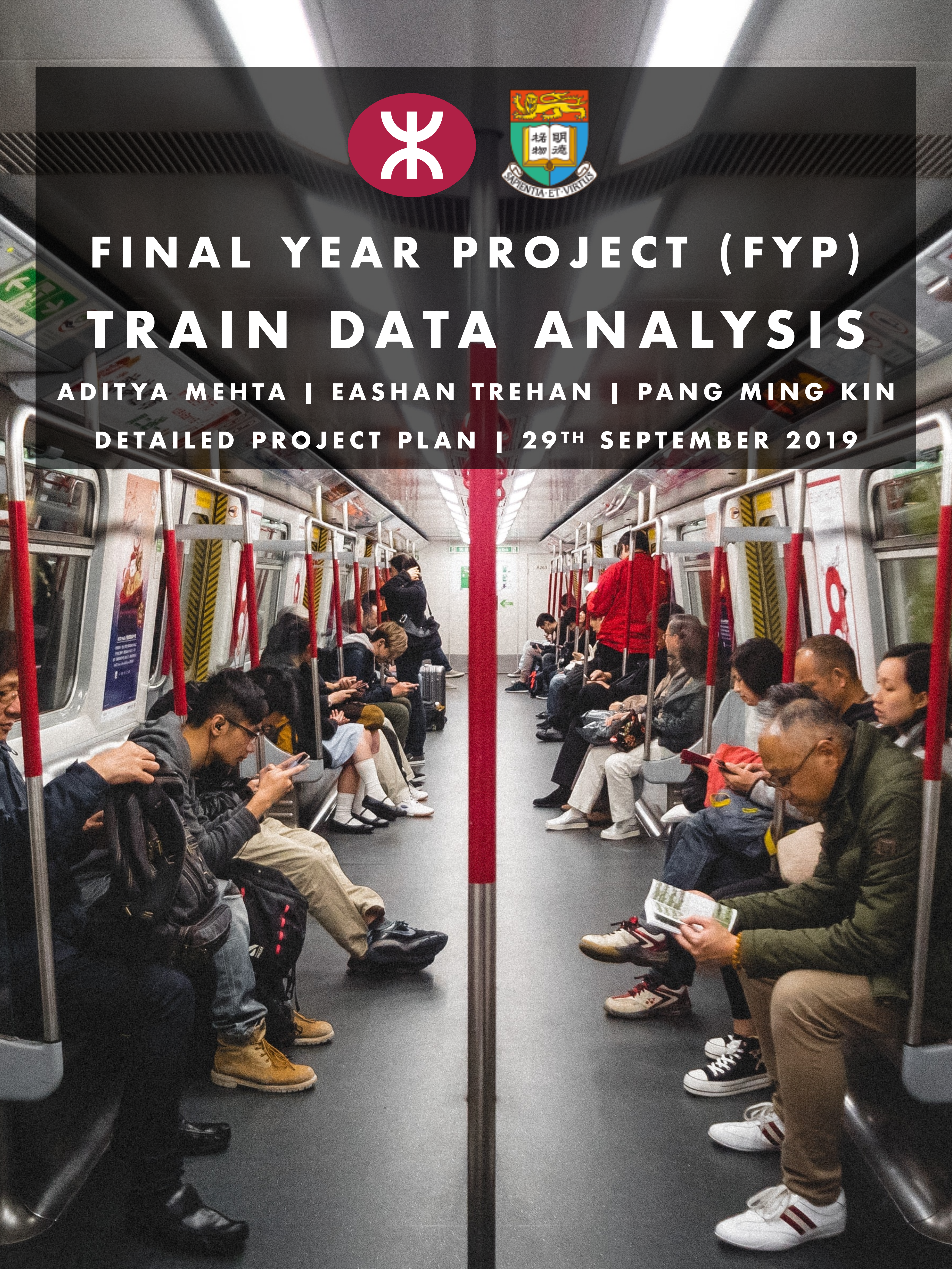
FINAL YEAR PROJECT (FYP)

TRAIN DATA ANALYSIS

ADITYA MEHTA | EASHAN TREHAN | PANG MING KIN

DETAILED PROJECT PLAN | 29th SEPTEMBER 2019



Contents

OUTLINE

This report introduces the project, exploring in brief - the background, relevant literature, objective, scope, deliverables, proposed methodology and potential benefits. The report also discusses certain risks and challenges that may affect the project and ways to mitigate them. A project schedule has also been created with key milestones to keep track of progress. The current status of the project, in terms of progress, as of the date of submission, is also included. A concise conclusion along with references and a supporting appendix are provided for further guidance, at the end of this report.

The introduction chapter discusses the various entities and people involved in the project and their relevant history. The background chapter covers an overview of the Mass Transit Railway Corporation Limited (MTR), certain problems that are currently being faced, certain technical aspects and illustrative examples and data. The following chapter discusses various works of literature that may offer important leads, followed by the objective that the project aims to achieve. The chapter on scope and deliverables clearly states factors that are outside and within the scope of the project, further listing out key deliverables. The chapter on proposed methodologies explores various approaches that may be followed during the course of the project, while the chapter on schedule and milestones lays out the timeline and key landmarks of the project. The benefits chapter discusses the potential, positive impact that the project may deliver at its completion. The report then discusses certain risks and challenges that may be faced during the course of the project and how the team plans to tackle them. In conclusion, the report details the current status of the project.

INTRODUCTION

This project is a collaboration between the MTR and the University of Hong Kong (HKU). A team comprising of students and professors from a variety of departments, including but not limited to Computer Science, Statistics and Mathematics has been assembled to provide technical expertise in diverse fields ranging from software development to data analysis.

The project is supervised by Dr. Reynold Cheng (Figure 1), Associate Professor of the Department of Computer Science, HKU. He was an Assistant Professor at HKU between 2008-11 and received his bachelors in Computer Engineering in 1998, and masters in Computer Science and Information Systems in 2000, from the Department of Computer Science, HKU. He then obtained an MSc and PhD from the Department of Computer Science, Purdue University in 2003 and 2005 respectively. Dr. Cheng served as an Assistant Professor in the Department of Computing at the Hong Kong Polytechnic University between 2005-08. Dr. Cheng's main research area is large-scale data management and has been working on the modeling, querying, cleaning, mining, and system development of uncertain databases.

As part of the project, Dr. Cheng is providing much valued advice and guidance to three final year students under his supervision, namely, Eashan Trehan, Aditya Mehta and Pang Ming Kin. Eashan Trehan (Figure 2) is pursuing a Computer Science and Finance double major at HKU and has attended exchange programs at the University of Toronto (UofT) and the London School of Economics and Political Science (LSE), with past work experience at J.P. Morgan and Deutsche Bank. Aditya Mehta (Figure 3) is a Computer Science and Finance double major, with experience in Machine Learning and has previously worked at Hong Kong and Shanghai Banking Corporation (HSBC). Pang Ming Kin (Figure 4) is a Computer Science student with experience in Data Analysis, and has previously worked as an intern at the Alibaba Group and China Merchants Bank (CMB).

Figure 1

Figure 2

Figure 3

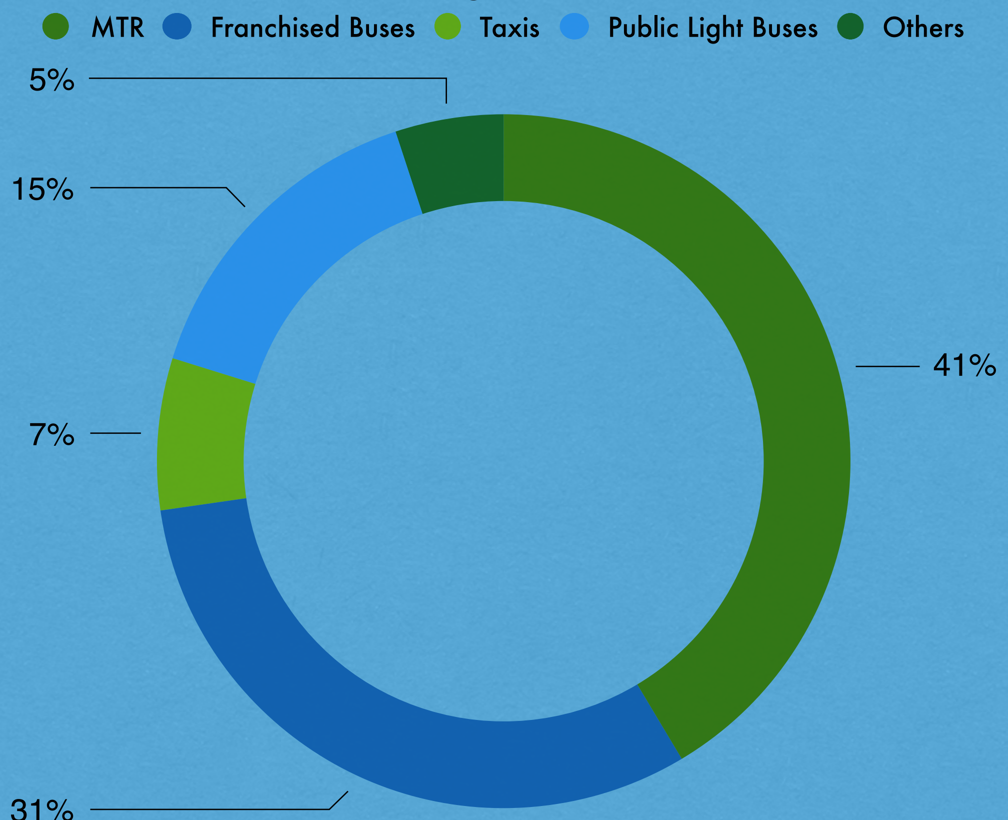
Figure 4

BACKGROUND

3.1 ABOUT MTR

The MTR is a forty year old organisation, offering railway service in Hong Kong, covering all eighteen districts and enabling over 12 million passenger journeys, every weekday. Figure 5 (Legislative Council Secretariat, 2017) highlights the market share of various means of public transport in Hong Kong, as of 2016, illustrating the market leading popularity of the MTR among commuters. Between February 2018 and January 2019, MTR achieved a stellar service performance with 99.9% of all passenger journeys reaching on time (Mass Transit Railway Corporation, 2017). However, during the same period of time, there were 10 separate incidents which caused a significant delay of over 30 minutes.

Figure 5



3.2 CURRENT PROBLEM

For instance, on 25th October 2018, train services at Po Lam station were delayed due to power supply problems, forcing MTR to arrange shuttle buses for affected passengers (Wong, 2018). On 5th August 2017, the Kwun Tong Line (KTL) suffered a delay of over 10 hours, due to signalling issues (Chung, 2017). Such incidents can cause significant inconvenience to passengers as well as the MTR. In the event of a delay of over 5 minutes, the MTR is required to pay a penalty to the Hong Kong government, while commuters also lose valuable time. Figure 2 below is a list of incidents which took place on the KTL, on 4th August 2019, it is notable that there were a total of 10 incidents with a delay of 5 minutes or greater.

3.3 CURRENT SOLUTION

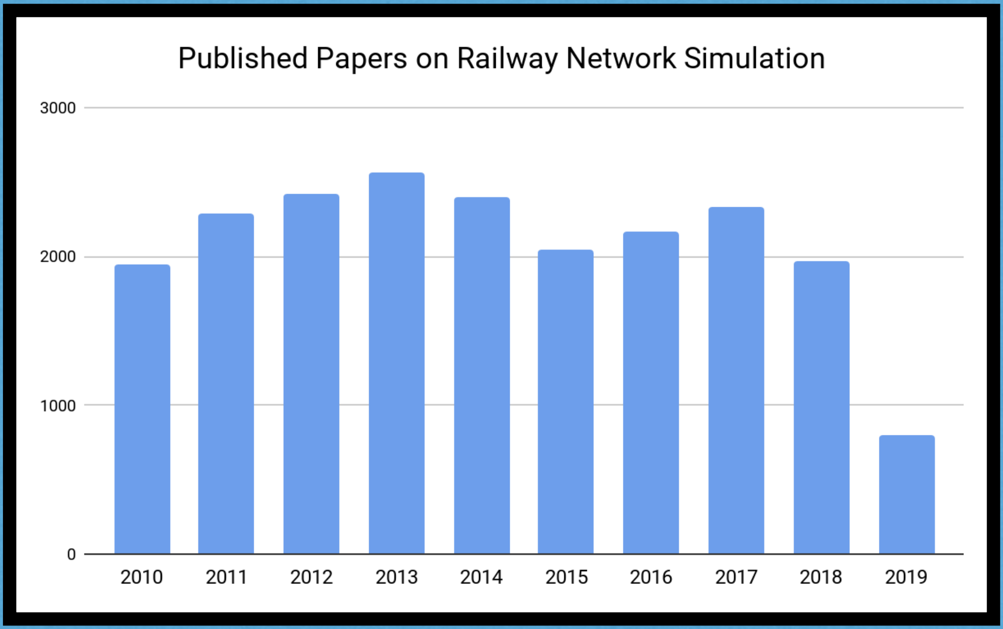
When facing such scenarios, train operators are required to generally rely on their past experience and technical knowledge, in order to minimise the impact of a disruption, therefore heightening the risk of human error.

LITERATURE REVIEW

In 1784, James Watt invented the steam locomotive. This was the seminal development that catalysed radical progress in the space of railroad transportation. Since the dawn of the first steaming train, railroads have become the lifeblood of countries (Schivelbusch, 1986). The challenge of simulating a railway network has the dual importance of having economic as well as social significance. In this simulation, there are two agents or stakeholders. The first is the rail operator and the most common parameter optimised for this agent is the fuel cost. The second agent represents the passenger for whom journey time is the minimisation parameter.

The simulation process is undertaken, most commonly for two distinct outcomes. The first is to find an efficient solution to the scheduling problem. This entails making optimal allocation of the trains across the network to minimise passenger transit time and fuel costs. The second kind of simulation is for decision-making during disruptions and other adverse events. This is evidenced by the volume of research that has been done on this topic. As seen in Figure 6, a burgeoning body of literature exists on these modelling methods. The approaches undertaken to this end can be broadly classified into three categories: those which propose analytical mathematical models, those which apply stochastic optimisation techniques, and the more recent approaches which use machine learning. For the scope of our project, we will not be dealing with the scheduling problem.

Figure 6



The focus of our project is how to predict the time delay caused by known disruptions to the rail network. The MTR is a passenger rail network, and as noted in (Jespersen-Groth J. et al., 2009), the train infrastructure manager/operator has a degree of influence on the performance, as they have discretion on how to react to a disruption. We will be investigating the utilisation of OpenTrack for our simulation (Nash & Huerlimann, 2004). It is a microscopic simulation model, i.e. it is not based on statistical averages, and models network behaviour at the individual train level. We also evaluated (Wang P. & Zhang Q., 2019) for their big-data driven approach to predict train delays. They utilised analysis of historic delay times, along with weather data sets. Based on the initial data extracts provided to us by MTR, this approach fits the data best. However, we will need to account for additional factors such as ridership, time-of-day, etc. since the MTR network is predominantly indoors and may be unaffected by weather.

OBJECTIVE

The goal of this project is to work with the MTR, in order to develop a software platform that analyses and visualises train schedule and movement, for example, the amount of time the train spends at a platform, the departure time of the train, the amount of time for which the train stops in a tunnel etc. The project aims to reduce the number of trains that suffer from a short delay, i.e. greater than or equal to 5 minutes and long delay, i.e. 30 minutes or above. The end deliverables of the project include a working simulation model, visualised to take into account various scenarios that a train may face, such as a signal fault, overcrowding etc. and a software/algorithm that can provide relevant recommendations to deal with the scenario in the most optimised manner possible.

SCOPE

The scope of the project is limited to a simple visualiser which can simulate a finite set of scenarios in a 2-Dimensional space, i.e. the visualiser cannot account for all the infinite possibilities and the exact scenarios which are being covered and shall be further determined as the project progresses. The recommendations that are made by the software/algorithm are planned to be in real-time and are limited to certain options and constraints that have been communicated to the team by the MTR, for instance, the software may recommend the train operator to offboard passengers from the train, however it cannot recommend changing the direction of movement of the train by 180 degrees. Furthermore, the project is specifically focused on the Kwun Tong Line (KTL) in particular and does not take into account other lines of service, such as the Tsuen Wan Line etc.

PROPOSED METHODOLOGY

As the project is currently at a nascent stage and involves learning several new concepts, the methodology proposed in this report is tentative in nature and may be altered, depending on the flow of the project.

7.1 VISUALISATION OF SIMULATOR

The visualisation created to showcase the simulation of the project is built through “Unity”, a platform and programming language which is often used for creating games, films etc. and is apt for creating a 2-Dimensional simulation model for the purpose of this project. Unity was chosen in particular, over other options, such as “Arena”, "Unreal Engine” and “Photon”, as it is widely compatible with a diverse range of platforms, the visuals produced can adapt to different devices and screens without compromising quality and there is extensive documentation and tutorials that are available for new developers to learn quickly.

7.2 DATA REQUIREMENTS

The project involves extensive data requirements, including but not limited to - the distance between various stations, the planned timetable for different times and days of the week, the dwell time at each station etc. These datasets are required to be available in .csv or Excel format in particular, for ease of use. The team plans to use “Python” and “MongoDB” to analyse the data. Python was chosen in particular, over “R”, as the latter is mostly used for statistical analysis, while the former can be used more widely for data science purposes in general and provides ease of replicability and accessibility, in comparison. MongoDB was preferred over “SQL server”, as the former offers better availability and scalability, due to auto-sharding and makes it easier to represent complex relationships. The project involves the use of Machine Learning, which will be executed by utilising libraries such as SciPy and Scikit-Learn, while Deep Learning can be implemented via Caffe, TensorFlow, Theano or Keras and be visualised through Pytorch or Keras-vis. Each of these options present their own respective advantages and disadvantages and a final decision will be taken after the team has received all the relevant data from MTR and studied it in detail.

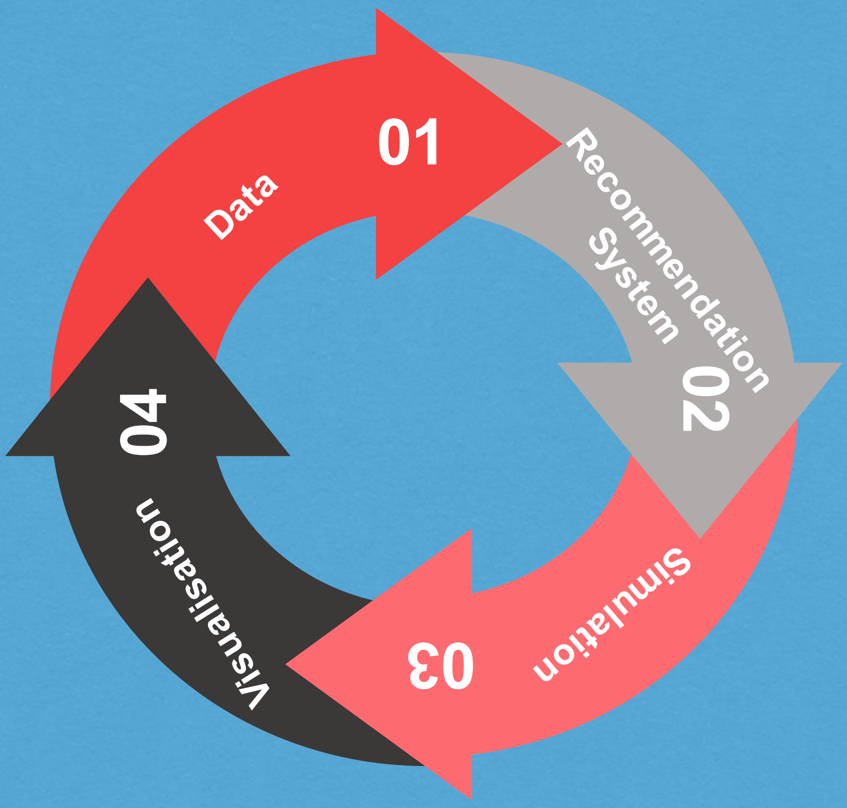
7.3 HARDWARE EQUIPMENT

GPUs with high processing capabilities are required in order to successfully process the vast amounts of data in a short time frame.

7.4 SOFTWARE DEVELOPMENT APPROACH

As illustrated by Figure 7 below, there are four key steps to developing a working simulation model that can be visualised and provide relevant recommendations. Firstly, the algorithm must be able to accept various forms of data, variables such as the train number, the track it is located at, the station it is leaving or arriving at, what kind of situation or incident is the train experiencing etc. Based on these key data points, a recommendation needs to be produced, to optimise the handling of any extraordinary situation. The recommendation is to be simulated, to judge whether it can produce optimal results and the selected recommendation is then simulated visually.

Figure 7



SCHEDULE & MILESTONES

Figure 8, attached below, showcases a "Gantt Chart" outlining the proposed software development cycle.

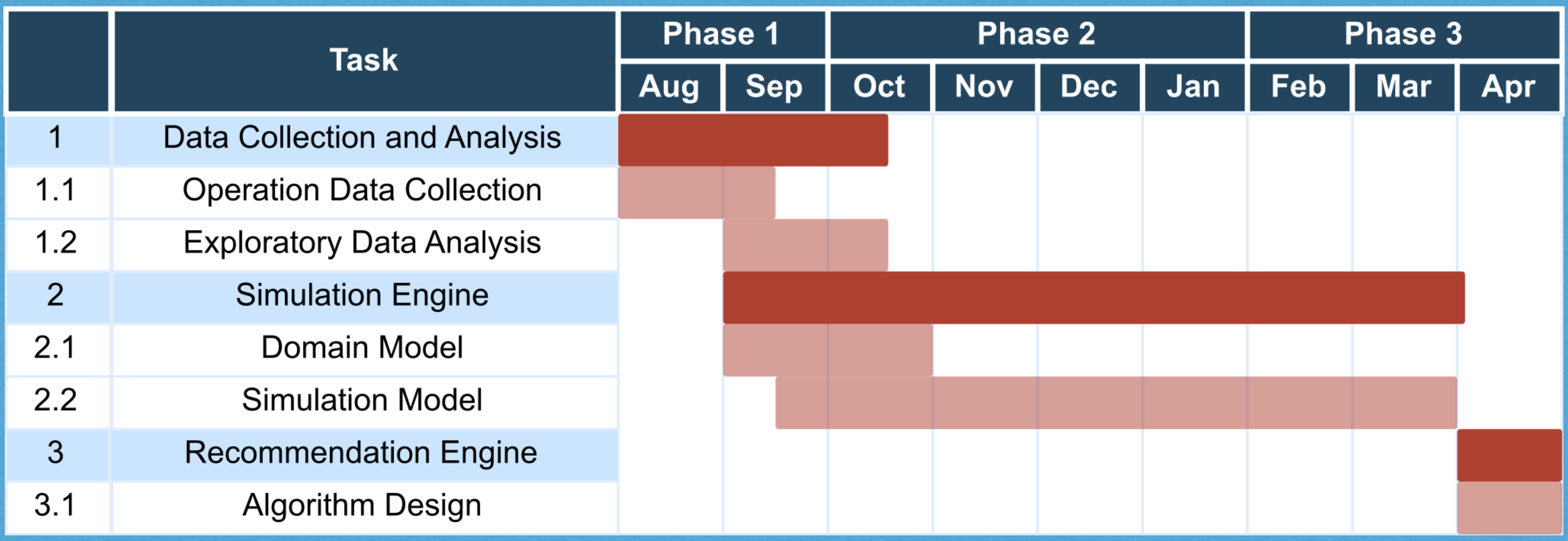
The key milestones involve creation of a visualised simulation model, a software/algorithm for providing recommendations and linking the simulation model with the recommendation system to provide real-time recommendations.

In Phase 1, from August to September, exploratory data analysis has been carried out on the dataset provided by MTR to understand the problems. The deliverables are this project plan, the project website, and a prototype of the simulator for demonstrations.

In Phase 2, from October to January, major components of the simulator will be implemented, with a simulation model validated by actual data. The deliverables are the simulator and the interim report.

In Phase 3, from February to April, the implementation of the simulator will be finalized, with comprehensive testing and validation. A preliminary prototype of the recommendation engine is also expected to be developed as the foundation of the upcoming stage of the research project. The deliverables are the final implementation of the simulator, the prototype of the recommendation engine, and the final report.

Figure 8



	Task	Phase 1		Phase 2				Phase 3		
		Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
1	Data Collection and Analysis	🟥	🟥	🟥						
1.1	Operation Data Collection	▨	▨							
1.2	Exploratory Data Analysis		▨	▨						
2	Simulation Engine		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
2.1	Domain Model		▨	▨						
2.2	Simulation Model		▨	▨	▨	▨	▨	▨	▨	
3	Recommendation Engine									🟥
3.1	Algorithm Design									▨

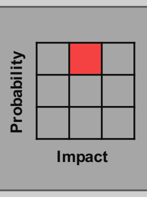
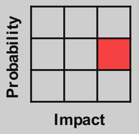
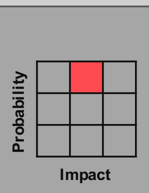
BENEFITS

As aforementioned, the objective of the project is to minimise the delays suffered by the train. The potential benefits from the impact of this project include, but are not limited to, a reduction in delays, improved efficiency and time management, lower probability of the MTR being penalised by the government for extensive delays and time being saved for commuters.

RISKS, CHALLENGES & MITIGATIONS

The project involves certain ambiguities and new skills that need to be learned, for instance, in order to gain a good understanding of the datasets, one needs to have obtain the necessary domain knowledge and a sound understanding of the concepts. As highlighted in Figure 9, there is a high probability of there being technical terms and concepts that are new to the members of the team, for instance, "dwell time" refers to the time that a train waits at a particular station, to allow passengers to onboard and offboard. Such key pieces of information may be learned by interviewing train operators, reviewing existing literature and reaching out to experts who are familiar with such concepts. Another key challenge is gaining access to data, since a lot of the data is confidential and proprietary, hence the team needs to be aware of what forms of data are necessary and negotiate with the MTR officials for the required access. For example, data on the distance between various stations is required in order to calculate the required speed and time for travelling. Given that the students involved are new to the field of railways, the project is exploratory in nature, this poses a key risk which can be overcome through an emphasis on teamwork, maintaining flexibility, regular communication and adopting an agile approach to the development process. These are some of the main challenges that may be faced by the team, during the course of this project, certain other challenges, such as multitasking, ability to meet deadlines etc. may also pose a risk.

Figure 9

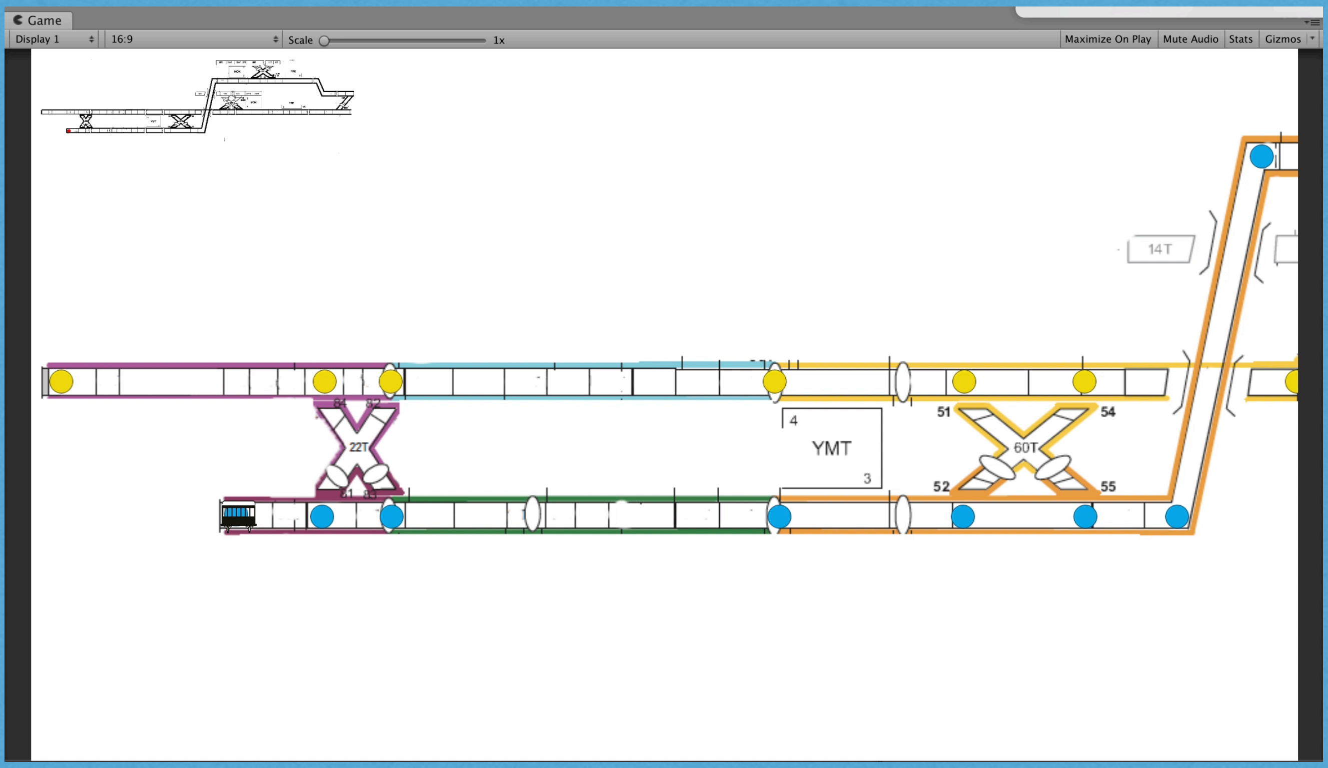
	Risks & Challenges	Impact-Probability	Mitigations
1.	**Lack of Domain Knowledge**		• Review existing literature on relevant topics • Reach out to experts to gain knowledge • Interview train operators
2.	**Limited Access to Internal Data of MTR**		• Negotiate for access to necessary data • Understand the various data requirements
3.	**Exploratory Nature of Project**		• Strong emphasis on teamwork • Maintain flexibility • Regular communication and feedback • Agile approach to software development

CURRENT STATUS

At present, the project is on schedule, a preliminary simulator has been created, which simulates the train movement on 25% of the overall KTL track, under normal conditions, as shown in figure 10 below, which is a screen capture of the simulation running. The team has received data from the MTR and is currently analysing it. The next step is to complete the simulation for the entire KTL track and attempt to draw out relevant conclusions from the data that has been received.

Afterwards, the team shall attempt to simulate the movement of multiple trains on various tracks together, using actual past data and thereafter simulate various scenarios such as accidents, technical failures etc. and form an effective recommendation system to deal with these issues.

Figure 10



CONCLUSION

In sum, the MTR, albeit a highly efficient organisation, is suffering from delays, which are causing an inconvenience to its passengers and incurring itself a penalty owed to the government.

This project may help reduce delays by creating an accurate, visualised simulation model and recommendation system, which can effectively deal with various extraordinary scenarios, as explained in the scope of the project.

Should the project be successful, there are several further prospects for development, potentially leading to the creation of autonomous driverless trains which can take quick action even during emergencies.

REFERENCE LIST

1. Chung, N.K. (2017). *MTR apologises for recent Hong Kong service delays and promises to invest in system maintenance*. South China Morning Post. Retrieved from https://www.scmp.com/news/hong-kong/economy/article/2106320/mtr-apologises-recent-hong-kong-service-delays-while

2. Jespersen-Groth J. et al. (2009). *Disruption Management in Passenger Railway Transportation*. In: Ahuja R.K., Möhring R.H., Zaroliagis C.D. (eds) *Robust and Online Large-Scale Optimization. Lecture Notes in Computer Science*, vol 5868. Springer, Berlin, Heidelberg.

3. Legislative Council Secretariat. (2017). *MTR Train Service Performance*. Retrieved from https://www.legco.gov.hk/research-publications/english/1718issh07-mtr-train-service-performance-20171220-e.pdf

4. Mass Transit Railway Corporation. (2019). *MTR Train Service Performance*. Retrieved from http://www.mtr.com.hk/en/customer/main/MTR-train-service-performance-jan-2019.html

5. Nash A. & Huerlimann D. (2004). *Railroad simulation using OpenTrack*. Swiss Federal Institute of Technology.

6. Schivelbusch, G. (1986). *The Railway Journey: Industrialization and Perception of Time and Space in the 19th Century*. Oxford: Berg.

7. Wang P. & Zhang Q. (2019). *Train delay analysis and prediction based on big data fusion. In: Transportation Safety and Environment*, vol 1. Oxford.

8. Wong, O. (2018). *MTR delays hit Hong Kong passengers for second time in October*. South China Morning Post. Retrieved from https://www.scmp.com/news/hong-kong/article/2170088/mtr-delays-hit-hong-kong-passengers-second-time-october

APPENDIX

Abbreviation	Full Form
API	Application Program Interface
CMB	China Merchants Bank
GPU	Graphics Processing Unit
HKU	The University of Hong Kong
HSBC	Hong Kong and Shanghai Commercial Bank
KTL	Kwun Tong Line
LSE	The London School of Economics and Political Science
MTR	Mass Transit Railway Corporation
UofT	The University of Toronto

Term	Meaning
Auto-Sharding	Database partitioning into smaller parts
Caffe	A deep learning framework
Dwell Time	Time spent waiting at a station
GPU	Electronic device for processing graphics
Keras	Neural network API
Keras-vis	Toolkit for visualising neural net models
MongoDB	Document database program
Python	A particular programming language
Pytorch	Python based scientific computing package
R	A particular programming language
Scikit-Learn	Python based machine learning platform
Sci-Py	Open source Python library for scientific computing
SQL Server	Relational database management system by Microsoft
TensorFlow	Open source library for machine learning
Theano	Python based library for mathematical operations