MATH70103: Unstructured Data Analysis – MLDS
Autumn 2024-2025

Dr. Anthea Monod
GTAs: Paul Lezeau, Yueqi Cao

Final Project

Congratulations on making it to the end of this module on unstructured data analysis! Now is the time to put everything that you've learned over the past 10 weeks into real, hands-on practice.

For this project, you may either:

(i) Select your own imaging, network, or text dataset(s) and apply the techniques that you've learned from the relevant topic via the lecture notes, videos, and homeworks to your chosen dataset(s); or

(ii) Explore further some related concepts that we have touched on in discussions in office hours and live sessions that were not formally covered in the material (e.g., graph neural networks, graph convolutional networks, adversarial machine learning, persistent homology, etc.)

Project Criteria and Mark Allocation

1. **Data Selection (20/100 marks):** You may choose any data type and real dataset(s) you wish, from work, a hobby, or the internet. You may choose several datasets but please ensure that each dataset does not exceed 100 MB in size in order to be able to submit your dataset along with your project report. For a project of the first type (i) above, simulated datasets will not be accepted, but may be used for a project of the second type (ii) (e.g., exploring persistent homology).

   You will be given credit for originality and complexity of your dataset (10/100 marks). For example, if you are working with a digitized image of an old black-and-white photograph from your family valuables, you will get more credit for originality and complexity rather than using a famous benchmarking dataset from a publicized challenge. A series of datasets from an online challenge will get more credit for originality and complexity rather than a single, pre-cleaned dataset made available online for a project of a similar nature from another similar course/module.

   Please be sure to properly source your data and give a thorough description of your data (10/100 marks).

   Please use good judgment and common sense when selecting your dataset. Keep in mind that some datasets will not be accepted, e.g., the famous "Lena" photograph, although classically used in image processing and analysis, is now considered inappropriate. Society of Industrial and Applied Mathematics (SIAM) journals will not accept any submissions involving the Lena photograph to demonstrate proposed methodology, and neither will we. Similarly, if you are choosing a text dataset in a language other than English, please be sure that you are sufficiently familiar with the structure and grammar of that language in order to be able to assess the output of your analyses. Please describe any peculiarities specific to that language in your report, e.g., in Hebrew, there is no separate word for the article "the" or the conjunction "that" (the noun or connecting word is simply preceded by the Hebrew letters "heh" and "shin," respectively, but the situation is tricky because there are also words beginning with these letters, so they are not always indicative of a stop word).

   It is your full responsibility to ensure that you have permission to use your chosen dataset. We will not take any responsibility for permission-related oversights and their consequences. We will also not assist with setting up any data sharing agreements. For this reason, if you choose to use a dataset from online, we very strongly recommend that you use open access data.

2. **Problem Statement or Topic Background and Overview (15/100 marks):** For a project of type (i) above, given your chosen dataset and its characteristics, please be sure to clearly and concisely state the problem you aim to address or the area you wish to study. Depending on the complexity of your chosen data, this may be a single task, but it may entail more. For example, if you have a very large and challenging network, you may choose to do only community detection. If you have a very noisy image, we have seen that the tasks of image denoising, edge detection, and segmentation are all closely related so you may try to take on all three tasks. The goal here is to choose task(s) of suitable

complexity for the data you have chosen; there should be a clear desire to "get through the thick" of your data and uncover what you can from them, rather than just choose an easy task to "get it over and done with."

For a project of type (ii) above, this criterion corresponds to introducing and providing the background on the new topic area you would like to explore further.

3. **Description and Justification of Methods and Analysis (20/100 marks):** For a project of type (i) above, reflect on the most appropriate techniques and approaches from what we've seen in the module to apply to your data, given the problem you want to study. For any technique you use, be sure to motivate its use properly and justify it as an appropriate approach.

   If you used a benchmarking or challenger dataset from the internet, you will also need to do a literature review on other approaches used and their results on these same data. Specifically, if you are straightforwardly applying one of the methods we've seen in the module to a famous dataset, it's extremely likely that this same analysis has been done previously by others. You should cite these works and summarize the results that others have achieved.

   For a project of type (ii) above, discuss the appropriateness of the technique to the dataset you have chosen to work on. Alternatively, if you are using simulated data, describe here how you plan to simulate data that reveals the most insight to the technique you are interested in studying.

4. **Interpretation and Reflection on Output (20/100 marks):** Describe and interpret your findings from having applied your chosen techniques to your dataset. Are the results what you expected to see? Why or why not? Did you try several techniques? Which ones worked better, why, or why not? Are there any characteristics of the data that made one approach better than others?

   If you used a benchmarking or challenger dataset from the internet, you will also need to do a comparison on your results to those established by others who worked on the same data. How does your implementation compare to theirs and what can you say about that?

5. **Report Presentation and Clarity (15/100 marks):** Please pay close attention to detail as well as overall structure and format of your report. Please use appropriately-titled sections and subsections, references, clearly-labeled mathematical statements (e.g., definitions, propositions, etc.) where appropriate. Writing style and grammatical correctness are important.

6. **Code Presentation and Clarity (10/100 marks):** Similarly, a well-organized Jupyter notebook with clear code comments and a description of all functions and purposes for cells will be very important. Open and transparent research and reproducibility are becoming increasingly important; in fact, the top statistics and data science journals are now requiring code to be submitted with manuscript submissions and dedicated reproducibility editors are being engaged on editorial boards to ensure reproducibility of results. Please give clear instructions on how to run your code (e.g., in terms of a README markdown on your GitHub repository). Clearly label which functions and cells produced which figures in your report. Please also give clear indications on the hardware and software used in your analyses (what resources did you use; did you run the code on your laptop, a desktop, a cluster? Did you run parallel jobs, using e.g., SLURM or OpenPBS? How many nodes and cores, CPU/GPU, memory per CPU?) and an indication of runtime.

General Advice

- There will be no new content covered in the final week (Week 10) of the module. Instead, we will be available for project advice and discussions.
- The office hours leading up to and including Week 10, as well as Week 10's live session, are a good opportunity for you to get feedback from the teaching team on the suitability/remit of your chosen dataset and the appropriateness of your proposed problem statement. Keep in mind, though, that depending on the data you are interested in, we may not be domain experts. This is especially true for datasets coming from your work or hobbies, you are the expert in this area and not us! We are very interested and happy to learn about new application areas, so we will be happy to discuss with you and give you advice on scientific approaches but cannot provide domain specific advice.

- You are expected to carry out the actual analyses and write the report individually, i.e., without support from the teaching team. You are very welcome to discuss and work together with your peers, but your project report should be your own work.
- The teaching team will not look at or help you with your code, or read any drafts of your project report before submission.
- We hope you will think of this project holistically and view it as an opportunity to learn something new that you have a hand in designing and deciding, rather than a "box-ticking" exercise. While the criteria above are required components of your final report and marks will be allocated according to those criteria, the overall structure and approach is quite a good simulation of how real research problems look like and how to approach real data-centric problems, both in academia and industry. This is ultimately what you will have to encounter, certainly in the near future with the research project component to the program, and most likely in your future professional and career plans upon completion of the program as well.

Information and Technical Instructions

- This final project is worth 40% of your final score for the module. It is due by **Friday 3 January 2025 at 15:00 UK**.
- This final project is designed to be completable in 2 weeks and you are expected to spend 20-25 hours on it. You are given more time than this to submit the project but you are not expected to spend all that time working on the project. The extra time and notice is to allow for access to technical support for submitting your work and, most importantly, scheduling/flexibility around the holidays, professional/personal commitments, as well as assessments for other modules you are simultaneously taking.
- Please submit your project as a PDF document (max 12 pages long, font size 11). Please also submit your dataset and accompanying Python code as a Jupyter notebook with clear code comments; your code does not count towards the 12-page limit on your project submission. Your code and dataset should be submitted separatels. Please give appropriate file names to your project submissions, e.g., `UDA_FinalProject_Lastname.pdf`.
- The preferred method for submitting your data and code is via a GitHub repository with the link included in your report, but if you don't know Git (not required to create a repository, but makes things a lot easier) or don't know how to create a GitHub repository, you may also share data and code as you did for Homeworks 1 and 2 on Images and Networks. If you don't know git and are submitting your data and code by a drop box, and if your project works on a very large dataset, please submit only a sample set of the original dataset for testing.
- Please submit code written only in Python for the coding requirements of the project. You may use whichever packages you wish to complete the assignment, however, please clearly indicate which ones you have used in your notebook.
- Please include a statement indicating that you have worked independently.
- Any questions on the project should be asked only during contact time during Week 10 and on the Ed Discussion forum. In line with the previous instruction, please use your best judgment when posting/responding to questions about the project and especially using content from posts on Ed Discussion. While we certainly encourage discussion on the content of the module amongst yourselves, we would like for the project reports to be written up entirely independently and reflect your own understanding and interpretation of the content or topic at hand.
- While you are welcome to communicate with each other on the Ed Discussion forum right up until the project due date, the teaching team will only be available to respond to questions about the project until the end of term on Friday, 6 December, 2024 at 17:00 UK. The College re-opens from Thursday, 2 January, 2025 at 09:00 UK and you will be able to get technical support for help with submitting your projects from then until the due date.