Machine Learning and Statistical Methods in Transport Problems

Hao Wu
and David Levinson
School of Civil Engineering
University of Sydney

Research Questions

- Performance of Machine Learning models compared to statistical methods, in solving transport problems

- Methods in combining different model predictions
    - 'the best model' vs an ensemble of models

Applications

1. Fare and travel time forecast
    - Recreate pooling choice scenarios
    - Populate sparse OD time matrices using ride-sharing vehicles as probes

2. Pooling decisions (mode choice)

3. Flow (number of trips) between places
    - Estimate travel demand between zones

Machine Learning and Statistical Methods for Regression/Classification

Statistical Methods:

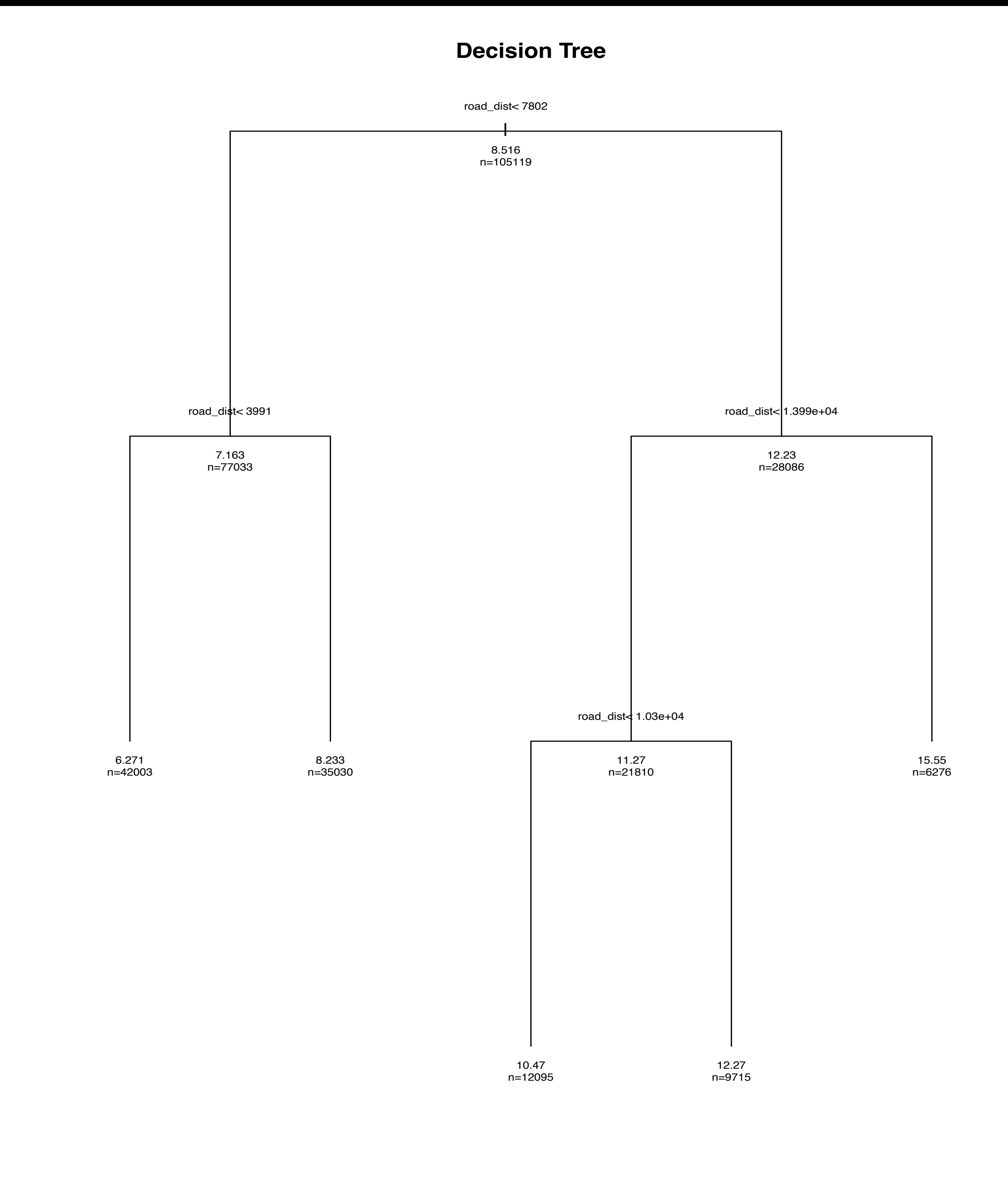
- Linear regression
- Discrete choice modeling (e.g. logit)
- Trip distribution
    - the number/estimate of trip generation/attraction required for all zones within a study area

Machine Learning and Statistical Methods for Regression/Classification

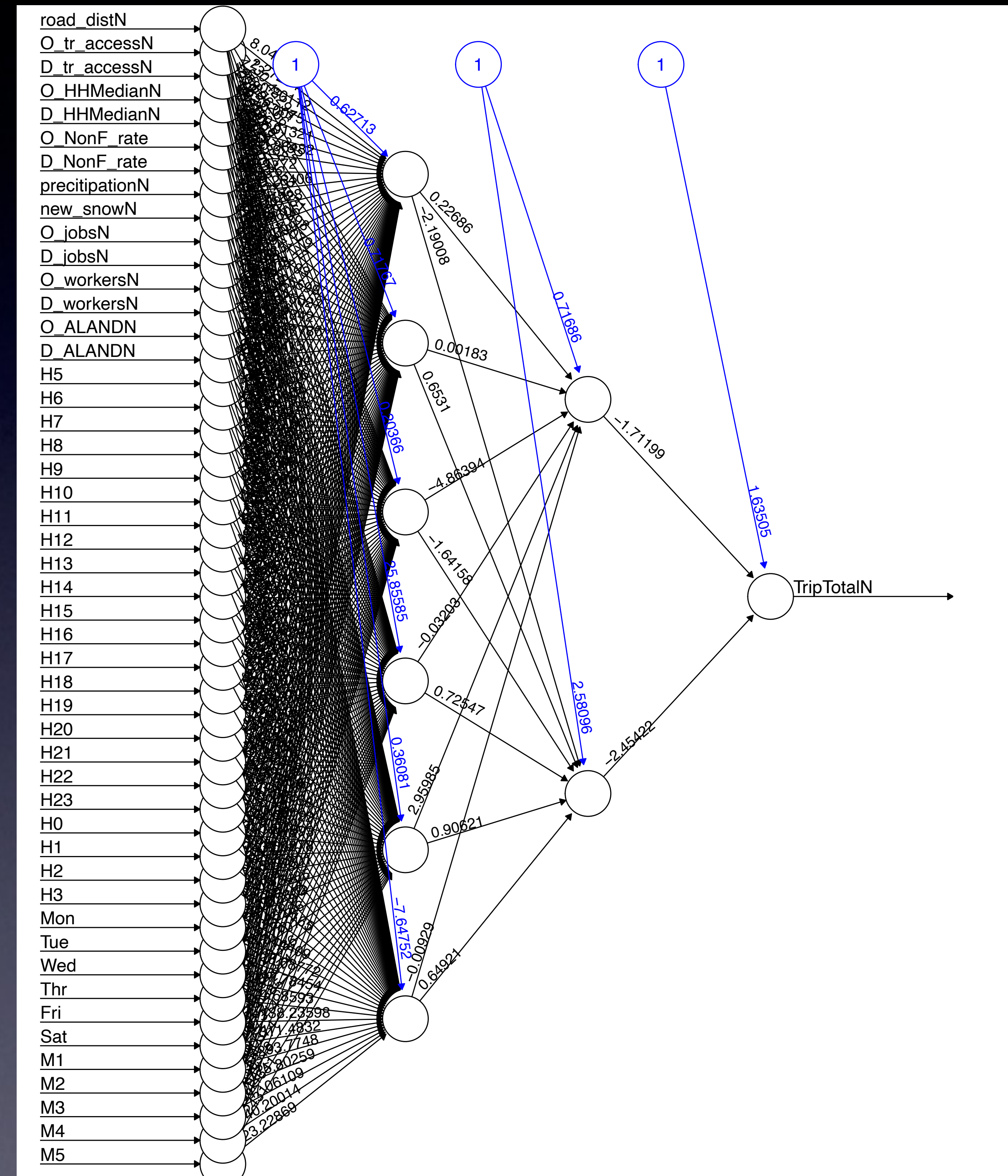
Machine Learning Algorithms:

- Decision Tree
- Random Forest
- Gradient Boosting Machine
- Artificial Neural Network
- Support Vector Machine
- K-Nearest-Neighbor

Machine Learning and Statistical Methods for Regression/Classification



Decision Tree



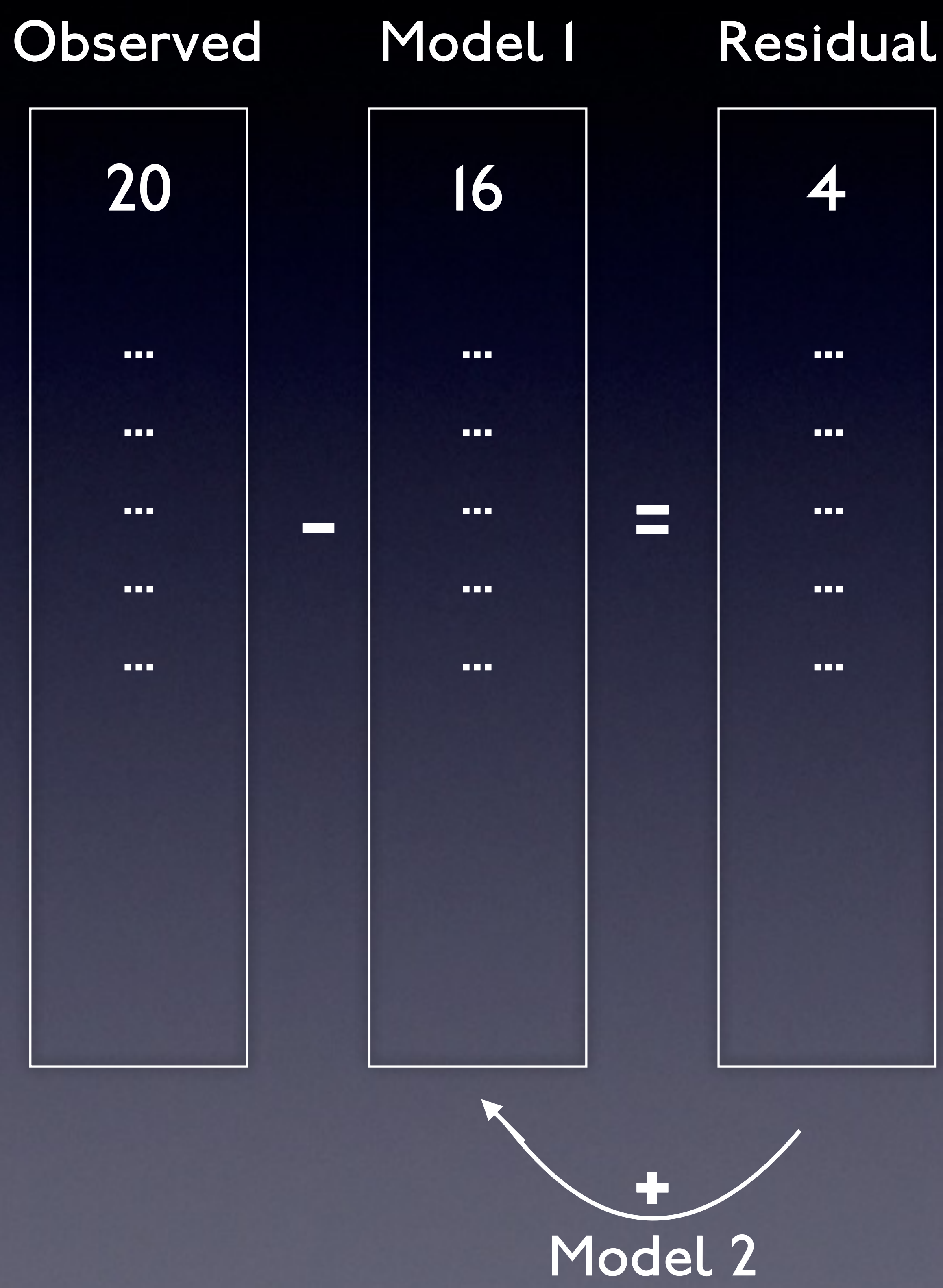
Neural Network

Combining Forecasts

- Boosting (but using different models)
    - Sequentially adding models, predicting the residual from the previous model

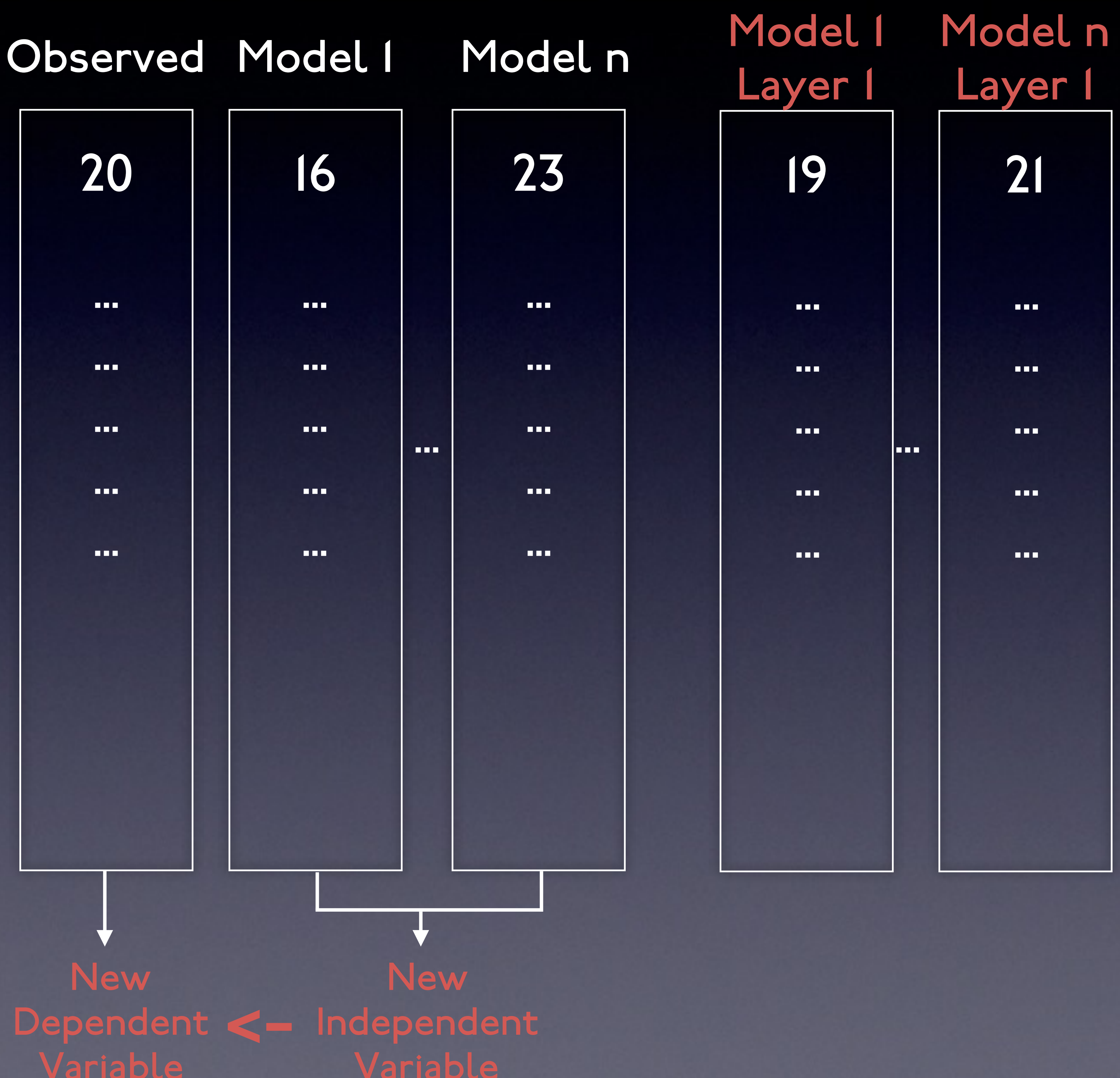
- Deep Modeling (combine, and iterate forecasts from different models)

Boosting



- Sequentially add models to predict the residual from the previous model
- A small testing data used to select the best model for predicting residual in the next round, based on RMSE
- Use the same explanatory variables, as the initial model 1

Deep Modeling



- Combine model predictions from the previous layer of models

Deep Modeling

- Based entirely on the training data
- Testing data validates improvement in model accuracy

Evaluating Model Performance

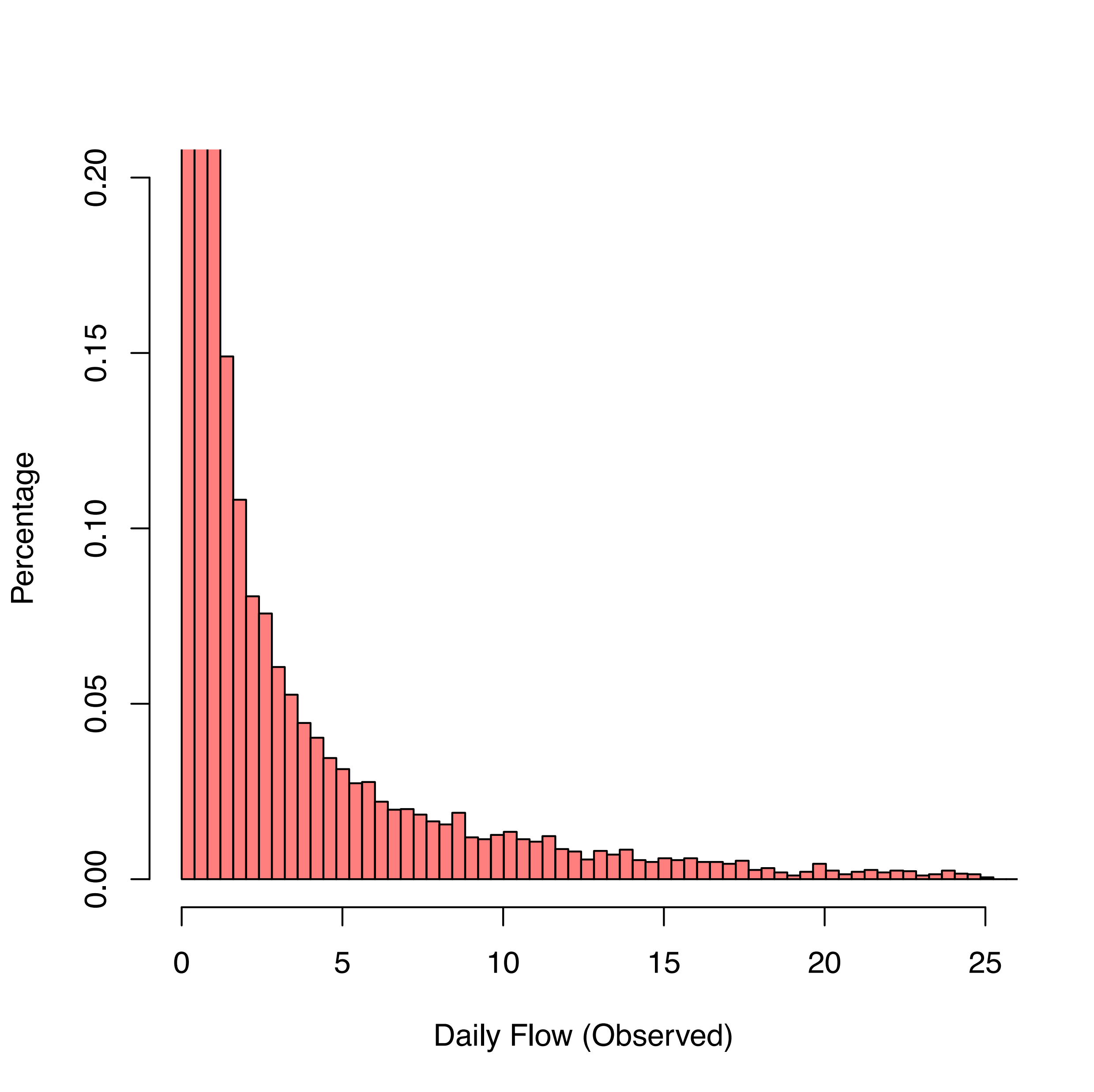
1. Root Mean Square Error (RMSE) $$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}$$

2. Mean Absolute Error (MAE) $$MAE = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|$$

- Both RMSE and MAE measure the magnitude of error; smaller is better
- RMSE $\geq$ MAE
- RMSE penalizes larger errors

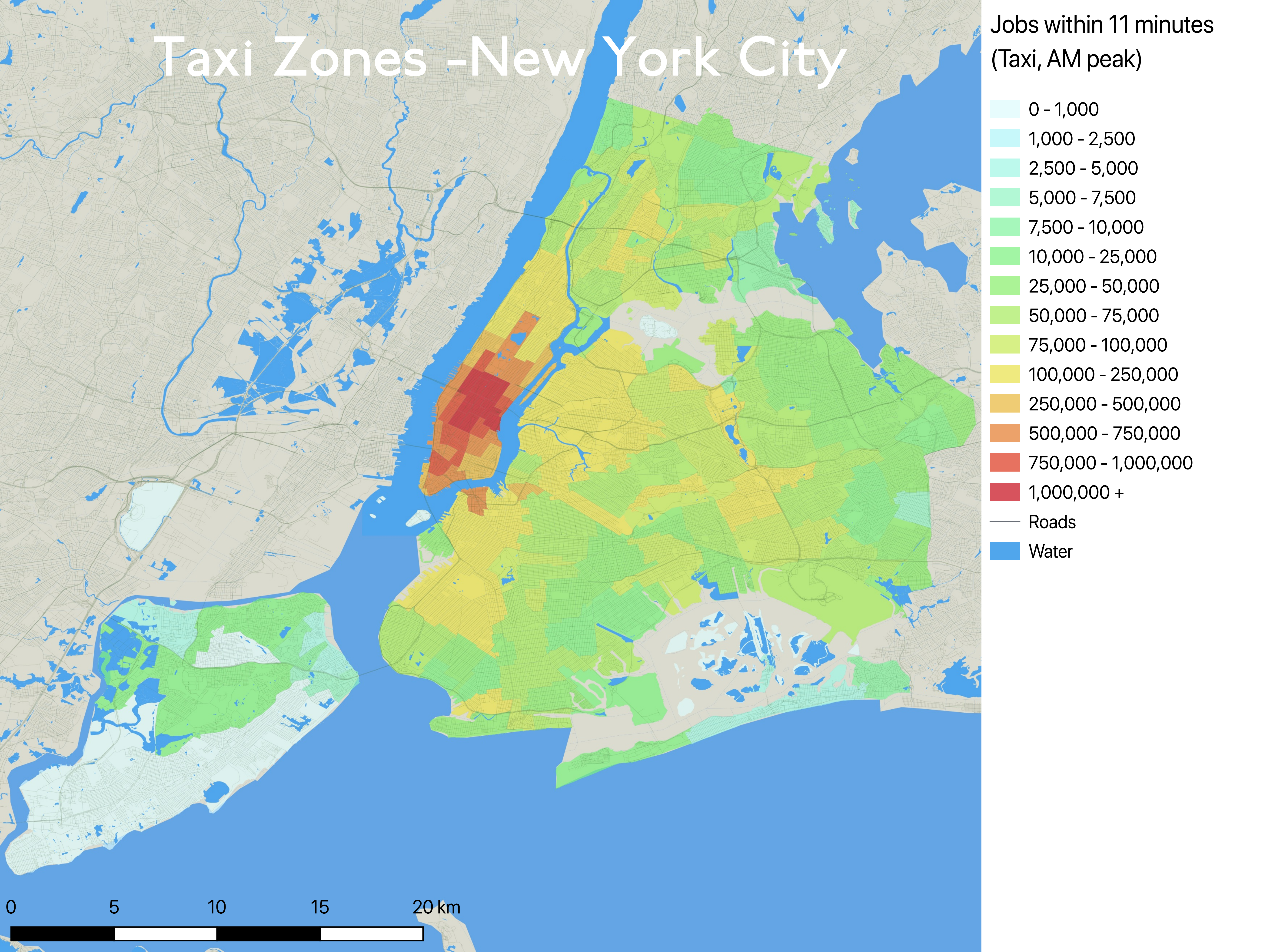
Flow Prediction

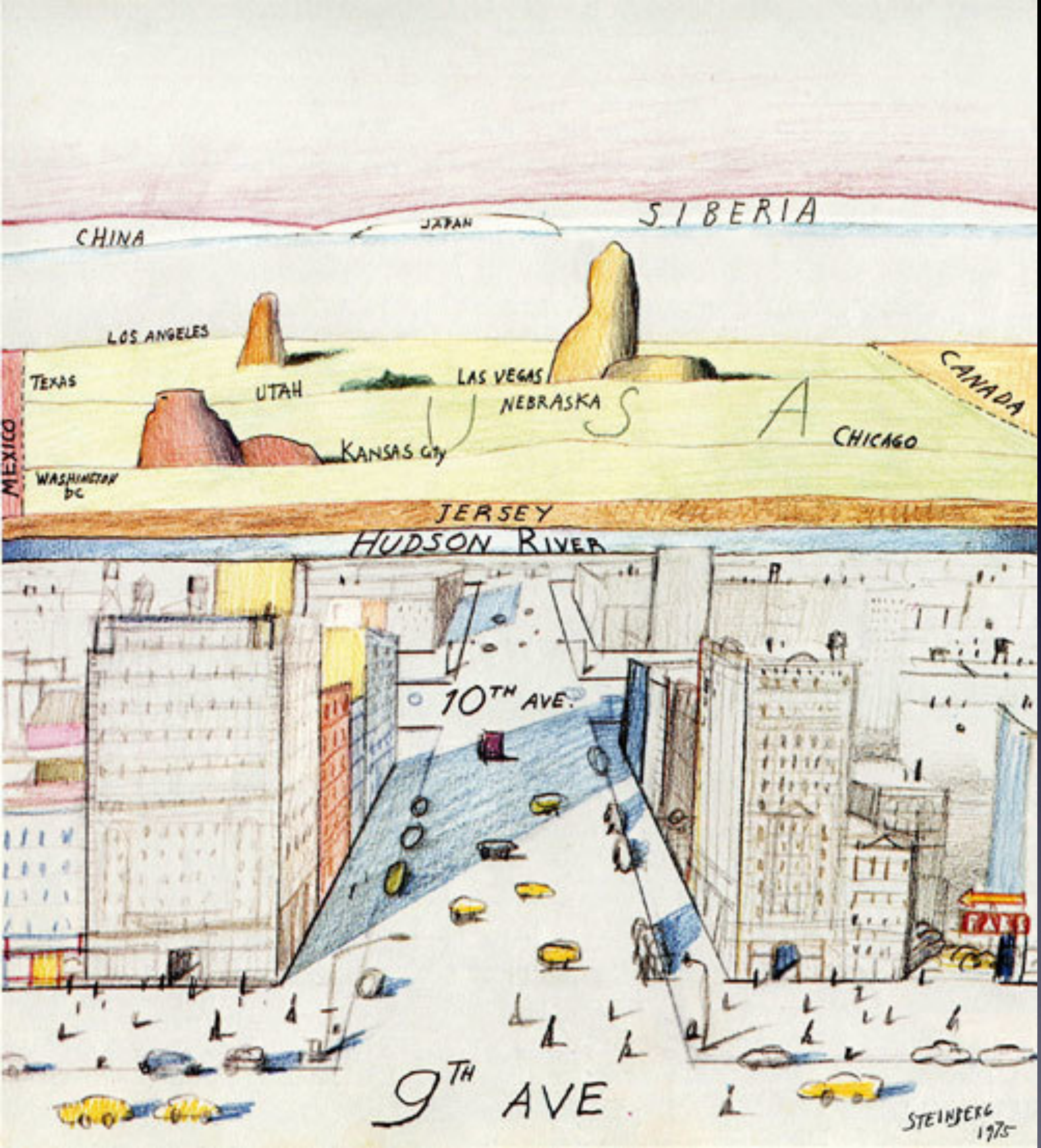
Flow Prediction



Observed Flow, Testing Data

- New York City ride-sharing vehicle data
- Daily trips between 263 taxi zones, for every Wednesday, June - December, 2017.
- Standard Deviation 23.48 trips/day


Taxi Zones -New York City
Jobs within 11 minutes
(Taxi, AM peak)
0 - 1,000
1,000 - 2,500
2,500 - 5,000
5,000 - 7,500
7,500 - 10,000
10,000 - 25,000
25,000 - 50,000
50,000 - 75,000
75,000 - 100,000
100,000 - 250,000
250,000 - 500,000
500,000 - 750,000
750,000 - 1,000,000
1,000,000 +
Roads
Water
0
5
10
15
20 km

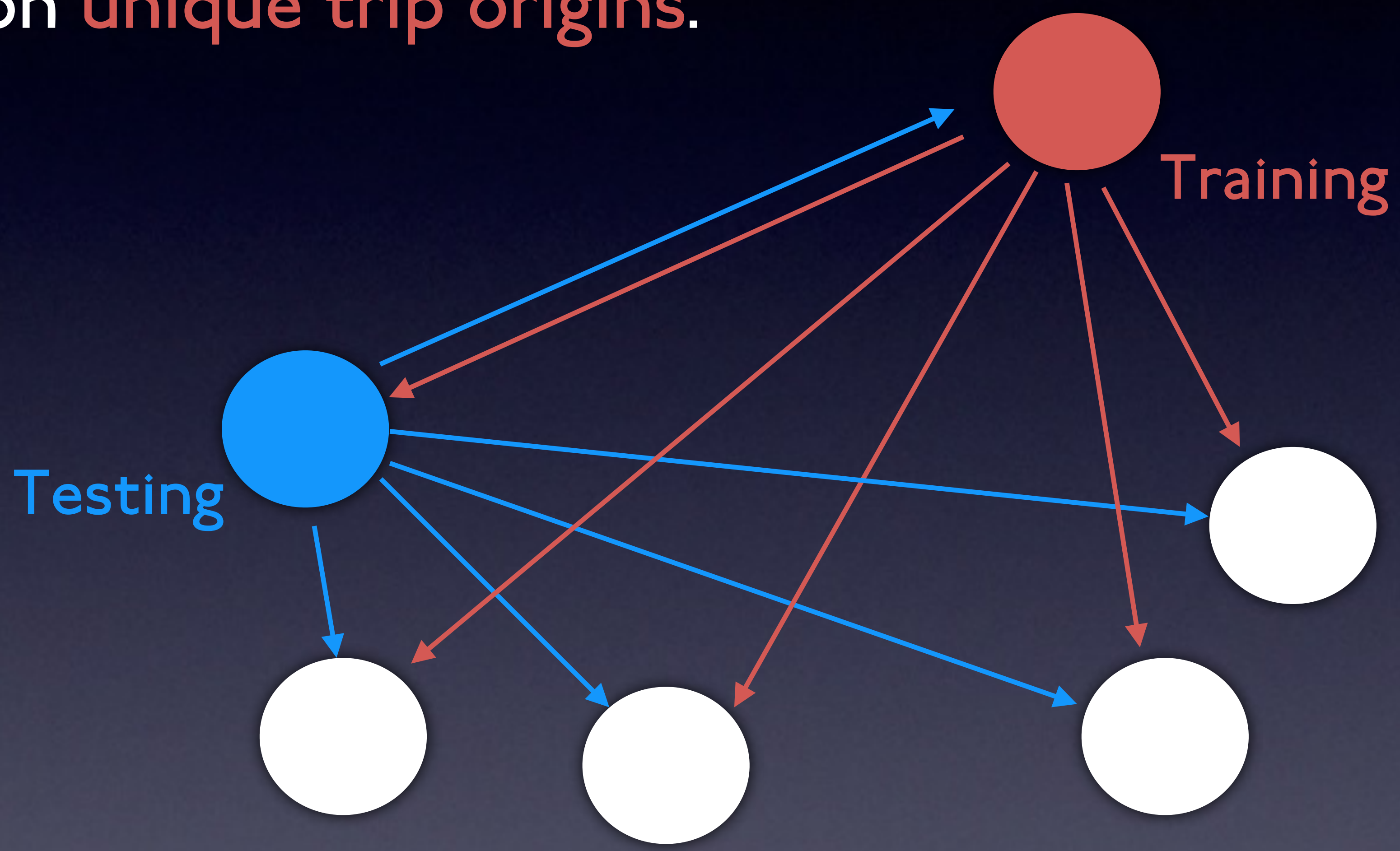


A New Yorker's view from the 9th Avenue

Source: https://curiator.com/art/saul-steinberg/view-of-the-world-from-9th-avenue

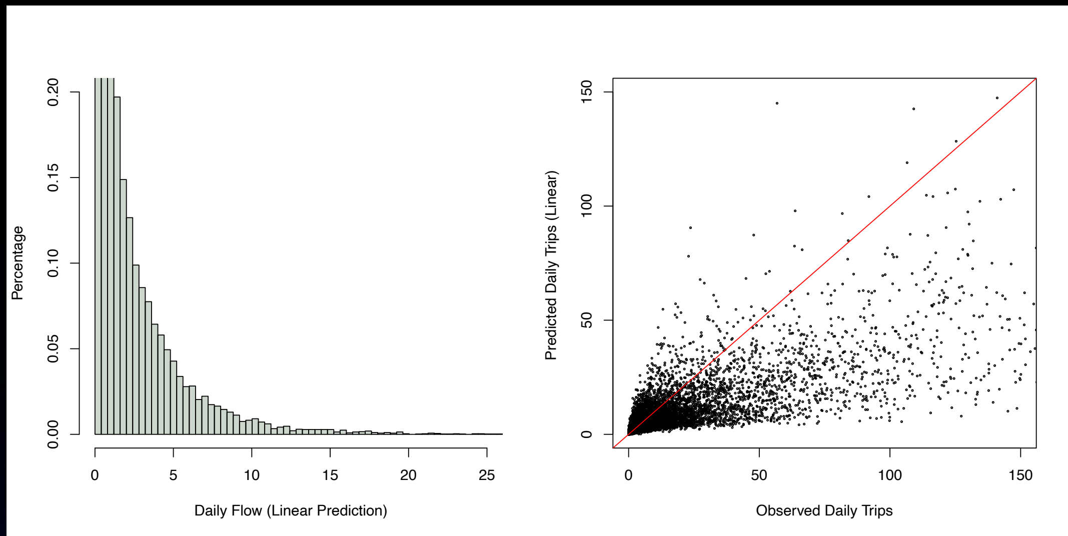
Flow Prediction

- To prevent the models from 'memorizing' the data, trips between OD pairs are separated into training, and testing data, based on unique trip origins.

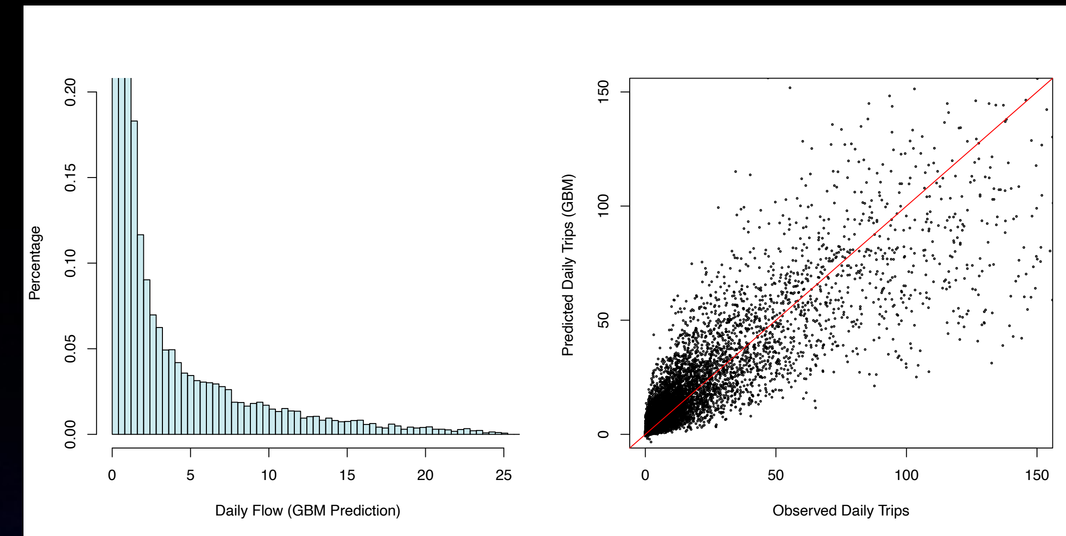


- Trips between 263 taxi zones,
- Over 10 million individual trips on 69,169 origin destination pairs.

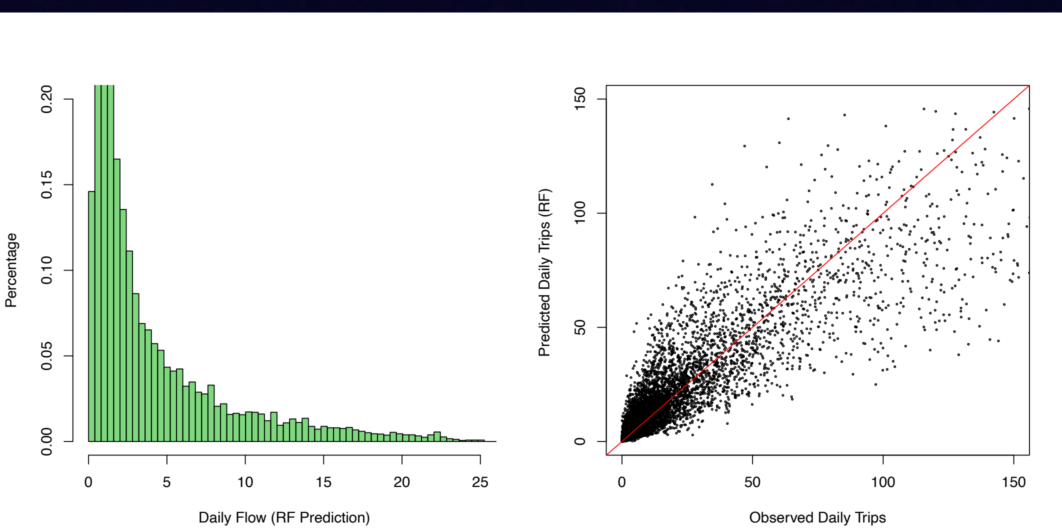
Linear Model (RMSE 16.61; MAE 4.50)



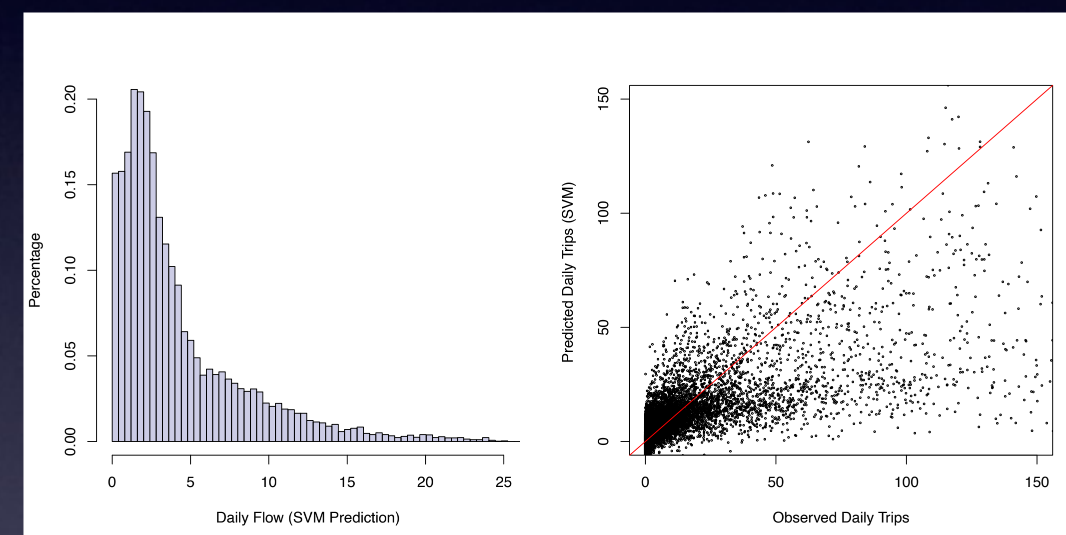
Gradient Boosting Machine (RMSE 11.05; MAE 3.37)



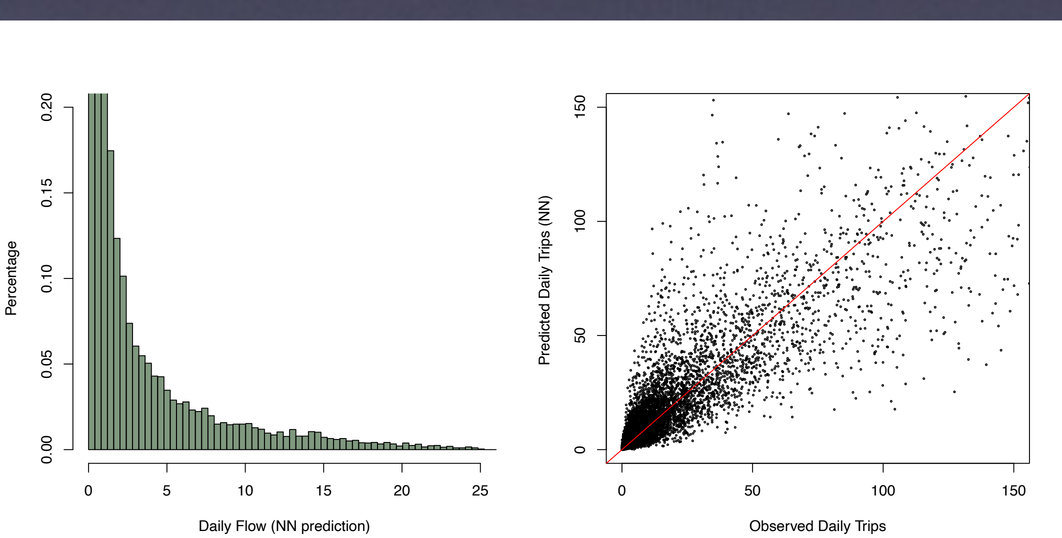
Random Forest (RMSE 11.24; MAE 3.13)



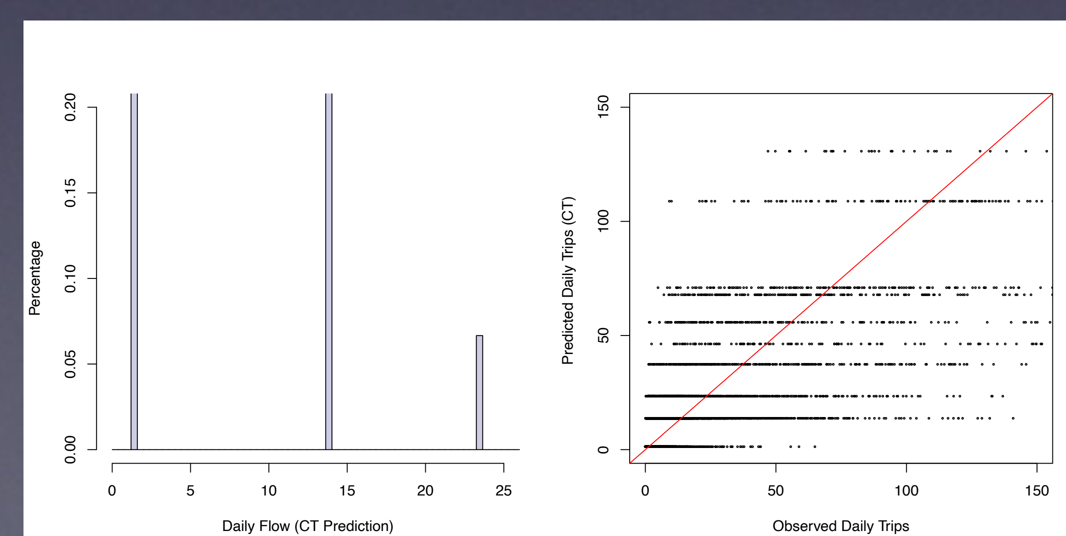
Support Vector Machine (RMSE 17.39; MAE 5.85)



Neural Network (RMSE 11.08; MAE 3.19)



Classification Tree (RMSE 14.85; MAE 5.11)



Not Entirely a Black Box

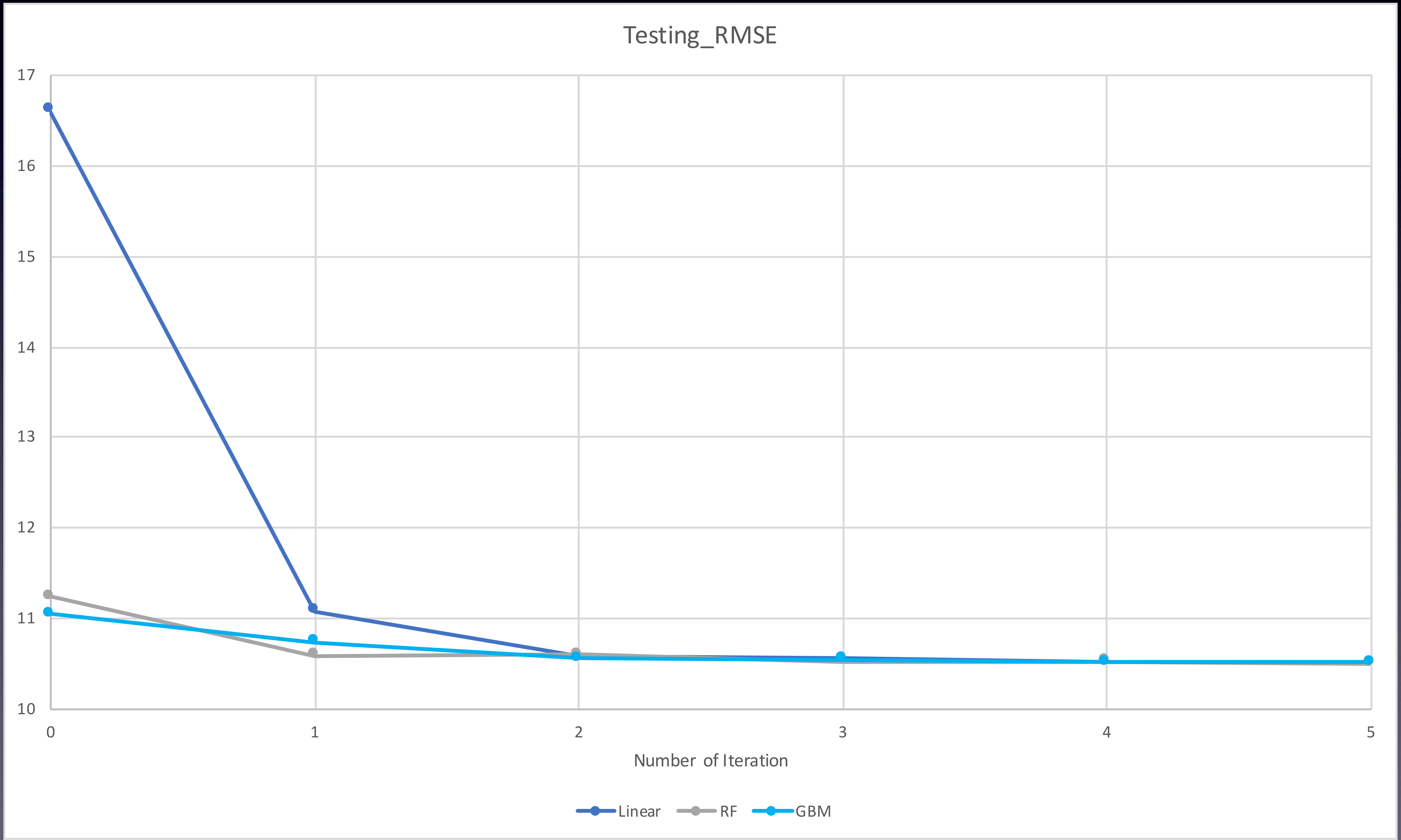
var	rel.inf
road_dist	60.8458037
D_jobs	9.8597064
O_jobs	6.6286250
D_workers	6.2706125
O_workers	5.2278565
O_NonF_rate	3.0864752
O_sqrkm	2.0824650
D_NonF_rate	1.7268486
D_sqrkm	1.6291277
O_tr_access	1.2924369
D_tr_access	0.6150512
D_HHMedian	0.4135431
O_HHMedian	0.3214482

Relative Influence of Variables (Gradient Boosting Machine)

- Variables can be added/dropped to test the sensitivity of the model's explanatory power.
- Some models are more interpretable than others

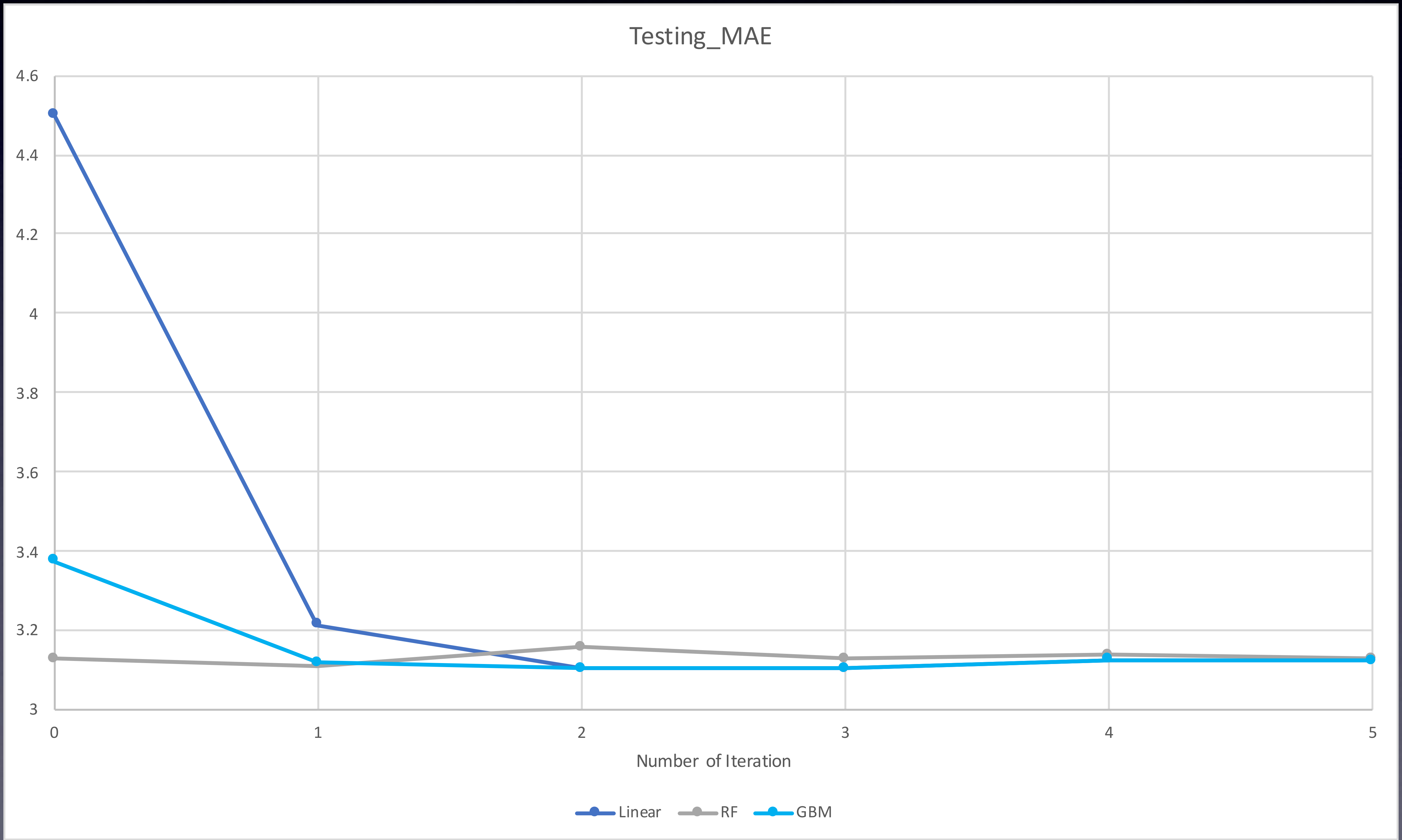
Boosting

5.10% lower RMSE than best performing model



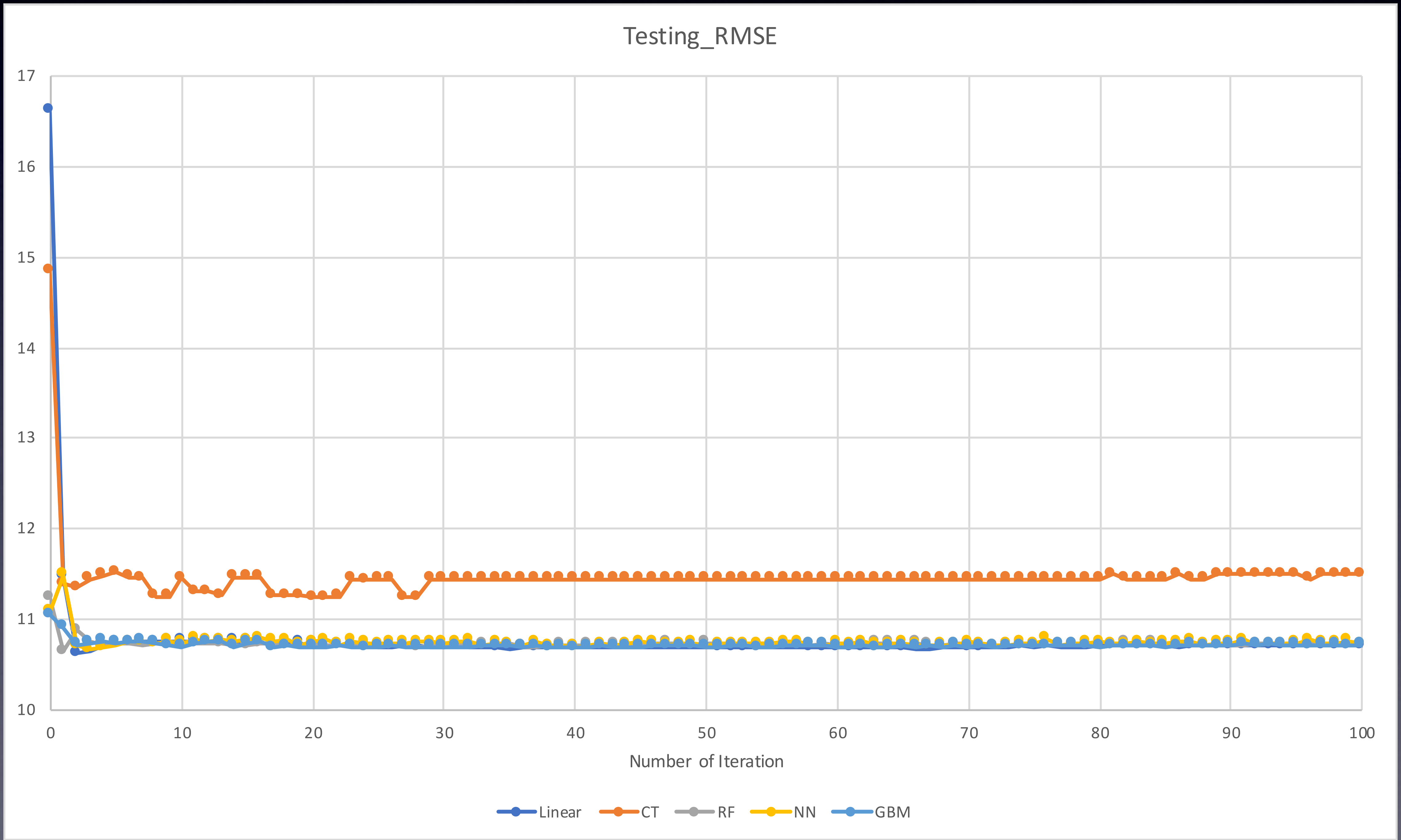
Boosting

0.38% lower MAE than best performing model



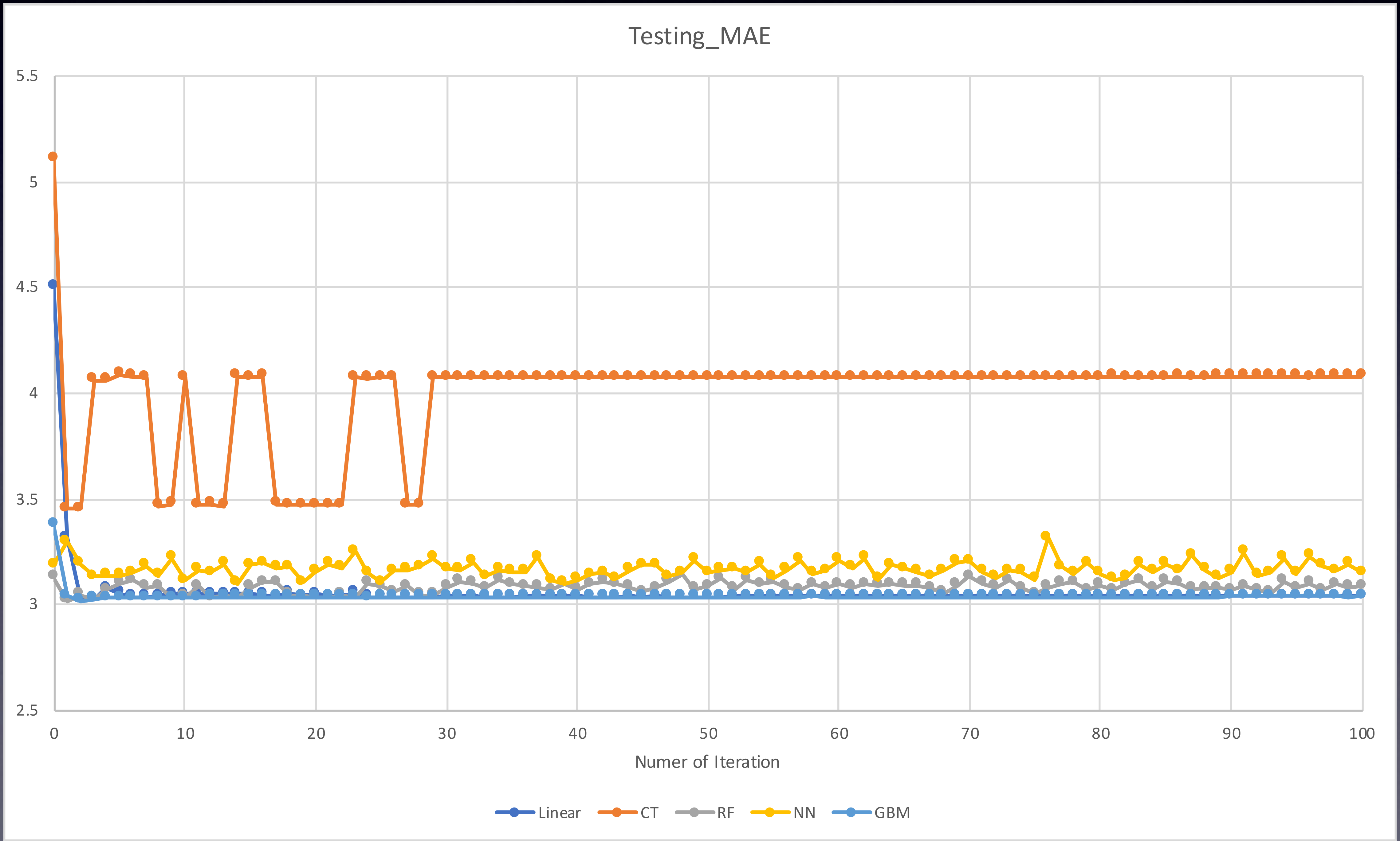
Deep Modeling

3.09% lower RMSE than best performing model



Deep Modeling

3.21% lower MAE than best performing model



One Cautionary Tale

- Noisy training data can cause the combination model to go after the noise in the training data, so lowering the accuracy of the combined forecast in the testing data

Answering the Research Questions

- Machine Learning models are well equipped to deal with complex transport problems

- An ensemble of models better utilize existing data

Questions ?

Fare and Travel Time Forecast