Thesis Proposal Data Science & Society

Version F2023 – July 2023

PLEASE MAKE SURE TO READ THE THESIS QUICK START GUIDE AS WELL

0. Objectives of the Master thesis

With the Master Thesis project “Data Science in Action”, students of the Data Science & Society (DSS) program demonstrate their mastery of the data science methodology (cf. the Edison Data Science Framework[1]). The core of a DSS thesis is a **machine learning approach** (including deep neural networks) to data science, focused on exploring how existing or adapted features and algorithms contribute to **regression or classification problems**. Different algorithms are quantitatively contrasted in combination with multiple feature sets or pre-processing steps to arrive at the best predictive model, and model generalization is examined on hold-out test set(s). Students interpret the models and explore the error patterns to discuss the scientific and societal impact of their work.

For their projects, students use the R and/or Python programming language with accompanying libraries in an appropriate and correct manner. They are expected to approach the data scientific problems and questions pertaining to their project with curiosity, creativity, and as analytical thinkers. Students are required to translate complex and often extensive practical requirements (for instance, those of a commercial or governmental organization, or a research institution) into a work plan for developing, improving, or extending a data science solution. The proposed solution will support specific decision making and problem-solving processes and generalize to other, similar contexts and new data.

Using an existing data set approved by their supervisor, students identify a substantive research question that can be addressed using the selected large data set(s). In order to formulate an appropriate research strategy, students will produce a project definition that outlines the research goal and actively develop in-depth knowledge about existing solutions for the specific application area that will be discussed in the theoretical background for their thesis. Students are supported by experts in the domain provided by the data set owner (internal or external supervisor) and are advised to build on their prior expertise in a particular domain (e.g., their Bachelor studies) as much as possible.

The first stage in the project should be a well-crafted individual thesis proposal that provides the evaluating staff members with a clear view on the feasibility of the project. The thesis proposal is presented both in writing and orally during a presentation session organized by the evaluating staff members. If the thesis proposal and its presentation are successful (receive a “pass”), students continue with the thesis project. The end-product of the Master Thesis Data Science in Action (DSiA) project is the Master thesis.

[1] https://edison-project.eu/edison/edison-data-science-framework-edsf/

1. General

The thesis project proposal and proposal presentation jointly form a summative assessment in the Master Thesis course and determine the outcome of a Go/NoGo decision for the project in question. The goal of the thesis proposal is to provide a roadmap to the final thesis submission that can be evaluated on its proposed data science methodology, its scientific and societal relevance, its novelty, its feasibility, and its planned experimental rigor.

The project proposal consists of a well written document (1250 ± 20% words, excluding the title page, references, tables, and figures). The proposals should be written in correct Academic English and adhere to the APA7 or IEEE Style. Proposals with spelling, grammar or style mistakes will not be evaluated; instead, you will be asked to resubmit a corrected version. An Overleaf template for the Thesis Proposal (recommended) is available. A separate template is available for the eventual full thesis. If you do not use the Overleaf template, make sure that the title page displays provide your full name, email address, thesis cohort, the name of your internal supervisor and the contact information of your external supervisor, if applicable. Submit your proposal as a pdf file in the Canvas Assignment.

Your first deliverable will be a **draft version** of your thesis proposal. You will receive a formative assessment of this draft version from your supervisor, as well as peer reviews from other students. It is expected that you make use of the feedback on the draft proposal for your final thesis proposal and for shaping the content of your Thesis Proposal Presentation.

Your **final proposal** will be evaluated by your supervisor of the project and, if necessary, a second reader. You will receive a summative evaluation of the proposal. Thesis proposals that receive a NoGo (fail) can be submitted as a resit for this assignment, but this will not change the deadlines for the final thesis product. Note that your project plan may need to be adapted as you learn more about the data. This is fine as long as your overall goal (the task you are addressing, the data set you are using, the methods you are using) remains generally the same. Should your project change in a major way from what you initially proposed, you need to get renewed approval from your supervisor.

During your **Thesis Proposal Presentation**, you present your proposal to your peers and answer questions that may arise from your project presentation. You will need to pass both the written proposal and the proposal presentation to get a "Go" to continue with your thesis.

2. Avoiding Plagiarism

As with all assignments, you have to make sure that you do not commit plagiarism. Plagiarism is considered a serious case of fraud that, when suspected, will be reported to the Examination Board. Committing fraud can have serious consequences. At the minimum, when fraud is established by the Examination Board, the assignment is declared invalid and, in the case of a thesis, a new thesis will have to be written. Please see Article 16 of the Rules and Guidelines for TSHD (see below) for the procedure and sections in case of fraud. Note that TiU defines plagiarism as: "Using parts of a text written by someone else, or the reasoning or ideas of others, for a thesis or other assignment, without due acknowledgement." (Source: https://www.tilburguniversity.edu/students/studying/regulations/fraud/whatisplagiarism – this text contains a more elaborate explanation of what is plagiarism).

Note: Specific up-to-date guidance on the use of chatGPT and other AI solutions is presented in the Canvas course for the thesis.

2.1. Overlap Detection – Feedback

To prevent accidental plagiarism, we want to ensure that students can gain experience with the overlap detection mechanism implemented by TurnItIn. TurnItIn is one of the tools that supervisors use to assist them in detecting potential cases of academic fraud. Please note that establishing fraud is a decision that is always made by the Examination Board of the School, not just by the TurnItIn algorithm. To provide you with this experience, the Canvas assignment for the *Draft* Thesis Proposals will give complete TurnItIn feedback to students as well as supervisors, so that both can learn from the feedback provided by this system. In addition, this Canvas assignment is open for multiple resubmissions, so that students get the opportunity to repair a draft proposal with regard to overlap before submitting their full proposal.

Useful Resources:
- TSHD Education and Examination Regulations (EER), including the Rules and Regulations: https://www.tilburguniversity.edu/students/studying/regulations/eer/humanities
- What is plagiarism? https://www.tilburguniversity.edu/students/studying/regulations/fraud/whatisplagiarism

3. Outline and Contents

In the following subsections, a description of the general contents of the sections that will be assessed with the Assessment Rubric (see chapter 6) is laid out. This is a general description and should not be thought of as a complete "recipe" for the proposal. Your proposal can differ in order or grouping from this outline, as needed.

3.1. Project Definition, Motivation & Relevance, Research Question

Provide a clear description of context of your thesis project, including a problem statement. Briefly explain why this problem is worth addressing, both from a societal and scientific point of view. Make sure that the problem you address has not been solved already.
Next, outline the concrete, overarching Research Question(s) that the thesis project will answer. Good examples:

- Can capsule networks improve the accuracy of MRI-based brain tumor segmentation, in comparison to U-Nets?
    - This is a very specific research question, and it is immediately obvious what will be done in this thesis.
- Which machine learning model performs best in forecasting electric vehicle supply equipment availability?
    - This is a more general research question, and requires more detail outlined in the research strategy on which models will be compared/considered.

You can sketch briefly what will be contributed to the literature by answering these research questions. Avoid very general statements ("is it feasible to…") but try to formulate concrete research questions. The RQ should follow logically from the problem statement.

3.2. Literature review (initial) & State of the Art

Provide a summary of what is known in the scientific literature about this problem. This should be based on at least 5 relevant recent sources and, if appropriate, some more classical sources. These recent sources need to satisfy the following requirements: (1) recency (published in the last 5 years), (2) quality (published in scientific peer-reviewed journals or conference proceedings), and (3) usefulness (they should help you frame the theoretical background of your project). At this stage of your project, a full literature review is not expected but it will be expected by the time of the final thesis.

From the literature, you should extract what the current State of the Art (SOTA) is for the prediction task at hand. This will be your concrete baseline against which to compare your models. Especially for publicly available datasets, there might be specific leaderboards or competitions that outline the SOTA. If you are using a new dataset, specify what would be reasonable expectation from related work when defining your state-of-the-art performance expectation. Describe the contribution to the literature this project will make.

3.3. Dataset Description

Describe the dataset(s) that you will use in your project (size, format, accessibility, class (im)balance or distribution). Provide rationale as to why you are choosing these data. **If, at the point of proposal submission you do not yet have your complete dataset (e.g., in a project with an external partner) there is a very real risk the project might fall through, and students are advised to look for a backup solution.**

3.4. Research Strategy for answering the Research Question(s)

In your research strategy, you outline how you will approach the prediction problem outlined in your RQ, based on the state of the art as outlined in the Literature Review.

If you look at the evaluation rubric for the full thesis, it includes these elements that are essential for a DSS data science thesis:

- Model optimization, evaluation, and comparison
- Error analysis / Disparate Impact
- Out-of-sample generalization

Usually, the following elements are also included in a good research strategy:

- Feature engineering and/or selection (this could include data augmentation)
- Sampling, Stratification, and methods for dealing with Class Imbalance

You can divide your Research Strategy into sub-questions that address each of these elements. Consider the following content suggestions for structuring this section:

Feature engineering and/or selection (this could include data augmentation)

- Are there separable feature sets in the dataset(s)? Will additional features be generated or joined from other datasets? Which different methods will be contrasted for feature selection/ranking or model interpretability?

Sampling, Stratification, and methods for dealing with Class Imbalance

- Does the dataset/your target variable contain large class imbalances/non-normal distributions/influential outliers?
- Are there hierarchical levels in the dataset that should be used when splitting the data to avoid data leakage (i.e., repeated measures from the same individual)
- Which different methods will be contrasted to deal with class imbalance (e.g., under-/oversampling, synthetic data generation, using class weights) or
- How will you stratify your data splits to promote similarity between splits/folds? Besides the target variable, are there other important groups that should be equally represented in train/val/test data splits?

Model optimization, evaluation, and comparison

- Based on the literature review, which models should/will be considered?
- Are there any particular hyperparameter settings or model modifications that will be contrasted in the model optimization step?
- Which evaluation metric is best suited for the current dataset given its potential class imbalance/skewed distribution?
- Which method will be used for model comparison? For example, will the modelling strategy include resampling of the data to provide error estimates on the metrics? Are there various separate validation datasets across which performance metrics could be combined?

Error analysis / Disparate Impact

- How will the error patterns be interpreted (e.g. confusion matrix, residuals plot)?
- How will the performance and potential impact of the models be analyzed with respect to (protected) groups in the data? For example, as part of the model optimization (loss function), or as contrasts across groups made from the overall error patterns?

Out-of-sample generalization

- When the model comparison step has been completed, which models will be considered for an out-of-sample generalization test on a held-out or separate test set?
    - E.g. top-3, or best model configuration per algorithm/architecture

You can write a short motivation leading up to a subRQ, for example: “previous research has shown that a larger proportion of men failed in X compared to women. Therefore, model performance and error analysis will also be split according to gender”.

4. Milestones and Plan

Sketch out what you think will be the major intermediate milestones that you will need to achieve. Give a general idea of your planning.

5. Assessment Details:

Relationship to program learning outcomes

The Thesis Proposal assignment relates all the learning goals for the thesis, but specifically to the ILO KU1-2 and MJ2 from the DSS Program:
Knowledge and understanding (KU):

1. Students of the program: “Have broad knowledge and understanding of data science theories, methods, and techniques concerning data from socially relevant domains”.
2. Students of the program: “Are able to formulate novel ways of producing and processing information with the help of data analytics using existing knowledge in socially relevant domains”.

Making Judgment (MJ):

3. Students of the program: “judge the appropriateness of use for statistical and coding techniques employed in data analysis for a specific domain.

Learning goals (for the thesis proposal)

After finishing this assignment, the student can:

1) Motivation, Relevance, Research Question
   a) Illustrate the scientific and societal relevance of the thesis research goal
   b) Formulate an empirical research question that aligns with the research goal
2) Literature, State of the Art, Contribution
   a) Summarise existing literature of methods and results applied to a particular data science problem or analysis (research goal).
   b) Identify or define a relevant baseline for the research goal
3) Research Strategy
   a) Formulate a clear and specific research strategy based on identified gaps in literature that lead to solving the research goal.
   b) Organise the proposed work in a logical and feasible research strategy, with the help of sub-questions.
   c) Argue why chosen data science method(s) is/are most appropriate, in contrast to other methods, to approach RQs
   d) Illustrate how model comparison, error analysis, and out-of-sample generalization will be implemented in the data science approach of the thesis
4) Form and Presentation
   a) Explain the research goal using their own words without relying on quotations
   b) Relate the literature, the research question(s), and methods that make up the thesis proposal in a coherent structure

6. Assessment Rubric for the thesis proposal (pass/fail)

Item	Sufficient (Pass)	Insufficient (Fail)
Motivation, Relevance, Research Question	Proposal clearly presents societal and scientific relevance of the project’s research goal Proposal features an empirical and feasible data science research question that aligns with the research goal	Proposal does not clearly present scientific or societal relevance of the project’s research goal Proposal does not feature an empirical and feasible data science research question that aligns with the research goal
Literature, State of the Art, Contribution	Proposal provides succinct summary of at least 5 relevant sources from scientific literature Proposal clarifies SOTA and highlights gap(s) in literature that motivate the research question(s) Proposal illustrates the expected contribution to the literature of the proposed work	Proposal omits or does not clearly illustrate relevance of cited literature Cited literature does not clarify SOTA or research question(s) do not address a gap in the literature Proposal does not address expected contribution to the literature
Research Strategy	Proposal motivates choice of data science methods in light of the dataset and literature review Proposal clearly outlines a scientifically rigorous evaluation strategy to answer the RQ(s), including: - model comparison - analysis of errors - out-of-sample evaluation - (if appropriate) disparate impact/bias across classes/groups	Data science methods chosen are not appropriate for a data science thesis or not appropriate given the literature and dataset Research Strategy is unspecific, unattainable, or does not answer RQ. Research Strategy does not include steps addressing: - model comparison - analysis of errors - out-of-sample evaluation
Form, Structure & Presentation	Proposal conforms to guidelines regarding formatting, style, sections, and length, including citations and bibliography Proposal is written in a cohesive and structurally sound manner. Information content is placed in the appropriate places	Proposal does not conform to the necessary formatting of a TSHD Master thesis, including approved citation styles Structure of the proposal is incohesive or information content is placed in inappropriate places
Originality of writing	Proposal makes proper uses of citation, quotation and paraphrasing to avoid plagiarism	Thesis contains improperly paraphrased material, improper/incomplete citations, or improper attribution of direct quotations

All items in the proposal rubric must receive a Pass mark for the Thesis Proposal to receive a Pass mark. In case of a fail, a resit opportunity will be scheduled by the supervisor.