Group Project Proposal

Topic:

Pick your own topic:

You need to justify that the topic is interesting, relevant to the course, and is of suitable difficulty.

Three major required components:

i) at least one relatively large, real dataset;
ii) some **non-trivial** analysis/algorithms/computation performed on the dataset (e.g., computing basic statistics, like average, min/max will not be enough); and
iii) visualizations that visualize the discoveries.

Please be mindful that the class project is NOT the best venue to express personal or subjective opinions — rather, it is for students to practice performing objective analysis based on facts and science, you tell the stories told by the data and analysis.

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, 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. The larger the harder the problem, the more thinking you will need to do, and the more you will learn.

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.

I suggest you to consider datasets that have at least hundreds of thousands of rows/records.

Though we encourage students to use larger datasets, due to the submission limitation, please do not use very large dataset, say larger than 20M. If the dataset is too large, please do not upload your dataset to UBLearn, instead, save it in a cloud storage and upload the link in your submission, please make sure that it can be opened by anyone with the link.

Proposal Requirements

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., regressions, classification, clustering or multiple methods), 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 will take, etc.

9 Heitmeier questions:

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 document must be no more than **2 letter-size pages long, excluding references.** In other words, only the references do **NOT** count towards the page limit; everything else — including the literature survey — counts. **Use at least 1-inch margin for each page (top, right, bottom, left).** It must use **11pt font (or larger)**. The document must be in **PDF format. You may create the document using any software that you want).** Include any figures, charts, tables, captions, etc. whenever useful — they count towards the page limit (they may include text whose font size is smaller than 11pt, but such text must be legible). Your document should be self-contained. For example, do not 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.

Grading scheme & Submission instructions

- [60%] Literature survey
    - Your literature survey should have **at least 3** papers or book chapters per group member (outside of any required reading for the class).
        - Using “long” papers, see below for description.
        - 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.
    - Make sure to cite your references in your literature survey.
    - The literature survey can be in its own section, or be integrated into the answers of relevant Heilmeier questions (e.g., #2 and #3).

- [30%] Expected innovations
- [10%] Plan of activities

    - Using either a **Gantt chart** (example) or a **table**, describe

        - the activities each member has done and will do; and
        - each activity's start and end time (or start time and duration).

    - [-5% if not included] Provide a **statement** that summarizes the distribution of team members' effort. The summary statement can be as simple as "all team members have contributed a similar amount of effort". Place this statement immediately after the Gantt chart (or table). If effort distribution is too uneven, we may assign higher scores to members who have contributed more.

- [-5%] For every Heilmeier question that's **not** mentioned.
- Some teams organize their proposals based on the Heilmeier questions (e.g., each section addresses one question). Some teams organize theirs using section headings from the final report (e.g., "Introduction", "Literature Survey"). The exact organization is up to you, provided that your answers to the Heilmeier questions are easy for us to spot.
- Include your team project's title, team number, and all team member names (at the top of the first page)
- Team's contact person submits a softcopy, named **teamXXXproposal.pdf** (i.e., that person submits for the whole team), where XXX is the team number (e.g., team001proposal.pdf for team 1). Submit via Canvas.

How to write the literature survey without using too many words?

- Multiple papers may share similar themes, use similar methods so they may be summarized and discussed together.
- Note that the literature survey accounts for 60% of the proposal's grade, so your literature survey should be substantial!

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.

Should papers be peer-reviewed? How to tell if a paper was peer-reviewed?
Yes, they should be peer-reviewed, unless there is a strong reason for it not to be (e.g., a book chapter).

You can usually find out whether a journal or conference proceeding is peer-reviewed by checking its submission and reviewing process (or lack thereof), by visiting the conferences or publishers' websites. A quick way to look up the venue at where a paper was published is to plug in the paper's title into Google Scholar. A **very short** of these for data analytics/Machine Learning/AI include: KDD, IEEE ICDM, IEEE Vis, UbiComp, SIAM SDM, ICDM, NeurIPS, ICML, AAAI, IJCAI, AISTATS, and many more.

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). Searching on Google Scholar can also help you to find relevant papers.