Starting a project/thesis

Table of contents

Official info 1

1 Preamble 2
- 1.1 Goal . . . 2
- 1.2 Libraries . . . 3
- 1.3 SMART criteria . . . 4
  - 1.3.1 Example . . . 4
  - 1.3.2 Example . . . 5
  - 1.3.3 Example . . . 5

2 Timeline 6
- 2.1 Deadlines . . . 6
- 2.2 Progress Forms . . . 6

3 Additional Thesis meetings 7

4 Supervisory Dissolution 7

5 Outcomes of project 7

6 Expectations For Writing 7
- 6.1 Resources . . . 8

7 Proposal evaluation criteria 9

Official info

The official information about your proposal for your Master's of Science in Data Science (MSc) is available on moodle. The major items to think about during the proposal are:

- Committee (thesis) or second reader (project)
- Forms
- Proposal Document

If there's ever a discrepency between the information here and what's on moodle, assume the moodle one is more correct.

1 Preamble

For project students, this is *your* proposal for your project, not a project that is designed by an instructor and assigned to you. It's up to you to come up with a project that's suitable after we've agreed that I'm an appropriate supervisor for that kind of project. I will approve whether the specifics are at an appropriate level for a MSc and you will tell me the criteria for which I will hold you accountable during the terms you are completing your project. Only rarely should you get stuck or need help because you've given sufficient thought and research to your proposal.

A thesis involves novel work which will challenge the student to the point that they need help and advice even with a well-designed proposal. It will have the requirements of the project plus the added challenge of creating/justifying something new to the field.

To not stifle creativity, there are no official criteria for a project but generally, a project must convey and describe an appropriate theoretical data science method, or uncover interesting insights from non-trivial data. This would disquify some projects that are more of a coding project, such as coding an AI to automatically respond to emails, even though it heavily uses machine learning. An acceptable variation of this is if your project is to learn about natural language processing, and create your own method of training and testing a machine learning algorithm.

1.1 Goal

Your proposal documents that you have done enough research and have the skills to perform the tasks that you are assigning yourself without needing additional help except in rare circumstances. Any possible changes in your plan should be anticipated and the contingency appropriately outlined. You should not have to figure out how to do anything after the proposal is written because you've planned well.

If you plan things well, you will not spend very much time at the computer coding the project itself. By that, I mean a well-planned proposal should be able to be coded up within a week, perhaps two. You should *not* have a mindset of diving in and trying things to see what happens, or figuring it out as you go. If you have such plans, then you haven't actually proposed your project, you've proposed a proposal.

I strongly recommend you think of this as writing a set of background knowledge, and then a concise set of instructions for two other MSc DSc students as their part-time job. Each of these hypothetical students should be able to follow your proposal independently of each other, with no real questions on what to do, and come back to you with the same set of results.

The proposal's literature review will give an overview of the mathematics and algorithms that the student must learn through your document. You should imagine that the student would need to learn the math as much as I require you to learn the math in courses, and would be required to code it all from scratch using no libraries so you are responsible for the mathematical details that will have to go in the final document. This also disqualifies some interesting projects as it can lack a necessary theoretical component. In your lit review, implementation details should focus on what the algorithm does and uses very limited analogies/metaphors unless used as an introduction mechanism. You should not discuss how the functions or libraries work, the parameters you pass, etc..

In your methodology, you will explain how your methods in your lit review are implemented and there should be no questions about how something should be done, what process to use, or a decision to make. This means the proposal requires considerable thought to outline for different possible ways the that the project can ultimately go. That may sound like more work, and it is, but the more thought you put into it, the more you will realize which possibilities can actually occur and which ones won't.

1.2 Libraries

You are permitted to use any existing library if you could code the whole thing from scratch but you're choosing not to, to save time. At any point, I should be able to present you with a small version of the data set and you can show me using on paper and a calculator the exact steps and calculations which lead to to the final answer that matches the library as well as corresponding analysis.

There are some exceptions to this, such as if an algorithm that's one small step of your project works faster but gives you the same answer for that one small step as a standard algorithm that you understand. In this case, you may use the faster algorithm without knowing the details. If, however, a major component of your project is to compare how the two algorithms perform, or show that they give different results, then you *are* responsible for knowing how the faster algorithm works. Similarly, it is unacceptable to simply use a library with default parameters: you should be able to justify why the default parameters are applicable in your scenario.

An easy way to to know if you're doing things at an appropriate level is to assume you're doing everything with slow/standard algorithms that you understand. If you can complete your proposal and project that way, then you may make substitutions simply to make your project go faster or more accurate. If you cannot make a proposal/project without mentioning the faster or more accurate algorithm (e.g., because the more accurate algorithm is the point

of the comparison) then you must understand the more accurate method as that's a key aspect of your project.

1.3 SMART criteria

In creating your proposal, you should keep in mind the SMART criteria:

- Specific
- Measurable
- Achievable (for your time frame/skill/resources)
- Relevant (to our program)
- Timeline

These are not 5 separate points/sections, but at some point in the proposal each of these things will be addressed and satisfied. Before writing your proposal, I strongly recommend you create a single (run on) sentence which satisfies all of the SMART criteria for your proposal, and then build your proposal from that. It's very common to put the SMR criteria in the last paragraph of your introduction.

Very often, at least one of these criteria are violated when the concept of a proposal is rejected or needs more work.

1.3.1 Example

Bad example: This project will determine whether naïve Bayes performs better than a classification tree on large data sets.

I would respond with things like

- "better" defined and measured how? faster computationally? accuracy? f1 score? mean squared error?
- What kind of naïve Bayes and what kind of tree? Using greedy optimization like making full trees? If you prune, how much will you know to prune (assuming this isn't in the methodology). Which distribution will you use and why? Do you have any proof that the defaults in the code are appropriate for your data?
- etc..

Better This project will assess whether naïve Bayes performs better than a classification tree on three commonly used large data sets (data set 1, data set 2, and data set 3). Performance will be measured through mean squared error when employing 10-fold cross validation.

In this example, it wasn't **specific** or **measurable**. In your proposal, you should spell out both of these things to the point that the theoretical data science student would not have to

think about how to do this, because there is a clear instruction/decision/choice made for this in the proposal.

1.3.2 Example

Bad example: This project will create a new machine learning library that will analyze one million incoming emails, and their corresponding responses, to try generating an AI that can automate responses.

My biggest issue with this is whether this project can convey knowledge at a Master's level. Despite being an interesting project, it could be inadequate as the MSc final project if it remains vague about what the theoretical aspect that the student will convey. The second issue I have is how would you know whether the responses are correct?

Better: This project will use an existing library to analyze emails to determine which require responses. Of the ones that require responses, the sentiment analysis will analyze incoming emails, and their responses, to try generating an AI that can automate responses.

In this example, the project was not **achievable**, may not have been **relevant**, and was not **measurable**.

As a side note: I would generally not be involved with this as this is not my area. I can chat with you to help you find a good supervisor that is in this area, though.

1.3.3 Example

Bad example: I will conduct further analysis to see which variables should be included to make the best possible model.

This is not **specific**, and therefore, not reproducible, which indirectly means there is no way of knowing if it is actually **achievable**. There's no sense of how many sources will be used to look for variables, when to stop looking, how to determine whether the variable "should" be included in the model, etc.. You have to propose *how* you will do each of these things and leave no doubt for how other hypothetical students would do it, and ensure they will come up with the same answer as someone else who does this proposal.

Better: This project will consider the following 11 variables outlined from [reference] as potentially influential in the outcome. Variables will be included by including any variable that remained in the fnial model when performing forward stepwise regression, as outlined in the literature review in section XX, as well LASSO, as outlined in the literature review in section YY. The union of variables that comes out of the final models from either method is included in time series analysis, described below.

2 Timeline

Timelines should generally be performed backward. For thesis students, check the official documentation since there are some institutionally-dictated ones. For project students, this is slightly more flexible but the official documentation has the necessary deadlines. Include all of the pertinent deadlines in your proposal.

Dates should be set for the following (put them in chronological order but when you are brainstorming you'll do this backward):

2.1 Deadlines

Plan for at least the following deadlines. You will have *many* more deadlines, however.

- Nov 15: implementation complete
- Jan 15: first complete draft*
- Feb 15: second draft
- Mar 15: last day to submit final draft to circulate to committee/second reader
- Apr 15: presentation/defence

In addition to the above, prior to the implementation, you should state your own milestones for the implementation leading to the completion above. I will leave this to you but hold you accountable to it.

2.2 Progress Forms

Both project and thesis students will conduct supervisor-only progress reports every 2 months on approximately the following dates

- Oct 01
- Dec 01
- Feb 01
- Mar 01

At these meetings, the supervisor evaluates if the student is making appropriate progress according to the timeline set out in the proposal. While infrequent setbacks can occur, the student is still generally expected to follow the timeline.

The results of progress reports are usually one of 3 or 4 categories:

- satisfactory progress
- concerning progress
- unsatisfactory progress

3 Additional Thesis meetings

The thesis students will have supervisory committee meetings at least once a term. The committee will decide whether it should be the first, second, or third month of each term at the inaugural committee meeting.

4 Supervisory Dissolution

The student will state as an additional section in their proposal that they agree the thesis or project will be dissolved if any of the following happen:

- Two consecutive progress reports are unacceptable
- Three consecutive progress reports are concerning/unacceptable
- An academic integrity violation is suspected by the supervisor and suspected by at least one other faculty member.

5 Outcomes of project

I like the idea of there being clear expectations of what happens to the final result of the project. If the project is being published (e.g., an R package, or a journal publication) what is the order of authorship, who owns data, etc..

My personal approach is that students are first author if they've done the writing. If I have to heavily edit or rewrite your thesis to make it acceptable quality for publication, you lose the prestigious status of first-author, which I would have preferred you to have. I also may not have the time or interest to do that so it may simply not be published at all and what could have doubly benefited you has been wasted. This is another reason the quality of writing is important.

This section does not have to be in the proposal, but I like it here so that we know it's officiall documented and signed off on.

6 Expectations For Writing

For documents like the proposal and the final document, you'll be given feedback. The feedback is usually in terms of content, or presentation/writing. Ideally you only need feedback for the former, and small amounts for the latter.

In practice, I find that some students underestimate the standard of writing expected for the document that ends their MSc degree and so the timeline I propose reflects that.

I will typically limit how much time I spend reading/annotating a document you submit (e.g., an hour for a proposal, three hours for final documents). If there are so many issues that I do not make it through your document due to poor writing, you've lost your opportunity for me to give you feedback on the later contents of the document, where I may catch additional major issues which themselves require additional revisions. If you require more revisions, they will simply occur *after* your original timeline, which means your degree completion will be set back by a term (or more), and you will be responsible for paying the corresponding extension fee.

6.1 Resources

If you believe there is a difference in how you write and the quality of writing in technical documents like academic journals, you'll be expected to change your writing style to match. If you anticipate this is going to be a problem, begin working on this early, such as in your classes for assignments and projects, as well as your seminar reports. These are all excellent opportunities to obtain feedback about your writing (even if there's not a mark associated it for that particular document). Instructors likely don't have the time to edit your documents for you, but likely will be happy to give suggestions on how to improve your writing. It is up to the instructor's discretion on how much (if any) weight to assign to writing quality in courses and seminar. Please do not make the logical fallacy of concluding that a good mark in an assignment (not expressly meant to evaluate your writing quality) implies that level of writing would be appropriate for your final project/thesis. The institution has a writing centre that will help you in improving your writing. I have observed students improve tremendously over several sessions when the session is approached correctly. The writing centre can teach you how to write better, which requires a commitment to learn how to critically appraise your own writing with their guidance to find where you need improvement and fixing your own writing. Treat the staff as educators that can teach you the skills you need to help you write better depending on your particular needs. Do not expect them to be editors that simply fix your writing for you.

The one caveat I give is that they become *very* busy in the winter term when most students are approaching their deadlines. I would strongly suggest that your final project is not when you seek them, but much earlier than that such as during your coursework if you receive feedback that the quality of writing needs improvement.

If I suggest from your proposal that you would benefit from the writing centre, I expect you to learn from any feedback I give you, and seriously consider making an appointment with the writing centre to rectify writing issues so that you are not held back a semester.

7 Proposal evaluation criteria

While not a binding document, you're encouraged to look (in the menu) at the proposal checklist of things that I look for when evaluating a proposal or final project or thesis.