# Final project guidelines

Wesley Tansey*

September 11, 2023

Statistical modeling is a skill. Skills require practice. The final project is therefore central to this course. You are generally free to choose your topic and project. The requirements are that it fundamentally relies on or extends the concepts we will cover in class. Ideally, it will be a project that is relevant to your PhD and may even lead to a chapter in your thesis.

Most projects will likely involve an **application of data science** to a dataset you are interested in analyzing. For these projects, the data can be publicly available or internal to a lab you are in. For the latter case, use of the data is subject to the approval of the PI who owns the data and their consent for you to write about your findings in a report that only I will see. Data analysis projects typically will have some stated scientific question you want to ask about the data, like "how does x influence y?" or "what are the latent subtypes of y and are they predictive of response to treatments a, b, and c?" For these projects, you will need to hypothesize a statistical model of the data, find a way to fit it, investigate the results, and write up both your scientific findings and your criticisms of the model you chose. You will also need to provide a brief background of the scientific problem and a short review of how people have addressed (or failed to address) the problem to-date, such as previous experiments or analyses of different data modalities. You will likely go through two or three iterations of the model during the course of the project, as your familiarity with the dataset grows.

Another route is a **methods project** where you propose a new machine learning or statistical method for solving a class of problems. For these projects, the typical approach is to survey a few papers from the recent literature and then propose an extension, variation, or enhancement to a method that you hypothesize will improve upon the current state of the art. Methods must be benchmarked against at least two baseline methods on two different tasks or datasets. You will then have to write up your method, including the justification, motivation, background review, benchmark results, and discussion of any strengths and weaknesses you observed. As with modeling projects, you may go through a few iterations of the method through the course of the project. It is totally fine if the conclusion of the project is that the method does not beat the baselines; in that case, you should investigate why, pose a possible reason, and suggest what could potentially improve the method.

The project involves three assignments. For each, please use LaTeX, a 12-point font, and 1-inch margins. Page limits are without figures; include as many pages of figures as needed. Put the figures inline where appropriate rather than at the end of the paper.

## Proposal abstract and EDA (Due: October 2)

The first assignment is to write your proposal in the form of an *aspirational abstract*: if everything goes well, this will be the synopsis of the paper, including the discovery, insight, or success of the model or method you

*Largely stolen from David Blei's machine learning course project guidelines: `http://www.cs.columbia.edu/~blei/fogm/2022F/materials/project.pdf`

are proposing. This should be justified by exploratory data analysis (EDA). For application projects, this means you should have made several plots investigating different variables and investigated where potential latent signal or structure may exist in the data (e.g. through looking at PCA or UMAP plots, running off-the-shelf models, etc.). For methods projects, you should have run at least one of your baseline methods on your benchmark datasets and identified an area where the baseline fails. As noted above, the abstract and report is likely to evolve as you work on the project. The abstract is still helpful as a way to organize your thoughts, plan the project, and think about what success would look like.

You are encouraged to refer to computer science conferences (such as *Neural Information Processing Systems*, *International Conference of Machine Learning*, *Artificial Intelligence and Statistics*, and *International Conference on Learning Representations*) or journals (such as *The Annals of Applied Statistics*, *Journal of the American Statistical Association*, *Journal of Machine Learning Research*) to get a sense of how to write an abstract.

The proposals are graded as 15% of your project grade and should not be longer than 2 pages, not including figures.

## Milestone report (Due: October 30)

The milestone report describes the problem you are addressing and discusses some preliminary results. For application papers, you should have reached the point of trying a very basic model on your data. For methods papers, you should have reached the point of your first prototype method. Include what you have completed and what you plan to finish by the end of the semester. By the time of the milestone report, you should have invested substantial effort into your report; budget at least 40 hours of work to reach this point. It is your responsibility to contact me if you get stuck along the way. If you delay working on this until the last week, do not expect me to be available for substantial time to help you reach the deadline.

The milestone project is 2–4 pages long and counts for 30% of your project grade.

## Final report (Due: December 4)

The final report should include all of your results, as well as a discussion on how your project evolved (and why), any final takeaways, and thoughts on next steps that could extend the analysis or method. The report is up to 5 pages, not including figures. You are encouraged to have a lot of figures and analyses. Additional details that do not fit in 5 pages can be placed in an appendix; there is no length restriction on the page length.

You are also encouraged to make your code open source (e.g. on github). You can submit your code for your project as either an ipython notebook, colab link, or github repository. For private github repositories, add the github user `tansey` as a collaborator.

The final report counts for 45% of your project grade.

## Final presentation (December 11–13)

The last 1–2 weeks of class everyone will be required to briefly present their project. The talk is 10min plus 5min for questions. It counts for 10% of your project grade.

### Project evaluation

We evaluate the project on ambition, significance, originality, technical depth, results, relevance, and writing quality. A good book on writing quality is Williams (1981). A useful app for editing is `https://hemingwayapp.com/`.

## References

Williams, J. (1981). *Style: Towards Clarity and Grace*. University of Chicago Press.