Final Project

Your goal in this project is to use what you've learned in class to address an interesting real-world research problem. This project is meant to be a substantial **independent** research effort involving a topic of your choice that will give you some experience with doing original research in data mining and writing up your results in a conference paper format. You will formulate an idea/task that you will describe clearly, relate to existing work, implement, and test on a dataset. This will involve writing code, using your code to run experiments on a dataset, making figures to communicate your data/methods/results, reading some background papers, and writing a report about your task, the algorithm(s) you used, and the results you obtained.

Your research projects may be one of several types of projects, for example:

- Design a new algorithm or modify an existing algorithm to improve performance on some task or dataset (must compare to at least one existing baseline algorithm)
- Empirical evaluation of multiple algorithms for an interesting real-world data mining problem
- Come up with a solution to solve an open real-world problem or research challenge
- Implement a data mining method for an existing task/dataset that has not been tried for that problem before

Whatever you choose to work on, it should be interesting. Perhaps the problem is novel or your approach is novel. Perhaps it is a topic you are highly interested in. You should make sure that your analysis is not trivial. For example, spending an hour applying a method in Scikit-learn to a dataset and then doing a quick write-up would be considered trivial. That said, you do not want to make your problem so complex that you do not have time to finish it. Keep in mind that you will have about 6 weeks (October 26 to December 9) to work on your project.

Ground rules

You may not work in groups. All projects should be completed individually.

You may work with an advisor (e.g., a professor or senior graduate student) to come up with an interesting topic, for example that contributes to a larger research project. However, the work must be done entirely on your own.

You may work on a topic that is useful for you in the future (e.g., related to your thesis). However, you will be graded on new work that you do as part of this class project (i.e., starting October 26) and **you may not** re-use work that you have done for other classes or projects.

You may end up using the same dataset or working on the same or a similar problem as another student. This is inevitable in a large class, but you must work on your projects independently.

You may discuss your project with other students, but **you may not** share written work, code, or other details of your implementation.

You may use any third-party libraries or code as long as it is publicly available and you reference it appropriately.

You may use algorithms, techniques, libraries, etc. that were not covered in class (e.g., self-organizing maps for clustering). However, your chosen methods must be related to the course topics.

You must properly provide references to any work that is not your own.

Although your final report will be in a conference paper format (specifically for ACM conferences like KDD), there is no expectation whatsoever that your final report will be a publishable paper. That said, some students may produce research that results in publishable papers, which will be pretty cool!

Deliverables

There are four deliverables for the final project with requirements outlined in the table below. The expectations for each deliverable are described under the table.

Deliverable	Format	Expected length	Due date
Proposal / plan	PDF	1-2 pages	November 4, 2022
Progress report	PDF (using ACM template, available in Word or LaTeX)	2-3 pages	November 21, 2022
Written report	PDF (using ACM template, available in Word or LaTeX)	6 pages	December 9, 2022
Presentation	Slides (PPT, Google Slides, PDF) **and** recording (mp4)	5 minutes	December 9, 2022

Proposal/project plan

The proposal should consist of 1-2 pages describing the problem you plan to solve, outlining how you plan to solve it (i.e., methods/algorithms, datasets), and describing what experiments and results you will include in your final report/presentation. A proposal template is provided for you to fill in; you are required to use this template. The proposal is due by November 4 but you are encouraged to submit the proposal earlier if possible; proposals will be reviewed as soon as

possible after submission so this may allow you to get early feedback on your idea and scope of planned work.

Written report
You will prepare a report on your work in the style of a conference paper. The maximum length is 6 pages (not including references or appendices).

You must use the template for conference papers submitted to the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) conference. Word and LaTeX templates can be found here. Your report should introduce and motivate the problem your project addresses, describe related work in the area, discuss the elements of your solution, and present results that quantitatively (and perhaps qualitatively) measure the performance of your solution (with comparisons to other methods as appropriate). Think critically about the metrics you choose to measure performance; you should choose (or create) metrics that are appropriate for your problem and dataset. You must include the following section headings in your report: Abstract, Introduction, Related Work, Methods, Results, Discussion, and Conclusion; you may choose to include additional section or subsection headings as appropriate.

- **Progress report:** The progress report should be an intermediate draft of your final written report. This will use the same format as the final report and should contain the same section headings as your final report. The progress report is meant to demonstrate the progress you have made towards your final goals by about halfway through the project period and outline the work you still have to complete. It is expected that your Abstract, Introduction, and Related Work sections should be mostly complete and the subsequent sections should have initial details of your methods, experiments, and results.

Presentation
You will record a 5 minute presentation that visually presents the work you did for your final project. You will submit a recorded video (mp4 format) and the slides you used to present (powerpoint, Keynote, PDF, or Google Slides format). Your presentation should concisely describe the problem, related work, your approach, and your key results. When grading the presentations we will stop watching after 5 minutes, so do not exceed this time! Think of this as a "lightning talk" that presents the key highlights of your work to convince someone (e.g., your classmates or a future advisor) to read your written report. You do not need to (and should not) include every detail of your project in the presentation—this is what the report is for. The presentation will be shared on Canvas with all of your classmates and it should distill the highlights and key findings of your work. You can use any software to record your presentation, as long as the video and audio are clear; some options are Zoom, Powerpoint, or Quicktime.

Grading

The final project will make up 40% of your grade in this class. Please take this project seriously and do not procrastinate! The breakdown of this 40% is as follows:

- Final project proposal (5%)
- Final project progress report (5%)
- Final project presentation (15%)
- Final project written report (15%)

Grading for all deliverables will be based on the following criteria (adapted from Stanford CS221):

Criterion	Exceeds expectations	Meets expectations	Progressing	Below expectations	Weight
Justification	A really interesting problem	A justified problem	Superficial justification	No explicit justification	10%
Real world problem	Novel formulation of algorithm, task, or dataset	Clear and reasonable formulation of existing problem	Unclear or unreasonable formulation of problem	Unclear and unreasonable formulation of problem	15%
Algorithm choice and implementation	Implements new or challenging algorithm	Choose and run a reasonable algorithm	Chooses an algorithm but unable to get it working	Chooses a non-applicable or unsuitable algorithm	30%
Evaluation and results	Clever analysis (e.g., domain- specific metric or uncertainty analysis)	Show numerical results that measure how well your approach worked	Numeric results that do not sufficiently measure how well your approach worked, or only qualitative results	Incorrect evaluation or results	30%
Contextualize with related work and future work	Contextualizes idea with existing research and proposals moving directions for future work	Any reasonable contextualization with prior work and directions for future work	Does not contextualize with prior work or propose future work directions in relation to results	Not explicitly stated	15%

Note: You can earn up to 5 points extra credit or lose up to 5 points based on writing quality, neatness, and visual appearance. For example, if you exceed expectations in all criteria for your presentation but your slides are messy and your presentation runs over 5 minutes, you may lose points. Conversely, if the content of your written report is below expectations but it is eloquently written, reads easily, is well formatted, and has nice figures, you may gain extra credit on your total points.

Project ideas

Many of you likely already have some ideas about what problem you would like to address in your project. If you are not sure or would like to discuss some preliminary ideas, you may come to TA or instructor office hours to discuss or contact the TA or instructor. You may also browse some of the following sources to get ideas for project topics:

- Example projects for Stanford CS221 - scroll down to Project Ideas section here
- Sample projects from Fordham University CISC 4631 - starts on page 4 here

- Conference papers from Proceedings of data mining conferences such as KDD, AAAI, Canadian AI Conference
- Search Zenodo for datasets with keywords of your interest, e.g.:
    - Mars dust images dataset
    - Martian frost image dataset
    - Mars fresh impacts images dataset (published by guest speaker Kiri Wagstaff)
    - Mars rover surface images dataset
    - Mars orbital images dataset
    - Conference abstracts from Lunar and Planetary Science Conference
    - Folktables - datasets derived from US Census
    - Galaxy classification dataset
- Search Google Datasets Search for keywords of your interest, e.g.:
    - Sentiment analysis of movie reviews search results
    - Climate change search results
    - Malaria search results
    - COVID-19 search results
- Kaggle datasets and competitions
- Radiant Earth ML Hub datasets and competitions
- Data.gov - US Government's open data site
- Datahub.io
- NASA Earth Data portal
- NASA Planetary Data System
- CERN Open Datasets
- World Health Organization Global Health Observatory data repository
- UN Food and Agriculture Organization data repository
- British Film Industry data
- FBI Crime Data Explorer
- City or state open data repositories; search "<city> open data" in your browser for your city of interest, e.g.:
    - Tempe data catalog and on data.gov
    - Phoenix data catalog
    - Denver data catalog
    - Maui County public information
    - Washington DC open data catalog