EECS 498-001, Fall 2018: Data Mining (4 cr.)

CS-Eng: ULCS / DS-Eng: Advanced Technical Elective, Application Elective

Instructor: Danai Koutra, Assistant Professor, EECS
(short OH after class + appt by request)
Teaching Assistant: TBD (check Canvas for OH)

The **office hours** of the teaching staff will be available on the calendar that is shared on Canvas. Before you head to the office hours, please check the calendar for exceptions.

Class Hours

- Thu 3-6pm @ G906 COOL
- **Discussion:** Friday, 11am-12pm @ 2246 SRB
  Friday, 1pm-2pm @ G906 COOL

Prerequisites

Prerequisites EECS 281 or graduate standing in CSE: You are required to have the background from a data structures course (e.g., lists, hash tables, arrays, search trees) and strong programming experience.

Advisory Prerequisites: MATH 214 or equivalent, STATS 250 or equivalent

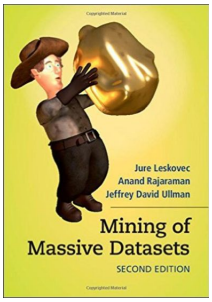
Course Description

Unprecedented amounts of data are being generated daily everywhere -- on the web, social networks, mobile apps, supermarket transactions, movie and music services, traffic sensors, smart home devices, healthcare, and more. Methods for extracting “nuggets of information from mountains of data” are transforming the world: data-driven approaches are changing the scientific and decision-making processes and solving various societal problems. This course covers the fundamental concepts in data mining, focuses on methods and algorithms and, at the same time, aims to equip the students with practical skills for mining of large-scale, real data. The topics that will be covered include big data systems, frequent itemsets, similarity and cluster analysis, mining of networks / time series / data streams, and applications, such as recommendation systems, social network analysis and web search.

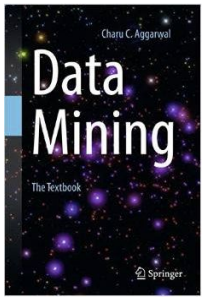
By the end of this course, the students will know:

- The various steps of the data mining process.
- The fundamental ideas behind data mining tools and methods.
- How to explore and analyze different types of data by applying the appropriate methods.
- How to use big data systems (e.g., Hadoop, Spark) for mining massive datasets.

Textbook (recommended)



For most topics, we will follow the book “Mining of Massive Datasets” (2nd edition), by Jure Leskovec, Anand Rajaraman and Jeff Ullman, Cambridge University Press, 2014. The book is available in pdf form (http://infolab.stanford.edu/~ullman/mmds/bookL.pdf) and has a supporting website (http://www.mmds.org/) with additional material.



For some lectures we will be following the book: “Data Mining: The Textbook”, by Charu C. Aggarwal, Springer, 2015 (online: http://rd.springer.com/book/10.1007/978-3-319-14142-8). UM is a subscribing institution (use VPN or download while on campus).

For topics that are not covered in these books, we will provide pointers to additional resources on Canvas.

Course Online Material and Posting Questions

- **Canvas**: We will use Canvas for posting announcements, assignments, lecture slides, updated schedule etc. Make sure you monitor Canvas at least twice a week. Website: https://canvas.umich.edu.
- **Piazza**: We will use Piazza for Q&A (i.e., homework, projects and technical questions). **Sign-up link**: Make sure to sign up, since you will not be automatically added to the course Piazza site: link TBD.
    - You should first consider posting your question there, AFTER **searching to see if a similar question has already been answered**.
    - You can ask a "**Private question**" that is only accessible to instructors or a question that is visible to all. **You should strictly refrain from sending private questions UNLESS** posting your question is going to be a violation of Honor Code (e.g., you are posting something from your solution). With non-private questions, other students will also benefit from your answers.
    - **Giving answers**: We encourage you to give answers to conceptual questions by other students (it counts towards your participation grade, 1% of the total grade). The teaching staff may "endorse" your answer or give their own answer. Of course, be careful of the Honor Code---you don't want to solve a homework problem for someone, only help them conceptually so that they are on the right track.
- **Personal requests and special accommodations**: For questions that are definitely not relevant to other students, send email to eecs498dm-f18@umich.edu. This will reach all the teaching staff. Please **do not send requests to our work emails**, as your email is likely to be overlooked (we do get a lot of emails) and you may get a very delayed response or even no response.
- **Regrade requests**: Send an email to eecs498dm-f18@umich.edu, **at the latest 5 days after graded exams or assignments** are distributed by the instructors or the TAs. Include all the relevant files and the **Regrade Request Form** (in ‘Files’ on Canvas).

Flux Accounts

Some projects require you to use Hadoop / Spark. We will provide instructions on obtaining a Flux account in the discussions.

Course Requirements and Grading

Midterm Exam on **Oct 25 (in class)**	**23%**
Final Exam **During finals week**	**30%**
Projects: 3 programming assignments, 10% each	**30%**
Homeworks: 4 written assignments, 3% each	**12%**
In-class Exercises	**4%**
Participation on Piazza	**1%**
Course Evaluation	**+1%**

The % are subject to slight adjustments based on the discretion of the instructor.

To get a C or higher, a passing performance is expected on both exams and all the projects. In general, this means that exams are also important.

Please make sure that you have no conflicts on the days of the midterm and final.

- **Homework & Projects:** Unless otherwise specified, the homework and projects must be completed individually. The homework is meant to familiarize you with fundamental concepts in data mining. The projects will require coding and applying the methodologies that you learned in practice.
- **In-class Exercises**: During some lectures (chosen at random and without notification) we will give out short problems to check your understanding. These will be graded for effort.
- **Participation on Piazza:** Constructive contributions to answering other students' questions on Piazza during the semester count as participation on Piazza.
- **Exams:** The examinations are **cumulative**. The final will cover all the topics, but will have slightly more emphasis on the topics covered after the midterm.

Late Days

For the homework and project submissions, check out the schedule on the Canvas. Typically, they will be due on Friday, 5:00pm.

For homework and projects, you can use up to 3 late days with 5% penalty for each 24-hour period (e.g., if the project is due on Friday, 5:00pm, a late submission is due before Monday, 5:00pm)**. You will have 10 minute grace period**, but beyond that the corresponding penalty will be applied. **Late days are rounded up to the nearest integer.** For example, 11 minutes late means 5% off, 1 day and 11 minutes late means 10%, 2 days and 11 minutes or 3 days late means 15% off. Submissions after the 3 late days will get a zero.

Please submit at least 30 minutes before the regular deadline as a safety measure. It is common to experience delays close to the submission deadlines (e.g., due to slower servers).

For extreme circumstances, like medical emergencies, no-penalty extensions will be granted. Email eecs498dm-f17@umich.edu with written documentation (e.g., doctor's note, email from the Office of Student Support and Accountability).

Honor Code

All students (including LS&A and Engineering) are required to observe the Engineering Honor Code in all assignments and exams. A copy of the honor code can be found at http://ossa.engin.umich.edu/honor-council/. The University takes honor code violations seriously, and penalties can be severe.

Any suspected violation of the honor policies appropriate to each piece of coursework will be reported to the Honor Council, and, if guilt is established, penalties may be imposed by the Honor Council and Faculty Committee on Discipline. Such penalties can include, but are not limited to, letter grade deductions or expulsion from the University. **If you have any questions about this course policy or you are not sure in any specific case, please consult the course instructor.**

Among other things, the Honor Code forbids plagiarism. To plagiarize is to use another person's ideas, writings, etc. as one's own, without crediting the other person. Thus, you must credit information obtained from other sources, including web sites, e-mail or other written communications, conversations, articles, books, etc.

Policy for homework: You may discuss the homework assignments with your fellow students at the conceptual level, but must complete all calculations and write-up, from scrap to final form, on your own. Verbatim copying of another student's work is forbidden. You may not consult homework solutions from a previous term (even for similar courses at UM) unless we provide a pointer in this class (no unfair advantage can be sought).

Policy for projects: All programming projects in this course are to be done on your own. Any violation will result in initiation of the formal procedures of the Honor Council. We will be using a sophisticated automated program to correlate projects. We do encourage students to help each other learn the course material. You may give or receive help on any of the concepts covered in lecture or discussion and on the syntax specifics of a programming language. You are allowed to consult with other students in the current class to help you understand the project specification.

However, you may not collaborate in any way when constructing your solution – the solution to the project must be generated by you working alone. You are not allowed to work out the programming details of the problems with anyone or to collaborate to the extent that your programs are identifiably similar. You are not allowed to look at or in any way derive advantage from the existence of project specifications or solutions of similar courses prepared in prior years (e.g., programs written by former students, solutions provided by instructors, project handouts).

If you have any questions as to what constitutes unacceptable collaboration, please talk to the instructor right away. You are expected to exercise reasonable precautions in protecting your own work. **Do not leave your program in a publicly accessible directory (e.g., github public repositories),** and take care when discarding printouts.

Handling Data with Integrity

You may not falsify or misrepresent methods, data, results, or conclusions, regardless of their source.

Unfair Advantage

You may not possess, look at, use, or in any way derive advantage from the solutions of homework, exams or papers prepared in prior years, whether these solutions were former students' work products or solutions that have been made available by University of Michigan faculty or on the internet, unless this section's faculty expressly allows the use of such materials.

Disabilities and Conflicts

If you have any disability as defined under the Americans with Disabilities Act that might interfere with your ability to participate in class, or to turn in assignments on time or in the form required, please contact your instructor and the Office of Students with Disabilities at the start of the term so that arrangements can be made to accommodate you.

Tentative Schedule

(subject to change)

Lect. # or Disc. #	Date	Topic
L1	Sep 6	Introduction & course logistics
L2		Data preparation (collection, storage, cleaning, …)
D1	Sep 7	Hands-on Hadoop / Spark HW1 out
L3	Sep 13	Big Data Systems (Hadoop, Spark, ...)
L4		Big Data Systems cnt'd
D2	Sep 14	Project 1 Intro; Hands-on Hadoop / Spark Project 1 out
L5	Sep 20	Frequent Itemsets & Association Rules
L6		Frequent Itemsets & Association Rules (cnt'd)
D3	Sep 21	Project 1 discussion & exercises HW1 due
L7	Sep 27	Similar Items, Locality Sensitive Hashing & Applications
L8		Similar Items, Locality Sensitive Hashing & Applications - Part 2
D4	Sep 28	HW1 solutions; Hands-on Exercises Project 1 due; HW2 out (Oct. 3)
L9	Oct 4	Similar Items, Locality Sensitive Hashing & Applications - Part 3
L10		Scalable Cluster Analysis
D5	Oct 5	Project 2 intro Project 2 out
L11	Oct 11	Cluster Analysis / Outlier Analysis
L12		Outlier Analysis
D6	Oct 12	Project 2 Q&A; Hands-on Exercises
L13	Oct 18	Midterm Exam Review HW2 due
		Q&A / OH for the midterm
D7	Oct 20	HW2 solutions; Midterm Exam Review
Midterm Exam	Oct 25	During class time
D8	Oct 26	Midterm solutions Project 2 due
L15	Nov 1	Classification (interpretable models: k-NN + decision trees)

L16		Classification (decision trees, perceptrons + scalable models: SVMs)	
D9	Nov 2	Review, exercises, OH	HW3 out
L17	Nov 8	Classification (SVMs) / Matrix Methods & Dimensionality Reduction	
L18		Matrix Methods & Dimensionality Reduction	
D10	Nov 9	Project 3 logistics	Project 3 out
L19	Nov 15	PCA + Network Analysis (Power laws)	
L20		Web Search and Link Analysis (PageRank / HITS)	
D11	Nov 16	Graph mining review; Project 3 Q&A	
--	Nov 22	Thanksgiving (no class)	
--	Nov 23	Thanksgiving break (no discussion)	
L21	Nov 29	Similarity in Graphs + Spam detection	HW4 out (Nov 29)
L22		Large-scale Graph Algorithms (community detection)	
D12	Nov 30	HW3 solutions; Project 3 intro	
L23	Dec 6	Time Series Analysis (models / classification / clustering)	
L24		Final Exam Review	
D13	Dec 7	Discussion: Final Exam Review	Project 3 due; HW4 due
Final exam	**TBD**	Location TBD	