KDDCUP 2017 Volume Prediction

Step by step modeling for travel volume prediction

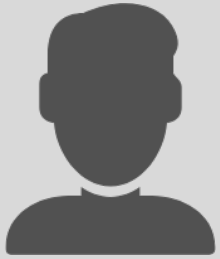
Team：Convolution

Huan Chen

Beihang University(Master Candidate, CS)

Pan Huang

Bing Ads, Microsoft



Ke Hu

Bing Ads, Microsoft

Peng Yan

Meituan-Dianping

Outline

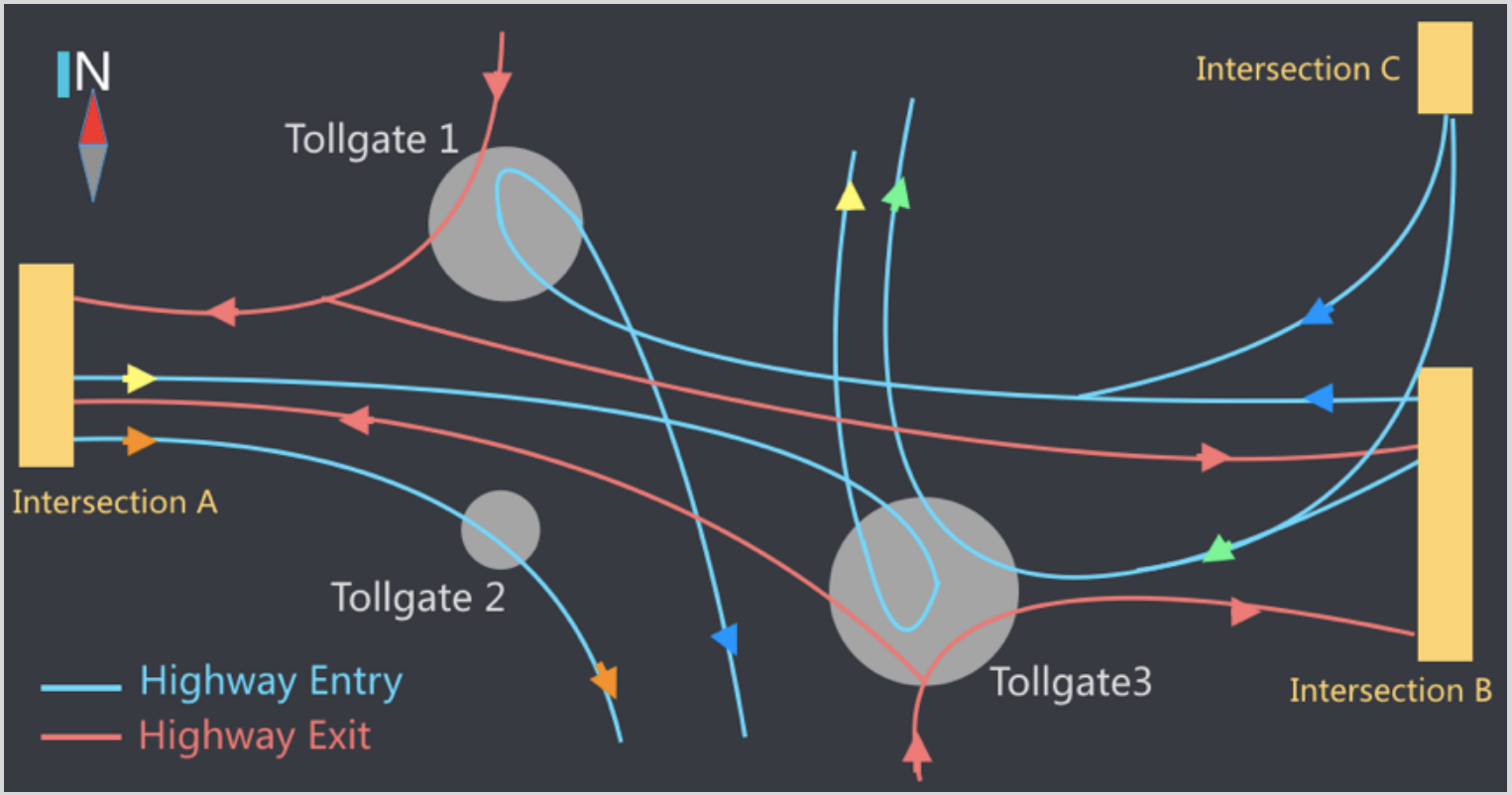
- Problem understanding
- Features
- Models
- Conclusion

Problem definition

Task 2: Volume Prediction

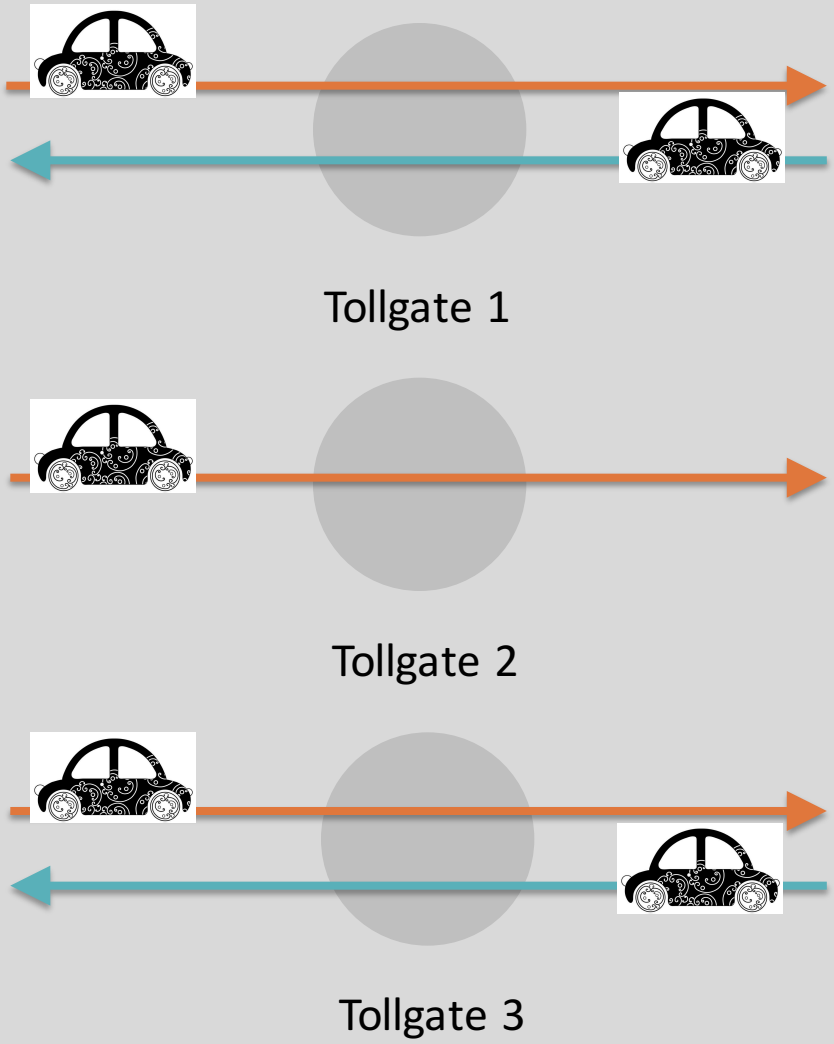
For every 20-minute time window, predict the entry and exit traffic volumes at tollgates 1, 2 and 3.

$$MAPE = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{T}\sum_{t=1}^{T}\left|\frac{d_{rt}-p_{rt}}{d_{rt}}\right|\right)$$



Problem understanding

Problem definition

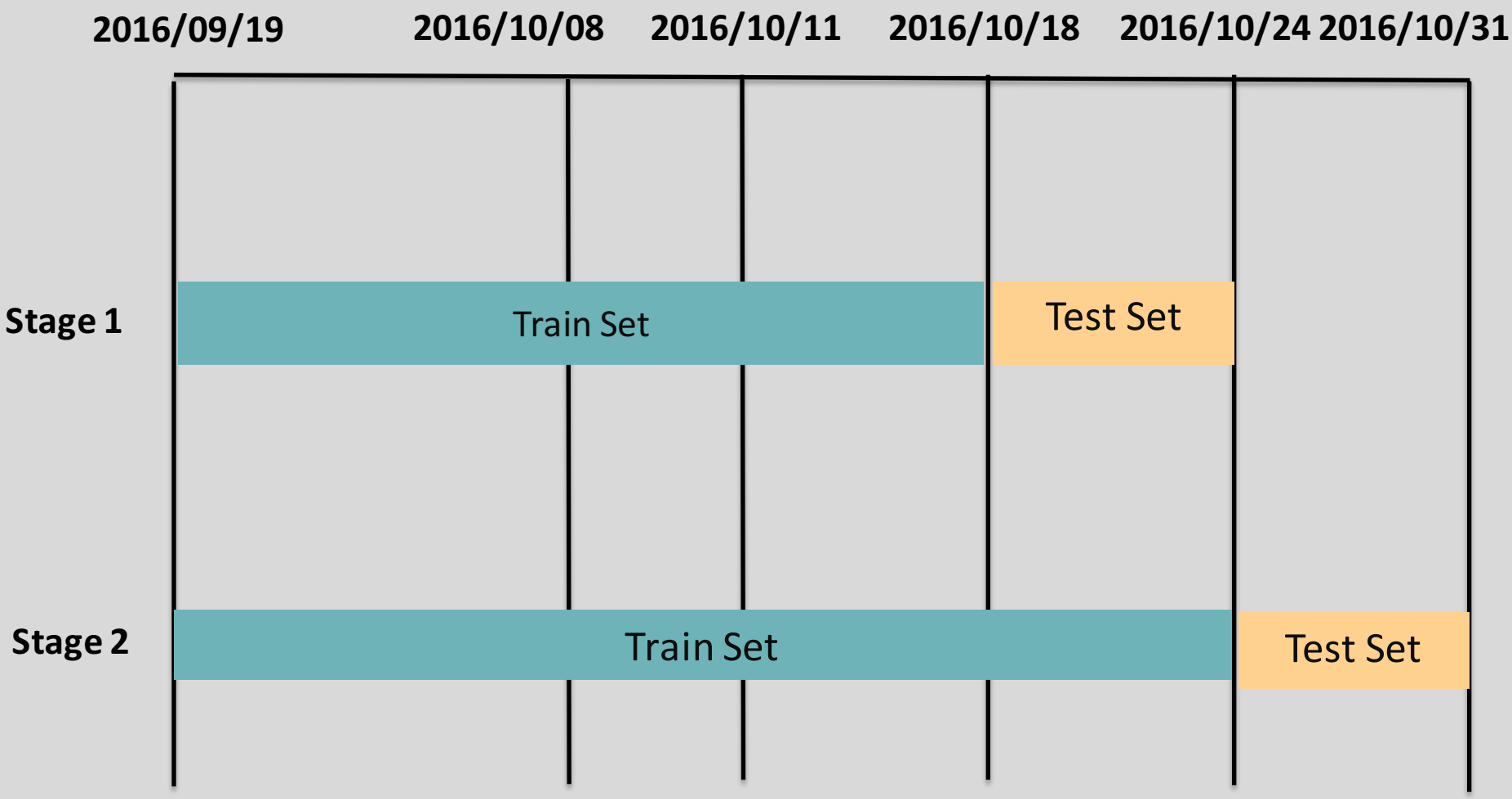


Total:

5 tollgate-direction

Problem understanding

Problem definition

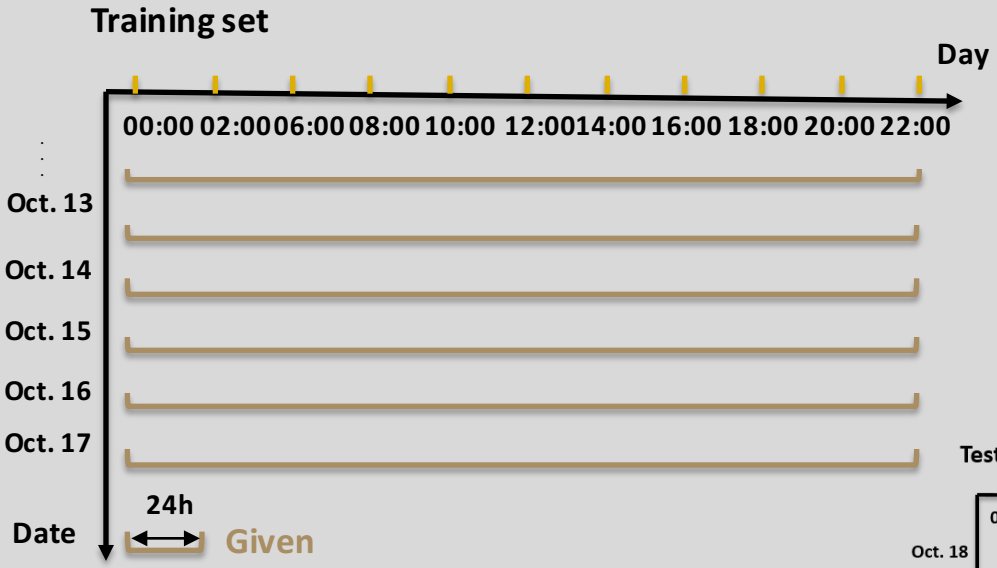


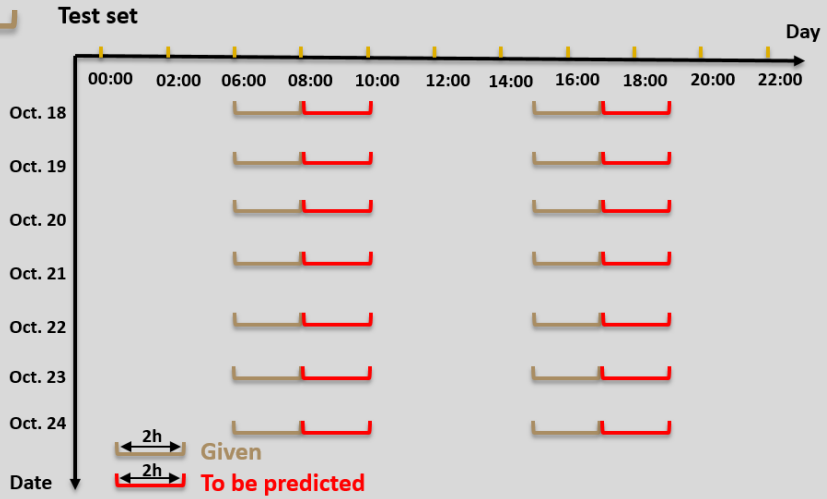
Problem understanding

Problem definition

Stage1: 4 weeks training data + **1 week** test data

Stage2: 5 weeks training data + **1 week** test data





Problem understanding

Problem definition

Differences from Average Travel Time Prediction Task

1. Travel volume values are much noisier comparing to average travel time
2. Dataset is much smaller(Stage1)
   - **3 month** training data for task1
   - Only **4 weeks** training data for task2, and holidays' traffic is abnormal

Strategy

1. Multiple offline validation sets
2. Using multiple **simple models for ensemble**
   - Complex models are unstable on such dataset
3. Utilizing as much data as possible
   - **Step by step modeling** to "squeeze information from test data"

Problem understanding

- Select validate set

Slide Window

Stage 1 (total : 4 sets)

2016/10/08 2016/10/17

Stage 2 (total: 7 sets)

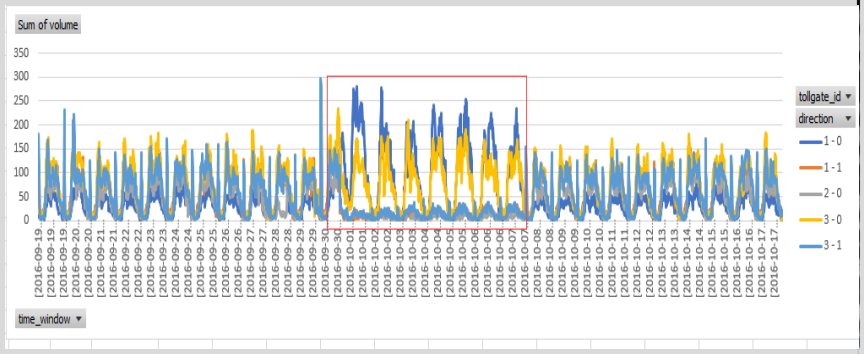
2016/10/08 2016/10/17

Problem understanding

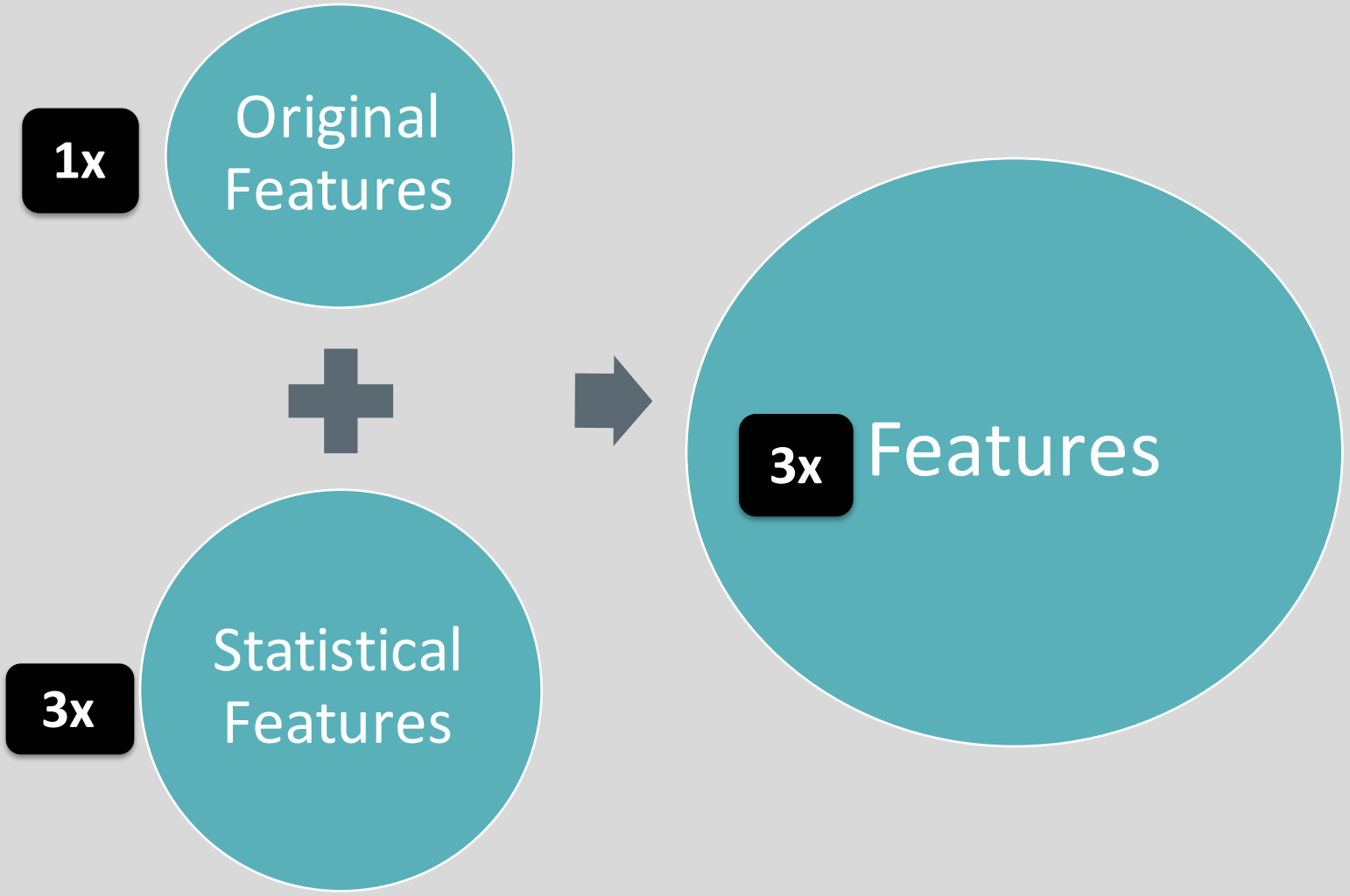
Noise reduction

- Mid-Autumn Festival
- National Day

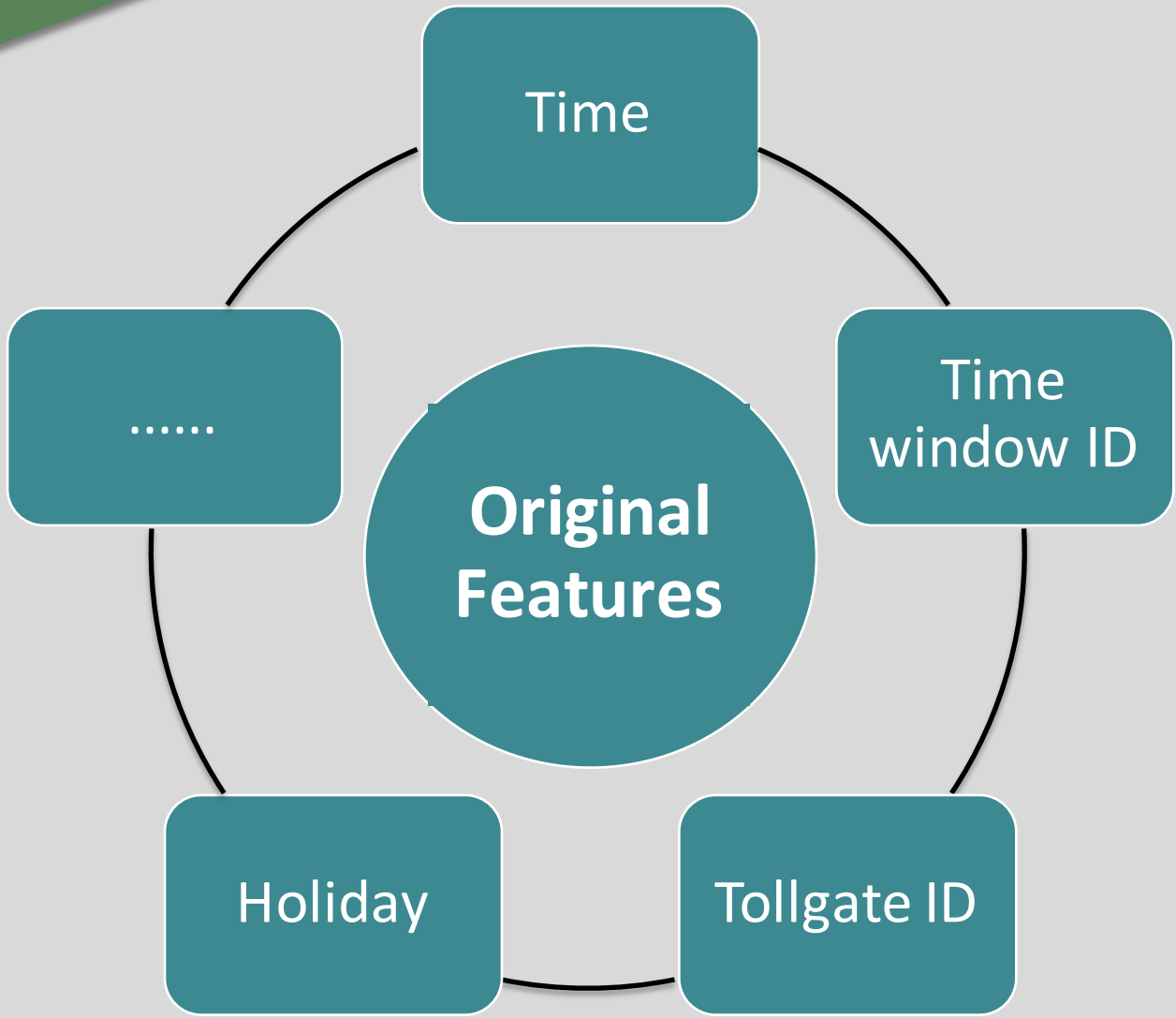
Every day's data in 23:00-01:00



Features



Features



Features

Statistical Features

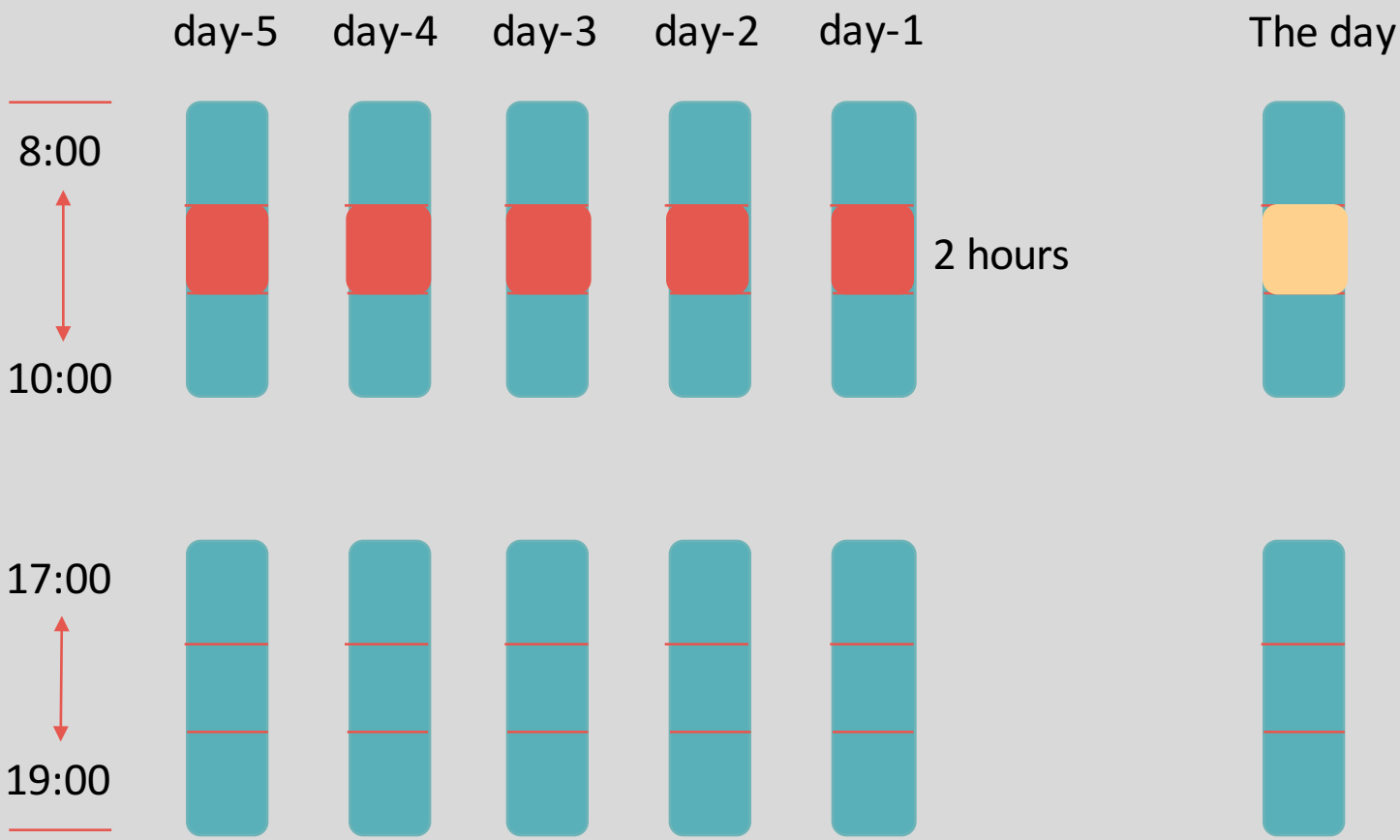
1. Statistical features of the corresponding window in the last n days
2. Statistical features of all samples in the last n days
3. Statistical features of rush hours in the last n days

Features

■ Example (n = 5)

7 statistical features of the corresponding window in the last n days

Blue: unused data
Red: used data
Yellow: target sample



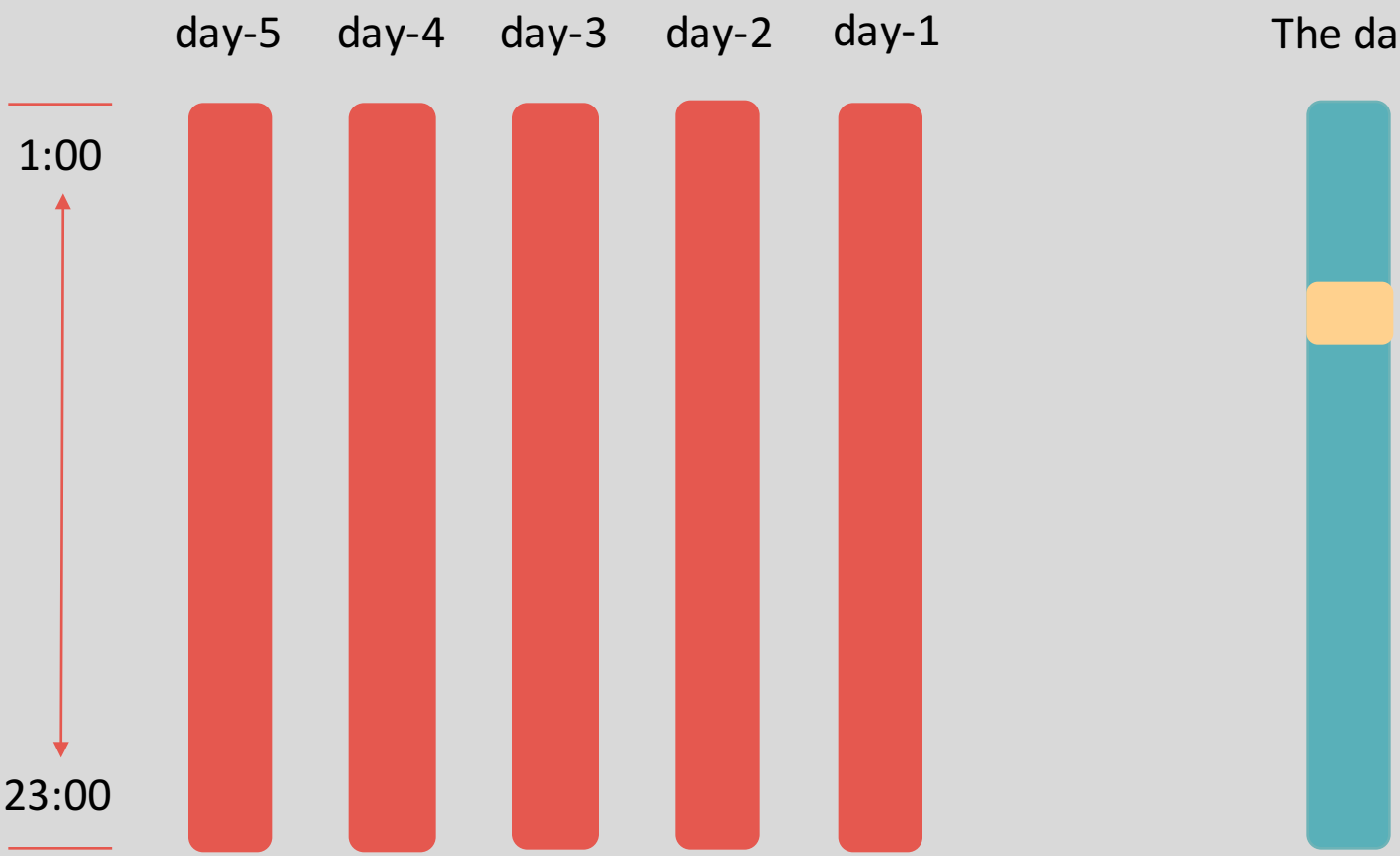
Features

- Example (n = 5)

2

statistical features of all samples in the last n days

Blue: unused data
Red: used data
Yellow: target sample

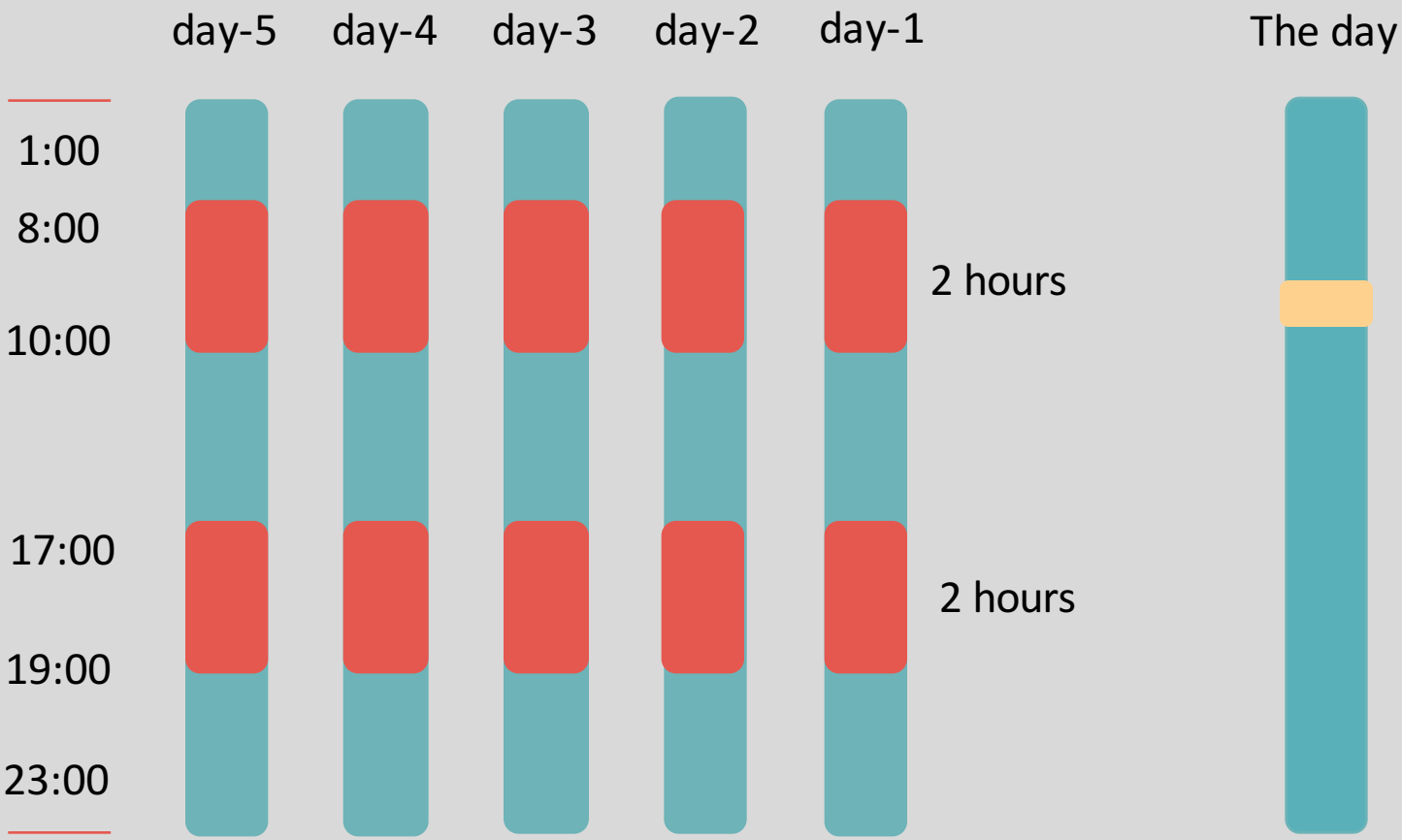


Features

■ Example (n = 5)

3 statistical features of rush hours in the last n days

Blue: unused data
Red: used data
Yellow: target sample



Features

Statistical features:

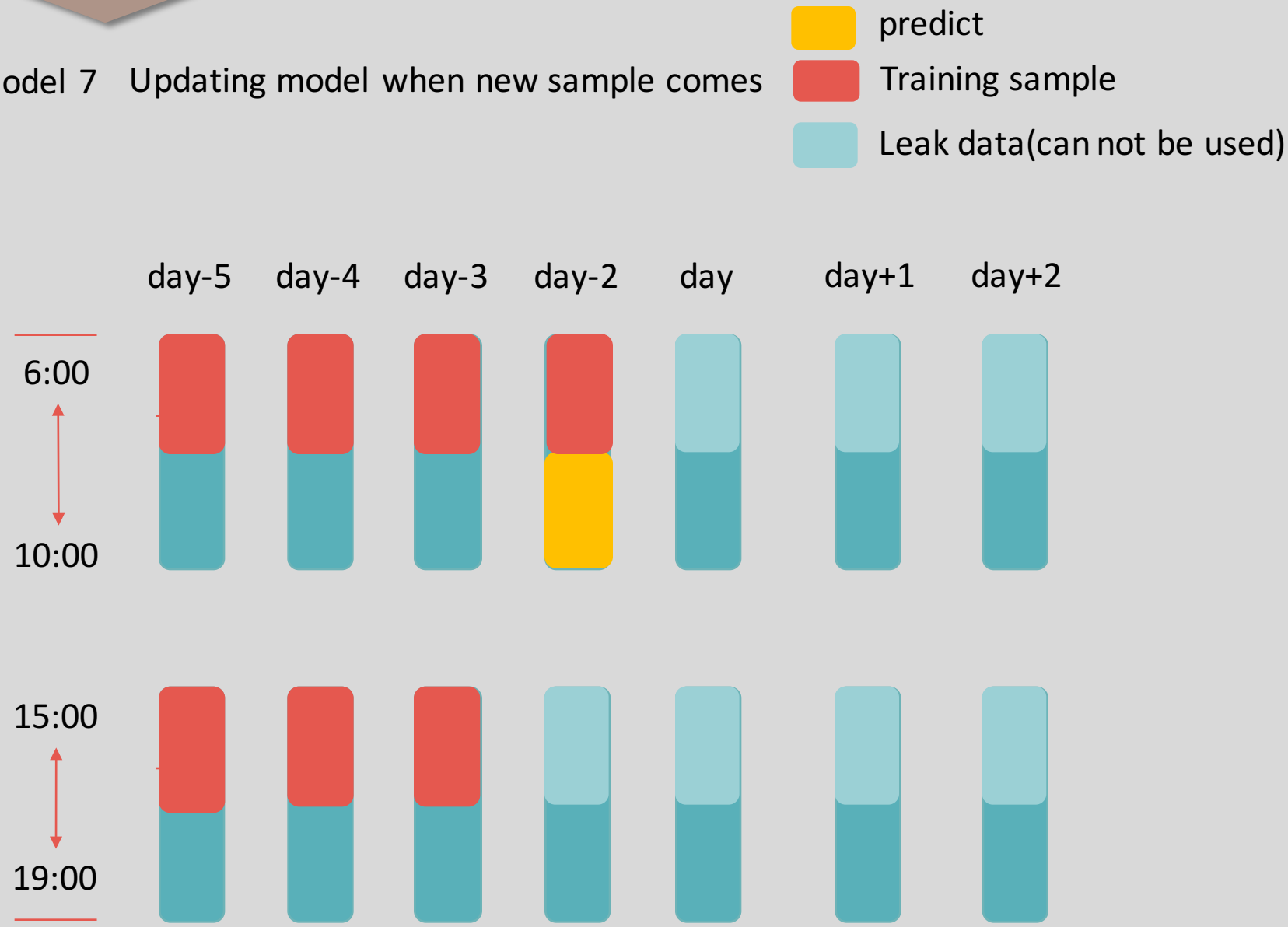
1. Statistical functions: max, min, average, medium, slope, skewness, ratio, diff, percentage, rank etc.

2. Statistical dimensions: forward/reverse direction, car type: etc/not etc type …

Models

Step by step modeling

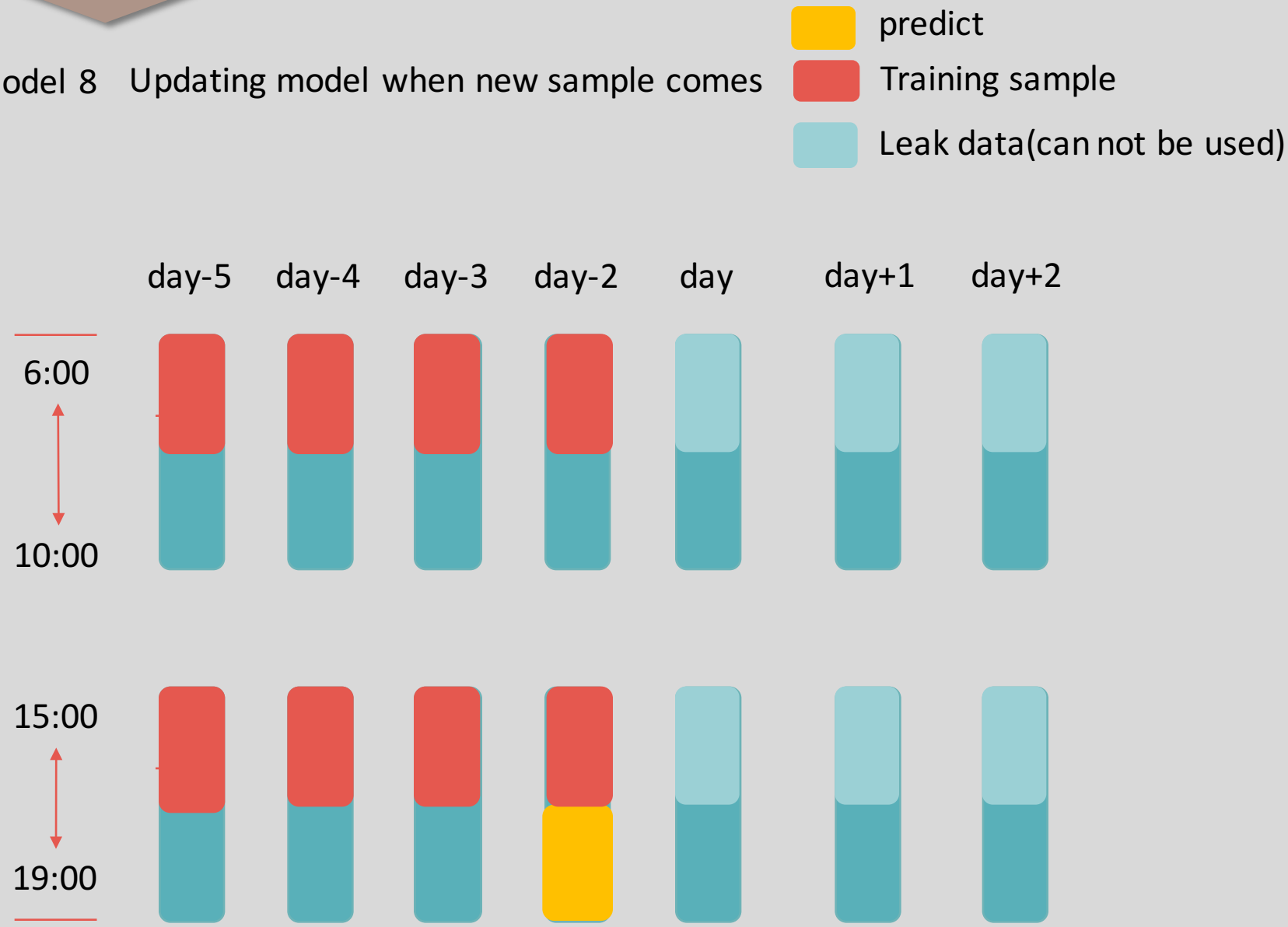
Model 7 Updating model when new sample comes



Models

Step by step modeling

Model 8 Updating model when new sample comes



Models

Step by step modeling

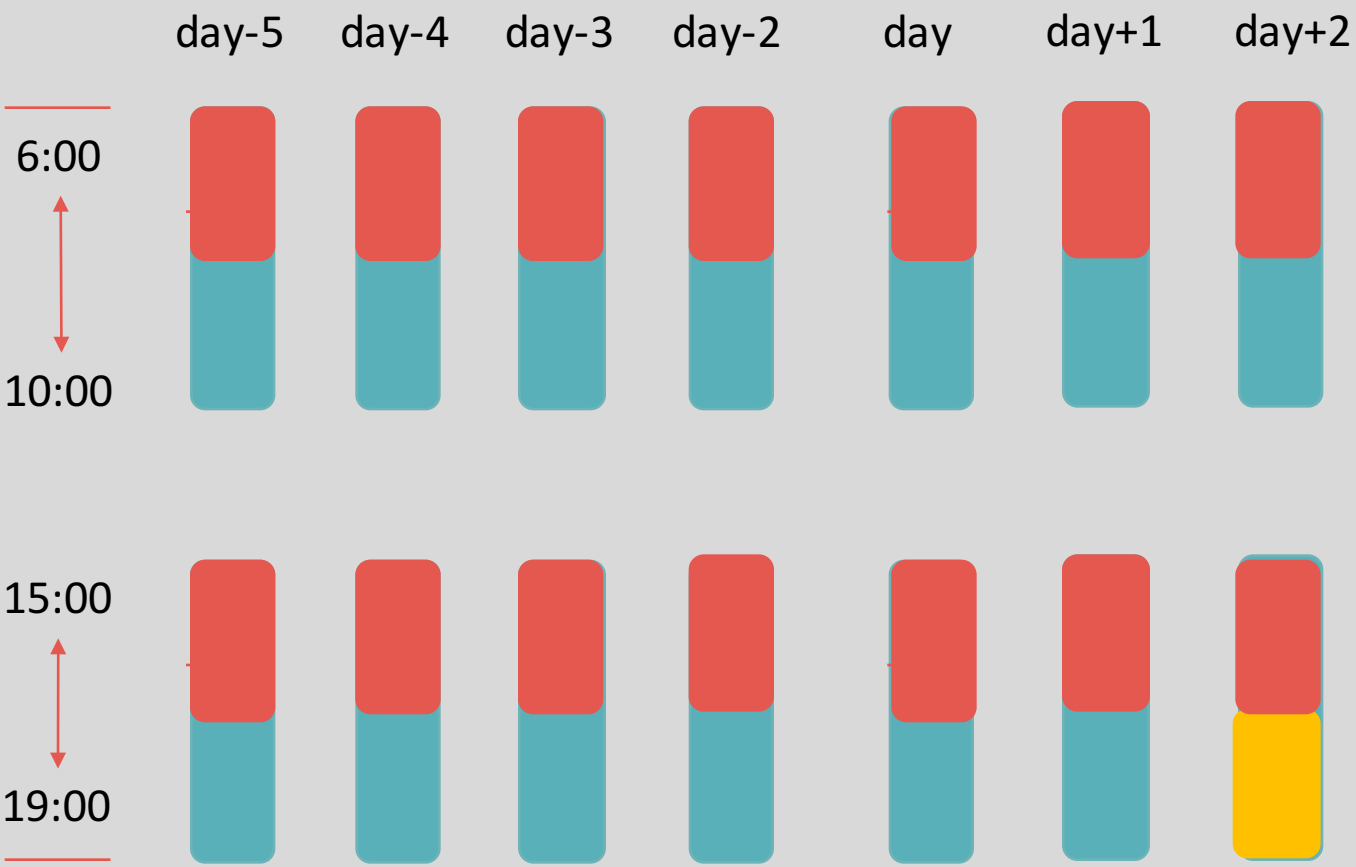
Model 14

1. “Updating” model when new sample comes
2. Better utilize test data

predict

Training sample

Leak data(can not be used)



Models

Xgboost

Fully-connected NN

LightGBM

Models

Label

Use **log** and **boxcox** to Reduce the influence of outliers

Objective function

In Xgboost:

Use **fair**[1] instead of rmse as the objective function

Weights of samples

Increase the weight of **small** samples:

weight=1/sqrt(y)

y is the label of the sample.

Model ensemble

Calculate weighted mean

[1] M-estimators:http://www.statisticalconsultants.co.nz/blog/m-estimators.html

Conclusion

Simple model to handle noisy data

Produce diversity from model, data, feature, loss function

Fully utilize all the information in data

Don’t waste any useful data points

Thanks

Q&A