Q1.

UBC Master of Data Science (Vancouver option)
Capstone Proposal Form
2024-2025

For help or further information, please contact Tiffany Timbers and Joel Ostblom at tiffany.timbers@stat.ubc.ca and joel.ostblom@ubc.ca

Q31. **Partner organization information:**

Q2. **Your name:**

Tiffany Timbers

Q32. **Your position in the organization:**

Associate Professor of Teaching, Department of Statistics & Co-Director, Master of Data Science Program (Vancouver)

Q3. **Your email:**

tiffany.timbers@stat.ubc.ca

Q4. **Organization name:**

FixML team, University of British Columbia

Q41. **Organization website:**

This question was not displayed to the respondent.

Q25. **Organization phone:**

[redacted]

Q20. **Organization address:**

Department of Statistics, Faculty of Science, Room 3152, Earth Sciences Building, 2207 Main Mall University of British Columbia, Vancouver, BC Canada V6T 1Z4

Q5. **About your organization:**

Please write a short description about your organization. This description should be a 2-4 sentence, high-level overview.

The FixML team is led by Dr. Tiffany Timbers, an Associate Professor of Teaching in the Department of Statistics, and includes collaborators from the University of Madison-Wisconsin (Dr. Simon Goring), the University of Toronto (Dr. Rohan Alexander), and several MDS-V alumni (John Shiu, Orix Au Yeung, Tony Shum, and Yingzi Jin). The aim of the team is to communicate about and build tools to support best practices for creating robust and trustworthy applied machine learning projects. The team's first contribution toward this is the development of the fixml software package for context-aware evaluations of machine learning project code bases using a checklist-based approach.

Q7. **Short & snappy project title:**

Checklists and LLM prompts for efficient and effective test creation in data analysis

Q6. **Description of the problem/question:**

Please write a description about the data problem/question your organization is challenged with. The more details provided here, the better understanding we will have of your data problem/question. Please use accessible language without domain-specific jargon. Limit your response to 500 words.

Note: If your problem/question leads to multiple projects, please submit a proposal for each project.

Checklists have been shown to decrease errors in safety critical systems (Gawande 2010), and the use of a reproducibility checklist at the Machine Learning NeurIPS 2019 conference led to an increase in the percentage of authors submitting the code for their work (from 50% to 75%; Pineau et al. 2021). Thus, in an effort to make applied machine learning software more trustworthy by increasing its robustness, we created a general and robust checklist for software tests for applied machine learning code. This checklist includes tests for data presence, quality and ingestion at the beginning of the analysis, the model fitting and evaluation, as well as tests for the artifacts (presence and quality) which are created by the analysis. Our checklist can be used by data scientists and machine learning engineers to guide the manual writing of tests. It can also act as a source for engineering large-language model (LLM) prompts that act as reliable starting points for evaluating the quality of existing applied machine learning code, as well as for engineering reproducible test data and software tests themselves for each item on the checklist. The former has been demonstrated as a proof of concept in last year's MDS capstone project by the creation and evaluation of the fixml software package (https://ubc-mds.github.io/fixml/report/final_report.html). The fixml software package additionally has a generate function that uses LLM's to generate very general test function specifications and docstrings for missing test cases using the checklist. In this upcoming year's MDS capstone project, we would like to assess whether we can improve the specificity of the fixml LLM generated tests for a given machine learning project through prompt engineering (e.g., providing additional context for the machine learning project codebase to fixml) and experimentation on an existing benchmark codebase. Experimentation on the benchmark codebase will allow us to evaluate which prompts, LLM's themselves and LLM settings work best in regards to accuracy and consistency of the generated test suites.

Q38. **Problem/question impact:**

Describe how this problem/question impacts your organization. Limit your response to 250 words.

Improving the LLM powered test suite generated feature to fixml to be more specific for a given machine learning project would have positive impacts on data science educators and students, machine learning practitioners, as well as the quality of applied machine learning projects created in the future. Data science educators and students would benefit from this fixml feature, as educators could use it to demonstrate best practices for creating robust and trustworthy software for machine learning applications. Students could also use this feature in their course projects. Machine learning practitioners could use this feature of fixml to decrease the amount of time they spend writing software tests from scratch, and better focus their time on code review and code improvements. Finally, with new data science students, and seasoned machine learning practitioners regularly using this tool, more machine learning project code will become more robust and trustworthy - an critically important goal as machine learning applications are now used broadly in many ways that touch human lives. Evaluation of any improvements to fixml's LLM powered test suite generate feature on a benchmark is critical. None of the impacts described will be possible without thorough evaluation of their impacts on accuracy and consistency.

Q33. Problem/question keywords:

Please select the problem/question keywords that best describe the work required for this project.

☐ Descriptive (e.g., you wish to summarize characteristics of a dataset)

☐ Exploratory (e.g., you wish to generate new hypotheses to test in the future)

☑ Inferential/explanatory (e.g., you wish to try to explain patterns or relationships in your dataset and generalize them to the larger population)

☑ Predictive/machine learning (e.g., you wish to be able to predict future categories or values for an individual)

☐ Causal (e.g., you wish to establish a directional effect of one, or more, variables on another)

☐ Other

Q8. Summary of available data sources:

Please write a short summary description of the data that will be provided to the students for the project. This description should be a 2-4 sentence, high-level overview. More detailed questions will follow.

For the benchmark, the students will be given a toy/example machine learning project, for which many mutant versions have been made. Each mutant project will contain a error in the code base that corresponds to an item on the applied machine learning test checklist (https://github.com/FixML/test-checklist-for-machine-learning).

Q28. Detailed dataset description:

Describe the subjects of the dataset (e.g., sensitive data about people, data about places and objects, synthetically generated data) as well as the variables/features recorded for the dataset subjects. An annotated snapshot of the dataset could be used to answer this question.

The toy/example machine learning project, for which many mutant versions have been made, is available in public a GitHub repository (https://github.com/FixML/breast_cancer_predictor_py). The toy/example machine learning project is a reproducible machine learning pipeline for a classification model which can use breast cancer tumour image measurements to predict whether a newly discovered breast cancer tumour is benign (i.e., is not harmful and does not require treatment) or malignant (i.e., is harmful and requires treatment intervention). For each applied machine learning test checklist item (https://github.com/FixML/test-checklist-for-machine-learning), there exists a subfolder in that GitHub repository with a copy of the project, where an error has been injected into the code that leads to the requirement for that item to not be satisfied. A high-quality test for that test checklist item should fail for that mutant version of the code-base, but not for the original, or others. There exists a `mutants.json` file in the root of the toy/example machine learning project repository that maps each applied machine learning test checklist item to the folder containing the mutant code base for that item, as well as the path to the mutated file(s), line(s) of the code mutation, and the original and mutated code.

Q27. Dataset sensitive attributes:

For any data sources that will be used for the project, please describe any human and other sensitive attributes. If no attributes are sensitive, please write "Not applicable".

Not applicable

Q29. **Dataset provenance:**

For any data sources that will be used for the project, please describe the methods of how the data were collected (e.g., API, surveys, scraped or crawled, artificially generated, etc).

The toy/example machine learning project, and it's mutants were created by hand.

Q30. **Dataset owners:**

For any data sources that will be used for the project, please provide the name, affiliation and contact information for the data owners.

Tiffany Timbers, Associate Professor of Teaching, Department of Statistics, University of British Columbia tiffany.timbers@stat.ubc.ca 604-803-4962

Q9. **Data product:**

Please write a description about the data product that will help your organization overcome the described problem, or answer the described question. Limit your response to 500 words.

Examples of possible data products include:

1. A dashboard, such as a Shiny or Dash app, to explore an aspect of your data
2. An R or Python package with documentation to simplify an analysis
3. A data pipeline that includes some data science model
4. A technical report outlining student findings

The minimal data product would be a reproducible report documenting the experimentation with the fixml software packages test generation feature, with regards to prompt engineering, LLM choice, and LLM settings on the benchmark toy/example machine learning project. The experimentation should report on how these affect consistency and accuracy of the generated test suite. Stretch goals could include modifying the fixml software package to more easily incorporate codebase context into the generate function, setting the fixml generate feature defaults to the ones found to work best from the experimentation (as well as updating the documentation of the fixml software to reflect all this).

Q37. **Data product impact:**

Describe how such a data product would positively impact your organization. Limit your response to 250 words.

The improvement of the LLM powered test generation feature to the fixml software, and an evaluation of it's capabilities and limitations on a benchmark, would be the basis for an manuscript so that the work could be disseminated and shared via scholarly publication. The feature would also be merged into the fixml codebase and available open source on GitHub, and for user download via PyPI.

Q34. **Data product keywords:**

Please select any data product keywords below that are relevant to this project.

☐ Hypothesis testing

☑ Inferential modelling

☐ Bayesian modelling

☐ Supervised machine learning

☐ Unsupervised machine learning/Clustering

☐ Neural networks/deep learning

☑ Generative artificial intelligence

☐ Time series analysis

☐ Spatial analysis

☐ Survival analysis

☐ Natural language processing

☐ Computer vision

☐ Data visualization

☐ Dashboard

☐ Data engineering

☐ Big data

☐ Other

Q36. **Additional computing resources:**

Describe the computing resources needed to complete the proposed project. If they are beyond the standard laptop students will be equipped with (details here), describe how your organization will provide these resources to the students. If no additional computing resources are needed beyond that of the students' standard laptop, please write "Not applicable".

Students will require access to tokens for interacting with LLM API's. We will provide some funding for this.

Q26. **Are you an UBC-affiliated sponsor?**

Sponsors who hold any type of UBC appointment (e.g. clinicians with teaching appointments) or startups that have strong relationships to UBC (e.g. currently "incubating" within e@UBC, Hatch, etc.; or with directors who are UBC faculty/staff).

● Yes

○ No

Q10. **Confidentiality and IP for non-UBC capstone partners:**

Projects which require agreements to be put in place to protect confidential background information and/or to define ownership of new project intellectual property can be bound by the UBC-provided Mutual non-disclosure agreement (NDA) and/or IP assignment agreement. Details and templates of these can be found here: https://ubc-mds.github.io/capstone/guide-to-mutual-nda-ip/

We strongly recommend that partners show the UBC template documents to their legal counsel and get their agreement to use these documents before submitting the capstone proposal. We cannot sign alternate agreements, nor amend our agreements in any way.

This question was not displayed to the respondent.

Q16. **This project will require the UBC mutual non-disclosure agreement (NDA)**

This question was not displayed to the respondent.

Q17. **This project will require the UBC IP assignment agreement**

This question was not displayed to the respondent.

Q24. **I confirm that I have shown the UBC template documents to our legal counsel and gotten their agreement to use these documents.**

This question was not displayed to the respondent.

Q18. **Confidentiality for UBC capstone partners:**

For projects that have confidentiality requirement that are sponsored by UBC faculty or by companies (startups) that are closely related to or incubated by UBC, any new inventions will be handled by the UBC’s University-Industry Liaison Office - the UILO. More details can be found here: https://ubc-mds.github.io/capstone/guide-to-mutual-nda-ip/

Q19. **This project will require the handling of confidentiality and IP through UBC's UILO**

○ Yes

● No

Q22. **Additional security requirements:**

Is there anything else that students will be required to do to work on this project (e.g., complete a background check, etc)? If there are no additional security requirements, please write "Not applicable".

Not applicable

Q23. **Student communication of work:**

We understand that you may require some restrictions to be put in place, but we also would like for our students to have some freedom to talk about the work they've done, particularly when applying for jobs. We want our students to know about these restrictions up-front so that they can make an informed decision about the projects they choose.

How do you anticipate students will be able to share aspects of their work with others? Examples can include listing the project on their resume, discussing it in a private job interview, writing a blog post about the experience, open-sourcing the code they write, etc.

This project is an open source and open science project. The students will be able to share their work in any way they see fit.

Q11. **Potential conflicts of interests:**

Do you have any potential conflicts of interest to declare? For example, if a current MDS student or family member is involved with your organization on a professional or personal level, this should be declared along with a short explanation. These situations are generally not problematic, but we prefer to disclose them to the students before they rank the projects. If there are no conflicts of interest, please write "Not applicable".

Dr. Simon Goring, and myself, Tiffany Timbers, are currently instructors in the MDS-V program. I am also currently a co-director of that program.

Q12. **Space for students at your organization:**

Do you have space available for students to work on site?

No

Q35. **Additional market research questions:**

Q13. Do you anticipate having data scientist job opening(s) after the project?

In the past we have had funding to hire students for some part-time work after capstone.

Q14. How did you hear about the UBC MDS Capstone program?

I work for the program!

Location Data

Location: [(49.4645, -122.84)]

Source: GeoIP Estimation

Vernon
Merritt
Kelowna
Vancouver Island
Vancouver
Chilliwack
Nanaimo
Abbotsford
Bellingham
Victoria