Learning Outcomes

This course will explore approaches to extract insights from large-scale datasets. The course will cover the complete analytical funnel from data extraction and cleaning to data analysis and insights interpretation and visualization. The data analysis component will focus on techniques in both supervised and unsupervised learning to extract information from datasets. Topics will include clustering, classification, and regression techniques. Through homework assignments, a project, exams and in-class activities, students will practice working with these techniques and tools to extract relevant information from structured and unstructured data.

Extended Course Description

This course explores the application of data science techniques to unstructured, real-world datasets including social media and open data sources. The course will focus on techniques and approaches that allow the extraction of information relevant for experts and non-experts in a wide range of areas including smart cities, transportation or public safety.

This course will explore approaches to extract insights from large-scale datasets. The course will cover the complete analytical funnel from data extraction and cleaning to data analysis and insights interpretation and visualization. The data analysis component will focus on techniques in both supervised and unsupervised learning to extract information from datasets. Topics will include clustering, classification, and regression techniques. Through homework assignments, a project, exams and in-class activities, students will practice working with these techniques and tools to extract relevant information from structured and unstructured data.

After successfully completing this course you will be able to:

- Collect and clean large-scale datasets
- Articulate the math behind supervised and unsupervised techniques
- Execute supervised and unsupervised machine

Jennifer Proctor
jeproc@terpmail.umd.edu

Class Meets

Monday, Wednesday & Friday
10:00am – 10:50am
MCB #1207

Office Hours

HBK #4105J
Mon & Fri
Noon-1:00pm
and by appointment

Prerequisites

INST 314 - Statistics for Information Science

Course Communication

I will use ELMS announcements for time-sensitive communication.

To reach me, email is best. Please put "[INST 414] at the beginning of the subject line.

I can NOT guarantee prompt answers to questions sent on nights or weekends. If you have not received a reply within 2 business days, please message

Data Science Techniques

learning techniques
- Select and evaluate various types of machine learning techniques
- Explain the results coming out of the models
- Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach

Required Resources

Course website: ***elms.umd.edu***

The Data Science Handbook
Cady, Field
First edition (2017).
Available on Amazon

You need to bring a laptop to each class. If you do not have a laptop you can bring, contact the instructor before the first week of class ends. Over the course of the semester, you will need to install free software including ***Weka, Python, SublimeText, OpenRefine,*** and ***Tableau***. If you encounter problems with these tool, some of them will be available over the ischool Virtual Computing Lab as an alternative option you may use.

me again - I get a lot of messages and sometimes one gets lost in the shuffle.

To notify me you had an excused absence, email me as promptly as possible. Work missed for excused absences may be made up for full credit.

Campus Policies

It is our shared responsibility to know and abide by the University of Maryland's policies that relate to all courses, which include topics like:

- Academic integrity
- Student and instructor conduct
- Accessibility and accommodations
- Attendance and excused absences
- Grades and appeals
- Copyright and intellectual property

Please visit go.umd.edu/ug-policy for the Office of Undergraduate Studies' full list of campus-wide policies and follow up with me if you have questions.

Accessibility and Learning Support

Students with disabilities should inform me of their needs at the beginning of the semester. Please also contact the Accessibility and Disability Support Office (http://www.counseling.umd.edu/ADS/). ADS will make arrangements with the student and me to determine and implement appropriate academic accommodations. Inclusion is one of the

iSchool's core values, and we have attempted to make all materials and assignments accessible to people with varying abilities. However, if there is something else I can do to make the class more accessible please schedule a time to come talk to me. This will benefit not only yourself but also future students!

Writing and English Help

Even though all UMD students are either native English speakers or have passed the TOEFL, it is normal to struggle to grasp the tone and style expected in advanced professional writing and speaking. The best resource to help with this issue on campus at this time is the Undergraduate Writing Center which offers help with writing. Appointments can be made online at www.english.umd.edu/academics/writingcenter.

Student Health and Time Management

For your own health and sanity, please ***stick to the hours of daylight*** for as much of your work for this class as possible. If you have to turn in an assignment late, I would rather you go to sleep and submit it at lunch time the next day than stay up late and get it in at 3:30 am. Your grade is the same either way and YOU DESERVE TO SLEEP.

Furthermore, ***DO NOT COME TO CLASS IF YOU ARE ILL. This is VITAL because your instructor is unusually susceptible to illness.*** Additionally, y*ou will recover much faster if you get your rest* and you will *avoid exposing your classmates and teacher* to infection. As this class meets three times a week, you may miss two consecutive classes with a <u>self-signed sick notice emailed</u> to me within 48 hours of missing class. For further consecutive classes, a doctor's note will be required. The only exceptions are the ***<u>midterm and final presentation days which require a doctor's note regardless of duration of illness</u>***. There is no limit to the total number of absences that may be excused via official doctor's notes, allowing you to make up all missed work without late penalties.

Additionally, the entire course is published on Canvas already so that you can plan ahead and manage your time. We may need to adjust the schedule (postponing deadlines for snow days, for example) and if the majority of the class is struggling with a particular issue, I may assign an additional assignment on it, but this will be pretty rare. Please take the time to look over the course calendar. You may even want to use project management software or a digital calendar so you can set yourself reminders to start each of these assignments and reminders to submit them.

Course-Specific Policies

For this course, some of your assignments will be collected via Turnitin on our course ELMS page. I have chosen to use this tool because it can help you improve your scholarly writing and help me verify the integrity of student work. For information about Turnitin, how it works, and the feedback reports you may have access to, visit Turnitin Originality Checker for Student.

Get Some Help!

Taking personal responsibility for you own learning means acknowledging when your performance does not match your goals and doing something about it. I hope you will come talk to me so that I can help you find the right approach to success in this course, and I encourage you to visit tutoring.umd.edu to learn more about the wide range of campus resources available to you. In particular, everyone can use some help sharpen their communication skills (and improving their grade) by visiting ter.ps/writing and schedule an appointment with the campus Writing Center. Finally, if you just need someone to talk to, visit counseling.umd.edu.

Everything is free because you have already paid for it, and **everyone needs help**… all you have to do is ask for it.

Names/Pronouns and Self Identifications

The University of Maryland recognizes the importance of a diverse student body, and we are committed to fostering equitable classroom environments. I invite you, if you wish, to tell us how you want to be referred to both in terms of your name and your pronouns (he/him, she/her, they/them, etc.). The pronouns someone indicates are not necessarily indicative of their gender identity. Visit trans.umd.edu to learn more.

Additionally, how you identify in terms of your gender, race, class, sexuality, religion, and dis/ability, among all aspects of your identity, is your choice whether to disclose (e.g., should it come up in classroom conversation about our experiences and perspectives) and should be self-identified, not presumed or imposed. I will do my best to address and refer to all students accordingly, and I ask you to do the same for all of your fellow Terps.

Data Science Techniques

Grades

Grades are not given, but earned. Your grade is determined by your performance on the learning assessments in the course and is assigned individually (not curved). <u>If earning a particular grade is important to you, please speak with me at the beginning of the semester</u> so that I can offer some helpful suggestions for achieving your goal.

If you feel you are struggling at any point in the course, on an individual assignment or topic or with getting work completed in general, please TALK TO ME as soon as possible so we can come up with a plan to help you do better.

All scores as well as comments and feedback (except on Midterms and Final Presentations) will be posted on ELMS. If you would like to review any of your grades (including the exams), or have questions about how something was scored, please email me to schedule a time for us to meet in my office.

Unless you have reported an excused absence to me <u>in advance of the due date</u>, late work will be subject to a markdown of 10% per day late so please plan to have it submitted well before the scheduled deadline. I am happy to discuss any of your grades with you, and if I have made a mistake I will immediately correct it. Any formal grade disputes must be submitted in writing and within one week of receiving the grade.

Category	Assignment	Points
Quizzes	Syllabus Quiz	5
	"Race, Ethnicity and Criminal Justice" Reading Quiz	5
	Problem Set #1	10
	Problem Set #2	10
	Problem Set #3	10
Critical Reading Assignments	"Information Needs for Software Development Analytics"	10
	"The Midwest is Getting Drenched, And It's Causing Big Problems"	10
	"The President on Twitter: A Characterization Study of @realDonaldTrump"	10
Exercises	WHO Cleaning Homework	5
	Weka Activity Predictive Classifier & Evaluation	5
	Bayes Theorem Script	5

	Titanic Information Gain Problem	5
	Multilayer Perceptron Exercise in Weka	5
Portfolio Projects	Criminal Profiling Article	15
	Regular Expressions Exercise	5
	Web Scraping with Python	10
	Recommender System in Python	10
	TBA	10
Final Group Project	Problem Statement	10
	Data Collection, Formatting, and Cleaning	10
	Model	10
	Paper	10
	Poster and Presentation	10
	Peer Evaluation	25
Exams	Midterm	80
	Total	300

Final letter grades are assigned based on the percentage of total assessment points earned. To be fair to everyone I have to establish clear standards and apply them consistently, so please understand that being close to a cutoff is not the same as making the cut (89.99 ≠ 90.00). It would be unethical to make exceptions for some and not others.

Final Grade Cutoffs									
A+	97-100*	B+	87-89.99	C+	77-79.99	D+	67-69.99		
A	93-96.99	B	83-86.99	C	73-76.99	D	63-66.99	F	<60.0%
A-	90-92.99	B-	80-82.99	C-	70-72.99	D-	60-62.99		

* Note: To receive an A+ you must have demonstrated significant contributions to the class in addition to achieving this numeric grade.

Extra Credit

If you received a grade lower than an 80% on any assignment from the Portfolio Projects and Final Project Parts 1-3 ONLY, you may revise it based on feedback (and/or assistance from the instructor during office hours) and resubmit ONCE PER ASSIGNMENT provided it is submitted before the extra credit deadline listed on the syllabus. If the quality of the resubmitted work is

Data Science Techniques

good, your grade on that assignment can be raised as high as 90%. If the quality is not good, it is possible to LOSE points by resubmitting. No additional, unlisted assignments will be offered for extra credit purposes.

Course Schedule

Week	Date	Topic	Activities	Readings (Due BEFORE Class)	Homework (Due AFTER class)
1	1/28	Syllabus, Introductions	Student Introductions		Syllabus Quiz
	1/30	What is Data Science?	Skills search	DSH Intro and Chapter 1	
	2/1	The Data Science Process	Requirements Analysis	DSH Chapter 2	Install OpenRefine
2	2/4	Data Cleaning	Watch OpenRefine Tutorial	"Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes Magazine"	
	2/6	Data Cleaning	Cleaning the US Homicide Data set	DSH Chapter 4	Reading Assignment Worksheet #1
	2/8	Data Cleaning	Cleaning the US Homicide Data set		HW: Cleaning WHO Data
3	2/11	Cognitive Bias, Critical Thinking	Group-share Cognitive Biases	Links in Bias Module on ELMS	Reading Quiz (Open Book)
	2/13	Data Formats and Manipulation; Regular Expressions	Adding Structure to Unstructured Data; Regex101.com demo	DSH Chapters 4.6 & 7	Install Tableau
	2/15	Feature Engineering	Working with US Homicide Data set		Install Weka

4	2/18	Data Visualization and Storytelling	Exploring examples/Tableau Demo	Links on ELMS module and DSH Chapter 5	Reading Assignment Worksheet #2
	2/20	Data Visualization and Storytelling	Exploratory Data Analysis on US Homicide Data set		Regex HW
	2/22	Machine Learning and Artificial Intelligence	Exploratory Data Analysis on US Homicide Data set; Weka Tutorial Video	DSH Chapter 6	Problem Set #1
5	2/25	Machine Learning and Artificial Intelligence	Hands on Model Building with Weka	ELMS articles	
	2/27	Metrics, Overfitting	Weka accuracy video; Evaluating models	ELMS articles	Install Python & SublimeText
	3/1	Metrics, Overfitting	Evaluating models		Weka Predictor and Evaluation
6	3/4	Data Collection	Web Scraping with Python	DSH Chapter 3.3	
	3/6	Data Reuse and Data Management	Web Scraping with Python		
	3/8	Decision Trees, Information Gain	Web Scraping Q&A, Installing and Enabling Python Syntax Checkers	DSH Chapter 8	Criminal Profiling Article
7	3/11	Probability Review and Bayes Theorem	Web Scraping Troubleshooting	DSH Chapters 18 & 19 (Skim)	HW: Titanic Information Gain
	3/13	Bayes Theorem	Bayes Practice	"James May's Man Lab" - Bayes with Beer Video (link on Canvas)	Web Scraping with Python; Problem Set #2
	3/15	Final Projects - Review of Classifiers,	Final Project Group Sign-up		Bayes Theorem Script

	4/24	Text Analytics, NLP	Trump Tweets Analysis	DSH Chapter 16	
	4/26	Text Analytics, Word Vectors	Trump Tweets Analysis (Continued)		Final: Model Due
13	4/29	Network Analysis	Analyzing Social Media Networks	Excerpts from "Analyzing the Social Web" (on ELMS)	Reading Assignment Worksheet #3
	5/1	Network Analysis	Analyzing Social Media Networks		Final: Paper DUE
	5/3	Network Analysis	Analyzing Social Media Networks		All extra credit due - No exceptions!
14	5/6	Ethics Re-cap	Terrible Ideas in Data Science	Articles Assigned on Canvas	Problem Set #3
	5/8	Presentations		Final: Posters & Presentations - Day 1	Final: Posters
	5/10	Presentations		Final: Posters & Presentations - Day 2	
15	5/13	Presentations	Constructive Criticism Advice/ Peer Review Instruction	Final: Posters & Presentations - Day 3	Final Project Peer Evaluations DUE

Note: This is a tentative schedule, and subject to change as necessary – monitor the course ELMS page for current deadlines. In the unlikely event of a prolonged university closing, or an extended absence from the university, adjustments to the course schedule, deadlines, and assignments will be made based on the duration of the closing and the specific dates missed.