Applied Data Science

Building a predictive model for New York City Taxi drivers to maximise their revenue

Group 7

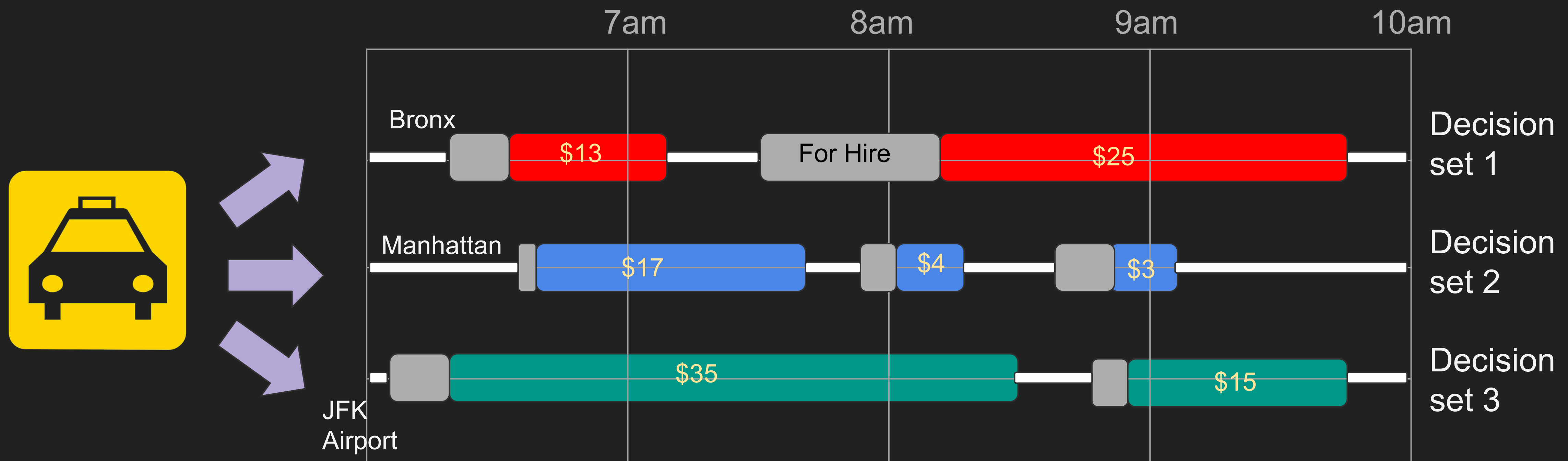
- Xuanken Tay
- Geng Yuxiang
- Li Shangqian
- Yin Zhou Zheng

Progress Report

Presentation Outline

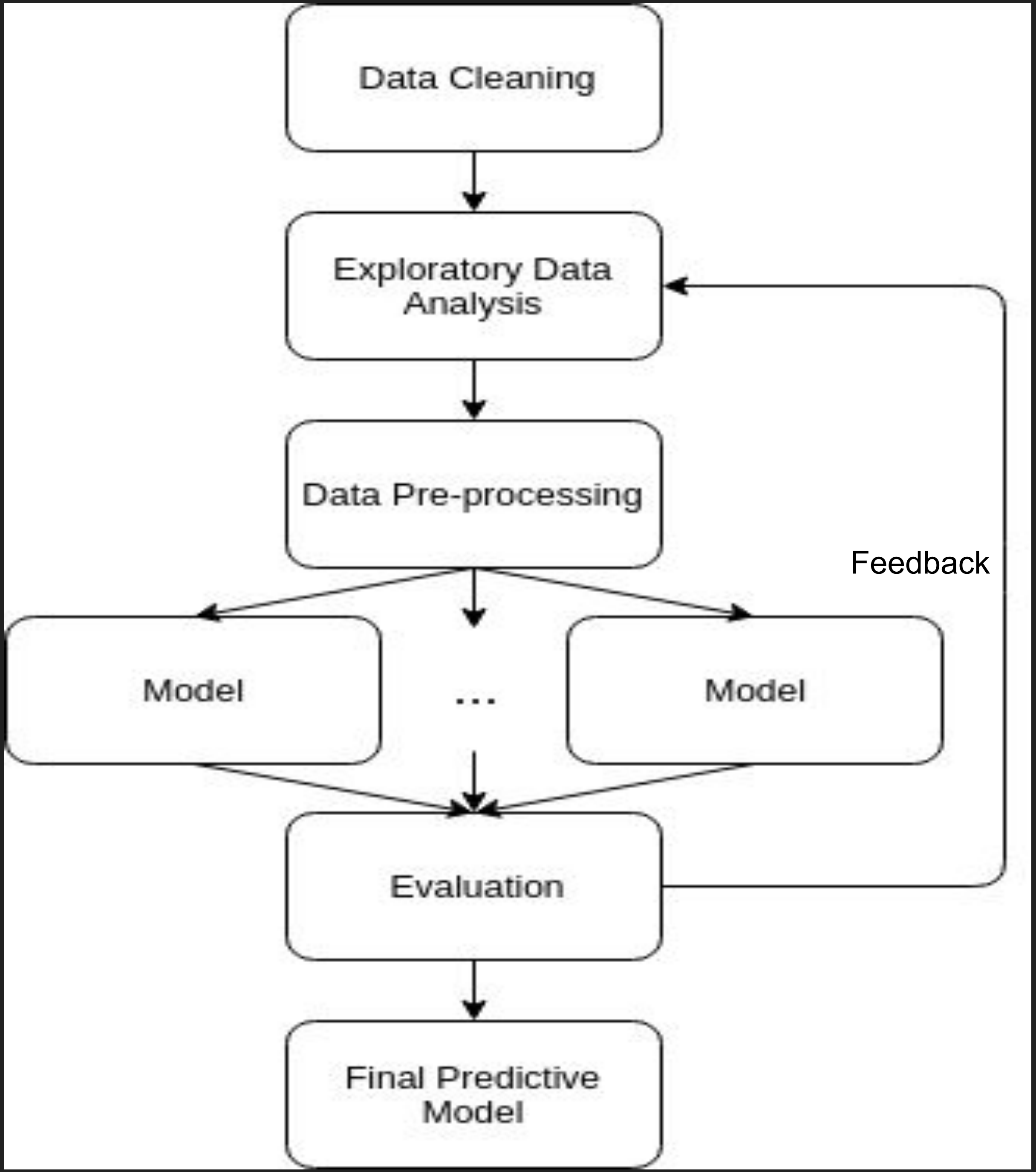
- Problem description
    - Workflow
- Feature Engineering
- Classification Model
    - Decision Making
- Evaluation of models
- Allocation of tasks

Problem Formulation



- Problem: What are the decisions that result in maximum revenue?
- Aim: Given the current time and location of taxi driver, find the next best decision to maximise revenue

Workflow


Data Cleaning
Exploratory Data Analysis
Data Pre-processing
Model
...
Model
Evaluation
Final Predictive Model
Feedback

FEATURES

Profitability = ...

Average Total Earnings Rate [AVG(TER)]

- Calculated as

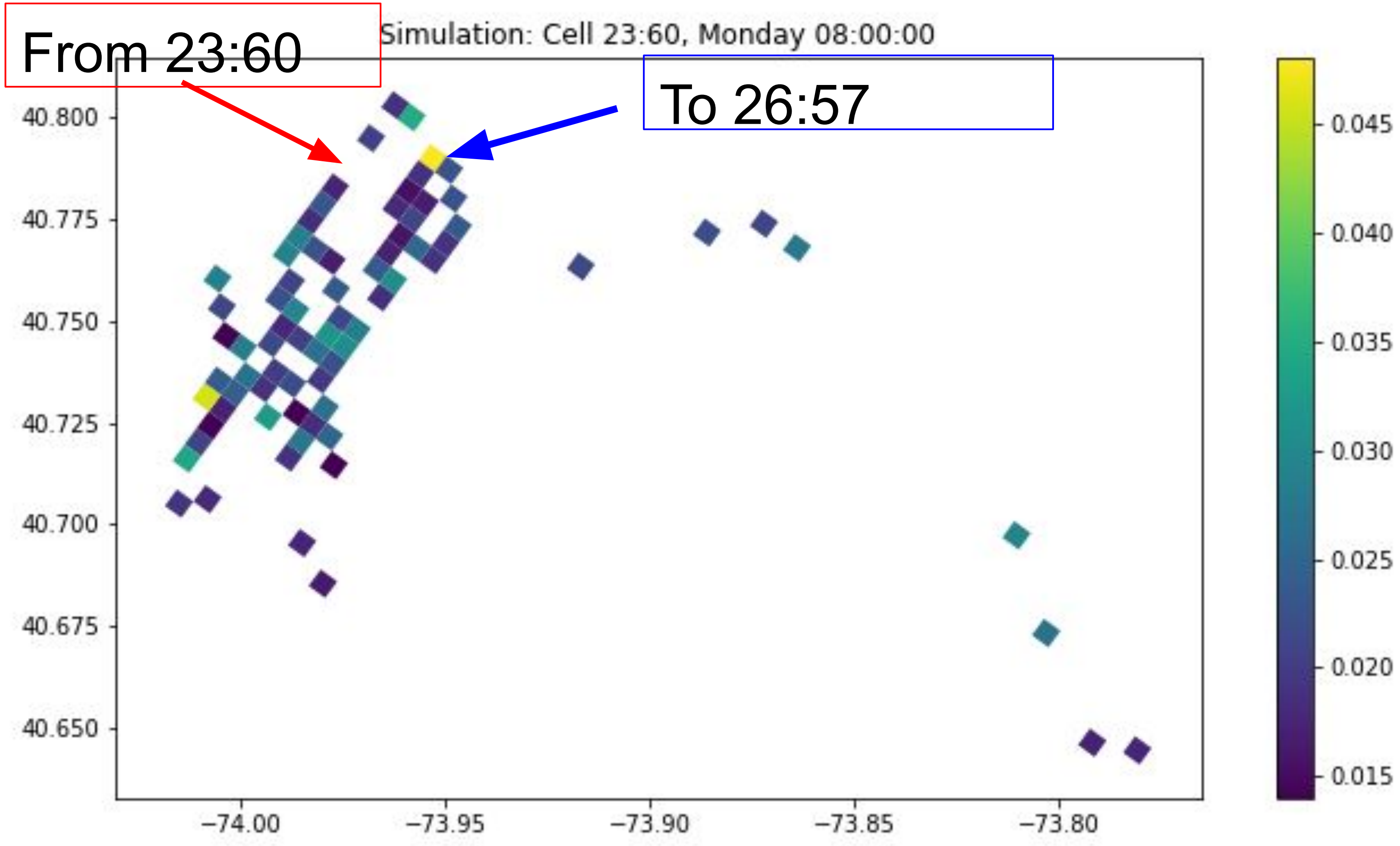
$$\mathrm{AVG(TER)} = \frac{\text{Averge Total Earnings [AVG(TE)]}}{\text{Average Trip Duration [AVG(TD)]}}$$

- Grouping Factors (Predictors)
  - Cell Id (e.g. “23:60”)
  - Day of the Week (e.g. Saturday)
  - Time of the day (e.g. 21:52:00)

Complementary Features (Penalized)

- Driving Duration [DD]
    - Time taken for driving from origin cell to destination cell.
    - Determine Shortest Cell Path between current cell and destination cell using Breadth First Search
- Average Waiting Time [AWT]
    - Average waiting time before next trip

$$\text{Profitability} = \text{AVG}(\text{TER}) + f(\text{DD}) + g(\text{AWT})$$



From 23:60

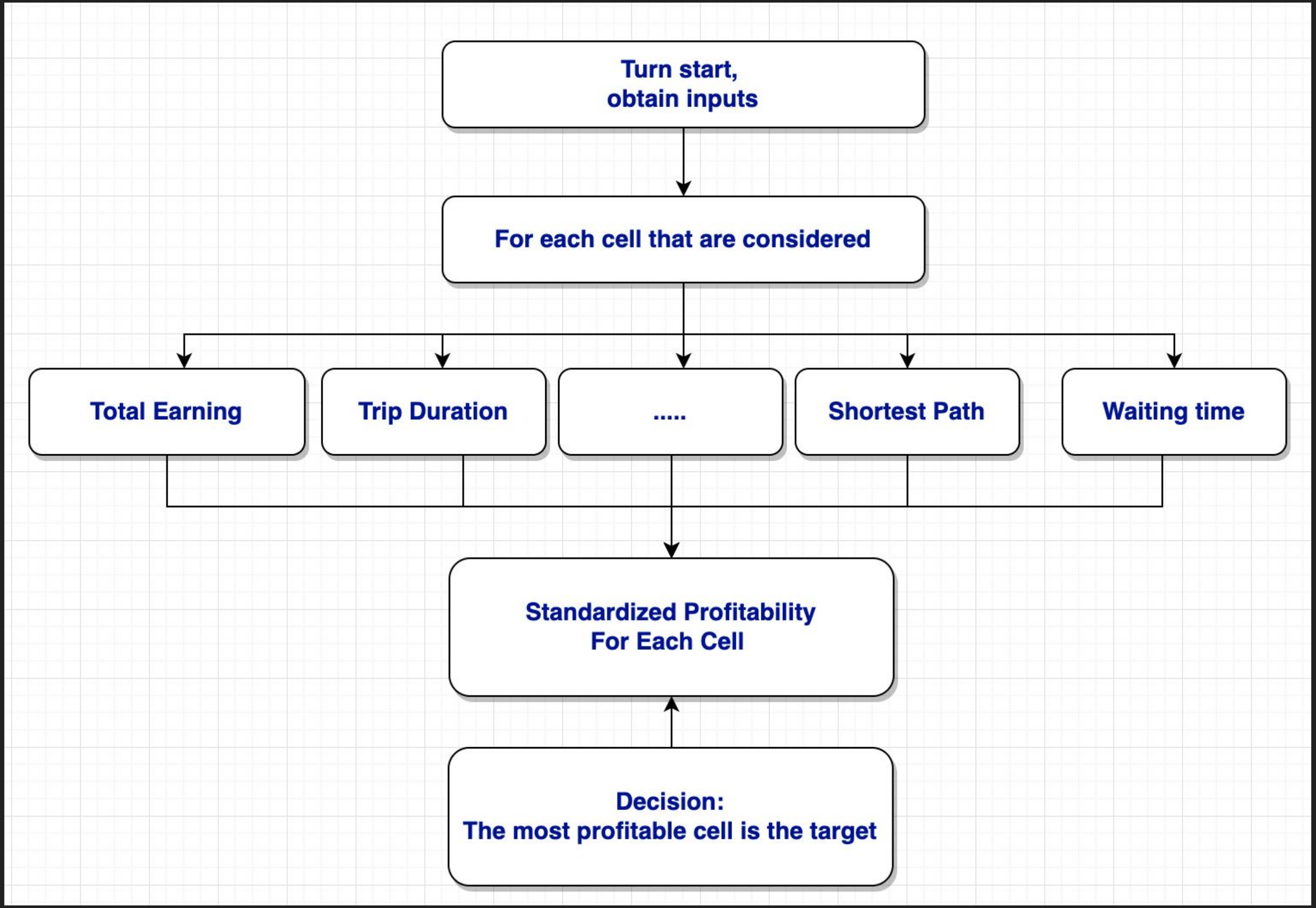
To 26:57

Method Implementation

- Different models to predict features:
  - Linear Modelling:
    - Predicting Average Total Earnings
  - Generalised Linear Model (Poisson, Gamma, Negative-binomial)
    - Waiting Time (Trip Frequency)
  - Clustering Method
    - Average Trip Duration

Classification Process

- Combine the predictions from the small models.
- Calculate standardised profitability of each cell.
- Choose cell with maximum as next cell.
- Goal-driven algorithm to prevent deadlock.



Evaluation

- Model Implementations:
  - Different combinations to produce players.
- Running on local game platform.
  - Observe total earnings over simulated week.
  - Compute statistics.
- Choose “best” player.
- Review logs to address uncertainty.

Job Allocation

- Partition:
  - Avoid overlapping.
  - Covering all avenues.
- Initial Ideas:
  - Models trained by different members.
  - Report sections split among members.
- Tools:
  - Meeting Minutes.
  - GitLab issue assignments.

Thank you for your time.

- Questions.