

Project*

Final Report Due: Monday, April 8
Turn in report at the start of class.

1 Overview

Your project will consist of five elements.

- Project Proposal : Due February 4
- Data Collection Report : Due February 25
- Intermediate Report : Due March 20
- Final Report : Due April 8
- Poster Presentation : April 24

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 will be allowed with special permission from the instructor. All students will need to have clearly defined roles as demonstrated in the final report and presentation.

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 led 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 homework 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 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 February 4

Prepare an at most **1 page** document detailing your plan. It can be less than 1 page, usually half a page will suffice.

This does not need to be too detailed, but needs to 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 quite likely 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 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 needs to be clear it is more substantial than that for a single student.

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

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.

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://data.geocomm.com/catalog/>
- http://meta.wikimedia.org/wiki/Data_dumps
- <http://ngrams.googlelabs.com/datasets>
- <http://kdd.ics.uci.edu/>
- <http://www.cs.utah.edu/~lifeifei/datasets.html>
- <http://www.cise.ufl.edu/research/sparse/matrices/>
- <http://webscope.sandbox.yahoo.com/>
- <http://www.google.com/publicdata/directory>
- <http://www.infochimps.com/tags/twitter>
- <http://lib.stat.cmu.edu/datasets/>

If you have an advisor, they may also be good sources of problems and data.

3 Data Collection Report (10 points)

Due February 25

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?
3. In what format are you storing your data (be precise)?
4. Did you need to process the original data to get it into an easier, more compressed format?
5. How would you simulate similar data?

In many cases you will not store the data in the original format. That is you will need to process it to be in some abstract representation (e.g., a matrix, a graph, or a point set). This is usually the most challenging aspect for students. These decisions should be discussed to answer steps 2 and 3 (above) fully.

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. This 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 March 20

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?

The instructor will attempt to provide feedback to make the final 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 April 8

Your report will be **at most 4 pages per student**, single columned at 11 point or larger font. 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.

If you work in a group of more than 1 student, then your report will be 4 pages per student, plus unlimited space for references and appendices. 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. 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.
2. 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.
3. 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!

4. Explain what you did. Did you prove something? Did you implement something? Did you compare several things? Did you extend something?
5. 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.

5.2 Feedback

If possible, the instructor will return feedback on reports before the end of the semester. If you lost points in the section, then **one half of the points** can be made up by turning in a revised version.

6 Poster Presentations (20 points)

Presented April 24 **MEB 3105 is booked from 4pm to 7pm**

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.

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 wall space and tape, and there will also be easels with foam board that can support posters. I suggest two approaches to actually make the posters:

beginner — slides: Make 8 slides in PowerPoint (or equivalent) and post print-outs of the slides. Make sure you use large enough font that it is easy to read. The slides will help you structure your “poster” linearly to tell a story and help you modularize the presentation (both good things).

advanced — one-big-image: You can use applications like Powerpoint, Keynote, Omnigraffle, Illustrator, or IPE to create a single large slide. If you make this to scale at 3 feet by 2 feet (either orientation), then the department will print it 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 (see below) and you are satisfied, to print the poster, send email to Chris Coleman at coleman@cs.utah.edu with the subject line “DATA MINING POSTERS” and he will print it for you, and it will be in MEB 3105 by 5pm on April 24. Make sure the poster is in pdf format and is 3x2 feet, otherwise it will not be printed. This email must be sent by April 22 at noon, or it will not be printed. Early submissions are greatly appreciated.

6.3 Poster Outline

A poster outline will need to be turned in (via email, see below) to or discussed with the instructor **by April 17**. Successful completion of this component is worth **5 points** towards the poster presentation. Early submissions are much appreciated.

This is best accomplished by stopping by during office hours or a set up appointment. 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 can make concrete suggestions to improve the poster.

Email outlines to the instructor with subject line "DATA MINING POSTERS". Please only send me pdf files (this can include power points converted to pdf).

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 poster 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 10 points, and the second top vote getter will earn a bonus 5 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 **2 minute** overview of the project. Prepare and practice this *short* oral presentation.