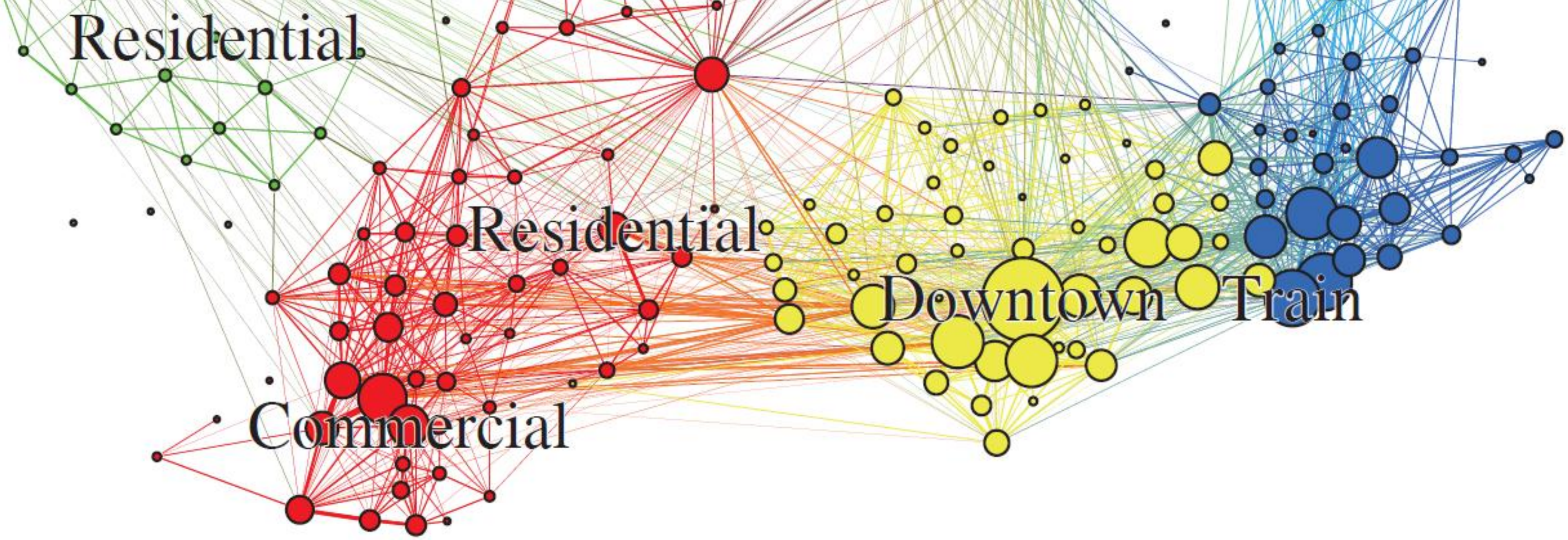


Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales

Desheng Zhang & Tian He — Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu

University of Minnesota, USA — Shenzhen Institute of Advanced Technology, China

ACM MobiCom 2014, Maui, HI

Outline

- **Introduction**
- Design
- Evaluation
- Application
- Conclusion

Human Mobility Patterns

- Mobile Networking
- Location Based Services
    - Real-time Navigation
    - Transit Services
    - Social Networking

Mobile Networking

Location Based Services

Real-Time Navigation

Transit Services

Social Networking

Human Location Tracking Devices

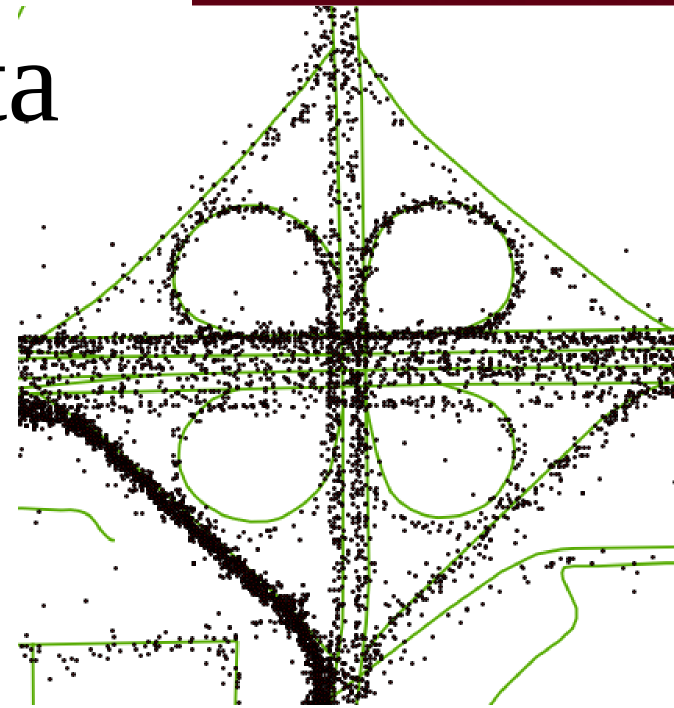
- **GPS Devices**
- **Cellphones** by Call Detail Records (CDR)
  - Cell Tower Levels
- **Automatic Fare Collection System** (AFC)
  - Station Levels: Subways, Buses, Taxicabs
- **Massive Empirical** Data Collection

Subway Station

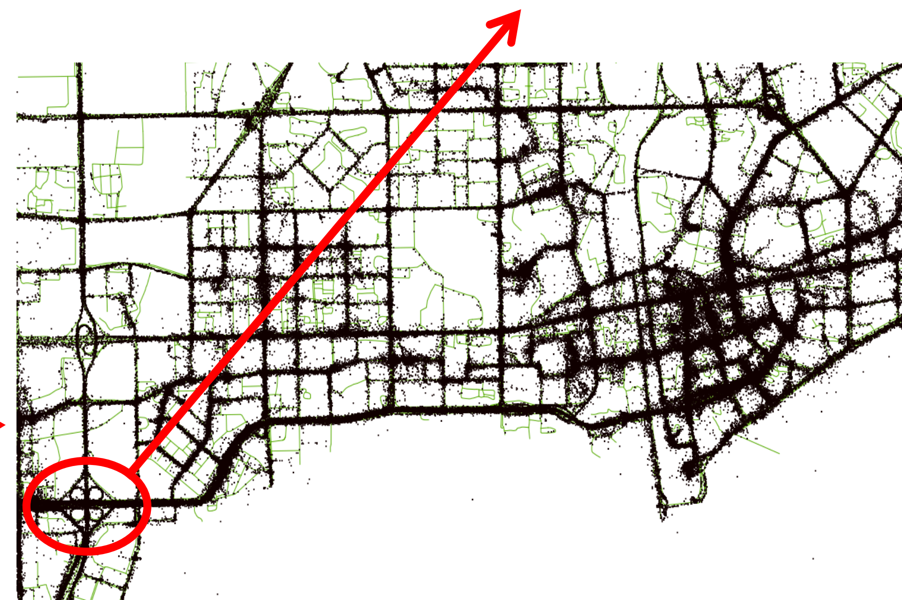
Bus

Taxi

Empirical Mobility Data



- Empirical Data for Mobility Modeling
  - Large Scale
  - Fine Granularity
  - Long Collection Period
- Taxicab Passengers in Shenzhen

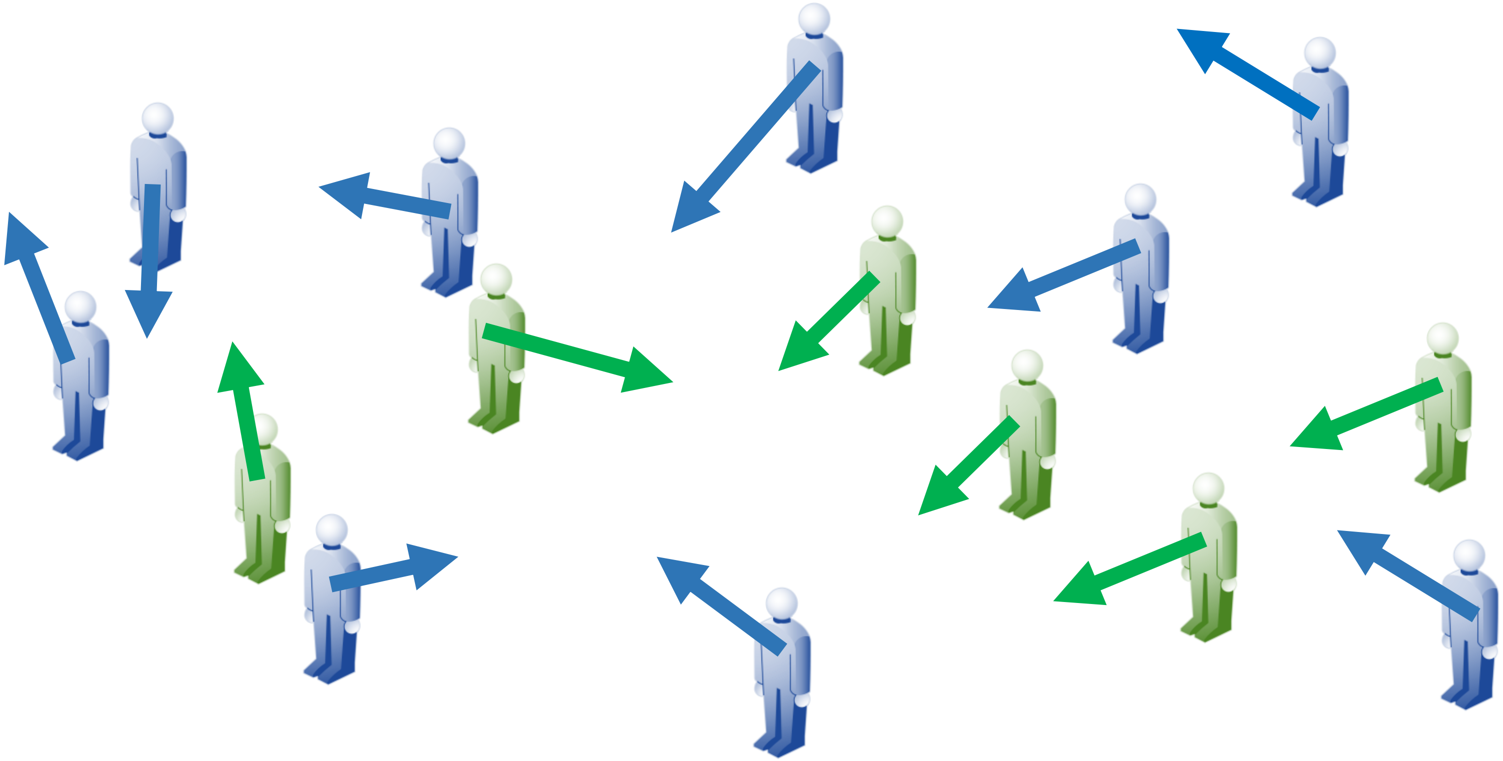


Human Mobility Models

- Legacy Mobility Models:
  - **MobiCom'03: Obstacles** based Mobility Model by Jardosh *et al.*
  - **MobiCom'04: Weighted Waypoint** Model by Hsu *et al.*
  - **MobiCom'07:** Mobility Modeling in **Bus-based DTN**: Zhang *et al.*
  - **UbiComp'11:** Mobility Modeling with **Smartcards**: Lathia *et al.*
  - **KDD'11:** Mobility in **Social Networks**: Cho *et al.*
  - **MobiSys'12: Cellphone** based Mobility Model: Isaacman *et al.*
  - **MobiCom'13: Residence Time** Prediction: Baumann *et al.*
  - **MobiCom'13: Ballistic** Model: Bogo *et al.*
- Models based on **Single-Source** Data
  - **Cellphone**
  - **One** Kind of Urban Transit
    - **Taxicab, Subway** or **Bus**

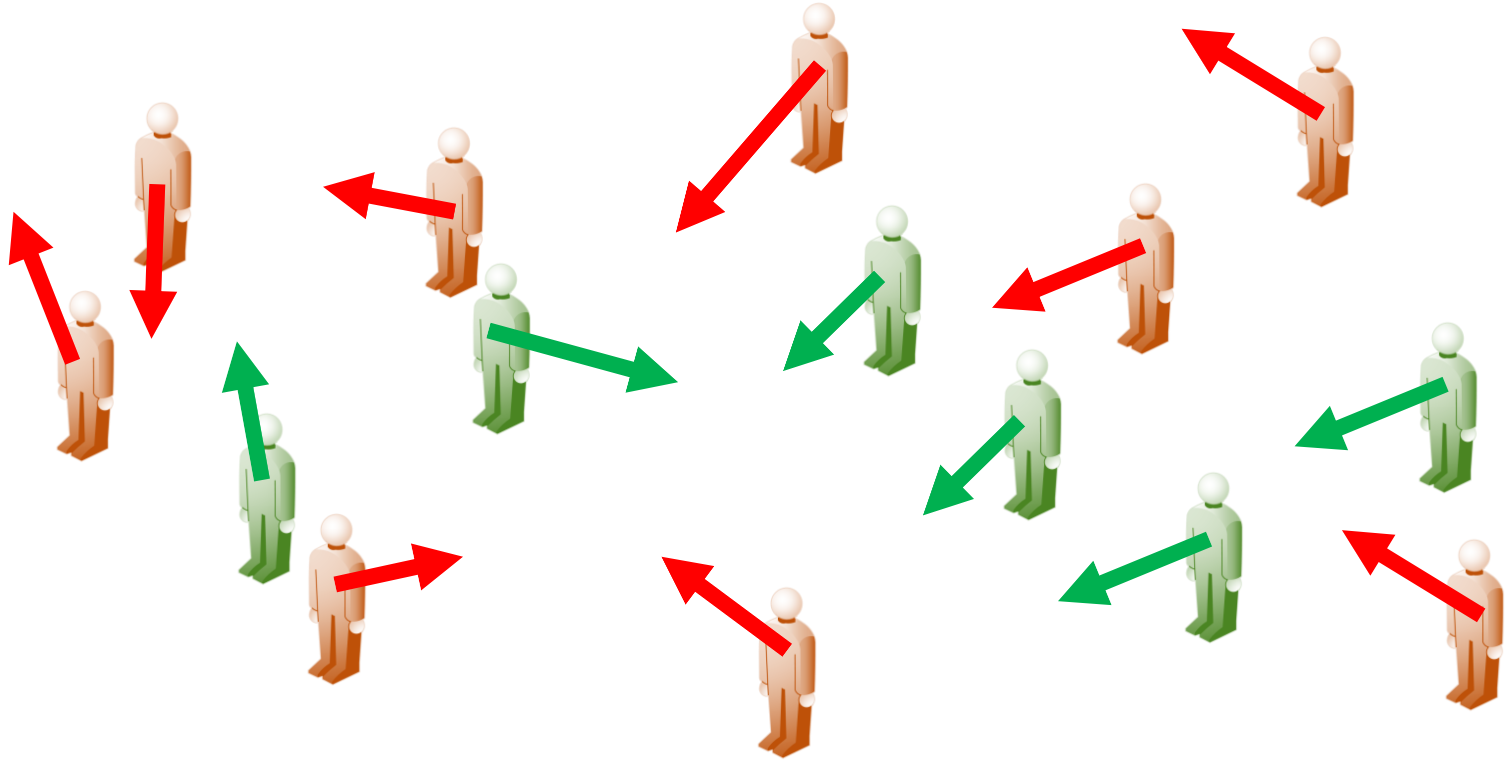
Common Drawback: **Biased Sampling**

- Using **Residents** in Single-Source Data as a Sample for ALL



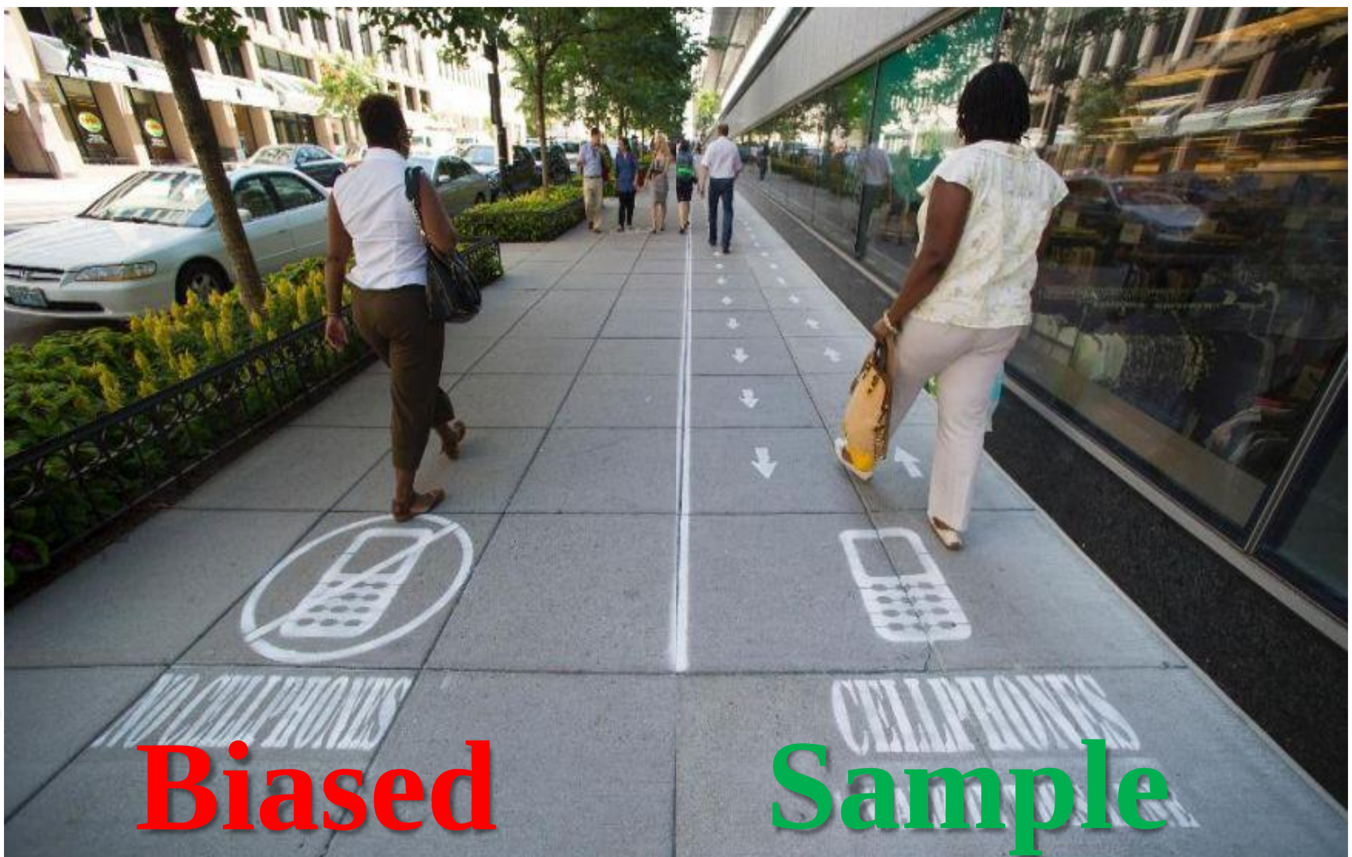
Common Drawback: **Biased Sampling**

- Using **Residents** in Single-Source Data as a Sample for ALL
- Introducing a **Bias** against **Residents not Involved**



Models Based on **Cellphone Data**

- Use **Residents** with cellphone activities as a **Sample** for all
- **Biased** against **Residents** without cellphone activities



Models Based on **Transit Data**

- Use a type of Passengers (e.g., taxicab) as a sample
- Biased against Residents using other transit

Taxi Passengers

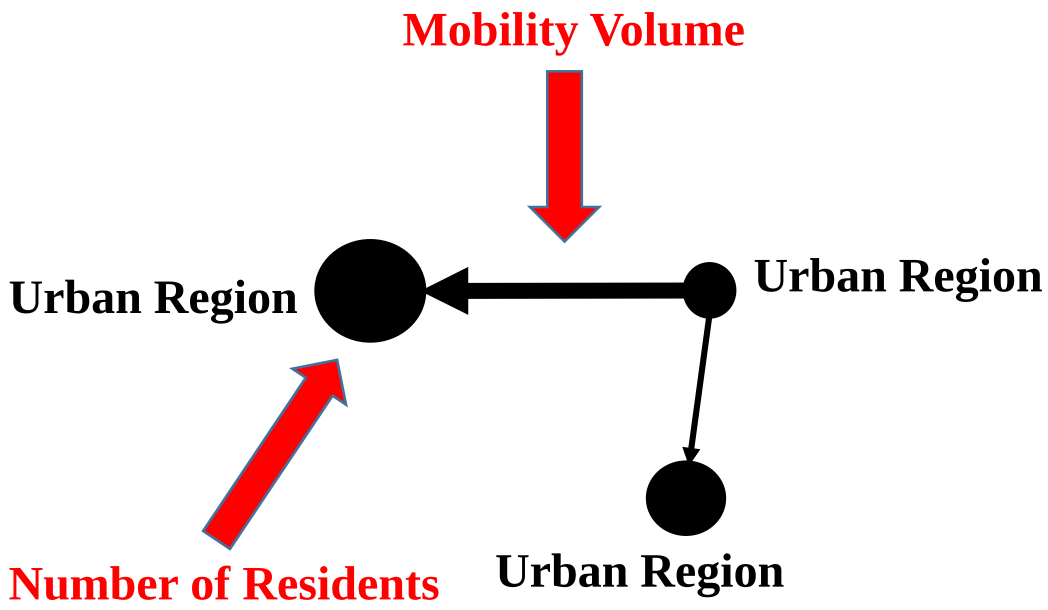
Bus Passengers

Subway Passengers

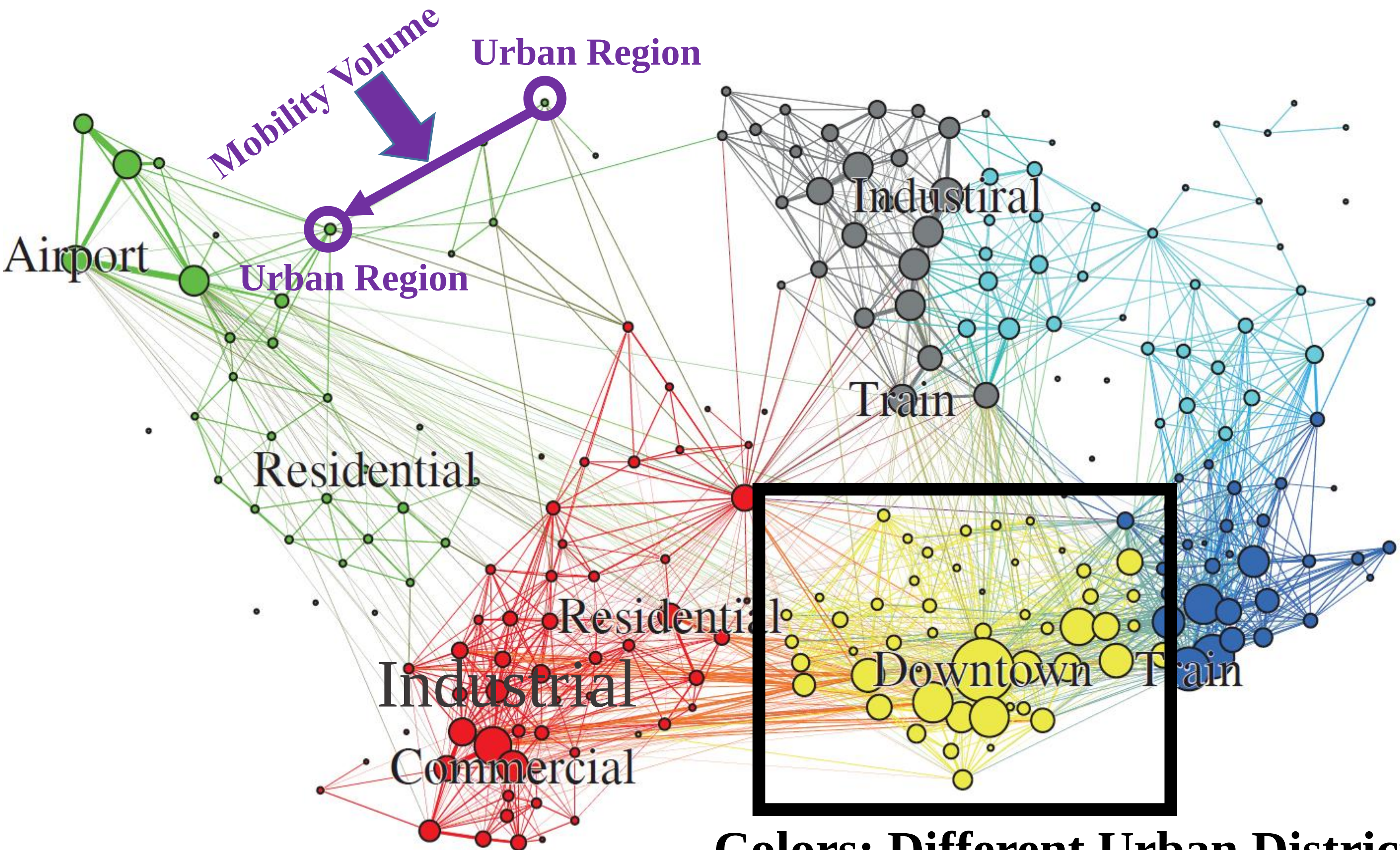
Private Cars

Visualizing Biased Sampling with Mobility Graph

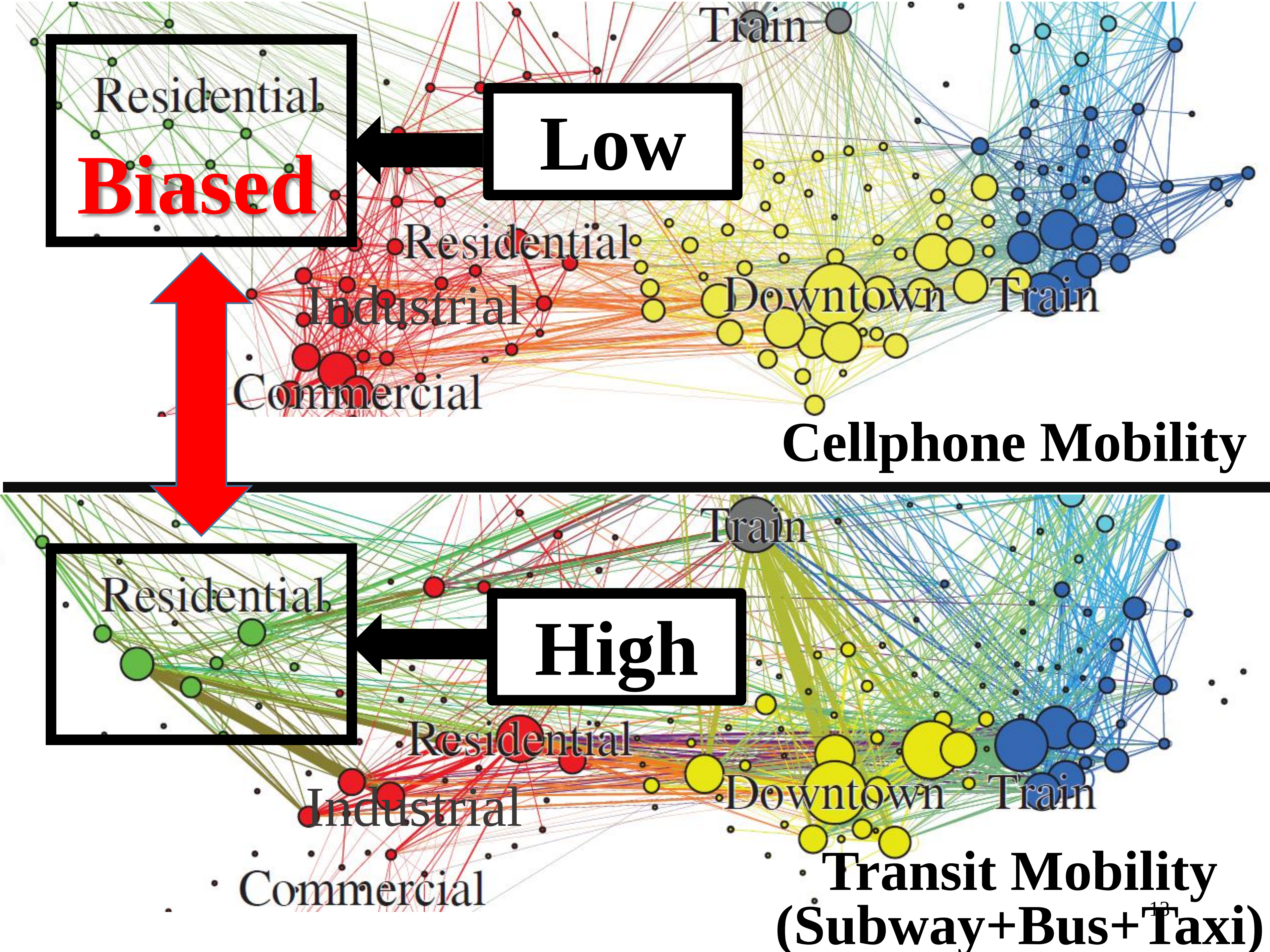
- **Vertex**: a Urban Region
- **Vertex Size**: Number of Mobile Residents
- **Edge**: Mobility between a Pair of Regions
- **Edge thickness**: Mobility Volume



Mobility Graph based on Cellphone Data

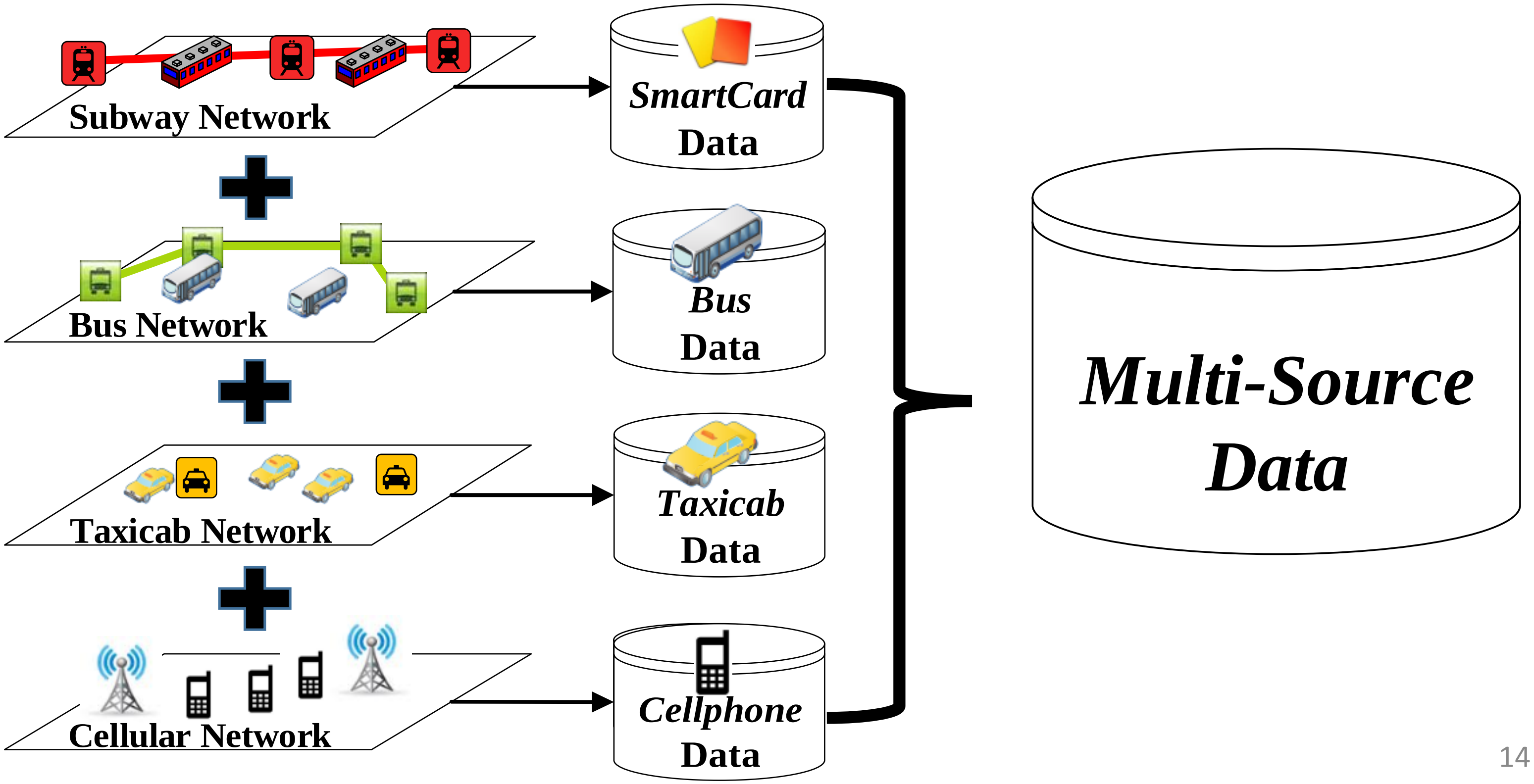


Colors: Different Urban Districts


Train
Residential
Biased
Low
Residential
Industrial
Downtown
Train
Commercial
Cellphone Mobility
Train
Residential
High
Residential
Industrial
Downtown
Train
Commercial
Transit Mobility
(Subway+Bus+Taxi)

Possible Solution: **Multi-Source Data**

- Quick Expansion of Urban Infrastructures
  - Enabling **Multi-Source** Data to address biased sampling
  - Integrating **Transit Networks** with **Cellular networks**



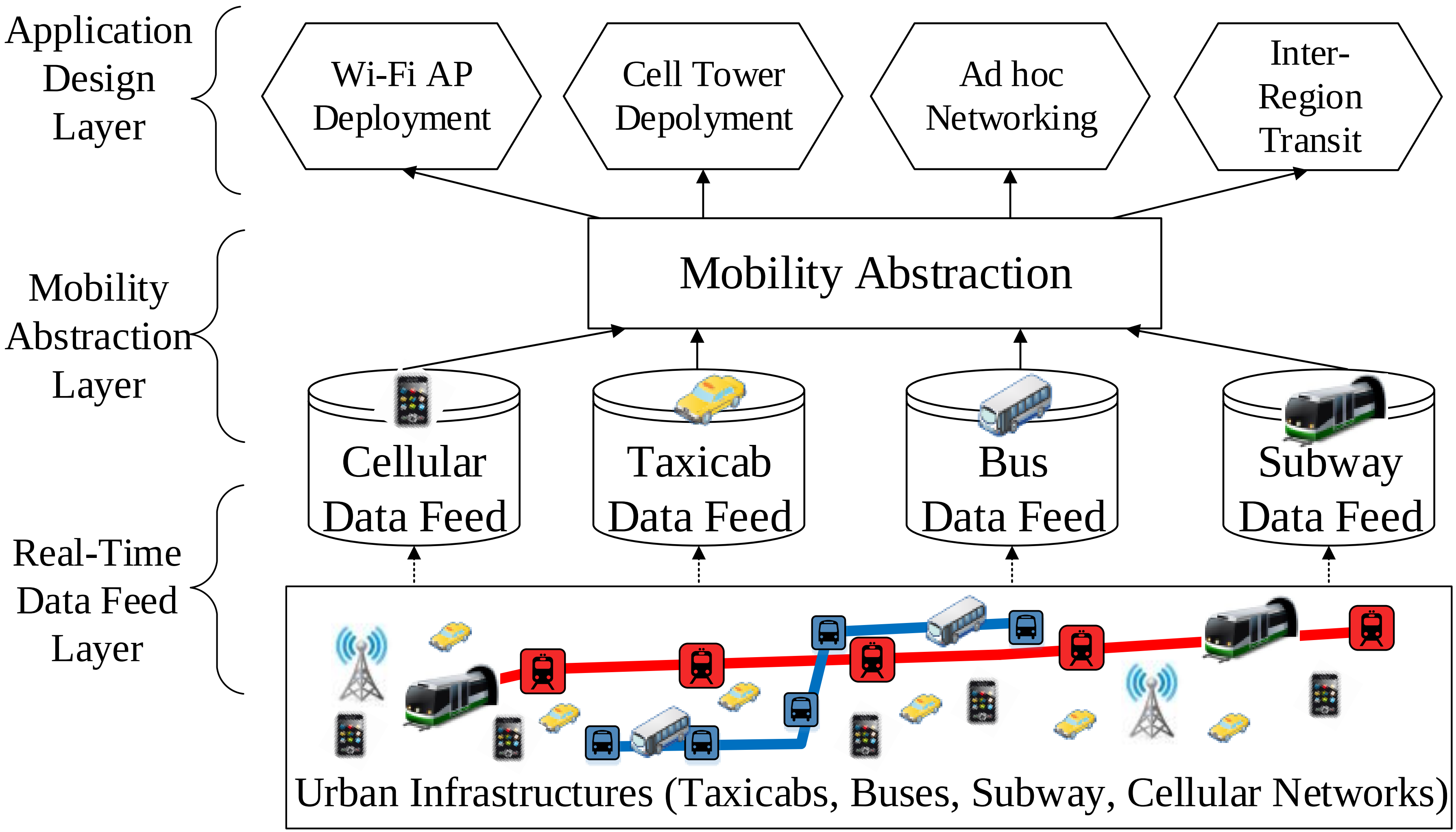
Contributions

- **Mitigating biased sampling** in single-source data by cross-referencing multi-source data
- Analyzing **spatial-temporal dynamics** of multi-source data to infer real-time mobility
- Designing the first **generic architecture** mPat for mobility modeling, separating low level data collection and high level service design
- Implementing **mPat** with extremely large-scale multi-source data **capturing 10 million residents** in Shenzhen
- Enabling an **inter-region mobility inference** with a 75% accuracy
- Developing a transit service based on inferred **inter-region mobility** to reduce 46% of passenger travel time

Outline

- Introduction
- **Design**
- Evaluation
- Application
- Conclusion

mPat Architecture

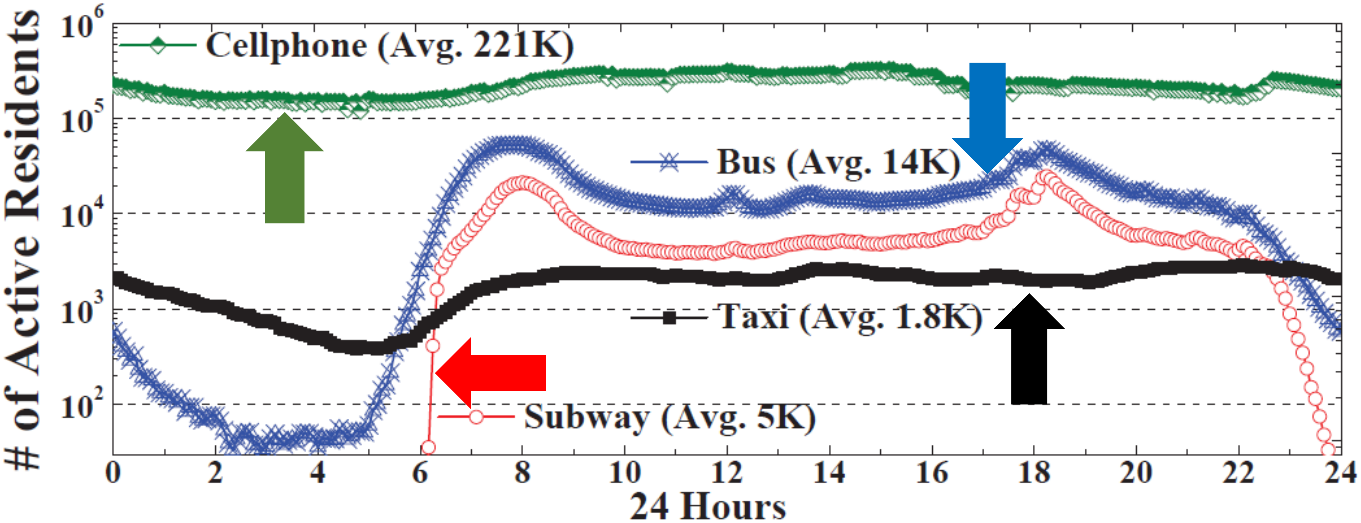


Data Feed Layer: Overview

- Data Feeding
- Data Managing
- Data Storing
- Data Cleaning
- Data Protecting

Data Feeding

- Close Collaboration
  - Shenzhen **Government Agencies**
- A Reliable Feeding Mechanism
  - **Cellphones**: CDRs for **10.4 Million** Users
  - **Smart Cards**: Fare Transactions for **16 Million** Users
  - **Taxicabs**: GPS for **14 Thousand** Taxicabs
  - **Buses**: GPS for **10 Thousand** Buses



Data Managing

- Hardware:
  - 11 Node Cluster with **34 TB** Storage
  - Node with **32 Cores** and **32 GB RAM**
- Software:
  - Hadoop Distributed File System (**HDFS**)
  - Pig and Hive

Cluster in Shenzhen

Data Storing

Cellphone Dataset	
Collection Period	10/01/13-Now
Number of Users	10,432,246
Data Size	680 GB
Record Number	434,546,754
Format	
SIM ID	Date and Time
Cell Tower ID	Activities

Taxicab GPS Dataset	
Collection Period	01/01/12-Now
Number of Taxis	14,453
Data Size	1.7 TB
Record Number	22,439,795,235
Format	
Plate Mumber	Date and Time
Status	GPS Coordinates

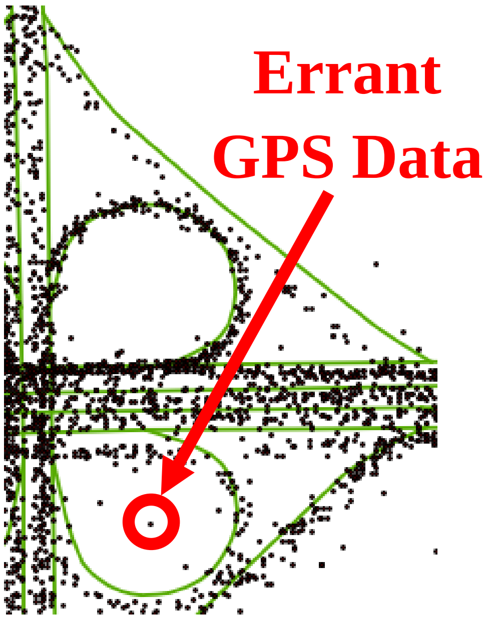
Bus GPS Dataset	
Collection Period	01/01/13-Now
Number of Vehicles	10,000
Data Size	720 GB
Record Number	9,195,565,798
Format	
Plate Number	Date and Time
Velocity	GPS Coordinates

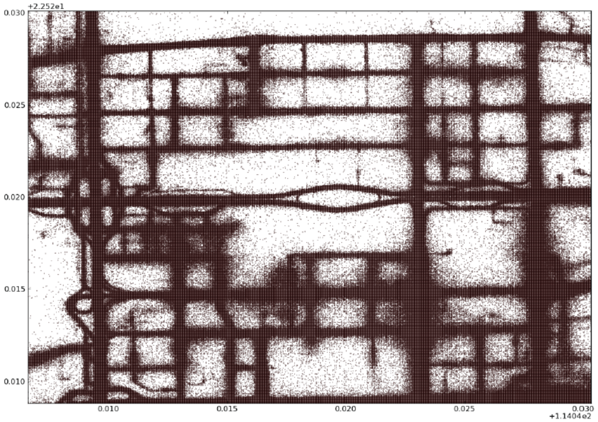
Smart Card for Subway & Bus	
Collection Period	07/01/11-Now
Number of Cards	16,000,000
Data Size	600 GB
Record Number	6,212,660,742
Format	
Card ID	Date and Time
Device ID	Station Name

University of Minnesota
Driven to Discover℠

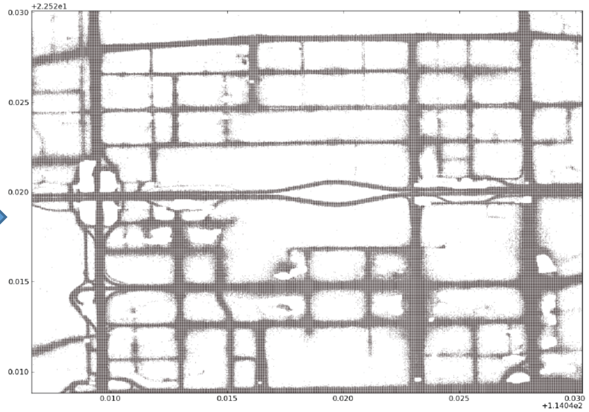
Data Cleaning

- Errant Data in mPat
  - Duplicated Data
  - Data with Logical Errors
  - Missing Data
- 11% of Data Removed

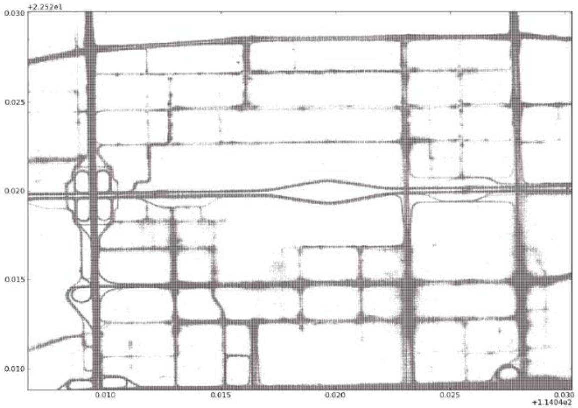




Raw GPS



Removing Errant and Duplicated GPS



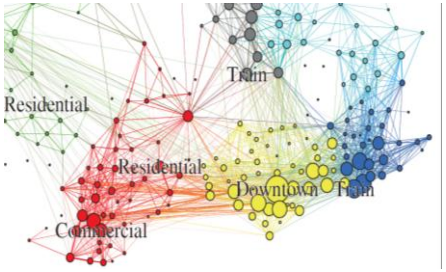
Map Matching

Data Protecting: Privacy

- **Anonymization**:
  - Anonymizing All Data
  - Replacing IDs with Serial Numbers

- **Minimal Exposure**:
  - Processing Mobility Info Only
  - Dropping Other Info

- **Aggregation**:
  - Presenting Mobility in Aggregation
  - Not Focusing on Individual Users



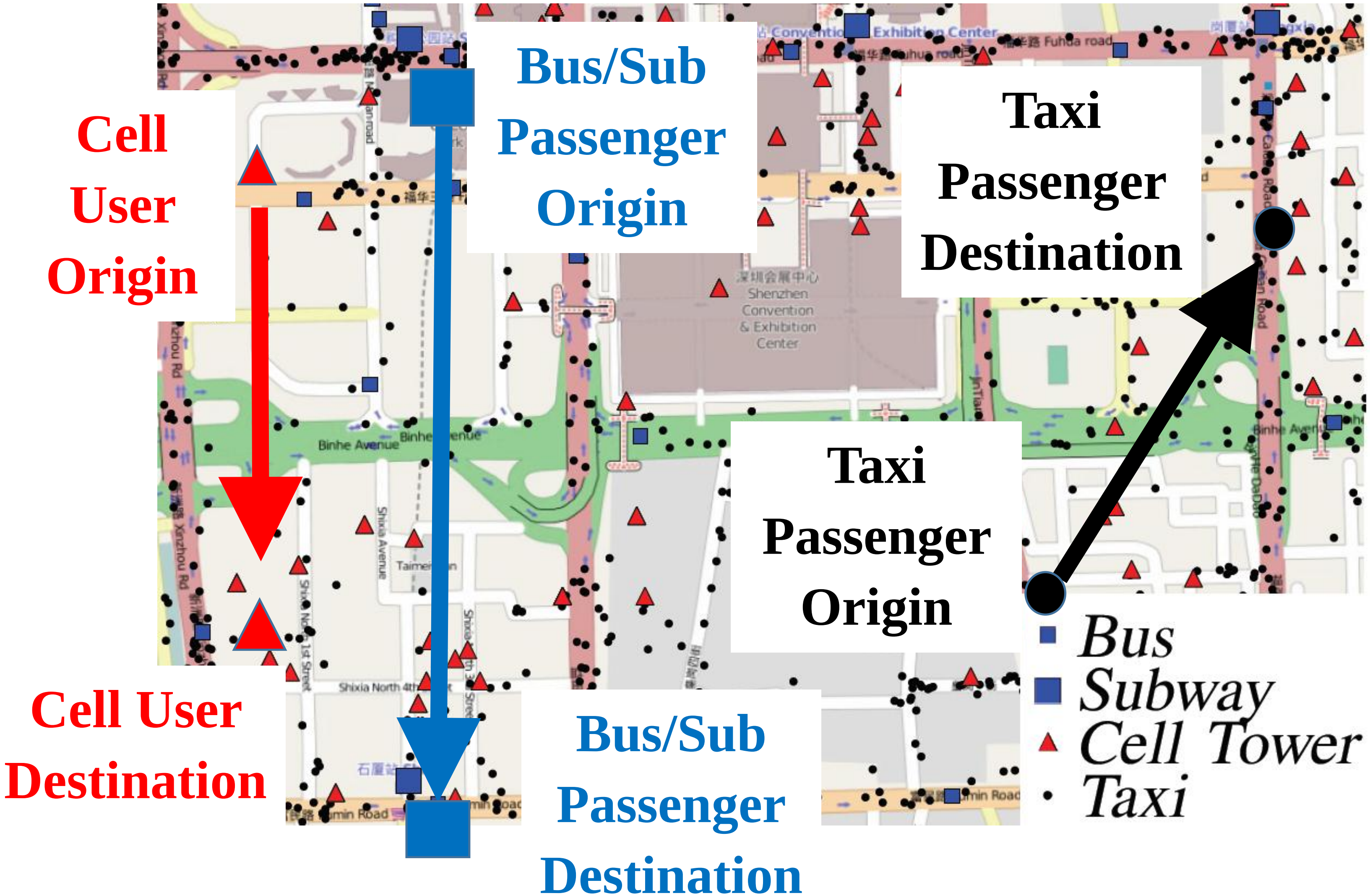
Mobility Abstraction Layer: **Overview**

- Trip Extraction
- Spatial and Temporal Characteristic Analysis
- Urban Region Partition
- Inter-Region Mobility Inference

Trip Extraction

- Cellphone User Trips
  - Obtaining trips by **a continuous trace** of cellphone towers associated CDRs for the same user
- Taxicab Passenger Trips
  - By finding **pickup and related dropoff** locations
- Bus Passenger Trips
  - By finding **boarding and alighting** bus stations
- Subway Passenger Trips
  - By finding **entering and exiting** metro stations
- Details in the paper

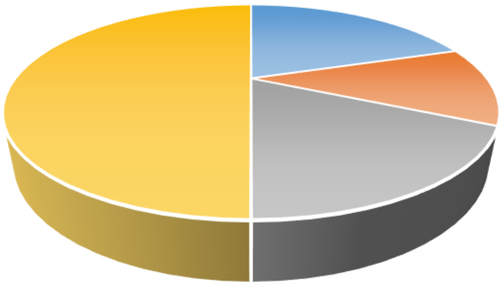
Trip Extraction



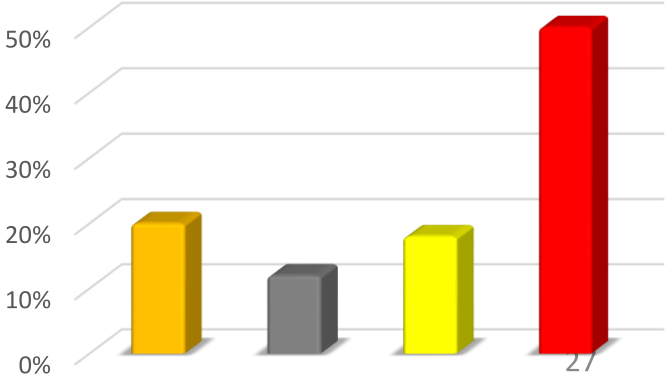
Characteristic Analysis

- Classifying All Trips
  - **Cellphone Trips**
  - **Transit Trips**

- **Spatial** Characteristic
  - Variety in **Lengths**

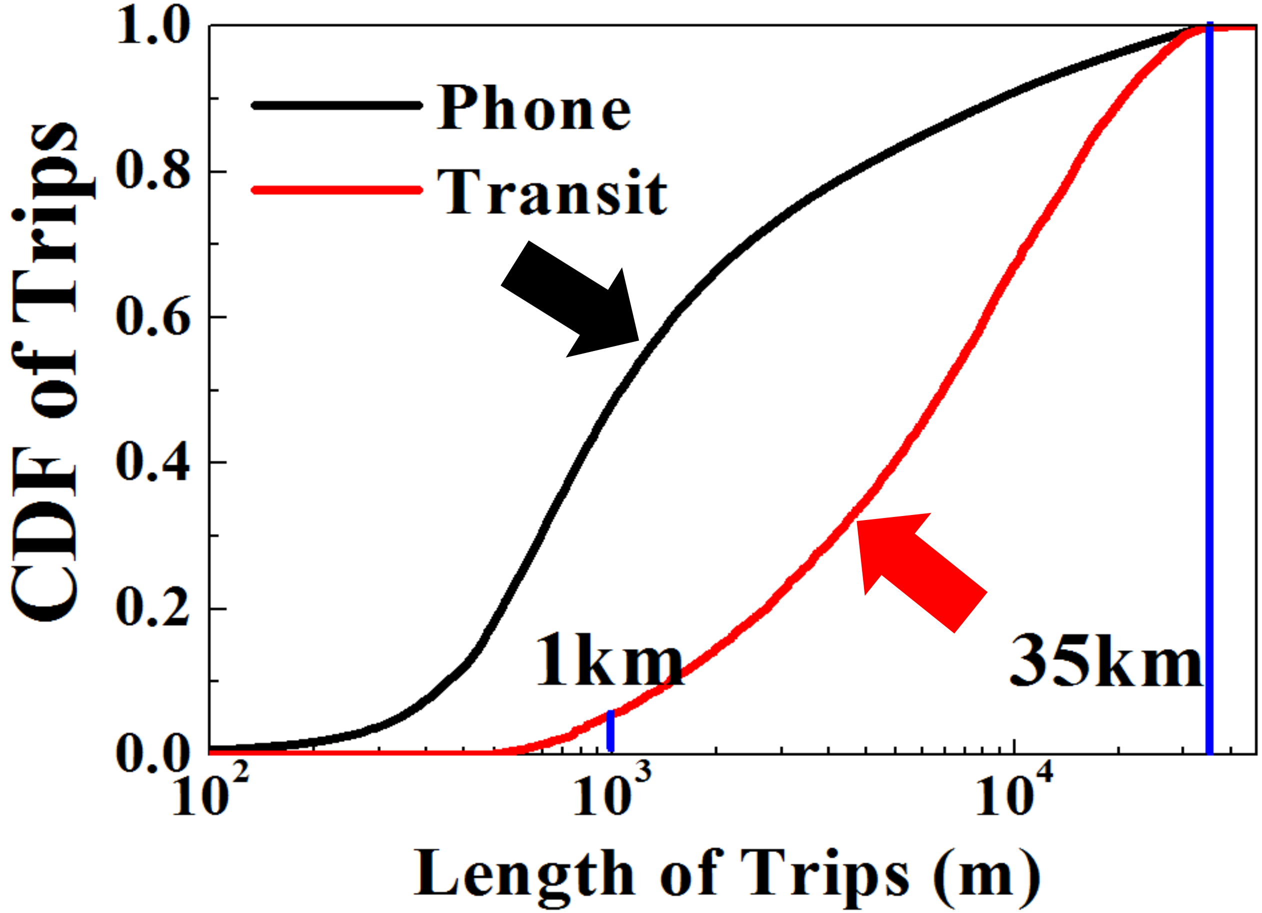


- **Temporal** Characteristic
  - Variety in **Time Periods**



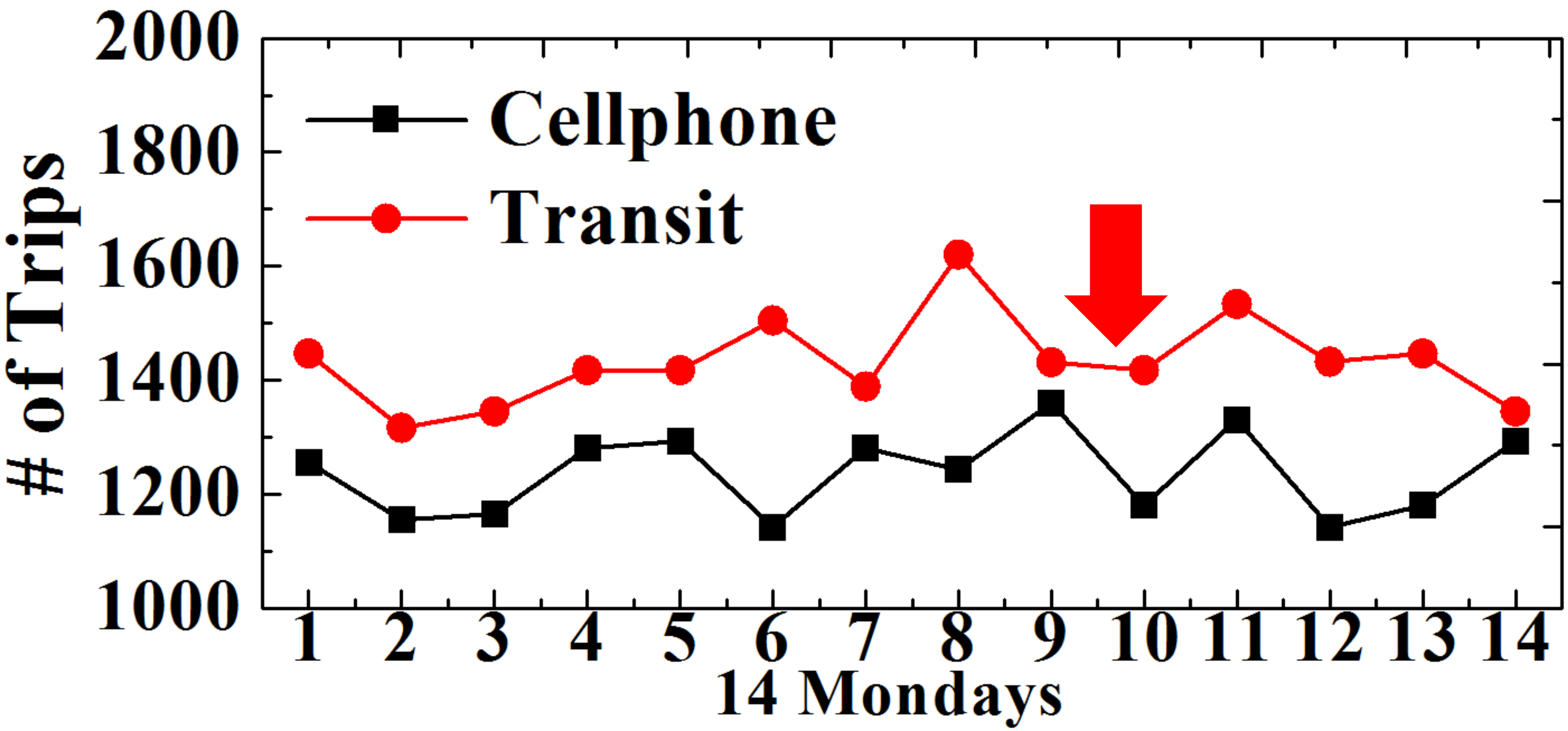
Spatial Characteristic

- Trips from **Transit Data**
  - Trips between 1 km and 35 km
- Trips from **Cellphone data**
  - Trips with various lengths



Temporal Characteristic

- Captured Trips in the slot 7-8 AM in **14 different Mondays**
  - **Fewer Trips** from Cellphone data
  - **More Trips** from Transit data

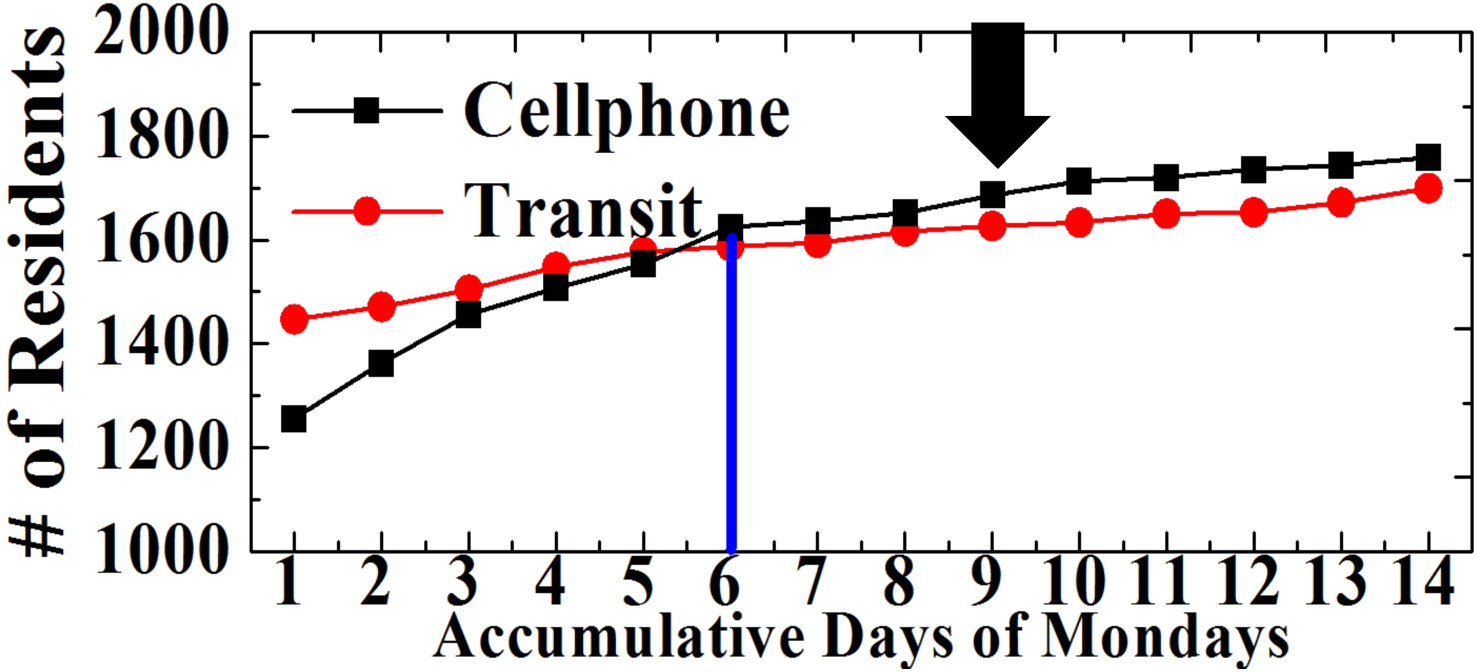


Empirical Insights

- Bias in **Transit** Data**:**
  - Capturing fewer **short (<1km) or long (>35km)** trips
  - **Difficult** to be mitigated
- Bias in **Cellphone** Data**:**
  - Capturing fewer trips in **a given time slot**
  - **Possible** to be mitigated
- **Mitigating** the Bias in Cellphone Data**:**
  - Urban trips are highly **repeatable**, e.g., daily commute
  - A traveling resident may use **cellphone before**
  - **Accumulatively** using historical data to capture residents

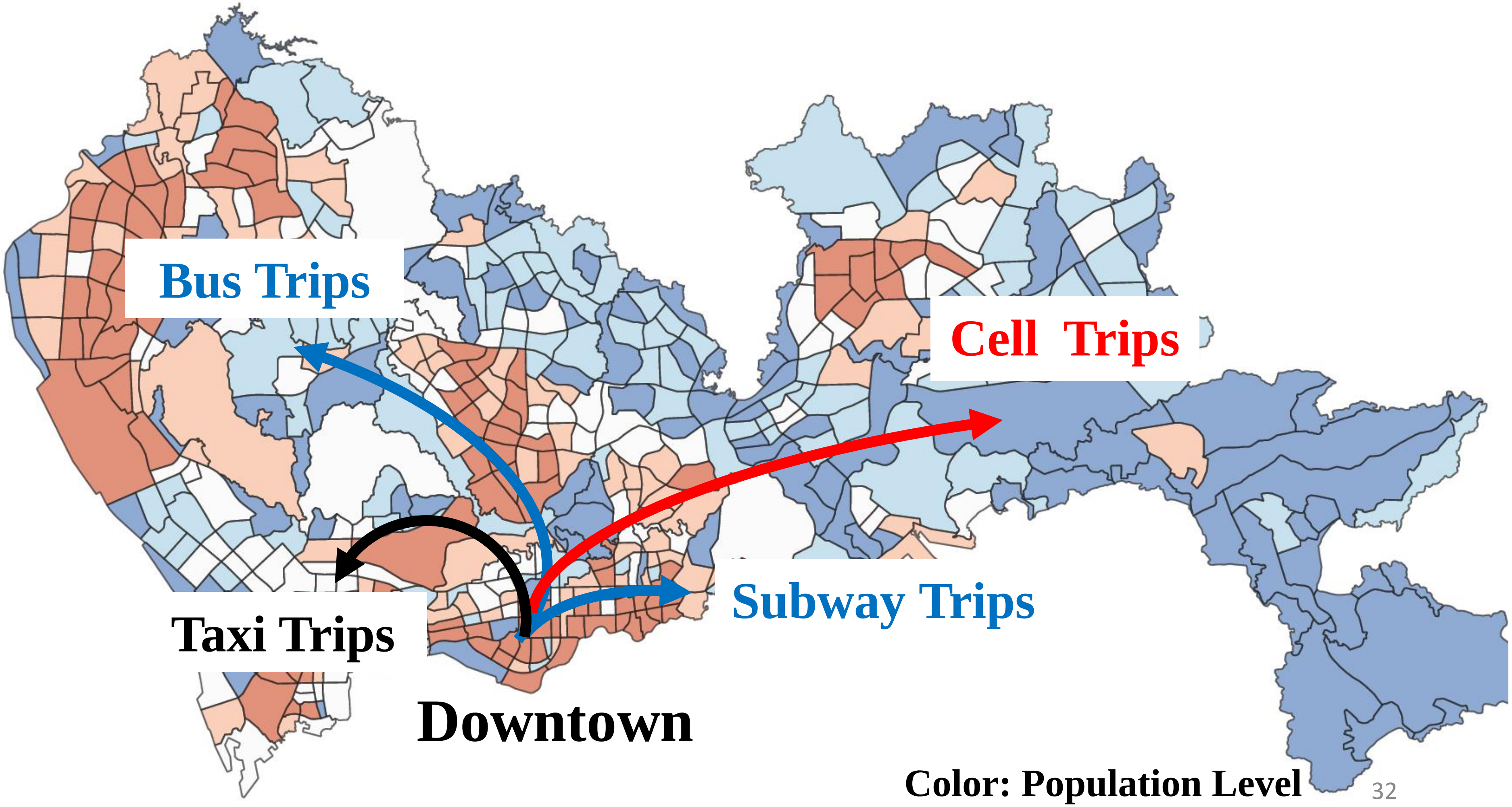
Cumulatively using Historical Data

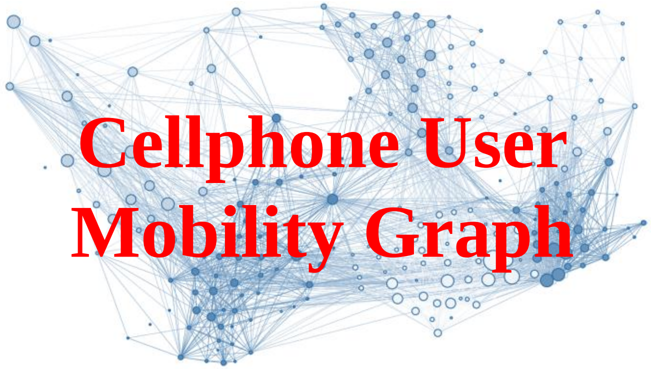
- **Captured Trips** from Unique Residents in Accumulative Mondays
  - **Cellphone Data** are better

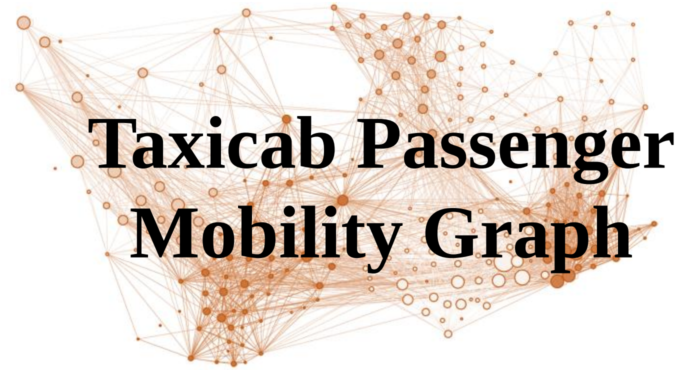


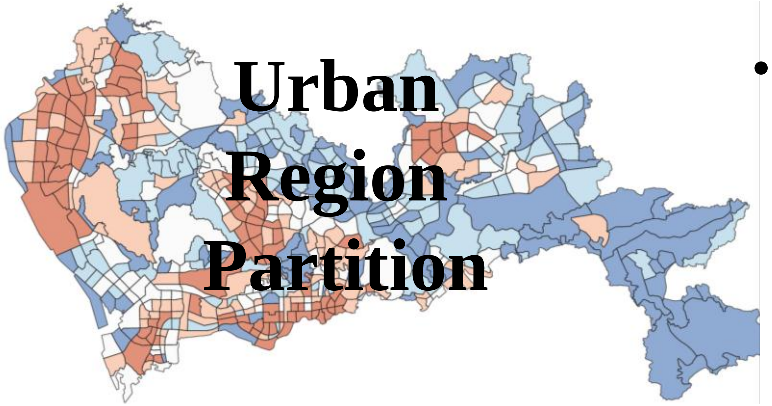
Urban Region Partition

- Utilizing **496 Shenzhen Urban Regions** as a spatial partition

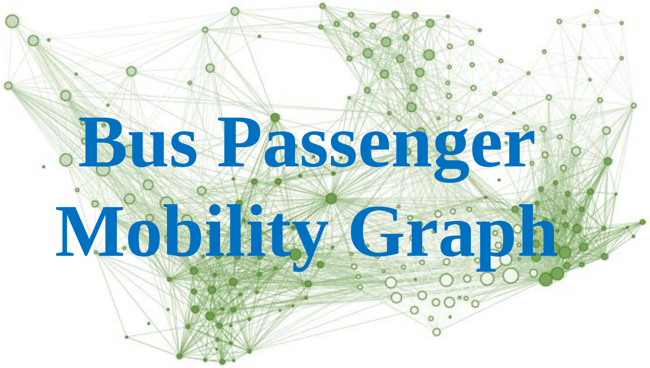


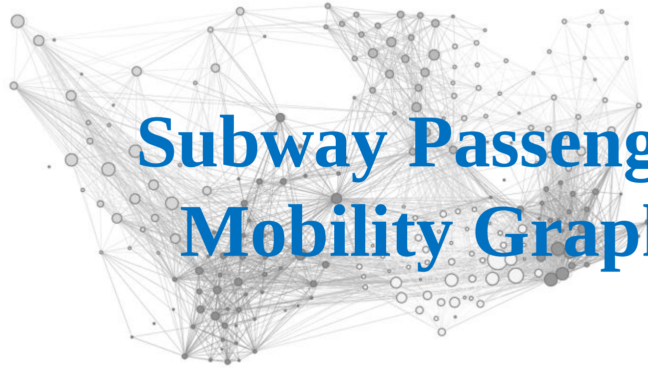






- **Mobility Graph**
  - **Vertex**: a Urban Region
  - **Vertex Size**: Number of Mobile Residents
  - **Edge**: Mobility between a Pair of Regions
  - **Edge thickness**: Mobility Volume

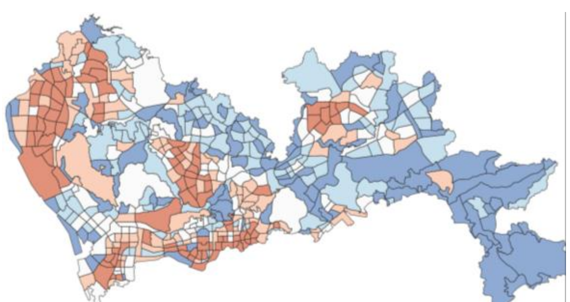




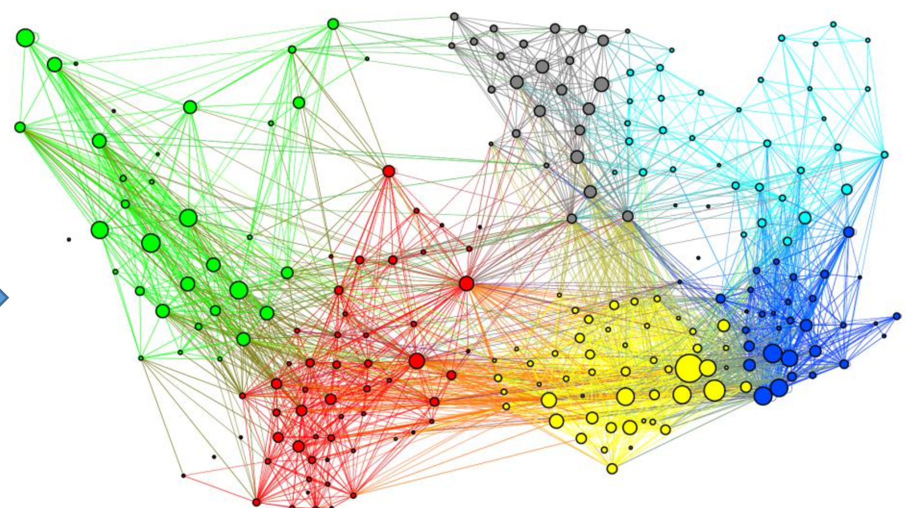
Online Inference:

- Objective: inferring the real-time mobility among different **urban regions** by a **mobility graph G** for the current slot
- Aggregating **individual mobility** to obtain **mobility volumes** for every **region pairs**

Trips from Data



Urban Region Partition

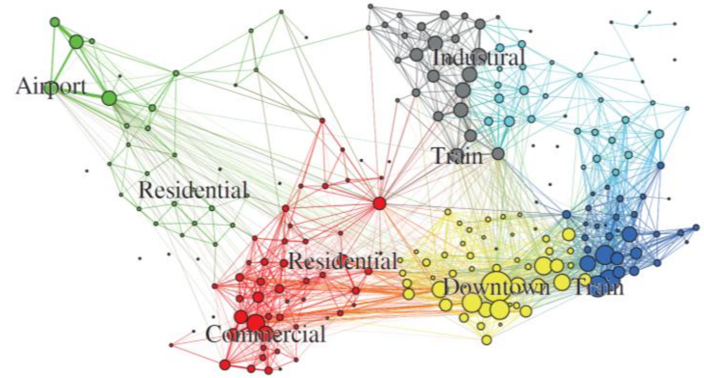


A mobility graph **G**

Online Inference by Cellphone Data:

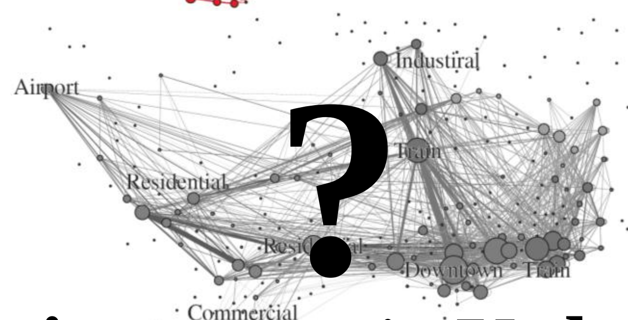
- 90% of urban residents have cellphones
- Infer $\boldsymbol{G = G^{c} + G^{\underline{c}}}$ in a slot $\boldsymbol{\tau}$
  - $\boldsymbol{G^{c}}$ for **active** cellphone users **with activities** in $\boldsymbol{\tau}$
  - $\boldsymbol{G^{\underline{c}}}$ for **inactive** cellphone users **without activities** in $\boldsymbol{\tau}$

Active Users



$\boldsymbol{G^{c}}$

Inactive Users

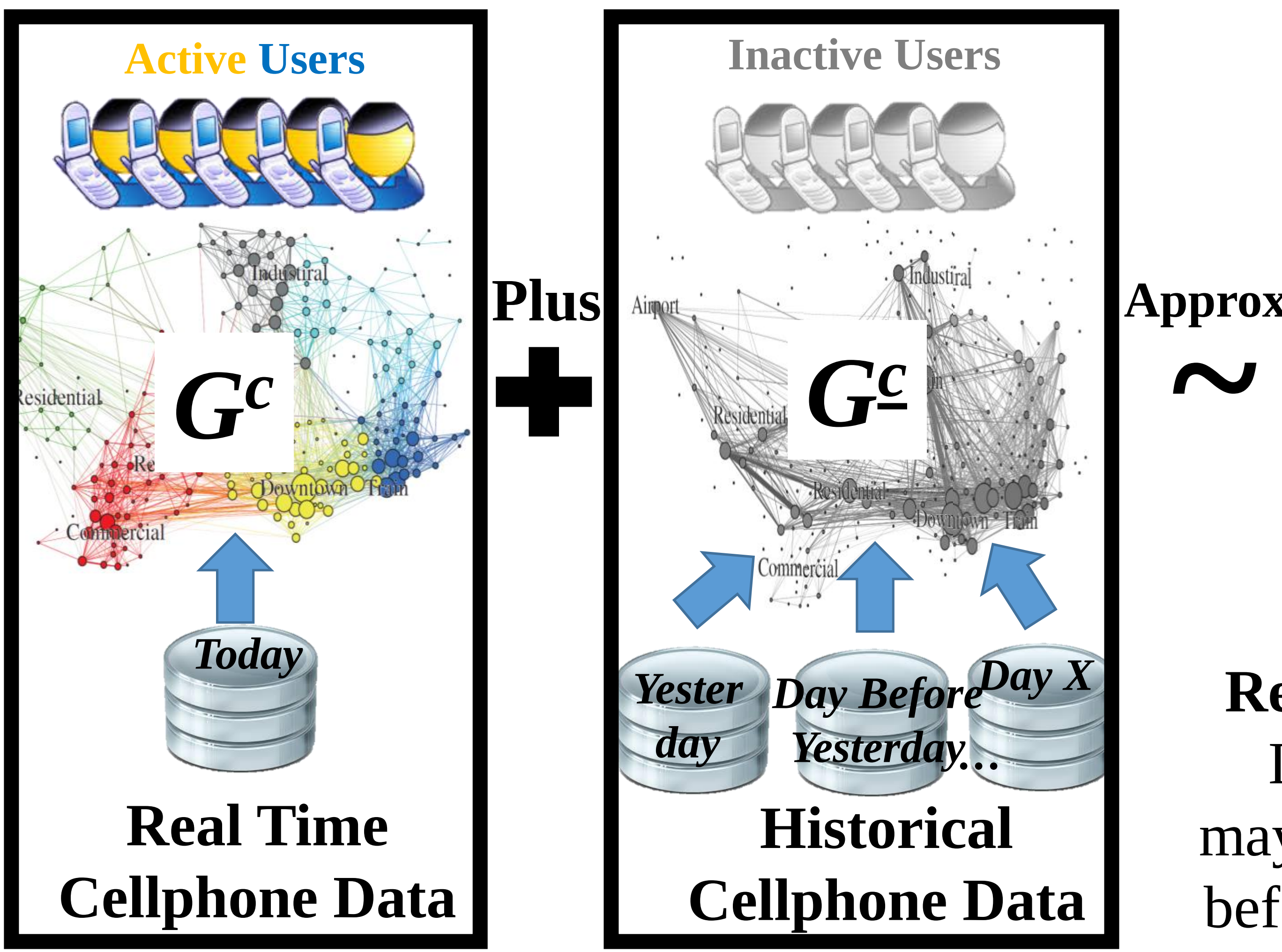


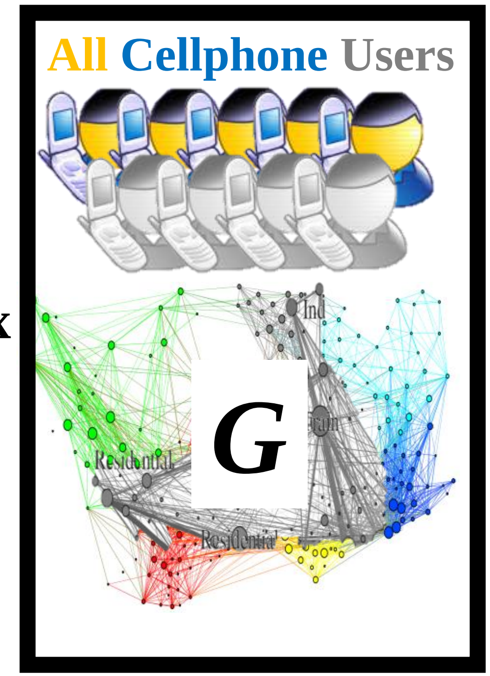
$\boldsymbol{G^{\underline{c}}}$

Key Challenge: $G^{\underline{c}}$ for **inactive users** is **Unknown**

Online Inference:

Solution: Infer $G^{\underline{c}}$ by **accumulatively** using historical data



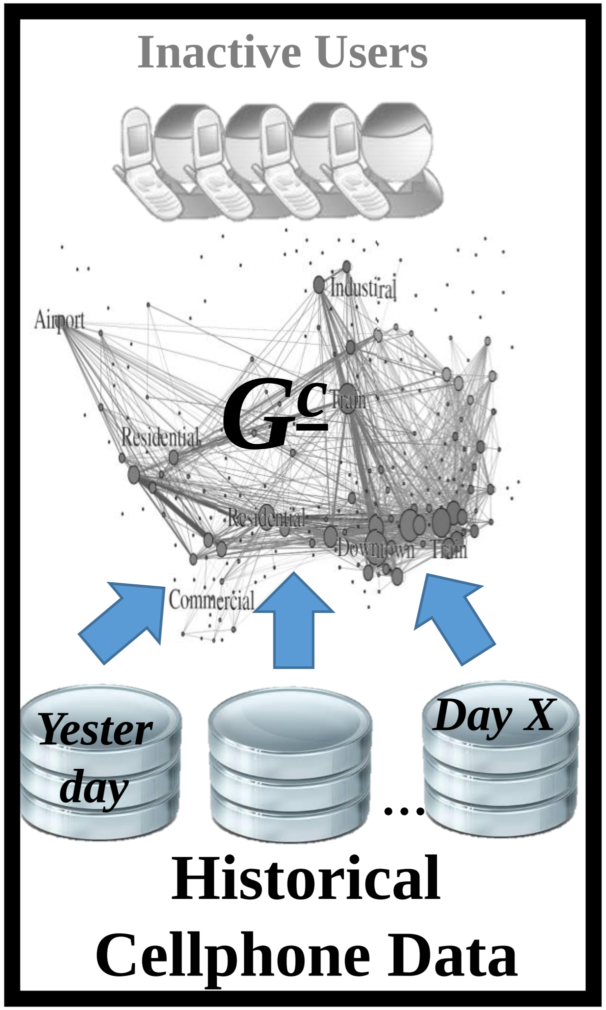


Repeatable Trip:

Inactive users may use cellphones before for same trip

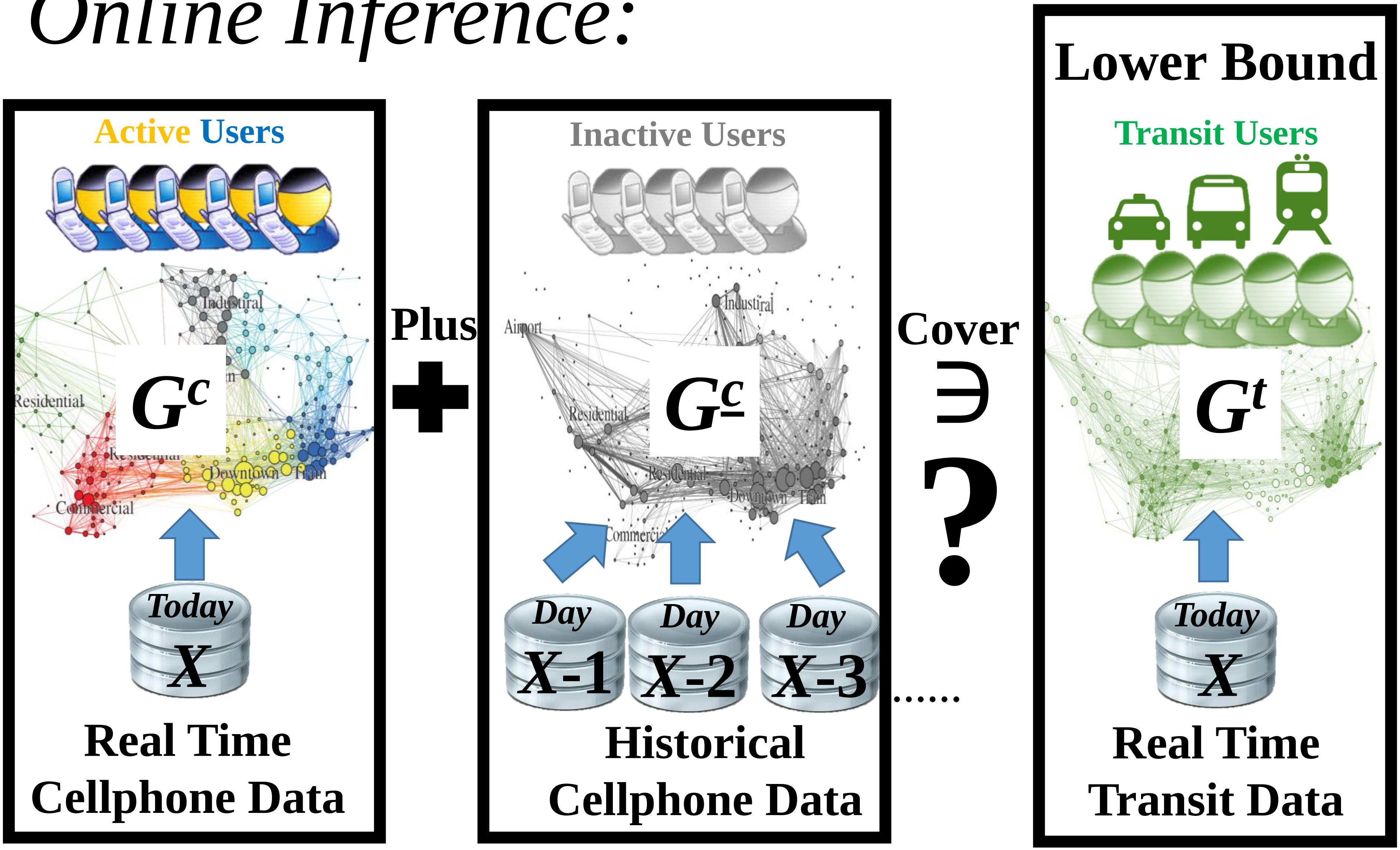
Design Issue: Accumulation

- When to Stop Accumulatively Using Historical Data?
  - One Day or One Week or One Month
- Avoiding **Under or Overestimated**
- Finding a Bound by another **Data Source** to stop the accumulation
- Using Mobility from Transit Data as a **Lower Bound** for Total Mobility



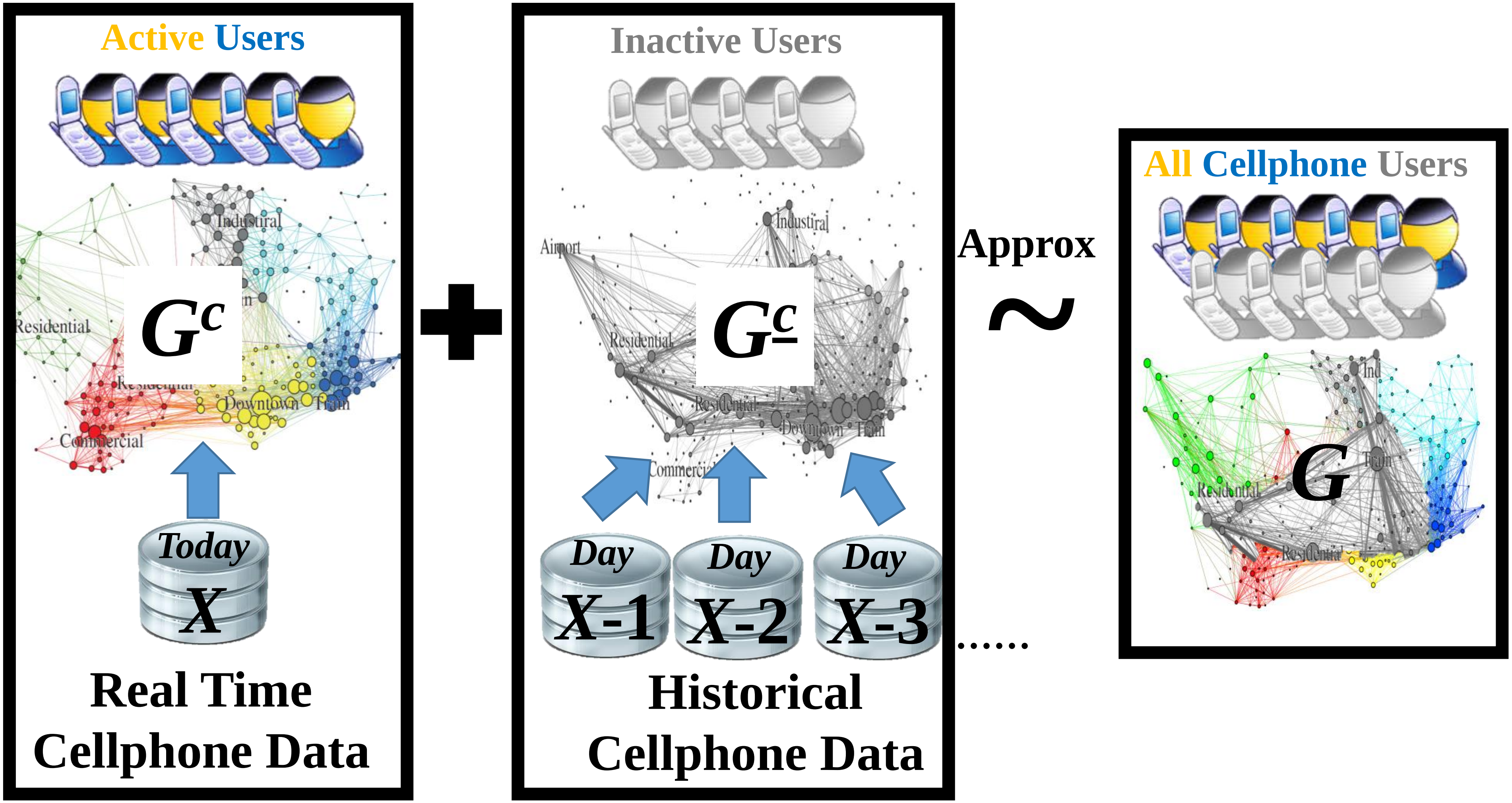
UNIVERSITY OF MINNESOTA
Driven to Discover

Online Inference:



Stop Accumulation, if G^c plus $G^{\underline{c}}$ covers G^t in terms of edge weights

Online Inference:



Using G^c plus $G^{\underline{c}}$ to approximate G for all inter region mobility

Outline

- Introduction
- Design
- **Evaluation**
- Application
- Conclusion

Evaluation Summary

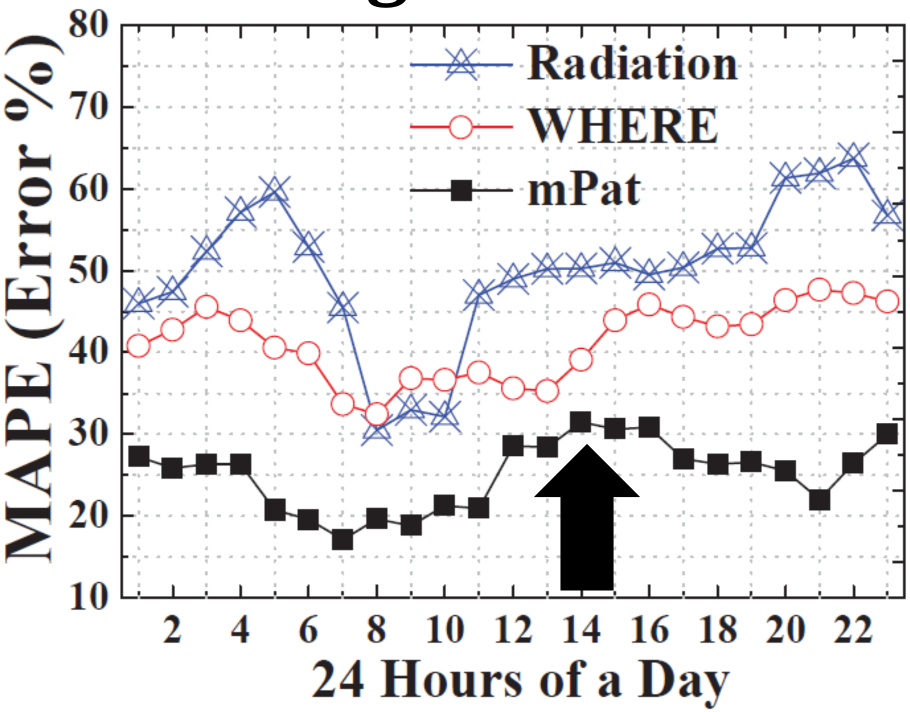
- **Comparison**:
  - **Radiation**: Statistical Model without Real-Time Data
  - **WHERE**: Single-Source Model with Cellphone Data
- **Metric**: Mean Average Percent Error (**MAPE**)

$$\frac{100}{n}\sum_{i=1}^{n}\frac{|\bar{\mathbf{T}}_i - \mathbf{T}_i|}{\bar{\mathbf{T}}_i}$$

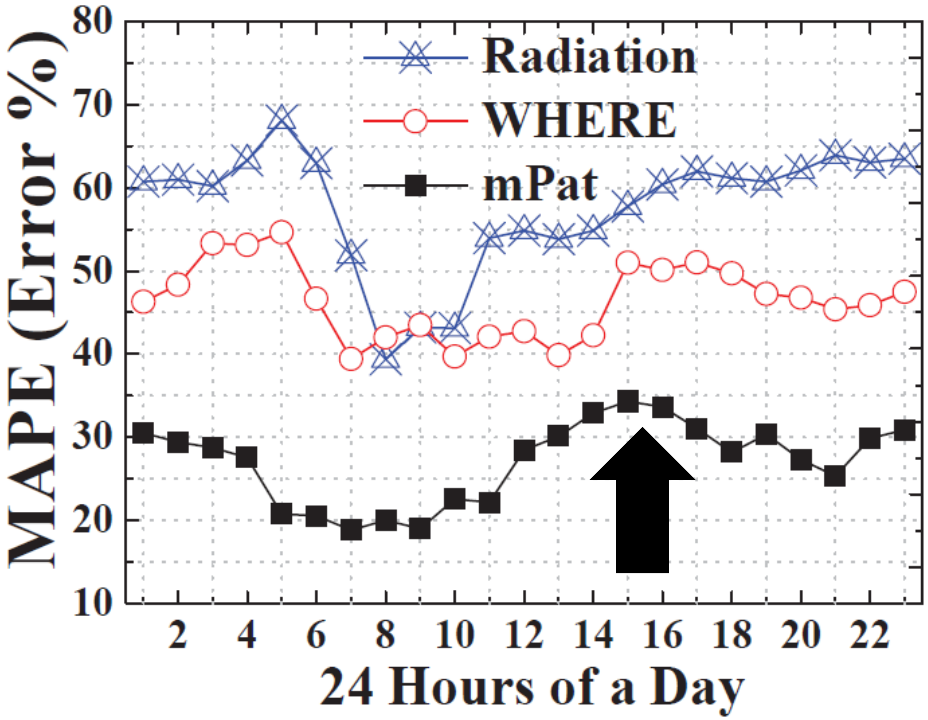
  - Among $n = 496 \times 496 = 246016$ region pairs, i.e., an OD pair
  - $\mathbf{T}_i$: **Inferred Mobility** in an OD pair i
  - $\bar{\mathbf{T}}_i$: **Real Mobility** in an OD pair i (**Ground Truth**)
- **Ground Truth:**
  - Obtained by a location updating dataset of 7 million cellphone users
  - Logging locations of all users in every 15 mins even without activities
  - Did not use in analysis since it cannot generalize to their cities, and need extra support in terms of software, hardware, and policies

Accuracy on different levels

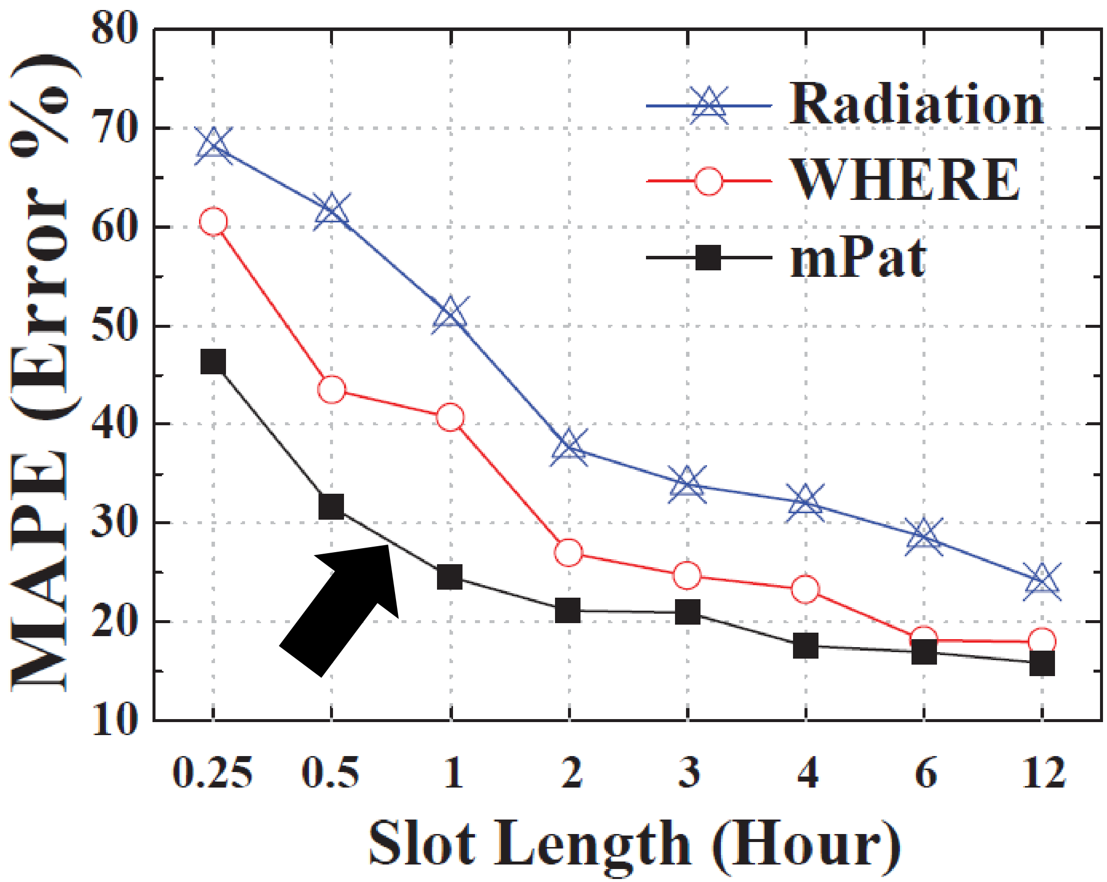
Region Levels



Street Levels

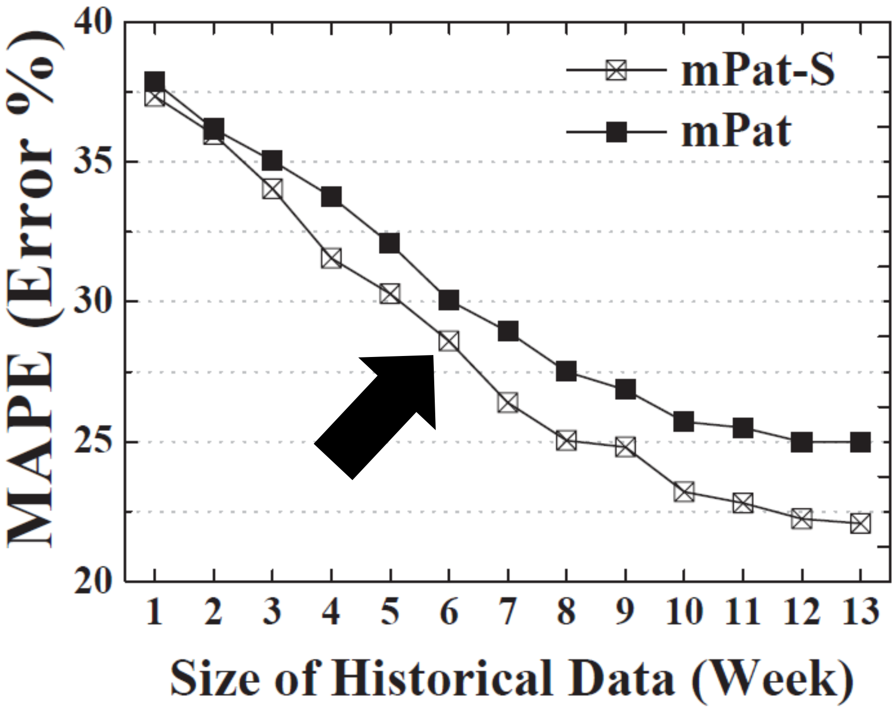


Impact of slot length

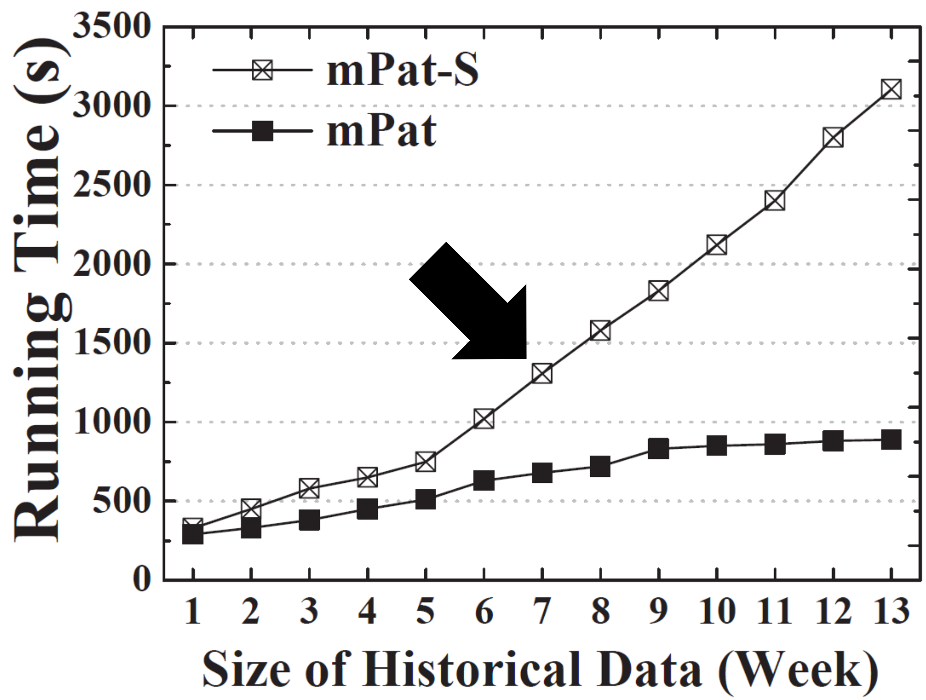


Impact of Historical Data

Accuracy



Running Time



mPat-S: using **all** historical cellphone data **without** analyzing the **correlation** with transit data

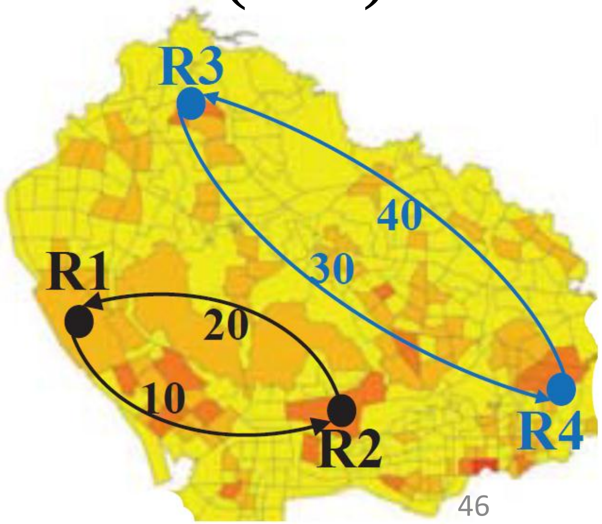
Outline

- Introduction
- Design
- Evaluation
- **Application**
- Conclusion

Inter Region Transit

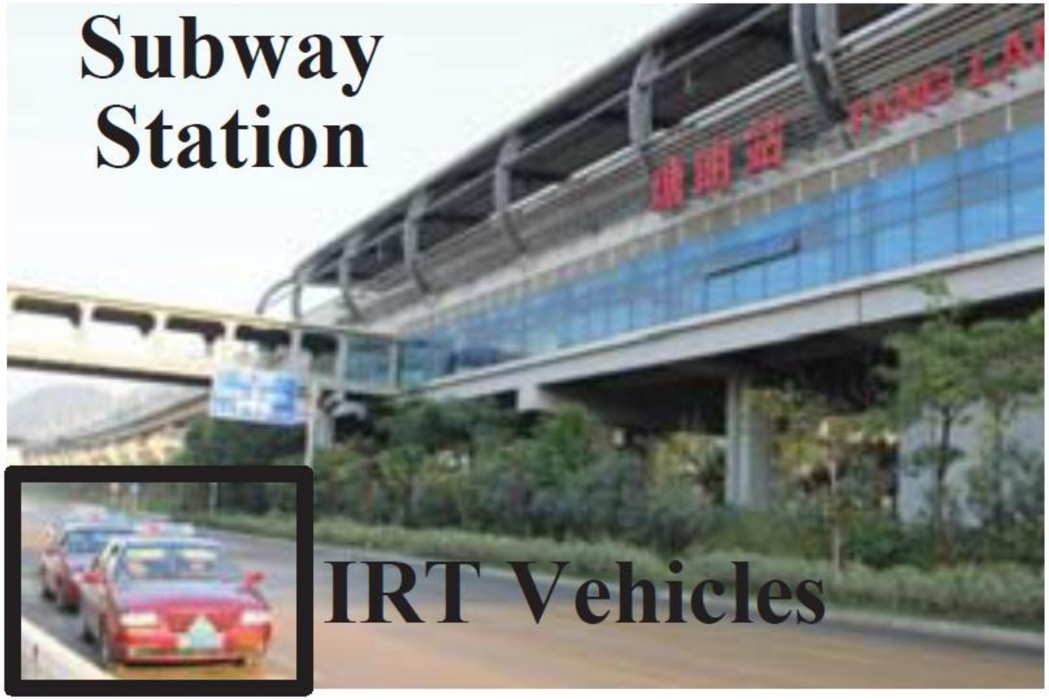
- Based on mPat, finding urban region pairs with
  - High human mobility (**Cellphone Data**)
  - Low public transit mobility (**Transit Data**)
  - Indicating **Inadequate** Transit Service

- Providing **non-stop express** inter region transit (IRT) services between these region pairs



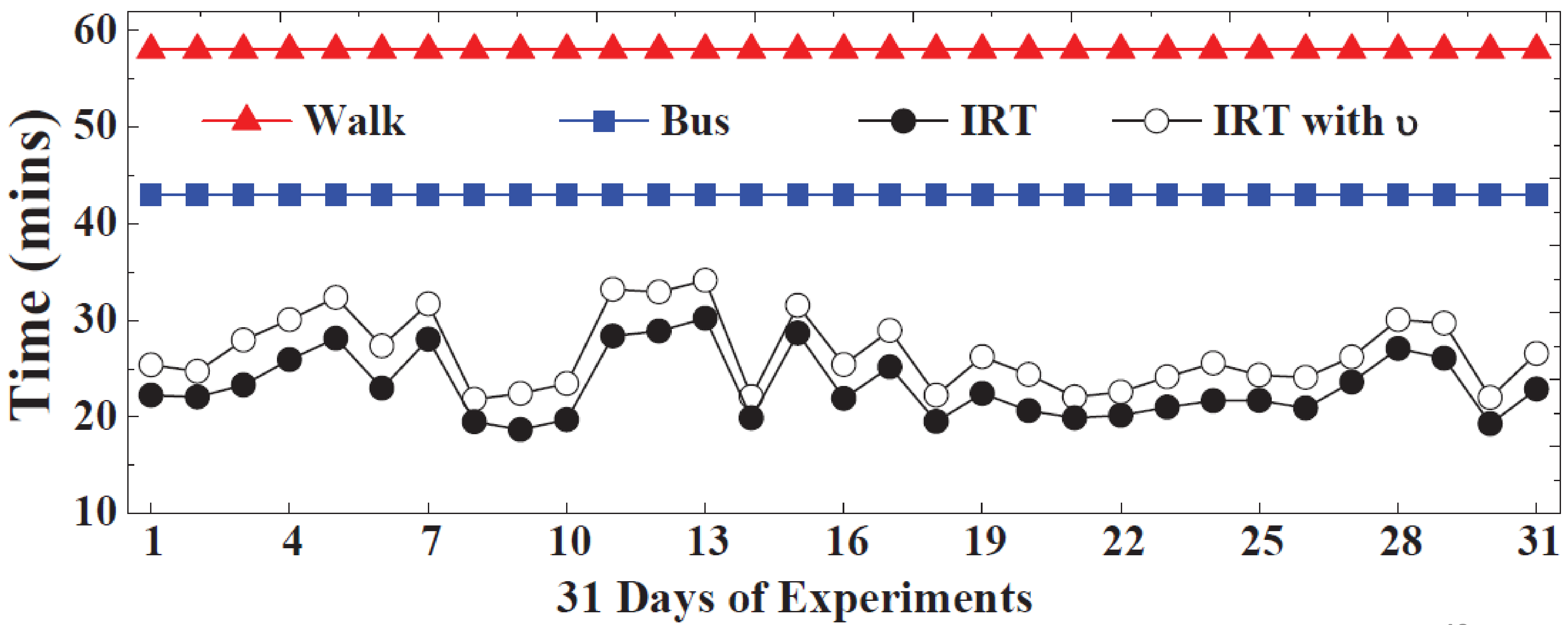
Real World Implementation

- Implementing between two urban regions
- Using **3 taxis as IRT Vehicles** to deliver **12 volunteers**
- Logging **Travel Time** for 30 days



Experiment Results

- Comparing IRT with **walking** and taking regular **bus**
- Quantifying **speed difference** between taxicabs and buses with a factor *v*



Conclusion

- Design an architecture **mPat** for the analysis and inference of the human mobility with a 75% inference accuracy
- Two key insights
  - models based on **single-source data** introduce **biases**, which can be mitigated by **multi-source data**
  - multi-source data can be used for **cross-referencing** to increase the performance

Thanks