EAS 508 Group Project Proposal

Fall 2024

Introduction to the project

The objective of this project is to explore a real-world dataset using statistical learning techniques taught in the course. The project aims to apply non-trivial statistical learning algorithms, provide insightful data analysis, and effectively communicate findings using visualizations. The project will include problem formulation, dataset selection, experiment design, model application, and performance evaluation.

Topic

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

The project must include the following three key components:Three major required components:

- At least one relatively large real dataset. You may use publicly available datasets from sources like Kaggle, UCI Machine Learning Repository, Google Dataset Search, etc..
- Clearly define the problem based on the dataset, such as prediction, classification, or clustering tasks. Justify why the problem is interesting and its relevance to statistical learning.
- Some non-trivial analysis/algorithms/computation performed on the dataset (e.g., computing basic statistics like average, min/max will not be enough); and
- Visualizations that visualize the discoveries.

This project offers an opportunity to practice objective, data-driven analysis. *The focus should be on the facts, discoveries, and the stories that the data tells, rather than personal or subjective opinions.*

What datasets are considered “large”?

The concept of a ”large” dataset varies. It could be defined by file size (terabytes or petabytes), the number of rows, or the complexity of the dataset (such as a large network or a significant number of records). For example, video datasets may be extremely large due to their data requirements. The challenge is to work with datasets that push you to apply advanced computational techniques and to develop creative visualizations, as this will enhance your learning experience.

If the dataset is too large, you can choose a subset to work with, but it should be big enough to necessitate non-trivial analysis. Small datasets, such as those with only a few hundred rows and limited attributes, may not offer enough learning opportunities. I encourage you to choose a challenging dataset that excites you, as tackling complex problems is where the most valuable learning occurs.

For submission, keep in mind the file size limitation (no datasets larger than **20M**). If your dataset exceeds this limit, please use cloud storage and submit a shareable link instead.

Proposal Requirements

Your proposal should answer Heilmeier's questions (all 9 of them; see list below). If a question is not very relevant, briefly explain why.

In other words, your proposal should describe:

Heilmeier's Questions:

1. What are you trying to do? Articulate your objectives using 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? How do you measure it (e.g., via user studies, experiments, ground truth data)?
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?

Remark: If some of items is not applicable, for example, if there is no cost to get the data to address Question 7, simply explain that you get data from somewhere with no cost.

Your proposal document must be **NO more than 2 letter-size pages** long excluding references. Use at least **1-inch margins** on all sides. The font size must be **at least 11pt**. The document must be in **PDF format**. You may create the document using any software. Figures, charts, tables, and captions should be included whenever useful, and they count toward the page limit.

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: Include at least 3 papers or book chapters per group member.

For each paper, describe:

- Using "long" papers, see below for description.
- Copying the abstract of the papers is obviously prohibited, constituting plagiarism.

- For each paper, describe
  1. The main idea,
  2. Why (or why not) it will be useful for your project, and
  3. Its potential shortcomings that you will try to improve upon.
- You may use any citation style (e.g., APA, Chicago). Be sure to cite your references in your literature survey.
- 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** or a **table**, describe:

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

-5% Provide a statement summarizing the distribution of team members' effort.

The summary statement can be as simple as, "All team members have contributed a similar amount of effort."

[-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 Brightspace.

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, a 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 where a paper was published is to plug the paper's title into Google Scholar. A very short list of venues for data analytics/Machine Learning/AI includes: 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.