Time Series Analysis of Mobile Data Usage Reveals Geographic Location

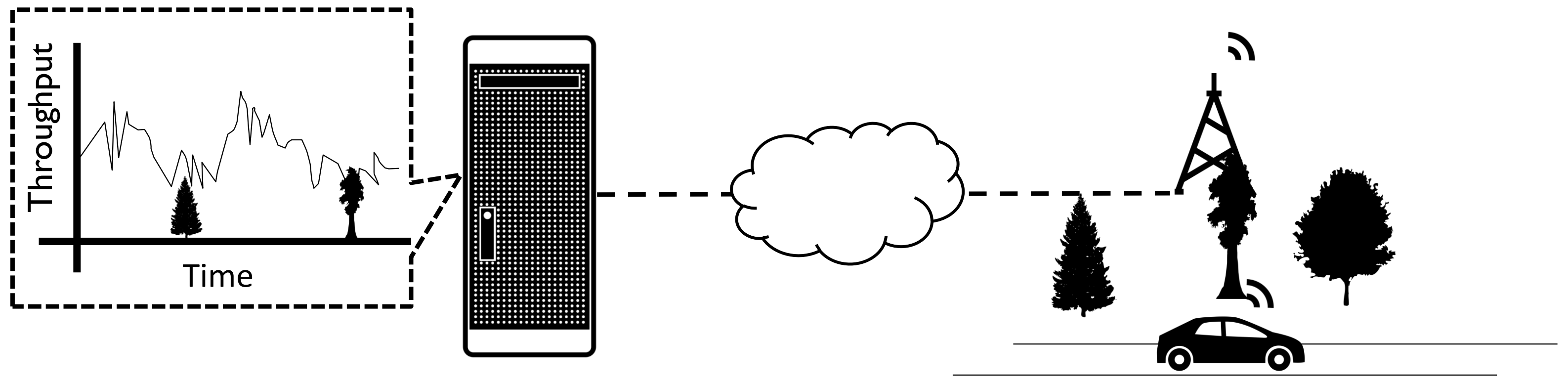
Keen Sung, Erik Learned-Miller, Brian Levine, Marc Liberatore

Mobile location can be deduced by looking at data usage

- Data throughput is correlated with cell phone signal strength
- Signal strength is affected by geographic location
- Location can be inferred using remote measurements of data throughput to a mobile device

Overview of attack

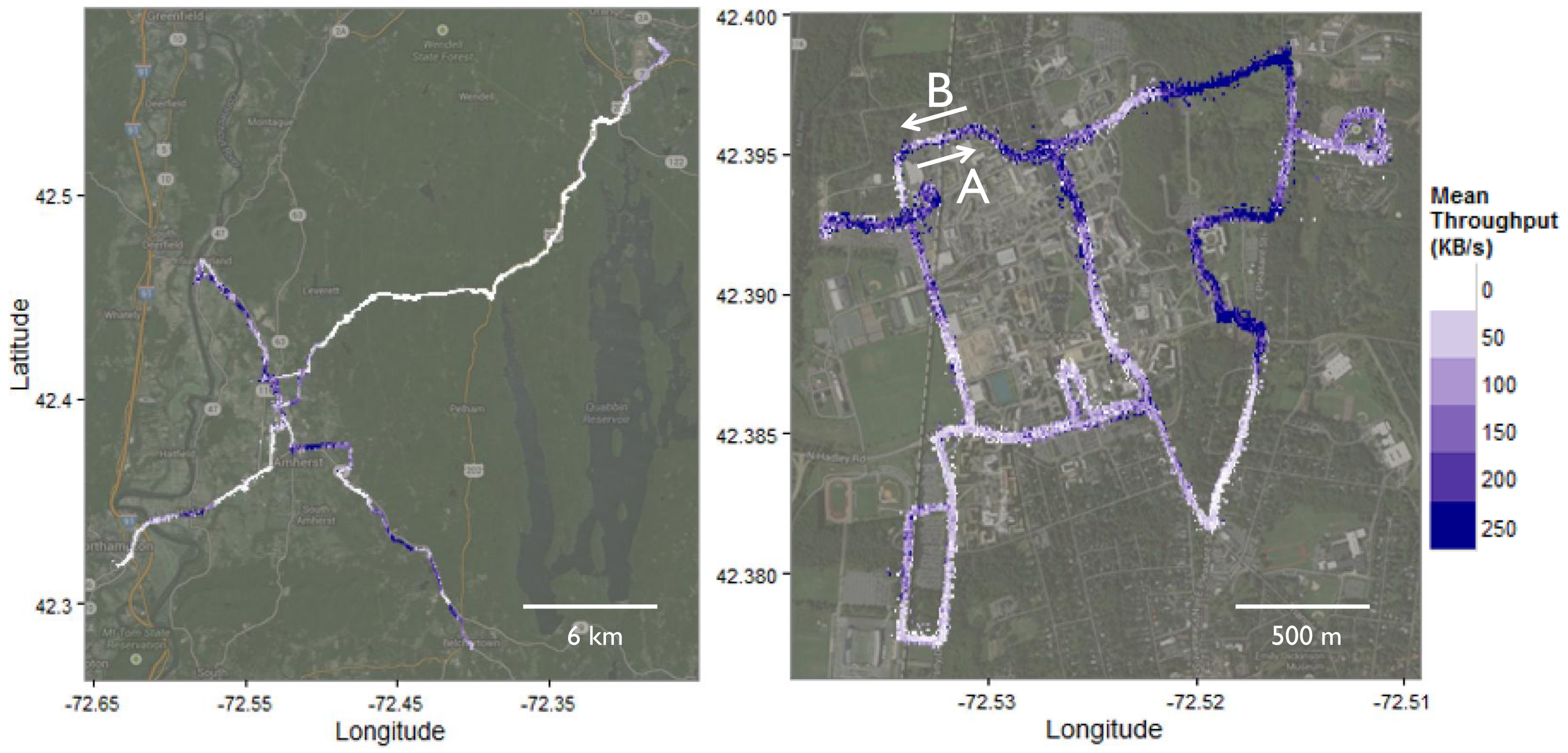
- Attacker serves a stream of data to a user in motion
- User receives data while travelling down a path
- Data throughput varies depending on geography
- Attacker remotely logs the throughput over time, and determines the path using an existing model of known paths



How precisely can mobile location be deduced by examining only throughput?

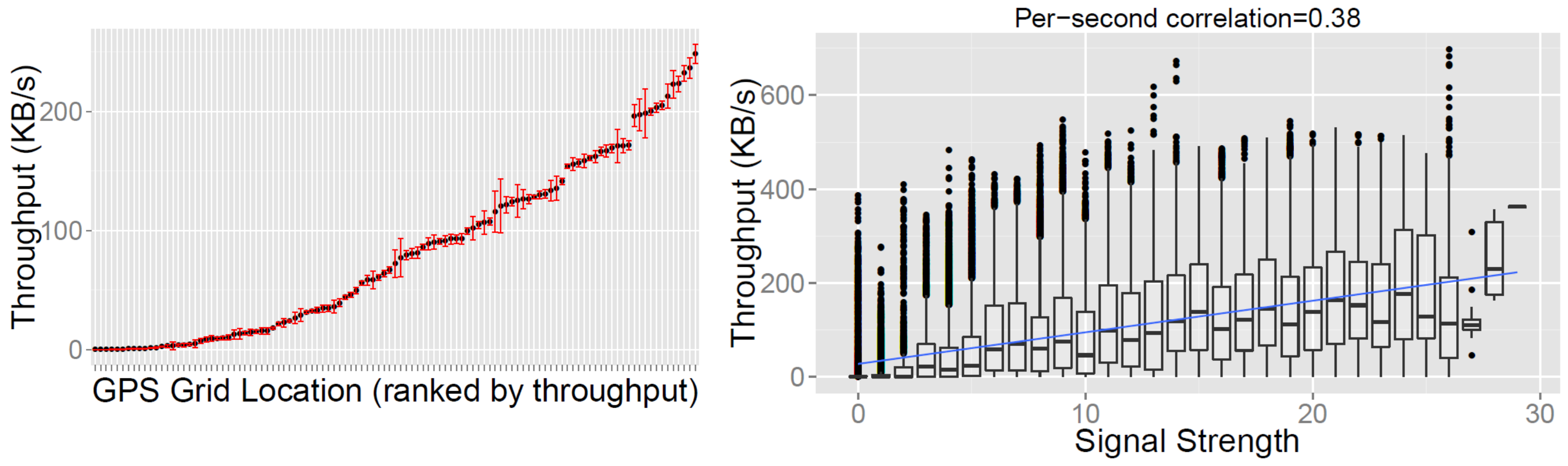
Datasets

- Mobile devices streamed music from an on-campus server via the cell network and recorded GPS location as it travelled down a path
- The on-campus server logged the TCP trace of data to and from the cell phone
- 295 traces recorded with phones travelling to one of four surrounding towns (15 – 30 km)
- 86 traces recorded within the campus bus loop (13 km)

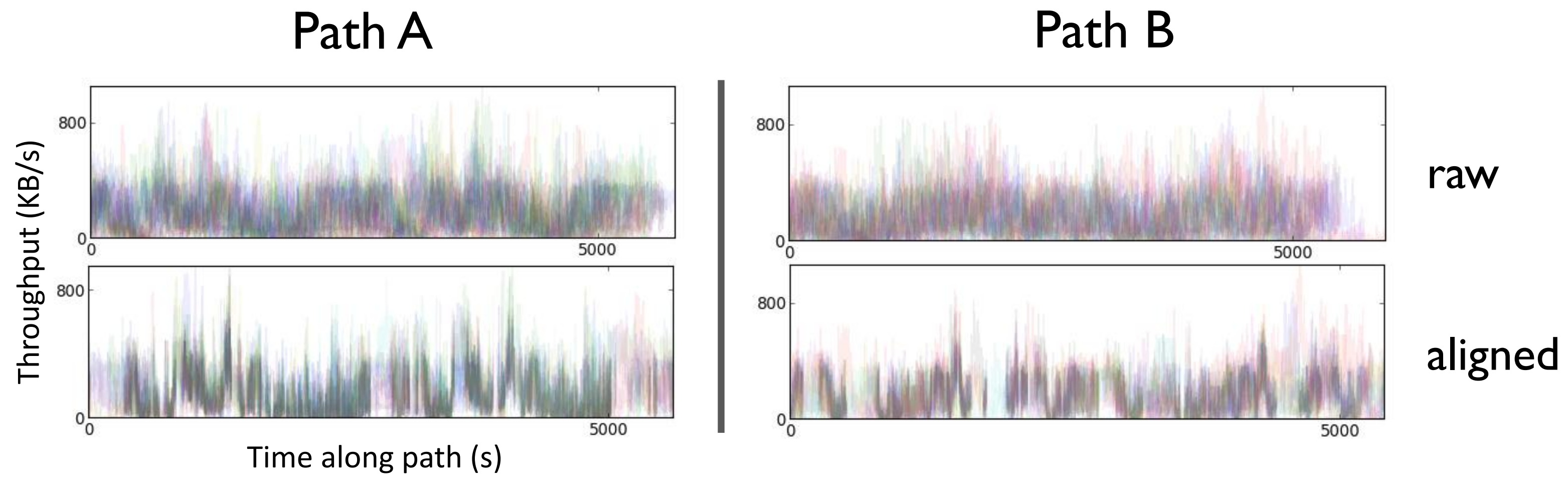


Exploratory analysis

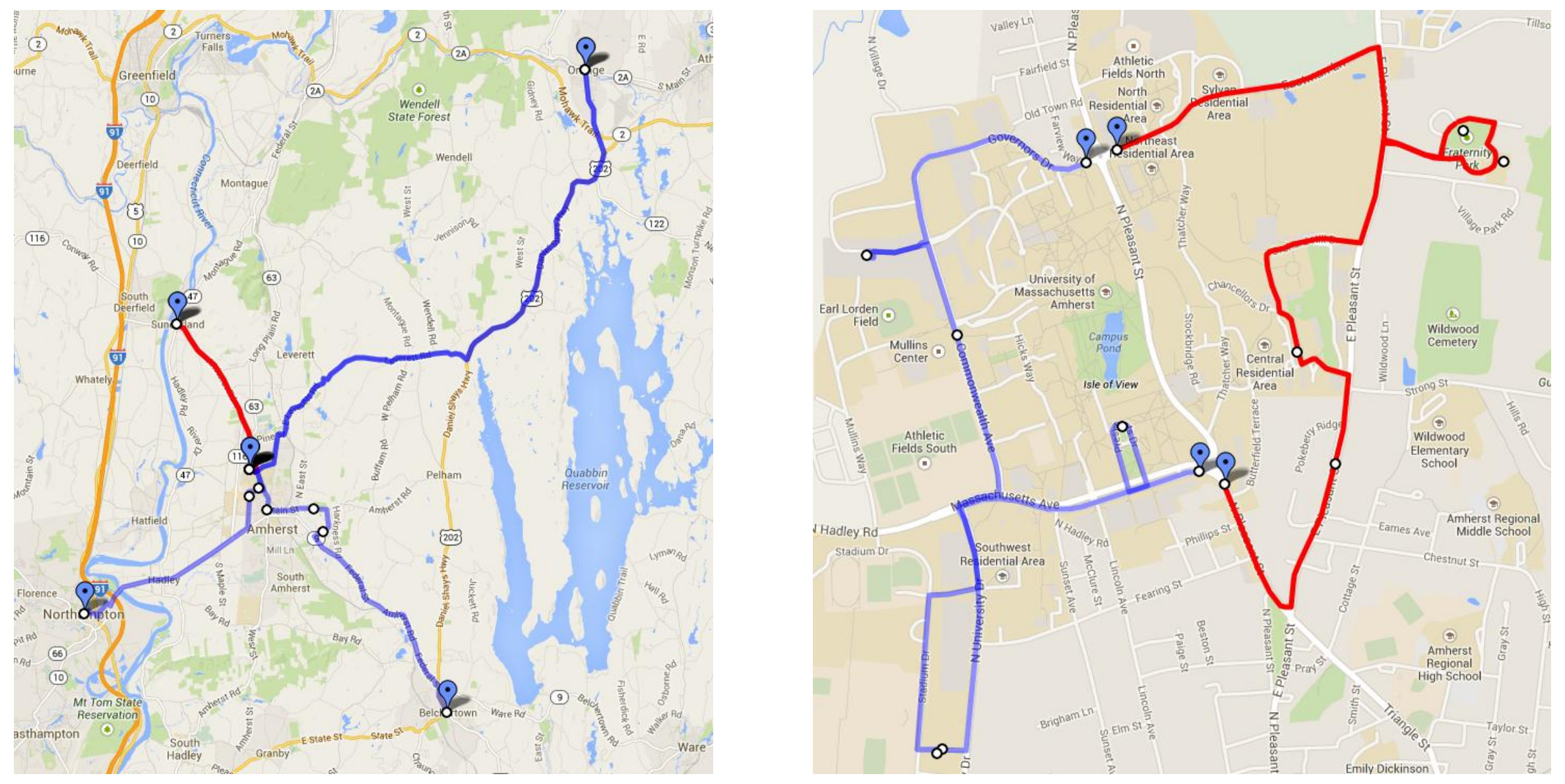
- The mean throughput of 95% of 0.9 km^2 areas is statistically different from at least 85% of the other areas



- Visualizations of traces reveal consistent variations in throughput over geography



Case I: Finding which path has been taken



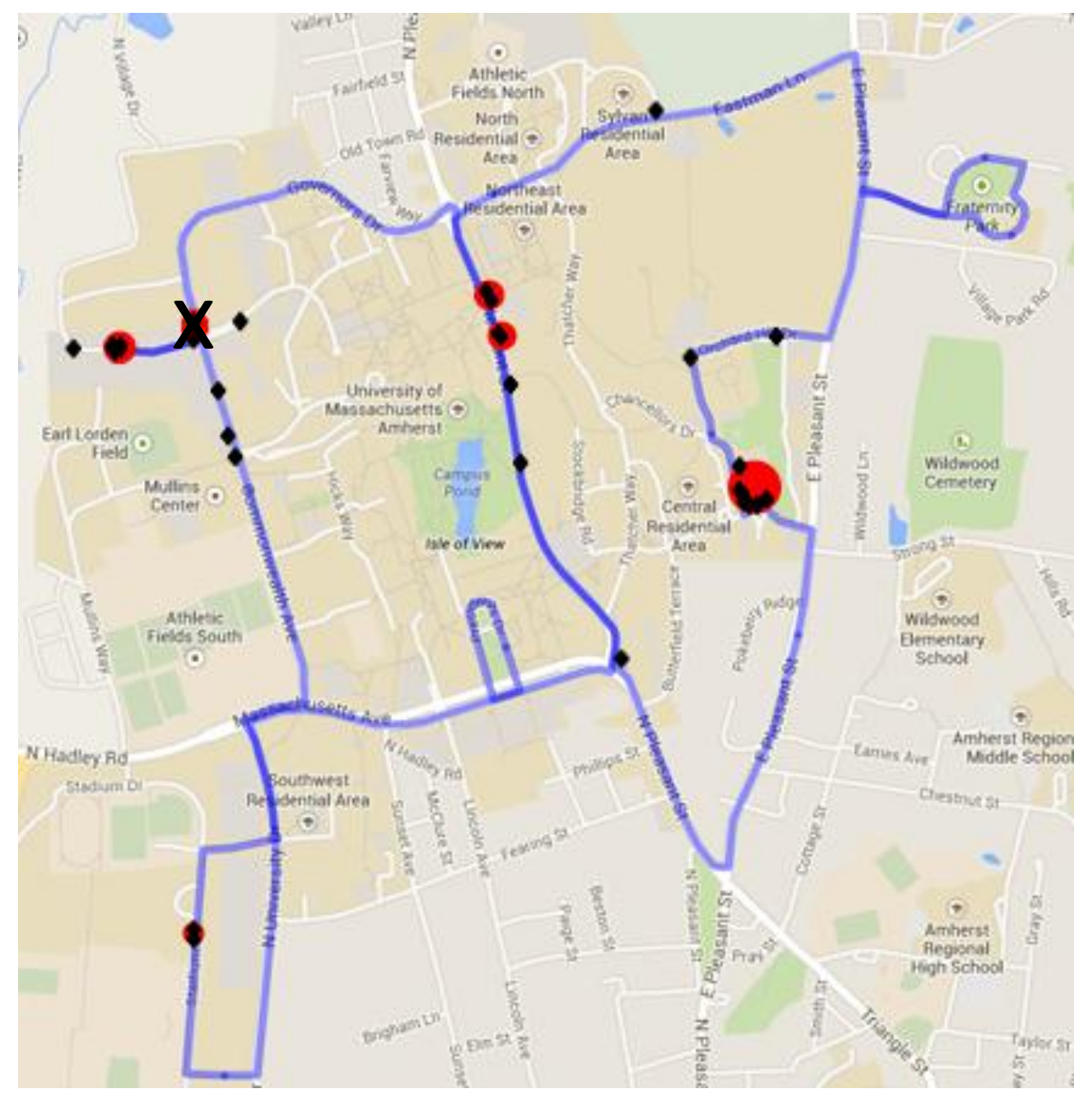
k-nearest neighbors

- Filter out low throughput traces
- Compare the test trace with each training trace by summing the differences in throughput between each corresponding time point
- Choose the most frequent class in the top *k* guesses

Results

- 78% along 4 long paths x 2 directions (15 – 30 km)
- 82% along 2 short paths x 2 directions (2 – 4 km)

Case II: Determining possible locations within a region



1. Score known locations

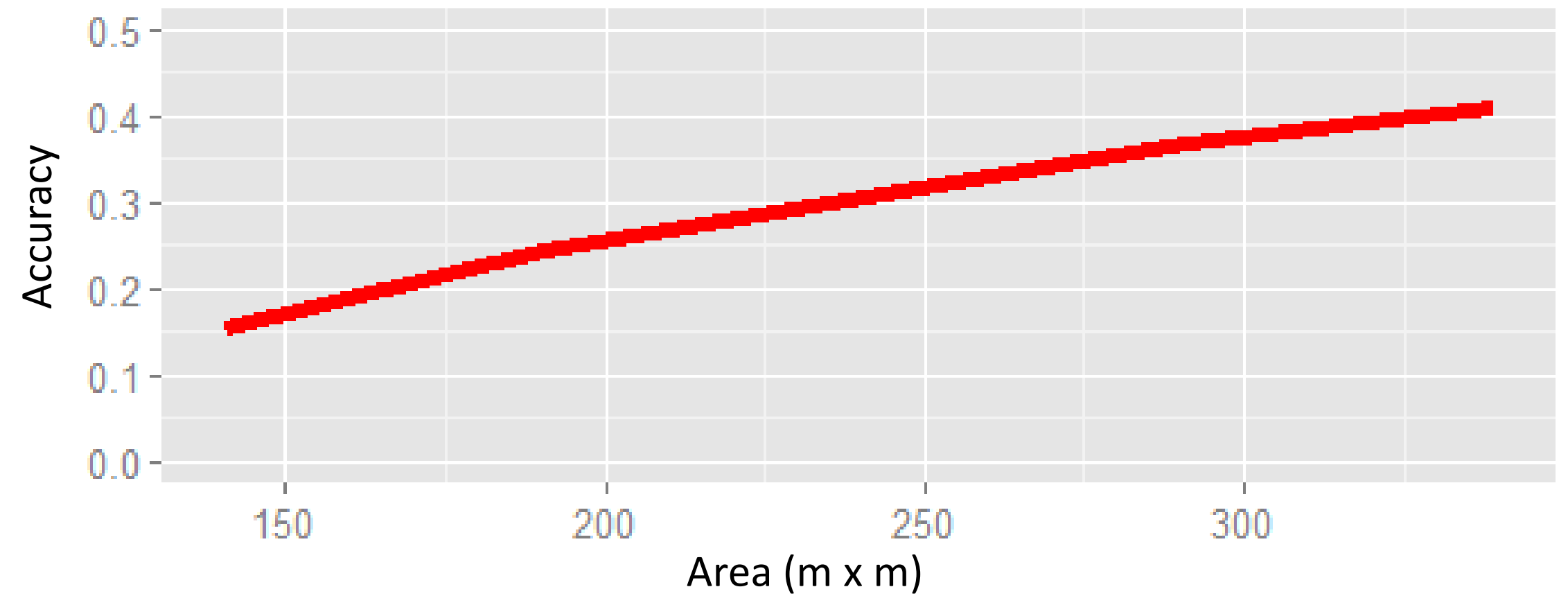
- Do a sliding comparison of a 5 min test segment against 2 – 8 min segments of all labeled traces
- Compute error using dynamic time warping of the throughput

2. Cluster

- DBSCAN to cluster the most similar points
- Predicted areas are circles encompassing each cluster
- Iterate until summed area of circles reaches a specified maximum

Preliminary results

- True location falls within predicted areas totaling 0.2 km x 0.2 km 16% of the time



Challenges

- Actual cell tower locations are unknown
- Unclear model of network traffic and variance between phones
- Travel speed is variable

Conclusions

- First demonstration of an attack of this nature
- Model may be improved using an HMM
- Can this attack be performed with sparse throughput information?
- How much must performance be degraded to defend against this?

Acknowledgements: This work was supported in part by NSF award CNS-0905349.

References: Soroush, H., Sung, K., Learned-Miller, E., Levine, B. N., & Liberatore, M. (2013). Disabling GPS is Not Enough: Cellular location leaks over the Internet. In *Privacy Enhancing Technologies* (pp. 103-122). Springer Berlin Heidelberg.

UMassAmherst Center for Forensics

forensics.cs.umass.edu
ksung@cs.umass.edu