Predicting Bus Arrival Time Based on Positional Data

Ljubljana, 9 October 2023

Matic Kladnik, Luka Bradeško, Dunja Mladenić

Outline

- Problem setting and data
- Description of proposed approach
- Overview of the architecture of the proposed system
- Machine learning models
- Evaluation application

Problem Setting

- Goal: predict arrival times of city buses to stations
- Live system: predictions are available at any time via the API
- Large city in EU with more than 800 buses

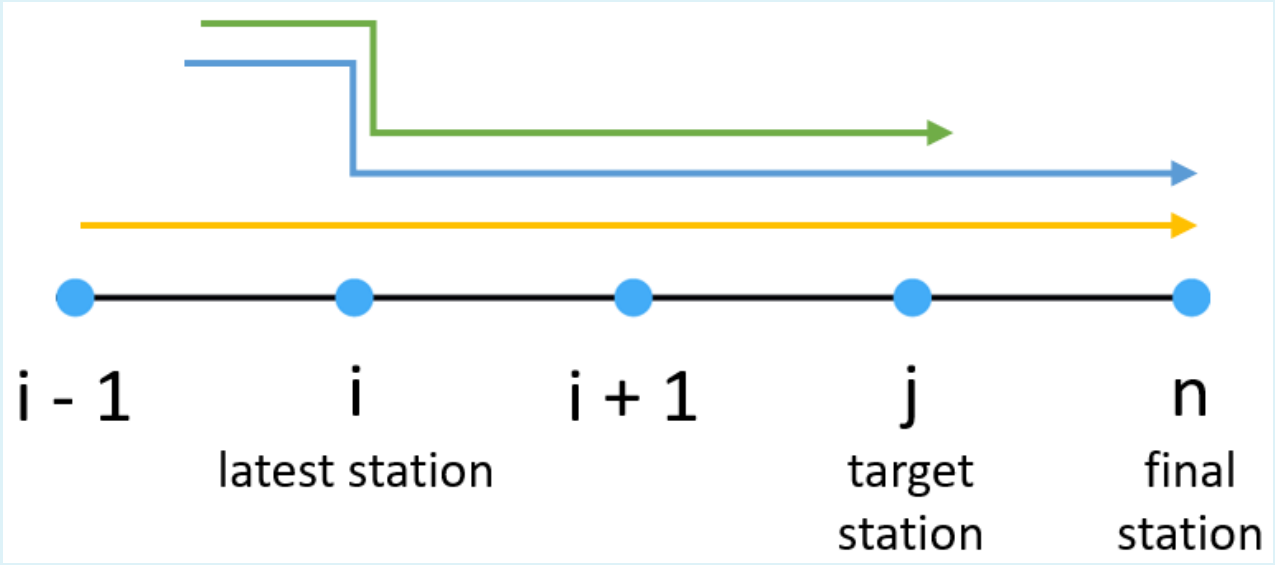
Data

- Latest bus position coordinates
- Details of routes: sequence of stations for each route
- Details of stations: positions, names
- Suboptimal level of detail: we do not know the exact departure time from the station -> we detect vicinity of a bus to a station

Proposed Approach Description

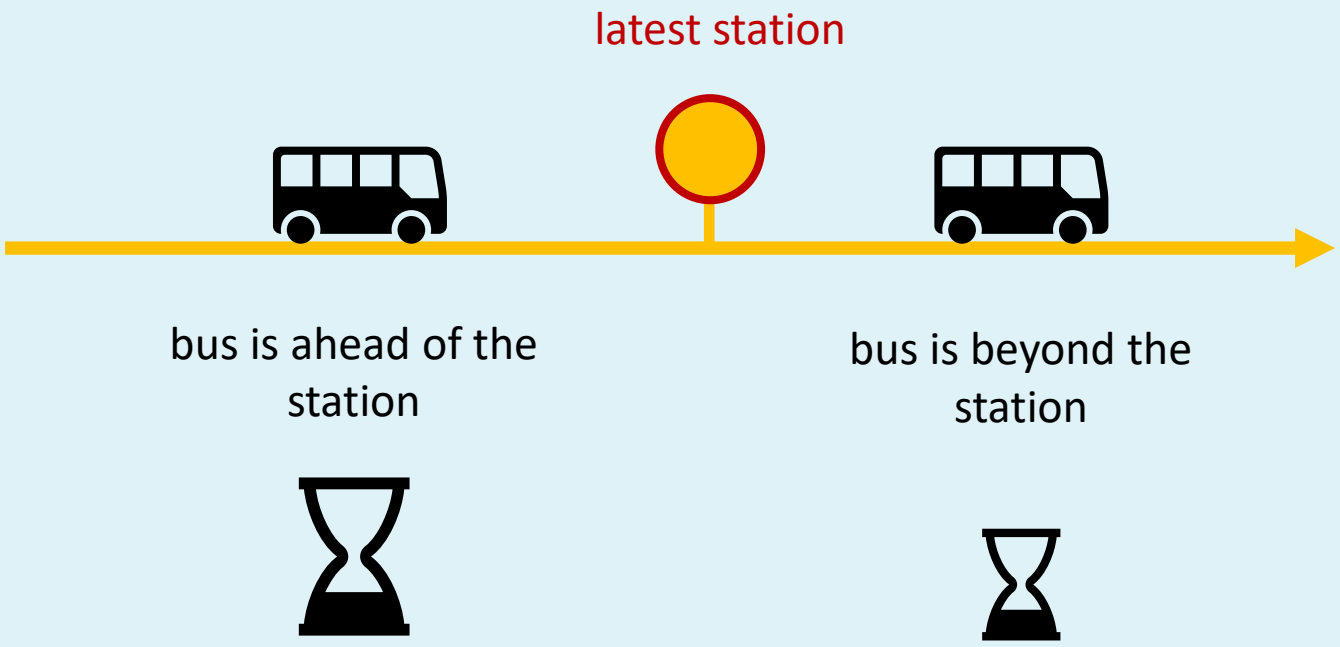
- Use recent historic data to estimate traffic flows
- Recent travel times on the same route

- Recent travel times on routes that share the path

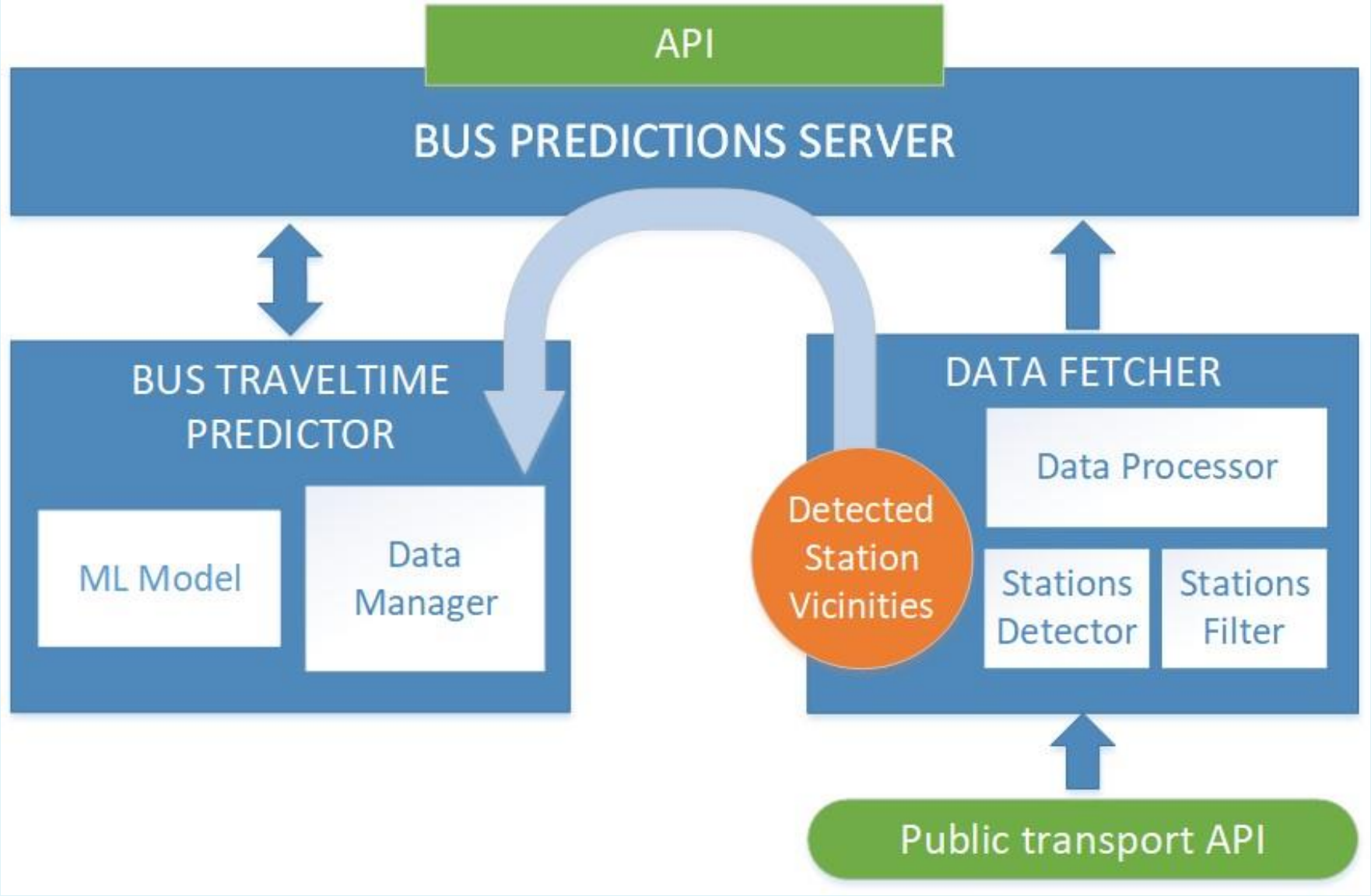


Proposed Approach Description

- Positional semantic context: relative position of the bus to the latest station



Architecture of the Proposed System



Architecture of the Proposed System

- Data fetching from the Public Server API several times per minute
- Most recent predictions about arrival times are stored in the data manager
- Latest predictions for specific stations, buses, or routes are returned immediately upon request through the proposed system’s API

Machine Learning Models

- Linear regression

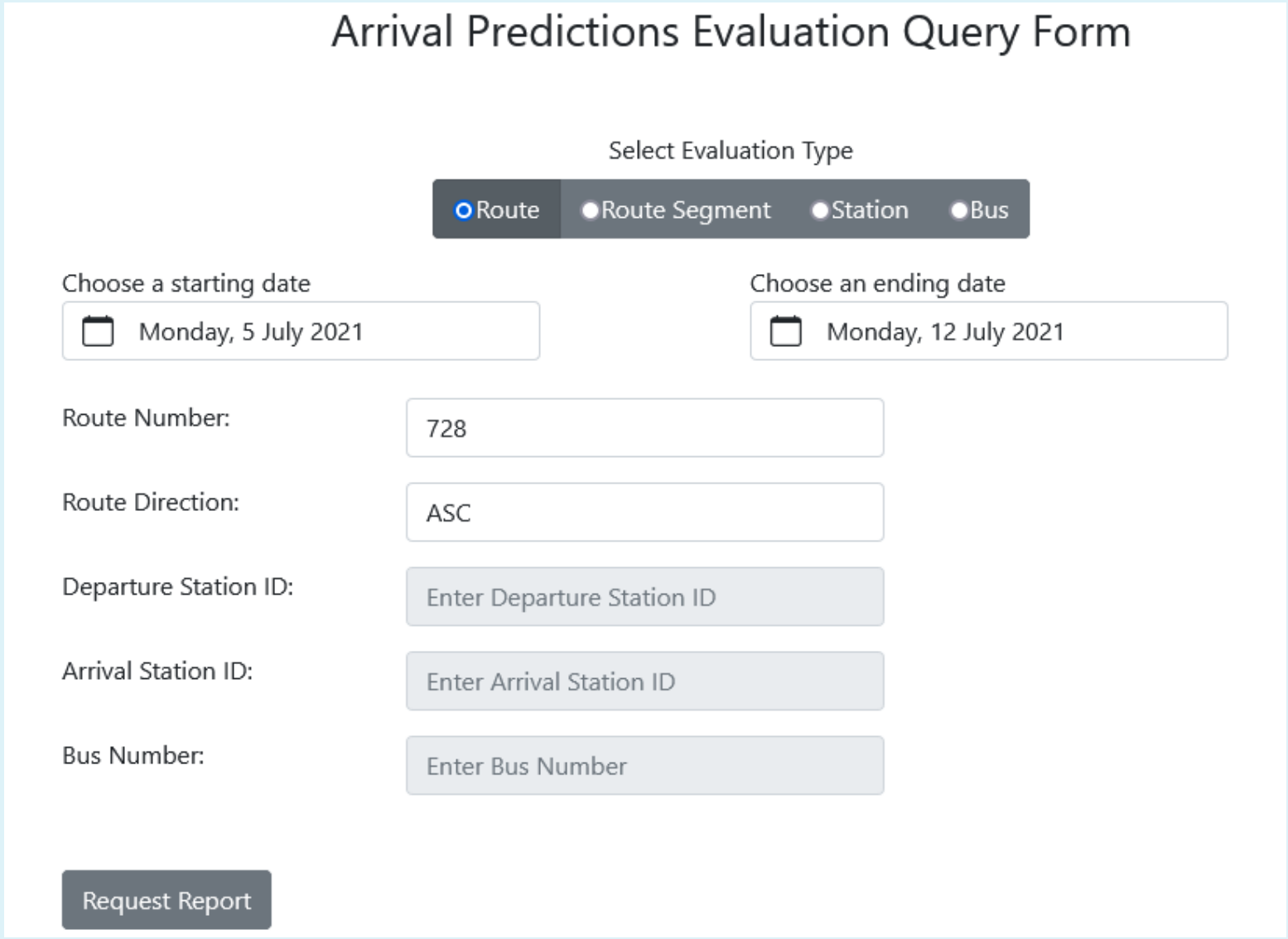
- **SVM (SVR)** using RBF kernel

- Neural network
  - Multi-layer perceptron: 2 hidden layers (15, 8)
  - L-BFGS for optimizing weights

Evaluation application

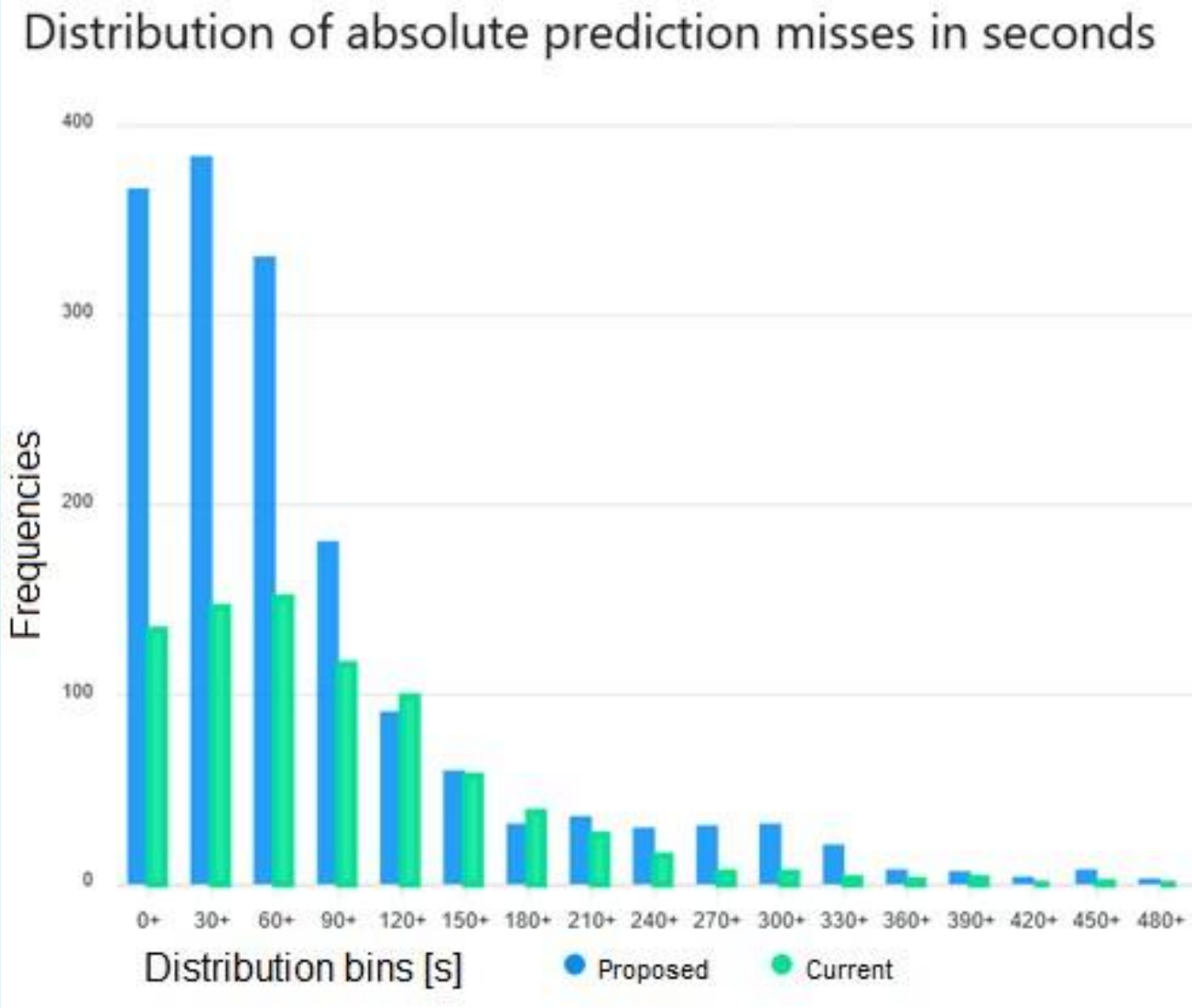
- Completed predictions (have detected arrival to target station) are stored in the DB
- Evaluation data is queried from the DB and sent to the web application upon request from the proposed system’s API
- Web application transforms data and shows visualizations

Evaluation Application



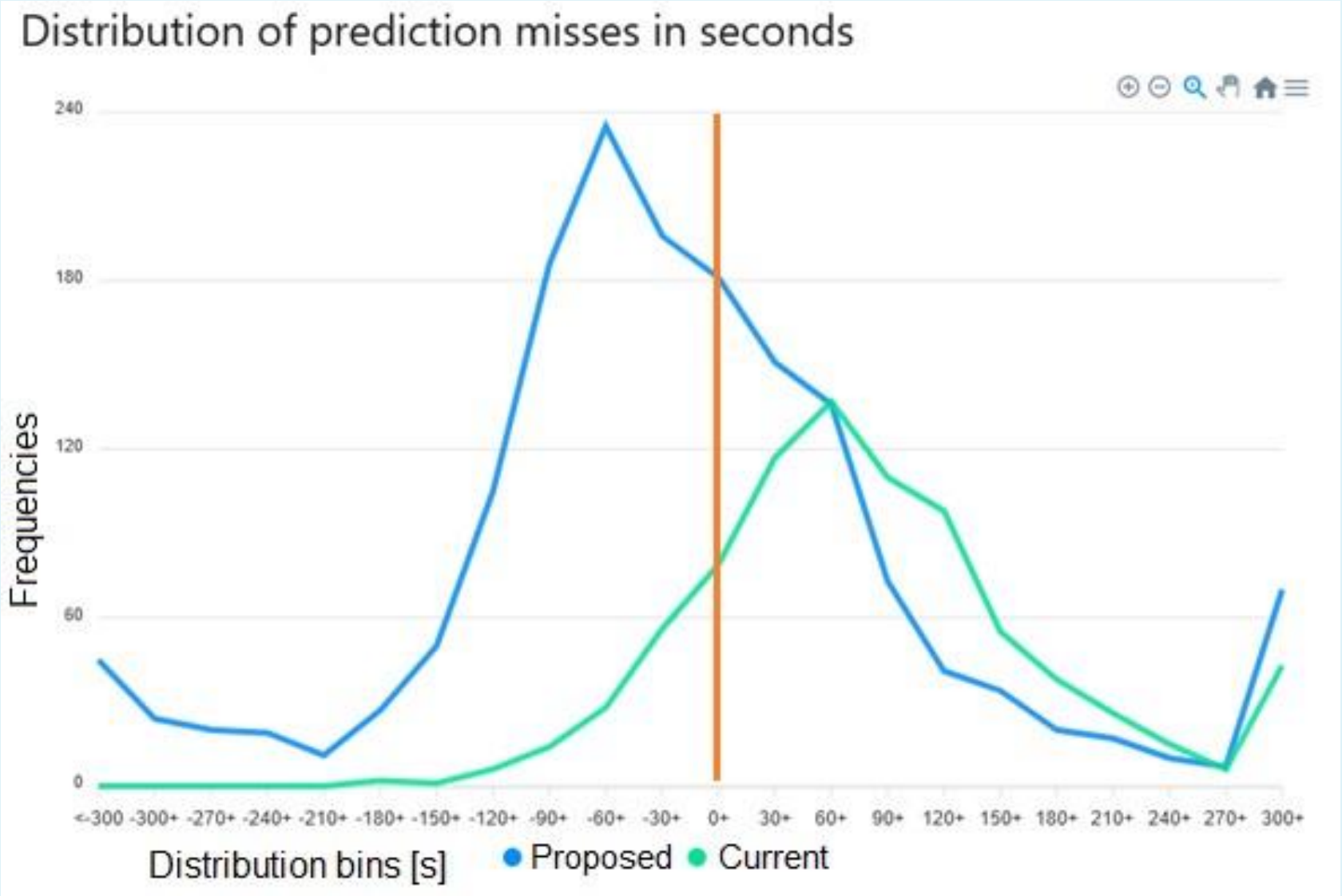
Evaluation Application

- Prediction errors are merged into bins of 30s
- 0+ bin is the best one



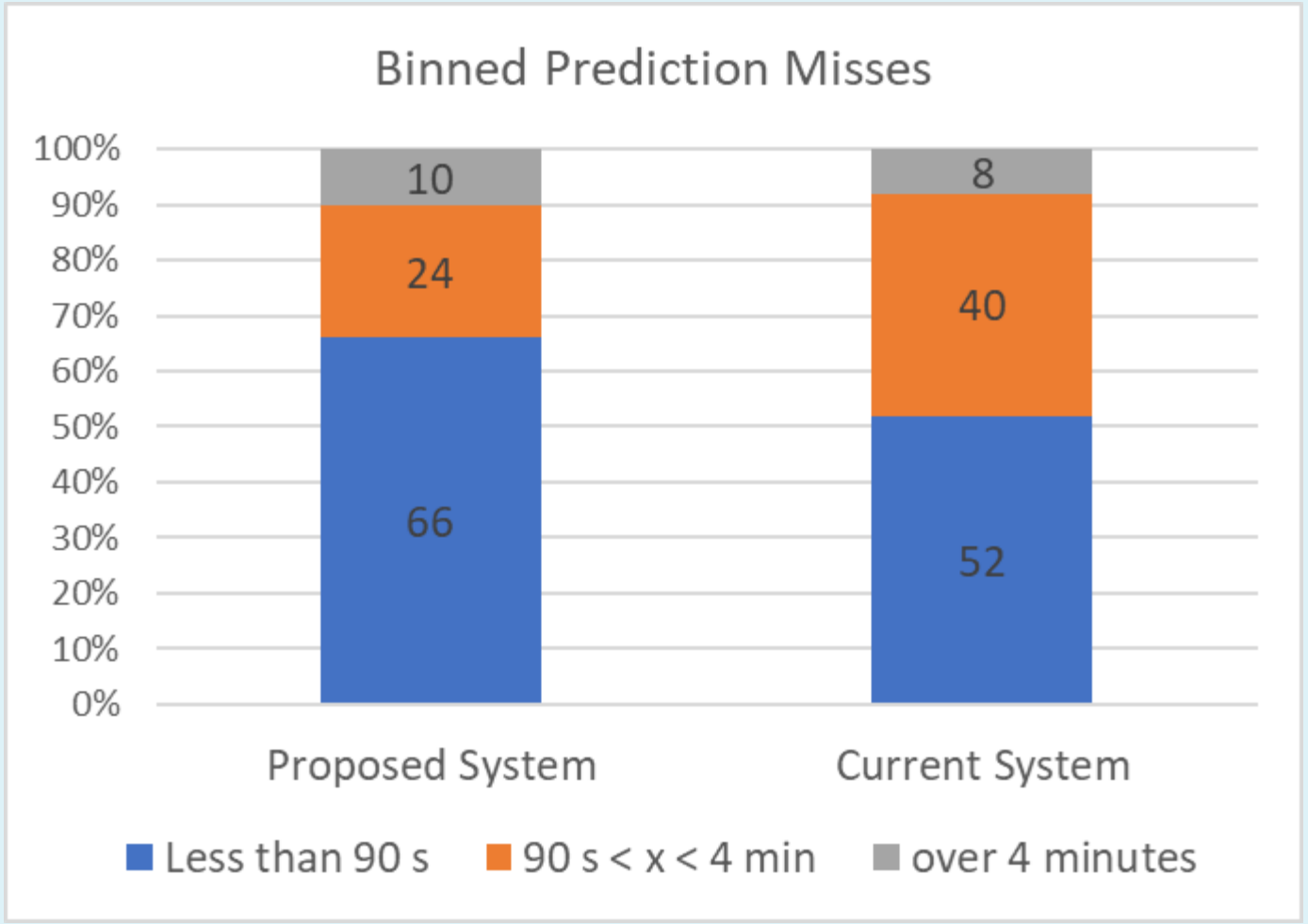
Evaluation Application

- Best bins:
  - -30+
  - 0+
- Negative bins represent predictions that undershoot



Evaluation Application

- Based on opinions of domain experts
- Blue bin is the best



Evaluation Application

- The system provides computed metrics of performances:
  - MAE
  - RMSE
- The proposed system performs better across various evaluation approaches
- Different evaluation approaches give us a more complete evaluation and comparison between various prediction systems

Conclusion

- Recent historic context of buses from the same and compatible routes
- Relative position to station as semantic context
- Architecture of the proposed live prediction system
- Evaluation of performance between systems with a web application
- Future work
  - Additional machine learning algorithms (random forest, XGBoost, other NN)
  - Improving the evaluation application to support showing relative share of distributions in bins instead of absolute numbers
  - More performance metrics, e.g. MAPE