M.Sc. Thesis Instructions for Data Science for Decision Making

Hand-in will be 3rd of July, 10.00

Presentation will be 6st of July (TBC) The final presentation time and date needs to be confirmed with the thesis challenge organizations and is therefore subject to change.

This document provides a framework for the MSc thesis for Data Science for Decision Making. This framework is set-up to provide students with enough freedom to develop their own research ideas while at the same time giving some common starting position for this work. Important points:

- Up to three students can cooperate when writing the thesis. Students form teams independently.
- Hand in one pdf and provide code on GitHub.
- The paper should be between 15 and 30 pages in length. However, appendices are allowed for example for robustness checks or longer mathematical proofs.
- Mandatory sections: Introduction, Related Literature, Data Section, Results Section, Conclusion
- Bibliography needs to signal that the students have done a proper literature research (see Research Hacks).

Additional points:

- Creativity and ambition will be rewarded. But you should use and build on tools provided in the MSc where possible. The tools need to be explained when used. This makes evaluation easier.
- I understand you are under severe time constraints. Please don't react to this constraint by being sloppy and overstretching. Simple, focused, correct arguments are better than wrong, bold claims.

- Always label axis in figures, give proper regression tables. Give summary statistics and exploration of corpuses on interesting dimensions. We will subtract points for improper tables and figures.

The key of the thesis is that the teams signal that they understood and are able to use the material and kind of arguments from the lectures/literature correctly. If you are working on a challenge some weight of the grade will be on whether you managed to solve the challenge. But we will follow the process closely so do not obsess about success.

The evaluation of the Master thesis will take into account this use of the material, the soundness of the arguments made, the originality and effort put into the three main parts of the thesis (literature, theory/conceptual, data analysis). These three parts need to be linked with each other.

Put effort into the presentation. This makes grading the Master thesis much easier for us because it communicates clearly where your team sees the strengths of your approach. The "work in progress" presentations can focus on partial aspects but the final presentation has to present the entire thesis.

Research Hacks

1) Good literature search and summary is completely underestimated by students. Typically two things arise: a) students come up with sources which are strange (unpublished or odd journals) b) the bibliography is over-reliant on very few sources. This leads to literature overviews which provide detailed summaries of articles but which are not really helpful in understanding the broader points. What you need to do is to think of issues/topics instead of specific articles when doing a literature summary.

2) Snowball like a pro: Don't just read one article but read the bibliography of the most relevant articles and follow it. Mark articles that seem relevant in the bibliography and look for them on Google Scholar. Once you have found them don't just download the articles blindly but read the abstract of each before you do. When you have ten articles start reading the intros. Stop reading if you think it's not relevant - keep reading if you think it is. Don't read more than the intro but make a list of articles that remain interesting after you have read the intro. Finish the article you started with, then go to the articles that seemed relevant. Repeat for several hours.

3) Important trick: if you have old articles that are relevant you can also forward snowball by typing them into Google Scholar and then looking at who cites

them. Don’t print out anything until you are sure you need to understand an article properly! Printing is a productivity surrogate and kills trees.

4) Even though this is a thesis in data science don’t underestimate the role played by "theory". Unless your challenge is explicitly not requiring it, your work should be linked explicitly to some concept you want to test. Descriptive analysis is an essential first step but making “sense” of the data is essential for a good grade. This is impossible without some "theory". Think of what the data-generating process looks like. If you can’t model this process explicitly think about which method is a good proxy.

5) Research is non-linear. This means you will not see progress sometimes - reading and looking at the data more will confuse you more or you will simply not have an idea. As long as you stay focused on the question and keep working on it an idea should come up. Don’t be impatient - creativity requires the combination of a lot of information. Get out of the non-linearity by doing simple data analysis first and asking yourself why you see the patterns in the data that you see.

6) Working in teams is difficult. There will always be different skill levels. It is fine to specialize. But at the very end of the project every member of the team needs to understand the entire project. If you have problems inside the team please reach out to Hannes or your supervisor.