Agenda

• Motivation
• How We Set Our Target Outcomes
• Data Sources
• Key Findings
• What's Next
Motivation
Four Key Reasons for Us to Start Estimating Traffic Counts from Our Big Data Resources

1. The Status Quo is Deficient
2. They’re Industry Standard
3. Whole Industry would Benefit
4. It’s Never Been Done Well
In addition, they’re a precursor to VMT

Segment Length

Average Trip VMT

Average Trip Distance + Distribution are Longtime StreetLight InSight Metrics
Add Zone

Draw New Zone: Grove Av

Zone Name: Grove Av

Zone ID: (optional)

Is Pass-Through Zone

Indicate whether the Zone is Pass-Through or not. Pass-Through should be Yes if the Zone is a road segment (so that trips that travel along the road are analyzed). Pass-Through should be No if the Zone is an ‘area’ such as a TAZ or County polygon (so that trips that start or stop in the zone are analyzed).

View Zone List  Save  Cancel
How We Set Our Target Outcomes
Our methods reflect our goals

1. Works better than 48-hour expansion
2. Productized for rural, arterial roads to maximize benefits.
3. Build an algorithm that is extensible:
   - To hourly, seasonal, truck v. car volume estimates
   - To be part of full, balanced origin, destination, routing Metrics

(NB – Real Time volume is not currently a goal)
## Beat Short Term Expansion and Modeled Data

### 1 Day Expansion Error CI by AADT Bin

<table>
<thead>
<tr>
<th>AADT Range</th>
<th>-95% CI</th>
<th>+95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;100</td>
<td>-53.6%</td>
<td>+53.1%</td>
</tr>
<tr>
<td>100-1,000</td>
<td>-28.3%</td>
<td>+34.5%</td>
</tr>
<tr>
<td>1,000-10,000</td>
<td>-16.7%</td>
<td>+20.2%</td>
</tr>
<tr>
<td>10,000+</td>
<td>-13.44%</td>
<td>+13.7%</td>
</tr>
</tbody>
</table>


It’s unclear why they didn’t publish the chart for 2day but we’ll assume it’s all 10% better.

Data Sources
What Data are We Working With?

LBS Data – 23% Coverage of Adult Personal Devices

GPS Data – 10% Coverage of Commercial Truck Trips
We Use Multiple Data Sources for AADT Metrics

1. **Big Data Input #1**
   Full Year of Navigation-GPS Data

2. **Big Data Input #2**
   Full Year of Location-Based Services Data

3. **Contextual Data Set**
   Census, Road Classification, Etc

4. **Calibration Data Set**
   Hundreds of Validated Continuous Count Locations
# Data Calibration (training set)

<table>
<thead>
<tr>
<th>Road Class</th>
<th>Reported AADT</th>
<th># Zones (permanent counters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50,000 +</td>
<td>660</td>
</tr>
<tr>
<td>B</td>
<td>25,000 - 50,000</td>
<td>370</td>
</tr>
<tr>
<td>C</td>
<td>10,000 - 25,000</td>
<td>500</td>
</tr>
<tr>
<td>D</td>
<td>5000 - 10,000</td>
<td>295</td>
</tr>
<tr>
<td>E</td>
<td>2500 - 5000</td>
<td>200</td>
</tr>
<tr>
<td>F</td>
<td>0 - 2500</td>
<td>250</td>
</tr>
</tbody>
</table>

We source our training data from permanent counters located in states across the United States:

- California (350)
- Illinois (115)
- Iowa (140)
- Georgia (180)
- Arizona (600)
- Utah (100)
- Washington (170)
- Florida (400)
- Minnesota (84)
- Montana (98)
- Ohio (140)
Key Findings
Four Key Findings

1. **We CAN hit our goal** of matching or out-performing short term expansion— and dozens of clients are taking advantage of that. (But we want to do better!)

2. Validation is tricky because **good comparison data is extremely hard to find**. Setting up a good validation set for a client-eval, especially for low volume roads, is far more time consuming than the data science.

3. **Neither GPS or LBS alone is adequate** – you need both! (Plus, Census data should be used to de-skew LBS data for local demographic bias)

4. Any equation with non-linearity can lead to problems. **Luckily, the linear algos performed as well as anything else.**
## Hitting Targets – Outcome of Example Validation Study in One State

<table>
<thead>
<tr>
<th>Min</th>
<th>Max</th>
<th>Bin Count</th>
<th>Internal R2</th>
<th>Average Client AADT in Bin</th>
<th>mean bias</th>
<th>median bias</th>
<th>Low 95% CI</th>
<th>High 95% CI</th>
<th>Target Low 95% CI</th>
<th>Target High 95% CI</th>
<th>RMSE</th>
<th>RMSE as percent Avg</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>50,000</td>
<td>1,000,000</td>
<td>68</td>
<td>0.75</td>
<td>91,985</td>
<td>-1.436%</td>
<td>1.01%</td>
<td>-4%</td>
<td>5.7%</td>
<td>-12.1%</td>
<td>12.3%</td>
<td>24,387</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>25,000</td>
<td>49,999</td>
<td>77</td>
<td>0.45</td>
<td>35,266</td>
<td>5.369%</td>
<td>2.49%</td>
<td>-3%</td>
<td>7.7%</td>
<td>-12.1%</td>
<td>12.3%</td>
<td>8,414</td>
<td>24%</td>
<td>22%</td>
</tr>
<tr>
<td>10,000</td>
<td>24,999</td>
<td>82</td>
<td>0.50</td>
<td>17,071</td>
<td>9.134%</td>
<td>6.93%</td>
<td>0%</td>
<td>13.8%</td>
<td>-12.1%</td>
<td>12.3%</td>
<td>4,964</td>
<td>29%</td>
<td>27%</td>
</tr>
<tr>
<td>2,500</td>
<td>9,999</td>
<td>58</td>
<td>0.33</td>
<td>5,907</td>
<td>-7.594%</td>
<td>-7.05%</td>
<td>-20%</td>
<td>5.6%</td>
<td>-15.0%</td>
<td>18.2%</td>
<td>2,063</td>
<td>35%</td>
<td>37%</td>
</tr>
<tr>
<td>0</td>
<td>2,499</td>
<td>22</td>
<td>0.23</td>
<td>1,709</td>
<td>2.753%</td>
<td>-0.08%</td>
<td>-15%</td>
<td>14.9%</td>
<td>-25.5%</td>
<td>31.1%</td>
<td>644</td>
<td>38%</td>
<td>65%</td>
</tr>
</tbody>
</table>

### Results

- **Sum**: 307
- **Overall R2**: 90%

<table>
<thead>
<tr>
<th>Weighted Average</th>
<th>Average</th>
<th>Weighted Avg RMSE as PCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>35,018</td>
<td>2%</td>
<td>7.4%</td>
</tr>
</tbody>
</table>

- Results better than target
- Results slightly worse than target (w/in 5p)
- Results worse than target
Training and Validation Data Needs Cleaning

1\textsuperscript{ST} CUT - STREETLIGHT AADT ESTIMATE VS. XDOT HIGH FIDELITY LOOP COUNTERS

Public agency reported AADT of 421, but StreetLight InSight Estimated >2,000

Expected location of counter

Actual location of counter
LBS Data Performs Best of “Alone Data” But Mixing in GPS-Truck + Context Makes It Better
## Multivariate Regression Performance is Competitive

<table>
<thead>
<tr>
<th>Model</th>
<th>Features (Current Best Model)</th>
<th>First Pass – Performance on Road Classes RMSE as percent avg</th>
</tr>
</thead>
</table>
| Multivariate Linear Regression | Scaled LBS trips  
Road type (motorway)  
Commercial trips (GPS)  
Personal Trips (GPS) | A 22%  
B 20% |
| Lasso Regression       | Scaled LBS Trips  
Weekend vs Weekday  
Peak traffic hour ratio  
Roadtype (OSM)  
# Commercial, Personal Trips (GPS)  
Census population data | A 19%  
B 21% |
| Random Forest Regression | | A 15%  
B 28% |
| Gradient Boosting Regression | | A 17%  
B 26% |
Extensible, Balanced Networks

• Example – a different algorithm for small and large roads would leave this intersection unbalanced.

• Example - non-linear algorithms can lead to the “bottom up” sum of all trips in a county ≠ the “top down” estimate for total county trips
What’s Next?
What’s Next?

• 2017 AADT
• Hourly, seasonal volume estimates
• Commercial truck/car split
• Full flow volume estimation improvements
• Expanded validation data and studies (especially in rural areas)

Questions?
www.streetlightdata.com for more information