



# CSCI 4502/5502

Data Mining - Fall 2023 - Lecture 1 - August 29, 2023

Ravi Starzl, PhD



# Welcome!

We're glad to have you here.



# I'm Ravi Starzl

- 1 BA Politics & Philosophy (Pitt)  
MS & PhD ML / AI (CMU)
- 2 Currently building novel forms of  
AI fusing logic and learning
- 3 Eclectic background, working with  
US Military, Academia, private firms
- 4 Lecturer at CU Boulder, Comp Sci  
Adj Professor at Carnegie Mellon





# Contact

- 1 Instructor:  
[ravi.starzl@colorado.edu](mailto:ravi.starzl@colorado.edu)
- 2 TAs:  
Samuel Pugh <[samuel.pugh@colorado.edu](mailto:samuel.pugh@colorado.edu)>  
Chintan Modi <[chmo5597@colorado.edu](mailto:chmo5597@colorado.edu)>  
Lakshmishree Chandra <[lach1767@colorado.edu](mailto:lach1767@colorado.edu)>
- 3 Graders:  
TBD





# Administrative Information



# CSCI 4502/5502 Data Mining Fall '23

- All sections share the same schedule
- All meetings open to all sections
- All office hours available to all sections
- Pre-recorded lectures will usually be made available in advance of the class. When that is the case, please watch those lectures at your convenience asap.
- Class meeting times will be used for clarifying questions, reviewing useful information, explaining use case scenarios, and working through interview questions.
- Instructor will conduct a majority of these class sessions. TA and Guests may conduct some of these class sessions too.



# CSCI 4502/5502 Data Mining Summer '23

- Class meetings are every T/Th from 12:30 - 1:45pm MT by zoom at:
- Course canvas is available at:  
<https://canvas.colorado.edu/courses/97317>  
For enrolled students only
- Piazza is available for discussions and questions (through canvas)
- Textbook: Jiawei Han, Micheline Kamber, Jian Pei. 3rd Edition, Morgan Kaufmann, 2012.
- Check canvas, particularly the syllabus and schedule, for latest course information.



# Course Overview



# Why Data Mining?

- Data, lots of data, and fast increasing
  - Discover interesting patterns from data
  - Example 1: Market basket analysis: Walmart, Amazon, NetFlix, ...
  - Example 2: Cluster analysis, classification: loan application, medical diagnosis, ...
  - Example 3: Time series analysis: social network, wind speed, stock, ...
  - And a lot more!





# Course Summary

- Data mining
  - concepts and techniques; quality vs. efficiency
  - discovering interesting patterns from large amounts of data
- Topics covered
  - data preprocessing, data warehouse, frequent patterns, classification, clustering, outliers
  - complex data, data mining trends





# Policies

- Please keep your computer on mute during course sessions, unless called upon by an instructor to speak. With a large group, it can get noisy and chaotic quickly.
- Course staff will endeavor to get a response to your questions within 24 hours of receipt. Please feel free to send a follow up if you have not heard back within 48 hours.
- Please be respectful, kind, and give the benefit of the doubt to all your peers, including your classmates and course staff.
- Please review the CU Policies - links on course canvas.





# Academic Integrity

- All work must be completed alone unless otherwise directed
- All submitted work should include the honor code pledge, either in writing or in commented code  
<https://www.colorado.edu/sccr/honor-code>
- Properly acknowledge other people's work, including information you find on the Web
- Cheating or plagiarism will NOT be tolerated!





# Grading

- Homework assignments (20%) (work alone)
- Midterm exam (20%) (work alone)
- Course project (40%)
- Final Exam (20%) (work alone)
- Late submission
  - at most 2-days, 20-point penalty each day





# Course Project

- A self-contained project related to this course's topics
- Team of 3-4 students
  - check with instructor for smaller or larger groups
  - can mix students in different sections
- Pick your own project idea
- Discuss project ideas with instructor, TA, and others
- Add this to your professional portfolio



# Project Proposal

- Submit a project proposal (~3 pages)
  - Motivation: why this problem?
  - Literature Survey: what has been done before?
  - Proposed Work: what do you plan to do?
  - Evaluation: what metrics? how to claim success?
  - Milestones: when to accomplish what?
- Project requires both Report & Presentation
- Proposal due September 26



# Project Checkpoint

- Submit a progress report (~6 pages)
  - updated, extended version of initial proposal, highlight progresses
  - Proposal Review: motivation, proposed work, evaluation, milestones
  - what have been accomplished so far?
  - what remains to be done?
- Checkpoint due October 31



# Project Final Report

- Follow the format of regular research papers (~10 pages)
  - title, authors' information, abstract
  - introduction, related work
  - main technique, evaluation
  - conclusions, future work, references
- Final report due December 7
- Selected projects showcased December 14



# Project Final Presentation

- A 10-minute presentation
  - motivation, literature survey, your work, evaluation, conclusions, future work
  - technical depth, evaluations, clarity, style
- Also submit source code & key results
- Enumerate contributions by individual team members



# REMINDERS



- Check the course canvas regularly
- Check the syllabus and schedule regularly
- Read chapter 1
- Watch for additional resources we give for each section
- Start thinking about your course project



# Thank you

A special thank you to Qin Lv for her slides,  
on which this lecture is based