Microsoft Research

# Mobility Modeling and Prediction in Bike-Sharing Systems

Zidong Yang, Ji Hu, Yuanchao Shu, Peng Cheng, Jiming Chen and Thomas Moscibroda

Microsoft

# Growing Dramatically Worldwide


# Growing Dramatically Worldwide

- > 500 bike-sharing systems
- > 50 countries
- > 1,000,000 shared bicycles

Canada
Hudson Bay
Sweden
Finland
Russia
Kazakhstan
Mongolia
Uzbekistan
Turkmenistan
Afghanistan
Iran
Iraq
Pakistan
North Atlantic Ocean
Morocco
Algeria
Libya
Egypt
Tunisia
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Yemen
Oman
Burkina Faso
Guinea
Ghana
Nigeria
South Sudan
Ethiopia
Somalia
Kenya
Gulf of Guinea
Gulf of Aden
Arabian Sea
Laccadive Sea
Gabon
DR Congo
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
Indian Ocean
South Atlantic Ocean
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Nicaragua
Caribbean Sea
Guyana
Suriname
Brazil
Chile
Sea of Japan
South China Sea
Philippines
Indonesia
Banda Sea
Arafura Sea
Papua New Guinea
Australia
Great Australian Bight
Google My Maps

# Background

## What's unique about bike-sharing?

| On-demand | Decentralized | Unattended | Concentrated |
| --- | --- | --- | --- |

Compared with other forms of shared-use mobility

1. Unlike conventional public transit (e.g., subways and buses) which follows a regular schedule and pre-determined routes, bike-sharing provides transportation on an **on-demand** basis with a **decentralized** structure.
2. Bike-sharing differs from classic ride-sharing (e.g., carpooling) and ride-sourcing (e.g., Uber and Lyft) in that bicycles are typically **unattended**. Also, during vacant hours, bicycles are **concentrated** at a group of stations.

# Challenges

# Challenges

## Uneven distribution of bikes across stations

- Caused by uncontrolled, uneven usage demand
- Making check in or check out service unavailable at some stations
- Bike redistribution is non-trivial

MATHEMATICS

### *Wheels when you need them*

Can special algorithms help keep shared bicycles rolling?

*By* **Chelsea Wald**, *in Vienna*

Grigorij Kuklin pulls his van over as far as he can while cars stream past.

### Balancing Bike-Share Stations Has Become a Serious Scientific Endeavor

Some top mathematicians and computer scientists are devoting time to the problem.

ERIC JAFFE | @e_jaffe | Aug 27, 2014 | 43 Comments

994 Shares | Share on Facebook | Tweet

# Challenges

## Uneven distribution of bikes across stations

- Caused by uncontrolled, uneven usage demand
- Making check in or check out service unavailable at some stations
- Bike redistribution is non-trivial

| | | | |
|---|---|---|---|
| Bike redistribution strategy design | Bike utilization balancing | Station location selection | Operation hour optimization |

# Challenges

## Network modeling is the key and foundation

To understand how people rent and return bicycles

To understand how bicycles move among stations

## Studies have been conducted

Extensive research on the nature of BSS, business models, how they have spread and adopted

Limited station clustering and coarse-grained rental volume forecasts

## No fine-grained modeling and prediction

# Main contributions

## Spatio-temporal mobility model

To model the bike-sharing system as a dynamic network
To take into account the interactions among all stations

## Traffic (check in/out) prediction mechanism

To jointly consider the spatio-temporal correlations and additional time factors and meteorology
On a per-station basis with sub-hour granularity

## Evaluation with world's largest public BSS

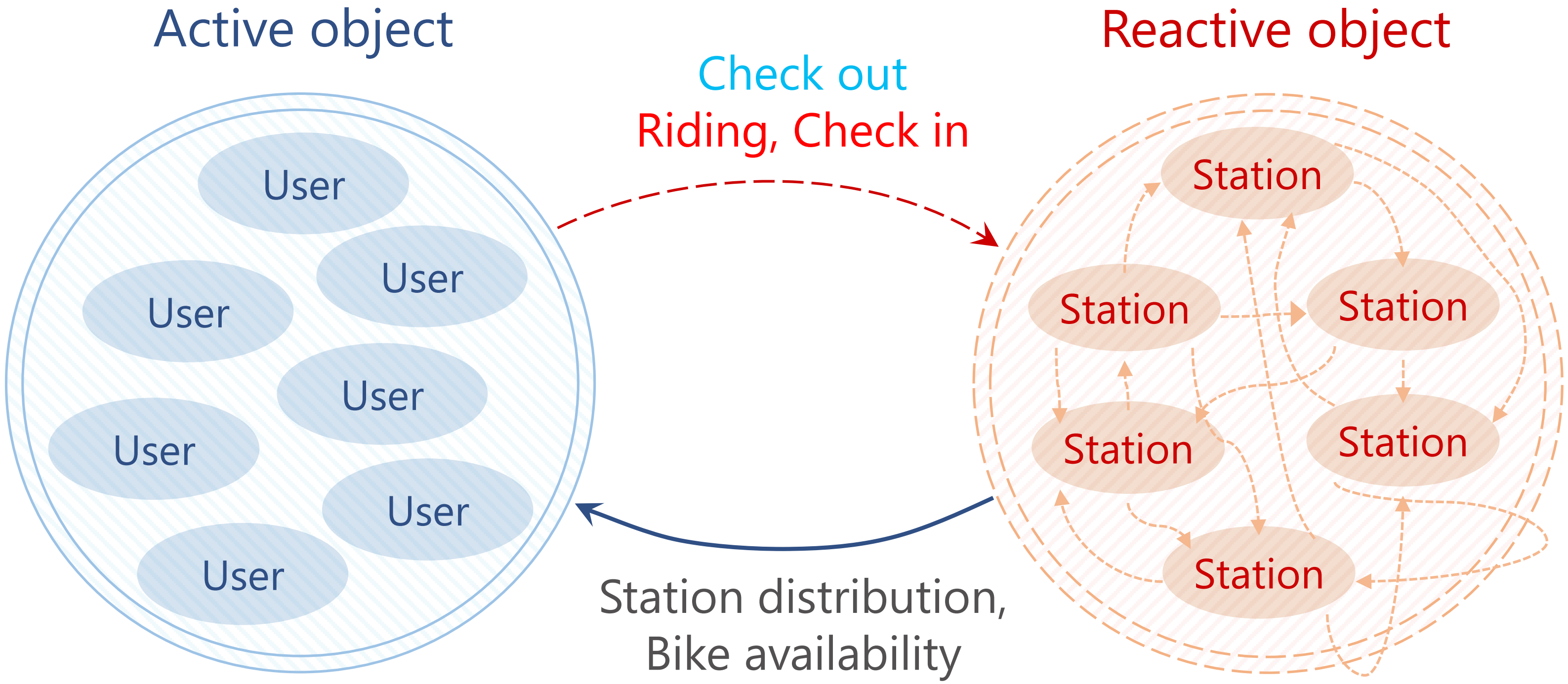
More than 2800 stations and over 103 million check in/out records
Best performance with an 85 percentile relative error of 0.6 for both check in and check out prediction

# Design

# Network modeling and flow prediction

## Problem formulation

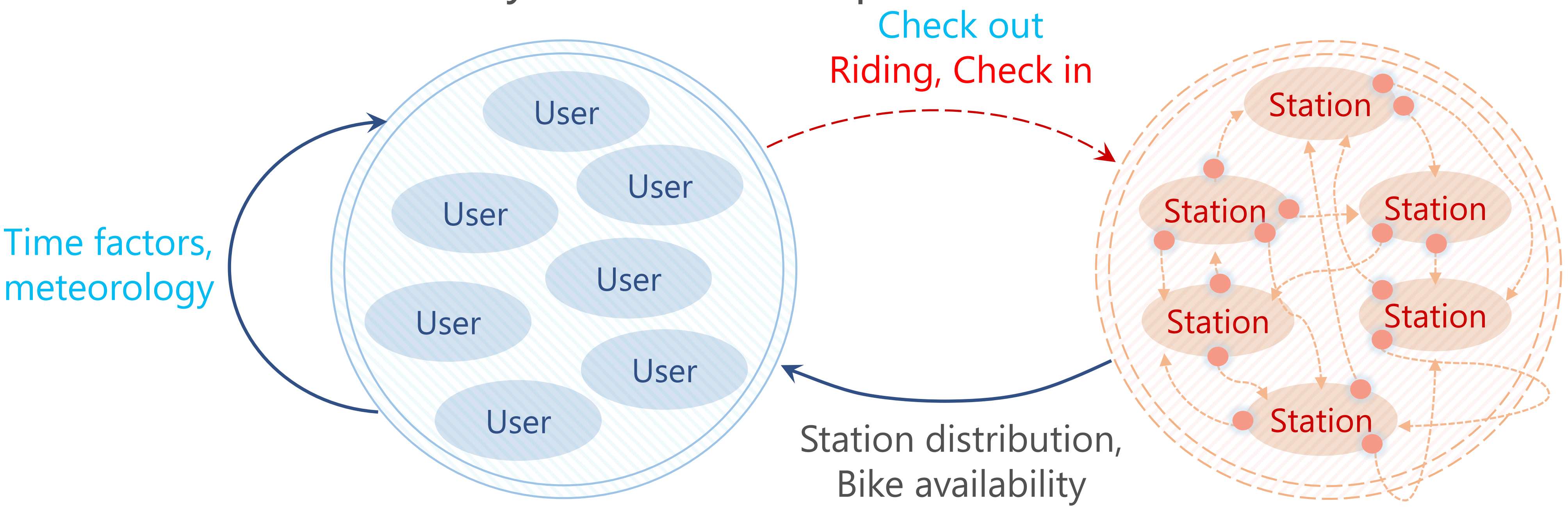
- Active objects (users) and Reactive objects (stations)
  - A shift instance (SI) = check out + movement + check in
- Coupled vs. Mutually independent



# Network modeling and flow prediction

## Design overview

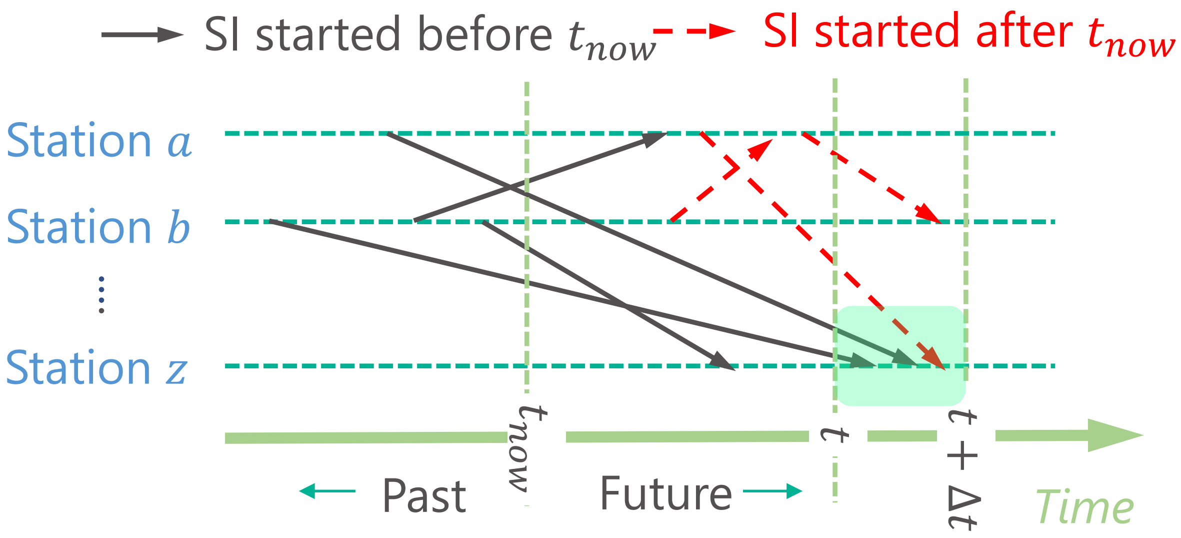
- Modeling the mobility of undocked bicycles
  - Probabilistic model based on historical data to describe the bike movements
- Modeling the check out behaviors
  - Random forest theory to model and predict the check out behaviors



# Network modeling

## Theoretical mobility model

Aim to quantify bikes that will be checked in at station $i$ during target period $[t, t+\Delta t]$ in the future




$$A_i = \sum_{j \in N} D_j \Gamma_{ji} P_t$$

- $A_i$: The number of bikes checking in to station $i$
- $D_j$: The number of bikes checking out from station $j$
- $\Gamma_{ji}$: The transfer probability from station $j$ to station $i$
- $P_t$: The probability that the bike will check in to station $i$ within the target period

# Network modeling

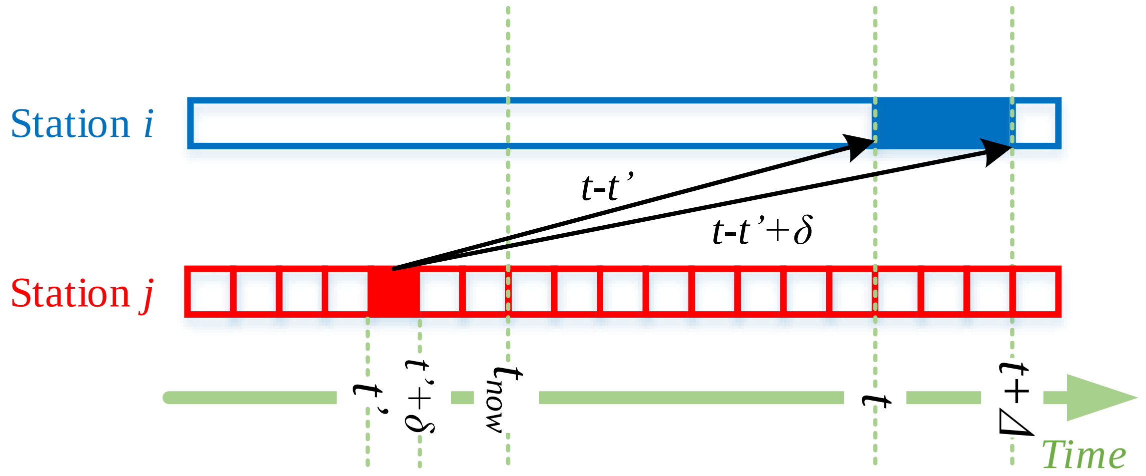
## Theoretical mobility model

Temporal discretization

$$n_{ji} = \sum_{k=1}^{\infty} D_j(t_k, \delta)\Upsilon_{ji}(t_k)\left(F_{ji}(t+\Delta-t_k) - F_{ji}(t-t_k)\right)$$

$\Upsilon_{ji}$ and $F_{ji}$ can be obtained based on historical SI data

Get the expression for $A_i(t, \Delta t) = \sum_{j \in N} n_{ji}$

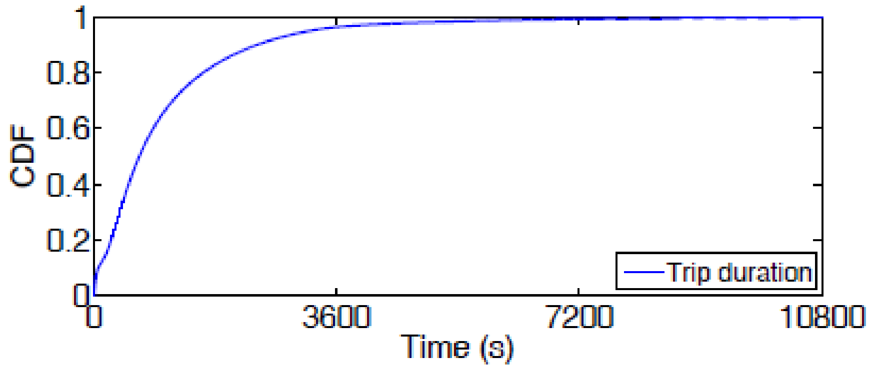


# Network modeling

## Pruning

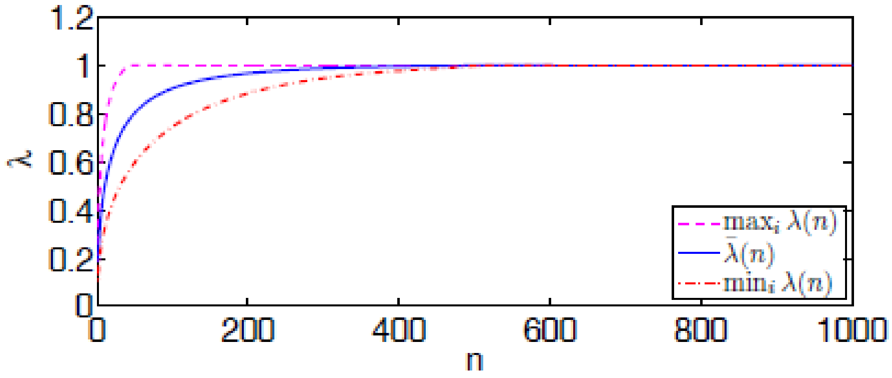
### Temporal pruning

- 99.6% SIs are completed within 3 hours
- $k \in [0, \infty] \Rightarrow k \in [0, 3]$




### Spatial pruning

- Top 200 stations contribute 96.6% of bikes on average
- $N \leq 2800 \Rightarrow N \leq 200$




### $\Upsilon_{ji}(t)$ Discretization and Calculation

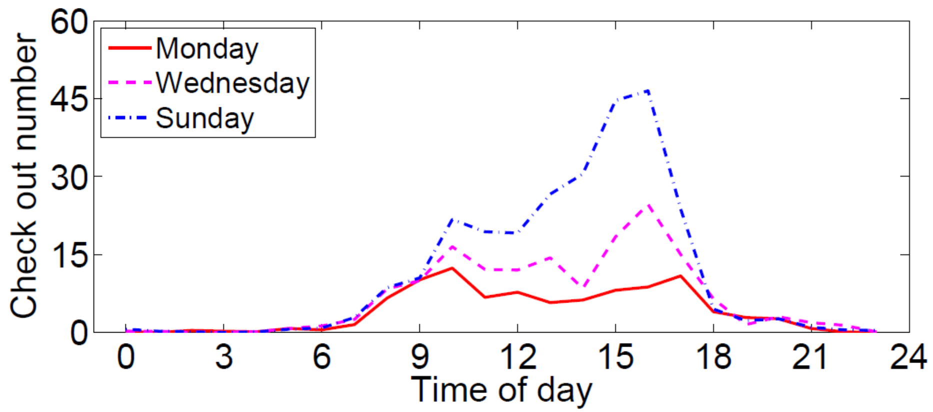
- Discretize $\Upsilon_{ji}(t)$ into a piece-wise function
- Compute its value within each time slot (0.5 hour) based on historical check in/out data

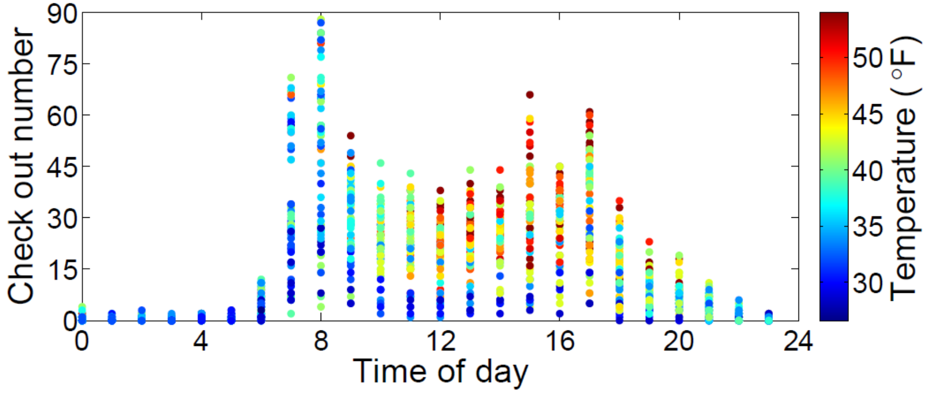
# Bicycle check out prediction

## Feature extraction

### Offline features

- Time factors (day of week, time of day, weekday, holiday)
- Meteorology (temperature, humidity, visibility, wind speed)



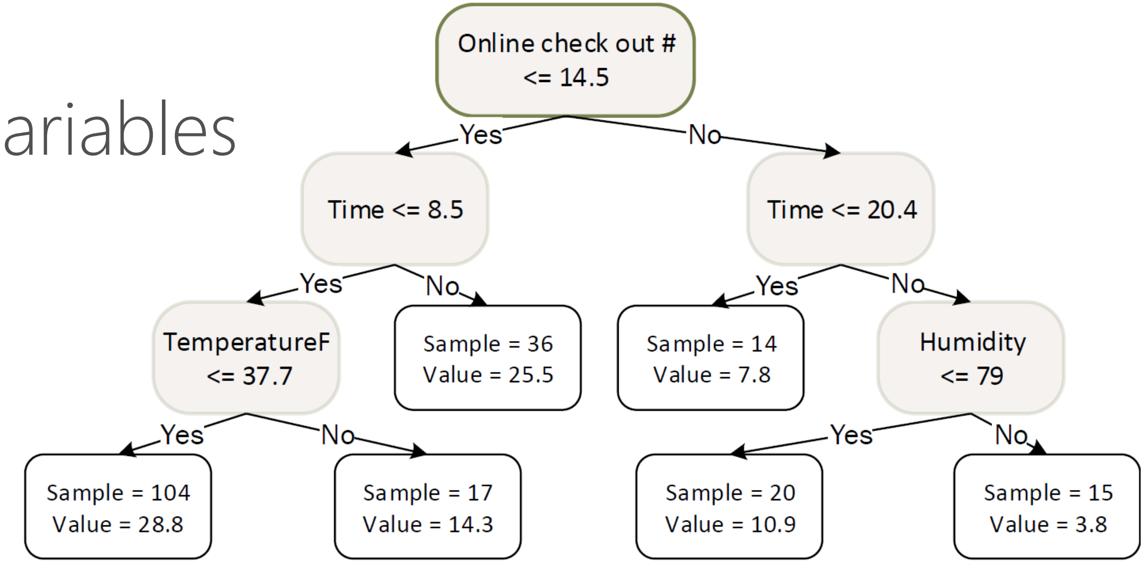

### Online features

- Online check out number from the previous time window

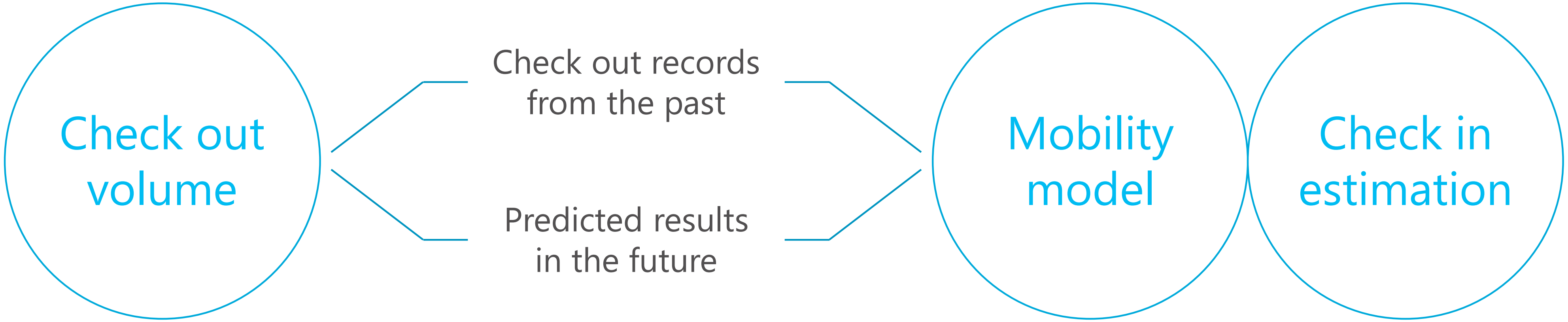
# Bicycle check out prediction

## Random forest model

- Deals with both categorical and numerical variables
- Provides importance of features
- Can be easily parallelized




## Put it all together



# Evaluation

# Dataset description

## The BSS dataset

World's largest public BSS in Hangzhou, China

Over 3300 stations and 84,000 shared bikes, 103,661,080 records

## The meteorology dataset

Weather conditions of Hangzhou with 17,520 (i.e., 24*2*365) records

| user_id | rent_netid | tran_date | tran_time |
|---|---|---|---|
| 6114381 | 4051 | 20130101 | 000152 |
| **return_netid** | **return_date** | **return_time** | **bike_id** |
| 4015 | 20130101 | 001547 | 913672 |

| Time (CST) | Temp (°F) | Dew Point (°F) |
|---|---|---|
| 12:30 PM | 100.4 | 69.8 |
| **Pressure (hPa)** | **Humidity (%)** | **Visibility (MPH)** |
| 29.65 | 37 | 6.2 |
| **Wind Dir** | **Wind Speed (MPH)** | **Conditions** |
| WSW | 8.9 | Partly Cloudy |

# Evaluation

## Baseline approaches

- Historical Average (HA) [Gast et, al. CIKM'15]
- Auto-Regressive and Moving Average (ARMA) [Vogel et, al. CL'11]
- HP-MSI/P-TD [Li et, al. SIGSPATIAL'15]

## Evaluation methodology

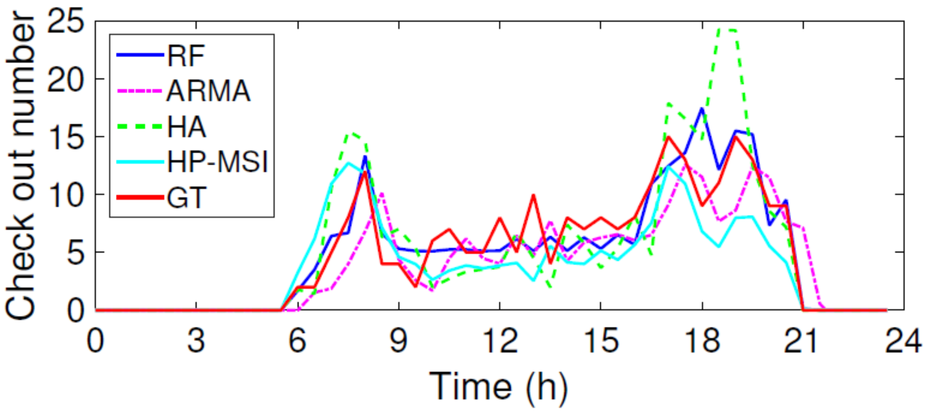
- Check out prediction
- Check in estimation

# Evaluation

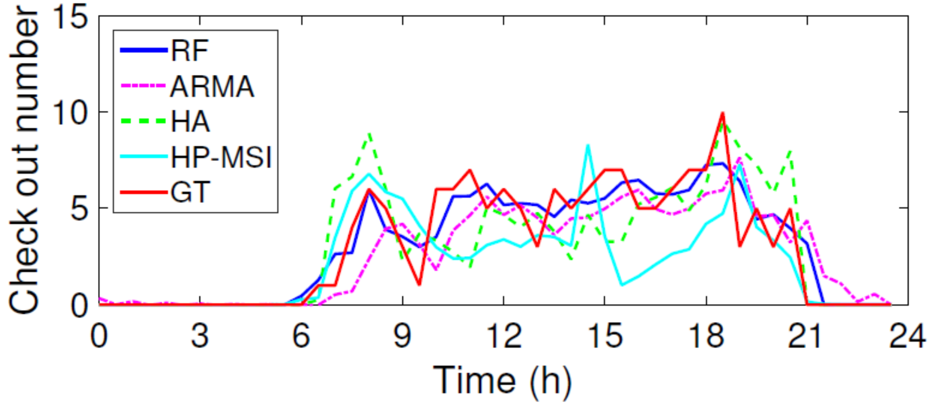
## Check out prediction

- Case study
  - Check out number over 24 hours
  - Summer > Winter
  - Different feature importance

| Day of week | Hour | Temperature | Humidity |
|---|---|---|---|
| 0.0288 | 0.1434 | 0.0846 | 0.0514 |
| Visibility | Wind speed | Holiday | Workday |
| 0.0332 | 0.0211 | 0.0030 | 0.0064 |
| Online check out number | | | |
| 0.6282 | | | |




(a) A rainy summer weekday.



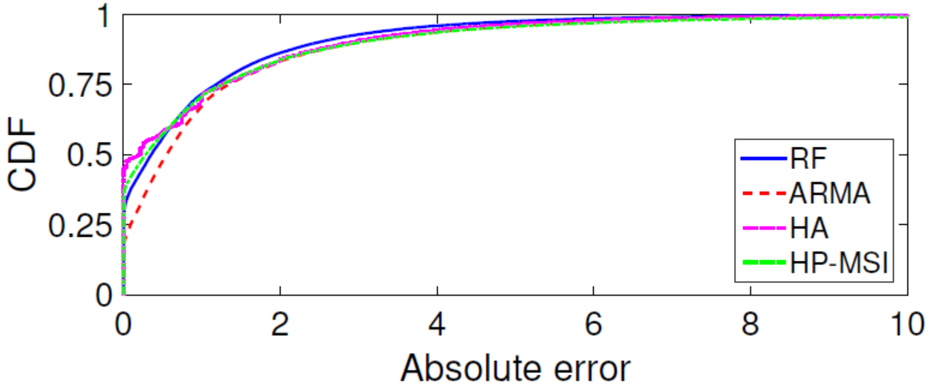

(b) A sunny winter weekend.

Check out prediction at station 3648

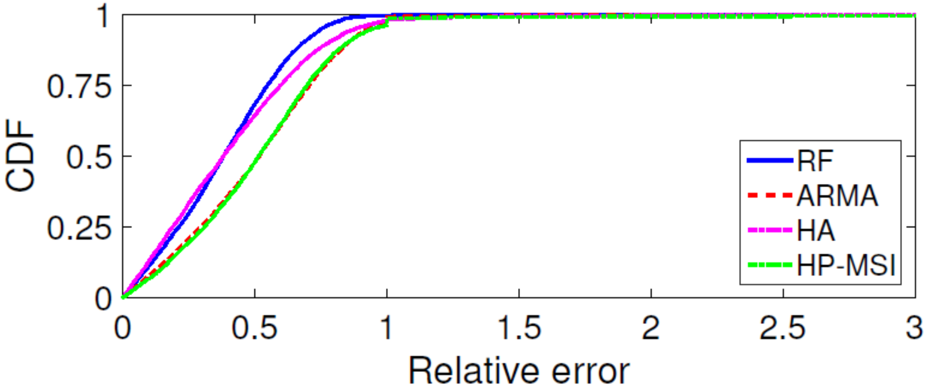
# Evaluation

## Check out prediction

- Overall performance
  - First 20 days of each month to train, and predict the numbers in remaining days
  - Absolute error: numbers of bikes
  - Relative error: dividing the absolute error by the ground truth.




(a) CDF of absolute error.



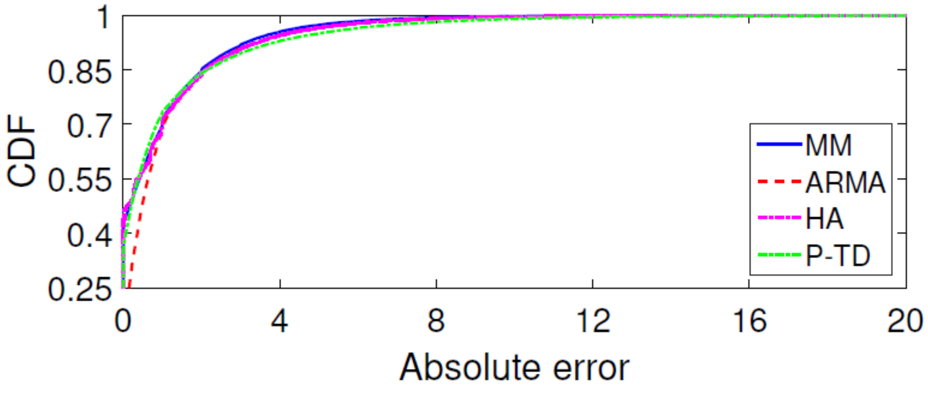

(b) CDF of relative error.

Overall performance

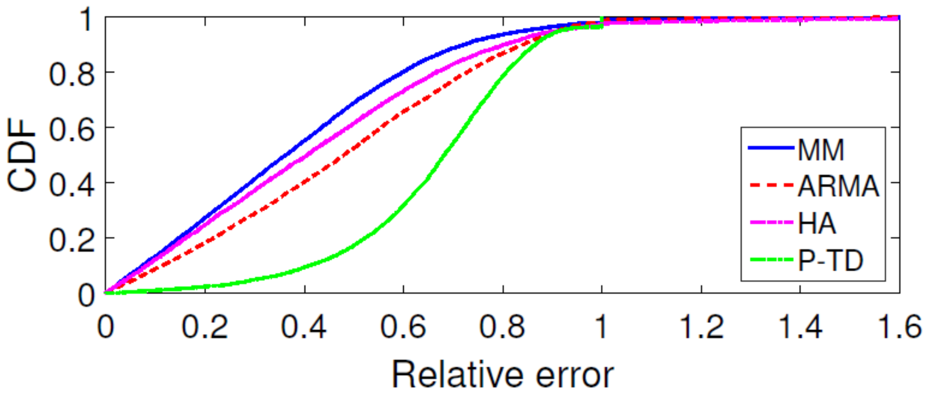
# Evaluation

## Check in estimation

- Overall performance
  - First 20 days of each month to train, and predict the numbers in remaining days
  - $\Delta$ = 30 (i.e., each approach is required to estimate check in number in the following 30 minutes)
  - Best relative error from MM



(a) CDF of absolute error.



(b) CDF of relative error.

Overall performance

# Insights

## Variation Among Different Scenarios

Better prediction from workdays and stations in business area

## User-centric Modeling and Prediction

Identify regular users and exploit their profiles

*n* routes of which the check out/in stations fall into two small circles

| $n$ | regular user(%) | $n$ | regular user(%) |
|---|---|---|---|
| 6 | 12.55% | 14 | 3.85% |
| 8 | 9.19% | 16 | 2.87% |
| 10 | 6.65% | 18 | 2.02% |
| 12 | 5.02% | 20 | 1.23% |

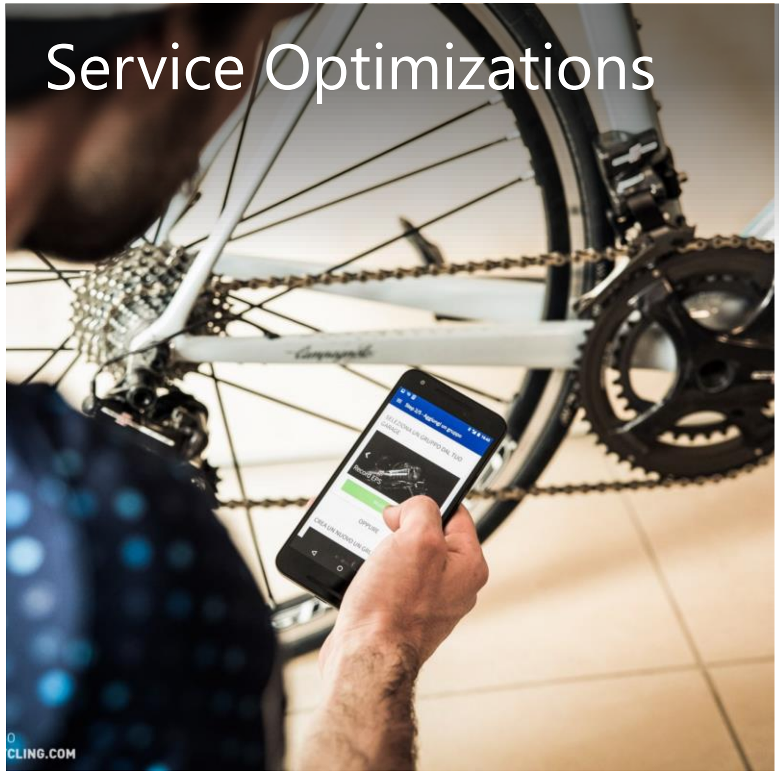
Regular user percentage

# Open research issues

## From a mobile system point of view

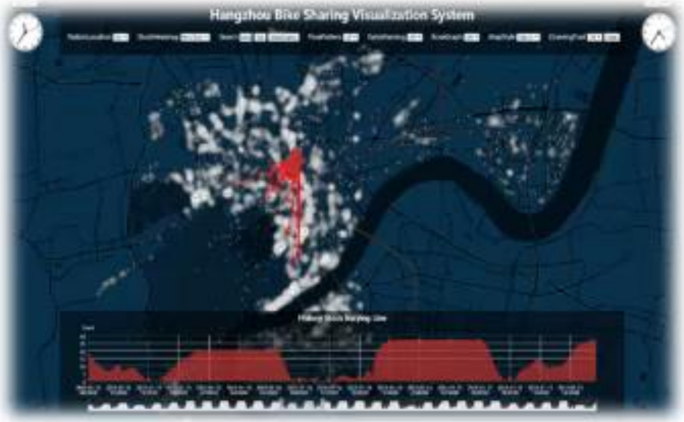

Exploit the inherent diversities from multi-source data (i.e., taxi, bus, subway)


Design an efficient and practical rebalancing algorithm




Station location optimization, service hour optimization, pricing strategy design etc.

# Demo: Data Analysis and Visualization in Bike-Sharing System




浙江大學
1897
ZHEJIANG UNIVERSITY

Microsoft
Research
微软亚洲研究院

# Thank you

yuanchao.shu@microsoft.com