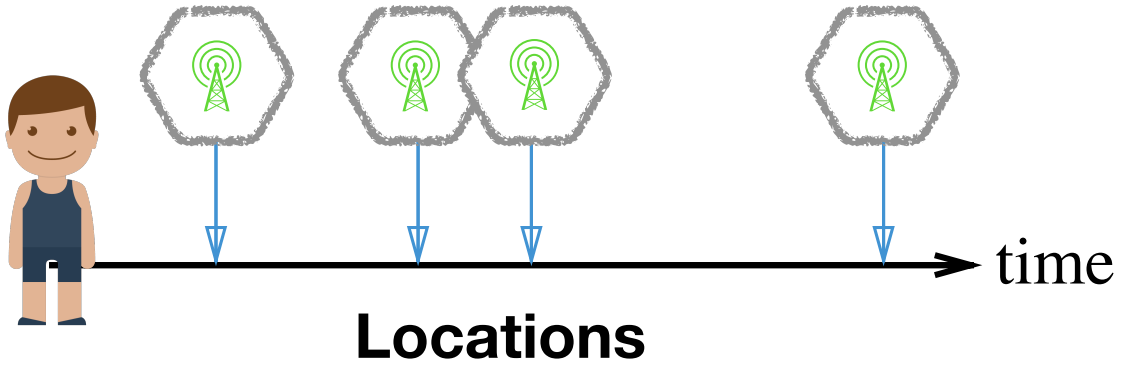
Takeaways in Large-scale Human Mobility Data Mining

Guangshuo Chen, Aline Carneiro Viana, and Marco Fiore

Human Mobility Investigation



General **Networking**

- Prediction
- Reconstruction
- Characterization

- Paging management
- Caching optimization
- Resource allocation

Human Mobility Investigation

1. **Data collection**

   Operator (CDR), App + Volunteers, WiFi, …

2. **Data preliminary/processing**

3. Data utilization

- Prediction
- Reconstruction
- Characterization

- Paging
- Caching
- Resource allocation

How to Obtain 100K Users' Locations?

WiFi? App + Volunteers? No! Only CDR!

- (Legacy) Call Detail Records
- (Now) Charging Data Records
  - 3GPP TS 32.240
  - Calls, SMS, data sessions, mobility updates, on/off, …
  - Necessary for billing
  - Large populations

CDR Data

Phone	Time	Location
User ID 1	2015-01-01 06:47:56	(19.028, -98.209)
User ID 1	2015-01-01 08:23:08	(18.993, -98.202)
User ID 2	2015-01-02 16:59:34	(19.025, -98.217)

Payload of communication events

How to Obtain 100K Users' Locations?

from CDR datasets?

- Collaboration
  - Operators
  - Data companies
  - Researchers

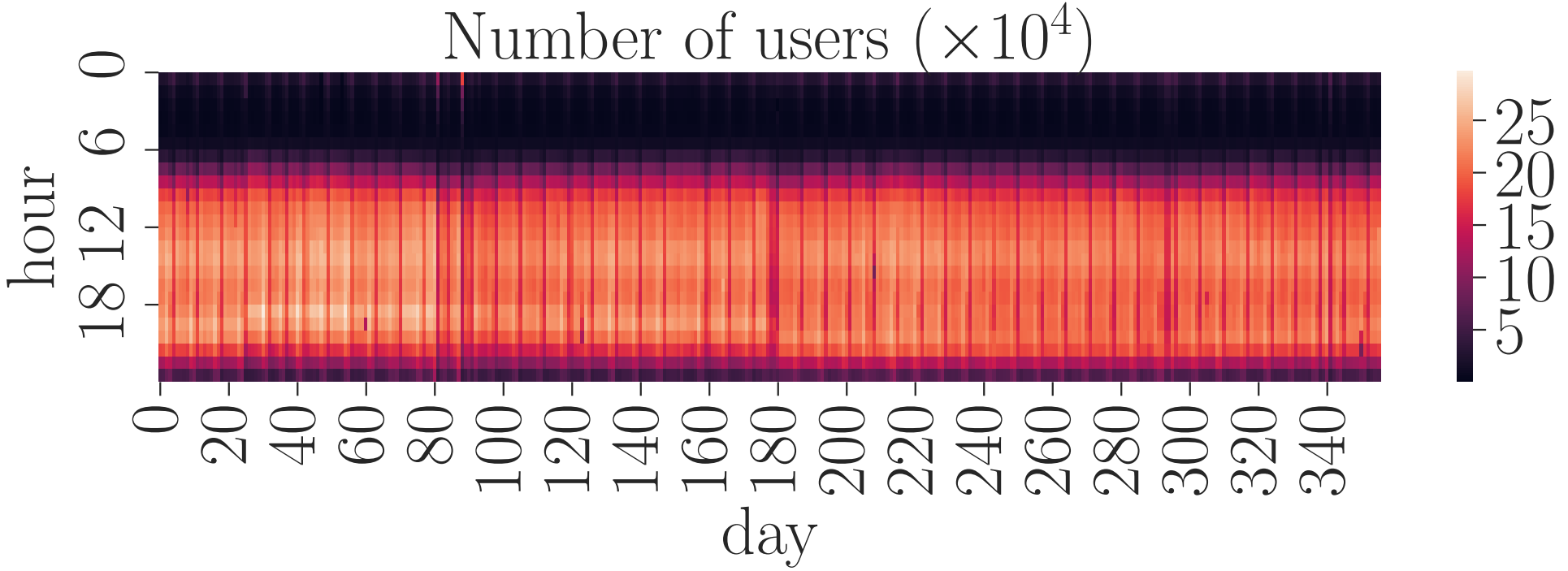
- **Internet!**
  - Publicly available datasets
  - Where to find?

Publicly Available CDR datasets

- Data competitions
  - ~~Orange D2D 2013~~
  - ~~TIM Big Data Challenge 2015~~
  - flowminder.org, South Africa
  - dandelion.eu, Milan, Italy
  - (**Chinese**) kesci.com Big Data Competitions
    - 2016-2017, China Unicom, Shanghai, 642K users (2016), 1 weeks
  - (**Chinese**) zjdex.com, China Mobile, Hangzhou, 7K users, 1 month

Data Preliminary Is Always Required

- Imperfectness
  - temporal heterogeneity
  - abnormalities



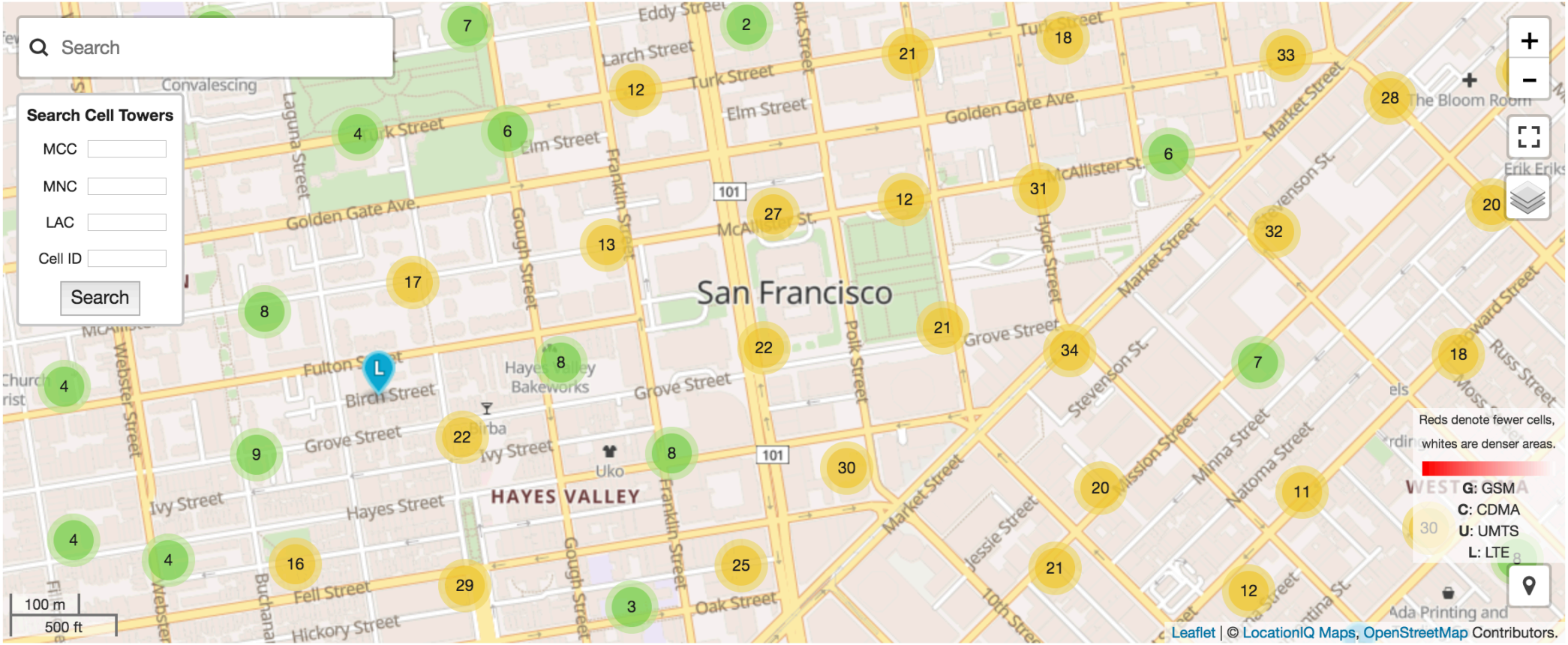
Data Preliminary: Common Practices

- **Extract location coordinates**
- **Filter out “bad” users**
- Reduce resolutions
- Segment observing periods
- **Correlate with mobility loss**
- Perform controlled experiments
- Fill spatiotemporal gaps

Extract Location Coordinates

- Locations in CDR
  - Geographical coordinates
  - **Cell tower IDs (extraction required)**
    - **MCC-MNC-LAC-CID**
- Reliable cell tower locations
  - **France OpenData, www.data.gouv.fr**
- Crowdsensing, third-party services
  - OpenCelliD, Google Geolocation, Unwired Labs, OpenSignal, and Mozilla Location Service

Unwired Labs

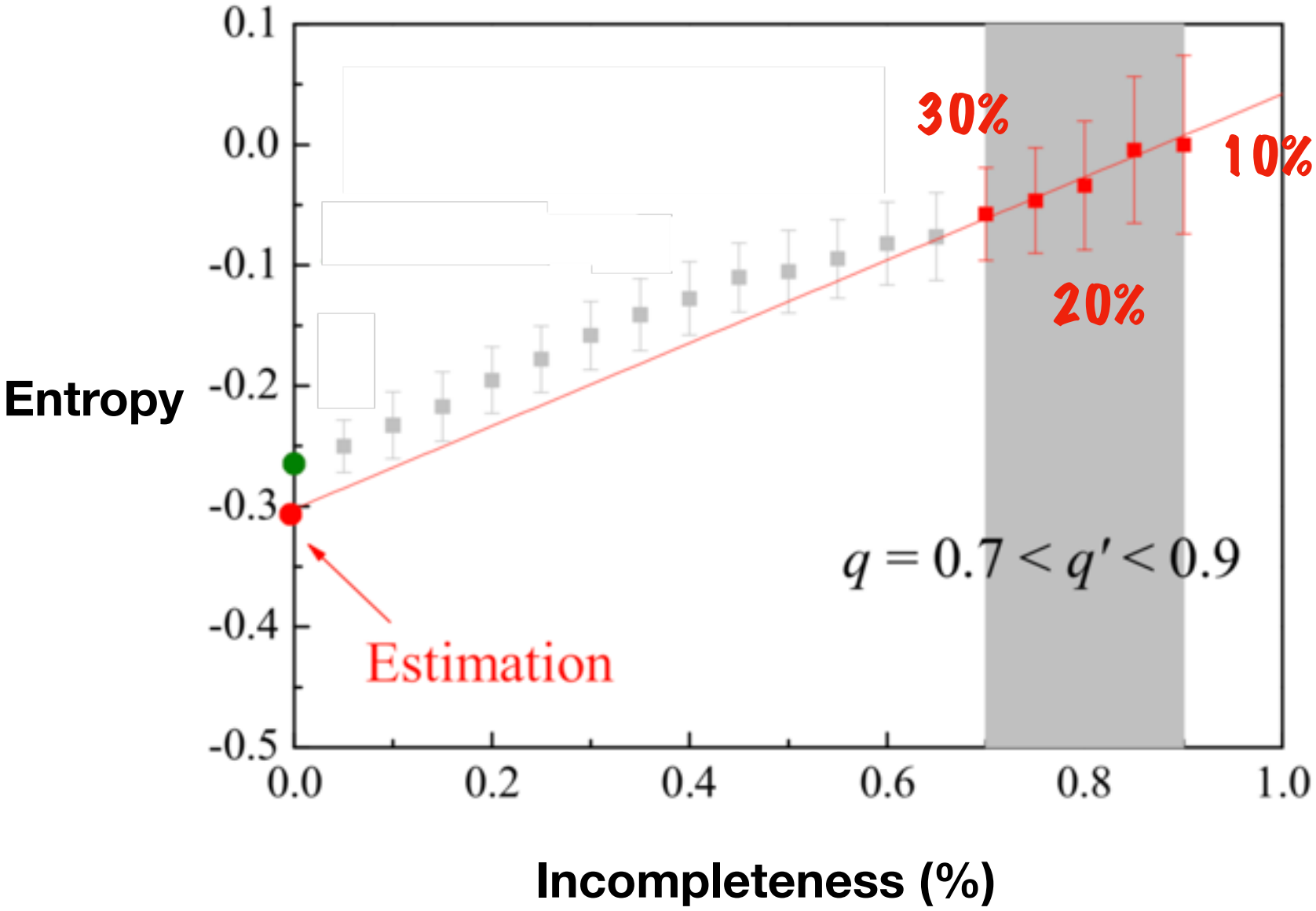


Filter Out Users

- Shrinking user population
  - 6M -> 100K
    - Gonzalez et al. "Understanding individual human mobility patterns," Nature, 2008
  - 10M -> 50K
    - Song et al. "Limits of predictability in human mobility," Science, 2010
  - 1M -> 700
    - Hoteit et al. "Estimating human trajectories and hotspots through mobile phone data." Computer Networks, 2014

- **To cut off, to let go, and to move on**

Collaborate with Mobility Loss

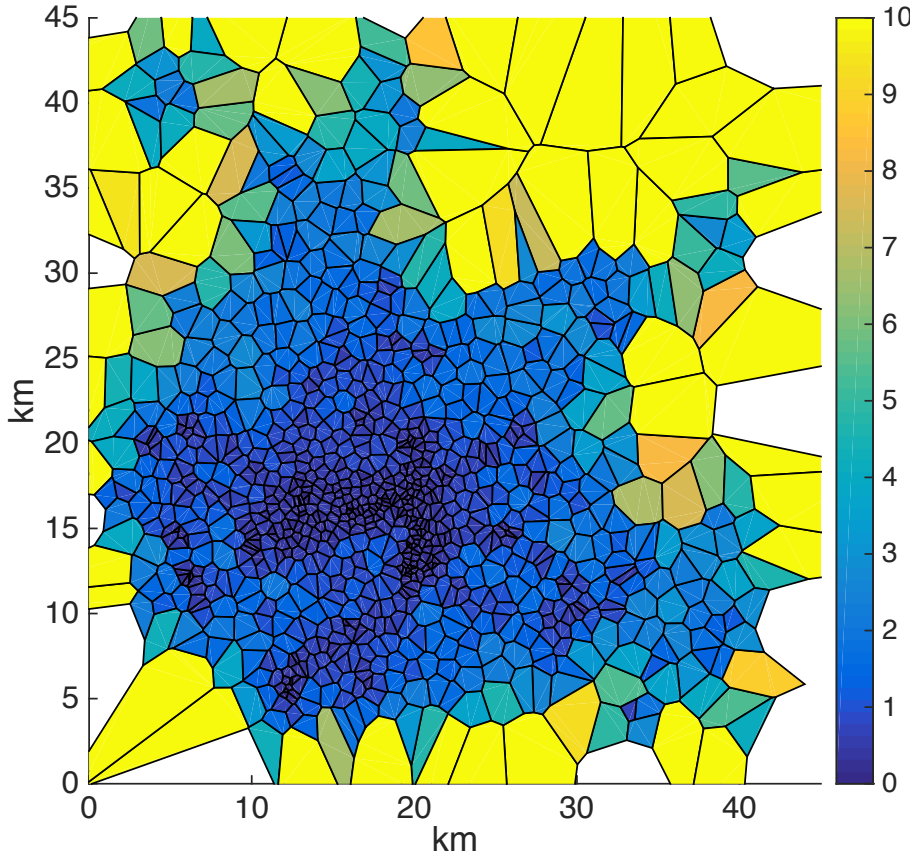


Before Filtering: Mobility Measurement

- Location
  - Cell coverage
  - Repetitiveness
  - Categories
- Individual
  - Displacement
  - Travelled distance
  - Radius of gyration

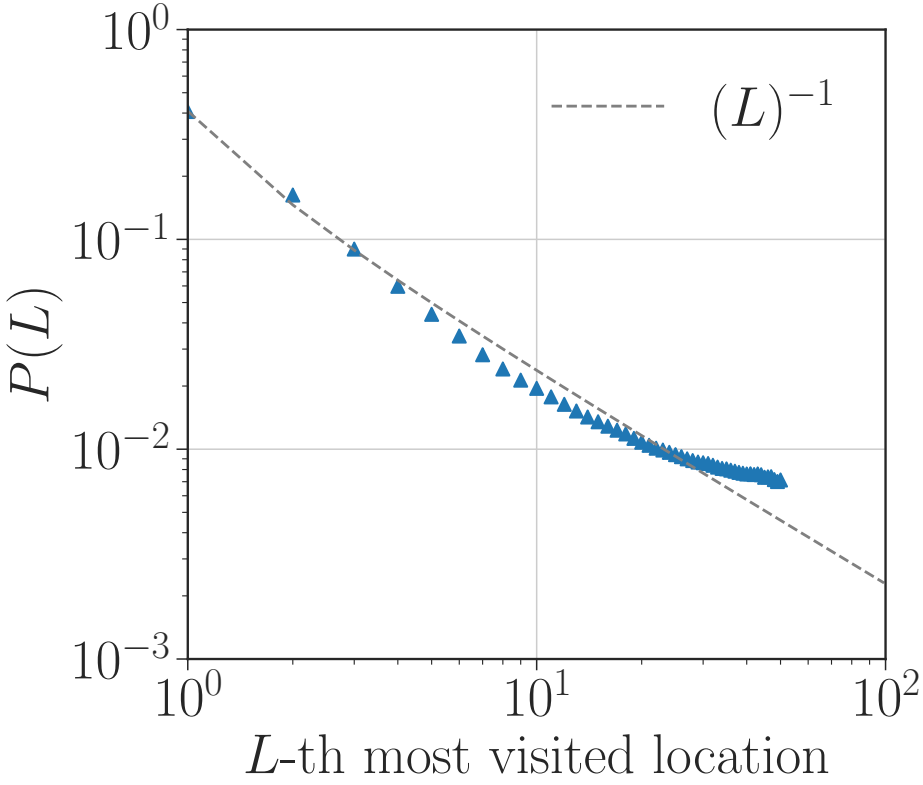
Before Filtering: Mobility Measurement

- Location
  - **Cell coverage**
  - Repetitiveness
  - Categories
- Individual
  - Displacement
  - Travelled distance
  - Radius of gyration



Before Filtering: Mobility Measurement

- Location
  - Cell coverage
  - **Repetitiveness**
  - Categories
- Individual
  - Displacement
  - Travelled distance
  - Radius of gyration

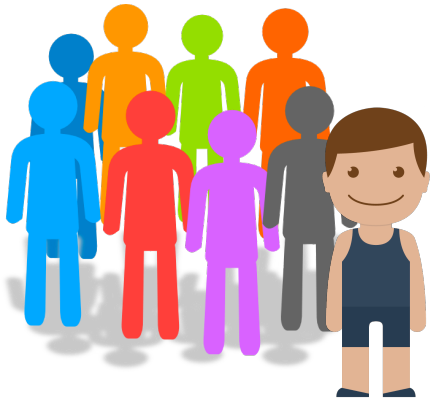


Before Filtering: Mobility Measurement

- Location
  - Cell coverage
  - Repetitiveness
  - **Categories**
- Individual
  - Displacement
  - Travelled distance
  - Radius of gyration

Home

Work

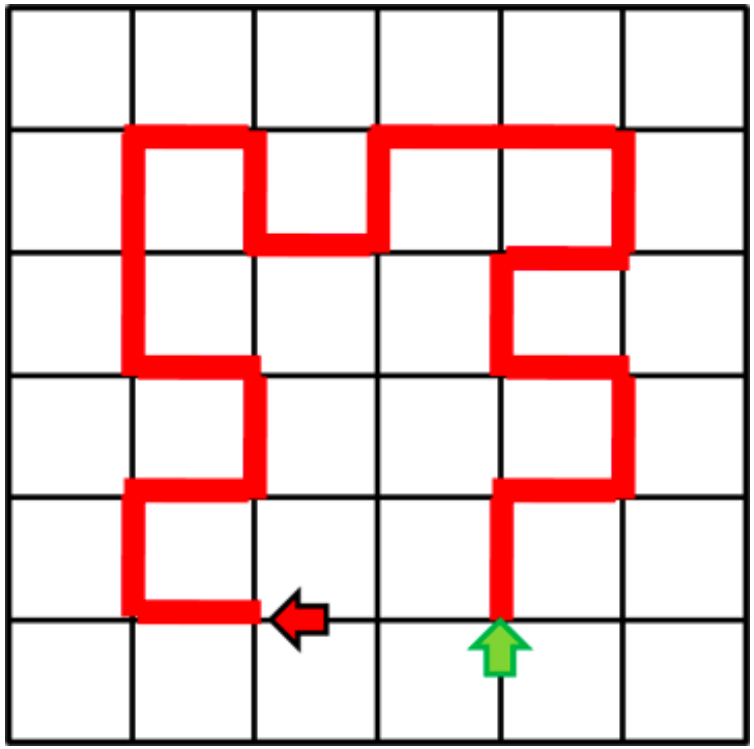


Before Filtering: Mobility Measurement

- Location
  - Cell coverage
  - Repetitiveness
  - Categories
- Individual
  - **Displacement**
  - Travelled distance
  - Radius of gyration

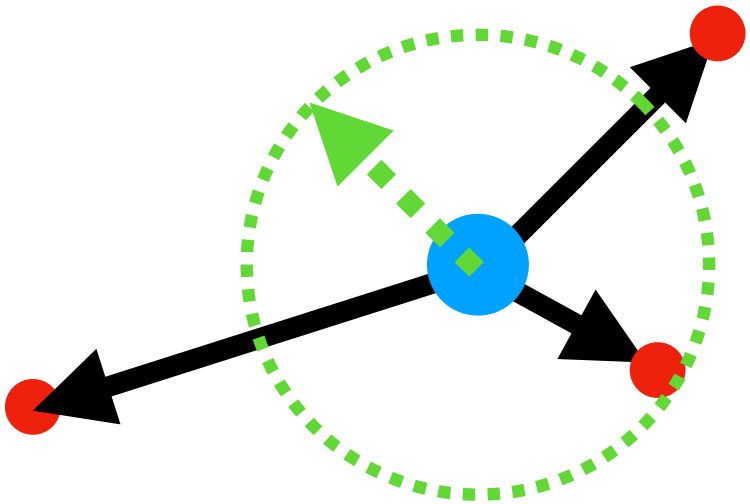
Before Filtering: Mobility Measurement

- Location
  - Cell coverage
  - Repetitiveness
  - Categories
- Individual
  - Displacement
  - **Travelled distance**
  - Radius of gyration



Before Filtering: Mobility Measurement

- Location
  - Cell coverage
  - Repetitiveness
  - Categories
- Individual
  - Displacement
  - Travelled distance
  - **Radius of gyration**



$$RG_{\text{time}} = \sqrt{\frac{\sum_{k=1}^{N} s_k \cdot (\mathbf{r}_k - \mathbf{r}_{\text{cm}}^{\text{time}})^2}{\sum_{i=1}^{N} s_i}},$$

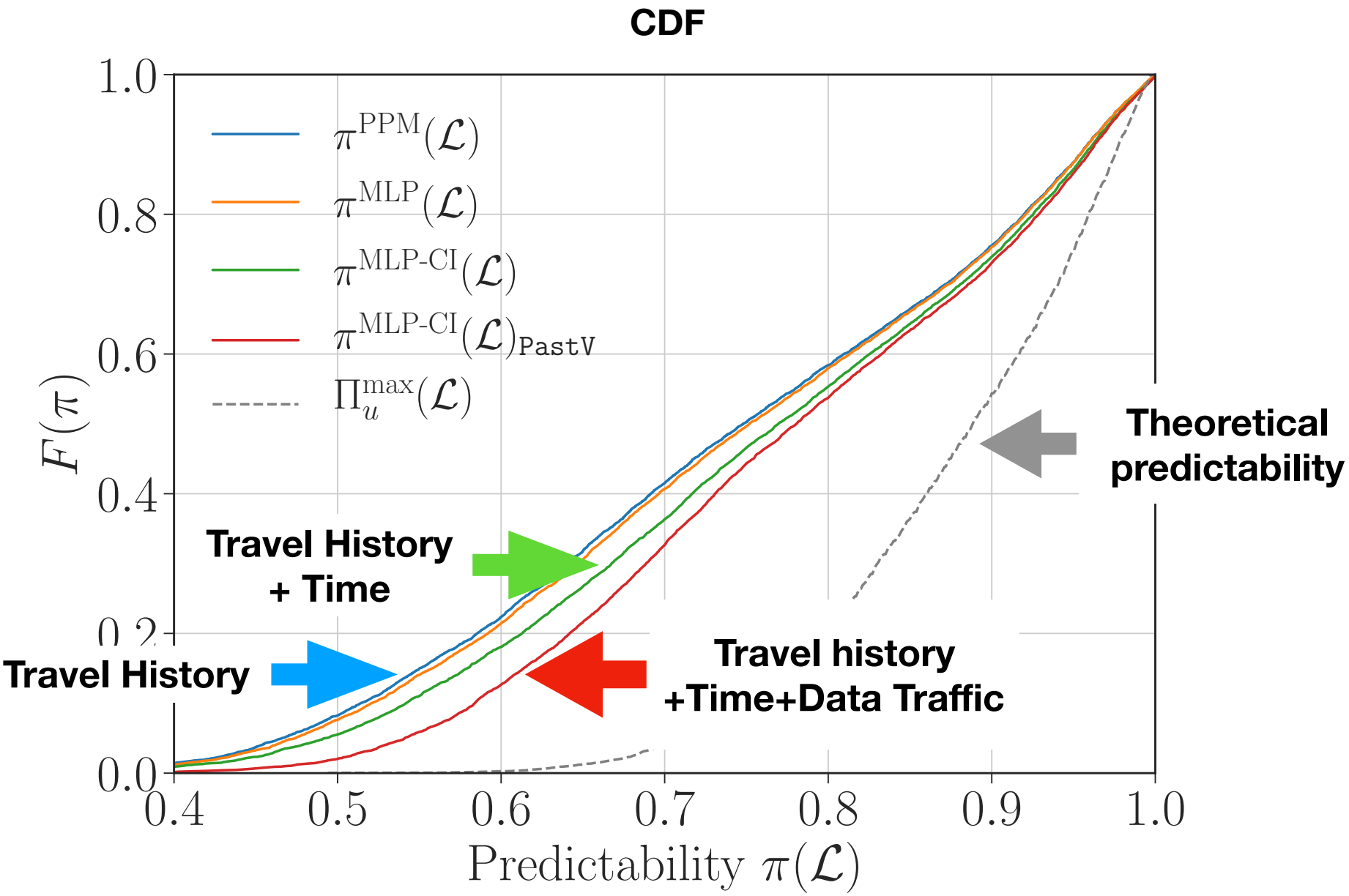
Case Study: Next-location Prediction

- Dataset: CDR (voice calls), ~1M users, 15 months
- User filtering:
  - Metropolitan area
  - Radius of gyration < 32km (urban, peri-urban)
  - Completeness > 20%
  - Weekdays only
  - Three or more unique locations
  - Data traffic > 1KB/day
  - Active days > 150 days
- Users of study: 7K
- Travel history: >150 days for each user

- **Problem:**

$$\hat{l}_t = \arg \max_{l \in \text{Cells}} P(L_t = l | l_{t-1}, l_{t-2}, \cdots)$$

- **Accuracy** Percentage of correct predictions
- **Theoretical** predictability
  - Entropy + Fano's inequality => Accuracy upper bound
- **Practical** predictive models
  - PPM: Prediction by partial matching
  - MLP: Multi-layer perceptron


CDF
1.0
0.8
0.6
0.4
0.2
0.0
$F(\pi)$
$\pi^{\mathrm{PPM}}(\mathcal{L})$
$\pi^{\mathrm{MLP}}(\mathcal{L})$
$\pi^{\mathrm{MLP\text{-}CI}}(\mathcal{L})$
$\pi^{\mathrm{MLP\text{-}CI}}(\mathcal{L})_{\mathtt{PastV}}$
$\Pi_u^{\max}(\mathcal{L})$
Theoretical predictability
Travel History + Time
Travel History
Travel history +Time+Data Traffic
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predictability $\pi(\mathcal{L})$

How to Improve Next-location Prediction?

- Still from travel history?
- (2006) Markov Chain, Text Compression Algorithms
  - Song, Libo, et al. "Evaluating next-cell predictors with extensive Wi-Fi mobility data." *IEEE Transactions on Mobile Computing* 5.12 (2006): 1633-1649.
- (2010) Matrix/Tensor Factorization
  - Zheng, Vincent Wenchen, et al. "Collaborative Filtering Meets Mobile Recommendation: A User-Centered Approach." AAAI. Vol. 10. 2010.
- (2012) ARIMA models
  - Li, Xiaolong, et al. "Prediction of urban human mobility using large-scale taxi traces and its applications." *Frontiers of Computer Science* 6.1 (2012): 111-121.
- (2016) Non-parametric Bayesian + MCMC
  - Jeong et al. “Cluster-aided mobility predictions." *INFOCOM 2016,* IEEE, 2016.
- (2016) Recurrent Neural Networks
  - Liu, Qiang, et al. "Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts." *AAAI*. 2016.
- **New techniques are limited**

- Travel history + Context + Deep learning

Travel history

AND

Shared trajectories

(2016) by Jeong et al.

Mobile Traffic

(2018) Call/SMS
by González et al.

(2018) Data Sessions
by us

?

Future

Detailed Traffic

Environment
(WiFi/Bluetooth)

POIs

Questions about mobility data processing?

guangshuo.chen@gmail.com