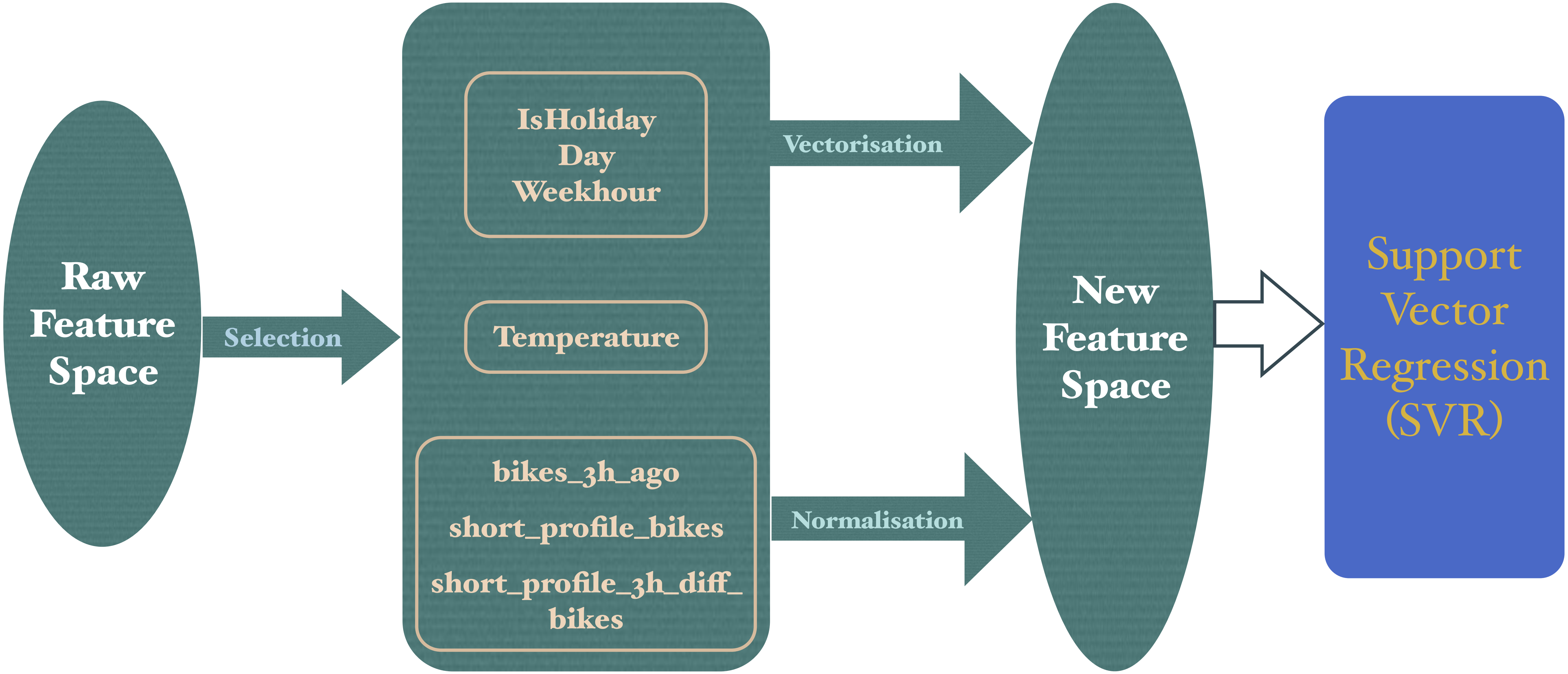
SVR-BASED MODELLING FOR THE MOREBIKES CHALLENGE

Analysis, Visualisation and Prediction

Yu Chen, Peter Flach



Raw Feature Space

Selection

IsHoliday
Day
Weekhour

Temperature

bikes_3h_ago

short_profile_bikes

short_profile_3h_diff_bikes

Vectorisation

Normalisation

New Feature Space

Support Vector Regression (SVR)

Model Structure

Raw Feature Space

Facts of Stations

Station ID
Latitude
Longitude
Number of Docks

fixed over time

Statistics

bikes_3h_ago

full_profile_bikes

full_profile_3h_diff_bikes

short_profile_bikes

short_profile_3h_diff_bikes

full profiles are
not aligned
2 years vs. several weeks

Weather

temperature

windMaxSpeed

windDirection

relHumidity

windMeanSpeed

precipitation

airPressure

shared by all stations,
linear models only
selected temperature

Temporal information

Timestamp

Year

Month

Day

Hour

Weekday

Weekhour

IsHoliday

Overlapping
tbc.

Relations between Temporal Information

Fixed year & month: Oct. 2014

Temporal information

Timestamp

Year

Month

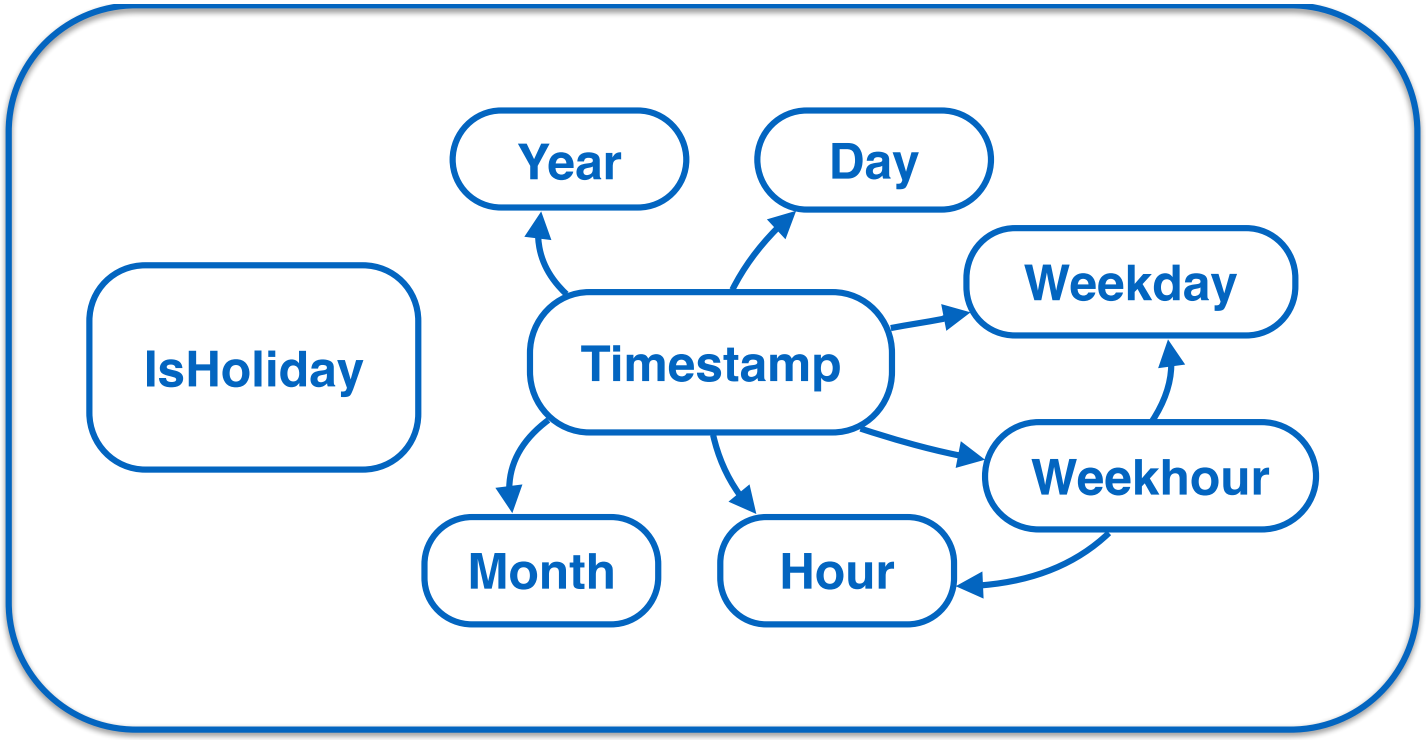
Day

Hour

Weekday

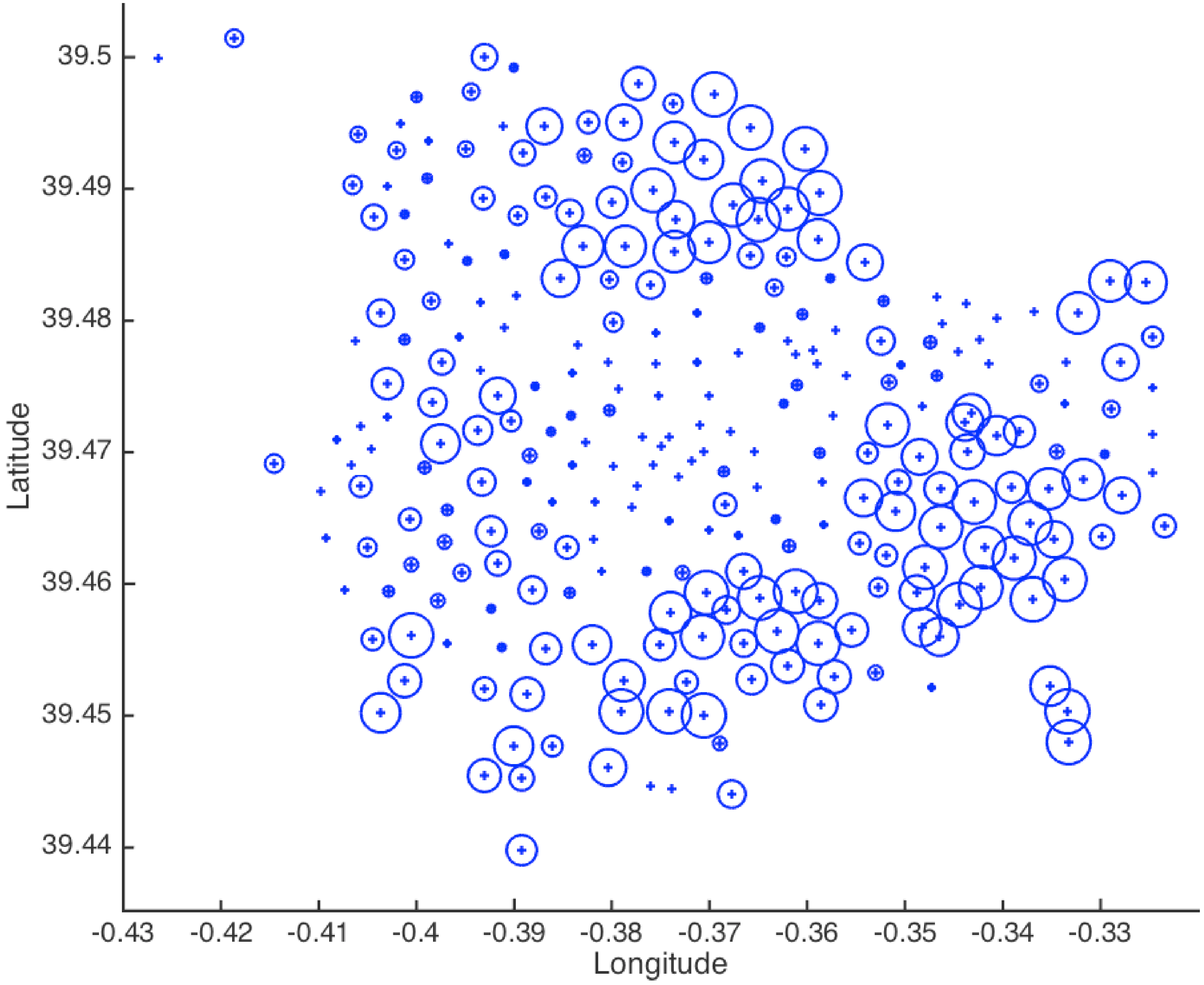
Weekhour

IsHoliday



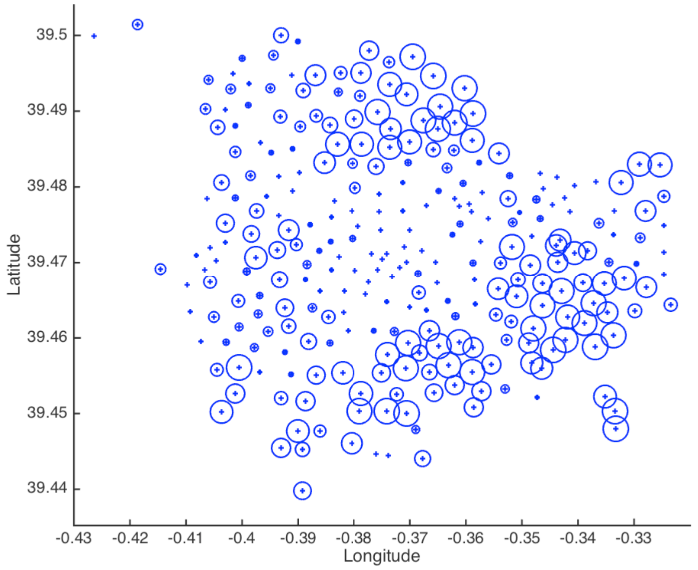
Timestamp?
Can not tell periodical similarity

Weekday + Hour vs. Weekhour ?
Unknown

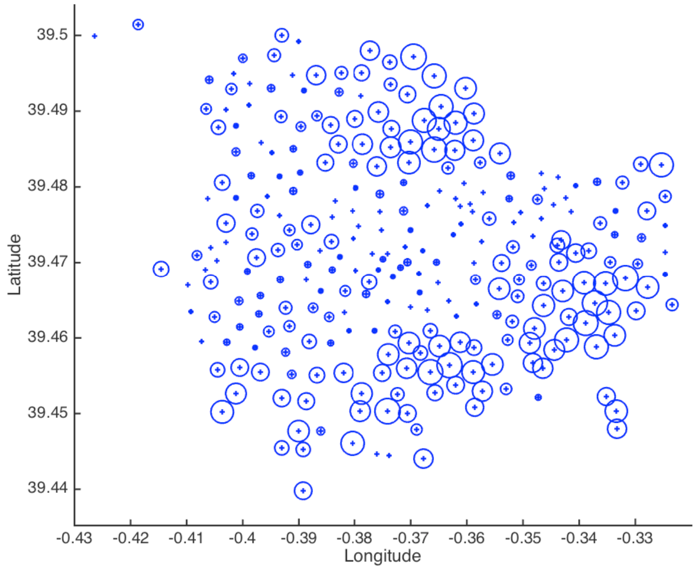


Wednesday, hour: 0

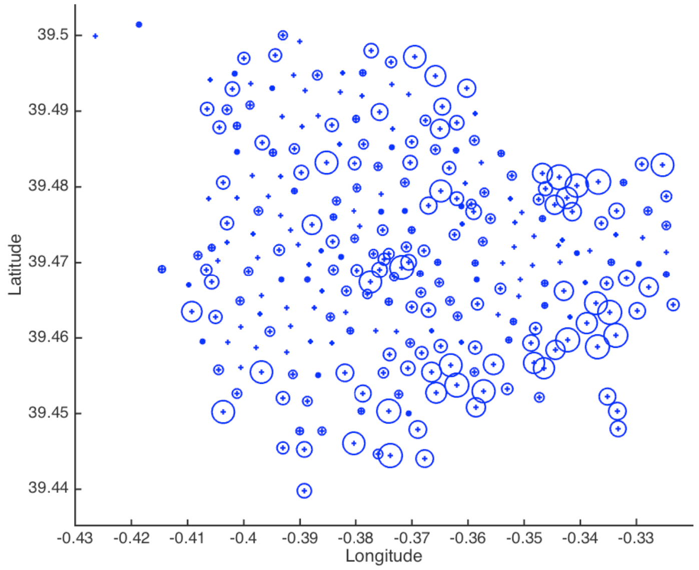
Changes of the bike storage over all stations in a workday:



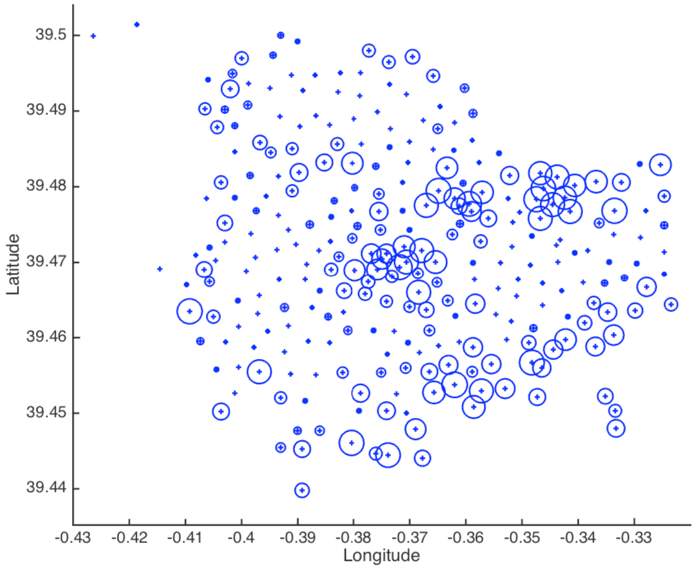
hour: 0



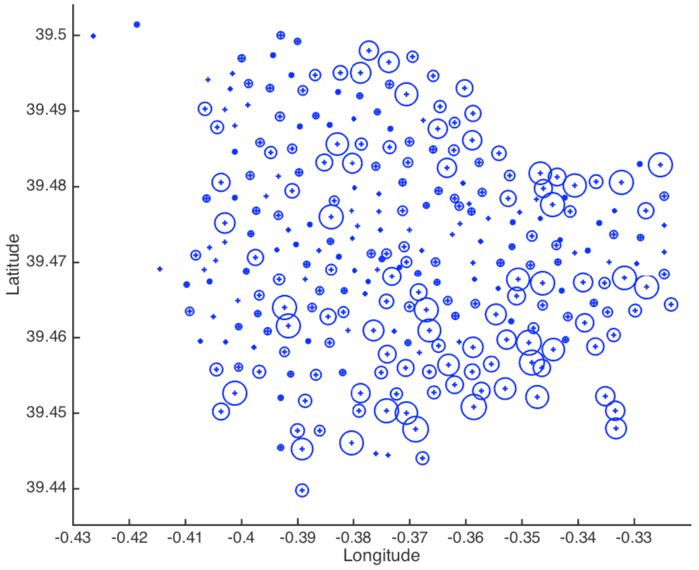
hour: 7



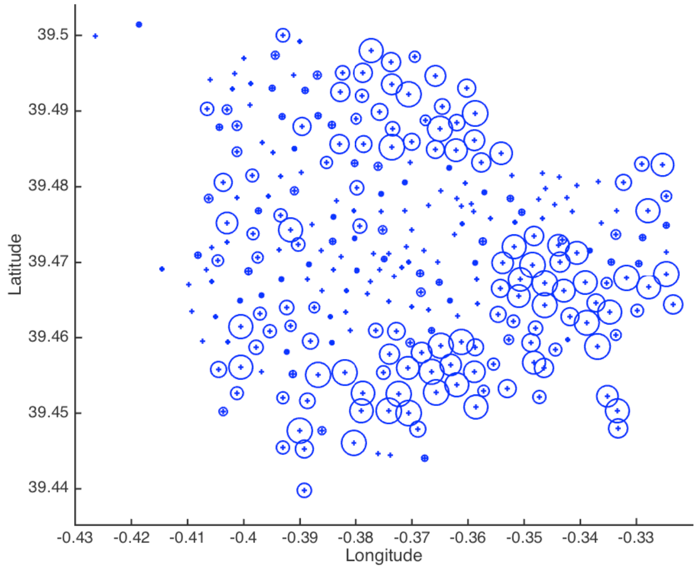
hour: 8



hour: 10

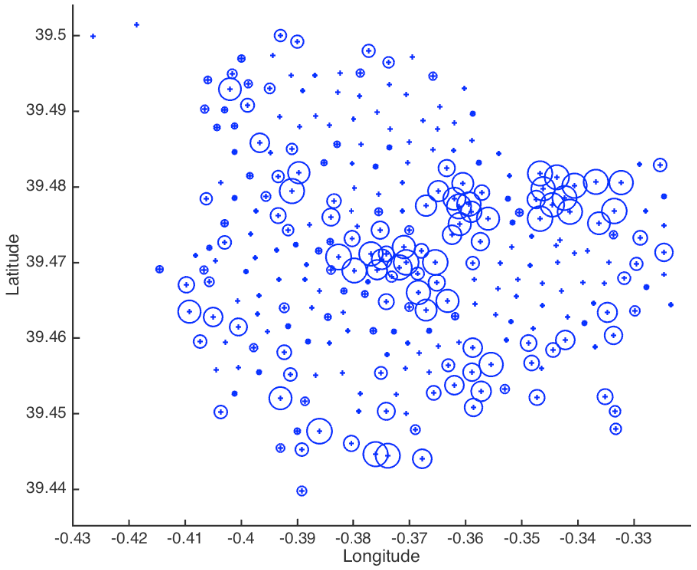


hour: 16

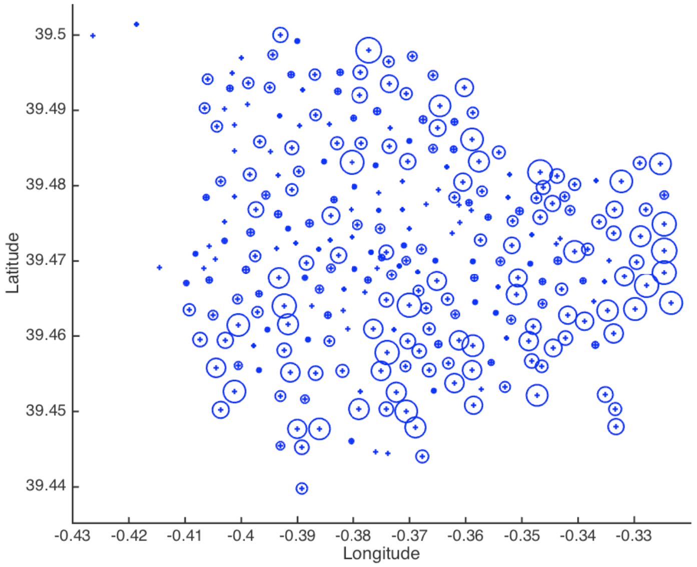


hour: 21

Workday

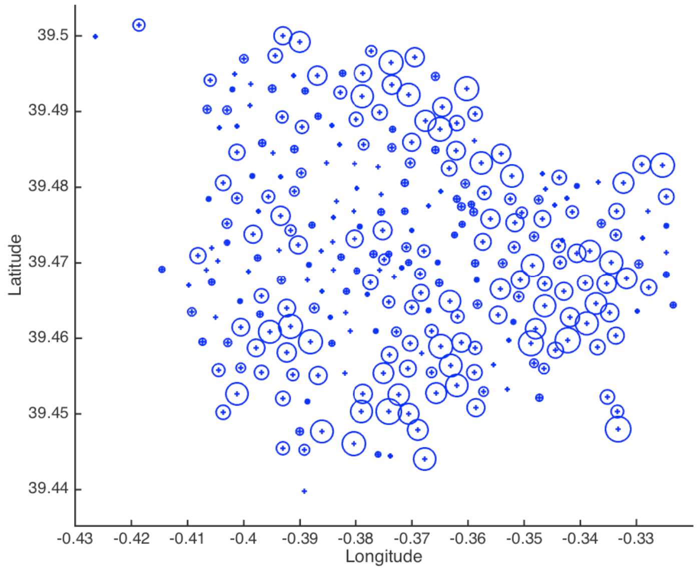


hour: 10

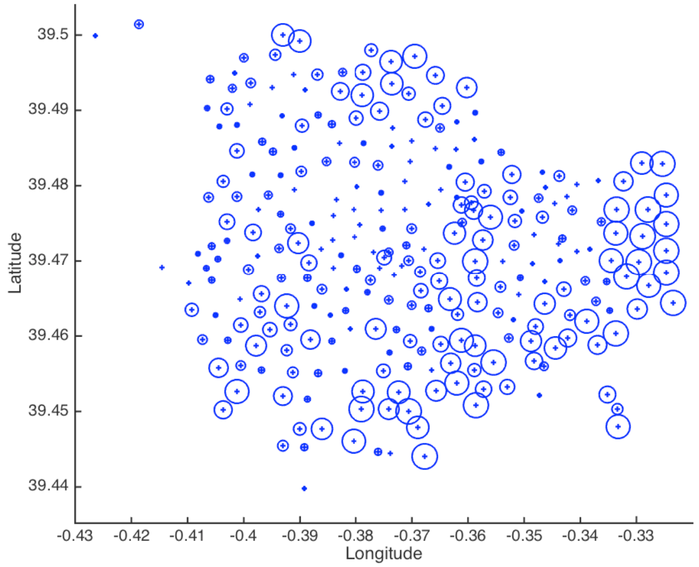


hour: 16

Weekend

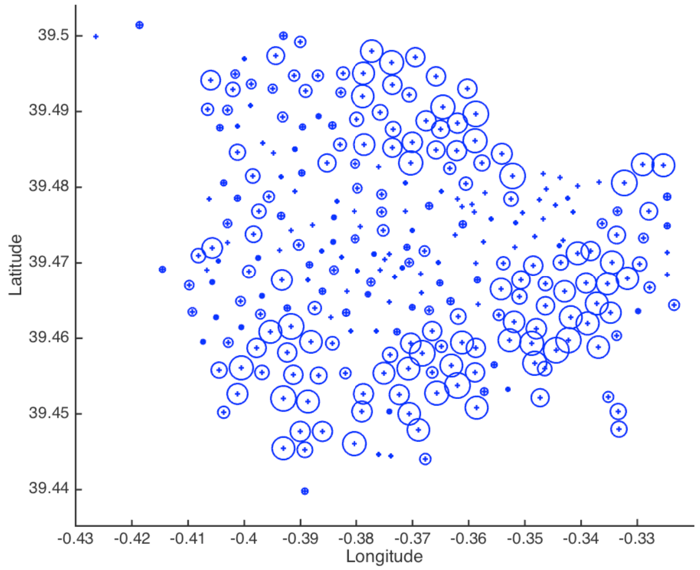


hour: 10

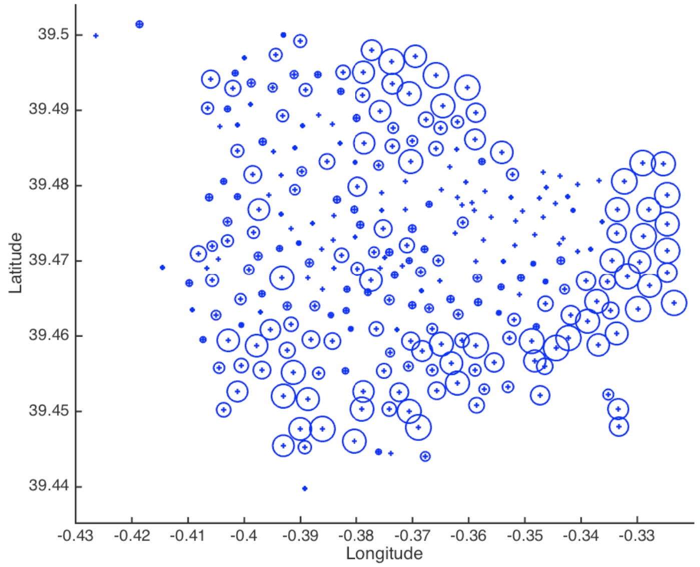


hour: 16

Holiday



hour: 10



hour: 16

- **Conclusion: Weekhour is more accurate than Weekday + Hour to represent the similarity between two time points.**

- **Conclusion: Location distance or station identifier can not represent similarity between two stations.**

Facts of Stations

Number of Docks

Weather

temperature

Statistics

bikes_3h_ago

short_profile_bikes

short_profile_3h_diff_bikes

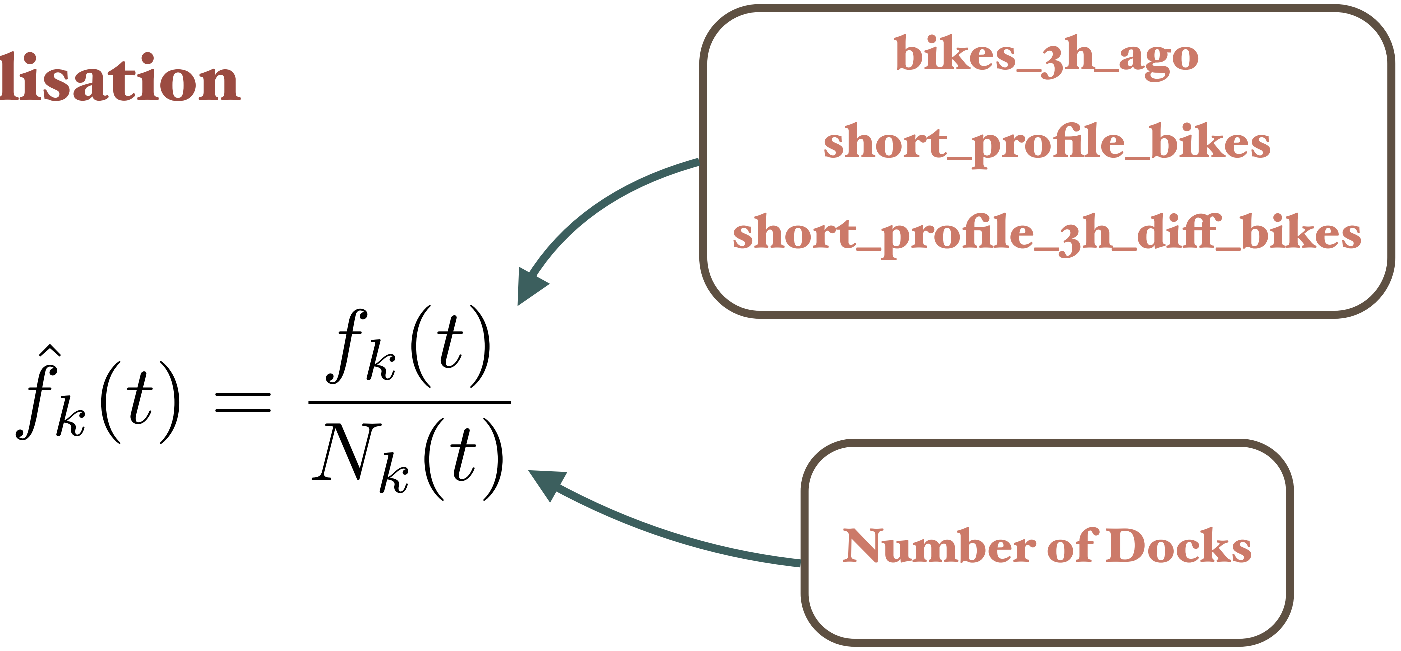
Temporal information

Day

Weekhour

IsHoliday

Normalisation



bikes_3h_ago

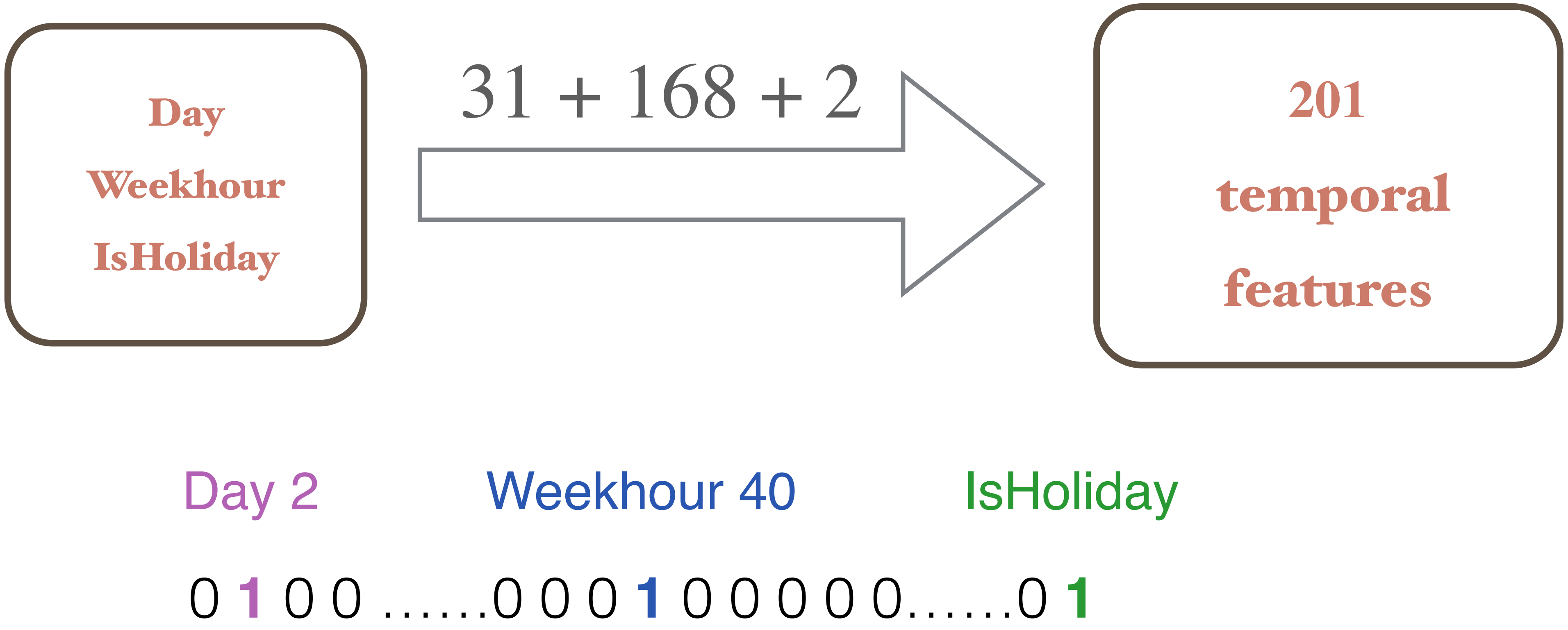
short_profile_bikes

short_profile_3h_diff_bikes

$$\hat{f}_k(t) = \frac{f_k(t)}{N_k(t)}$$

Number of Docks

Vectorisation



Day

Weekhour

IsHoliday

31 + 168 + 2

201

temporal

features

Day 2 Weekhour 40 IsHoliday

0 1 0 0 ......0 0 0 1 0 0 0 0 0......0 1

ε-Support Vector Regression Model

Target Value: $\hat{y}_k(t) = \dfrac{y_k(t) - y_k(t-3h)}{N_k}$

Kernel Function: $K_{ij} = \tanh(\gamma x_i^T x_j + c_0)$

Parameters: $C = 2, \epsilon = 0.02, \gamma = 0.25, c_0 = -1$

Configurations of SVR model are selected by fast test.
The implementation is from scikit-learn.

Fast test: choose a small subset of the data to train a SVR model for each station.

Training dataset for station i:

- the data of K stations which are nearest to the station i

$$Distance = \sqrt{\sum_{t=1}^{T} \left(\hat{y}_i(t) - \hat{y}_j(t)\right)^2}$$

for testing: K = 10;

for leaderboard submission: K = 20

Validation datasets:

- the data of 75 new stations in October 2014;
- the data of 10 old stations in November, December and January from 2012 to 2014

Leader board attempts

	Size of Training Set	Feature Options		
MAE	K	”Weekhour”	”Weekday”+”Hour”	Full Profiles
2.625	20		✓	✓
2.612	20	✓		✓
2.52	20	✓		
2.496	275	✓		✓
2.46	275		✓	
2.37	275	✓		

Final Model

Training set: data of 275 stations in October 2014

Features:

Normalised	Vectorised	Untouched
bikes_3h_ago	Day	temperature
short_profile_bikes	Weekhour	
short_profile_3h_diff_bikes	IsHoliday	

Thank You