Big Data Applications for Travel Demand Modeling

Tennessee Model Users Group

Ted Reinhold, Senior Solutions Engineer

Julie Hirni, Director, Public Sales Sector - Southeast

November 17, 2021

What is “Big Data?”

Big Data for transportation

1

Mobile Device LBS

Anonymized “opt-in” data from mobile device apps

2

Vehicle GPS

GPS devices installed in personal and commercial vehicles

3

Internet of Things

Connected vehicles, smart fare cards, RFID, geo-fencing, and more

Traditional Data Sources

- Household & Intercept Surveys
- Bluetooth & Other Sensors
- Aerial Photos & Videos
- Assumption-Based Modeled Data

Location-Based Services (LBS) and GPS data

MOBILE DEVICE DATA
from 110M+ devices of U.S. and Canadian adults

Example, San Bernardino, CA
Oct 8, 2017 24-hr snapshot

CONTEXT

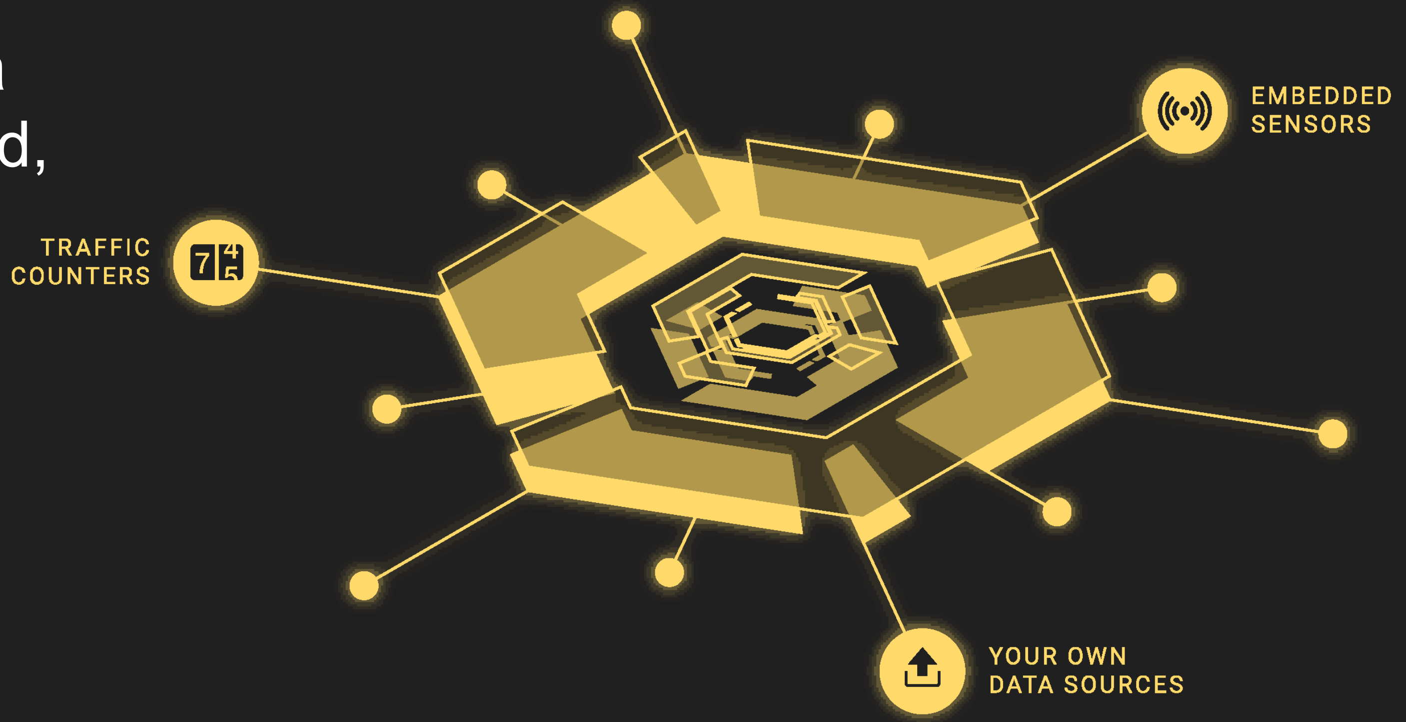
Parcel Data
Digital Road Network Data
U.S. Census

- Every month, we process over 100 billion anonymized location records from **smart phones** and **GPS navigation devices** in cars and trucks.

- **Route Science®** transforms them into **contextualized**, normalized and aggregated travel patterns.

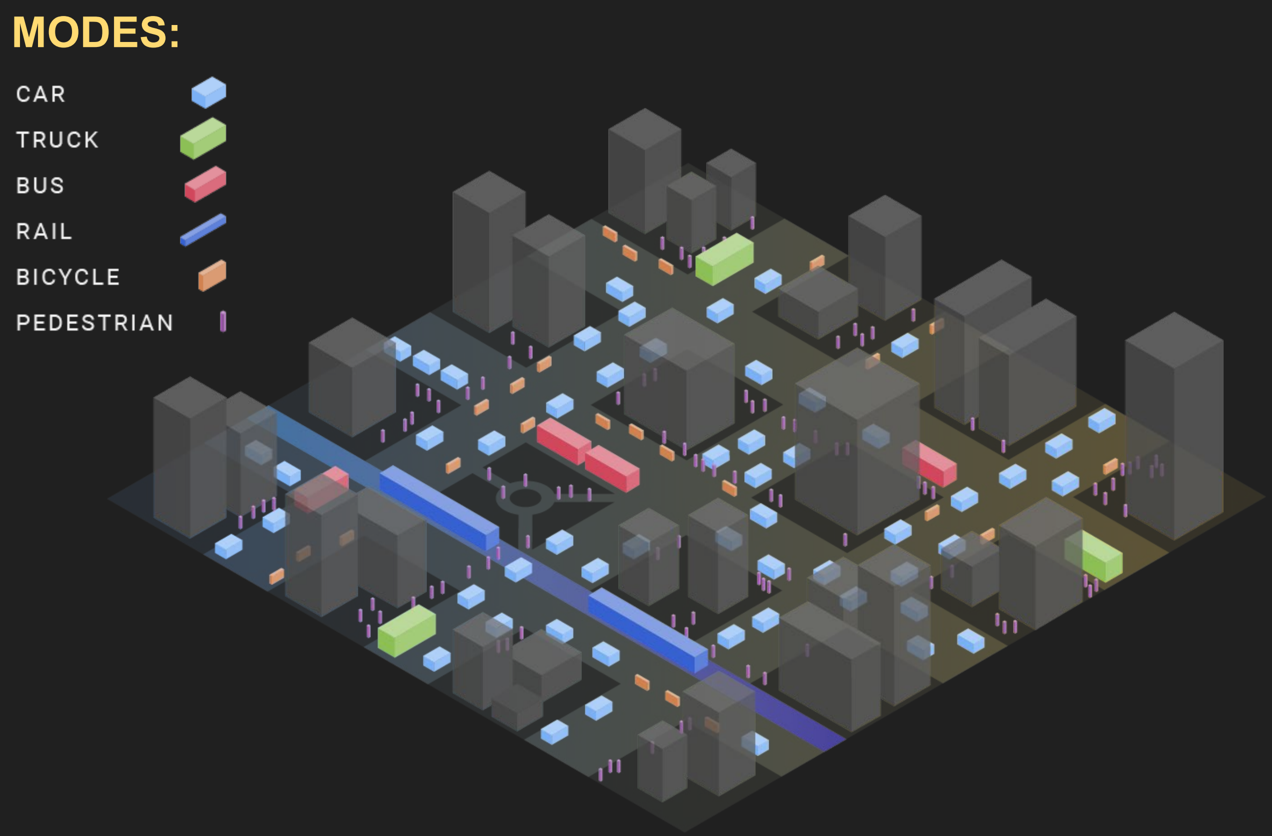
Data Processing

Proprietary algorithms and machine learning turn the data into contextualized, aggregated, normalized travel patterns.



Validation

Analytics providers validate results using counters, sensors, and other sources.



Which source data is the best?
A constantly evolving mix

Data is always changing…

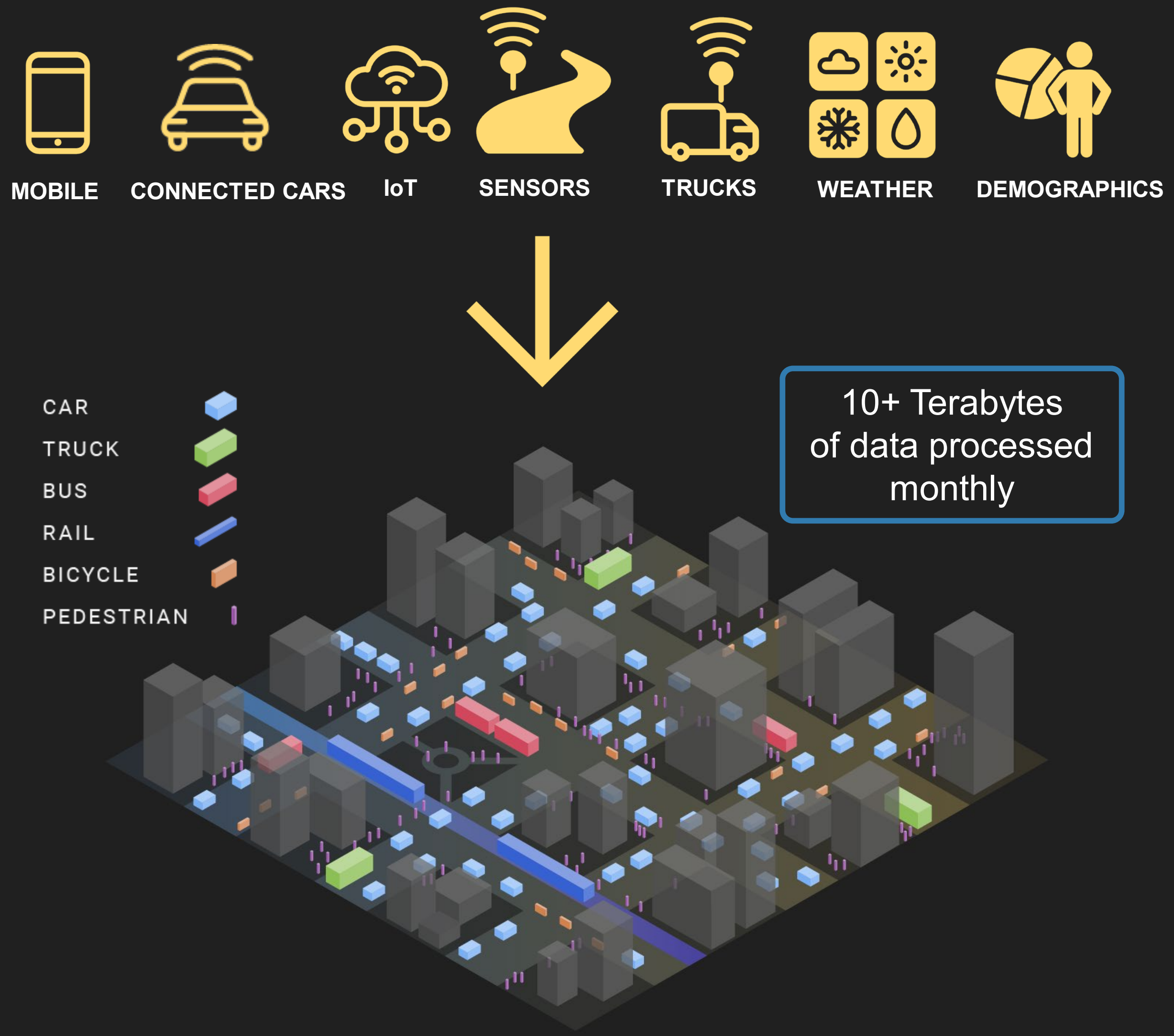
- *Historically we used cell tower data*
- *Today we use a blend of data sources →*
- *Tomorrow we'll add something new*

Framework for assessing Big Data sources

Data sample: size, representativeness, coverage (temporal/spatial), frequency of updates

Privacy protections: how does the process protect privacy of individuals?

Validation and uses: applicability to transportation-related use cases



Big Data Applications for Modeling

Benefits for Modeling

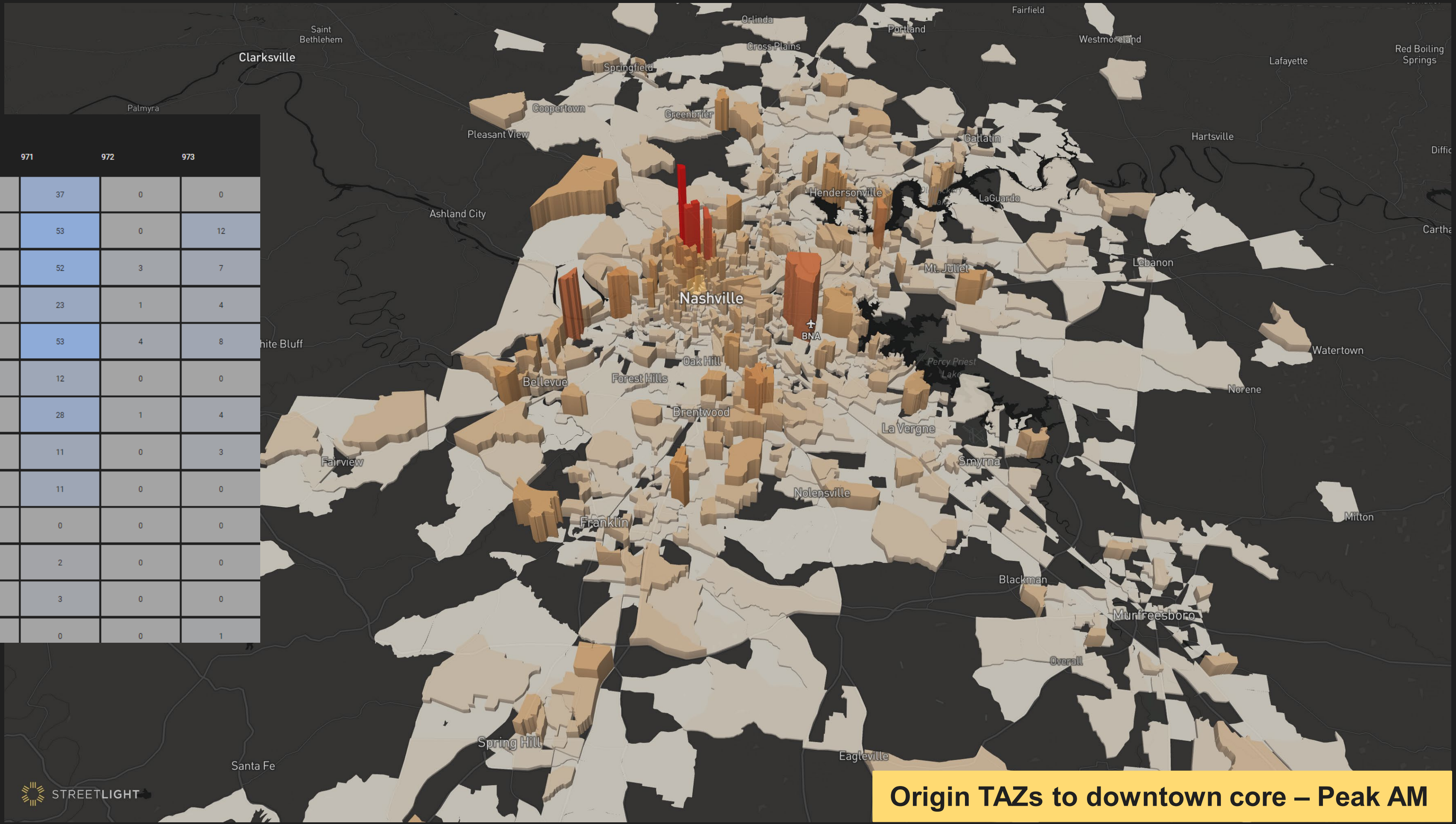
- Supplement traditional data sources
- Overcome modeling assumptions
- Continually refine and validate model

Model Development

Use Big Data O-D as a seed matrix to estimate full O-D trip table

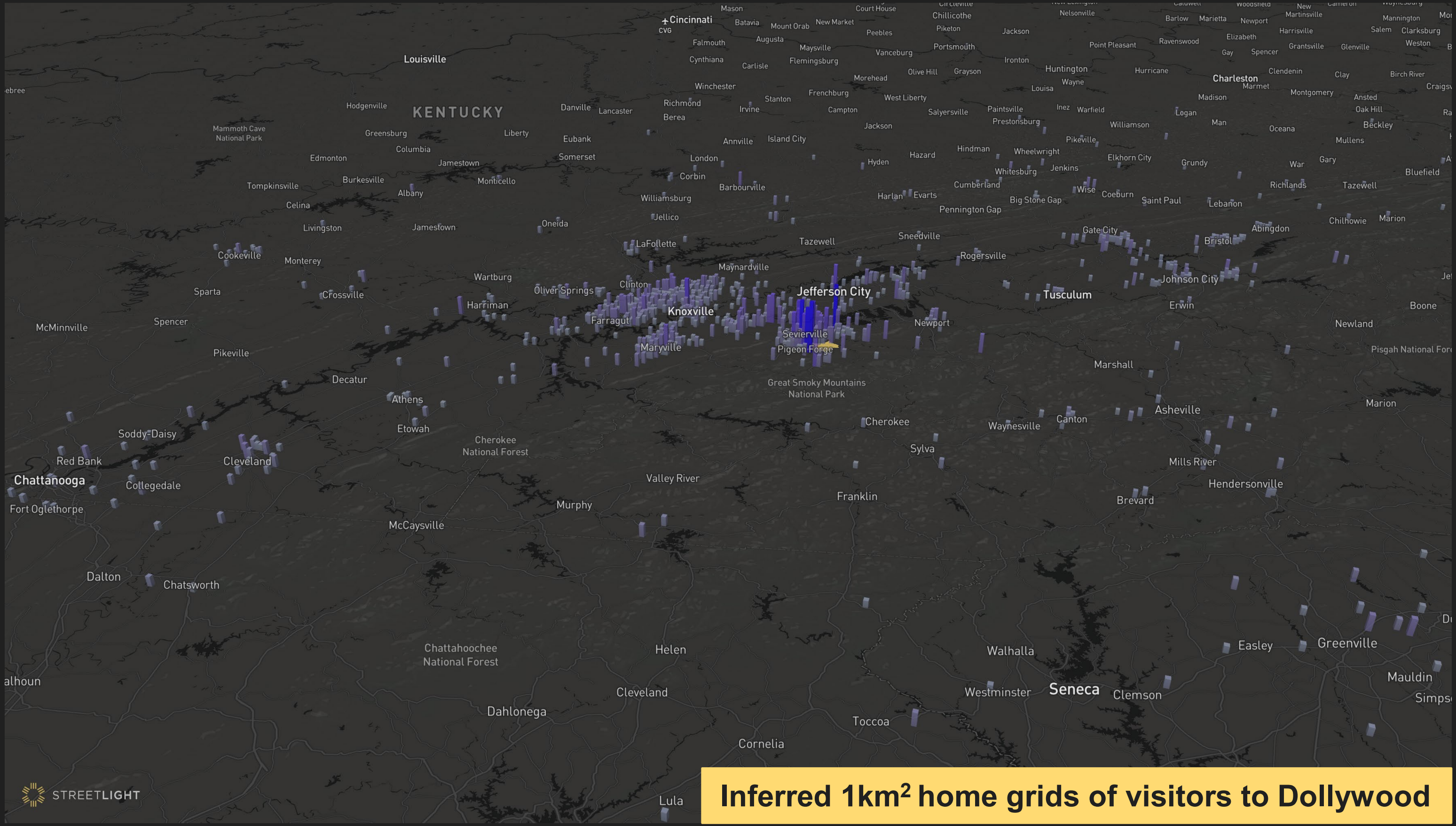
Destinations

Origins	949	950	951	957	968	969	970	971	972	973
1000	0	8	3	0	10	0	0	37	0	0
1001	0	1	7	0	36	1	0	53	0	12
1002	1	0	0	0	19	3	0	52	3	7
1003	0	2	9	0	12	0	0	23	1	4
1004	2	5	4	1	34	3	0	53	4	8
1005	1	0	1	0	3	0	1	12	0	0
1006	0	4	23	0	11	2	0	28	1	4
1007	0	0	14	0	7	0	0	11	0	3
1008	0	2	6	1	4	0	0	11	0	0
1009	0	0	0	0	1	0	0	0	0	0
1010	0	0	1	1	0	0	0	2	0	0
1011	0	3	1	10	2	0	0	3	0	0
1012	0	0	2	0	0	0	0	0	0	1



Origin TAZs to downtown core – Peak AM

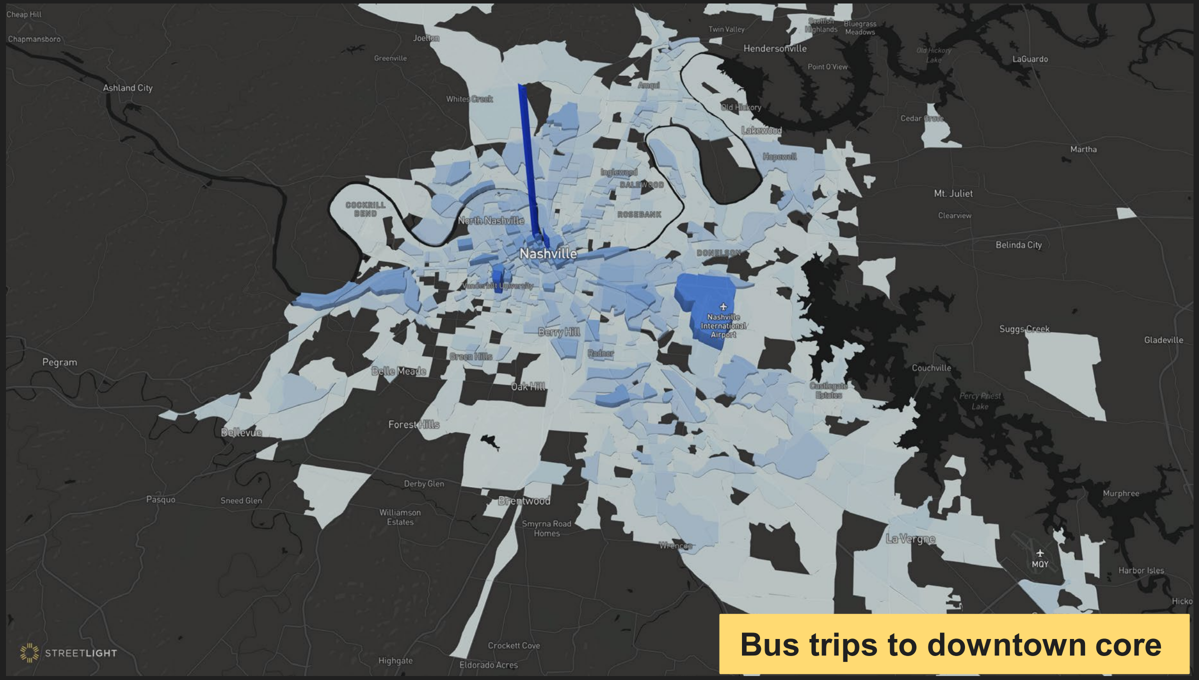
Supplement gaps in traveler surveys (e.g., tourists)



Inferred 1km^2 home grids of visitors to Dollywood

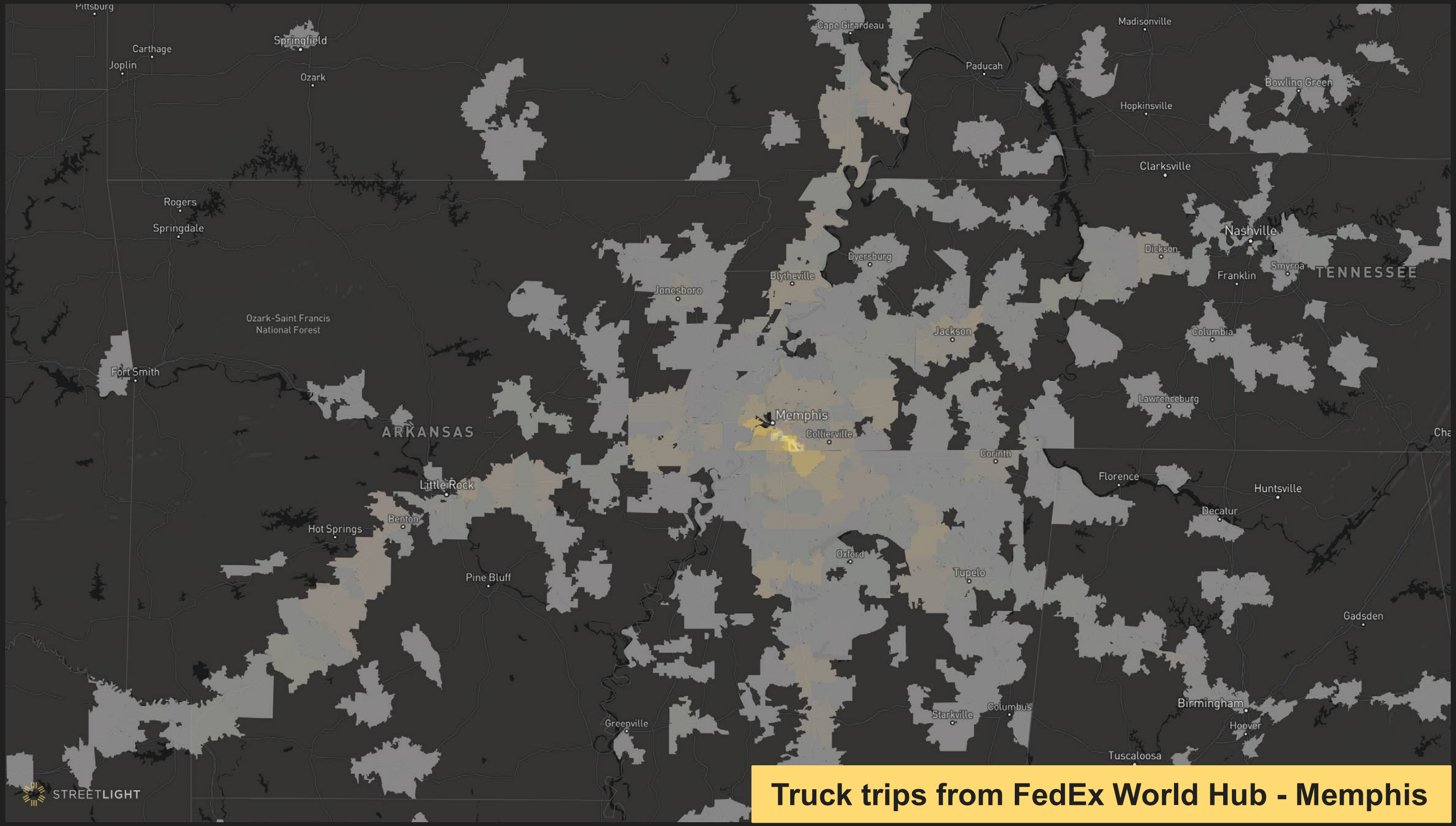
Granular metrics for non-motorized and transit modes

Pedestrian activity near downtown



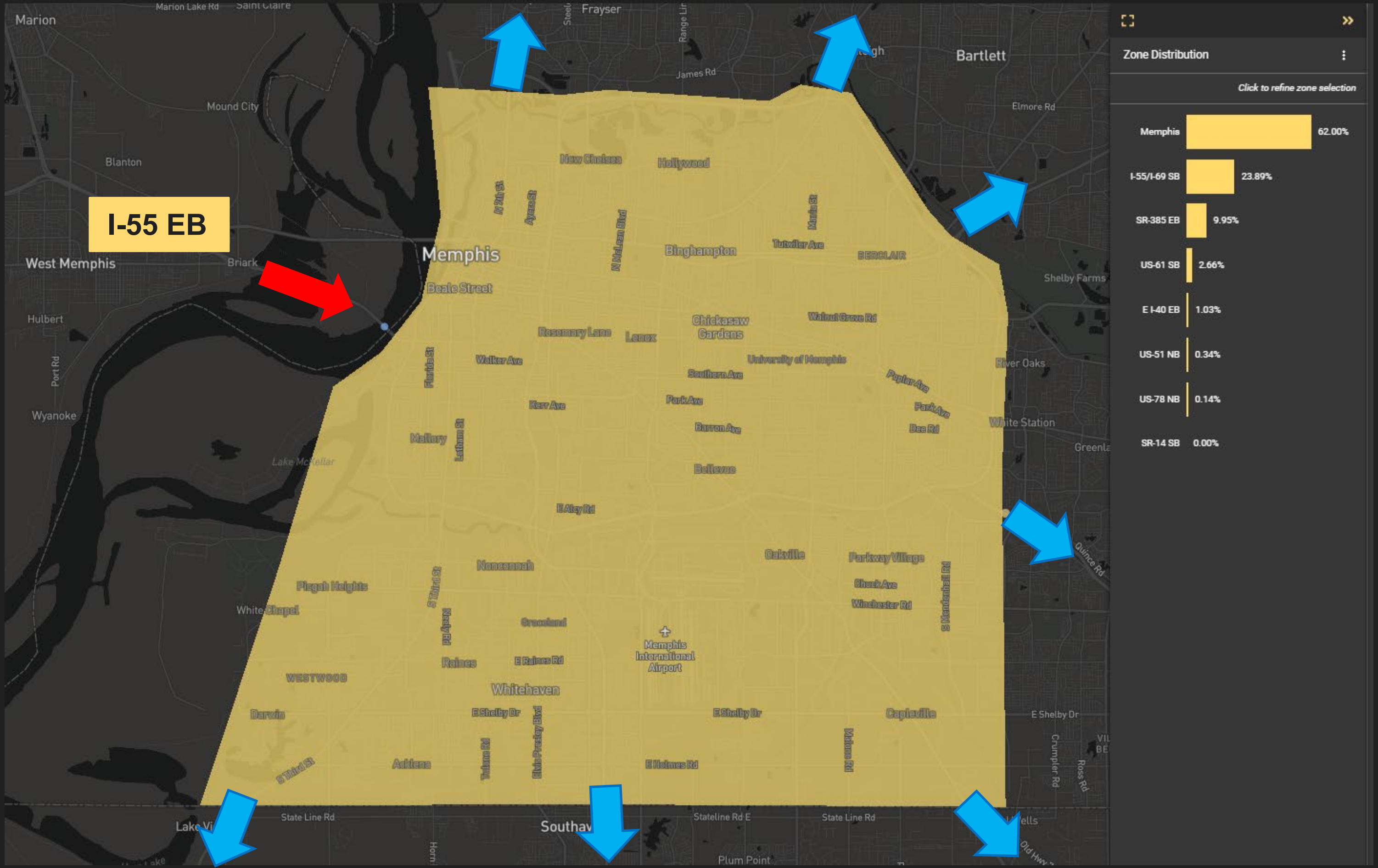
Bus trips to downtown core

Develop special modules for freight hubs, campus, airports, parks…

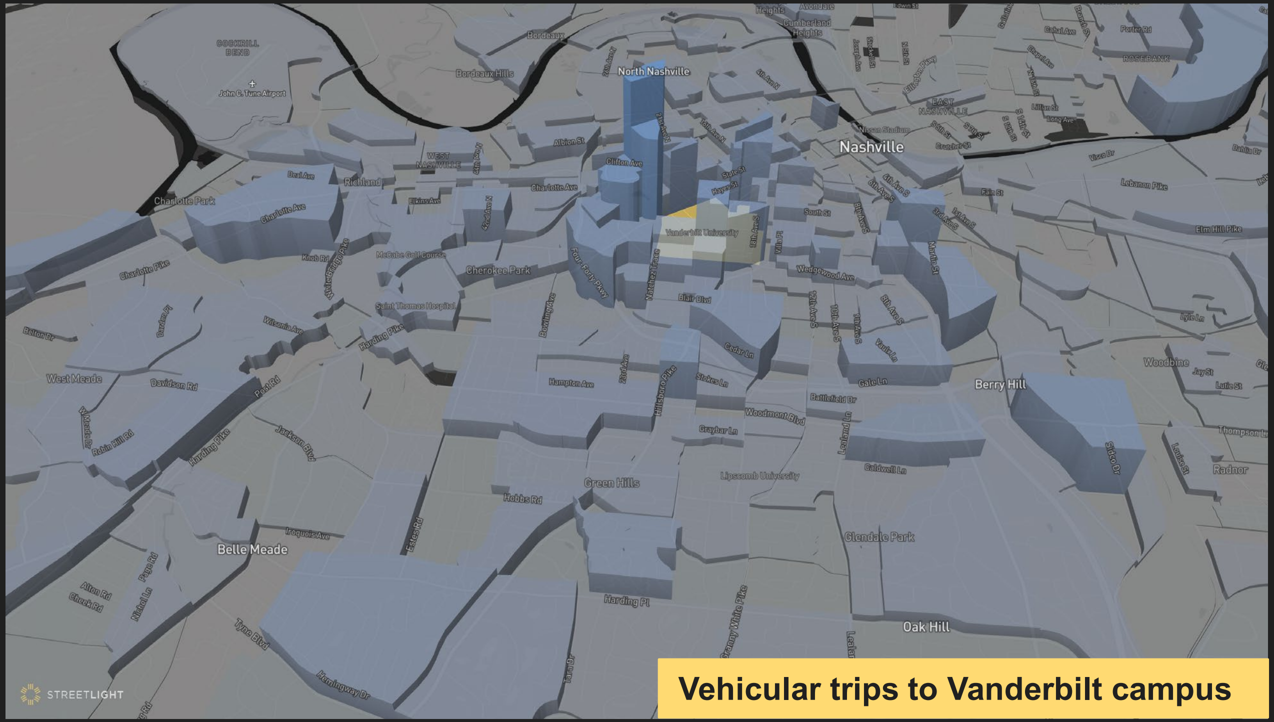


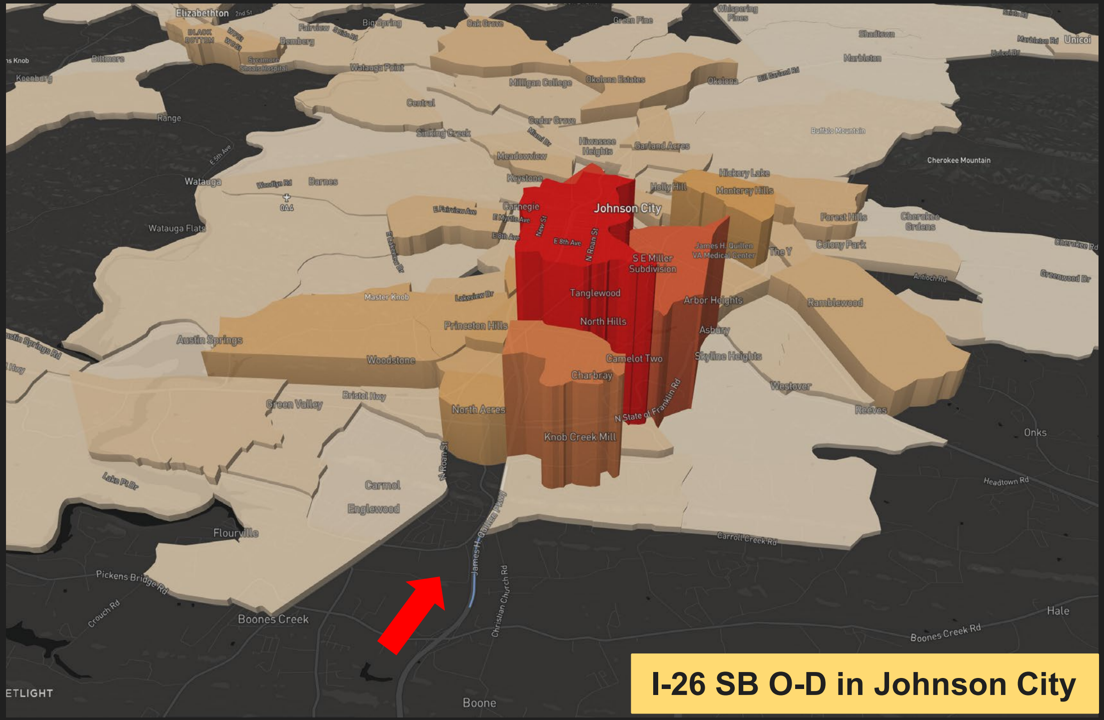
Truck trips from FedEx World Hub - Memphis

Internal-External demand (II / IE / EI / EE)

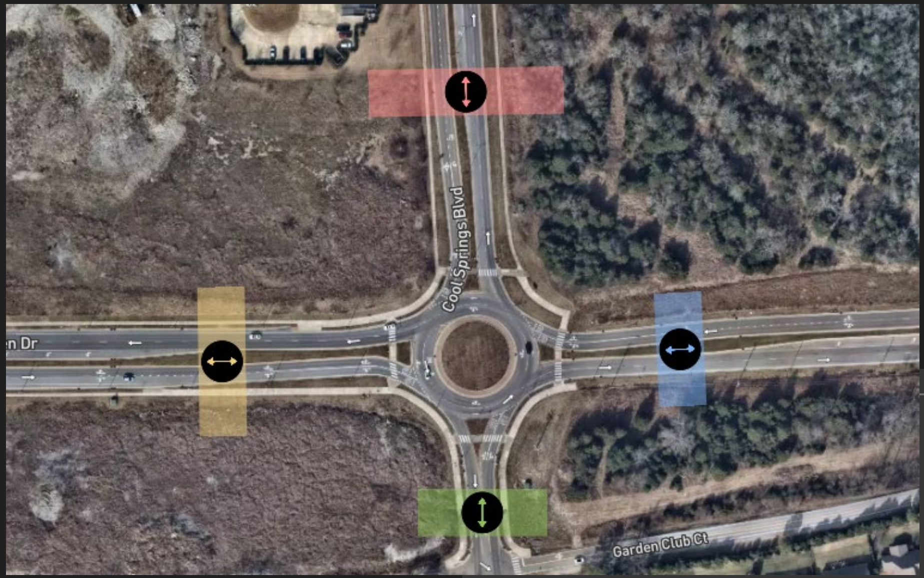


Develop sub-area models from regional travel demand model





O-D inputs to traffic simulation models, TIA, cut-through studies



McEwen Dr - Franklin

N Cool Springs Blvd

In	Total	Out
274	599	325
Right	Thru	Left
19	126	129
↙	↓	↘

W McEwen Dr

Out	317	Left	12	↗
Total	578	Thru	189	→
In	261	Right	60	↘

E McEwen Dr

↖	161	Right	440	In
←	256	Thru	790	Total
↙	23	Left	350	Out

S Cool Springs Blvd

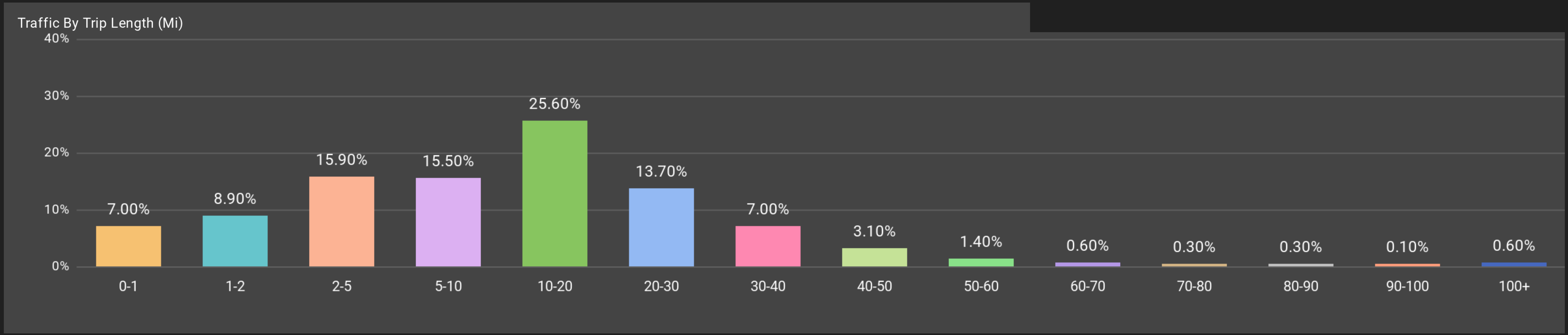
↖	↑	↗
42	152	32
Left	Thru	Right
209	435	226
Out	Total	In

Model Calibration & Validation

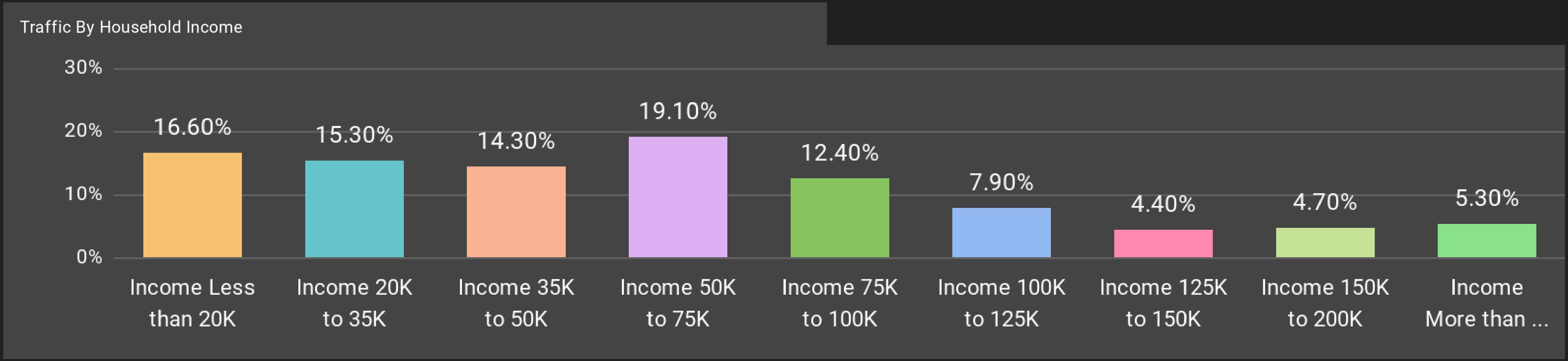
Total trip production and attraction

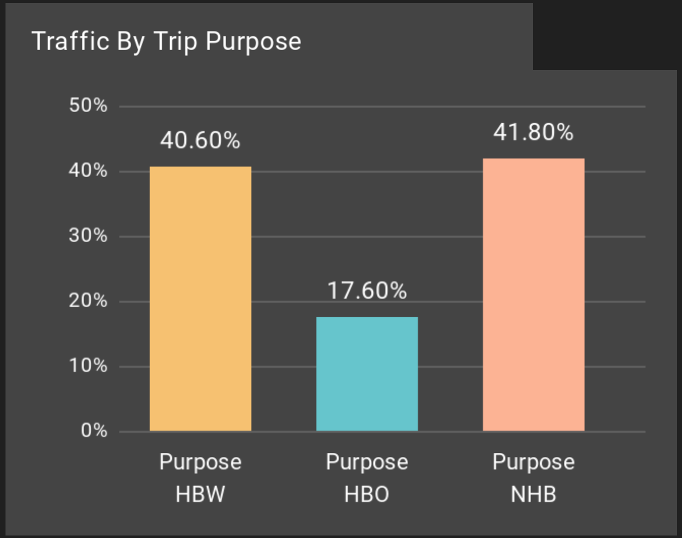
Vehicle trips destined for downtown, typical weekday AM Peak (6-10a)

Total trip production and attraction

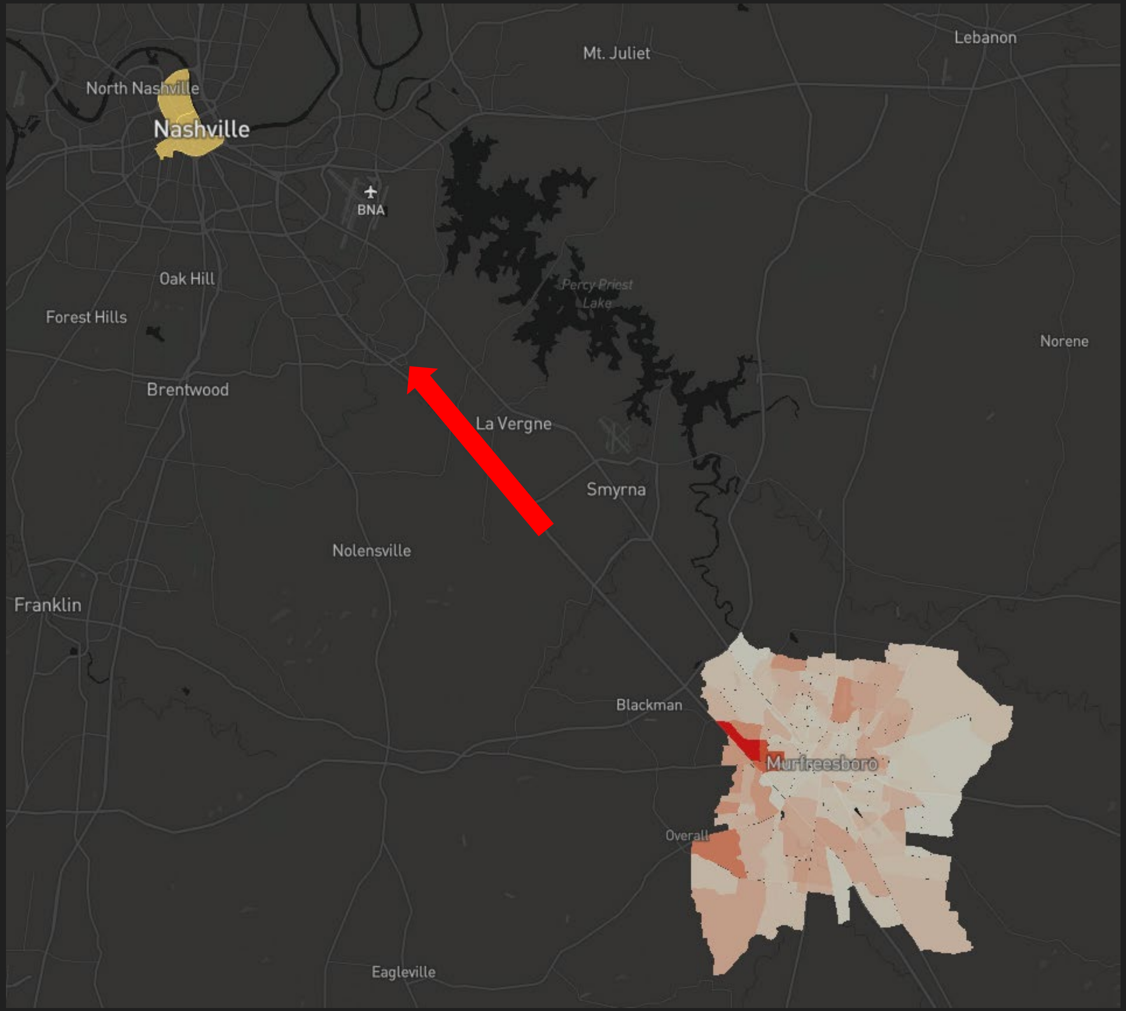


Total trip production and attraction

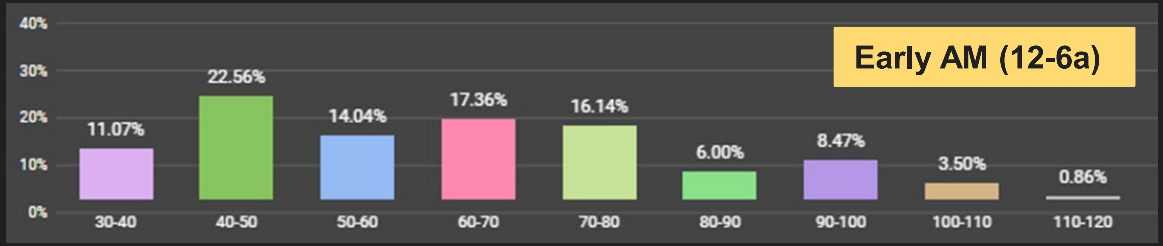


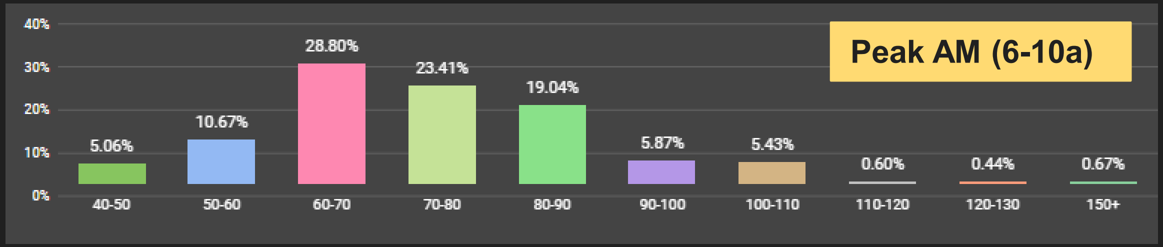


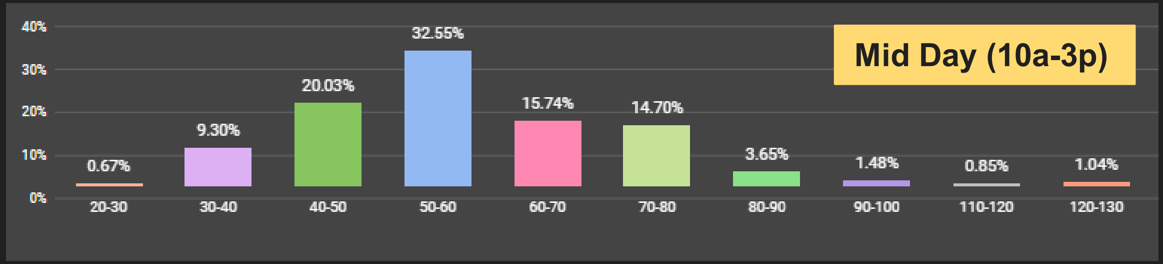
Travel time distribution by time of day

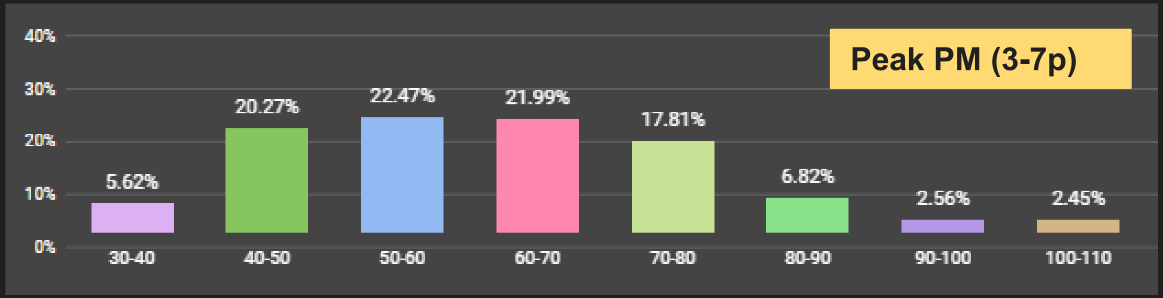


Travel time distribution by time of day





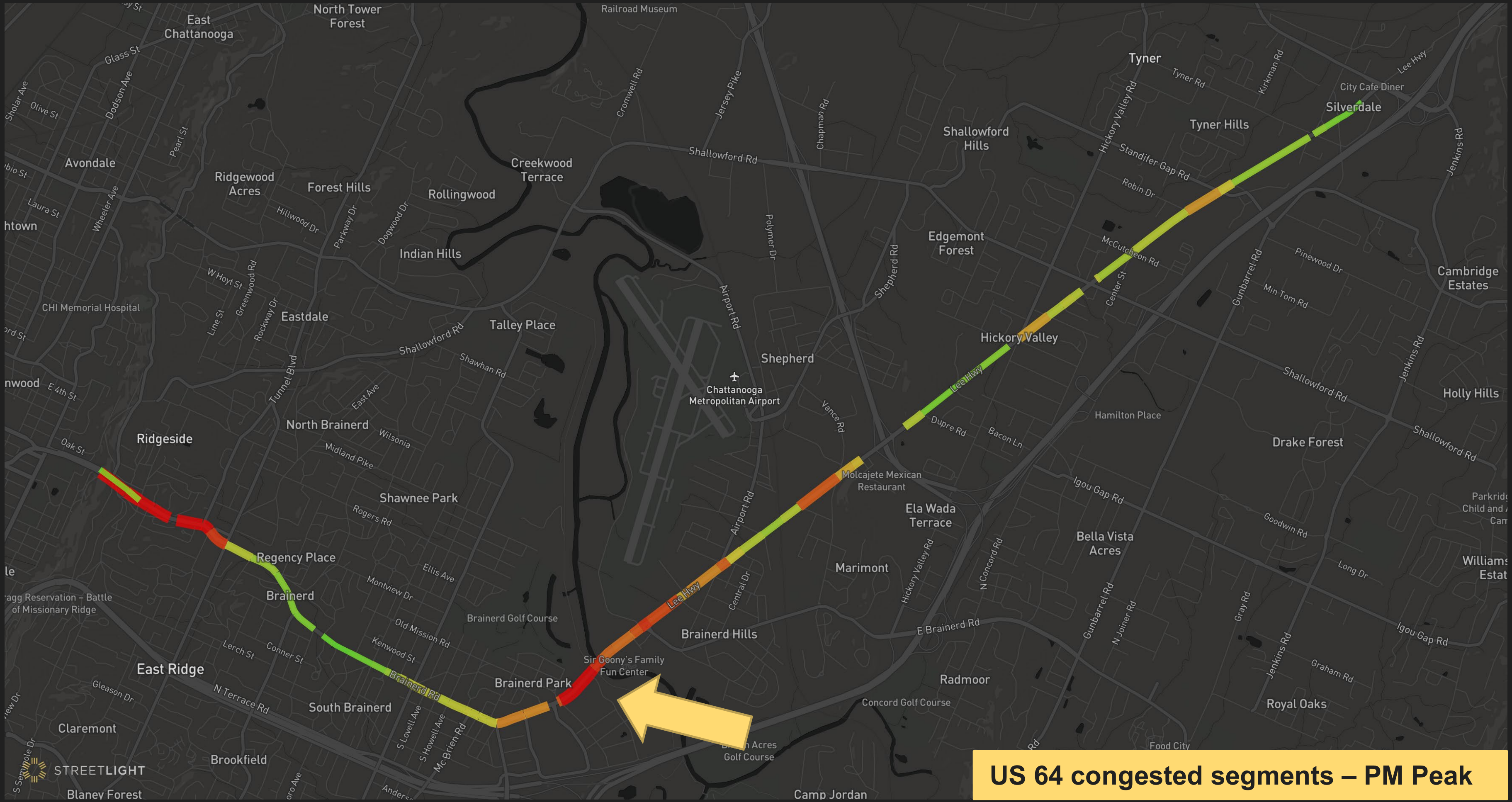




Calibrate network assignment – route choice

Airport to city center

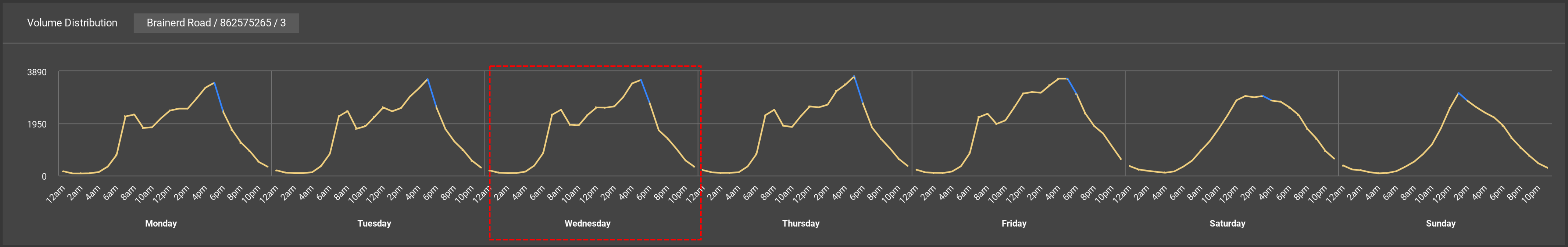
Calibrate network assignment – link delays/congestion



Calibrate network assignment – link volumes/speeds

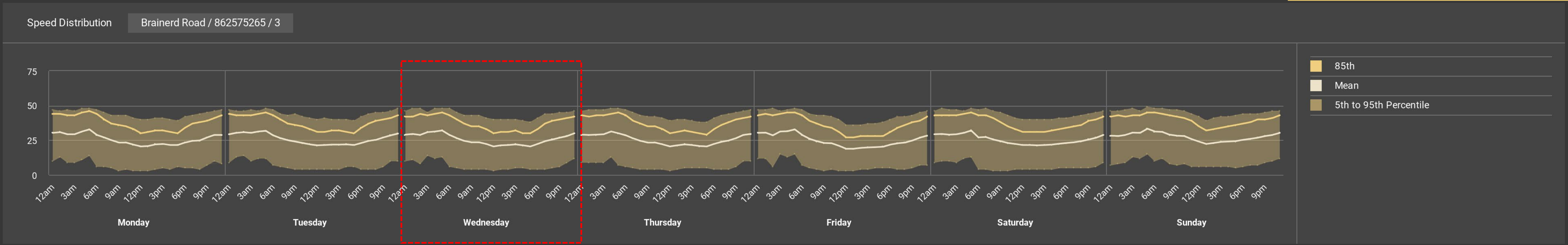
Time Distribution Metrics

US 64 WB @ S Chickamauga Creek - Volumes



Time Distribution Metrics

US 64 WB @ S Chickamauga Creek - Speeds



Calibrate model sensitivity to network or land use changes

Summary

1. Big Data can complement current sources of input, calibration and validation data for models
2. Opportunities to refine overlooked model components - I/E, freight, tourism, seasonality
3. Monthly data refresh facilitates frequent model updates and monitoring
4. Big Data enables a wide range of model applications and use cases

STREETLIGHT DATA

ted.reinhold@streetlightdata.com

julie.hirni@streetlightdata.com

Vetting Big Data Metrics Providers

A FRAMEWORK FOR QUESTIONS TO ASK AND BEST PRACTICES TO LOOK FOR

1 Data Sets and Sources

- Where does the data come from?
- How big is the sample size, and from how many providers?
- How frequently are the data sources evaluated and updated?

2 Processing Methods

- What algorithm and machine learning techniques are used?
- How granular are the metrics?
- What transportation modes are included?

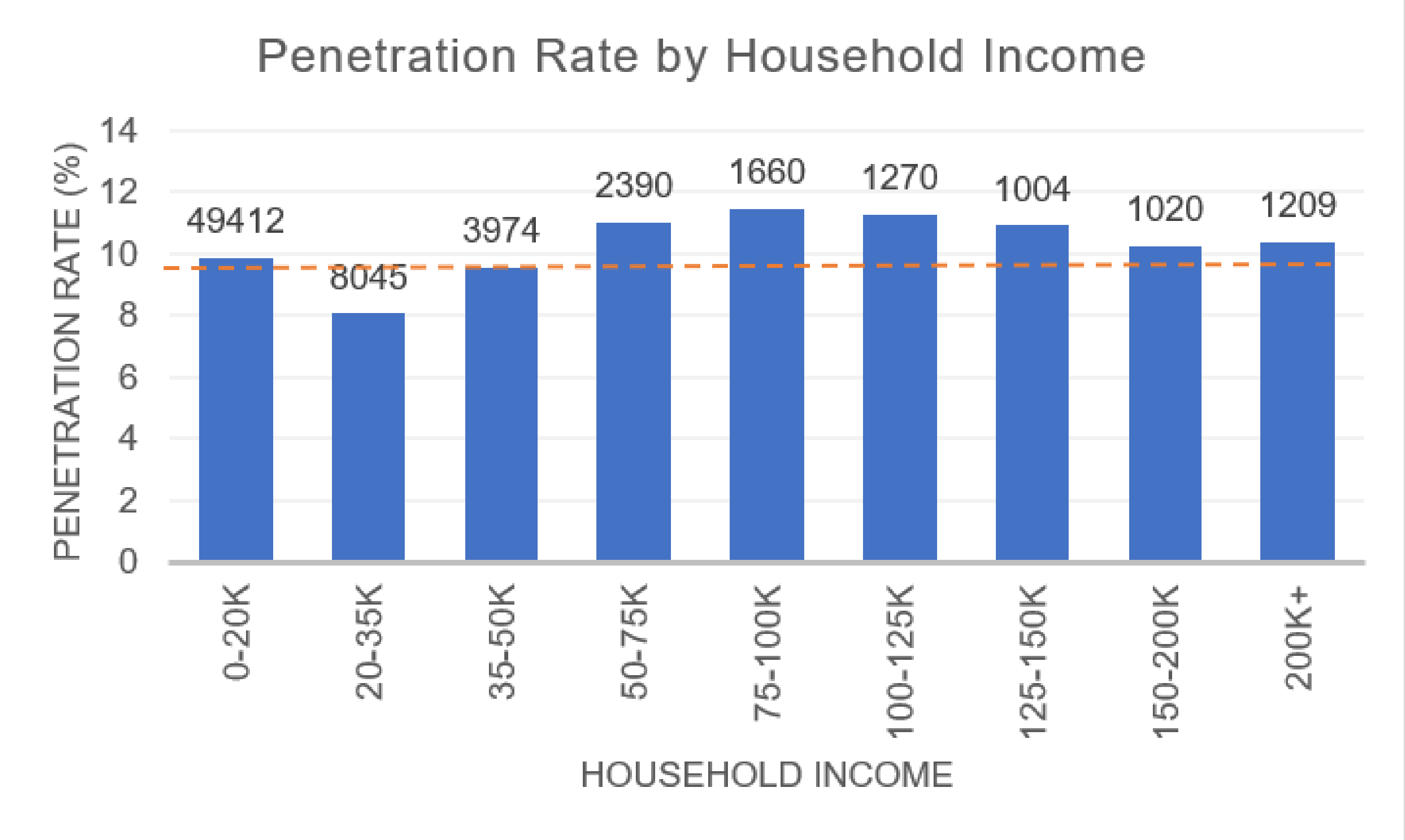
3 Privacy Protections

- How is the data collected, processed, and shared?
- How does the process protect privacy of individuals?
- Where are privacy practices built into the process?

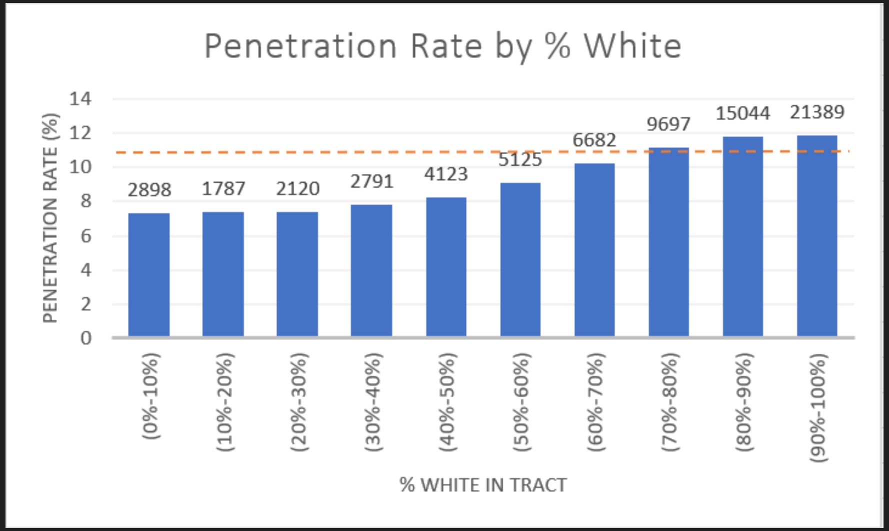
4 Validation and Uses

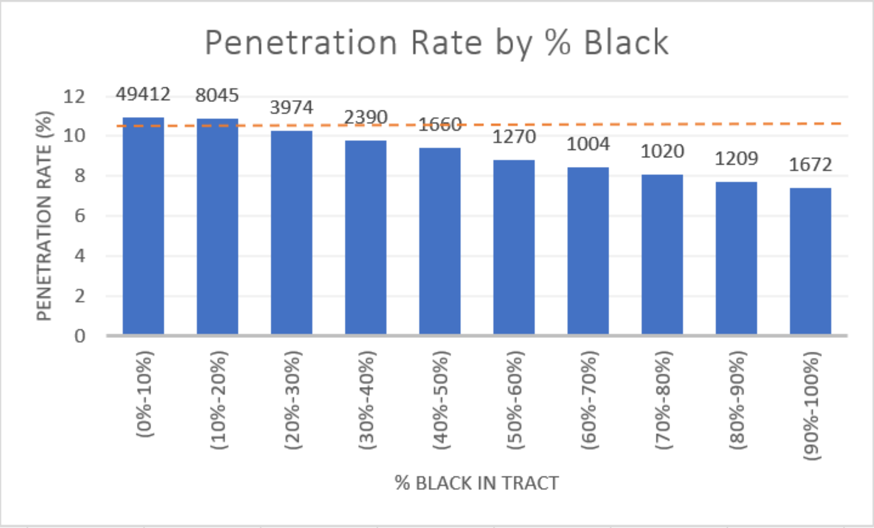
- How are the metrics validated?
- How have the metrics been used in real-world applications?
- How do customers access the metrics?

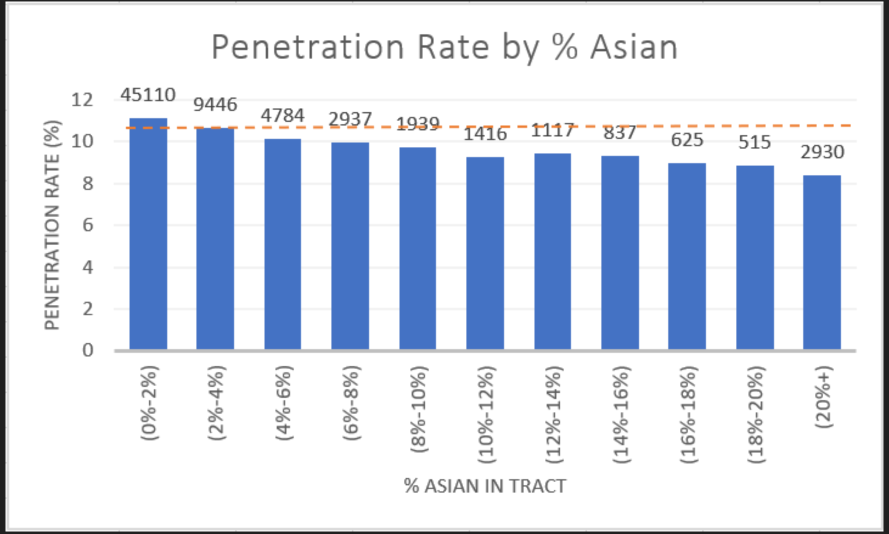
Sample representativeness

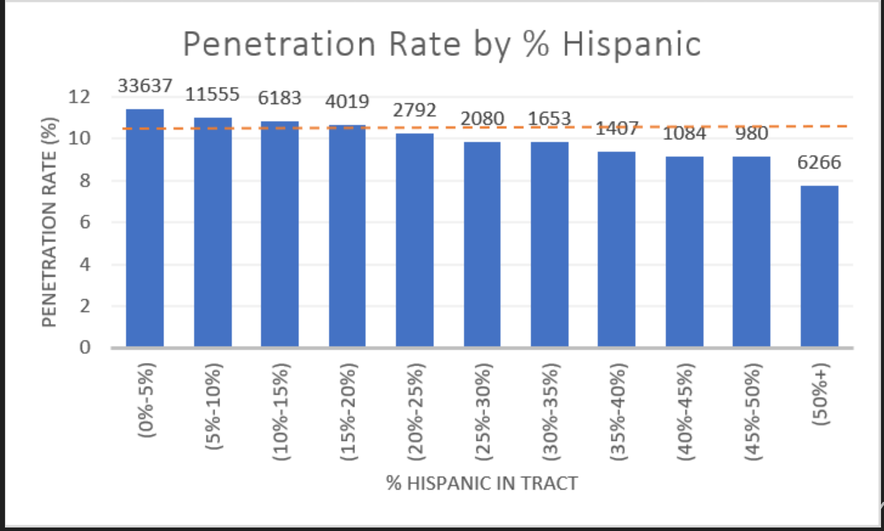


Sample representativeness







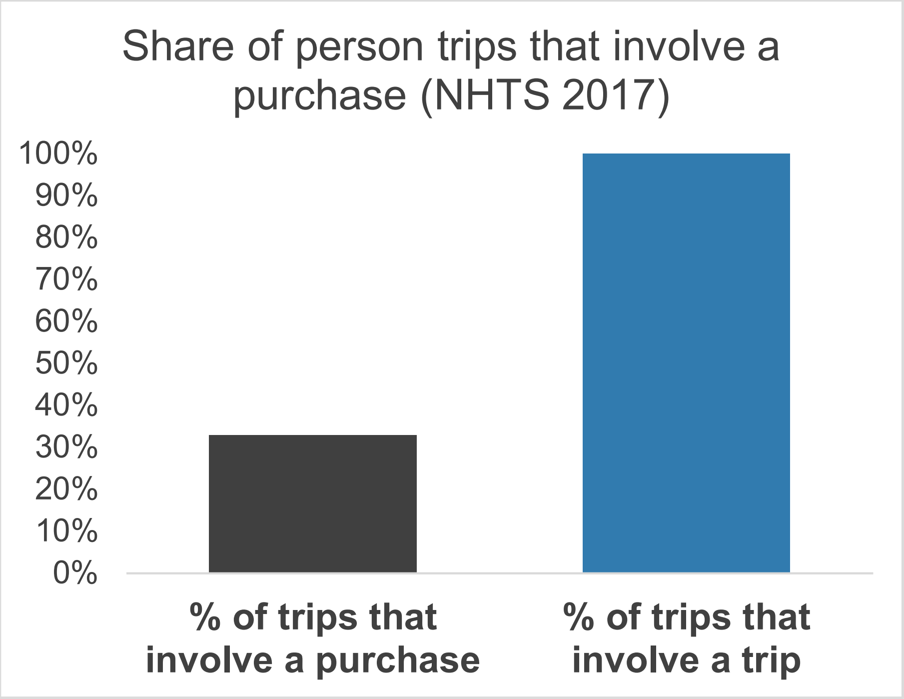


Sample representativeness

Income Bin	StreetLight Data	2017 NHTS Sample	StreetLight Data Bin compared to < $20K Bin	NHTS Data Bin compared to < $20K Bin
Less than $20,000	9.80%	0.07%	1.00	1.00
$20-$35K	8.10%	0.08%	0.83	1.14
$35-$50K	9.60%	0.09%	0.98	1.29
$50-$75K	11.00%	0.09%	1.12	1.29
$75-$100K	11.40%	0.09%	1.16	1.29
$100-$125K	11.20%	0.09%	1.14	1.29
$125-$150K	10.90%	0.09%	1.11	1.29
$150-$200K	10.20%	0.10%	1.04	1.43
$200K+	10.40%	0.09%	1.06	1.29

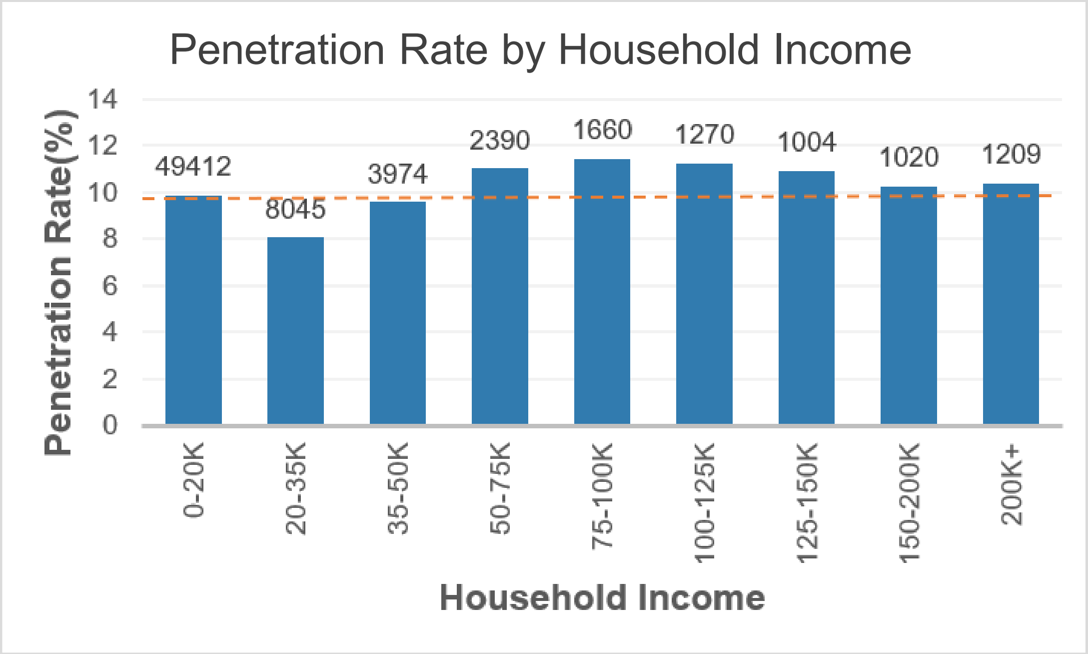
Density	StreetLight Data	2017 NHTS Sample
Urban	10.2%	0.08%
Rural	12.2%	0.11%

Is LBS data "unrepresentative"?
Not if you work with excellent suppliers.



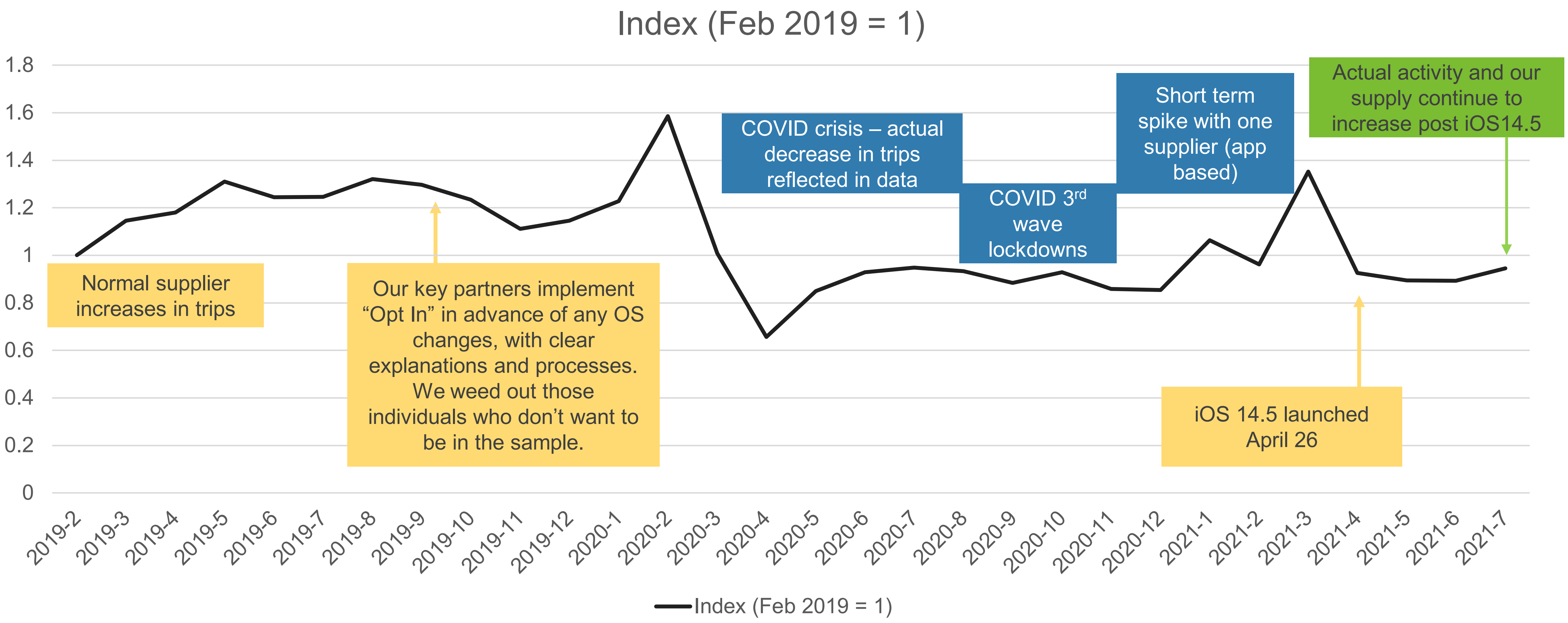
Biased trip purpose - hard to normalize

Well distributed - straightforward to normalize



Was LBS data impacted by the iOS14 updates?

Not if you work with excellent suppliers.



What is YOUR sample size?

110M

Usable devices in our current sample

1.5B

trips analyzed in January 2020*

Your project sample size

110 Million Usable Devices in U.S. + Canada

110M

Usable devices in our current sample

What are these devices?

- A mix of LBS smart phones, connected trucks
- We work with suppliers to increase presence of **highly active devices** that are making consistent trips throughout the year.
- We continually **de-duplicate devices,** and **remove devices with infrequent pings**
- These create over **half a trillion "pings" each year**

1.5+ billion usable trips analyzed per month

1.5B

trips analyzed in January 2020*

How do we do it?

- We continually vet the quality of apps and devices to achieve **maximum number of high-quality trips.**
- We work with suppliers to increase presence of **highly active devices** that are making consistent trips throughout the year.
- We continually **de-duplicate devices, and remove "inaccurate", unusable trips.**

*Last typical, pre-COVID month

Transparent Sample Size Information for All Projects

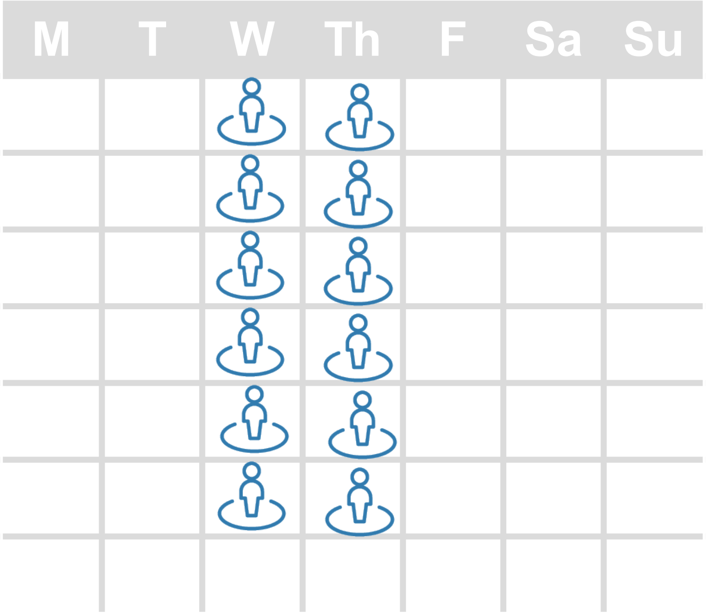
Your project sample size

- Sample size **varies** by time, data period, etc.
- For every StreetLight project you run, we'll share the **sample of devices and trips** that contributed to your analysis.
- Where feasible, we'll share **confidence intervals** too

Different approaches to sample size: Big Data vs. Traditional

Traditional (tube counter, etc.)

All people for a few days

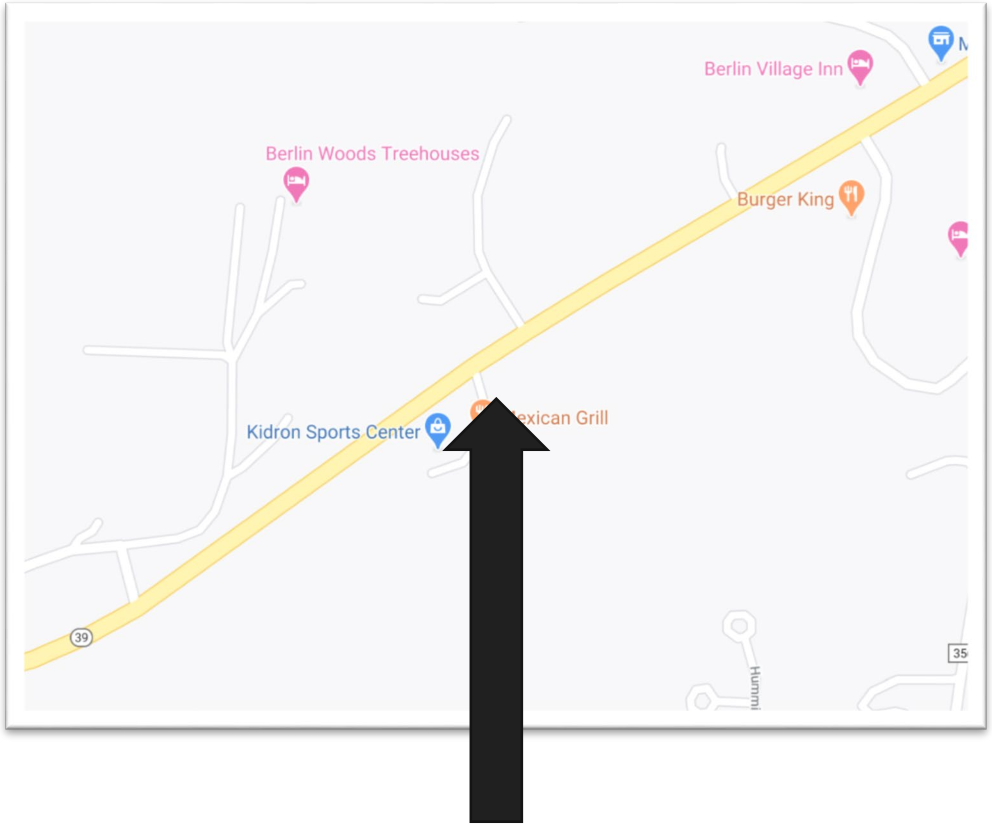


vs.

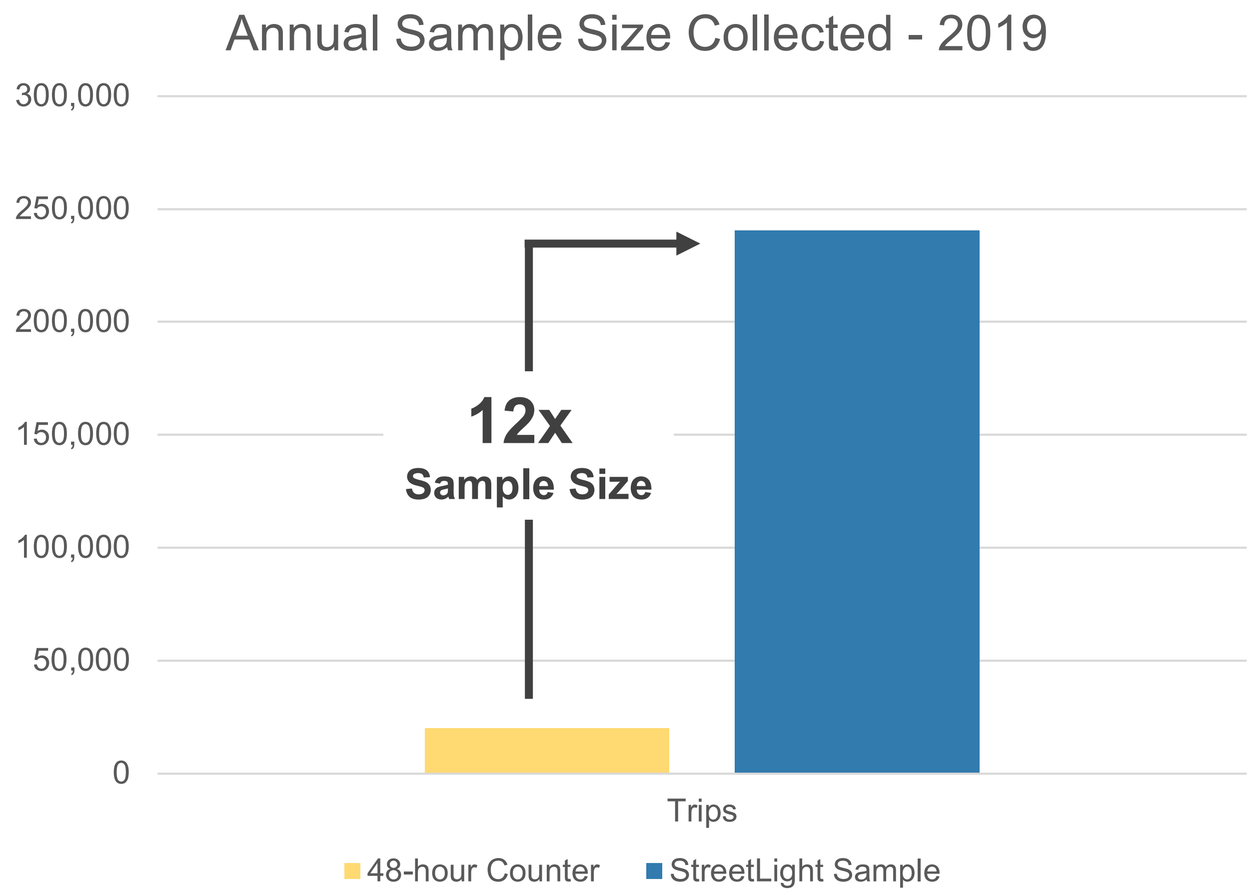
Big Data

Many people for all days

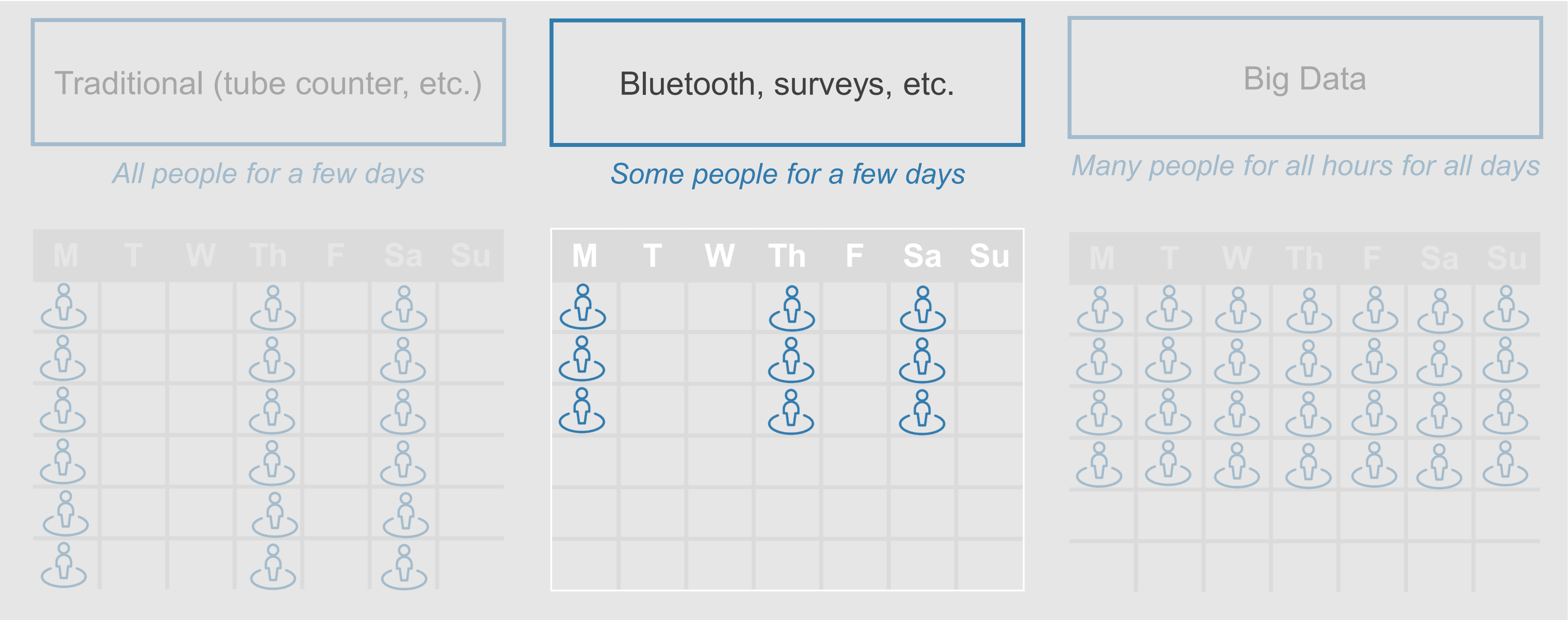
Different approaches to sample size: Big Data vs. Traditional



Location of permanent counter near Berlin, Ohio
2019 AADT = 10,028 (3.66M trips/year)



Different approaches to sample size: Big Data vs. Traditional



How we make our sample representative

A SIMPLIFIED SUMMARY

Population Adjustments

Each device carries an individual adjustment factor based on how many StreetLight devices live on its home block (versus actual population on that block).

IoT Counter Normalization

We normalize across space/time by comparing the number of trips we sense to 10,000+ permanent loop counters embedded in the infrastructure across the US

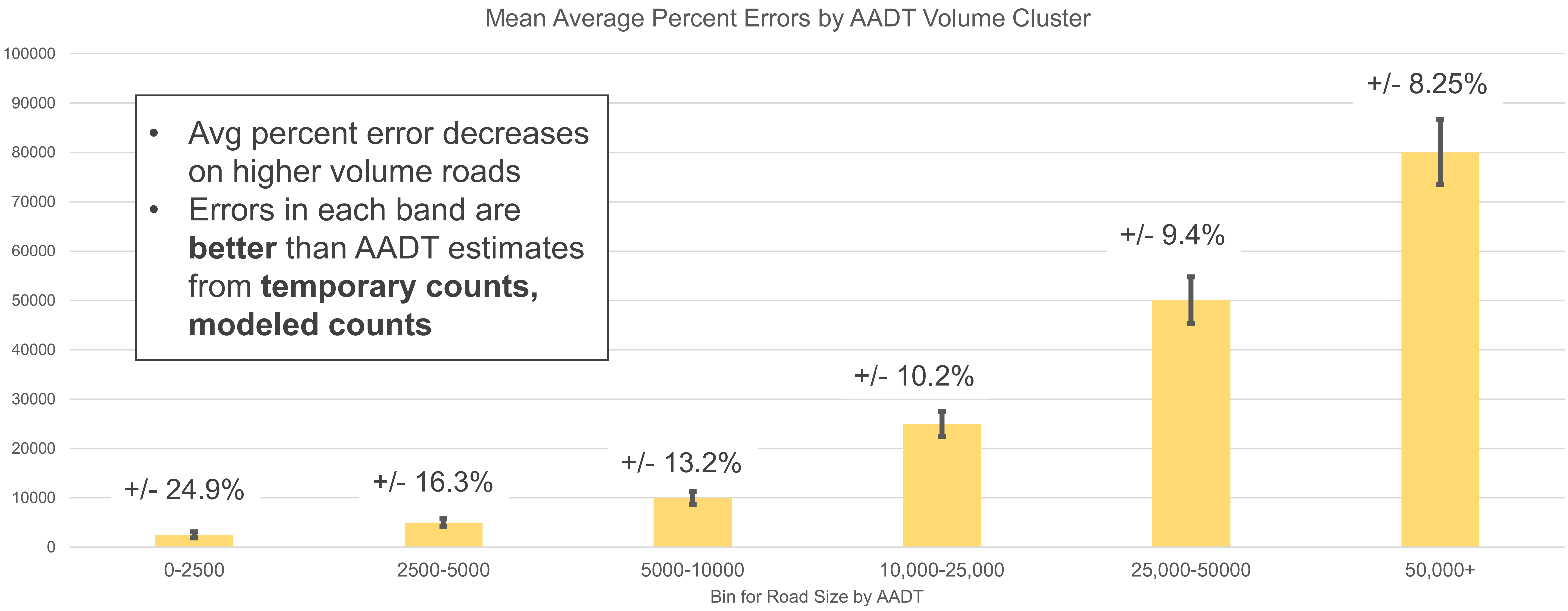
And More…

We check for major shifts in supplier data feeds, we use inputs about weather, road type, season and much more to finalize our normalization.

Repeat and update every month

How accurate are the metrics?

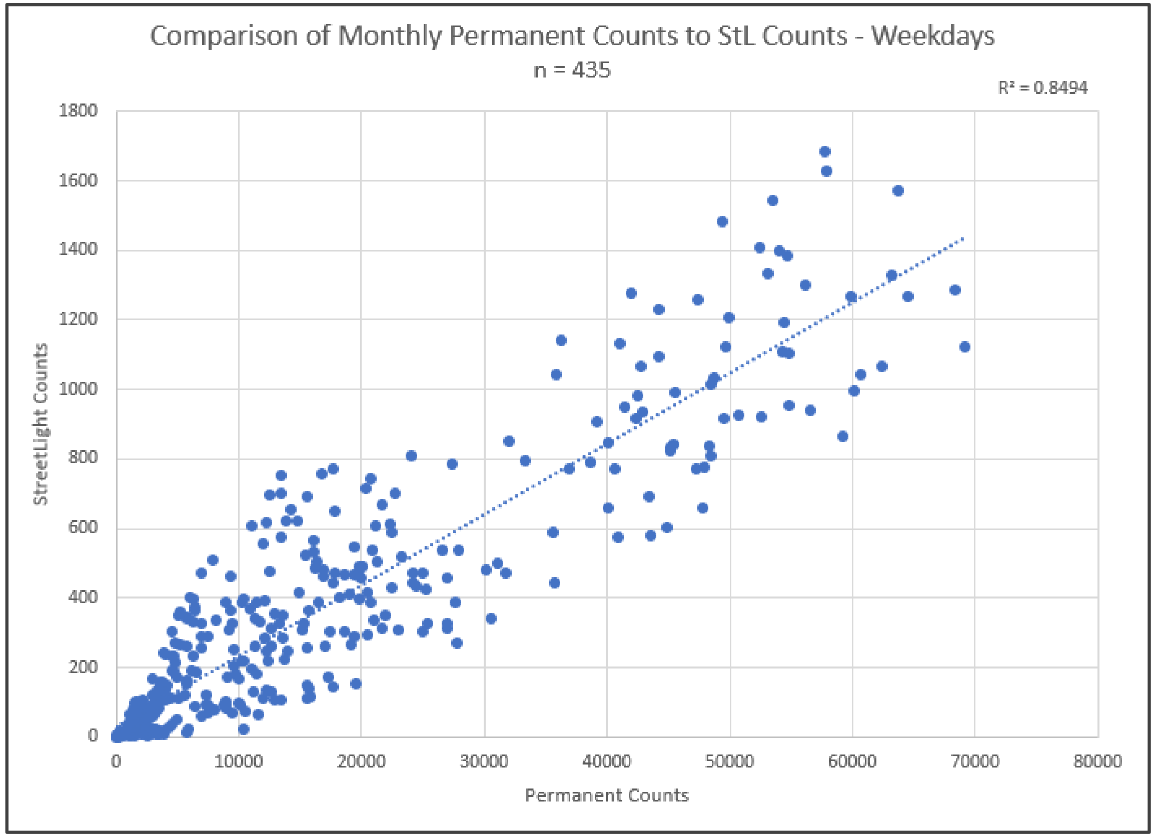
AADT accuracy by size of road (link to detailed validation paper)

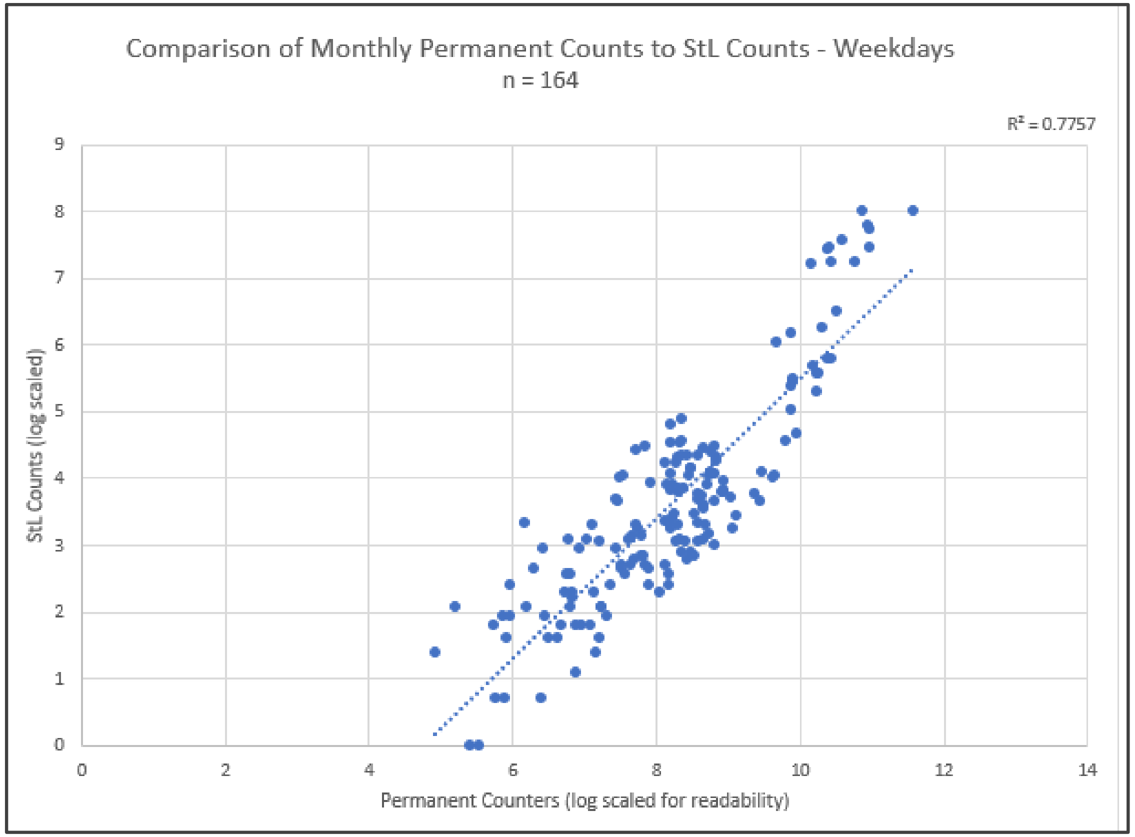


- Avg percent error decreases on higher volume roads
- Errors in each band are **better** than AADT estimates from **temporary counts, modeled counts**

How accurate are the metrics?

Bike and pedestrian Counts have extremely high correlation (link to detailed validation paper)



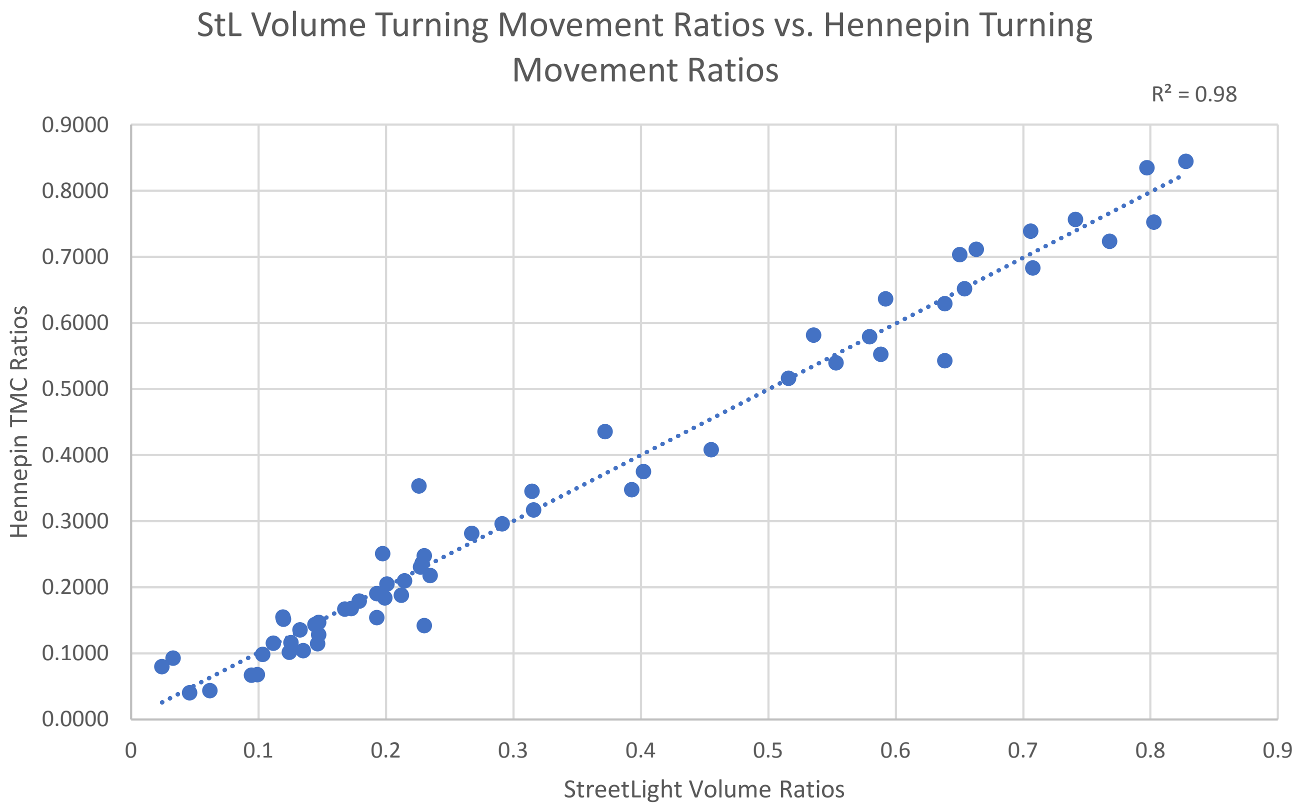


Correlation results comparing StreetLight to permanent counters for bicycle and pedestrian modes for weekdays show R^2 of 0.84 for bicycles and R^2 of 0.78 for pedestrians, **which means the actual and estimate values are highlight correlated**

These comparisons are based on permanent bicycle counters in San Francisco (SFMTA) and the Delaware Valley (DVRPC) and permanent pedestrian counters in Washington D.C. and the Delaware Valley (DVRPC) from 2018.

How accurate are the metrics?

Turning Movement ratios and demonstrate high accuracy correlation



Correlation between Hennepin turning movement ratios and StreetLight Volume ratios demonstrate **R^2 value of 0.98, indicating a very high correlation.**

These comparisons are based on permanent counter locations in Hennepin County, Minnesota, from 2017

Big Data for Modeling

- Trip Based Models
  - Trip rates can be customized beyond basing them on household size, household size and income level.
  - Gravity models and destination choice models can be refined. Different trip distribution models can be estimated since big data can be used to calibrate and validate them, without which validation data is lacking
  - Routing information from GPS and LBS data can be used to improve trip assignment algorithms

Big Data for Modeling

- Tour-Based Models
  - Trip stops along a tour can be better understood – where they are how often
  - Mode choice for tours and trips within them
- Activity Based Models
  - Reliable (anonymized) devices can be studied to better understand individual travel patterns to develop travel profiles
  - Long term (employment, school) and short term (daily routine) decisions impacting travel can be estimated

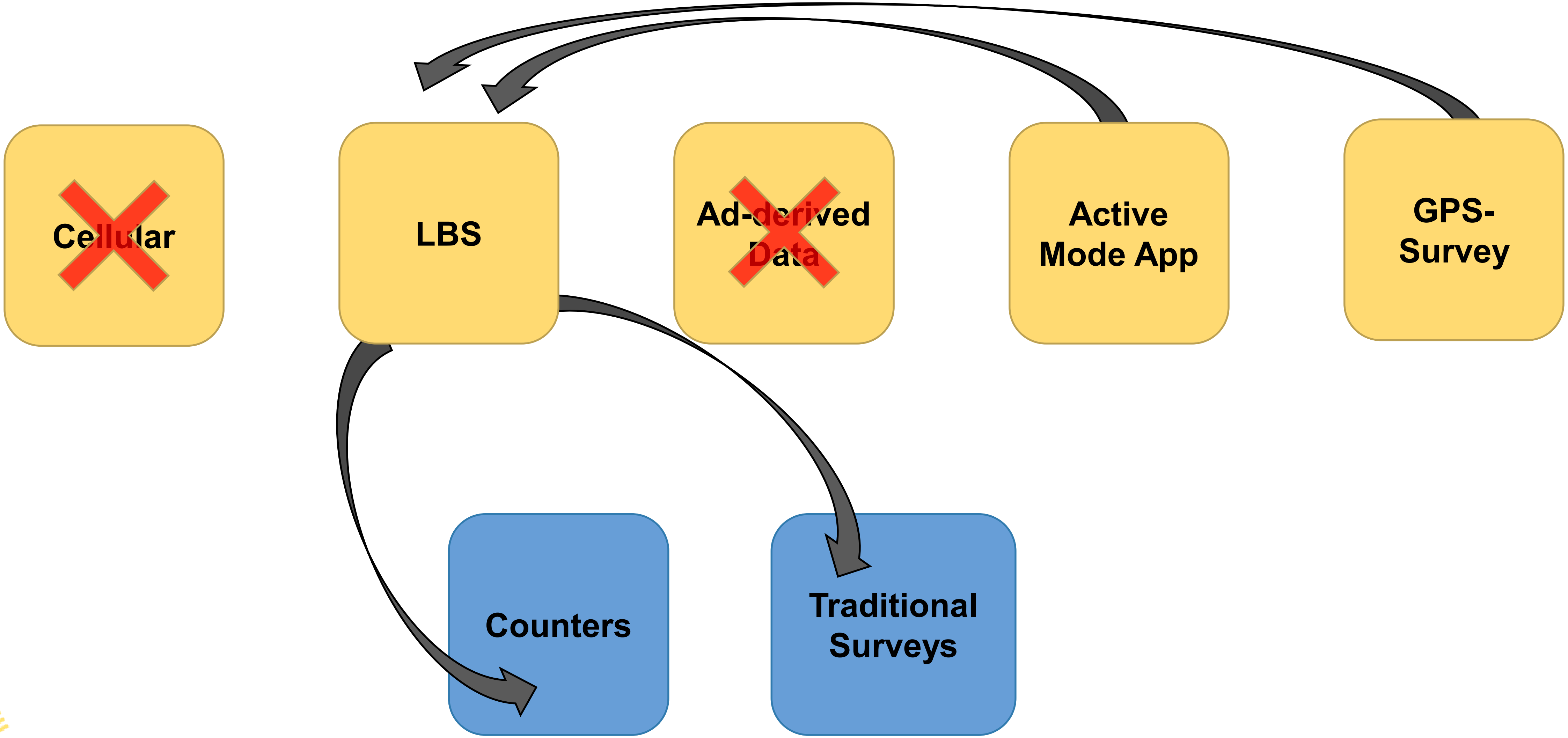
P-A vs O-D

Trip and tour based models start with Production-Attractions and convert them into O-D to get the actual direction of trip before trip assignment.

Production-attraction format of trips expresses the directions going from home-end of the trip (production) to non-home end of the trip (attraction). That does not reflect the real directions from origin to destination.

StreetLight creates trips in O-D format. Knowing the home location (zone) of the device making the trip, the trips can be converted to P-A format.

Step 1: Pick the Right Data

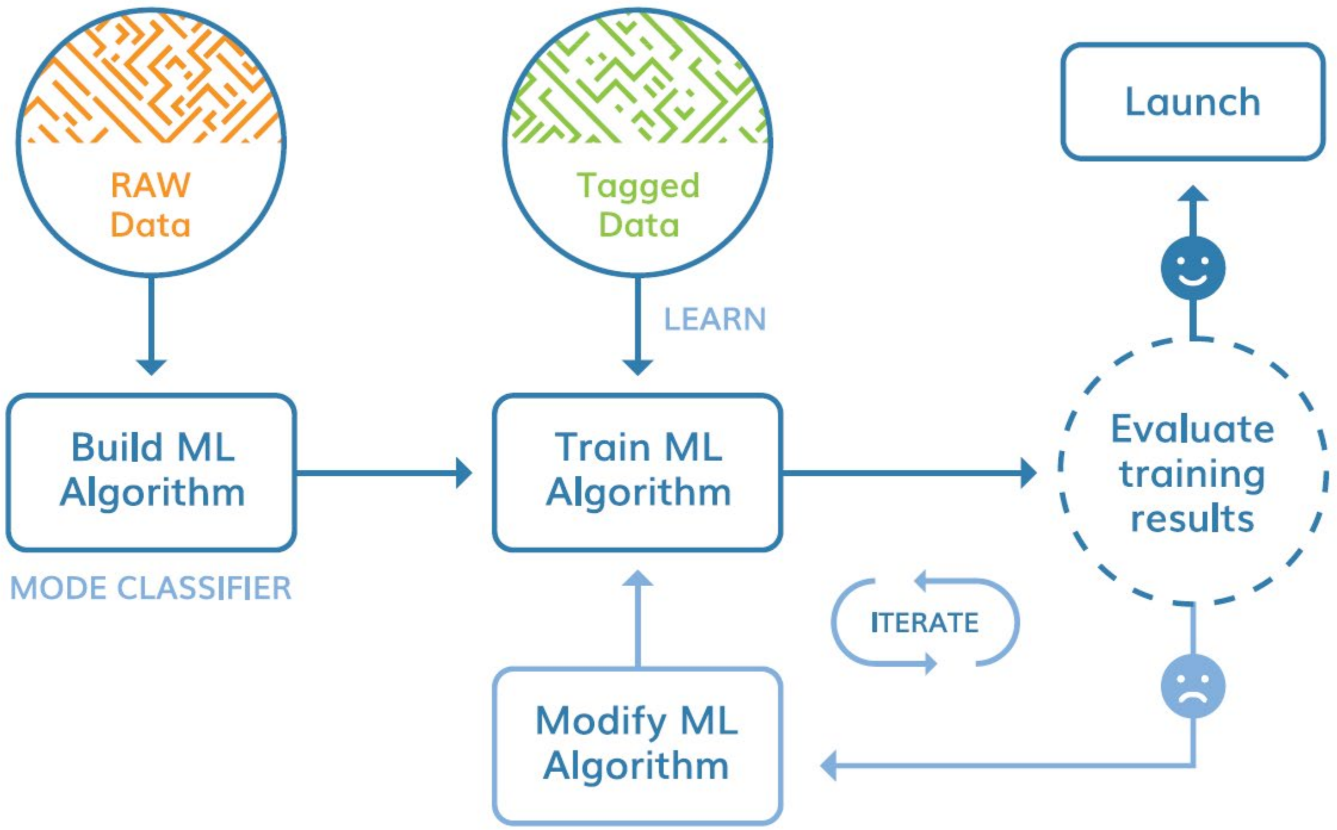


Step 2: Machine Learning to Recognize Modes at the Ping Level

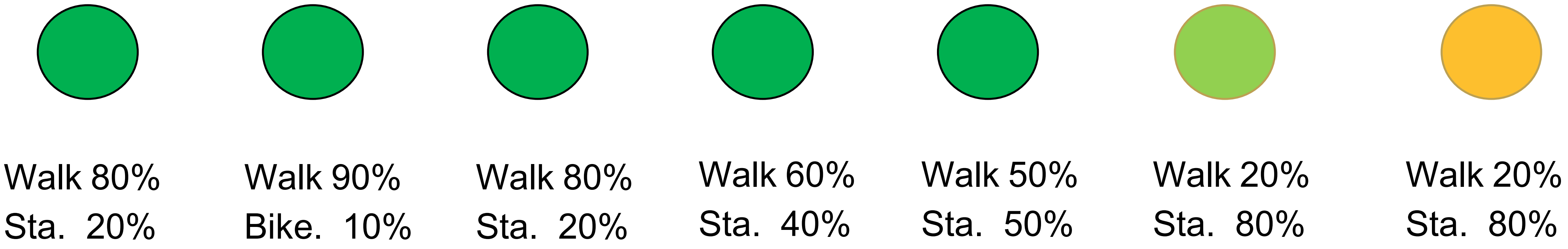
Training a Random Forest Classifier – Data Sets by Source

Source	Tagged Points	Harvested Points
Caltrans Travel Survey (NREL**)	26M	~500K
Atlanta Regional Travel Survey (NREL)	2.4M	~75K
Mid-Region Travel Survey – Albuquerque (NREL)	3M	~92K
Southern Nevada Household Travel Survey (NREL)	4.2M	~133K
Capital Bikeshare	334K	~48K
Beijing Pedestrian (Microsoft)	5.5M	~27K
Total Number of Points in Training Data Set		~900K

3 METHODOLOGY + VALIDATION



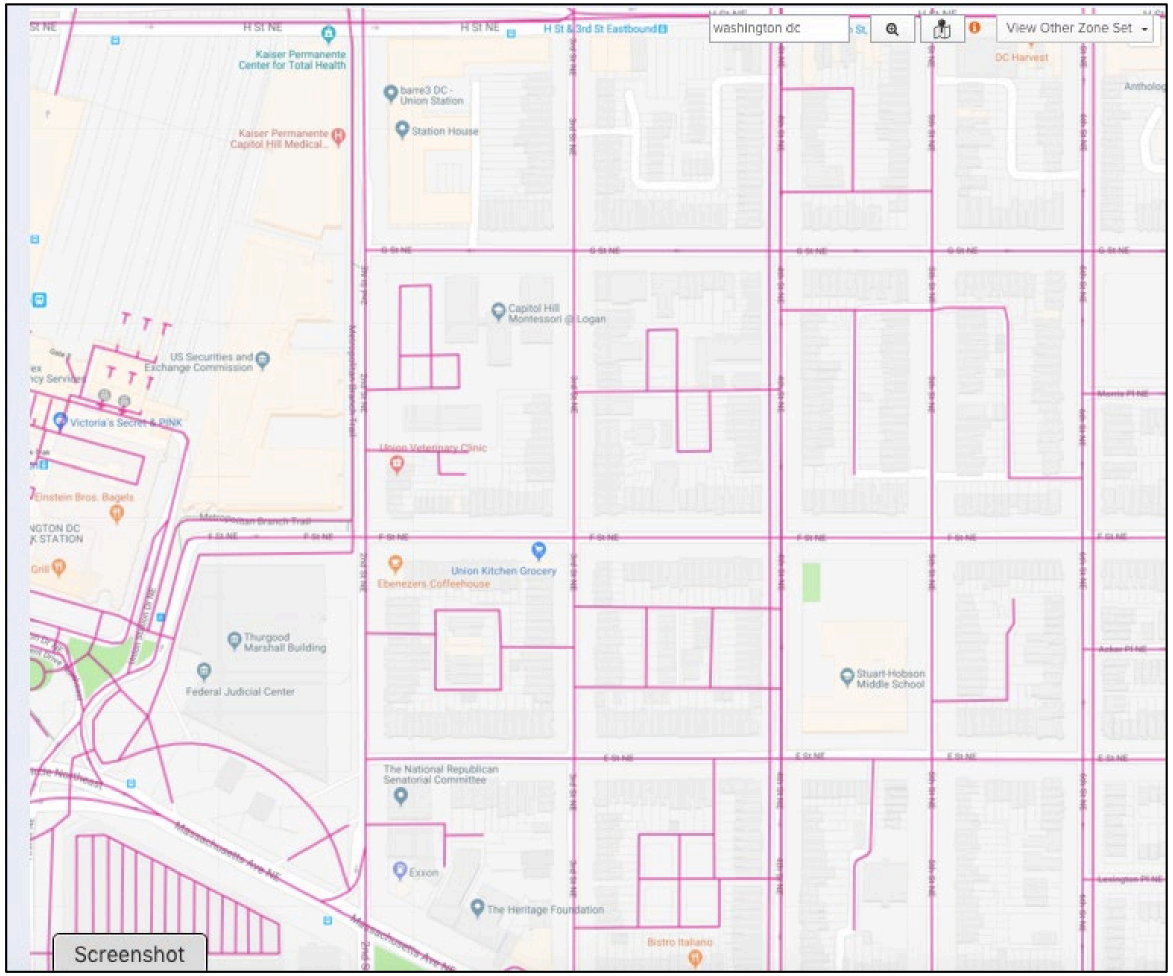
Step 3: Group “Pings” into Mode-Assigned Trips



- Apply intelligence from machine learning process to infer probability of mode choice for each ping
- Stationary is a “mode”

Step 4: “Lock” to Allowable Networks

- No geometry subtracted from car Open Street Map (OSM), only added
- Implications for Pedestrians
  - Complexities with pedestrians: jaywalking can be missed
  - Very large and spread out buildings, like conferences centers, can be confusing



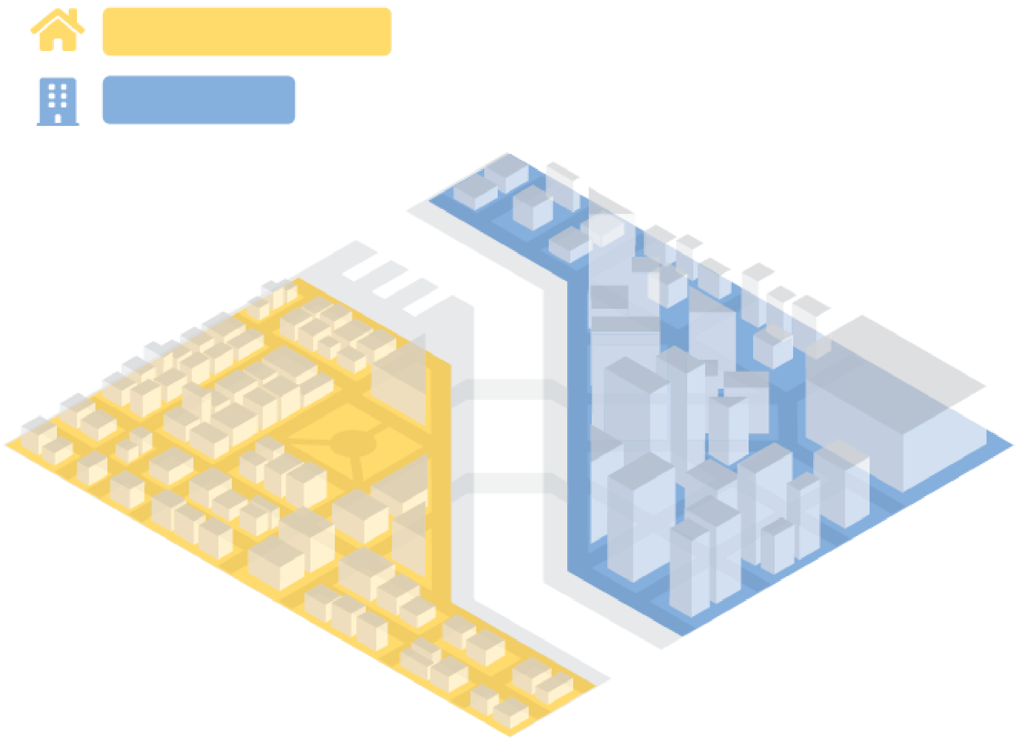
Inferred home and work locations

Identifying Inferred Locations

For home locations, we look at an entire calendar month, and identify the top five neighborhoods where a device pings during evening and night-time hours. Each of the five neighborhoods is assigned a probability weighting.

For example, a device may have a home location with 75% probability in Philadelphia and 25% in Cape May, N.J., distributed across three neighborhoods in Philadelphia and two neighborhoods in Cape May, N.J. Then, combining devices with a home neighborhood in Philadelphia and Cape May with their probability weighting would give us the Philadelphia/Cape May based composite travelers.

Similarly, for work locations, we look at the top five neighborhoods where a device pings during work hours only.



Contextualizing and Aggregating

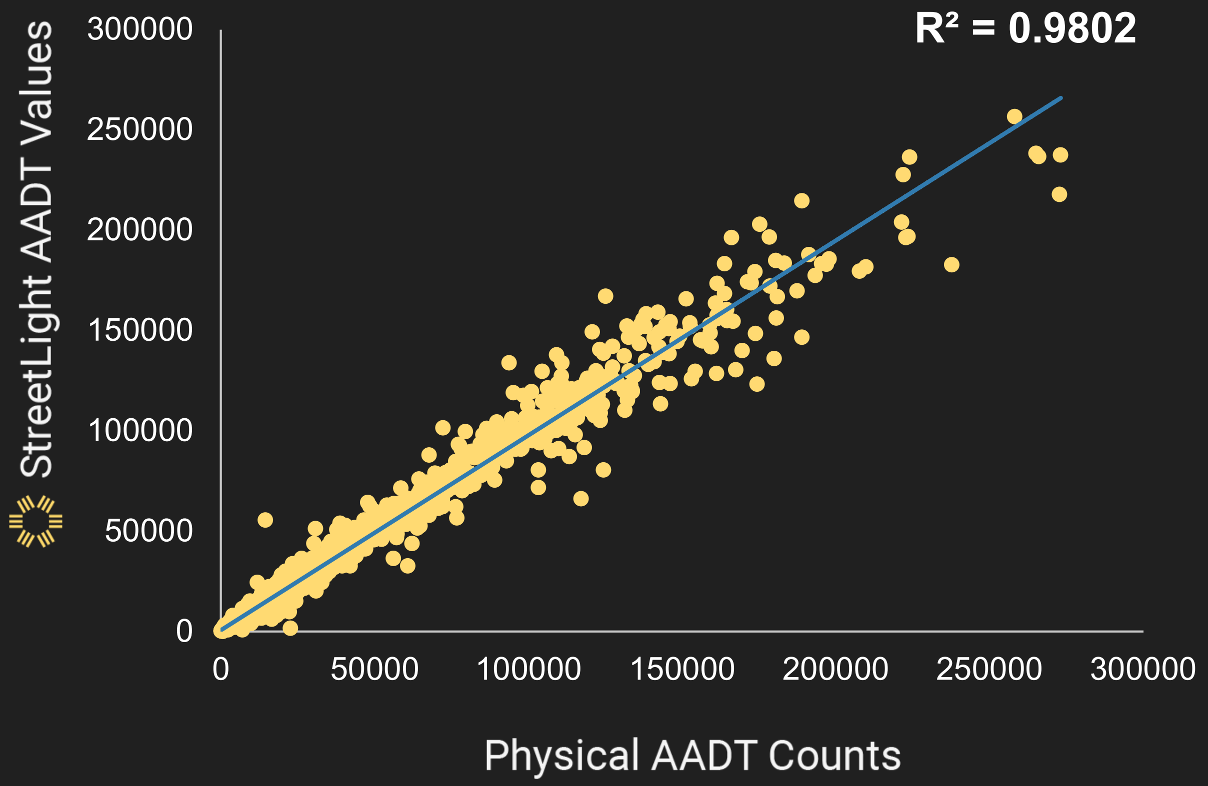
StreetLight adds context by incorporating demographic information based on the 2010 U.S. Census and aggregates to 1km x 1km grids. Metrics are provided through StreetLight InSight® and can be viewed in our interactive visualizations.

StreetLight can also provide output at other geographic unit levels such as Census Block Groups, Zip Codes, metropolitan areas, and states.

StreetLight AADT is trained on a Set of Permanent Counters

- StreetLight 2020 AADT data is **trained and validated** using 3,000+ unique permanent counter locations (6,600 permanent counts) across 25 U.S. states

- The R^2 between StreetLight AADT and 2020 AADT values from permanent counters is 0.98, **indicating a very strong relationship**

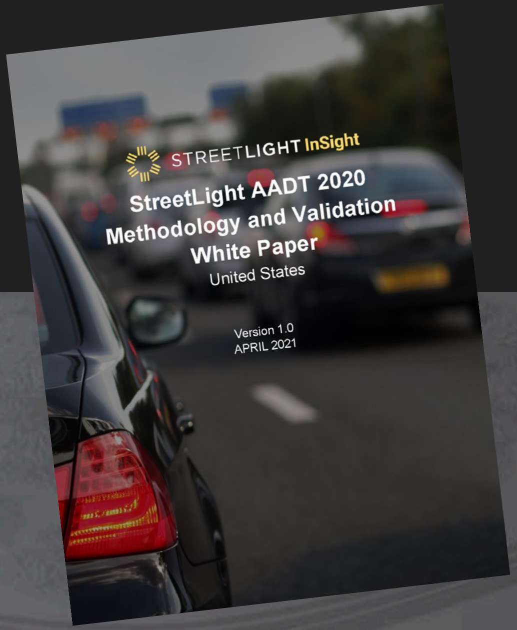


Continued year-over-year improvements to StreetLight's AADT Metric

AADT 2020 improves results for small traffic volume roads through large traffic volume roads, due to updated machine-learning model methodology

AADT Volume Range	StreetLight 2018 AADT	StreetLight 2019 AADT	StreetLight 2020 AADT
	MAPE (%)	MAPE (%)	MAPE (%)
A: <= 499	840.4	42.51	74.81
B: 500 - 1,999	27.29	27.76	25.48
C: 2000 - 4,999	21.14	19.3	14.98
D: 5,000 - 9,999	17.47	14.72	12.52
E: 10,000 - 19,999	16.21	12.7	11.71
F: 20,000 - 34,999	13.72	10.53	8.84
G: 35,000 - 54,999	13.98	9.36	7.32
H: 55,000 - 84,999	11.54	8.38	7.55
I: 85,000 - 124,999	11.58	8.63	7.09
J: > 125,000+	11.19	8.41	8.99

Comparison of cross validation results of mean percent error (MAPE) metric among StreetLight's AADT models for 2018, 2019, and 2020 (U.S.).



Download the full white paper to go deeper on the AADT methodology and validation.

StreetLightData.com /AADT2020

Accuracy exceeds industry standards compared to temporary counts

For medium and large roads, **the AADT 2020 model performs competitively with 48-hour same-year temporary counts across most accuracy metrics** and consistently better than the typical situations with no counts

AADT Volume Range	Method (n)	95% TCE Error Range (%)	Median Bias (%)	MAPE (%)	NRMSE (%)
0-499 (very low)	Same Year Temporary Counts (SY-TC)	Unknown	Unknown	Unknown	Unknown
	Typical “No Count” estimates (Typ-NC)	Unknown	Unknown	Unknown	Unknown
	AADT 2020 K-fold	**58.5**	**49.4**	**66.41**	**76.11**
500 - 4,999 (low)	SY-TC	**34**	**-0.1**	**10.2**	**18.0**
	Typ-NC	n/a	n/a	**50**	**65**
	AADT 2020 K-fold	**40.50**	**4.3**	**19.0**	**24.0**
5,000 – 54,999 (medium)	SY-TC	**28**	**1.1**	**8.6**	**14.2**
	Typ-NC	n/a	**2**	**18**	**27**
	AADT 2020 K-fold	**27.88**	**-0.7**	**10.6**	**15.4**
55,000+ (high)	SY-TC	24	**1.4**	**5.3**	**9.5**
	Typ-NC	n/a	**1.5**	**20**	**12**
	AADT 2020 K-fold	**15.65**	**-0.9**	**8.1**	**12.7**

Note: there are no targets available for roads with AADT under 500

StreetLight’s AADT 2020 cross-validated results compared to 48-hour same year temporary count expansion as represented by Krile et al (2015) and typical situations with no counts for key statistical indicators.

Source: Krile, R., Todt, F., Schroeder, J. (2015). Assessing Roadway Traffic Count Duration and Frequency Impacts on Annual Average Daily Traffic Estimation (FHWA-PL-16-008). United States. Federal Highway Administration.

Validation of AADT counts in Tennessee