Acknowledgement

• **Co-PI:** Cynthia Chen, UW

• **Students**
  – Graduate Students: Jingxing Wang, Feilong Wang, Nazib Siddique, Yiran Zhang
  – Undergraduate Students: Tim Adamson, Zack Ammer, Michelle Yeung

• **Sponsors/Collaborators:** FHWA, WSDOT, PSRC
Transportation big data

• “Emerging” data sources
  – Mobile phone data (cellular tower data)
  – Probe vehicle data (mobile sensing)
  – App-based data
  – Social media data
  – Connected/automated vehicles (CAV) data

• Traditional data sources
  – Loops, cameras, toll tags, travel surveys, ...
Promises: Large-scale Mobility Pattern

Beijing congestion map, taxi GPS data (Zheng et al., 2011)

Derived home locations vs. census population, app-based data (Wang et al., 2018)

Temporal travel pattern, app-based data (Wang et al., 2018)

Human mobility pattern, cellular data (Gonzalez et al., 2008)

Freight emissions by different policies, truck GPS data (Yang et al., 2014)
Promises: Fine-Grained Traffic Modeling

Real time queue length estimation of signalized intersections (Hao et al., 2013, 2015)

Trajectory-based traffic estimation and signal control (Li and Ban, 2018)
Issues

• Cellular data: **location accuracy** (a few hundred meters or more)
• GPS: Urban canyon (e.g., NYC)
• **Representativeness**
  – Data generated for *primary* uses (monitoring, navigation, searching, …)
  – Use of the data in transportation applications is a *by-product*
  – Data collected from a specific group of population; in some cases users select to (or not to) contribute to the data – *Selection Bias*
  – We (data users) have no idea of the details…
App-Based Data

462,401 users in the Puget Sound region (about 13% population), April 04 – June 05, 2017

App use activities (about 180+ apps): navigation, shopping, games, TNC, etc.
Analysis Framework

• Data properties revealed at travel activity locations
  – Zeroth-order properties (overall data properties)
  – First-order properties (single location)
  – Second-order properties (two locations/trips, e.g., OD)
  – Higher-order (multiple locations: deliveries, buses, taxis)

• Compare properties from big data with benchmarks (flow data, survey, common knowledge)
<table>
<thead>
<tr>
<th>Order</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Order</td>
<td>Study period</td>
</tr>
<tr>
<td></td>
<td>Number of unique IDs in study period</td>
</tr>
<tr>
<td></td>
<td>Sampling frequency</td>
</tr>
<tr>
<td></td>
<td>Pattern of observations</td>
</tr>
<tr>
<td></td>
<td>Number of day observed</td>
</tr>
<tr>
<td></td>
<td>Lifespan of unique IDs</td>
</tr>
<tr>
<td></td>
<td>Spatial distribution of the observations</td>
</tr>
<tr>
<td>1st Order</td>
<td>Location accuracy (mobile phone data only)</td>
</tr>
<tr>
<td></td>
<td>Location uncertainty (mobile phone data only)</td>
</tr>
<tr>
<td></td>
<td>Oscillation ratio (mobile phone data only)</td>
</tr>
<tr>
<td></td>
<td>Activity Duration</td>
</tr>
<tr>
<td></td>
<td>Spatial distribution of zone-level trips origins (generations) from a zone</td>
</tr>
<tr>
<td></td>
<td>Correlation of zone-level trip origins (generations) with the population/MPO results</td>
</tr>
<tr>
<td></td>
<td>Spatial distribution of zone-level trips destinations (attractions) from a zone</td>
</tr>
<tr>
<td>2nd Order</td>
<td>Distribution of trip rates</td>
</tr>
<tr>
<td></td>
<td>Distribution of departure and arrival times</td>
</tr>
<tr>
<td></td>
<td>Distribution of trip durations</td>
</tr>
<tr>
<td></td>
<td>Percentage of zero cells</td>
</tr>
<tr>
<td></td>
<td>Spatial distribution of cells with no trips 1) generated/ 2) attracted</td>
</tr>
<tr>
<td></td>
<td>OD collinearity</td>
</tr>
<tr>
<td></td>
<td>OD Compared with MPO OD</td>
</tr>
<tr>
<td></td>
<td>OD Sensitivity Analysis</td>
</tr>
</tbody>
</table>
- App-based data show two-peak pattern on weekdays, one peak on weekends, which is similar to GPS data. The mobile phone data do not show such pattern.
• Time interval between two consecutive observations

App-based data

GPS-based probe vehicle data
- **Location accuracy**

**Cellular Data**
- A few hundred to a few thousand meters

**App-Based Data**
- More than 80% of observations with an accuracy of 60 meters or less
- Nearly 90% of observations with an accuracy of 160 meters or less

**GPS Data**
- A few meters
No. of days observed (inter-day temporal sparsity)

- ID consistency is comparable with Buffalo mobile phone data, much better than Seattle GPS data.
Spatial distribution of inferred trip ends
App-based data

Trip ends cluster in major downtown areas: Seattle, Bellevue, and Tacoma, etc.
Inferred Home (a) vs. Census Population (b)

- Corr. Coeff. of **0.91** at the census tract level, compared with **0.43** obtained from the mobile phone data.
- The inferred residents accounts for about **3%** of the PSR population.
- Patterns from the app-based data is similar to that from the survey data
- Compared with the survey data, the long stays are underestimated
• Trip rate

App-based data

Buffalo mobile phone data

- **Missing trips** is common for big data
- An under-estimation bias of **trip rates** exists for three big datasets (app-based, mobile phone, and vehicle GPS data).
• Estimated # of trips originated from a TAZ

**App-based data**

Estimated trips originated vs. PSRC OD trip originated

- Correlation Coefficient: 0.505

**Buffalo mobile phone data**

Mobile phone trips vs. MPO trip originated

- Correlation Coefficient: 0.690
• OD travel demands estimated from app-based data

App-based data

Buffalo mobile phone data

Corr. Coeff: 364

Corr. Coeff: 0.665
• **Travel times**

App-based data, mobile phone data, PSRC survey, and Buffalo survey

- The curves of two surveys grow faster than the curves of two big datasets
Summary of Major Findings

- Big data can reveal useful travel patterns; discrepancies/biases exist
- Factors that may affect mobility analysis
  - Study area & duration of data collection
  - Updating frequency of device IDs
  - Location accuracy and uncertainty
  - The underlying data generation process: knowing how data are generated is important, e.g. mobile phone data are generated mainly for phone activities, may or may not be travel related
  - Data processing: different datasets require different techniques / algorithms to process data, which should be carefully determined
Implications

• Big data is very useful for transportation applications
• “Data → Solutions/Decisions” may not be enough
• We need to understand data first since representative data is crucial to transportation applications
  – Certain communities / population (low income, elderly, minority) are often less represented
  – Misled policies / investment
Implications

• Data bias modeling and correction
  – For the datasets that are available, how to model and address bias issues is critical
• One approach: *data-driven, model-based bias* modeling and estimation of *networked* transportation data (NSF #1825053, collaborator: Dr. Y.Y. Fan at UC-Davis)
## Comparison Across Four Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>App-based data</strong></td>
<td>Large size; higher observational frequency; mixed GPS and cellular tower positioning; presence of trace information; inexpensive; continuous</td>
<td>Non-probabilistic samples; missing trips/activities; no demographics information; unknown underlying population</td>
</tr>
<tr>
<td><strong>Mobile phone data</strong></td>
<td>Large size; presence of trace information; inexpensive; continuous</td>
<td>Lower observational frequency; lower positioning accuracy; Non-probabilistic samples; missing trips/activities; no demographics information; unknown underlying population</td>
</tr>
<tr>
<td><strong>Vehicular GPS data</strong></td>
<td>Large size; higher observational frequency; high positioning accuracy; presence of trace information; continuous</td>
<td>Non-probabilistic samples; missing trips/activities; no demographics information; unknown underlying population; only for vehicular travels</td>
</tr>
<tr>
<td><strong>Household travel survey data</strong></td>
<td>Probabilistic samples; representative; rich information on activity and travel patterns; with demographics and attitudinal information</td>
<td>Lack of trace information; Lack of information for non-residents; expensive; infrequent data collection; static information</td>
</tr>
</tbody>
</table>
Data Fusion

- Emerging data vary from their underlying population, generation process to their prosperities and issues. Data fusion can help correct the issues.
- Other traditional data sources (e.g. loop detector data, travel surveys) are also available. Data fusion is needed to integrate emerging and traditional data sources to provide data with better quality.

Big-ness

Representative-ness

GPS-based Survey

Now / Near – Future

Cellular

App-based

CAV/V2X

Previous / Now

Future

CAV/V2V only
Thank You!

- Contact: banx@uw.edu; qzchen@uw.edu
- iUTS Lab Website: faculty.washington.edu/banx
Backup Slides
• Departure times

App-based data

- Both travel survey data show two peaks (reflecting morning and afternoon commuting peaks)
- App-based data shows three peaks, mobile phone data shows afternoon peak, and GPS data only captures morning peak
More Challenges: Data Sharing

• Data is big but fragmented: owned by different parties who do not always want to share data (privacy/security, corporate secrets, etc.)

• One possible solution: UW Transportation Data Collaborative: https://www.uwtdc.org/ (PIs: Drs. Jan Whittington, Bill Howe, Mark Hallenbeck)
More Challenges: Data Evolution

• Big data is dynamic: the underlying population may change over time, the data generation process may change over time, ...

• Their properties may need to be investigated periodically
May 9th data shift

(a) Daily number of unique IDs
(b) No. of observations per ID

Guess what happened on May 9th?
May 9th data shift

- Before May 9th
- After May 9th
- PSRC Survey

Activity Duration (hours)

Percentage of user-day (%)

Before May 9th  Mean: 3.11
After May 9th   Mean: 3.37
Discuss May 9th data shift via down-sampling

Randomly remove certain amount of records/rows.
Discuss May 9th data shift via down-sampling

Effects on trip rate:

- Before downsampling (mean: 3.2)
- After downsampling (30% removed) (mean: 3.0)
- After downsampling (70% removed) (mean: 2.3)