon University's Policy on Cheating
(http://www.cmu.edu/academic-integrity/cheating/index.html) states the following:
According to the University Policy on Academic Integrity, cheating "occurs when a student avails her/himself of an unfair or disallowed advantage which includes but is not limited to:

- Theft of or unauthorized access to an exam, answer key or other graded work from previous course offerings.
- Use of an alternate, stand-in or proxy during an examination.
- Copying from the examination or work of another person or source.
- Submission or use of falsified data.
- Using false statements to obtain additional time or other accommodation.
- Falsification of academic credentials."

This policy applies, in all respects, to this course.

Carnegie Mellon University's Policy on Plagiarism (http://www.cmu.edu/academic-integrity/plagiarism/index.html) states the following: According to the University Policy on Academic Integrity, plagiarism "is defined as the use of work or concepts contributed by other individuals without proper attribution or citation. Unique ideas or materials taken from another source for either written or oral use must be fully acknowledged in academic work to be graded. Examples of sources expected to be referenced include but are not limited to:

- Text, either written or spoken, quoted directly or paraphrased.
- Graphic elements.
- Passages of music, existing either as sound or as notation.
- Mathematical proofs.
- Scientific data.
- Concepts or material derived from the work, published or unpublished, of another person."

This policy applies, in all respects, to this course.

Carnegie Mellon University's Policy on Unauthorized Assistance (http://www.cmu.edu/academic-integrity/collaboration/index.html) states the following: According to the University Policy on Academic Integrity, unauthorized assistance "refers to the use of sources of support that have not been specifically authorized in this policy statement or by the course instructor(s) in the completion of academic work to be graded. Such sources of support may include but are not limited to advice or help provided by another individual, published or unpublished written sources, and electronic sources. Examples of unauthorized assistance include but are not limited to:

- Collaboration on any assignment beyond the standards authorized by this policy statement and the course instructor(s).
- Submission of work completed or edited in whole or in part by another person.
- Supplying or communicating unauthorized information or materials, including graded work and answer keys from previous course offerings, in any way to another student.
- Use of unauthorized information or materials, including graded work and answer keys from previous course offerings.
- Use of unauthorized devices.
- Submission for credit of previously completed graded work in a second course without first obtaining permission from the instructor(s) of the second course. In the case of concurrent courses, permission to submit the same work for credit in two courses must be obtained from the instructors of both courses."

This policy applies, in all respects, to this course.

Carnegie Mellon University's Policy on Research Misconduct (http://www.cmu.edu/academic-integrity/research/index.html) states the following:

According to the University Policy for Handling Alleged Misconduct in Research, “Carnegie Mellon University is responsible for the integrity of research conducted at the university. As a community of scholars, in which truth and integrity are fundamental, the university must establish procedures for the investigation of allegations of misconduct of

research with due care to protect the rights of those accused, those making the allegations, and the university. Furthermore, federal regulations require the university to have explicit procedures for addressing incidents in which there are allegations of misconduct in research."

The policy goes on to note that "misconduct means:

- fabrication, falsification, plagiarism, or other serious deviation from accepted practices in proposing, carrying out, or reporting results from research;

- material failure to comply with Federal requirements for the protection of researchers, human subjects, or the public or for ensuring the welfare of laboratory animals; or

- failure to meet other material legal requirements governing research."

"To be deemed misconduct for the purposes of this policy, a 'material failure to comply with Federal requirements' or a 'failure to meet other material legal requirements' must be intentional or grossly negligent."

To become familiar with the expectations around the responsible conduct of research, please review the guidelines for Research Ethics published by the Office of Research Integrity and Compliance.

This policy applies, in all respects, to this course.

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you have questions about this or your coursework, please let me know.

Every individual must be treated with respect. The ways we are diverse are many and are critical to excellence and an inclusive community. They include but are not limited to: race, color, national origin, sex, disability, age, sexual orientation, gender identity,

religion, creed, ancestry, belief, veteran status, or genetic information. We at CMU, will work to promote diversity, equity and inclusion because it is just and necessary for innovation. Therefore, while we are imperfect, we will work inside and outside of our classrooms, to increase our commitment to build and sustain a community that embraces these values.

It is the responsibility of each of us to create a safer and more inclusive environment. Bias incidents, whether intentional or unintentional in their occurrence, contribute to creating an unwelcoming environment for individuals and groups at the university. If you experience or observe unfair or hostile treatment on the basis of identity, we encourage you to speak out for justice and support in the moment and and/or share your experience anonymously using the following resources:
Center for Student Diversity and Inclusion: csdi@andrew.cmu.edu, (412) 268-2150, www.cmu.edu/student-diversity
Report-It **online anonymous reporting platform:** www.reportit.net username**:** *tartans* password**:** *plaid*

All reports will be acknowledged, documented and a determination will be made regarding a course of action." All experiences shared will be used to transform the campus climate.