Electrical & Computer ENGINEERING

Carnegie Mellon

Data, Interference, and Applied Machine Learning
Fall 2020

Course Number: 18-785

Instructor: Prof. Patrick McSharry **Email:** mcsharry@cmu.edu

Course Discipline: Electrical and Computer Engineering
Course Level: Graduate
Type of Course: Onsite - Remote
Course Streamed to Pittsburgh: Yes
Course Concentration: Applied Machine Learning
Core/Elective: Core

Units: 12

Lecture/Lab/Rep Hours/Week: 3 Lecture Hours Per Week; 3 Lab/Rep Hours Per Week

Semester/Year Offered: Fall, All Years

Pre-Requisites: None
Zoom details:
ID: *926 6279 5445*
Passcode: *469303*
Direct url: https://zoom.us/j/92662795445?pwd=ZkFuR2pWaHA4STdWcERpKzZPYVlDZz09

Class Schedule:
Pittsburgh: *Hamerschlag Hall 1107*
Kigali: *CMR C525*

Lectures: *1 hour and 20 minutes*

Week	A	Pittsburgh	Kigali	B	Pittsburgh	Kigali
1	Tues, Sep 01	08:00 ET	14:00 CAT	Thurs, Sep 03	08:00 ET	14:00 CAT
2	Tues, Sep 08	08:00 ET	14:00 CAT	Thurs, Sep 10	08:00 ET	14:00 CAT
3	Tues, Sep 15	08:00 ET	14:00 CAT	Thurs, Sep 17	08:00 ET	14:00 CAT
4	Tues, Sep 22	08:00 ET	14:00 CAT	Thurs, Sep 24	08:00 ET	14:00 CAT
5	Tues, Sep 29	08:00 ET	14:00 CAT	Thurs, Oct 01	08:00 ET	14:00 CAT
6	Tues, Oct 06	08:00 ET	14:00 CAT	Thurs, Oct 08	08:00 ET	14:00 CAT
7	Tues, Oct 27	08:00 ET	14:00 CAT	Thurs, Oct 29	08:00 ET	14:00 CAT
8	Tues, Nov 03	08:00 ET	15:00 CAT	Thurs, Nov 05	08:00 ET	15:00 CAT
9	Tues, Nov 10	08:00 ET	15:00 CAT	Thurs, Nov 12	08:00 ET	15:00 CAT
10	Tues, Nov 17	08:00 ET	15:00 CAT	Thurs, Nov 19	08:00 ET	15:00 CAT

11	Tues, Nov 24	08:00 ET	15:00 CAT	Friday, Nov 27	08:00 ET	15:00 CAT
12	Tues, Dec 01	08:00 ET	15:00 CAT	Thurs, Dec 03	08:00 ET	15:00 CAT

Teaching Assistants:

Name:	**Email:**	**Zoom Meeting Id:**
Christian Iraduku	Christianiradukunda2@gmail.com	https://cmu.zoom.us/my/christian.iradukunda
Eliphaz Niyoduse	eniyodus@andrew.cmu.edu	https://cmu.zoom.us/my/eliphaz
Tanya Akumu	takumu@andrew.cmu.edu	https://cmu.zoom.us/my/tanya.akumu
Isaac Coffie	coffie@andrew.cmu.edu	https://cmu.zoom.us/my/pages.coffie
Innocent Mukoki	imukoki@andrew.cmu.edu	https://cmu.zoom.us/my/imukoki
Akem Aristide Tanyi-Jong	aaristid@alumni.cmu.edu	https://cmu.zoom.us/my/akem.aristide

Course Management Assistant:
Email Address:
Office Location: HH 1112

Waitlist Manager: Megan Oliver
Email Address: mvoliver@andrew.cmu.edu
Office Location: HH 1112

Course Description:
This course will provide the methods and skills required for utilizing data and quantitative models to automate predictive analytics and make improved decisions. From descriptive statistics to data analysis to machine learning the course will demonstrate the process of collecting, cleaning, interpreting, transforming, exploring, analyzing and modeling data with the goal of extracting information, communicating insights and supporting decision-making. The advantages and disadvantages of linear, nonlinear, parametric, nonparametric and ensemble methods will be discussed while exploring the challenges of both supervised and unsupervised learning. The importance of quantifying uncertainty, statistical hypothesis testing and communicating confidence in model results will be emphasized. The advantages of using visualization techniques to explore the data and communicate the outcomes will be highlighted throughout. Applications will include visualization, clustering, ranking, pattern recognition, anomaly detection, data mining, classification, regression, forecasting and risk analysis. Participants will obtain hands-on experience during project assignments that utilize publicly available datasets and address practical challenges.

Learning Objectives:
The objective of this course is to provide students with an overview of the use and potential of data analysis and machine learning in research, business and government. For example, the task could be to seek an answer to a practical challenge, segment customers, understand client behavior or formulate a new strategy for optimizing key performance indicators (KPIs), by applying data analysis techniques to real-world datasets. Participants will learn how to plan, design and implement an empirical research project

using statistical, computational and quantitative techniques. They will learn how to test for statistically significant relationships, construct a variety of models, communicate the outcomes from complicated analyses and build decision support tools. Practical skills will be strengthened by discussing project design, data collection, data quality and techniques to account for uncertainty and cope with measurement errors, missing values and outliers. The course will combine theoretical aspects of data analysis with visual examples and demonstrations of how to construct and utilize statistical models obtained from machine learning approaches in practice. There will be a strong emphasis on highlighting the challenges of working with real-world data, avoiding over-fitting and the risks of relying on traditional assumptions.

Outcomes

After completing this course, students should be able to:

- Design an empirical project in response to a specific objective
- Identify and collect relevant data for undertaking the project
- Acquire data and organize it into a structured format
- Visualize data, identify key characteristics and present a summary
- Describe the advantages and disadvantages of different models
- Decide which models are likely to work best for a given application
- Undertake feature construction and feature selection
- Apply machine learning techniques for estimation and evaluation
- Select an optimal model using statistical approaches
- Produce diagnostic information for investigating model properties
- Understand model weaknesses and where assumptions could fail
- Communicate model output and conclusions to end-users

Grading

The grades for this course will be based on students' performance on seven homework assignments, a final exam and class participation. Homework assignments will be done individually and turned in via Canvas by the designated due date. Late work will be acceptable until 24 hours past the deadline, but it will lose 10%. The assignments will be graded based on both a writing report and code used to achieve results presented in the report. Class participation will be evaluated based on 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. 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%
Kahoot Quiz	2.5%
Piazza Participation	2.5%
Homework Assignment 1	5%
Homework Assignment 2	10%
Homework Assignment 3	10%
Homework Assignment 4	10%
Mid-Semester Exam (Multiple Choice)	10%
Homework Assignment 5	10%
Homework Assignment 6	12.5%

Homework Assignment 7	12.5%
Kaggle Project	Bonus 2.5%
Final Exam (Multiple Choice)	10%

Important Dates:

Description	Release Date	Due Date	Days	Grade Date
Assignment 1	Tue, Sep 01	Mon, Sep 07	7	Fri, Sep 11
Assignment 2	Tue, Sep 08	Mon, Sep 21	14	Fri, Sep 25
Assignment 3	Tue, Sep 22	Mon, Oct 05	14	Fri, Oct 09
Assignment 4	Tue, Oct 06	Mon, Oct 19	14	Fri, Oct 23
Mid-Semester Exam	Thu, Oct 22	Thu, Oct 22		Thu, Oct 22
Mid-Semester Grades				Mon, Oct 21
Assignment 5	Tue, Oct 27	Mon, Nov 09	14	Fri, Nov 13
Assignment 6	Tue, Nov 10	Mon, Nov 23	14	Fri, Nov 27
Assignment 7	Tue, Nov 24	Mon, Dec 07	14	Sun, Dec 11
Kaggle Competition	Tue, Nov 24	Fri, Dec 11	3	Sun, Dec 16
Final Exam	*Tue, Dec 10	*Tue, Dec 10		*Tue, Dec 10
Final Grades				Weds, Dec 18

* The exam dates are provisional.

Deadlines:
It is expected that deadlines are respected and met 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 later than 24 hours after the deadline, will be regarded as unclassified and result in zero marks. In the case of a situation preventing a student from meeting the deadline (such as a medical condition), a student is required to write to the professor and TAs of the course in advance, copy his/her advisor and submit supporting evidence.

MATLAB:
In this course, we will be using MATLAB software. Download MATLAB software for your computer operating system from http://www.cmu.edu/computing/software/all/matlab/download.html . After unzipping the file, read Matlab-Licence_Instructions.pdf file for instructions for connecting to MATLAB server and running Matlab.

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)

Data, Interference & Applied Machine Learning 18-785			
Week	Activity	Lecture A	Lecture B
1	**Theme**	**Measurement**	**Data Collection**
	Challenge	Statistics	Capturing data

	Discussion	Hurricanes	Aggregating data
	Case Study	MindLeaps	True colours – mood rating
	Analysis	Data types and uncertainty	APIs, classification, stability
	Demo	Matlab as a calculator	Loading in data
2	**Theme**	**Data Manipulation**	**Data Exploration**
	Challenge	Measuring human dev	Accessing data in real time
	Discussion	Manipulating data	Exploring data
	Case Study	World Bank Indicators	QUANDL
	Analysis	HDI, MDGs	Interactive analysis, APIs
	Demo	Manipulating	Graphs and visualization
3	**Theme**	**Descriptive Statistics**	**Distributions**
	Challenge	Data for a fact based world	Data visually accessible
	Discussion	Truth and statistics	How to visualize data
	Case Study	GapMinder	DataWheel
	Analysis	Using descriptive statistics	Visualization techniques
	Demo	Descriptive statistics	Statistical distributions
4	**Theme**	**Statistical Hypotheses**	**Quantifying Confidence**
	Challenge	Monitoring quality	Measuring intervention effect
	Discussion	Making decisions	Quantifying impact
	Case Study	Guinness	A/B testing
	Analysis	Statistical hypotheses	Quantifying confidence
	Demo	Techniques for testing	Techniques for confidence
5	**Theme**	**Trends**	**Decision Making**
	Challenge	Detecting Trends	Real time decision making
	Discussion	Trends and importance	Testing and going live
	Case Study	Trend following	Quantopian, Collective 2
	Analysis	Statistics for detecting trend	Autoregression
	Demo	Techniques for trends	Human activity
6	**Theme**	**Forecasting**	**Model evaluation**
	Challenge	Understanding the past	An appropriate model
	Discussion	Price discovery of lego	Model selection
	Case Study	Sales and marketing	Kaggle
	Analysis	Linear regression	Model evaluation
	Demo	Techniques for linear reg	Evaluation techniques

Topic Outline (Weeks 7-12)

Data, Interference & Applied Machine Learning 18-785			
Week	Activity	Lecture A	Lecture B
7	**Theme**	**Statistical Learning**	**Statistical Learning**
	Challenge	Machine Learning	Learning from data
	Discussion	Learning	Sports prediction
	Case Study	APGAR score	Rugby
	Analysis	Sitting rising test	Fitting models
	Demo	Computer evolution	Curve fitting
8	**Theme**	**Linear Models**	**Linear Models**
	Challenge	Prediction using linear regression	Models for probabilities
	Discussion	Regularization	Modeling probability
	Case Study	Prostate Cancer	Heart attack / Credit risk

	Analysis	Parameter estimation	Logistic regression
	Demo	Techniques for linear regression	Logistic classification
9	**Theme**	**Nonlinear Models**	**Nonlinear Models**
	Challenge	Modeling nonlinear relationships	Monitoring
	Discussion	What causes nonlinearity?	Nonlinear structures
	Case Study	Heart rate dynamics	Growth charts
	Analysis	Feature selection	Nonlinear regression
	Demo	Features and ApEn	Nonlinear modeling
10	**Theme**	**Supervised Learning**	**Supervised Learning**
	Challenge	Prediction and classification	Data-driven model structure
	Discussion	Bayes Theorem	Density estimation
	Case Study	Spam detection	Bimodal distribution
	Analysis	CART	Non-parametric models
	Demo	Constructing and evaluating trees	KNN
11	**Theme**	**Unsupervised Learning**	**Unsupervised Learning**
	Challenge	Multiple sources of information	Clustering and segmentation
	Discussion	Noise reduction	Eurovision
	Case Study	Fetal Electrocardiogram	Customer Segmentation
	Analysis	Signal separation	Dendrograms and K-means
	Demo	Dimensionality reduction	Techniques for clustering
12	**Theme**	**Ensemble Approaches**	**Ensemble Approaches**
	Challenge	Coping with uncertainty	Model averaging
	Discussion	Wisdom of the crowd	Many versus the best
	Case Study	Ensemble forecasts	Netflix competition
	Analysis	Bootstrap	Bagging and Random Forests
	Demo	Bootstrap techniques	Ensemble techniques

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%, 63%] D - :]63%, 60%] F : < 60 %

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 review fully and carefully 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.