

Project

Final Report Due: Wednesday, November 26
Turn in report by 11:59pm (through Canvas).

1 Overview

Your project will consist of five elements.

- Project Proposal : Due September 10
- Data Collection Report : Due September 24
- Intermediate Report : Due October 29
- Final Report : Due November 26
- Poster Outline: Due December 4
- Poster Presentation : December 12 | (1:00pm - 3:00pm)

As in any research in order to get people to pay attention, you will need to be able to present your work efficiently in written and oral form.

You may work in teams of 2 or 3, but the amount of work you perform will need to scale accordingly. Teams of size 1 might be allowed under unusual circumstances with special permission from the instructor. All students will need to have clearly defined roles as demonstrated in the final report and presentation. I highly recommend groups of size 3. Although the project work will scale with students, the administrative parts will remain constant, so having a large group will make it easier for you.

Note that some topics will not be covered before many elements of the project are due. I realize this is not ideal. However, typically, most work on a project is crammed in the last week or two of the semester, which is also not ideal. In the past this has lead to much stronger projects without considerably more work required.

1.1 Scale of Project

The specifics of the project will be very flexible. I expect each student to perform data mining on some real data set. The goal is to gain more in depth experience in some aspect of the class, and to do so in a setting where the instructor can provide guidance.

The lectures will cover the material, the homeworks and quizzes will reinforce the material, and this project should provide deep understanding of some aspect of the material. Students who demonstrate a deep understanding of some aspect of data mining covered in the class through their project will receive high grades.

Project options may include:

- apply advanced techniques from the class towards a real data set
- compare several basic techniques from the class towards a real data set
- propose and test extension to techniques from class on a real data set

2 Project Proposal (5 points)

Due September 10

Prepare a **100 to 200 word** document outlining your plan. This should contain:

1. who is in your group

2. what data you plan to use and where you plan to get it from
3. what structure you want to mine from the data
4. why this problem is interesting
5. what is new, or what I (the instructor) will learn

The last bullet should try to make clear why this is *more* than just another homework assignment. Either you are applying data mining techniques to a specific data set in a way it has not (to your knowledge) been done before; you are comparing several techniques when you are not sure which is best, or what advantages each has; you are proposing some new variation to some existing idea.

It is not uncommon for the instructor will provide feedback and alter or modify your proposed plans. This can either happen by students stopping by to discuss with the instructor or TA before the proposal is due, or will come in feedback on the specific proposal. This step is most important when the topic is related to material that is covered later in the class.

To be clear, if there are multiple students on a project team, the project should scale accordingly. This may for instance involve trying three extensions to a technique and comparing to see which works the best (for 3 students). Or processing a data set in two phases using two consecutive techniques (for 2 students). The work does not need to be split as such, but it should be clearly more than for a single student.

If this is part of a larger project (e.g. part of your thesis work, extending your lab's work), please indicate its scope with respect to previous work.

As an arbitrary way to separate this class from the Machine Learning class, we will **not allow projects with central focus on building a classifier**. A classifier may be used to evaluate a result.

2.1 Topics

The best projects occur on topics that students have experience with and/or are passionate about. First look within your own experience, whether that is your research, your work, or your hobbies.

If you still have trouble finding a topic, I suggest exploring these resources for interesting datasets to explore. You cannot just explore a data set, you must propose to find some structure in it.

- <http://snap.stanford.edu/data/>
- <http://www.census.gov/>
- http://meta.wikimedia.org/wiki/Data_dumps
- <https://datasetsearch.research.google.com>
- <https://datacommons.org>
- <http://kdd.ics.uci.edu/>
- <http://www.cise.ufl.edu/research/sparse/matrices/>
- <https://github.com/huggingface/datasets>
- <https://aws.amazon.com/datasets>
- <https://dataforgood.facebook.com/dfg/tools>
- <https://docs.x.com/overview>
- <http://lib.stat.cmu.edu/datasets/>
- <https://nlp.stanford.edu/projects/glove/>

If you have a research advisor, they may also be good source of problems and data.

3 Data Collection Report (10 points)

Due September 24

Prepare an **at most 1 page** report describing what data you have collected (or for some larger settings like from Twitter) are continuing to collect.

Often a surprising amount of time in data mining is spent in acquiring and manipulating data. And demonstrating a technique on a real data set can have dramatic impact in convincing someone of its usefulness.

Please report:

1. How you obtained your data?
2. How large is your data? I mostly want number of data points (and dimensions) if relevant. This is more informative than just number of MB or GB.
3. In what format are you storing your data. Describe the *abstract data type*, not just the file format.
4. Did you need to process the original data to get it into an easier, more compressed format (e.g., convert from one format to another one)?
5. How would you simulate similar data?

For this report, I really want you to focus on which of the abstract data types (e.g., matrix, graph, point set, set of sets, or a combination of these) that you will think of this data set as. In many cases you will not store the data in the original format, and you will need to convert it to another format. I want you to think about this conversion along the idea of these abstract representations. This is usually the most challenging aspect for students. These decisions should be discussed to answer steps 2 and 3 (above) fully.

For instance, do not say data is a vector (which implies Euclidian distance) if the coordinates have different units, and are not compatible, unless the goal is only regression. Do not have each entry be a piece of text; it can be a count of the number of a word or set of words, but not a piece of text. This is usually not useful. Same goes for a matrix representation with different meanings for each column. Think carefully about this, and ask the instructor if you are not sure.

Step 5 is to make you think about how you would *model* your data. The structure you hope to find is likely correlated with how you model your data. Given a model, you can (at least theoretically) then generate random data from this model. Sometimes this is actually a hard open problem (e.g. for graphs of social networks). This is important for most statistical analysis of your data to validate how well your technique will actually work. And will also be important if you want to generate synthetic data to see how your technique scales beyond the real data set you have gathered.

4 Intermediate Report (15 points)

Due October 29 at noon.

*Unlike other components, this one will be **due at noon**. Then each project team will be assigned 2 other project teams. In class on October 29, we will spend half the class time with project teams meeting and peer evaluating each other's projects. Teams will have an option to update their reports before evaluated by the instructor. Under rare cases, remote students will be able to discuss over Zoom or other means if preferred.*

Prepare an **at most 1 page per student** report describing your progress so far towards your proposed goal. That is 2 students can turn in 2 pages, and 3 students can turn in 3 pages.

Basically this intermediate report should demonstrate to the instructor that you have made non-trivial progress towards your goal. For instance, perhaps you proposed to take a basic technique and slightly modify it. Then this intermediate report should note that you have gotten the basic technique working.

Some basic plots or numbers from experiments that you ran (mainly to convince yourself) that shows everything is working can be included. Perhaps you have finished all of the coding and setup already and

just need to run experiments at this point. Then note that and discuss what suite of experiments you plan to run for the final report.

Specifically describe:

1. What progress you have made towards your proposed goal? (just data collection is *not* an option)
2. If you tried some basic approaches: what worked well and what did not?
3. What could be done to improve the basic approaches?
4. What experiments have you run and are you planning to run to demonstrate the effectiveness?

Each student group should schedule a time (using this https://docs.google.com/spreadsheets/d/1z5W58QiHjhCtSAqomPxoTCKYHYP_7SALODbaV6356rg/edit?gid=0#gid=0) to meet with the instructor for **10 minutes** the week the intermediate report is due on specific times **Wed Oct 29 : 3-5pm** or **Thur Oct 30 : 9-9:20am** (costumes optional). Bring a **print out** of the intermediate report to this meeting — it can be updated after the peer review.

Be prepared for a short **10 minute** meeting where you describe your project goal and progress for about 5 minutes, and the instructor will attempt to provide feedback to make the final project and reports as strong as possible. Thus the more progress you have made and the more information you include, the more success you will likely have with your final report.

5 Final Report (50 points)

Due November 26

Your report will be **at most 4 pages per student**, single columned at 11 point or larger font. Margins should be at least 1 inch. (If it is shorter, but contains all needed information, it will make the instructor happy. He gives higher grades if he is happy.) However, you will be allowed an unlimited number of pages for references and appendices. The report will be graded on the first four pages (per student), but additional information to support the first four pages may be appended and referred to. The instructor will only read the appendix at his discretion. Putting *all* charts and figures in the appendix is not acceptable, and this will result in lost points.

If you work in a group of more than 1 student, then it should be painfully clear what part of the project was worked on by each student. If the distribution of work was complicated, a table in the appendix may be a good idea.

Why only 4 pages? A key aspect of scientific writing is efficiently conveying information. I expect students to easily generate more than 4 pages of information, but you need to convey this information to me efficiently. What are the key ideas? What are the key experiments to show me plots for? What is the relevant related work to highlight?

5.1 Content

1. Provide a succinct title for your project (and your names!). This is part of 4 pages/student. Your poster must have the same title.
2. Explain the problem and motivation. If you prepared a thorough proposal and intermediate report, then you may be able to borrow some material from there.
3. Explain what data you explored? Where did it come from, how did you process it? If you simulated to scale the experiments, how did this work? If your data collection report was thorough, you can likely reuse much of this material.

4. What is the **key idea** your project is built upon? If there is no interesting ideas, I will be a little disappointed. This should describe the rationale behind and what you were hoping to discover about the different approaches you are comparing, or what extension you are proposing to an existing technique, or how you are applying a technique to a dataset where it has not been explored before.

State this clearly in the beginning; try to make me excited to read the remainder of your report to find out how your idea played out!

5. Explain what you did. Did you prove something? Did you implement something? Did you compare several things? Did you extend something?
6. Explain what you learned. This is often greatly aided through charts of experiments. But you should also include what lessons you came away with in words; just charts or mathematics is insufficient.
7. An ethics evaluation of your project. This should be in a separate section that is clearly labeled. It should range from 1/2 page to 2 pages of the report.

What should be in the Ethics Evaluation? In this section, you should honestly evaluate if there could potentially be any negative societal impacts, including possible malicious or unintended uses. This could include privacy violations, security concerns, fairness issues, disinformation potential. Basically, could this work if released or taken further cause “harms” to groups or individuals?

There are no penalties for work that you now assess as having potential ethical issues or harms. This class is a sandbox to explore data mining, but this work could have consequences. This is a chance to reflect on this separate from that. This section should include your honest assessment.

If you do assess that your project may cause harms or lead to negative societal impacts, you may discuss if this is balanced by beneficial impacts. For many impactful products, there is some negative with the positive, and we often produce and release products where we feel the good outweighs the bad.

You can also discuss potential mitigation strategies (careful release, privacy/anonymization, careful monitoring, modification before release) that you feel may or may not be appropriate.

6 Poster Presentations (20 points)

Presented December 12 at 1pm - 3pm (in MEB hallway near KSoC front office)

The class will have a poster session where each group will present their projects. This will be open to a general audience. Each group is responsible for one poster. Posters will be available by around noon in the KSoC main office. [Here](#), we provided some guidelines for making posters.

6.1 Content

I expect to see four things on your poster:

1. What is the problem and data you worked on?
2. What were the key ideas in your approach?
3. What techniques from the class did you use?
4. What did you learn?

This is a great opportunity for the class to learn about a large variety of topics. If you approach this poster presentation as a teaching experience, you will be more likely to succeed.

There are two general types of posters: (A) ones that are meant to stand alone on a wall, with no interaction. (B) ones that are meant to serve as guides for a personal discussion on the topic. **Posters for the class should be type (B)**, since the main purpose will be a poster session with many people. As such, design your poster with this in mind: Add any image that you would want to draw on the board or have on a slide if you

are giving a presentation. Posters with just text are typically not very useful. All text should be large enough so that if you printed the entire poster on one piece of paper (standard notebook size) that it would be easy to read. (Moreover, outlining a poster on a single piece of paper is an excellent way to start organizing your thoughts.) Make the posters eye-catching, so that people from a distance want to come over and find out more.

Illustrations, pictures, and examples are the key reason to make a poster. Build your poster around these!

6.2 Making Posters

At the poster session there will be easels, foam boards, and clips that can support posters.

The TAs will print the poster for you. You do not need to print the poster.

You can use applications like Powerpoint, Keynote, LucidChart, Omnigraffle, Illustrator, or IPE to create a single large slide. The pdf file should be called group#.pdf (e.g., “group07.pdf” for the group number 7. You can see your group numbers in Canvas.) and the TAs (through the department) will print posters for you. This should still have a clear title and narrative flow. Other than including the key points, GO WILD, and do whatever you can imagine!

After your poster outline has been approved by the instructor or TA (see below) and you are satisfied, to print the poster, add note on canvas saying this is the FINAL version. Once you do this, you may not be allowed to make changes. This final version must be made by **TBA (probably noon on December 9)**, or it might not be printed (unless you have been granted an extension for extra revisions).

Your poster is not complete until it has been approved by me/TA Elisa! Make sure the poster is in pdf format and is 3x2 feet, otherwise it will not be printed. Early submissions are greatly appreciated.

6.3 Poster Outline : Due Dec 4

A poster outline will need to be turned in (via canvas) **by December 4 at 11:59pm**. This is to be a first full draft; but you can continue tweaking after this step. Successful completion of this component is worth **5 points** towards the poster presentation. Early submissions are much appreciated.

You are also encouraged to stop by during office hours to discuss with the instructor or TA Elise. The outline should include a draft of the poster (either as a set of slides) or as a single document. It should be near complete, so that the instructor/TA can make concrete suggestions to improve the poster.

6.4 Poster Session

Each student will be assigned two posters to write a small (250 words) summary on. During the poster session each student will need to visit their assigned posters and learn about it enough to write the small summary.

All students will vote one of their top 2 posters (judged by project accomplishments and quality of poster presentation). They cannot vote for their own. The top vote getter will earn a bonus 5 points, and the second top vote getter will earn a bonus 2 points. The past winners had very creative designs with lots of pictures or examples, *but still had all of the important points above*.

Also the instructor will visit each poster and ask for a **3-5 minute** overview of the project. Prepare and practice this *short* oral presentation.