HEATMAP ANALYSIS PROJECT PLAN

Mansoureh Foruzandehshahraky

Dona Joy Arimboor

Aniket Varbude

Stephen Maher

March 15, 2024

Executive Summary

The goal for the Heatmap Analysis Project (project) is to develop and deliver tools to the client, embedded in workflows, which simplify the process of analysing a general class of data comprising event sequences (called traces). These deliverables are workflow tools for exploratory data analysis and other statistical analysis (and visualisations of these), and modelling and model output visualisation of the event sequence data. The tools will provide "proof-of-concept" analysis for the client and will enable existing research into event sequences and provide a platform and guidance for further research and tool and workflow development. Time and cost for the Project are estimated at 1,000 person hours over twelve weeks, at $82,500. The value proposition for the client is for a marginal change in their existing resourcing plan, they receive of reporting on, and prototyped tools and workflows, forming a base their future research and development requirements.

1. Project Background and Description

Project Background

The client (sponsor) is a research institution seeking to identify techniques available to facilitate the meaningful analyses of a large dataset. This research is important as if successful, it will provide the sponsor with a workflow for the statistical analysis, visualisation, and modelling of their data. The sponsor derives a benefit, for the cost of mentor and sponsor time, from effectively managing their internal resources through outsourcing, of a proof-of-concept study into methods for researching directions for their data set.

For this project, the dataset comprises event sequence data. The data is of a general class where records (defined as "traces), comprise an ordered sequence of events (where an event is defined as a token) which are represented symbolically. The tokens can be viewed as distinct states for the trace at points in time. Of particular interest are the occurrence of "motifs" which are sub-sequences of tokens within traces which can be analysed for frequency and other factors.

These types of data are observed across a wide range of applications and there exists research which can inform analysis, modelling, and visualisation methodologies. One domain where this research is applied is in health records, while some other domain applications are network logs and website user activity. The volume of data is such that it offers real potential for analytical insight. However, analysis is complicated by data complexity and heterogeneity, which makes meaningful visualisation and effective modelling difficult. Using traditional analytical approaches, data volume and variety may result in time consuming and ineffectual analysis of the data, hence there is a need for developing tools which support analysis, visualisation, and modelling for meaningful insights.

Description

The data has been provided without context, has unknown provenance and no metadata. And while the data appears semi-synthetic, analysis and modelling can still be conducted. The sponsor has provided a subset of their dataset for analysis and has specified two key outcomes for this project.

The first key outcome is to develop methods for the statistical analysis, modelling, and visualisation of the sponsor's data and to evaluate proof-of-concept for the application of these methods. The second key outcome is to develop and implement workflows within some environment (such as R, Python or KNIME) based on preferred methods determined from the first key outcome. These workflows, along with a report on the project teams findings, comprise the deliverables to the sponsor. The sponsor has stated that the data may not have meaningful or interesting relationships, and that the data is being used to test the tools and workflows developed. Further, as the data provided is representative of a general class of problem, the tools and workflows need to be useable for other datasets within the problem class and not suited solely to the dataset provided.

Objectives

To achieve the implementing analytical, visualisation and modelling workflows, the project team has a series of intermediate objectives:

- Exploratory data and statistical analysis, basic visualisations, and associated workflow development.
- Modelling and model workflow development.
- Visualisation and visualisation workflow development.
- Workflows packaging and documentation.
- Reporting.

Further detail on each of these intermediate objectives (excluding project planning and acceptance) is provided in the following subsections. It is noted that the workflow implementation framework (such as R, Python or KNIME), will be determined as the project proceeds.

Throughout this project, there will be a continuous review of the existing literature for event sequence statistical analysis modelling, and visualisation techniques. No formal literature review is proposed as a project outcome.

Exploratory Data and Statistical Analysis (EDA) and Visualisation

This objective is for the team to explore and understand the data, with a focus on token occurrences, and token sequences and motifs. This includes a literature review and brainstorming for methods for trace analysis, modelling and visualisations, statistical analysis and preliminary visualisations and the development of workflows for these. Each discovery (such as patterns, associations, and statistical measures) will form the basis for discrete lines of workflow development in the subsequent objectives.

Model Development

Using the results from the EDA, this objective is to build proof-of-concept workflows which allow for modelling the patterns, associations, and other features in the data. The sponsor has not provided guidance on modelling approaches and work will be informed by the EDA and the literature.

Model Packaging and Documentation

The sponsor has specified model workflows as a deliverable. Moreover, these workflows need to be suitable for use with the data as a general class and the software implementation should be free. This objective is the creation of re-usable general purpose modelling workflows and supporting documentation.

Visualisation Development

Using the results from the EDA and modelling, this objective is to build proof-of-concept workflows which allow for visualisation of the patterns, associations, and other features in the data. The sponsor has not provided guidance on visualisation methods (apart from heatmaps) and work will be informed by the EDA, modelling, and the literature.

Visualisation Packaging and Documentation

The sponsor has specified visualisation workflows as a deliverable. Moreover, these workflows need to be suitable for use with the data as a general class and the software implementation should be free. This objective is the creation of re-usable general purpose visualisation workflows and supporting documentation.

Reporting and Delivery

Once the analysis has been finalised and findings determined, and the modelling and visualisation workflows have been packaged and documented, this is to be reported to the sponsor in the form of a report and packaged code and documentation.

2. Project Scope

The project is to develop proof-of-concept analytical, modelling and visualisation workflows suitable to use with the general problem class. As such, developed workflows will be laboratory prototypes and will require user familiarity with the subject matter and coding and will not be suitable for large scale deployment or use.

Inclusions

The following are in scope for the project:

- Data preparation.
- Development of analytical (statistical and model) workflows suited to use with the dataset as representative of a general class of data.
- Model interactivity limited to capacity to change code parameters, select tokens used, or equivalent.
- Development of visualisations workflows suited to use with the dataset as representative of a general class of data.
- Visualisation interactivity limited to capacity to change code parameters, select tokens used, or equivalent.
- Reporting on findings.
- Deployment of tools and workflows as a laboratory protype (local computers and including basic documentation).

Exclusions

The following are out of scope for the project:

- Data collection or cleaning.
- Results interpretation.
- Predictive or prescriptive analyses.
- Models are not “point and click” interactive.
- Visualisations are not “point and click” interactive.
- Production-ready tool deployment or full automation.
- Highly detailed documentation and use-case analysis.
- Real-time analyses.

3. Risks

This project possesses both general project risks and risks specific to the project. Understanding and proactively managing risks through using a risk management plan can positively influence the success and smooth execution of this project. The specific risks centre on the unknown provenance of the data and no metadata. This means that there is no context for the data or any results.

Managing risks as they arise will require engaging with the project sponsor. Overall, this risk management plan supports:

- Ensuring preparedness.
- Improving decisioning.
- Resource protection.
- Enhancing project quality and outcomes.
- Increasing stakeholder confidence.
- Promoting flexibility and adaptability.

Key risk factors identified for this project are detailed in the table below, along with their estimated probabilities, impact assessments and the proposed mitigation strategies. While not capturing all risks, major risks are covered. For risks not covered but realised, the mitigation strategy will be based on or be a modification of an aligned documented risk.

Risk is categorised using three levels (High, Medium, Low) to frame their relative importance. These can be considered as a 30%, 20% or 10% relative increase in time or budget, or a relative decrease in quality.

Risk	Probability	Impact	Mitigation Strategy
Scope creep	Medium	High (budget/time)	• Agree project scope with sponsors. • Clear project parameters. • Tracking activities against scope and objectives.
Time delay	Low	Med (budget/time)	• Monitor project schedule. • Understand the project lifecycle. • No resolution possible, full dependency.
Insufficient resources	Medium	High (time/quality)	• Define resources for each stage. • Assign tasks at each stage. • Track resource surpluses for re-assignment.
Operational changes	Low	Low (time/quality)	• Team communication. • Preparation for changes. • Adjust via team meetings.
Lack of clarity	Low	Med (time/quality)	• Continuously referencing requirements. • Team and sponsor agreement on requirements and objectives. • Scope clearly defined.

			• Regular team communication.
Technical: • Masked data • No data context • Data bias • Algorithm selection and tuning • Overfitting • Results accuracy • Results interpretation • Scalability • Resource intensity • Complexity	High	High (quality)	• Liaise with sponsor on data and results accuracy and interpretation. • Ensure algorithms and visualisations are effective and suit the data and results. • Use sampling to avoid overfitting. • Sponsor owns algorithm tuning (tool and workflow configuration). • Align resources to requirements. • Target simplicity.
Ethics and compliance	Low	Low (quality)	• Data masking.
Loss of team member(s)	Low	Low (time/budget/quality)	• Reassess and rescope with sponsor.
Stakeholder resistance	Low	Low (time/budget/quality)	• Proactively engage stakeholders. • Communicating project benefits. • Address concerns.

4. Budget

This section details the expected cost for the project. Cost risks stem from an increase or decrease in sponsor or mentor time.

Budget Item	Estimated Cost	Justification
Student time	$72,000	Assumed $75 p/h for four graduates for 20 hours p/w over 12 weeks.
Sponsor time	$6,000	Assumed sponsor salary and on-costs of $300 p/h for 20 hours
Mentor time	$4,500	Assumed sponsor salary and on-costs of $150 p/h for 30 hours
Software	$0	The sponsor has specified that software use in development is free to use and has no on costs.
Office supplies	$0	Provision “in natura” by the sponsor.
Total	**$82,500**	

5. Roles and Responsibilities

Project Stakeholders

This Section details the project’s stakeholders, and their role and interest in the project.

Name	Role	Interest in the project
Dr Jan Stanek	Project Sponsor	Proof of concept for re-useable applications for future research on dataset.
Prof. Georg Grossmann	Project Sponsor	Proof of concept for re-useable applications for future research on dataset.
Eric Lam	Mentor	Participate in weekly meeting to evaluate the project progress. Provide guidance and feedback on project development. Evaluate the performance of team members. Assist the team to overcome the unexpected issues that may arise during the project development.
Stephen Maher	Project Team Member	Finishing a research project for completion of a Capstone Unit.
Mansoureh Foruzandehshahraky	Project Team Member	EDA, modelling, and visualisation.
Aniket Varbude	Project Team Member	Creating tools and uncovering patterns for a masked dataset.
Dona Joy Arimboor	Project Team Member	Creating tools to visualise the patterns in the data.

Project Scope Allocation

Project Team

All project team members will collectively contribute to conducting the ongoing literature review. This provides diverse perspectives, comprehensive understanding and knowledge sharing within the team. Each group member will actively participate in all tasks, with a designated individual coordinating every task. Task supervision is as follows:

Preprocessing, EDA, Statistics and Pattern Identification – Dona Joy Arimboor

All team members will work on problems including pattern identification and data preprocessing. Dona will oversee the procedure and coordinate the workflow development,

Visualisation - Mansoureh Foruzandehshahraky

Mansoureh will oversee the visualisation component, leading the team in producing informative and significant visual representations resulting from the EDA, statistical and modelling. This includes workflow development, packaging, and documentation.

Modelling - Aniket Varbude

All team members will participate in the modelling phase with Aniket overseeing modelling. This includes workflow development, packaging, and documentation.

Documentation and Editing - Stephen Maher

To guarantee that all project materials are properly recorded, arranged, and presented in a professional and consistent manner, Stephen will oversee the editing and documentation processes.

Mentor – Eric Lam

Eric Lam servers as the mentor for the team provides guidance, support, and direction throughout the project.

Project Sponsors – Jan Stanek and Georg Grossmann

Jan and Georg serve as the assigned sponsors for this project with responsibility to define the project's goals deliverables, vet and advise on deliverables and confirm outcomes.

Communications Plan

This section provides an overview of strategies used for maintaining connectivity and information flow within the project team, and between the project team and the mentor and sponsors.

Team Communication

This includes scheduled meetings, communication, and creating procedures for dealing with updates and conversations. Coordinating team members to work together seamlessly is the goal. Multiple platforms are used for this:

- Weekly team meetings using the Microsoft Teams platform, every Monday at 7 p.m. These meetings cover task distribution, updates, and discussion of ongoing activities. Unscheduled meetings will be arranged when issues need more attention.
- Daily operational problems are resolved on the WhatsApp forum. This platform provides an immediate forum.

- All project -related information is kept on a shared SharePoint platform that is available to the group, including written materials, research papers, data, code, and analytical results.

Mentor and Sponsor Communication

There are weekly meetings scheduled with the mentor and fortnightly (or as required) with the sponsor(s) (Stephen Maher is the designated primary point of contact with the sponsors) on the Microsoft Teams platform to monitor the team's progress, to comment on progress and to provide guidance on the project. Email is also used for sponsor and mentor communication.

Plan Modifications

Procedures for handling modifications to the project plan are detailed below. Reasons for change may include the inability to complete some project objectives or supply deliverables. If the need arises, the team will first discuss any modification. If the team agrees on the need to modify the project plan, the team will prepare alternatives and the team will communicate the reasons for change and proposed variations to the sponsor and the mentor. The modifications will only be put into effect after evaluation by the sponsor and mentor and after agreement has been reached by all parties. This preserves the project integrity and guarantees transparency and expectations alignment.

Dispute Resolution

To guarantee smooth project progress, disagreements within the project team regarding any matter will be settled by open discussion and will be resolved using via a majority group decision.

6. Deliverables and Project Evaluation Criteria

This Section details the project's deliverables, a description of these and the evaluation criteria for deliverable fulfilment.

Deliverable	Description	Evaluation Criteria
EDA, Statistics and Visualisation	Conduct a literature review to identify methods and tools suited to analysing traces. Analyse data to identify patterns, trends, correlations, and other statistical measures and develop associated visualisations.	• Identified tools suited to analysing and visualising traces. • Effective use of appropriate statistical, data analysis, and visualisation methods. • Visualisation clarity, quality, and aesthetics. • Functional tools and workflow developed.
Develop Modelling Tools and Workflow	Create models and associated workflows suited to the data and ensure the efficiency and efficacy of the models.	• Functionality and performance of the models • Alignment of model capabilities with project analysis needs. • Model tools, workflow and documentation developed.
Develop Visualisation Tools and Workflow	Create visualisations and associated workflows suited to the data and ensure the efficacy of the visualisations.	• Functionality and performance of the visualisations developed.

		• Alignment of visualisations with project analysis needs. • Visualisation aesthetics. • Visualisation tools, workflow and documentation developed.
Ensure Tool and Workflow Flexibility and Repeatability	Incorporate into all tools the flexibility to handle datasets within the general problem class and allow for repeatability and consistent results.	• Robustness of tools when applied to other datasets. • Capacity for user interaction. • Consistency of results produced by the tools.
Detail project Outcomes in a project Report and Finalise Workflows	Compile all project results and findings into a report which covers the project from inception to conclusion and finalise tools, workflows, and documentation.	• Detailed and coherent report structure. • Inclusion of significant project milestones and findings. • Quality of analysis and synthesis of the project outcomes. • Quality and robustness of tools, workflows, and documentation. • Sponsor delivery.
Deliver a Stakeholder Presentation	Create and deliver a final presentation to stakeholders which summarises the project's objectives, processes, findings, and value, using clear business communication.	• Effectiveness of communication in the presentation. • Clear summary of the project's outcomes, achievements, and value.

7. Implementation Plan

This Section details the plan to achieve the Objectives. It is noted that there will be an ongoing literature review to further develop the initial review, and to supplement and support each implementation stage. The project report will be developed through each stage of this plan.

Exploratory Data Analysis

Objective: Conduct a literature review and brainstorming sessions to identify analytical and visualisation method for trace data, conduct EDA, statistical analysis and develop associated visualisations exploring symbol occurrences, sequences, and sub-sequences and develop workflows supporting this analysis.

- Literature review and brainstorming for approaches to the statistical, modelling and visualisation analysis of traces.
- Analyse full sequences and sub-sequences and identify patterns.
- Visualise analytical results.
- Visualise patterns identified.
- Develop tools and associated workflows.

Milestone: Developed tools and associated workflows and documentation.

Model Development

Objective: Develop and test models (tools) to capture the patterns identified in the EDA phase.

- Test various modelling techniques to determine the most effective approach.
- Evaluate models based on their ability to accurately represent data patterns.
- Test models against a new sponsor supplied dataset to support reusability.

Milestone: Developed and validated models ready for workflow packaging and documentation.

Model Packaging & Documentation

Objective: Create reusable modelling workflows with documentation.

- Package the chosen models for sponsor use as workflows.
- Prepare documentation on model usage, implementation, and limitations.

Milestone: Reusable modelling tools and workflows with supporting documentation.

Visualisation Development

Objective: Develop and test visualisations to represent patterns identified in the data through modelling.

- Test a variety of visualisation types (such as heatmaps, time series graphs and cluster visualisations) to determine effectiveness in representing patterns.
- Ensure visualisations are intuitive and informative for end-users.
- Test visualisations against a new sponsor supplied dataset.

Milestone: Developed visualisations that accurately represent data patterns.

Visualisation Packaging & Documentation

Objective: Create reusable visualisation workflows with documentation.

- Optimise visualisations for aesthetics.
- Package the chosen visualisations for sponsor use as workflows.
- Prepare documentation on visualisation usage, implementation, and limitations.

Milestone: Reusable visualisation tools and workflows with supporting documentation.

Configuration Development & Documentation

Objective: Develop tool and workflow configurability and documentation.

- Develop tool (analysis, modelling, and visualisation) configurability.
- Embed configurability in workflows.
- Document configuration options.

Milestone: Developed, deployed, and documented tool configuration options.

Bundle Workflows, Documentation and Final Report

Objective: The objective is to combine the work completed to date and to finalise the project report in preparation for sponsor delivery.

- Bundle developed workflows and documentation for sponsor use.
- Finalise project report.

Milestone: Ready to deliver workflows, documentation, and project report.

Report Submission and Package Handover

Objective: Submit the project report to the sponsor and deliver model and visualisation package to the sponsor. This may include a demonstration to the client.

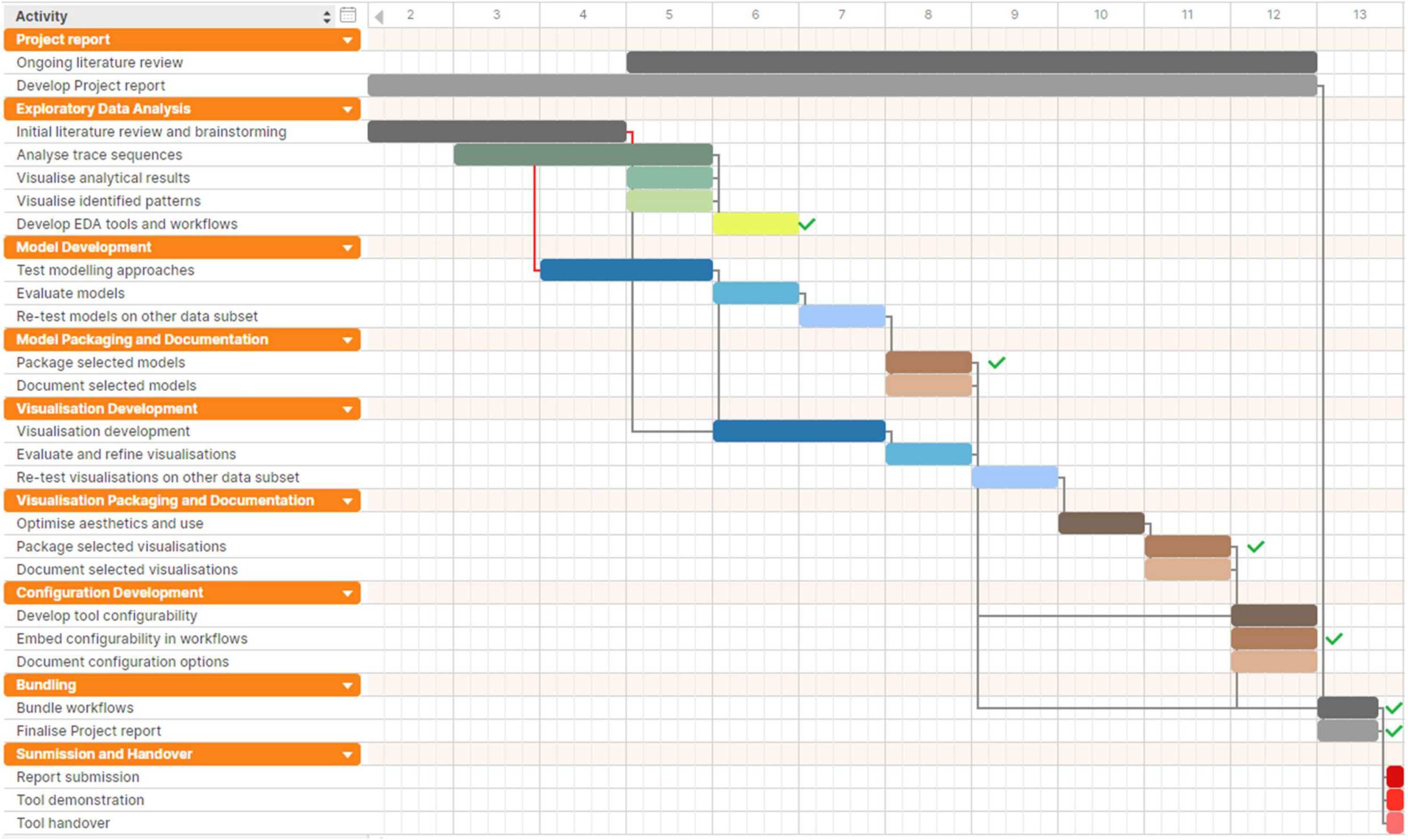
This will establish that the project has been completed and the project goals have been met. In summary:

- Submit project report, workflows, and documentation.
- Tool and workflow demonstration.
- Reusability demonstration.
- Comprehensive handover: Deliver all documentation and the codebase to the sponsor,

Milestone: Completion and sponsor acceptance, evidenced by the effective demonstration of the tools' capabilities and handover of all materials.

8. Project Schedule

The Gantt chart for the Project's implementation is shown below with columns representing the weeks in the first semester. The Gantt chart excludes the mid-semester break, and this time is reserved for as a buffer for Project delays. Lines denote dependencies and green ticks mark key milestones.



APPROVAL

We approve the Project as described above.

Name	Role	Signature	Date
Dr Jan Stanek	Sponsor		14.3.2024
Eric Lam	Mentor	Eric Lam	14/03/2024
Stephen Maher	Student	Stephen Maher	14 March 2024
Mansoureh Foruzandehshahraky	Student		14/3/2024
Aniket Varbude	Student		14/03/2024
Dona Joy Arimboor	Student	DonaJoy	14/03/2024