Starting a project/thesis

Table of contents

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 appropriate 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 perform 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 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 spend far less time at the computer coding than writing your proposal. I generally have not seen a project whose coding cannot be completed within a week or two if planned properly. You should *not* have a mindset of diving in and trying things to see what happens, or figuring it out as you go.

Your proposal should be written as if you are writing a project requirement for a good, but brand new data science student to execute for you, like for an independent study project of a course. In the proposal, you 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.

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..

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.

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.

It is absolutely unacceptable to simply use a library with default parameters: you should be able to justify why the default parameters are applicable in your scenario. There are some rare exceptions to this,

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.

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?
- et..

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.

2 Proposal format

After several issues, I now require all projects and theses to be written in LaTex.

Your proposal should have the following document sections

2.1 Introduction

Your introduction should be (1–2 pages).

- If the main goal is to analyze a particular data set, motivate the context (a few sentences) of how that data is generated and what kind of problem will be solved or what insight you wish to gain from it..
- If the main goal is to analyze performance (e.g., for comparison) you need to motivate what the method(s) are designed to accomplish, and why there are different methods that exist in the first place.

2.1.1 Data description

- Identify/describe of all variables. You must identify the type of variable. You must also know what each variable is in the context of the problem; it can't just be just another feature you use to train an algorithm.
- Continuous variables should indicate units
- Ordinal and nominal variables should at least identify the number of levels, and all levels if there are only a few and an example of a few if there are more a - If relevant to the investigation, an approximate distribution of each variable, using an appropriate method of summarizing. This may be extended to correlations

2.2 Lit review

Your background/lit review should be 4+pages.

The background/lit review will explain relevant methodological and algorithmic information to follow the next section (proposed methodology for the project/thesis). A bad habit of students is to rely on analogies of what's being calculated. You must understand and justify all of this. This section is usually the longest section of the proposal for this reason. Most of your citations should be from journals, though some websites and documentation are acceptable.

Methods should explain what any objective functions are, how optimization occurs, etc.. If a calculation (e.g., the objective function) is not the first, most intuitive calculation the brand new data science student would think of, you must justify why you use something less intuitive/more complicated to solve this problem.

If you elect to use a method, you are taking on the responsibility of explaining these choices at the Master's level. If you can't answer that, the topic/method is unsuitable for you and you must choose something you can understand and explain. Exceptions to this are rare. Many students underestimate how much they must understand the *math* behind algorithms to do a project, let alone a thesis.

What you discuss in your lit review should be required for the methodology. One notable exception is if you are doing a thesis and plan to provide an alternative approach to something, it is still worthwhile to explain what your approach is alternative to, here. This is especially important if you're performing some kind of comparison.

2.3 Methodology

Your proposed methods section should be 2+ pages.

Assuming the theoretical new data science student understands the literature review, they should be able to implement the methodology set out in the proposal with ideally no other material necessary. A good lit review will mean the methodology will be very concise, direct, and almost like a checklist of instructions.

The methods should include the specific implementation of the algorithm. E.g., a lit review will explain multiple regression with dummy variables, while the methodology will literally write out the equation of the regression variables listing each coefficient for the particular data set. Algorithmic details for your particular data set go here as well such as if you need special initial values for your algorithm, parameters you'll set in functions, or tuning parameters that you know ahead of time.

Simulation studies will identify exactly which parameters of the simulation will be investigated, which values will be fixed, and any corresponding assumptions.

A thesis which develops *new* methods will have a much longer methodology section because nothing exists to solve the problem at hand, so you propose to develop it in in this section, and the development is part of the instructions to the student. But because this is a proposal, you don't actually perform the derivations here, you write a skeleton of how you plan to perform the derivation using elements of the lit review.

2.4 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):

2.4.1 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.

2.5 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

2.5.1 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.

2.6 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.

2.7 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.

3 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 semester (or more), and you will have responsible for paying the corresponding extension fee.

4 Proposal evaluation criteria

While not a binding document, you're encouraged to look at the following (non-exhaustive) list of things that I look for when evaluating a proposal or final project or thesis.