cse6242-2019spring-online-project-description

Project

CSE6242 Data and Visual Analytics

Instructor: Prof. Polo Chau

Grading & Schedule

1. Proposal (7.5% of course grade)
2. Proposal presentation (5%)
3. Progress report (5%)
4. Final poster presentation (7.5%)
5. Final report (25%)

See Course Schedule for all deliverables' due dates.

Important: you will be submitting multiple files as part of your project deliverables. We will deduct 5% from a project deliverable for **every file** whose filename or file format that is different from what we have specified. It is time consuming to find "missing" files or to guess their names.

For example, suppose the final report requires README.txt and report.pdf; if your team submit README.doc and report.doc, 10% will be deducted from the final report's score.

Teaming and Choosing a Topic

The work will be carried out in teams of **4-6 persons**. You are welcome to decide on who to team with, since each team needs to decide on their topic. You may want to try Piazza's "Search for Teammates" pinned post.

During the semester, we will open up a shared spreadsheet that all teams will enter their team member information.

- **Pick your own topic:**
  - You need to justify that the topic is interesting, relevant to the course, and is of suitable difficulty.
  - Required components:
    1. at least one **large, real dataset**;
    2. some **non-trivial** analysis/algorithms/computation performed on the dataset (e.g., computing basic statistics, like average, min/max will not be enough); and
    3. an **interactive user interface** that interact with the algorithms (can be visual, voice-controlled, on tablet, desktop, etc.).
- **Harder way:**
  - Joint projects with other courses might be negotiable. You must obtain the instructor's approval, and you need to clarify exactly what will be done for this course that is on top of and different from what you will do for the other course.

- Projects related to your dissertation/master-project are also possible, as long as there is no '*double-dipping*', i.e., you clearly specify what the project will do, in *addition* to what you were planning to do for your thesis anyway.

Once you have selected a topic, you should do some background reading so that you are capable of describing, in some detail, what you expect to accomplish. For example, if you decide that you want to implement some new proposal for a multidimensional file structure, you will have to carefully read the paper that proposes similar structures, pinpoint their weaknesses, and explain how your approach will address these weaknesses. Once you have read up on your topic, you will be ready to write your proposal.

Why solo projects aren't offered?
The main reasons are: (1) most large-scale data analysis projects in industry are team-based; (2) many former students found the projects highly beneficial to them. If you strongly desire doing solo projects, this course unfortunately is not a good fit for you.

Some students mistakenly believe that the group projects reduce the amount of work that needs to be graded. Any instructors or TAs know that open-ended questions, and in this case projects with topics chosen by students, need a lot of thinking and time from the graders' end. In fact, open-topic group project is one of the hardest thing to scale to large classes. I thought hard about whether to remove it and do exams instead, which could have really saved us a lot of our time (TAs and myself)! I decided to keep the group project because of its benefits.

What datasets are considered "large"?
There are quite a few ways to define "large". It can be measured in size on disk, number of rows in a database, number of edges in a network, etc. One person's "large" could be another person's "small". For example, in my research group, we routinely work with million-edge graphs (say, just working with the graph structure; you'll work on such "large" graphs in HW3 using just your computer), which are considered "small" in the data mining community, but large in other communities and applications. If you're working with videos, a few million of them will take up terabytes of petabytes. For those of you working in industry, you likely would routinely work with datasets that are in terabytes or petabytes.

The main reason for requiring the use of a large dataset is so that you will learn to handle non-trivial computing and visualization problems. So the larger the better. The harder the problem, the more thinking you will need to do, and the more you will learn.

If you can run an algorithm or analysis on your computer and get the results in a few seconds, your dataset is likely too small (and you likely won't learn much from this experience). Similarly, if you can plot every single data point on the screen trivially and that doesn't create any visual complexity or interaction challenges, your data is also likely too small. In other words, you should "suffer" a little when analyzing the dataset, so that you would think about what the challenges are and how to tackle them!

If you have a large dataset and that makes the project too hard, you can always choose to work on a subset of it. But if your dataset is too small (e.g., a few hundreds of rows, each having only a few attributes), you will learn little.

I encourage you to pick an interesting topic and dataset (instead of a "safe" but boring topic) that would excite you -- this way, you would learn more. Be ambitious. It's OK if you end up getting negative results, as long as you make the best decisions you can and you are satisfying all project requirements. One of the nicest thing of being a student is that it's OK to try things out, so take advantage of this opportunity!

If you really need a rule-of-thumb guidance (for this course), I suggest you to consider datasets that have at least hundreds of thousands of rows/records, or at least hundreds of MBs (however, if that mostly contains "filler" information that you won't be using, then that's not a meaningful measure). Again, the larger the better. Use this project as a way to gain experience and knowledge in working with real datasets.

Can we see example project deliverables from previous courses?
Unfortunately, I do not have permission to share previous project teams' deliverables. Also, since all teams are welcome to choose topics most interesting to them, different teams' ideas and approaches can be quite different --- there are many different ways to produce good proposals. Based on the project description and guidelines, most teams in the past developed excellent projects. In academia, when submitting grant proposals, we are not provided with any examples. It is up to us to propose the topic, and to convince the proposal reviewers the significance of our problems, ideas and solutions. We are only provided with high-level format requirements and guidelines of our documents.

Below are two published articles that are based on previous projects from this course (campus section), the articles themselves are not project deliverables from this course, but are like extended, improved version of the teams' final project reports. For our OMS students, these projects are mentioned in *Week 1's "Course Introduction: Course Goals & Expectations" video, start at 4:16*.

Aurigo: An Interactive Tour Planner for Personalized Itineraries
https://www.cc.gatech.edu/~dchau/papers/15-iui-aurigo.pdf

PASSAGE: A Travel Safety Assistant with Safe Path Recommendations for Pedestrians
http://cc.gatech.edu/~dchau/papers/p84-garvey.pdf

To publish these articles, those student teams spent additional time and effort after the course has concluded to extend their project. For example, in Aurigo, the students design and conduct a formal controlled user studies; in PASSAGE, the students improved their methods of computing "safety" scores.

Proposal

Your proposal should answer Heilmeier's questions (all 9 of them; see list below); if you think a question is not very relevant, briefly explain why. In other words, your proposal should describe what you plan to do (the problem to address), why you want to do it, how you will do it (what tools? e.g., SQLite, PostgreSQL, Hadoop, Kinect, iPad, etc.), how your approach is better than the state of the art, why it may succeed, and when it does, what differences will it make, how you will measure success, how long it's gonna take, etc.

9 Heilmeier questions (source)

1. What are you trying to do? Articulate your objectives using absolutely no jargon.
2. How is it done today; what are the limits of current practice?
3. What's new in your approach? Why will it be successful?
4. Who cares?
5. If you're successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?
6. What are the risks and payoffs?
7. How much will it cost?

8. How long will it take?
9. What are the midterm and final "exams" to check for success? How will progress be measured.

Your proposal should be fewer than **1200** words, excluding titles, section names, reference list, tables, charts, images, table of contents, etc., but including the literature survey. It should use **12pt font**, **typed** in **PDF format** (can be created using any software, e.g., LaTeX, Word), and with figures, tables, etc. whenever useful. It should be self-contained. For example, don't just say: "We plan to implement Smith's Foo-Tree data structure [Smith86], and we will study its performance." Instead, you should briefly review the key ideas in the references, and describe clearly the alternatives that you will be examining.

An appendix is for optional, non-essential information. We may not read or even grade it.

Some teams, especially those that want to turn their project into a research publication, use LaTeX for type formatting. If your team chooses to go this route, you may consider using tools like Git (GT GitHub) or Overleaf to work on the article collaboratively. For the LaTeX template, we suggest ACM's standard template (sigconf). You may need to increase the template's default font size to 12pt (e.g., by changing "\def\ACM@fontsize{10pt}%" in the acmart.cls). You are also welcome to use other templates (e.g., IEEE, Springer).

How to write the survey without using too many words?

- See other articles' related work sections for inspiration, e.g., Apolo paper
- Multiple papers may share similar themes, use similar methods so they may be summarized and discussed together.
- Note that survey account for 60% of proposal's grade, so your survey should be substantial!

Grading scheme & Submission instructions

- 60% for the survey
- 30% for innovation
- 10% for plan of activities
- For every Heilmeier question that's **not** mentioned, deduct 5%.
- You may consider organizing your proposal based on the Heilmeier questions (e.g., each section addresses one question)
- Your survey should have **at least 3** papers or book chapters per group member (outside of any required reading for the class).
    - Short papers, like PNAS, Nature, Science papers, count as 0.5.
    - Copying the abstract of the papers is obviously prohibited, constituting plagiarism.
    - For each paper, describe
        - (a) the main idea,
        - (b) why (or why not) it will be useful for your project, and
        - (c) its potential shortcomings, that you will try to improve upon.
- You may use any citation style (e.g., APA, Chicago). Google Scholar supports a wide range of citation styles; it also provides BibTeX (needed if your team is using LaTeX).

- **Clear problem definition**: give a precise formal problem definition, in addition to a jargon-free version (for Heilmeier question #1).
- Provide a **plan** of activities and time estimates, per group member. **List what each group member has done, and will do.**
- Team's contact person submits a softcopy, named **teamXXproposal.pdf**, via Canvas (i.e., that person submits for the whole team), where XX is the team number (e.g., team01proposal.pdf for team 1)
- [-5% if not included] Distribution of team member effort. Can be as simple as "all team members have contributed similar amount of effort". If effort distribution is too uneven, I may assign higher scores to members who have contributed more.

Which papers are considered "long" (or "short")?
Long papers refer to typical papers published at top academic venues (e.g., KDD, CHI, ICML). They are usually at least 8-10 pages long, in 2-column format, which translate into 5000 or more words. Thus, short paper would be 4-5 pages or fewer. Example long papers:

- GAN Lab. VAST'18. 2-column IEEE format
- SHIELD. KDD'18. 2-column ACM format
- ShapeShifter. PKDD'18, 1-column Springer format

Should papers be peer-reviewed?
They should be peer-reviewed, unless there is a strong reason for it not to be (e.g., a book chapter).

What kind of papers are considered relevant?
A paper that you read and cite can be relevant to your project in different ways. You are welcome to cite a paper if you can justify its strong relevance to your ideas, problems (e.g., motivate the urgent need to solve them), or approaches (e.g., your approach improves on an existing method).

Proposal Presentation

1. Team's contact person submits
    1. A presentation slide deck via Canvas, called **teamXXslides.pdf**, where XX is the team number (e.g., team01slides.pdf for team 1). **PDF only**; no PPT or other formats.
    2. A **2-minute** video presentation (one presentation per team), called **teamXXproposal.mp4** (or .avi or .mov), where XX is the team number (e.g., 01 for team 1).
2. Your video should show your slides (e.g., as pdf on your computer screen via screen capture, say using Quicktime, MonoSnap, etc.) with voice narration; it is up to you whether to show your face. You should be able to create this recording quickly with little effort – no need to do any special video or audio editing.

Grading

- [45%] You must answer the Heilmeier questions. 5% for each question. If a question doesn't apply, say so.

- [15%] Brief literature survey. Can be combined with Heilmeier question(s).
- [10%] Expected innovation. Can be combined with Heilmeier question(s).
- [10%] Plan of activities
- [20%] Presentation delivery
- [-5%] Illegible text, tiny figures, bad color contrast, etc.
- [-5%] Overrun
- Your presentation does NOT need to strictly follow your project proposal document. For example, you can talk about ideas and materials that your team has come up recently.
- **Points will NOT be deducted or awarded based on the number of presenters.** We saw great presentations delivered by teams having various numbers of presenters.

Tips

- Use few slides. Less is more! Fewer slides mean less likely to overrun. Being succinct is hard.
- Practice timing and delivery! If you have several speakers, make sure you practice how to transition from one person to the next (e.g., passing the mic, passing control of mouse and keyboard, etc). PRACTICE! PRACTICE! PRACTICE!

Progress Report

This should be fewer than **2000 words, 12pt font, typed**.

It mainly serves as a **checkpoint**, to detect and prevent dead-ends and other problems early on.

It should consist of the same sections as your final report (introduction, survey, etc), with a few sections "under construction", describing the work performed up to then, and the revised plans for the whole project.

Specifically, the introduction and survey sections should be in their final form. The section on the proposed method should be almost finished. The sections about experiments and conclusions will have whatever results you have obtained, as well as “place-holders” for the results you plan/hope to obtain.

The progress report may be written based on your proposal. For example, the survey in the progress report is not required to be identical to the survey in the proposal. That is, you may update the proposal's survey as needed. Of course, the number of papers should not drop below the requirement (3 papers/team member), and the quality of discussion should still be equal or better than that in the proposal.

An appendix is for optional, non-essential information. We may not read or even grade it.

Grading scheme & Submission instructions

- [70%] for proposed method (should be almost finished)
- [25%] for the design of upcoming experiments / evaluation

● [5%] for plan of activities (please show the old one and the revised one, along with the activities of each group member)
● Clear **list of innovations**: give a list of the best 2-4 ideas that your approach exhibits.
● Team's contact person submits a softcopy via Canvas (progress report only), named **teamXXprogress.pdf**, where XX is the team number (e.g., team01progress.pdf for team 1)
● [-5% if not included] Distribution of team member effort. Can be as simple as "all team members have contributed similar amount of effort". If effort distribution is too uneven, I may assign higher scores to members who have contributed more.

Final Poster Presentation [Peer-graded]

Overview

1. Each team creates a single poster for the whole team.

2. Each team member separately prepares and creates a **3-minute** video presentation (i.e., one presentation per learner).

    2.1. Thus, every team member should know his/her project very well. Each team member should plan his/her presentation separately, and team members should not share presentation scripts.

    2.2. Your video should show your poster with voice narration (e.g., as pdf on your computer screen via screen capture, say using MonoSnap, native screen recording software on your OS). It is up to you whether to show your face. You should be able to create this recording quickly with little effort – no need to do any special video or audio editing. You may zoom into and out of the poster as you present, so the viewer can more easily see the poster content.

    2.3. Demo: optional but encouraged. Demo time counts towards presentation time.

3. Upload your video as an **unlisted YouTube video** (NOT “private” or “public”). Unlisted videos can be viewed by anyone (in this case, peer-graders who grade your presentation) with the link to your unlisted video.

    3.1. Submit the **URL (web link)** of your own unlisted YouTube video via Canvas. Your graders will use this URL to view your video. To double-check that your URL works, visit that URL using a separate web browser that has been fully logged out of Google services (e.g., all cache cleared, use “Incognito” mode in Chrome, etc.)

    3.2. Set the title of your YouTube video to **teamXXposter-YY**, where XX is the team number (e.g., 01 for team 1), and YY is the student's last name (e.g., smith). The video title will help your graders more easily recognize who they are grading, streamlining everyone’s grading effort.

    3.3. **IMPORTANT**: you need a Google Account to upload a video to YouTube. To access YouTube, depending on your geographic location, you may need to use VPN (e.g., Georgia Tech’s VPN). **Uploading a large video file can take a lot of time; VPN can further slows that down.** Make sure you finish creating and uploading your video early, so you have ample time to verify that your submission is successful.

4. Each learner will grade several other video presentations by learners from other teams.

    4.1. **Peer grading is NOT anonymous.** That is, a presenter knows who the graders are, and a grader knows who the presenters are.

5. If a grader does not finish all the assigned peer grading, that grader **may NOT receive all or part of the grader’s own final poster presentation grade** (i.e., up to 7.5% of final course grade), since the peer grading is an integral part of the project presentation.

Why unlisted YouTube video?

In previous semesters, students submitted videos via Canvas. That did not work well, creating significant challenges for both our students and the teaching team. A common problem was that graders could not view a video (while the video creator could), causing significant confusion for everyone involved, and overhead in fixing the problems. Some students also had difficulty uploading their video to Canvas or downloading them for grading.

Why peer grading is not anonymous?

Polo wants students to learn and practice delivering constructive criticism, for any concerns and weaknesses identified.

People rarely like to hear about negative comments, even if they are accurate and helpful. Giving negative news is always hard, but that is part of life! This means we should carefully phrase our comments as constructive criticism. For example,

- instead of saying “too much text and not enough figures”, you could say "Fig 1 to 3 are important figures in this project; currently they are not easy to see (images are too small; text is not legible). Suggest reducing the amount of text, e.g., into succinct, bullet points to create space for the figures".
- Similarly, avoid "I don't think that the visualization is anything new or how it is helpful," which is highly subjective. Instead, justify your comments; if the presenter did not clarify the novelty or significance of an approach (it is probably new, but just that the presenter did not point it out), you could say "it's unclear from the presentation and poster whether the proposed visualization is an improvement over the state of the art (it seems to be a standard design); more clarification is needed."

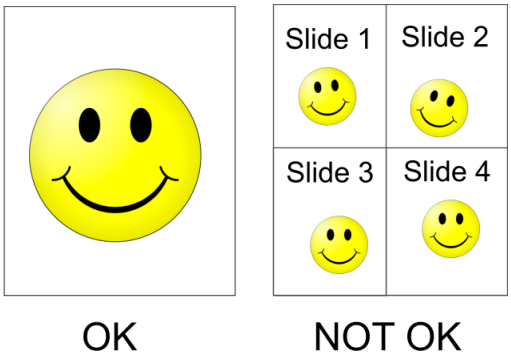
There are pros and cons for both anonymous and open review. It is still an open research problem. For example, one potential benefit for open review is reviewers could be more tactful and constructive, where anonymous reviewers could be more critical (sometimes not in a good way) and may do less work than they should.

Poster Design

Design your team’s poster ***well before*** the submission deadline, to avoid last-minute rush.

The poster must be in portrait orientation, **30 inches wide and 40 inches tall** (if printed). We suggest using **18pt** font size and larger.

A deck of PowerPoint slides is **not acceptable** as a poster. See the illustration below for what is allowed and what is not.



Grading

Your poster presentation should cover the following parts (point distribution shown on the left). Thus, the grading is about both your presentation delivery (e.g., what you say, where you direct the audience's attention), and the poster content.

10%	**Motivation/Introduction**: 5% What is the problem (no jargon)? 5% Why is it important and why should we care?
20%	**Your approaches** (algorithm and interactive visualization): 5% What are they? 5% How do they work? 5% Why do you think they can effectively solve your problem (i.e., what is the intuition behind your approaches)? 5% What is new in your approaches?
10%	**Data**: 5% How did you get it? (Download? Scrape?) 5% What are its characteristics (e.g., size on disk, # of records, temporal or not, etc.)
25%	**Experiments and results**: 5% How did you evaluate your approaches? 10% What are the results? 10% How do you methods compare to other methods?
10%	**Presentation delivery**: 5% Finished on time? 5% Spoke clearly and at a good pace?
25%	**Poster Design**: 5% Layout/organization (Clear headings? Easy to follow?) 5% Use of text (Succinct or verbose?) 5% Use of graphics (Are they relevant? Do they help you better understand the project's approaches and ideas?) 5% Legibility (Is the text and figures too small?) 5% Grammar and spelling

Possible software to create posters

1. Powerpoint/Word (save as pdf) -- GT's Office365 Powerpoint supports collaboration.
2. Apple Pages (FREE) supports real-time collaboration (via iCloud and desktop software)
3. Inkscape (free, cross platform)
4. Polo uses Affinity Designer (Mac and windows)
5. Google Drawings (File > Page Setup to set document size)
6. draw.io (File > Page Setup to set document size)

Example poster design

The following posters were for research projects conducted at the Polo Club of Data Science, and were not for projects from this class. They do not strictly follow the format described in our grading rubric.

- Apolo graph exploration
- Insider trading pattern discovery
- Comment spam detection

Final Report

It will be a detailed description of what you did, what results you obtained, and what you have learned and/or can conclude from your work.

Components:

1. **Writeup**: no more than **2800 words, 12pt font, typed**. Describe in depth the novelties of your approach and your discoveries/insights/experiments, etc.

2. **Software**: packaging, documentation, and portability. The goal is to provide enough material, so that other people can use it and continue your work, if you are to open-source it --- in other words, you should make it easy and attractive for others to use your work.

Grading scheme & Submission instructions

- Writeup
    - [2%] Introduction - Motivation
    - [3%] Problem definition
    - [5%] Survey
    - Proposed method
        1. [10%] Intuition - why should it be better than the state of the art?
        2. [35%] Description of your approaches: algorithms, user interfaces, etc.
    - Experiments/ Evaluation
        1. [5%] Description of your testbed; list of questions your experiments are designed to answer
        2. [25%] Details of the experiments; observations (as **many** as you can!)
    - [5%] Conclusions and discussion
    - [-5% if not included] Distribution of team member effort. Can be as simple as "all team members have contributed similar amount of effort". If effort distribution is too uneven, I may assign higher scores to members who have contributed more.
- [10%] Team's contact person submits one zip file, called **teamXXfinal.zip**, via Canvas, where XX is the team number (e.g., team01final.zip for team 1). The teamXXfinal.zip will contain the following 3 components:
    - **README.txt** - a concise, short README.txt file, corresponding to the "user guide". This file should contain:
        1. DESCRIPTION - Describe the package in a few paragraphs
        2. INSTALLATION - How to install and setup your code
        3. EXECUTION - How to run a demo on your code
        4. [Optional, but recommended] DEMO VIDEO - Include the URL of a 1-minute ***unlisted* YouTube video** in this txt file. The video would show how to install and execute your system/tool/approach (e.g, from typing the first command to compile, to system launching, and running some examples). Feel free to speed up the video if needed (e.g., remove less relevant video segments). This video is optional (i.e., submitting a video does not increase scores; not submitting one does not decrease scores). However, we recommend teams to try and create such a video, because making the video helps teams better think through what they may want to write in the README.txt, and generally how they want to "sell" their work.
    - **DOC** - a folder called DOC (short for "documentation") containing:
        1. **teamXXreport.pdf** - Your report writeup in PDF format; can be created using any software, e.g., latex, Word.

2. **teamXXposter.pdf** - Your final poster.

○ **CODE** - All your code should be added here. Make sure that your package includes only the **absolutely necessary** set of files.

Should datasets be included as part of our submission?

If you are referring to (small) toy data for a demo (that your graders will run), you are welcome to include them. Think about the open-source software libraries that you have seen or have used, they would often include some sort of "quick start" guide to get a demo running on a toy dataset.

For large datasets, please do **not** include them; if the dataset is public and can be easily downloaded, include the link to the dataset.

If getting a dataset requires writing scripts/programs, include those scripts, and write down the steps that people will need to go through (e.g., register for an account to get API key).

If you have processed the dataset in some ways, include the code you used, and the steps people will need to go through.

Version 1: 2018-12-17