Data Science, AI and HPC

Big Data Solutions for Mobility Planning

Jane Macfarlane, UC Berkeley ITS

Lawrence Berkeley National Laboratory

October 2019

BERKELEY LAB

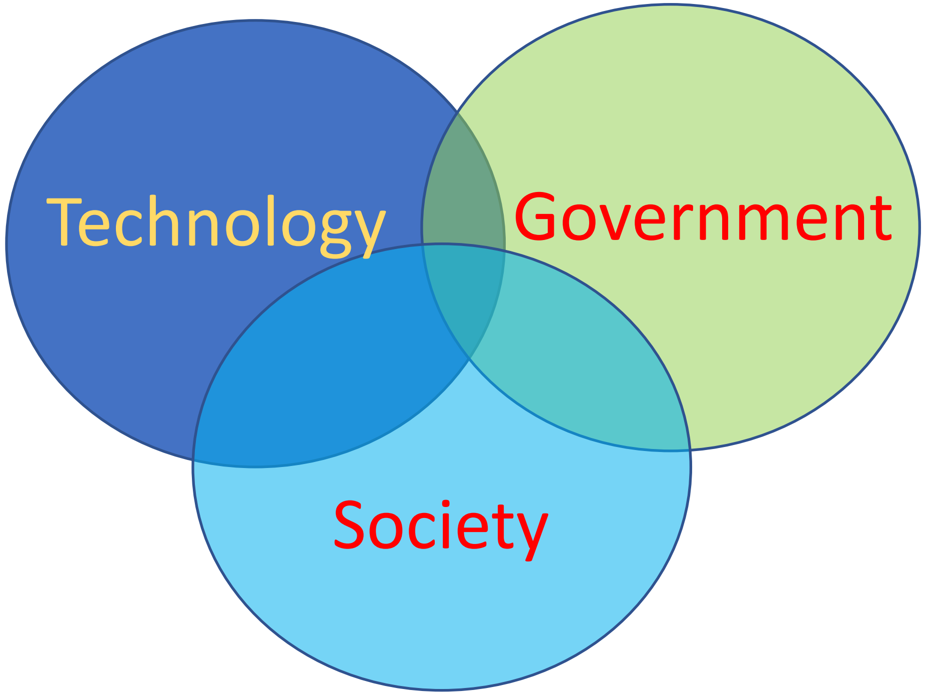
Argonne NATIONAL LABORATORY

Pacific Northwest NATIONAL LABORATORY

Cyber-Physical Transformation

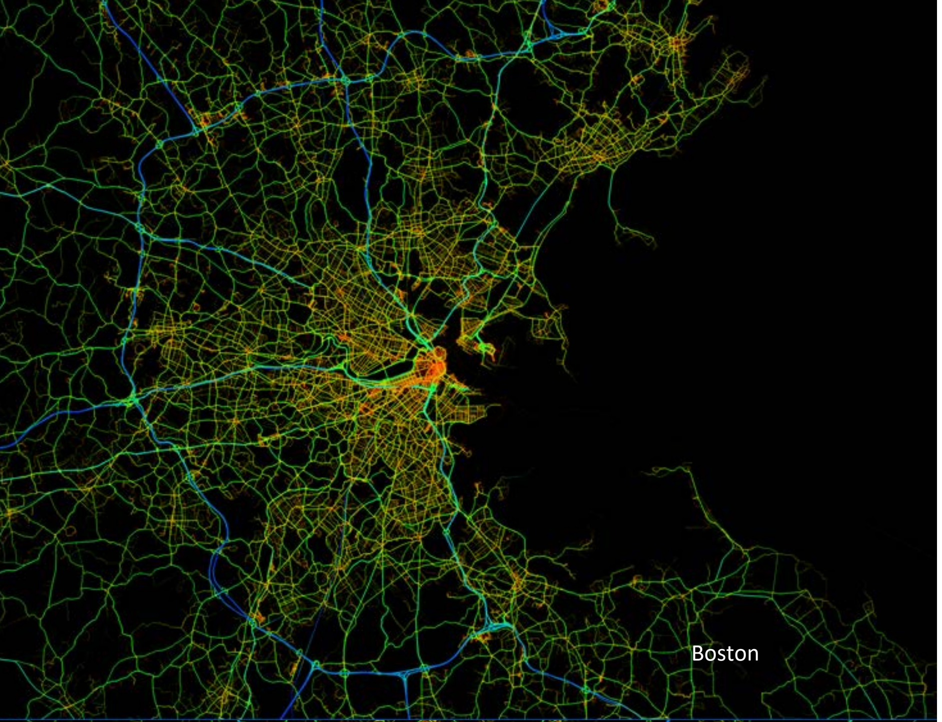

cricket

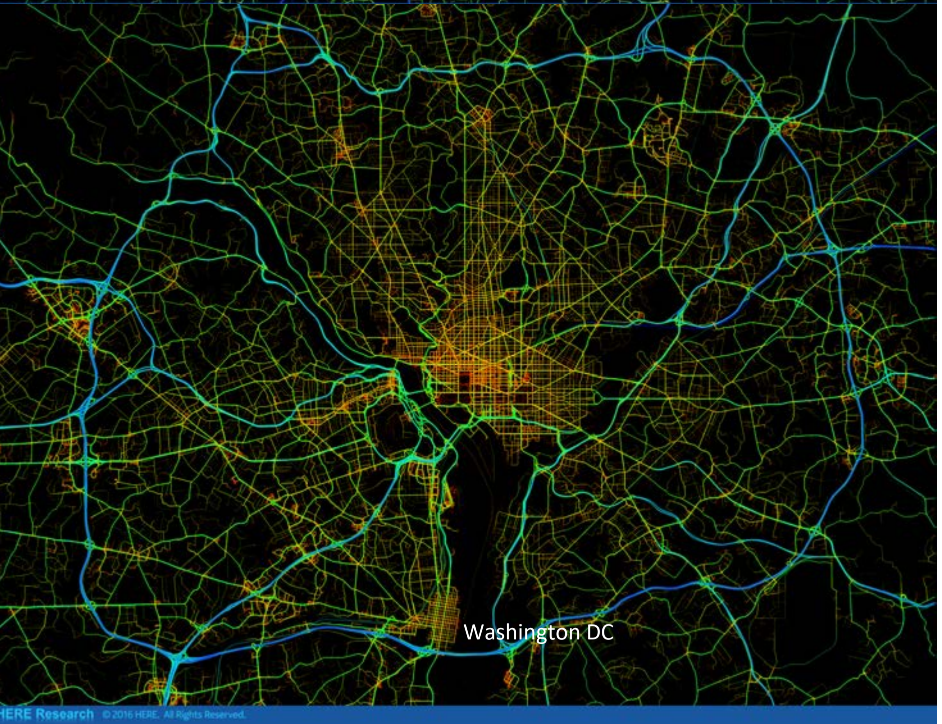
Convergence of bandwidth, network ubiquity, mobile devices/IoT, big data analytics



Fundamentally changing our social contracts

On Demand Society





Movement =>

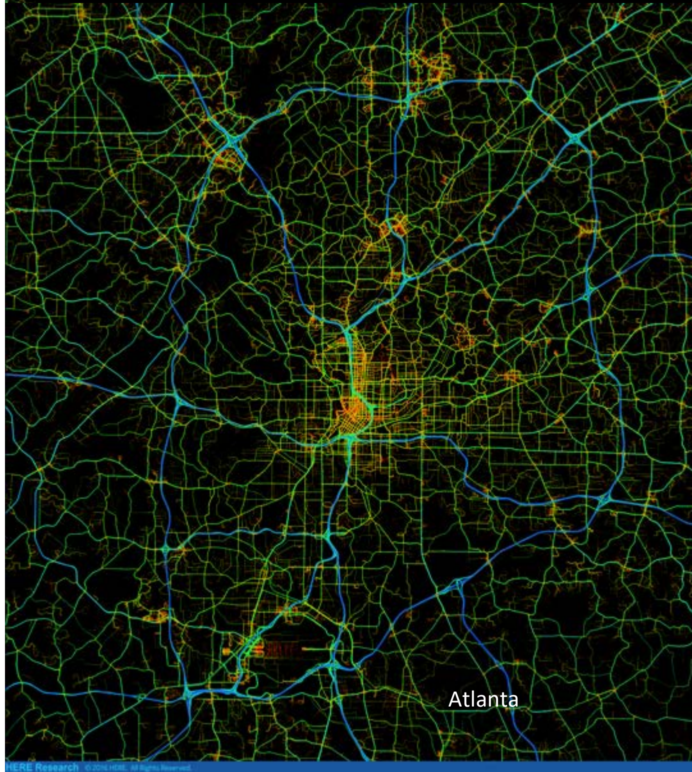
geospatial, temporal data

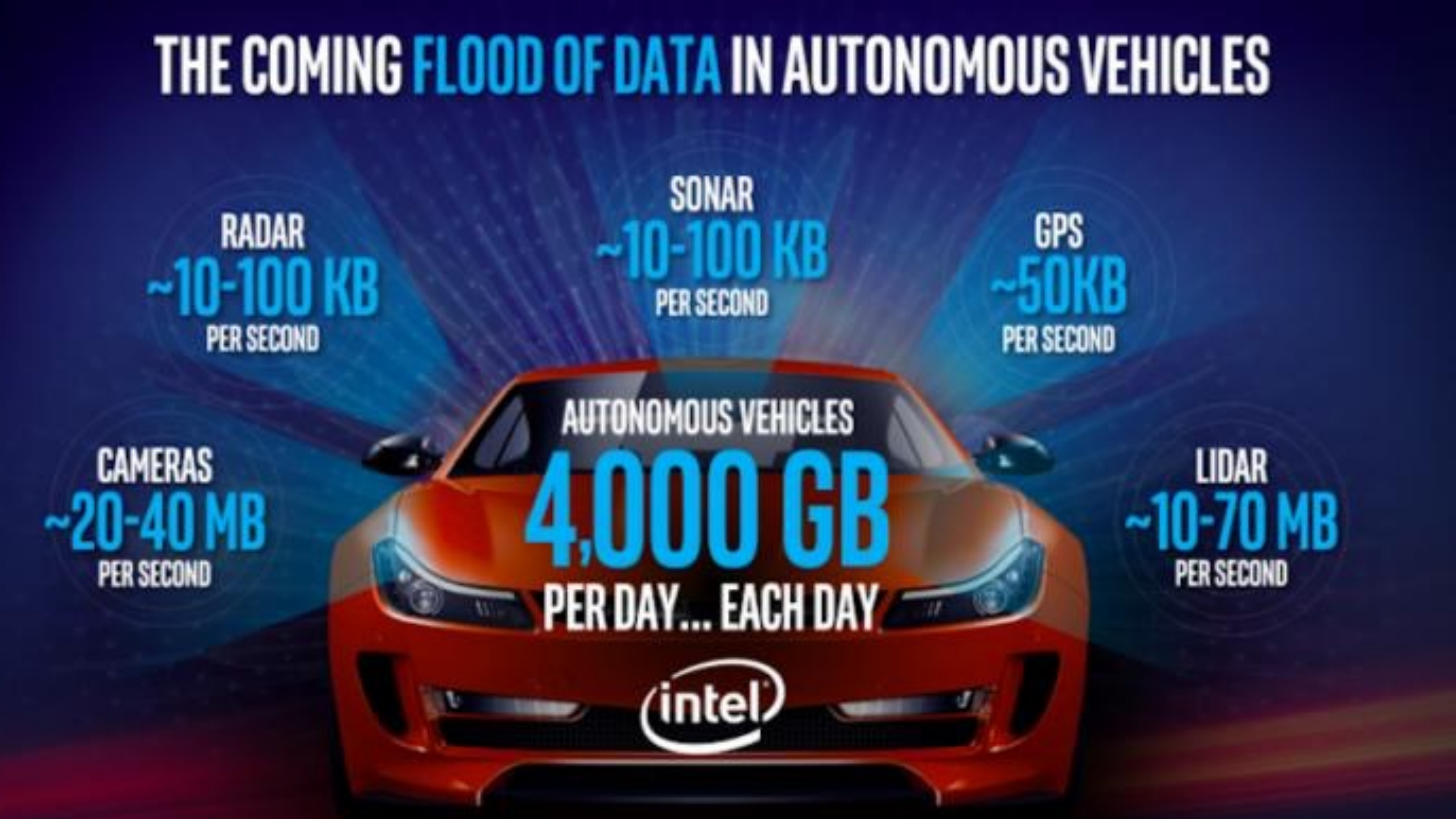
Smart phones
Vehicles
IoT Devices

GPS
Lidar
Images

Big Data

Images courtesy of HERE Research




THE COMING FLOOD OF DATA IN AUTONOMOUS VEHICLES
RADAR
~10-100 KB
PER SECOND
SONAR
~10-100 KB
PER SECOND
GPS
~50KB
PER SECOND
CAMERAS
~20-40 MB
PER SECOND
AUTONOMOUS VEHICLES
4,000 GB
PER DAY... EACH DAY
LIDAR
~10-70 MB
PER SECOND
intel

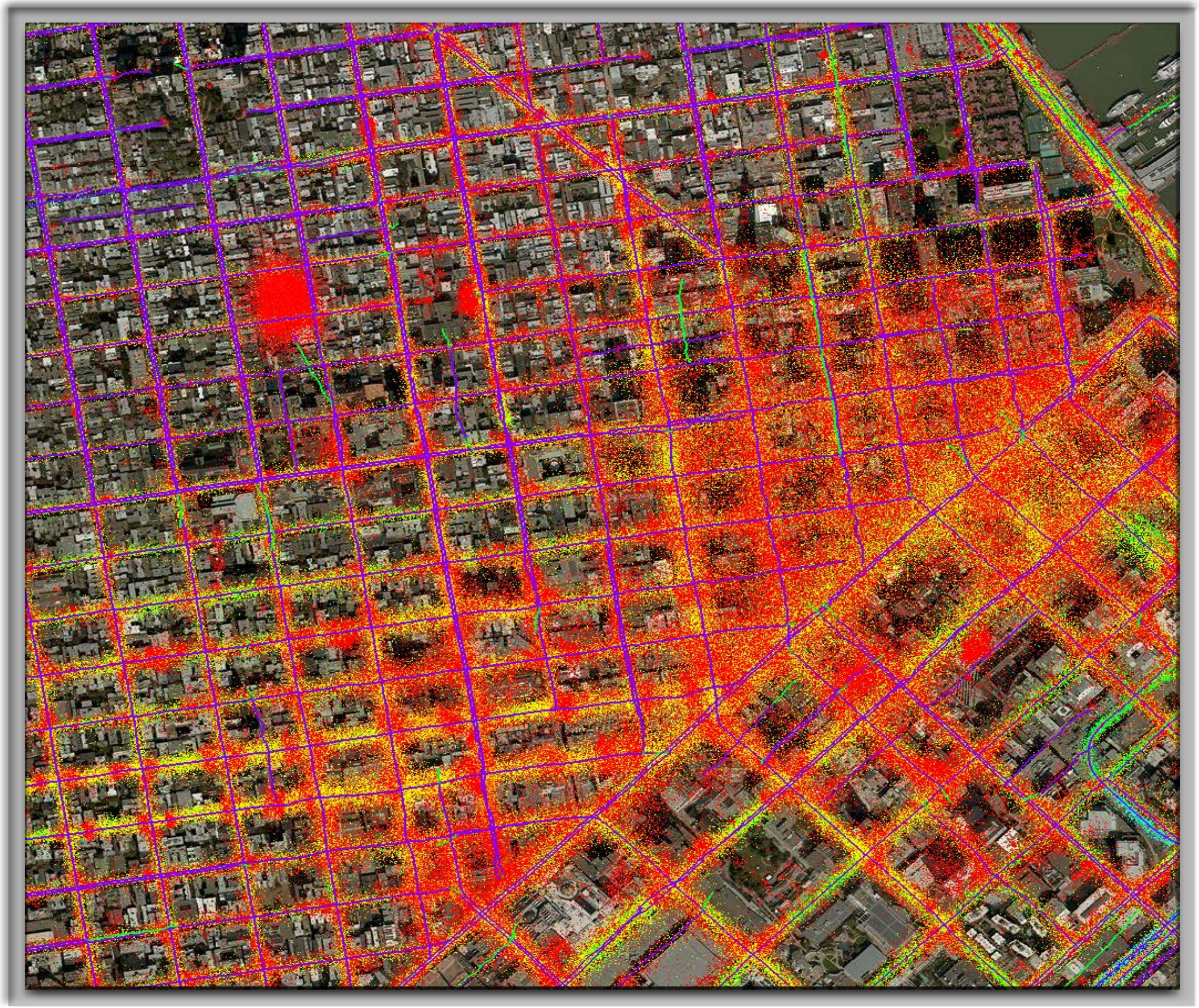
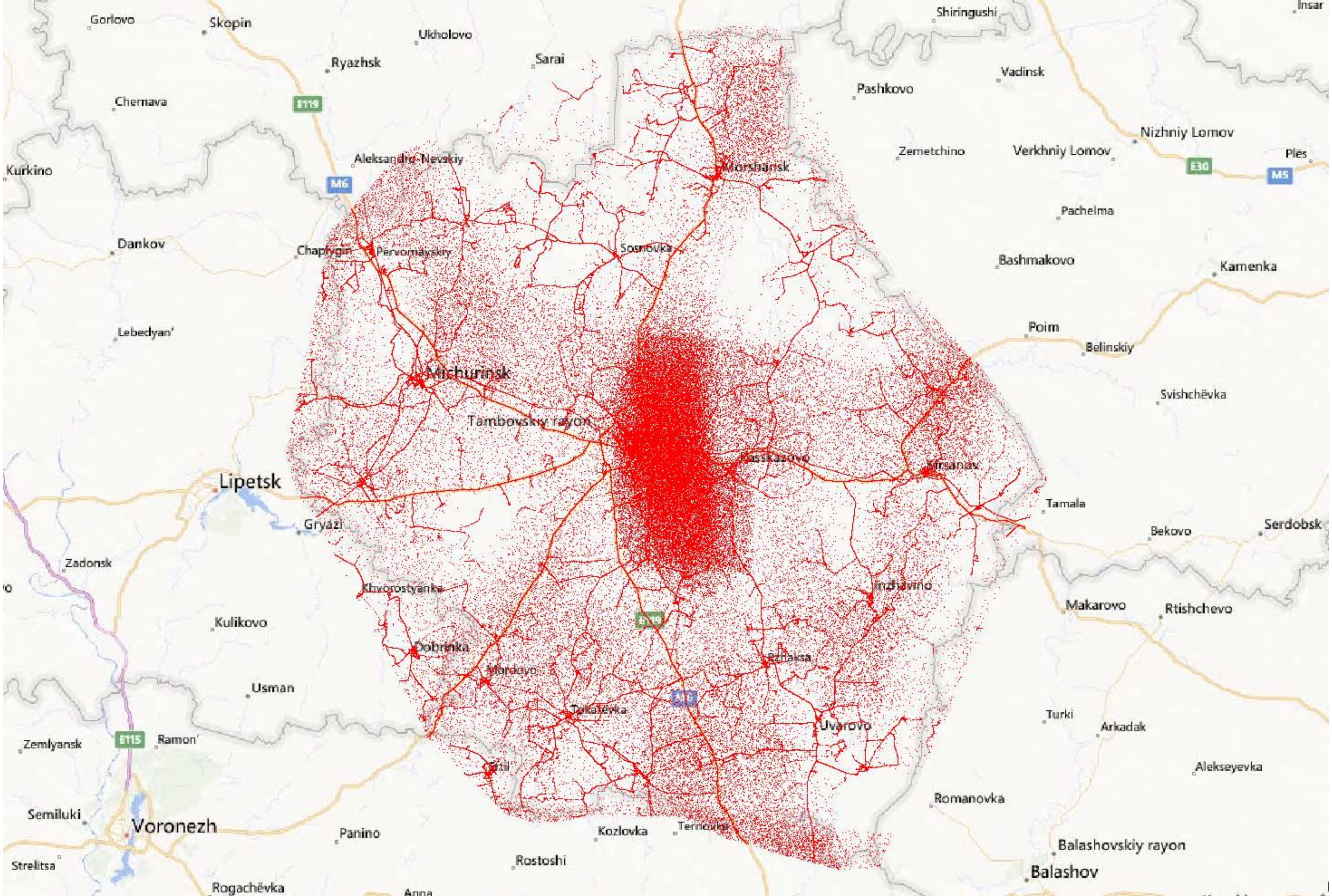
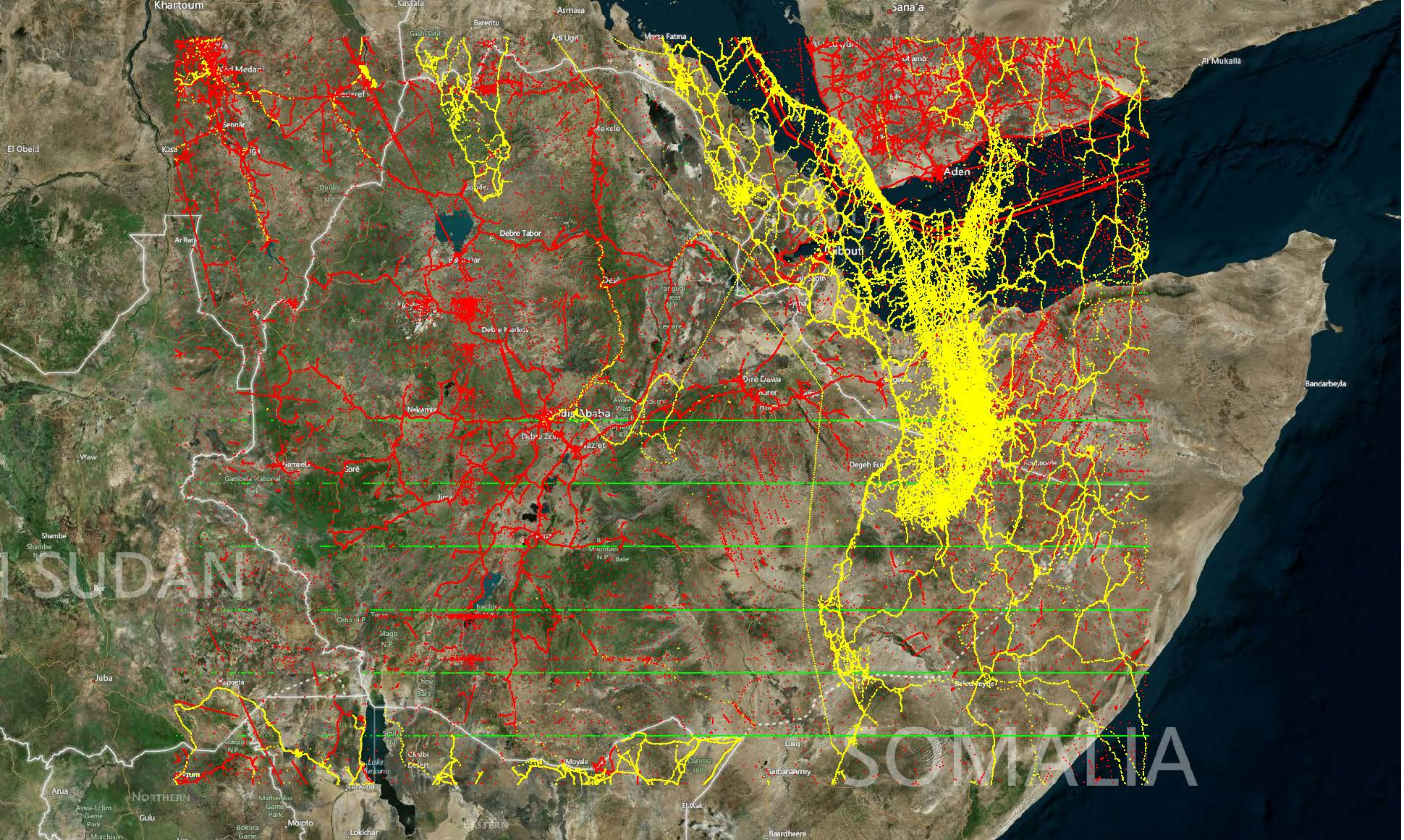
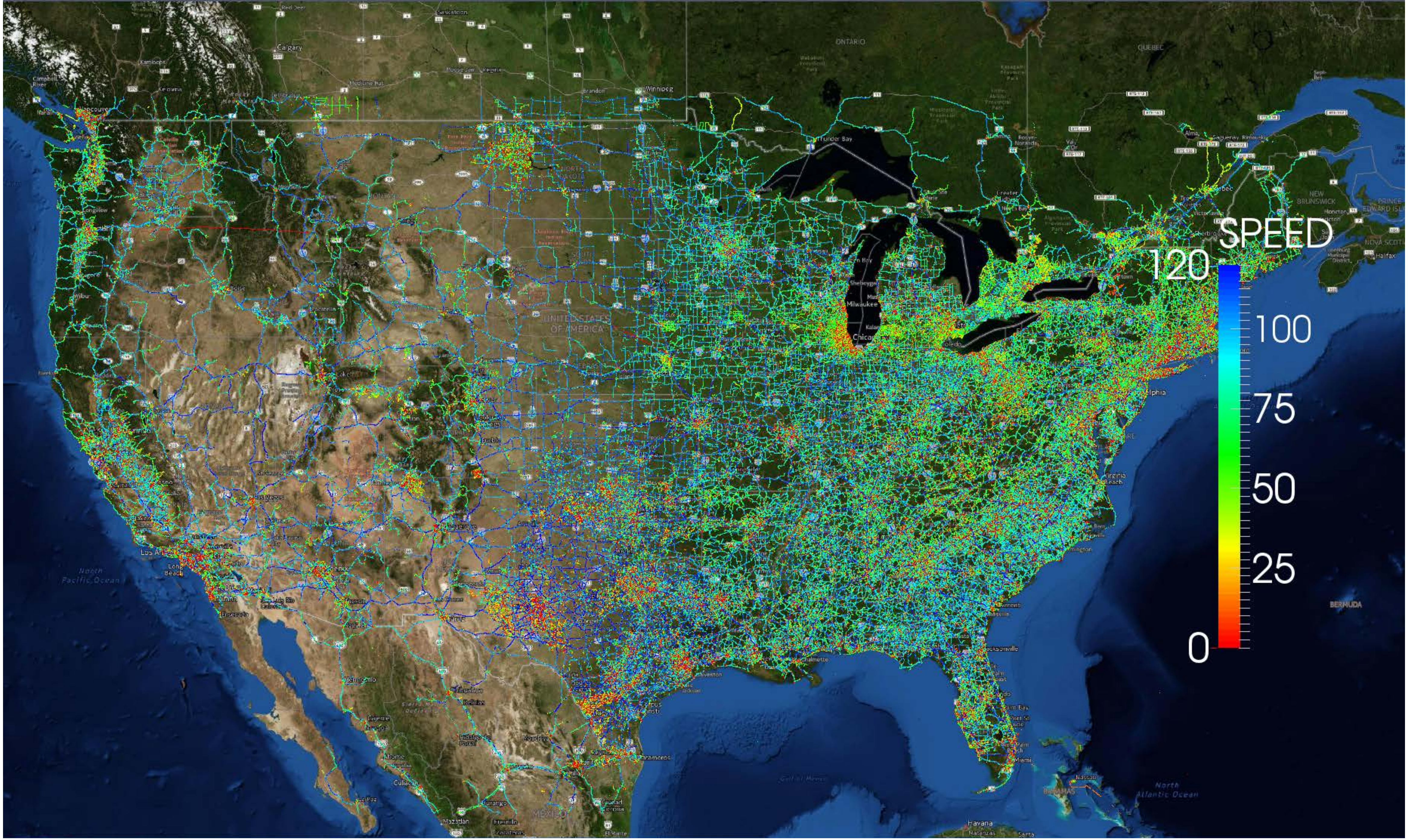


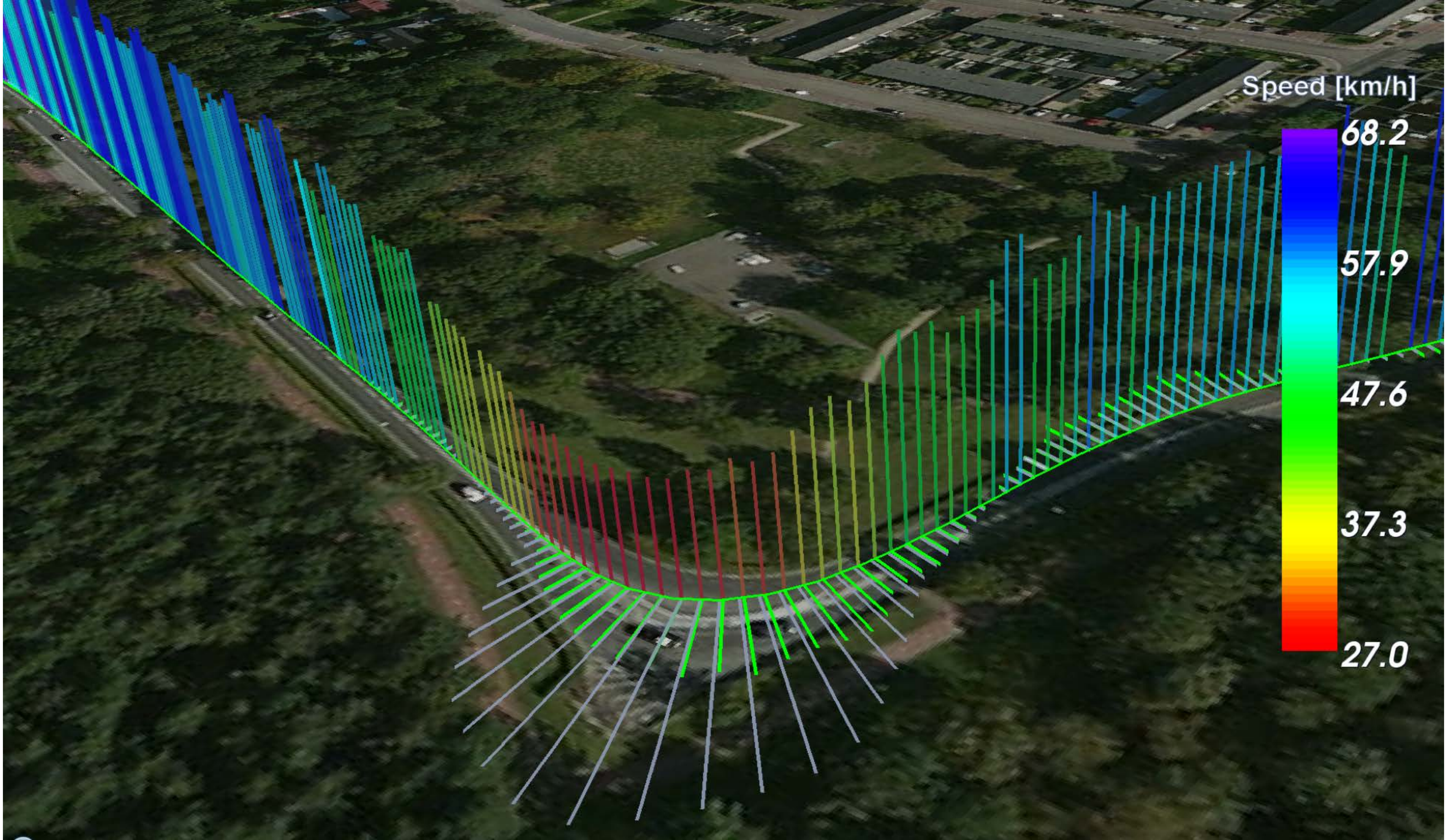
Figure Courtesy
of Here Research

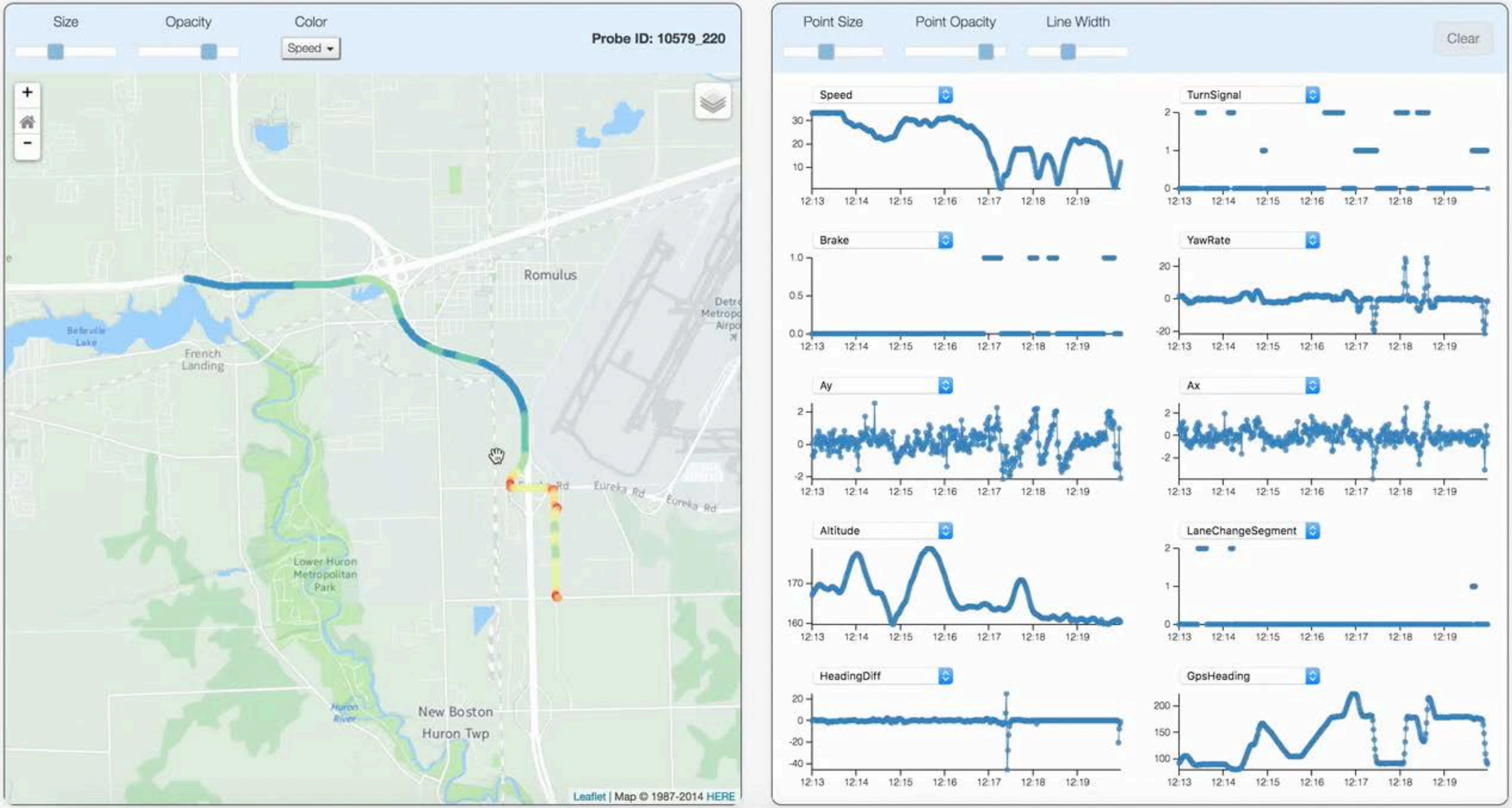

Gorlovo
Skopin
Ukholovo
Shiringushi
Insar
Ryazhsk
Sarai
Vadinsk
Pashkovo
Chernava
E119
Nizhniy Lomov
Zemetchino
Verkhniy Lomov
Kurkino
Aleksandro-Nevskiy
Morshansk
Pleš
E30
M5
M6
Pachelma
Dankov
Chaplygin
Pervomayskiy
Sosnovka
Bashmakovo
Kamenka
Lebedyan'
Poim
Belinskiy
Michurinsk
Svishchëvka
Tambovskiy rayon
Rasskazovo
Kirsanov
Lipetsk
Tamala
Gryazi
Bekovo
Serdobsk
Zadonsk
Khvorostyanka
Inzhavino
Makarovo
Rtishchevo
Kulikovo
Dobrinka
Mordovo
Rzhaksa
Usman
Tokarëvka
Uvarovo
Turki
Arkadak
Zemlyansk
E115
Ramon'
Alekseyevka
Romanovka
Semiluki
Voronezh
Kozlovka
Panino
Balashovskiy rayon
Rostoshi
Balashov
Strelitsa
Rogachëvka
Anna


Khartoum
Kassala
Asmara
Sana'a
Barentu
Adi Ugri
Marsa Fatma
Wad Medani
Gedaref
Al Mukalla
El Obeid
Sennar
Kosti
Mekele
Aden
Debre Tabor
Bahir Dar
Djibouti
Ali Sabih
Debre Markos
Dire Dawa
Harer
Nekemte
Addis Ababa
Debre Zeyit
Nazret
Bandarbeyla
Gambela
Gambela National Park
Gore
Jima
Degeh Bur
Wau
Shambe
SUDAN
Bor
Bale
Omo N.P.
Mago N.P.
Juba
Kapoeta
Kidepo N.P.
Lake Turkana
Moyale
Luuq
Garbaharrey
SOMALIA
Arua
NORTHERN
Gulu
Moroto
Lokichar
El Wak
Baardheere

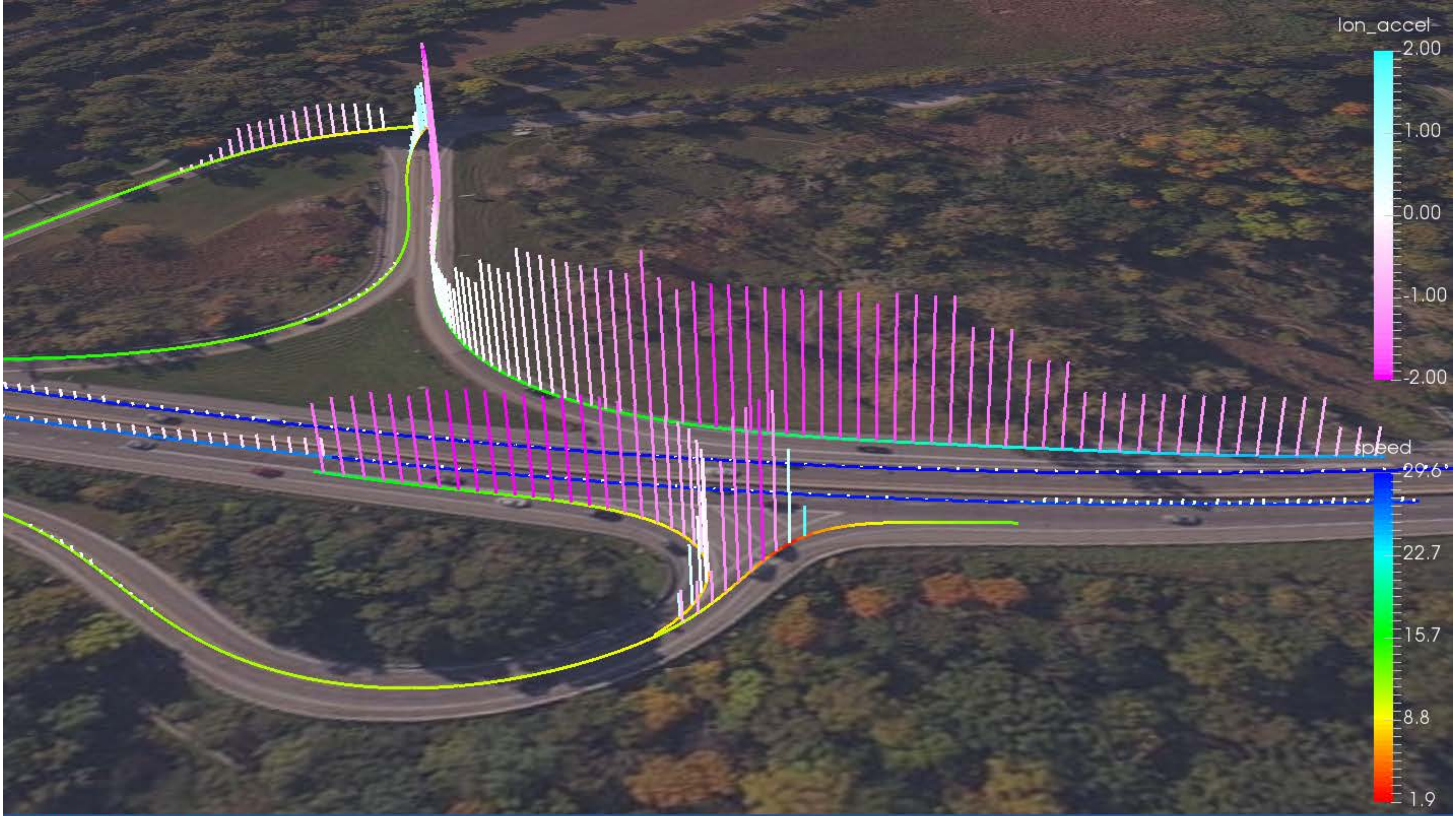

SPEED
120
100
75
50
25
0


Speed (km/h)
200.
150.
100.
50.0
0.000
2013-03-20


Speed [km/h]
68.2
57.9
47.6
37.3
27.0



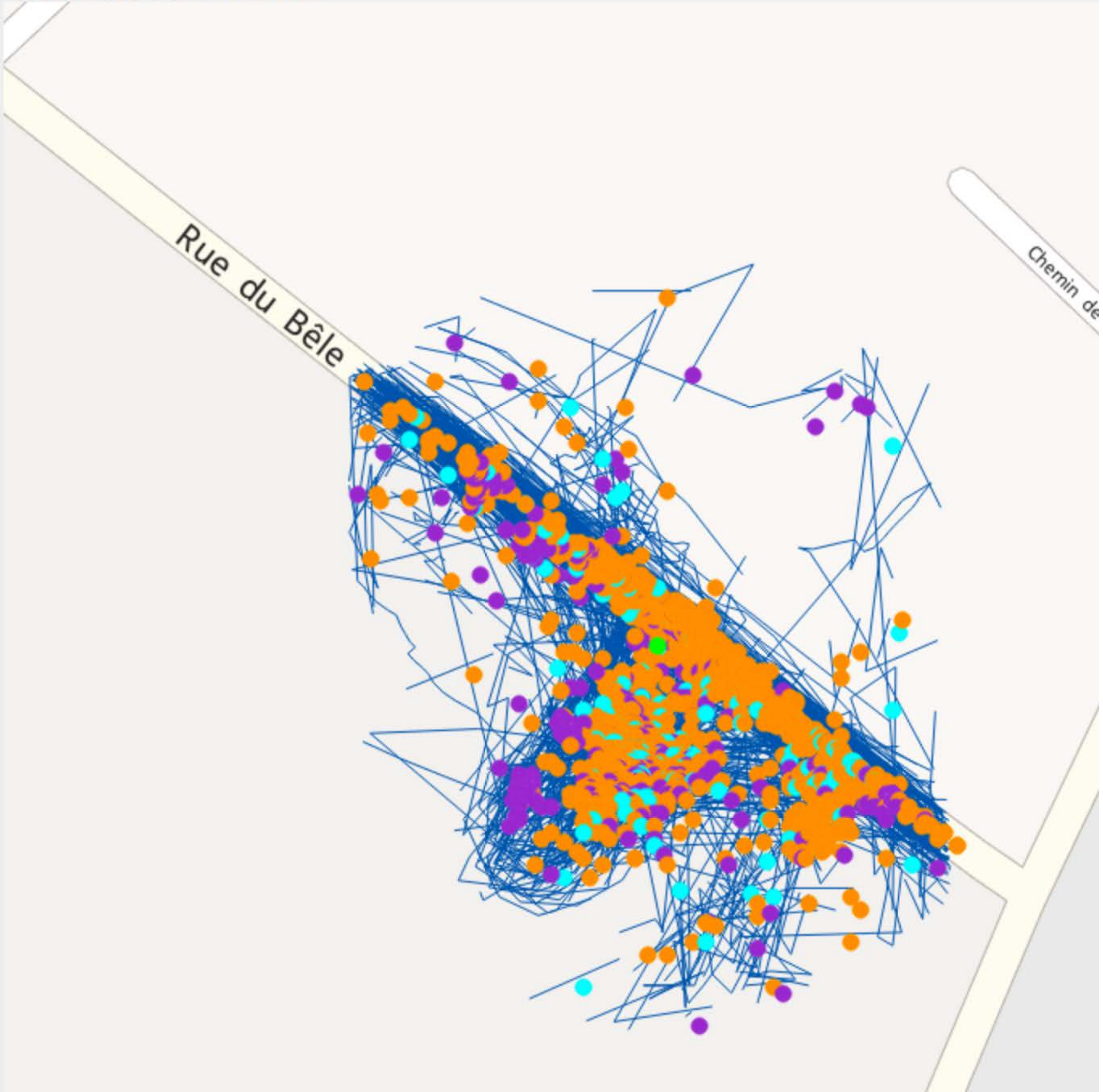
Size
Opacity
Color
Speed
Probe ID: 10579_220
Romulus
Belleville Lake
French Landing
Lower Huron Metropolitan Park
New Boston
Huron Twp
Huron River
Eureka Rd
Leaflet | Map © 1987-2014 HERE
Point Size
Point Opacity
Line Width
Clear
Speed
TurnSignal
Brake
YawRate
Ay
Ax
Altitude
LaneChangeSegment
HeadingDiff
GpsHeading
12:13
12:14
12:15
12:16
12:17
12:18
12:19

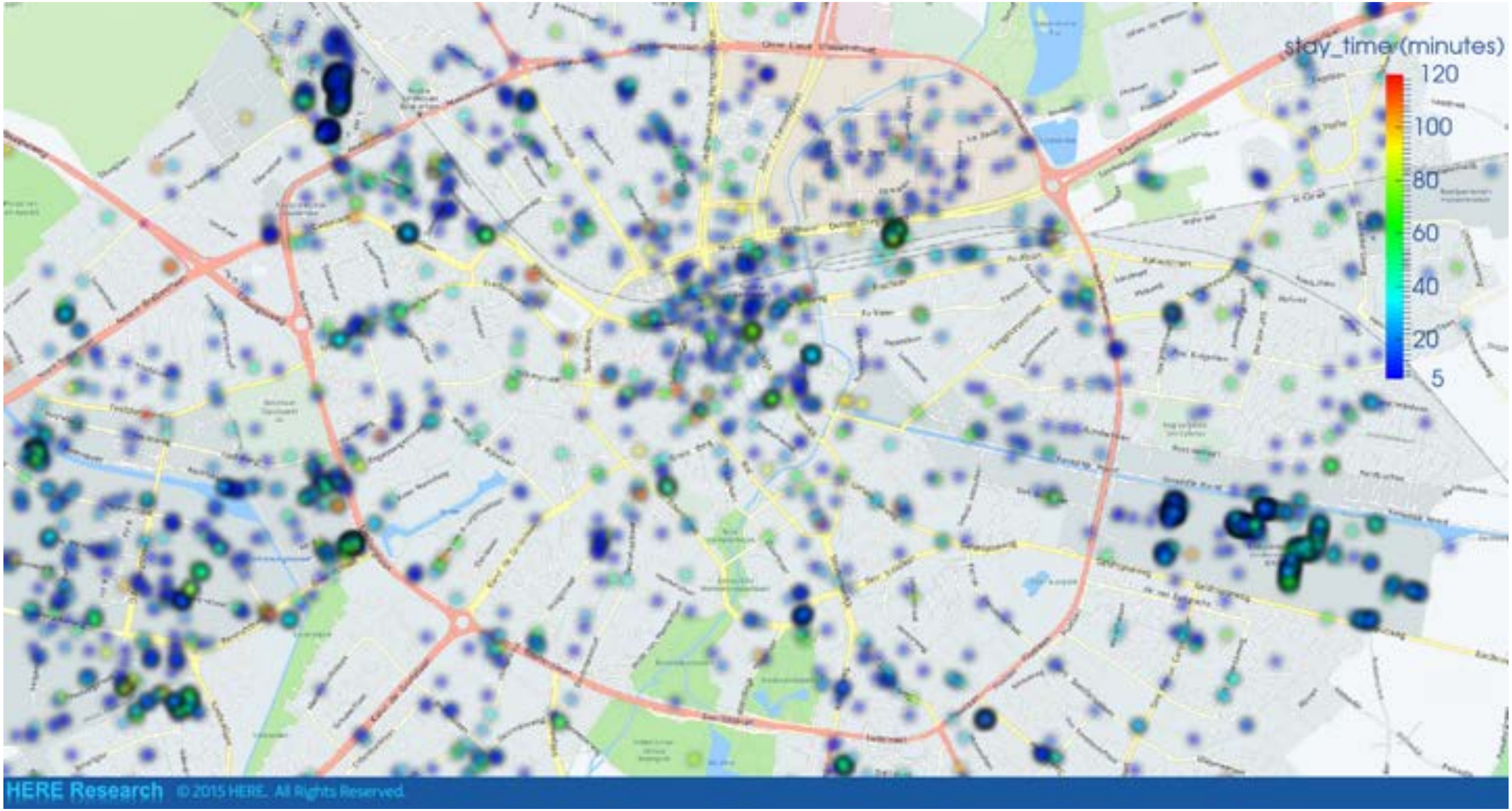

lon_accel
2.00
1.00
0.00
-1.00
-2.00
speed
29.6
22.7
15.7
8.8
1.9


speed
34
22
14
7.5


Eindhoven

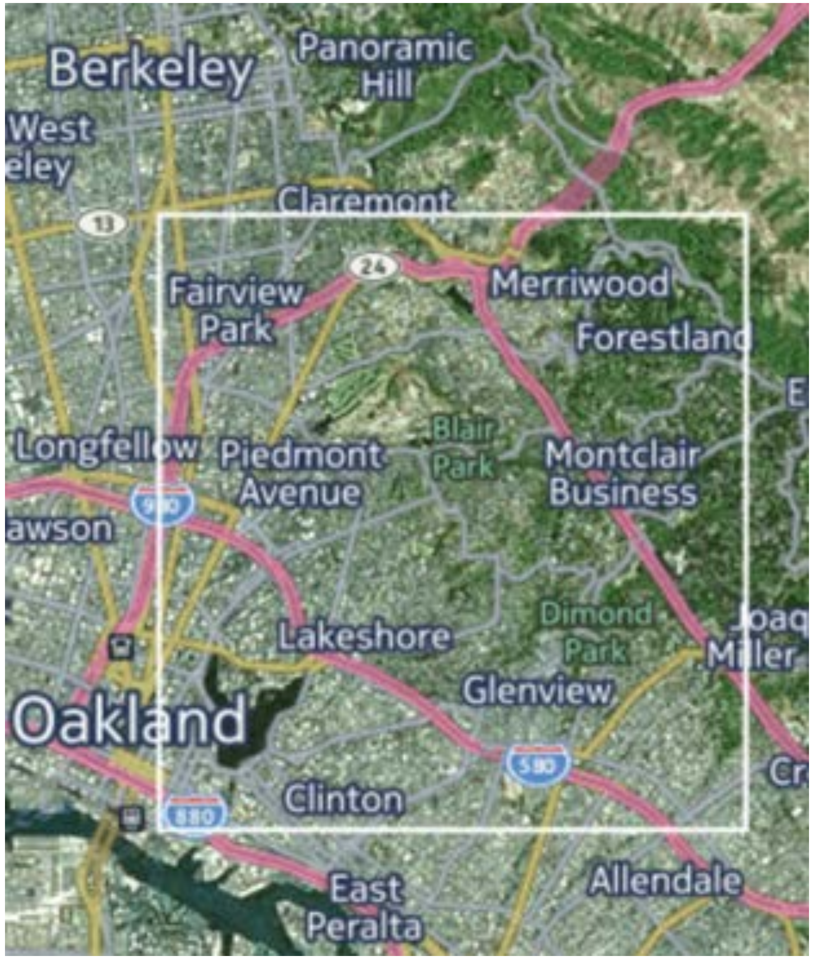
Petrol Station 2: EXIST

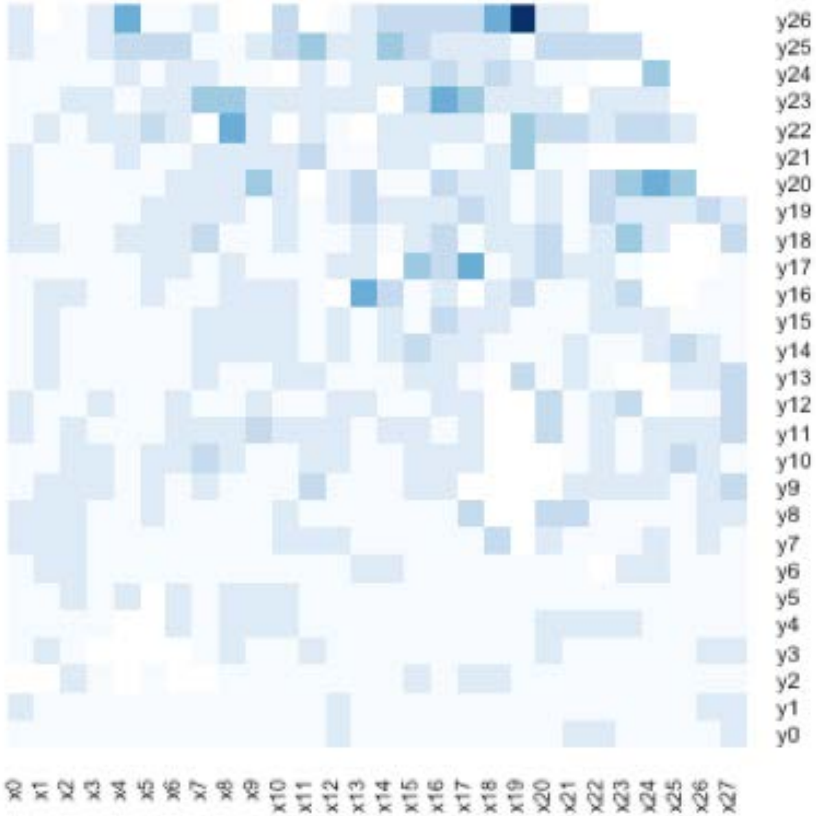



stay_time (minutes)
120
100
80
60
40
20
5
HERE Research
© 2015 HERE. All Rights Reserved.

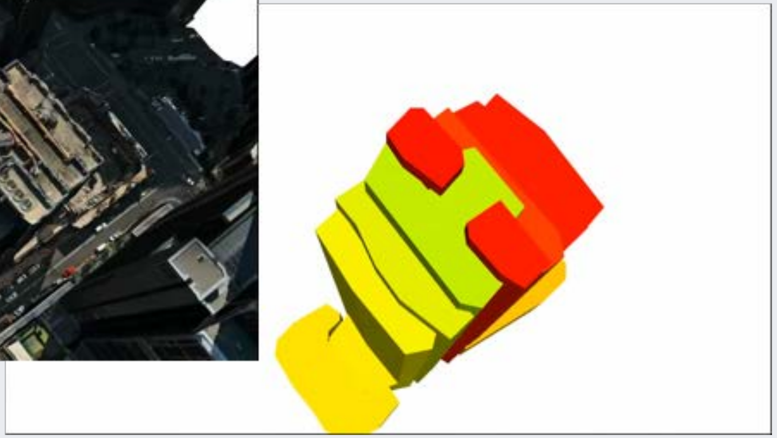
Map matching
to a “known”
representation of the
world

Images courtesy of HERE Research



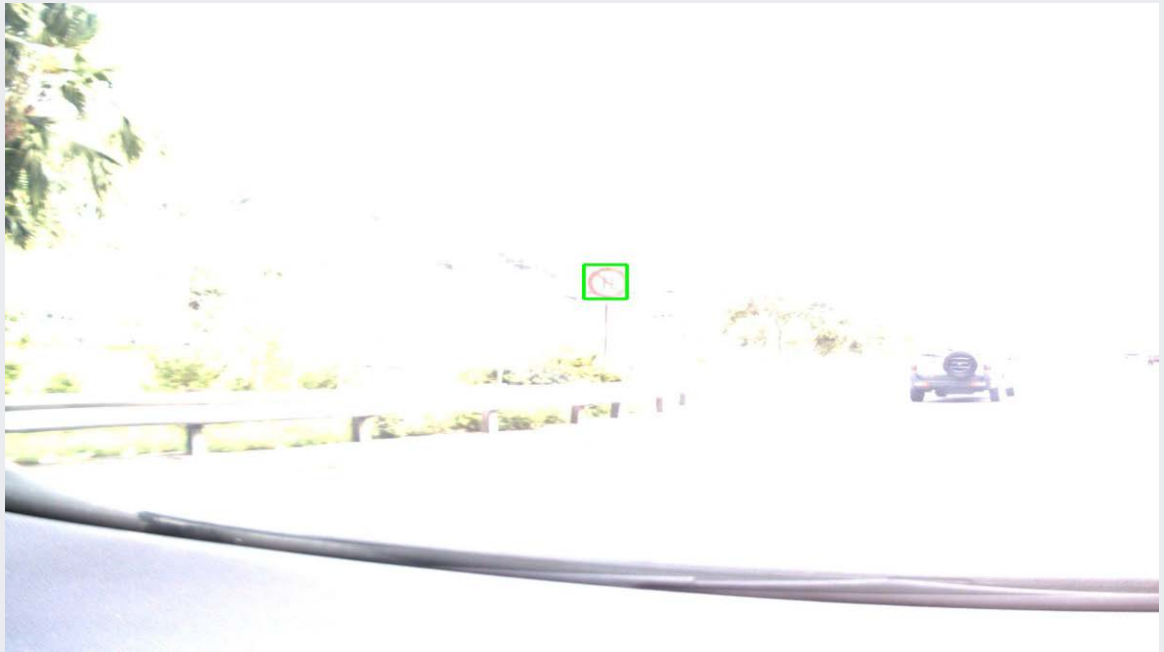


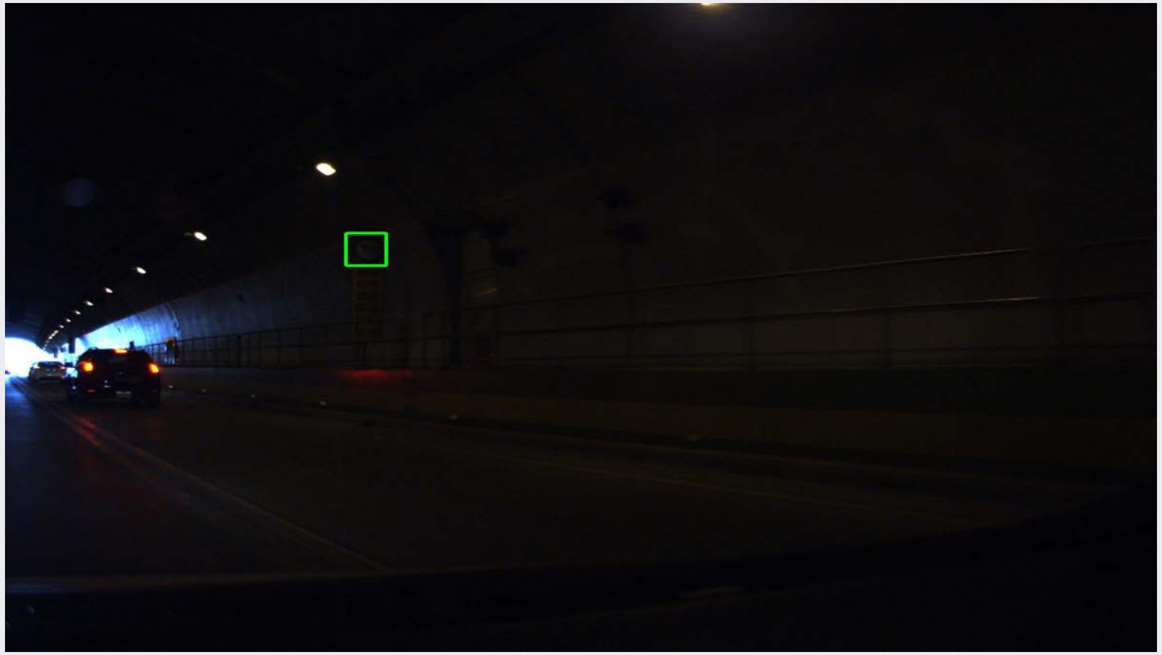
Does the big data have to be big?



Veracity

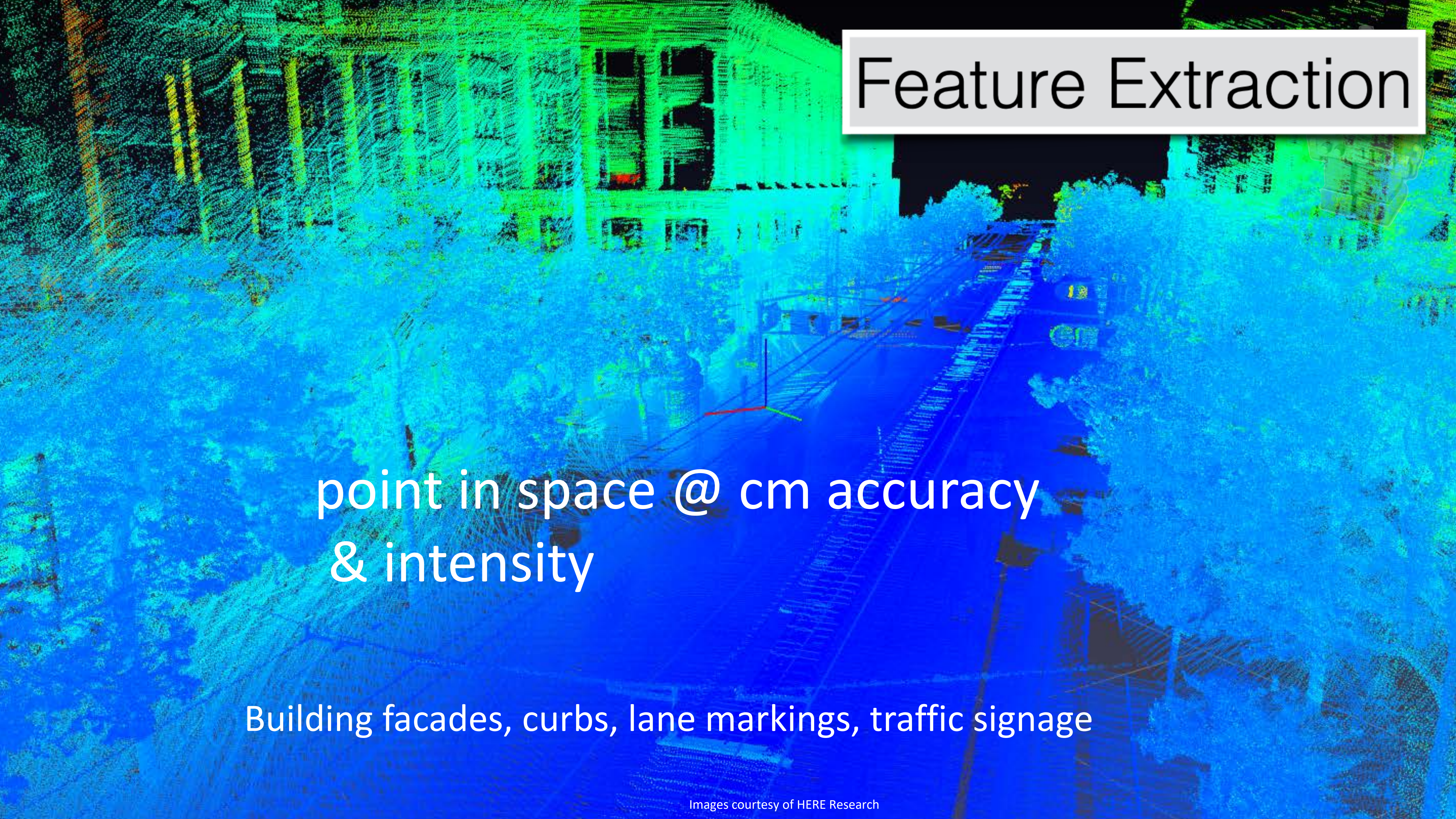
image analytics





Images courtesy of HERE Research

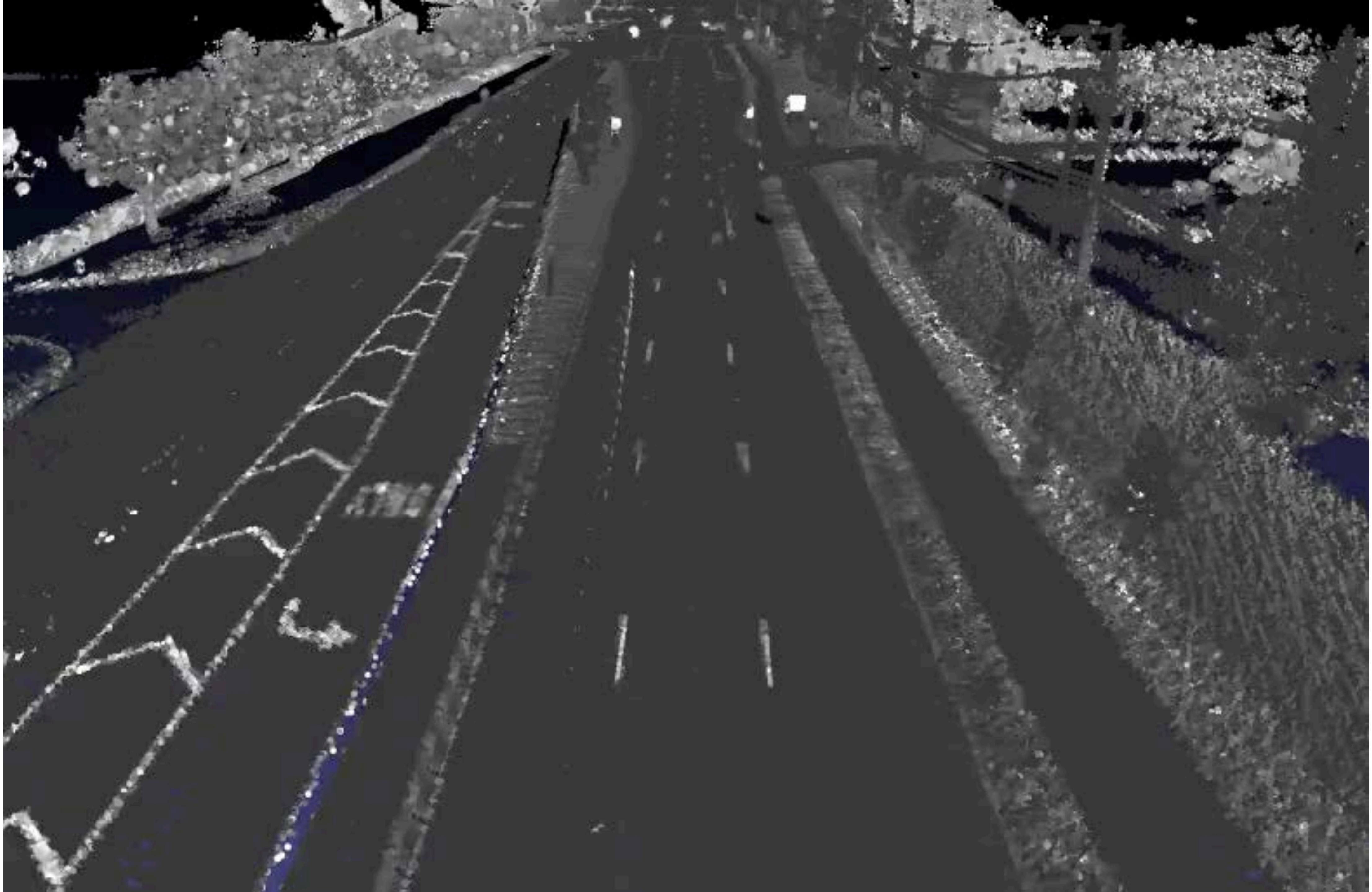
Feature Extraction

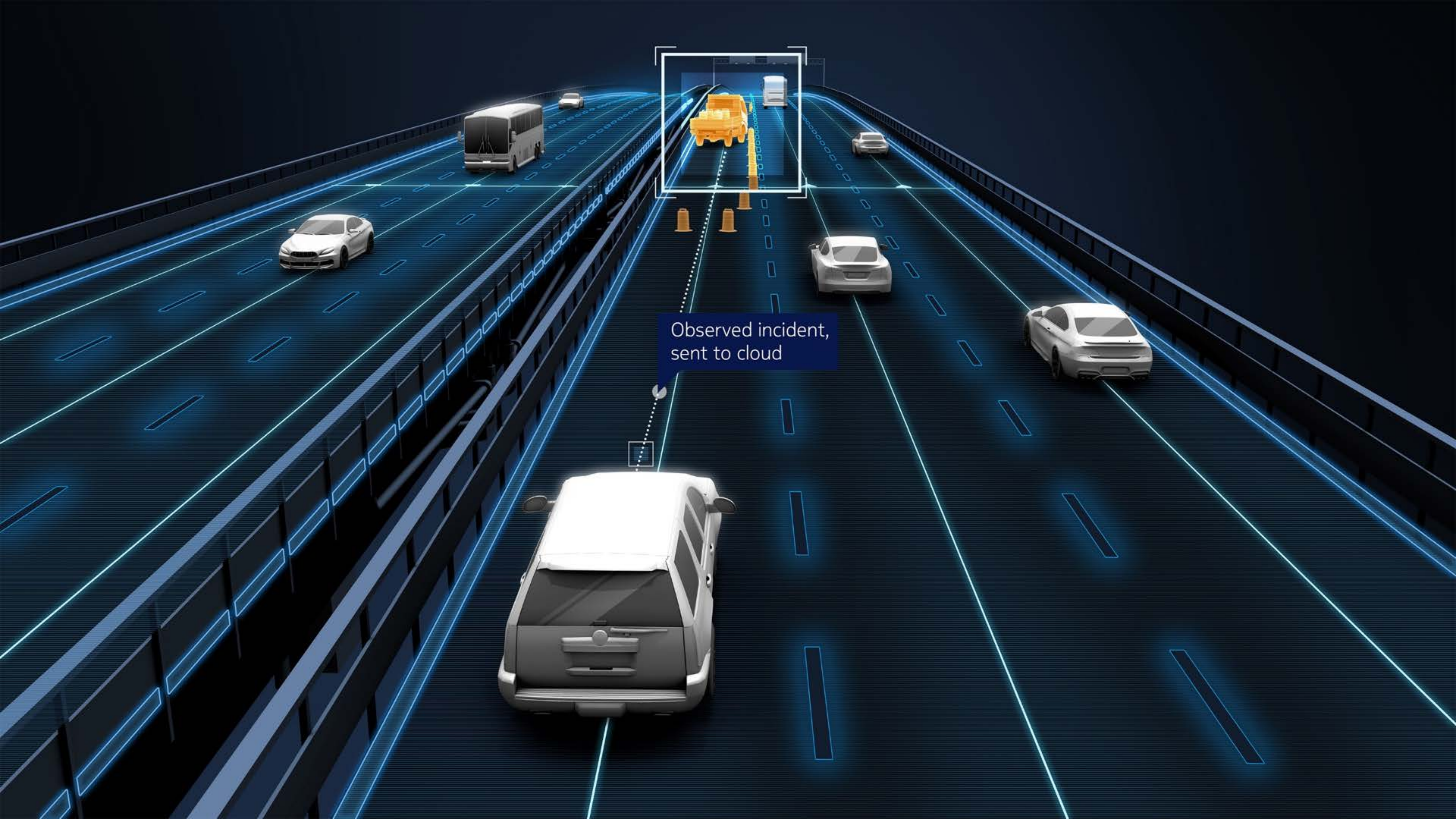


point in space @ cm accuracy
& intensity

Building facades, curbs, lane markings, traffic signage

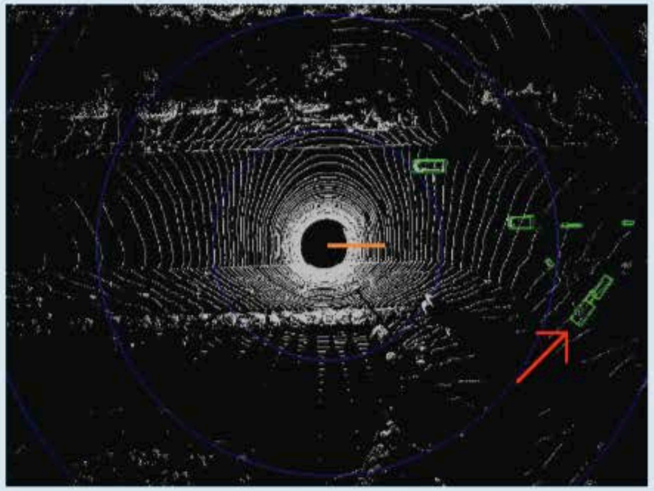
Images courtesy of HERE Research



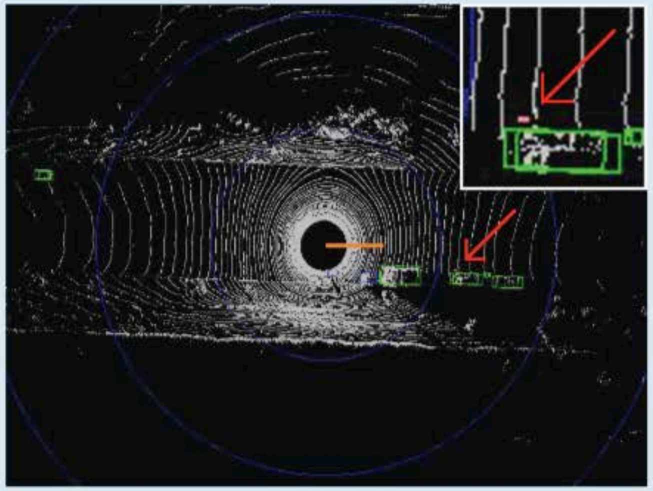

Observed incident,
sent to cloud


billions of investment
disruption

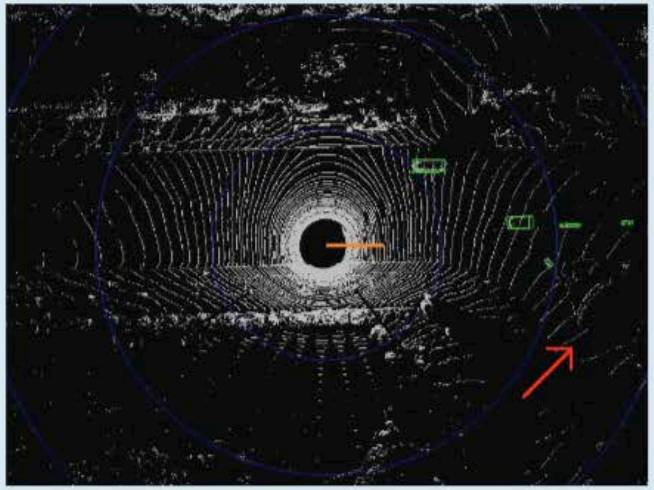
Figure 4. MT detected real-life fatal errors in LiDAR point-cloud data interpretation in the Apollo "perception" module: three missing cars and one missing pedestrian.



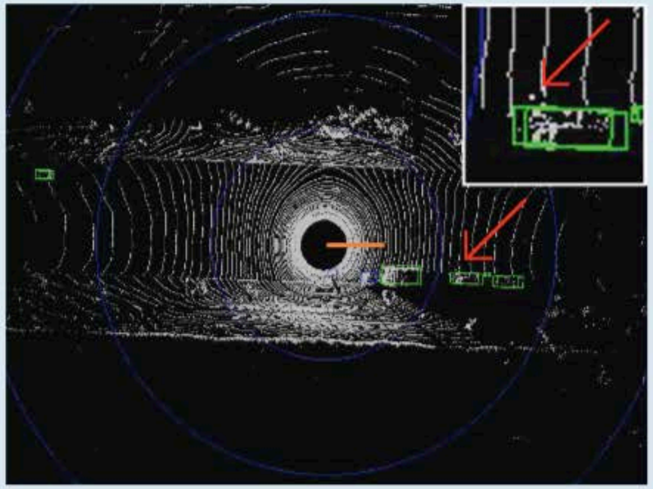
(a) Original: 101,676 LiDAR data points; the green boxes were generated by the Apollo system to represent the detected cars.



(c) Original: 104,251 LiDAR data points; the small pink mark was generated by the Apollo system to represent a detected pedestrian.



(b) After adding 1,000 random data points outside the ROI, the three cars inside the ROI could no longer be detected.

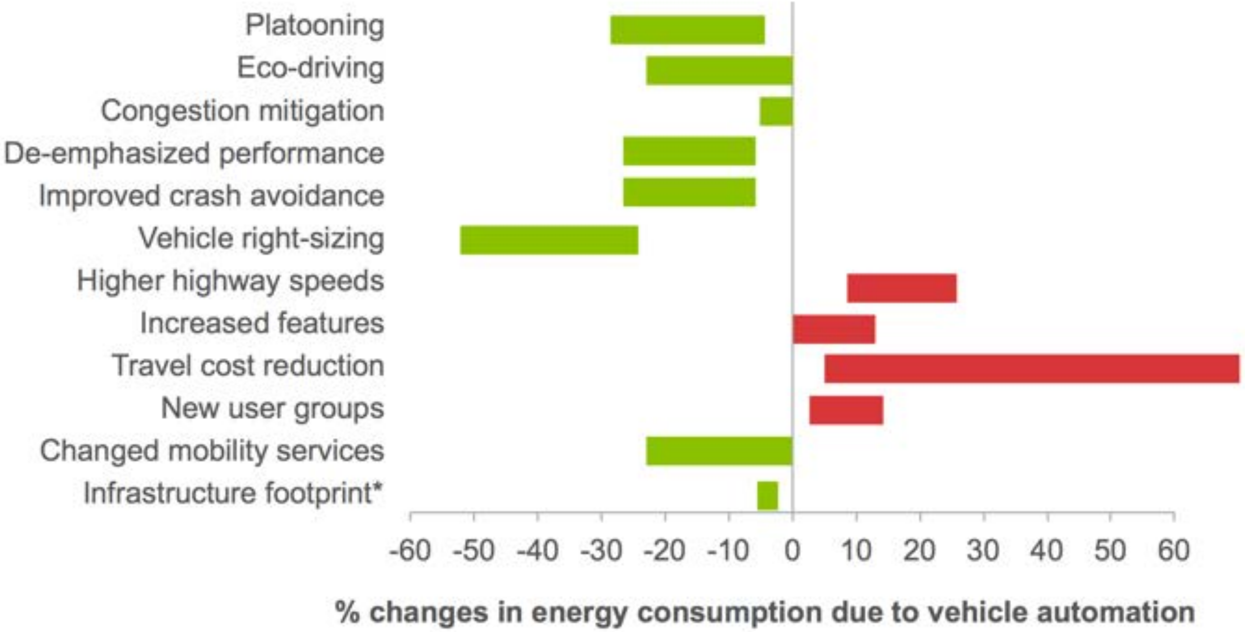


(d) After adding only 10 random data points outside the ROI, the pedestrian inside the ROI could no longer be detected.

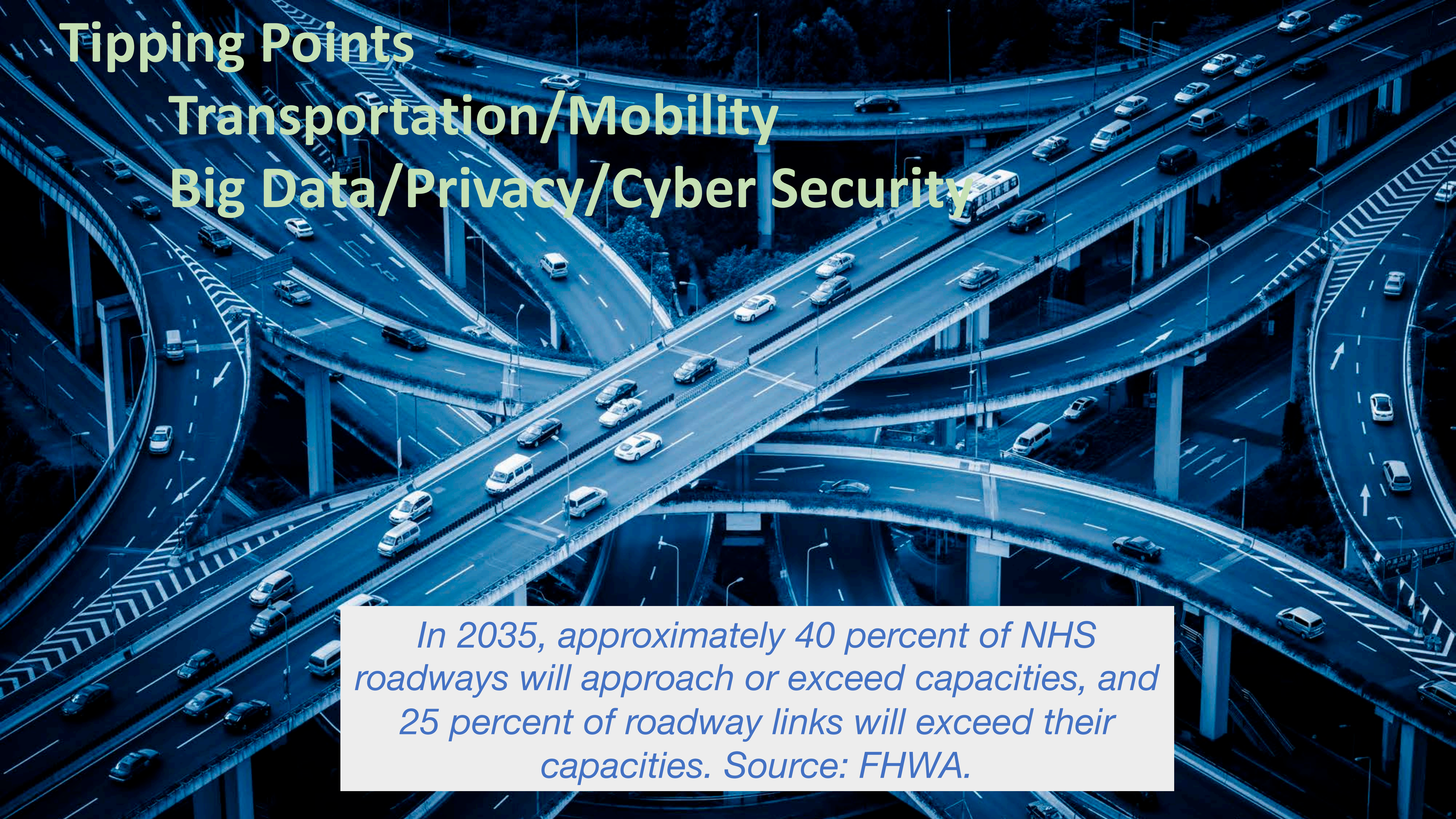
Metamorphic Testing of Driverless Cars, ZHI QUAN ZHOU AND LIQUN SUN

Impact?

Impact?



Data from: http://www.qualenergia.it/sites/default/files/articolo-doc/Automation%20manuscript-accepted.pdf

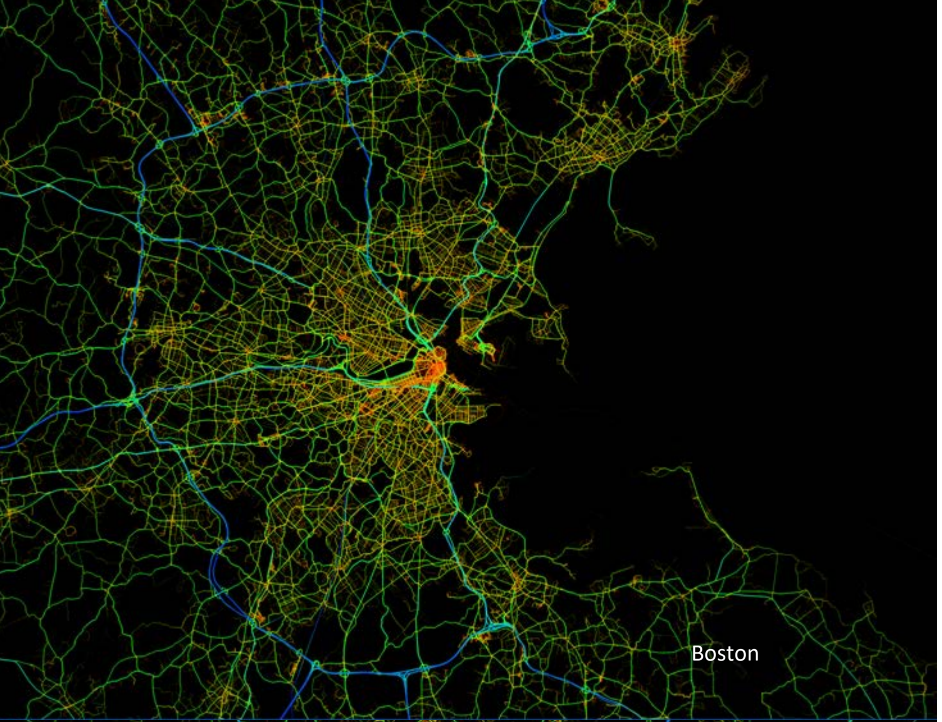


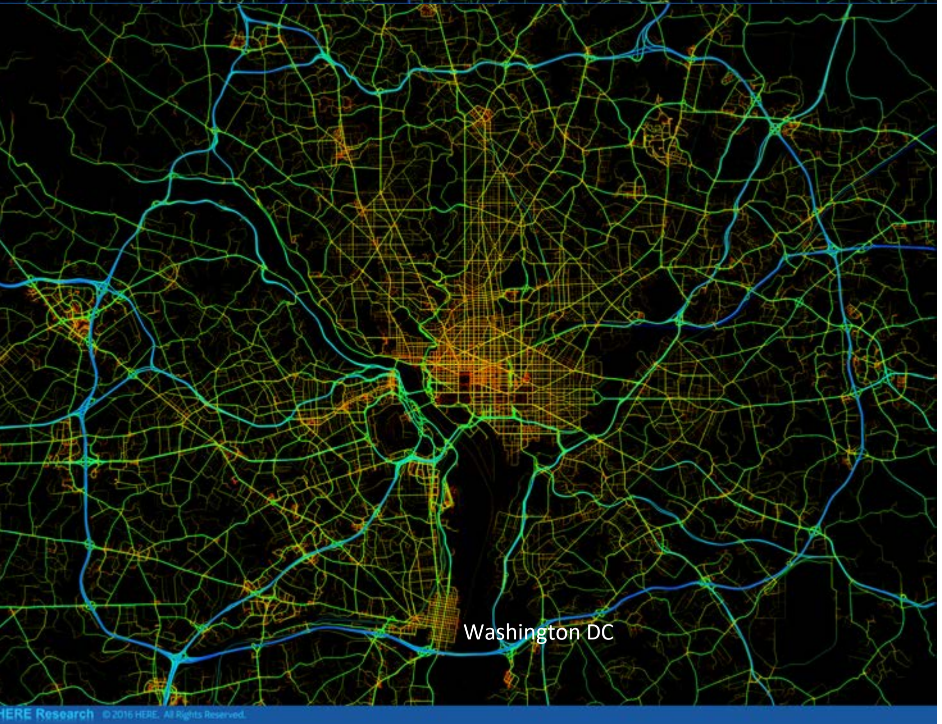
Tipping Points

Transportation/Mobility

Big Data/Privacy/Cyber Security

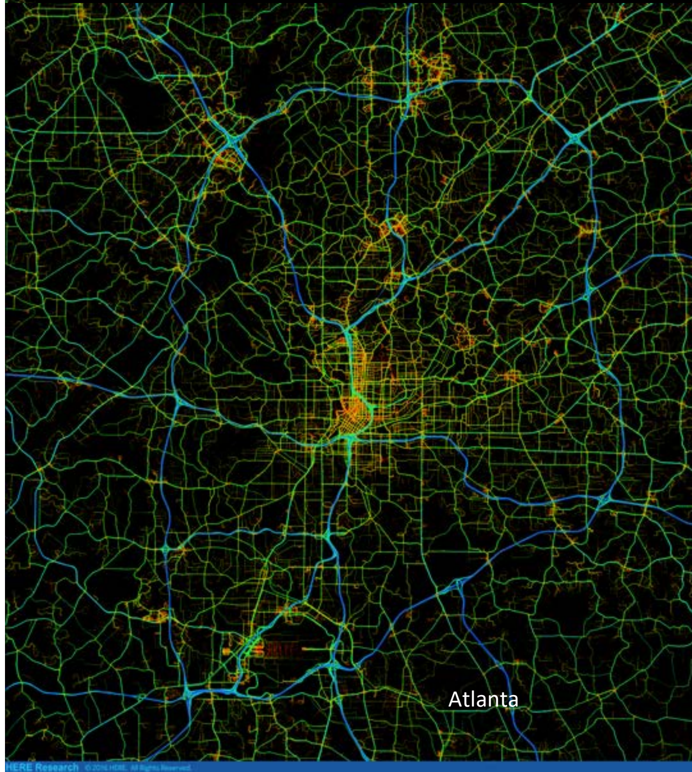
In 2035, approximately 40 percent of NHS roadways will approach or exceed capacities, and 25 percent of roadway links will exceed their capacities. Source: FHWA.





Challenge:

geospatially distributed, temporal data analytics



Images courtesy of HERE Research

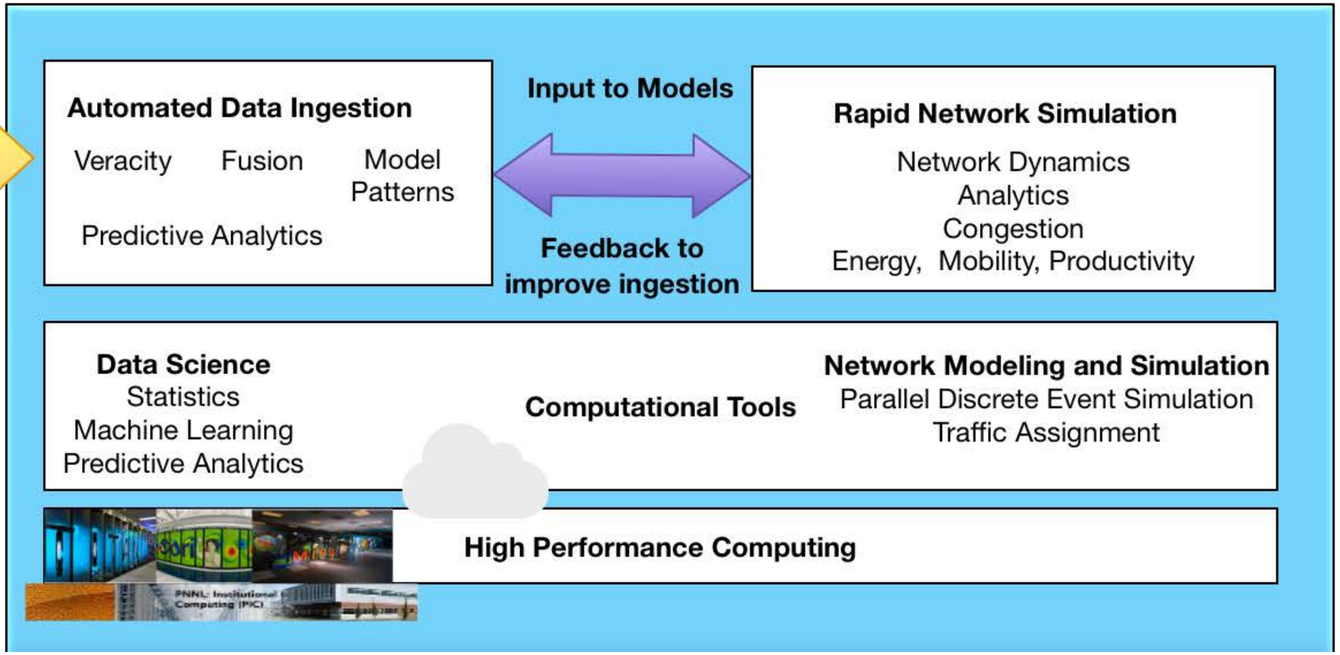
High Performance Computing

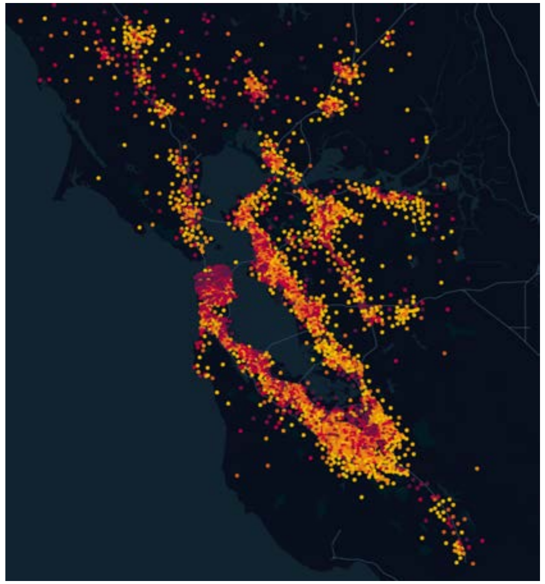
Big Data Solutions for Mobility

Develop high-speed HPC enabled tools that will create actionable control predictions at the network level

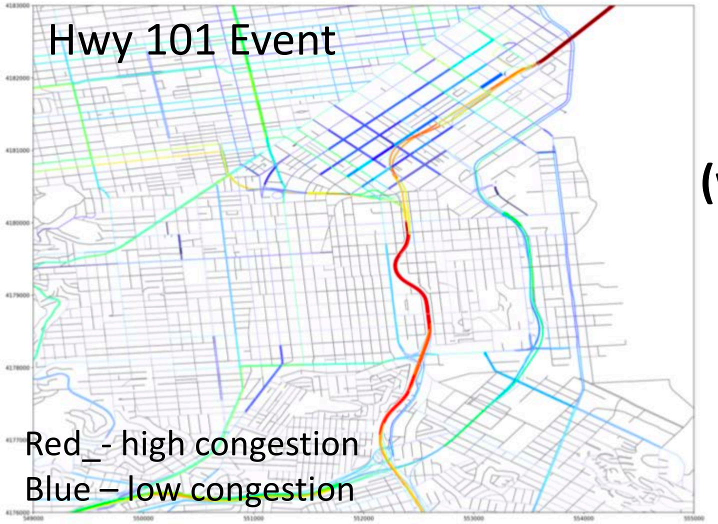
Urban-scale simulation, 22 Million trips with active routing in 3 minutes

Connected Corridor, Los Angeles UCBerkeley CalTrans LA Metro

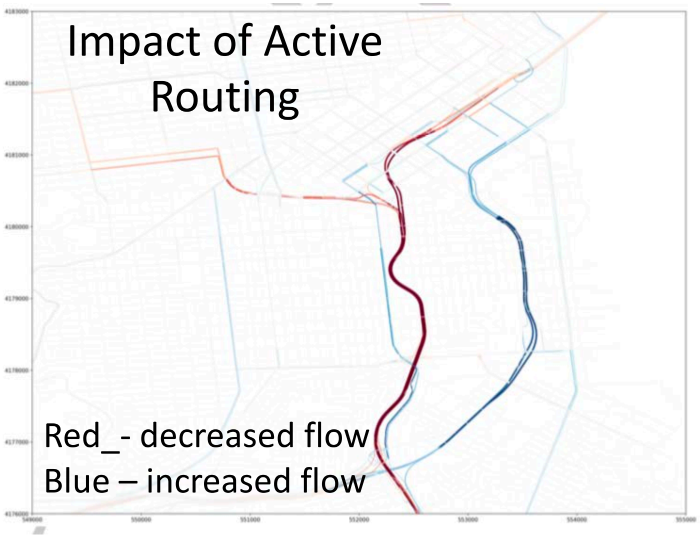




Demand



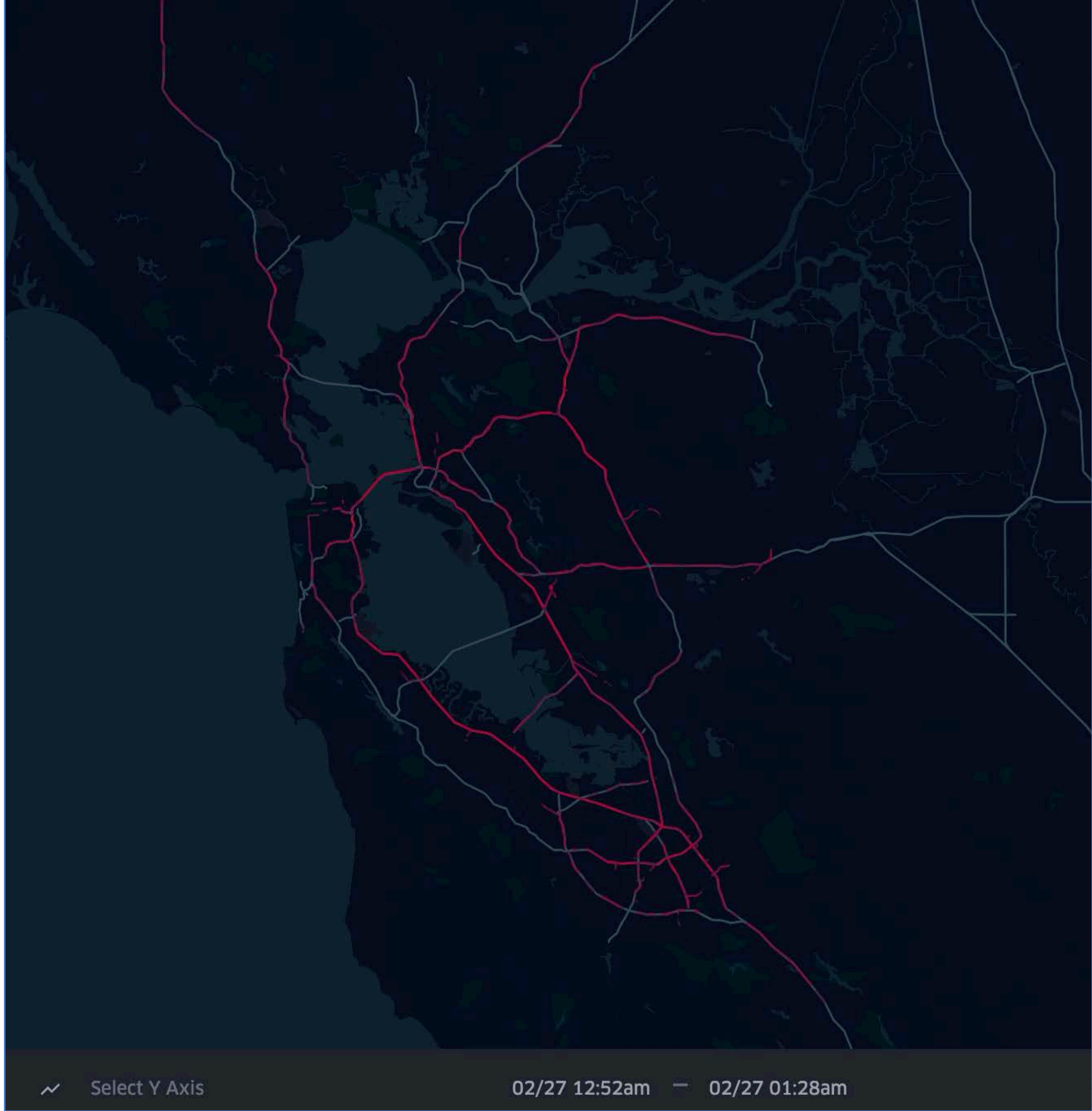
166K travel time hours
64K gallons of fuel
(with 25% vehicle modeled)
At cost of
368K extra vehicle km



Active control requires examination of the dynamics of our cities

- Mobiliti
- LBNL SuperComputer
- 22M trip legs
- ~2M link, 1M node road network
- With dynamic routing
- **3 minute run time**

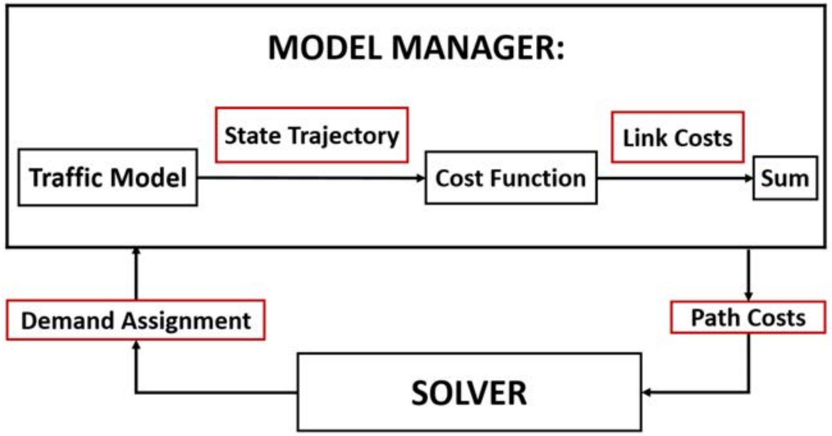
- Surrogate models



High-Performance Computing (HPC) Enabled Computation of Demand Models at Scale to Predict the Energy Impacts of Emerging Mobility Solutions

HPC4 Mobility

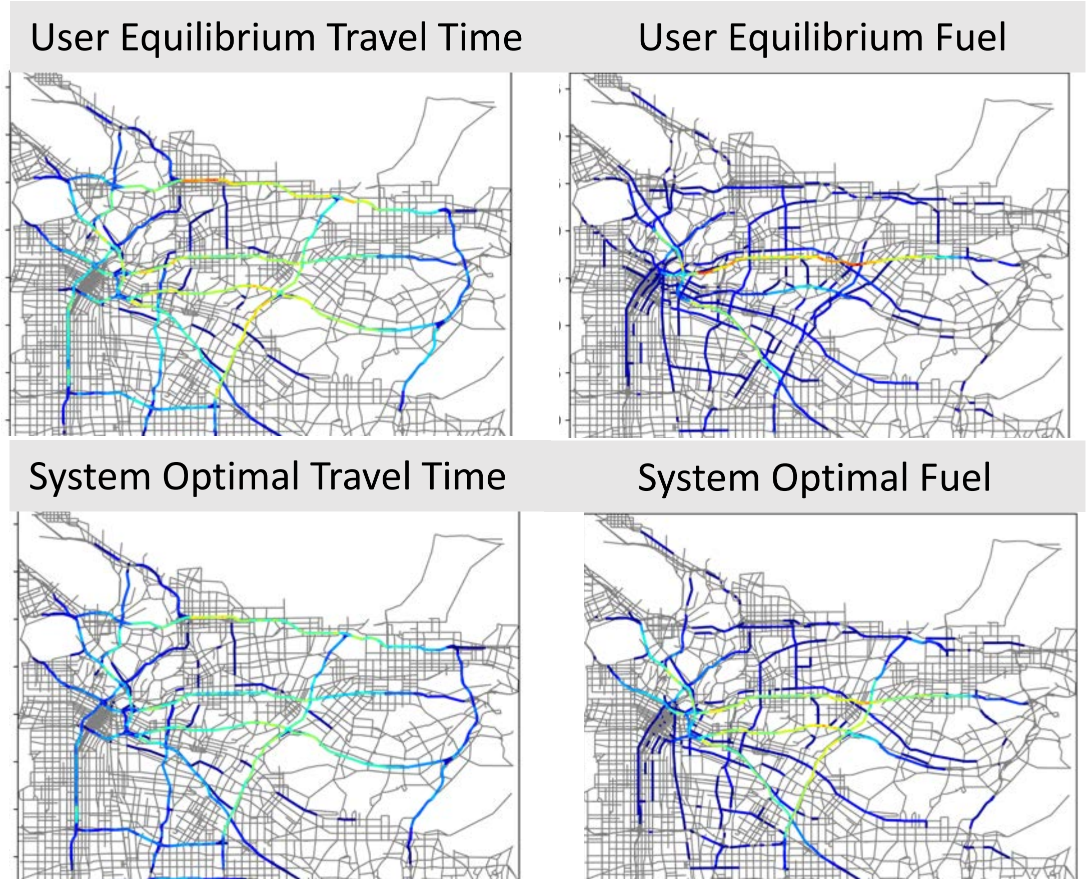
Traffic Assignment :



Optimize Travel Time for User
Optimize Travel Times Systems Level
Optimize Fuel for User
Optimize Fuel System Level

Given the window size=4h, the average leg duration of SOT is 192 s (8.47%) less than the UET case.

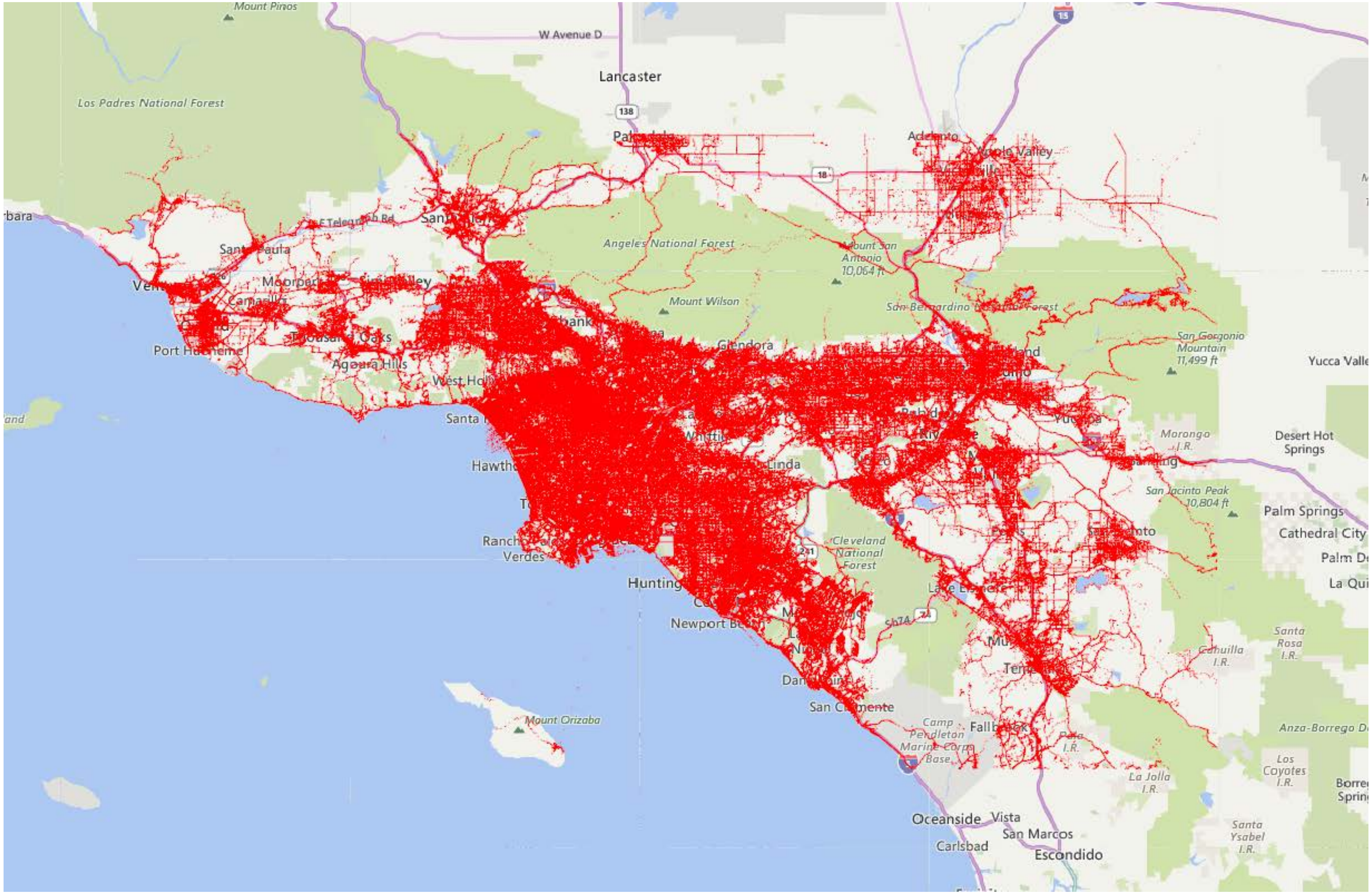
The average leg fuel consumption of SOF is 34 ml (3.59%) less than the UEF.



Compute time 5 hours

Parallelized on 16 nodes of Cori (32 processes x 31 threads per process). assigning 40 million trip legs for 12 (2 hour) time segments

Run time : ~10 minutes


Mount Pinos
W Avenue D
Lancaster
138
Los Padres National Forest
18
Angeles National Forest
Mount San Antonio 10,064 ft
Mount Wilson
Glendora
San Gorgonio Mountain 11,499 ft
Yucca Valle
Agoura Hills
Port Hueneme
Morongo I.R.
Desert Hot Springs
San Jacinto Peak 10,804 ft
Palm Springs
Cathedral City
Linda
Cleveland National Forest
241
74
Newport Be
San Clemente
Camp Pendleton Marine Corps Base
Cahuilla I.R.
Santa Rosa I.R.
Anza-Borrego D
Los Coyotes I.R.
La Jolla I.R.
Mount Orizaba
Oceanside
Vista
San Marcos
Carlsbad
Escondido
Santa Ysabel I.R.
15

Advance https://github.com/doctorjane/advance

- **Advance is a framework for building data transformation pipelines**. Advance allows you to concisely script your data transformation process and to incrementally build and easily debug that process. Each data transformation is a step and the results of each step become the input to the next step.
- **The artifacts of each step are preserved in step named directories**. When the results of a step are not right, just adjust the Advance script, delete the step directory with the bad data and rerun the script. Previously successful steps are skipped so the script moves quickly to the incomplete step. Similarly, when steps fail the results are preserved in directories prefixed with "tmp_". This isolates incomplete step data and ensures that the step is re-processed when the problem is resolved.
- Your project utilizing Advance contains, which we will call "your Advance script." **a primary ruby script that imports Advance and includes your data transformation steps**Each step describes a command to be run on your data. These **commands can be one of the prepackaged Advance scripts, unix commands (like split, cut, etc), or scripts/commands that you create in whatever language is convenient for you**. Advance invokes these scripts one by one much like you would at the command line. Advance logs the exact command that is invoked so that you can run it yourself to check the output manually and to debug failures.

Steps in Advance

- Advance steps are composed of a step processing type function, followed by a slug for the step, followed by the command or script. For example:

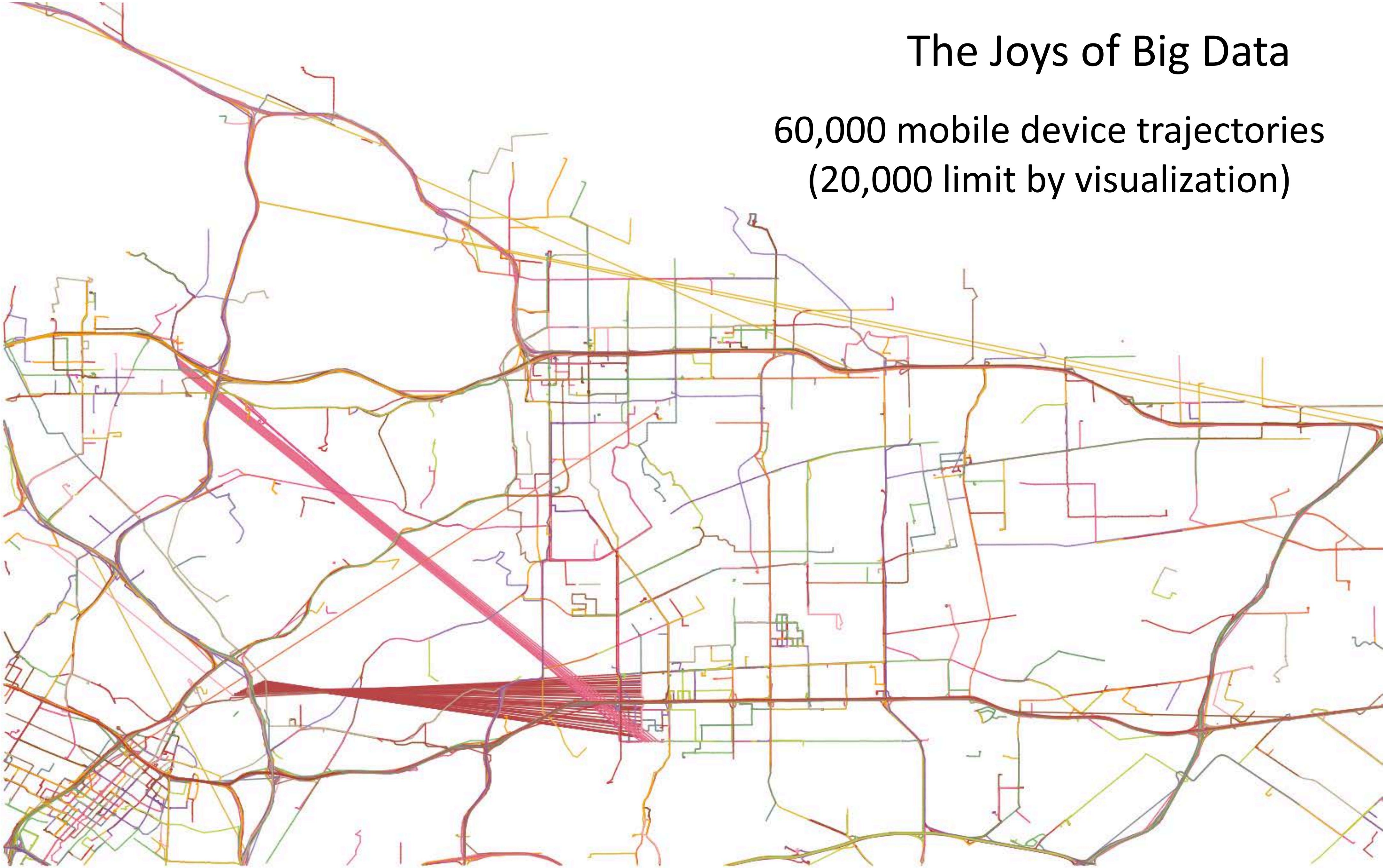
single :unzip_7z_raw_data_file, "7z x {previous_file}"

single :split_files, "split -l 10000 -a 3 {previous_file} gps_data_"

multi :add_local_time, "cat {file_path} | add_local_time.rb timestamp local_time US/Pacific > {file}" # ...

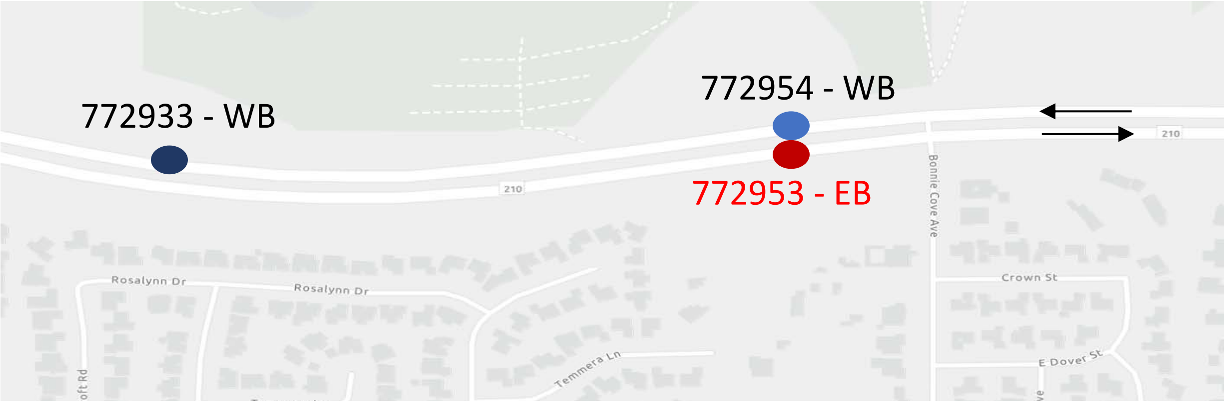
The Joys of Big Data

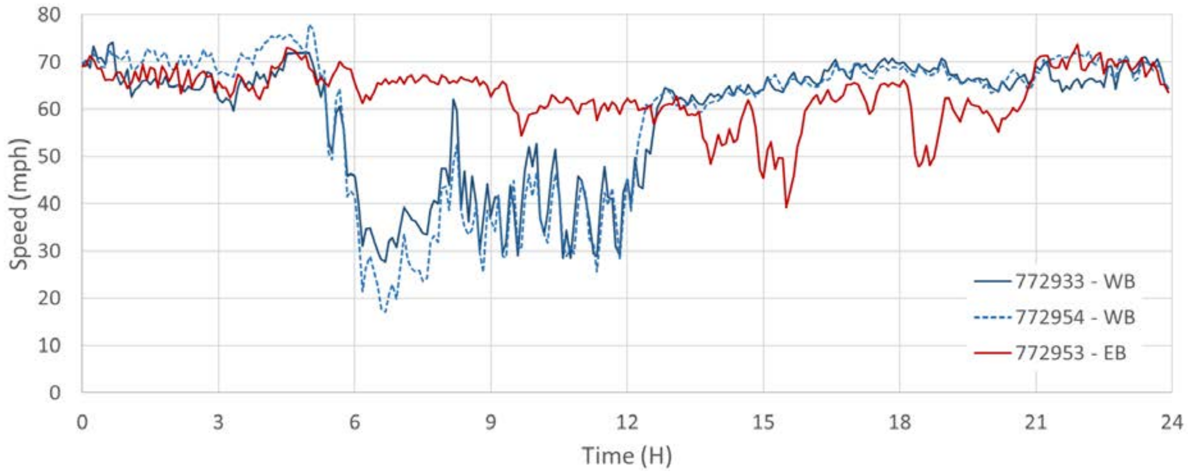
60,000 mobile device trajectories
(20,000 limit by visualization)



Challenges with Sensor Data Modeling

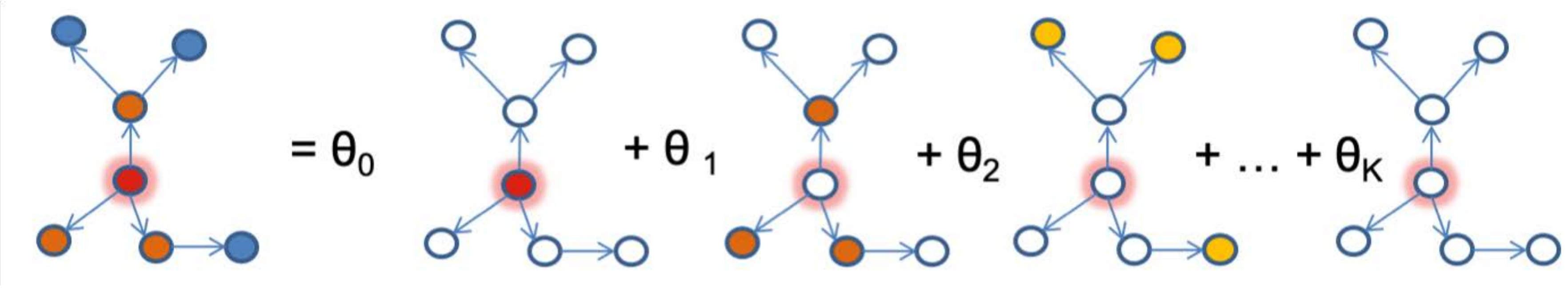
PeMS Data : Inductive loop sensors in major highways

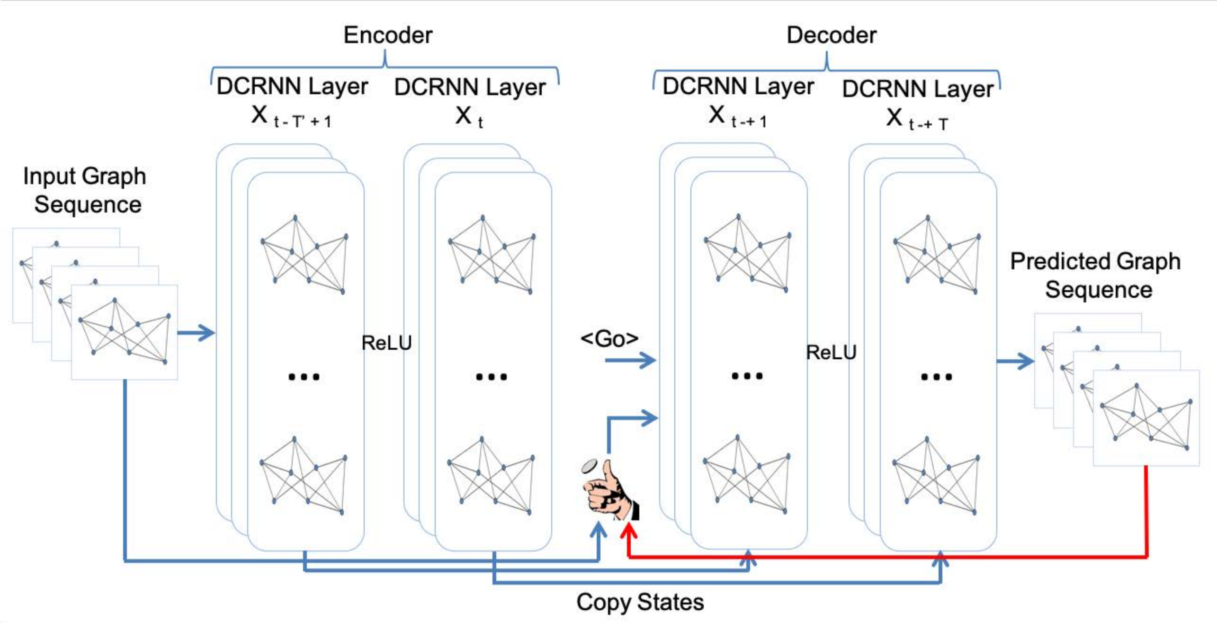


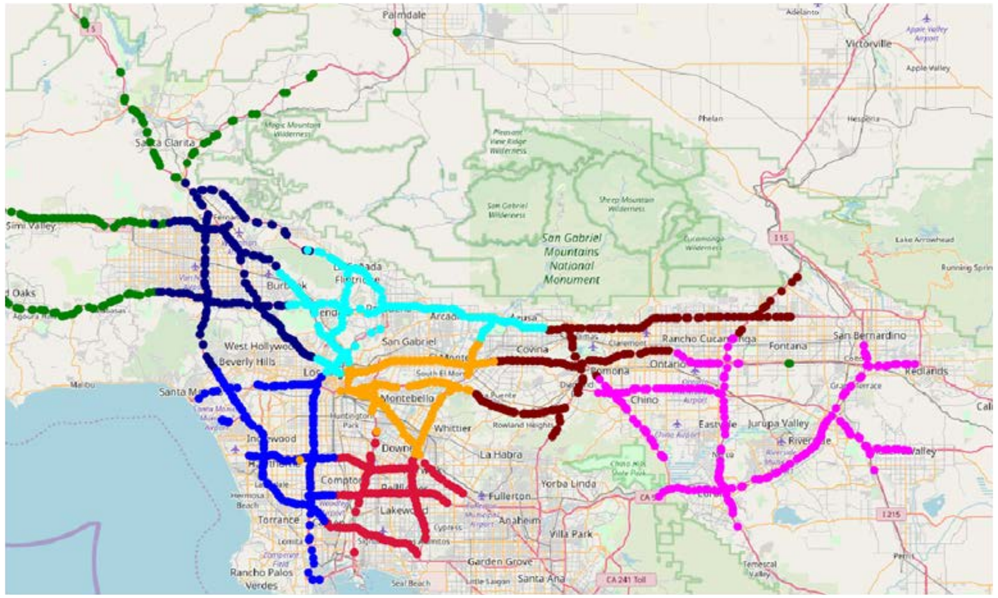


- Complex spatial dependency
- Non-stationary temporal dynamics
- Non-Euclidean spatial geometry
- Modelling each sensor independently fails to capture the spatial correlation

Forecasting Vehicle Dynamics Using DCRNN

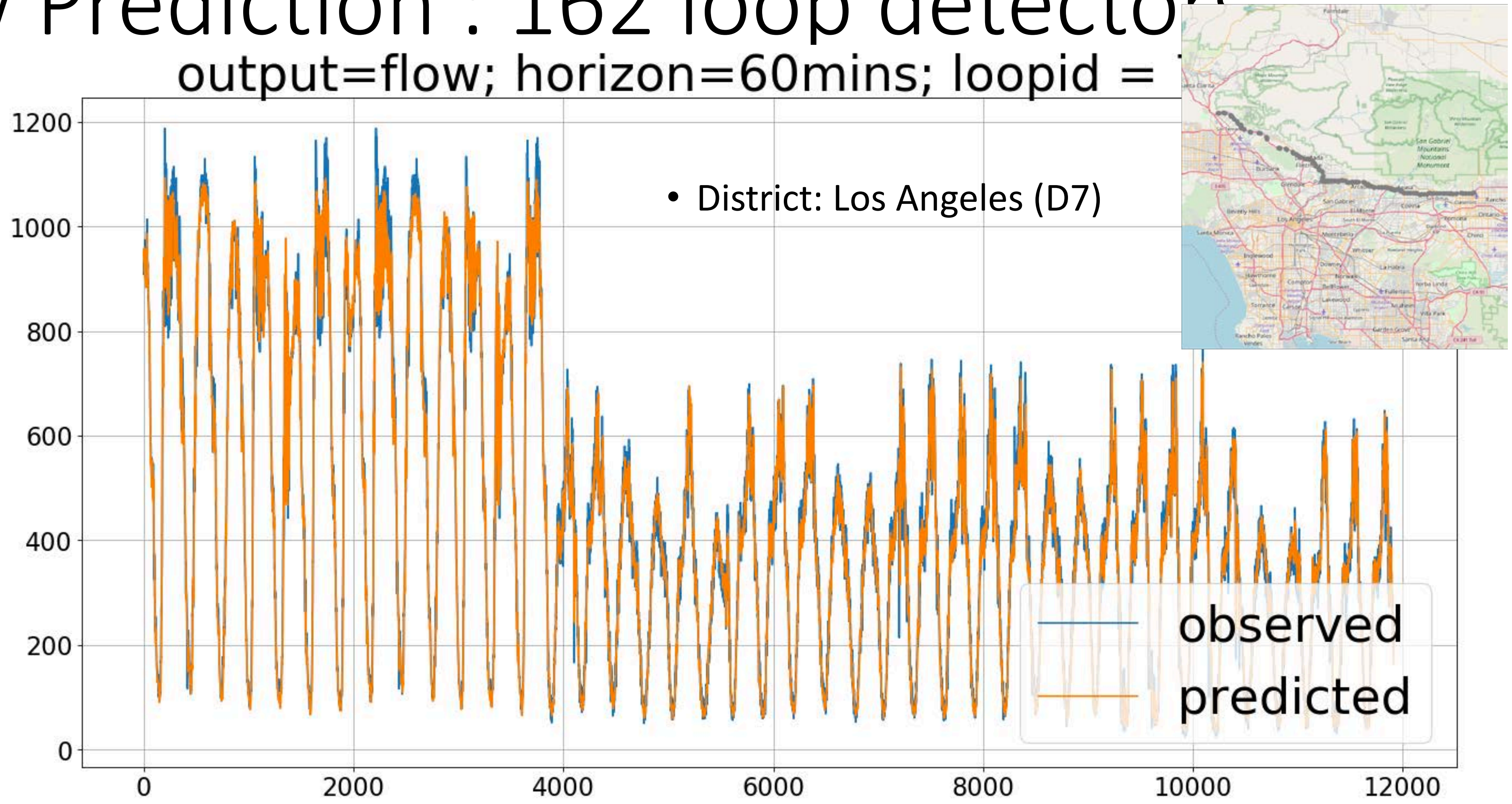






Combining the Diffusion Convolution with a Recurrent Neural Network into a Diffusion Convolutional Recurrent Neural Network (DCRNN) allows for predicting speeds and flows from inductive loop sensors.

Flow Prediction : 162 loop detectors

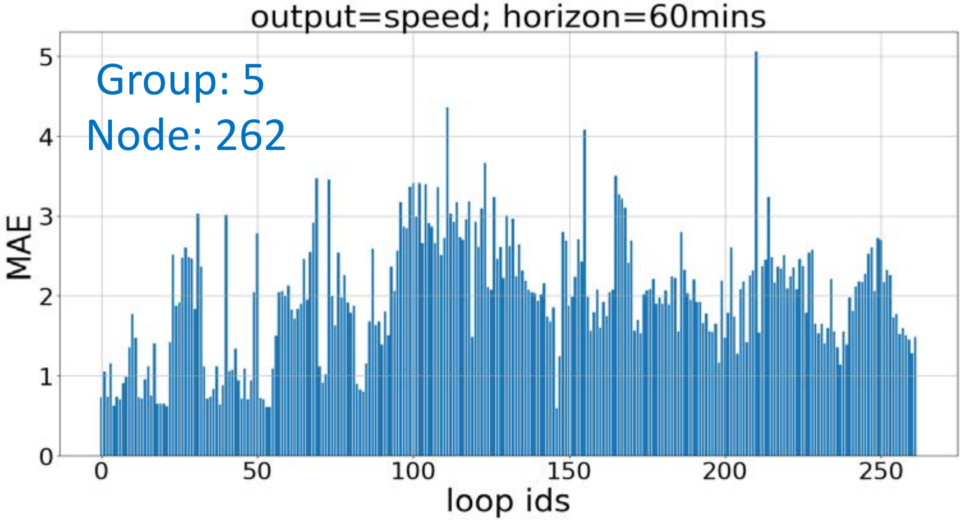


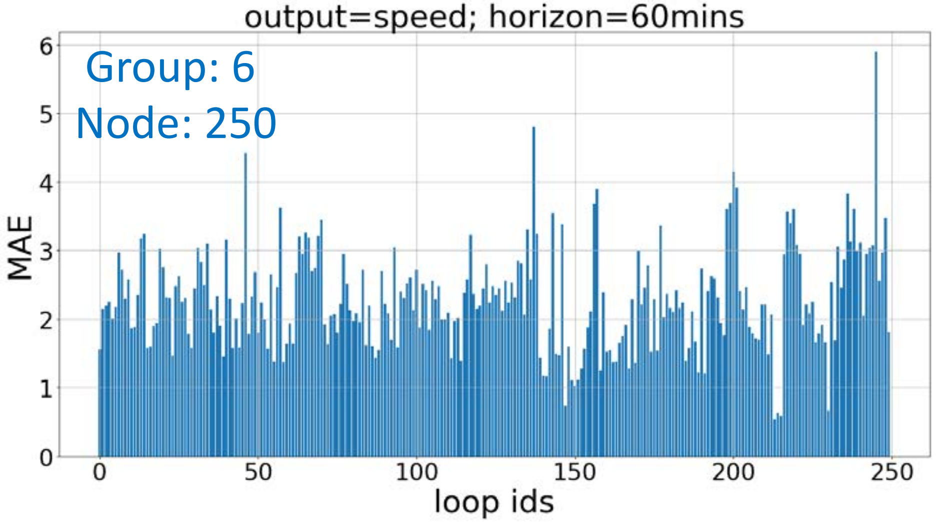
- District: Los Angeles (D7)

DCRNN tracks the real-world flows

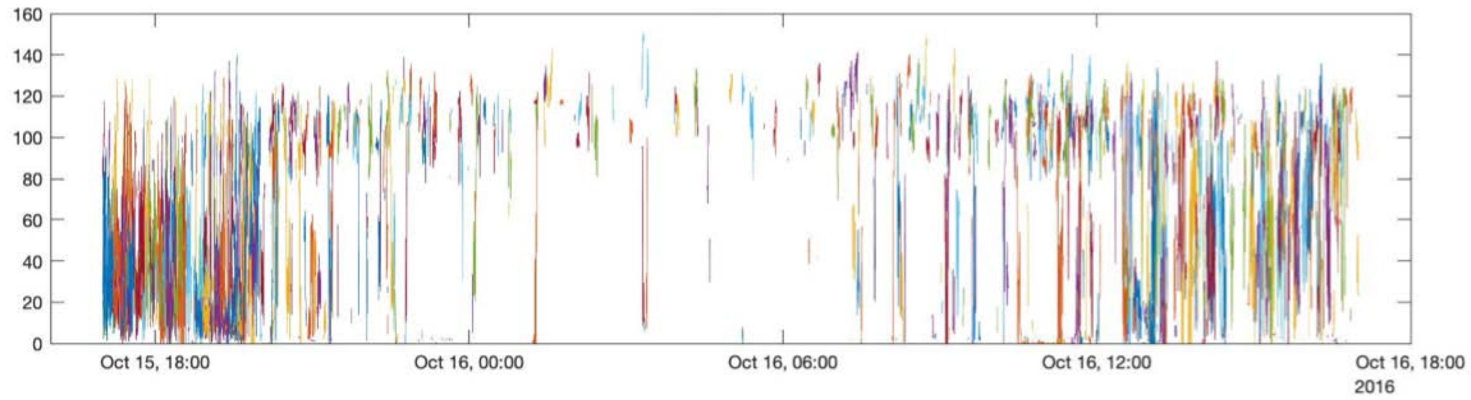
DCRNN Results : Next Step Mobile Device Integration

11,160 detectors



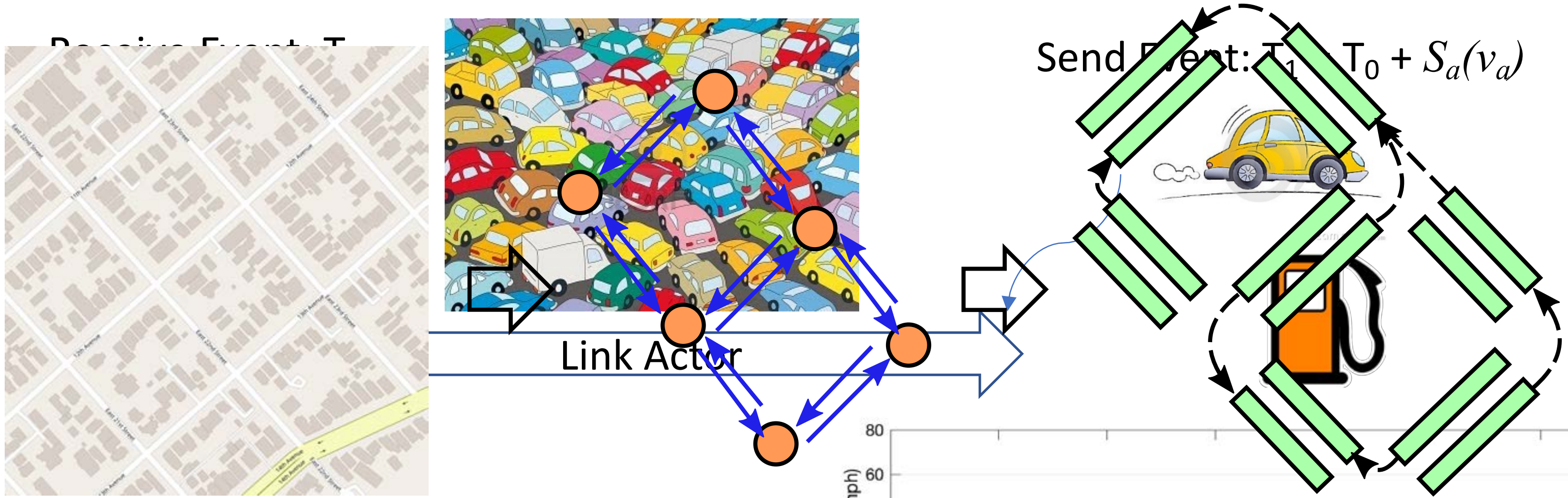


MAE is usually under 3 miles per hour

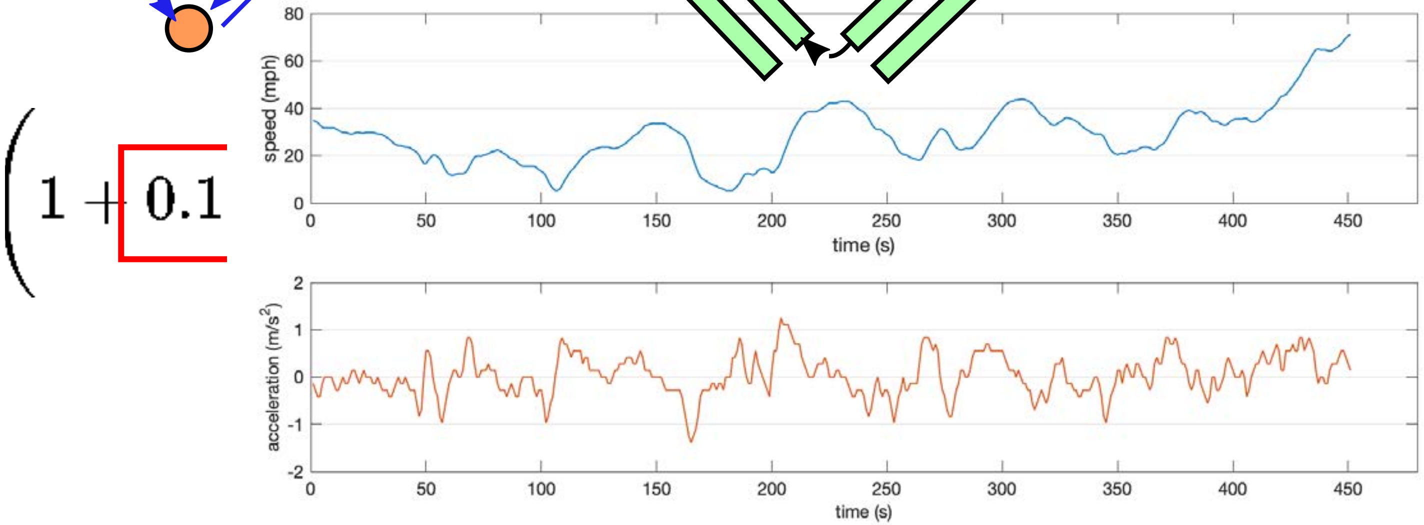


Mobile device trajectories for 1210 segment

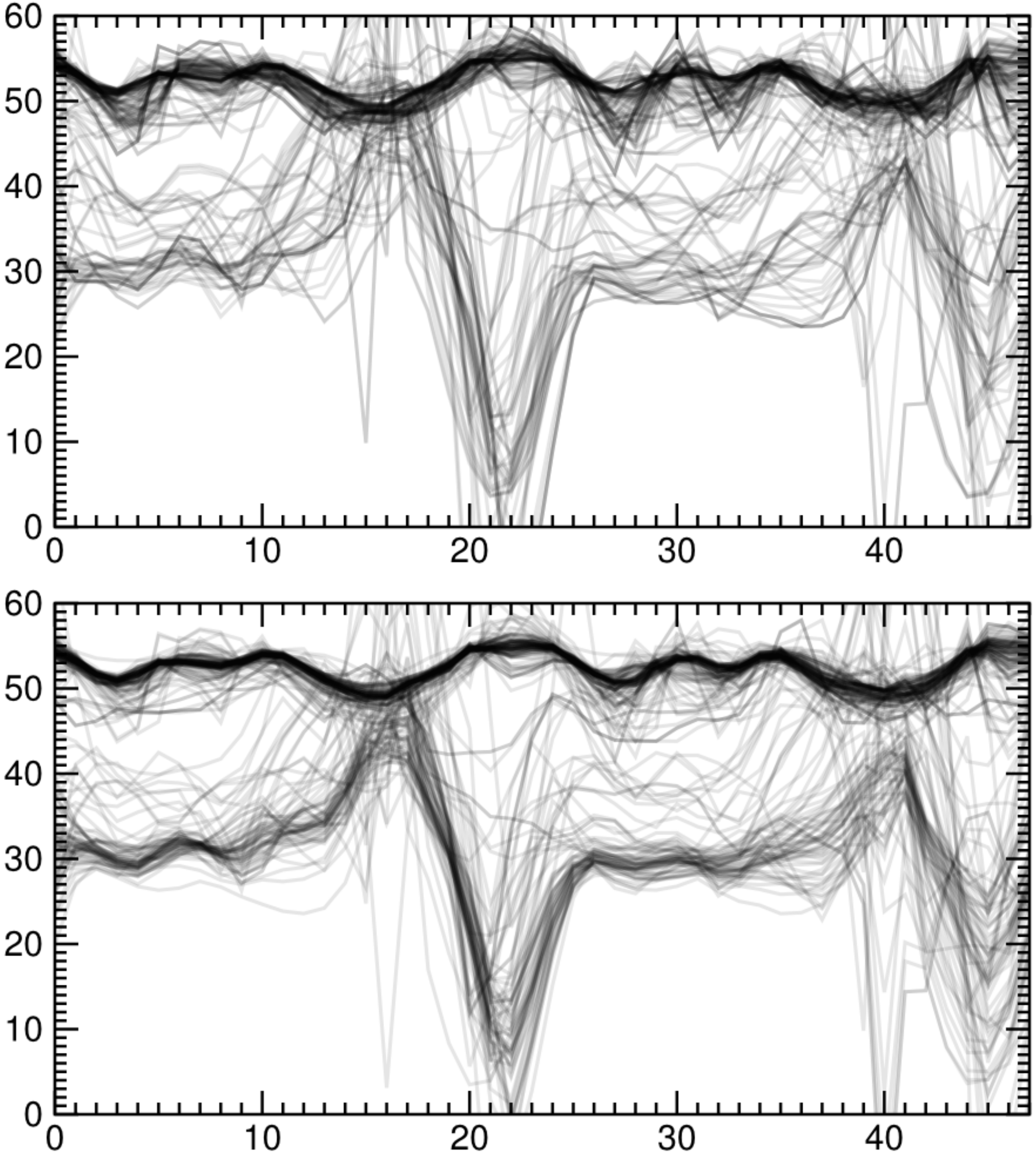
Link Actor Model Provides Foundation



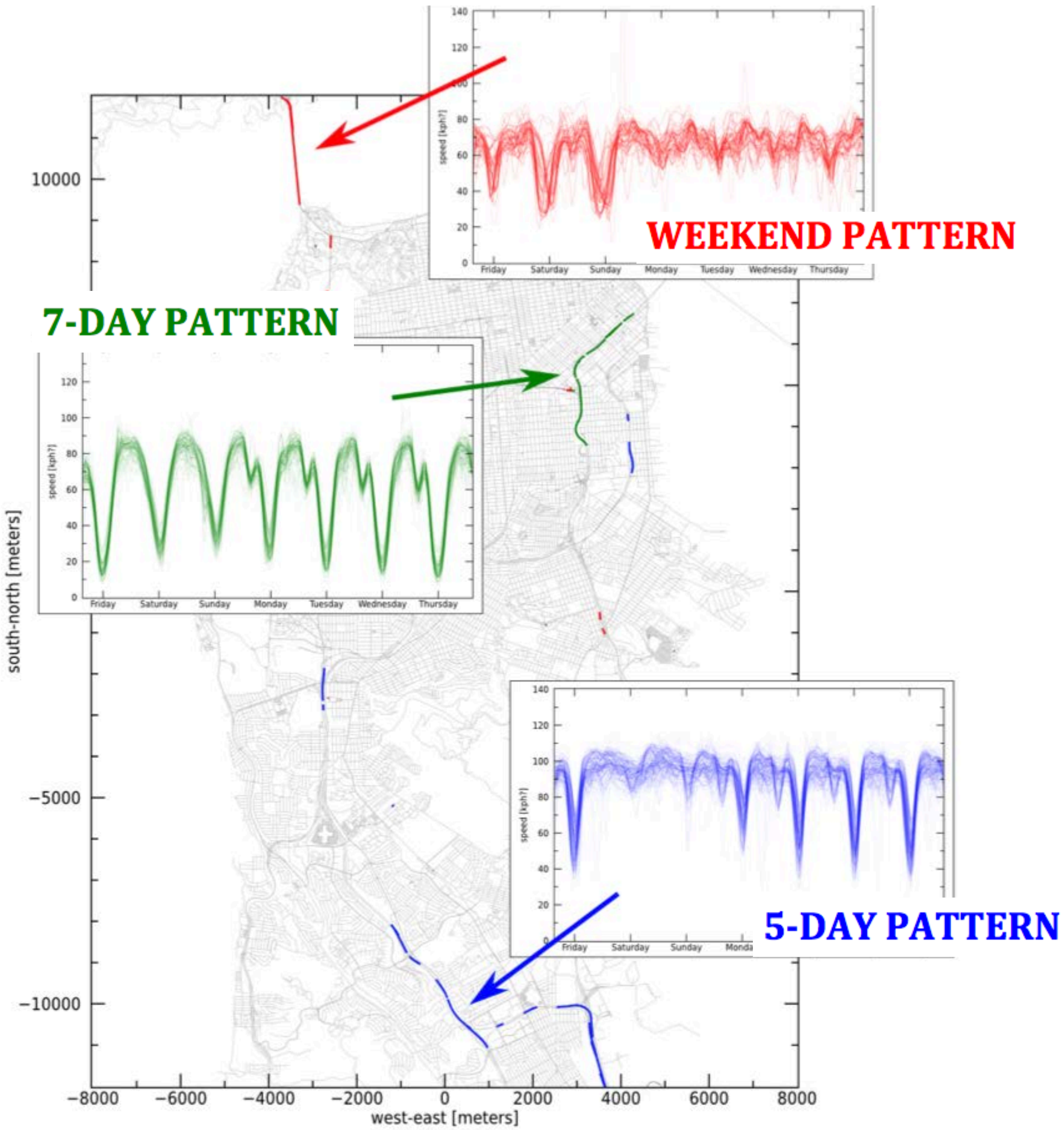
Replace with link/multi-link models learned from real-world dynamics



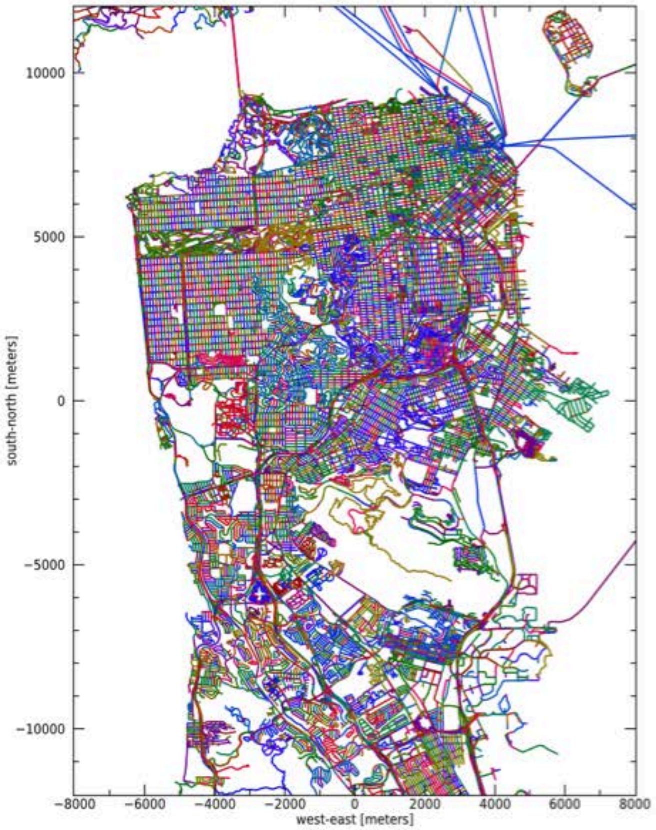
- Focused on the dynamic evolution on traffic networks – we are not modeling demographics, choice of transit over personal vehicle, lane level dynamics



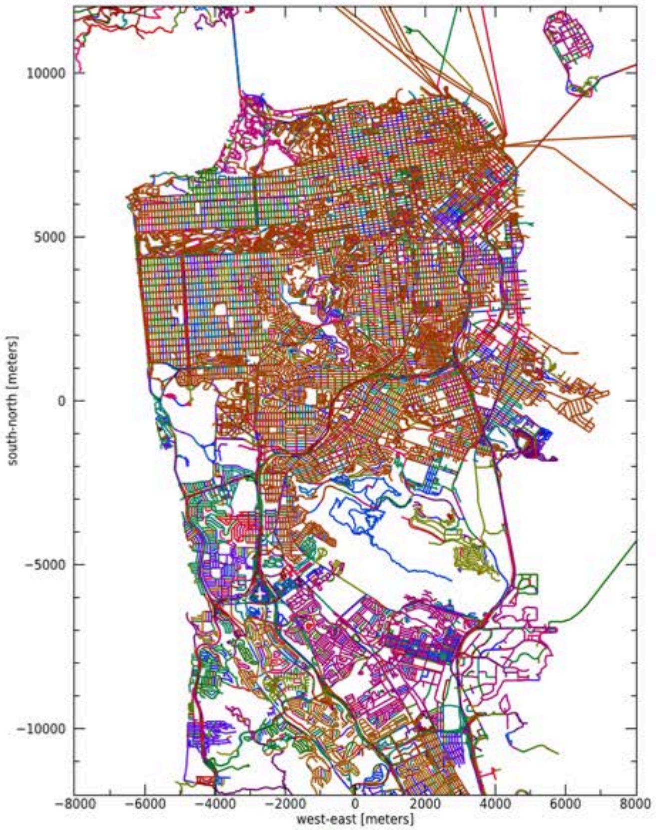
60
50
40
30
20
10
0
0
10
20
30
40
60
50
40
30
20
10
0
0
10
20
30
40



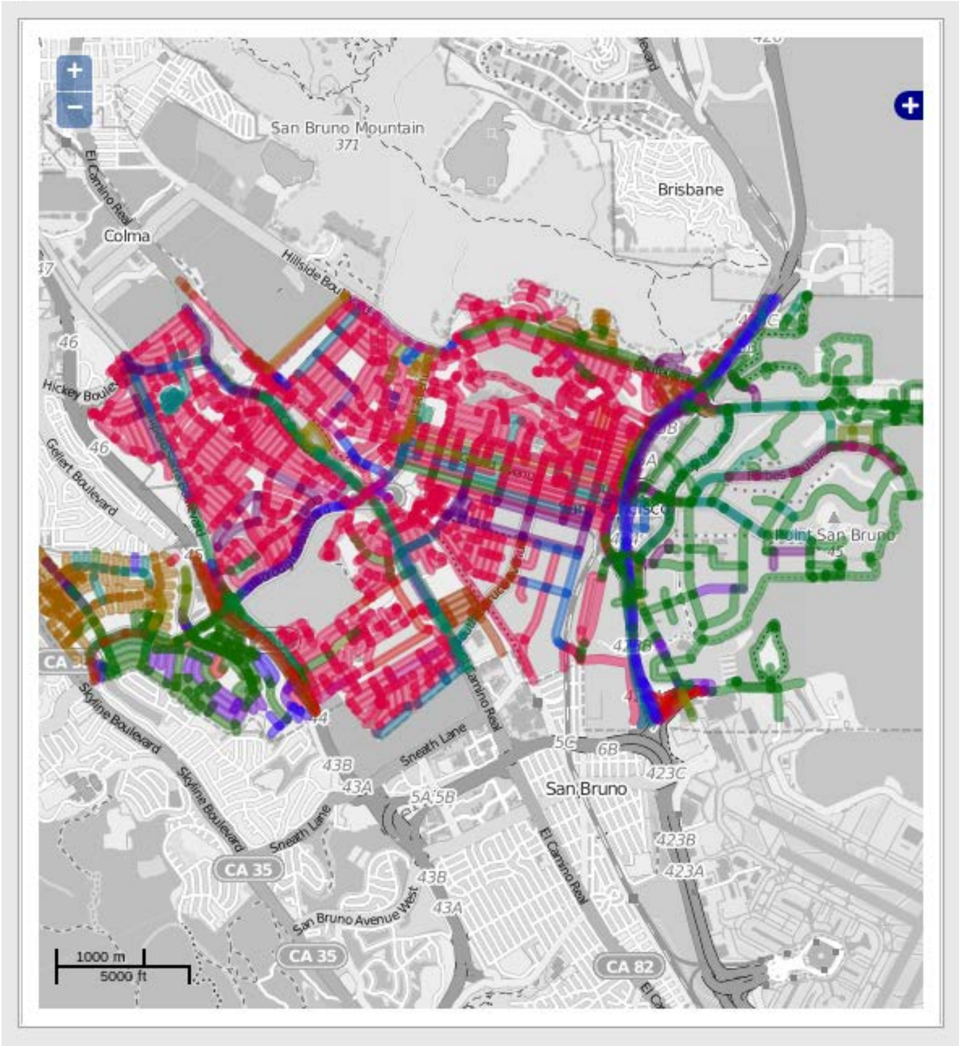
WEEKEND PATTERN
7-DAY PATTERN
5-DAY PATTERN
speed [kph?]
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
south-north [meters]
west-east [meters]



(a) $\chi^2 < 1000$



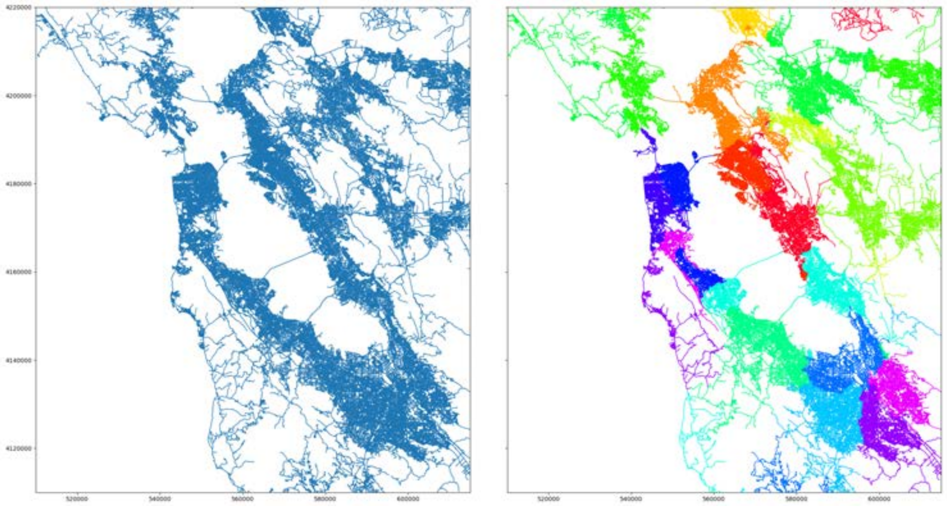
(b) $\chi^2 < 2500$



Optimistic Parallel Discrete Event Simulation

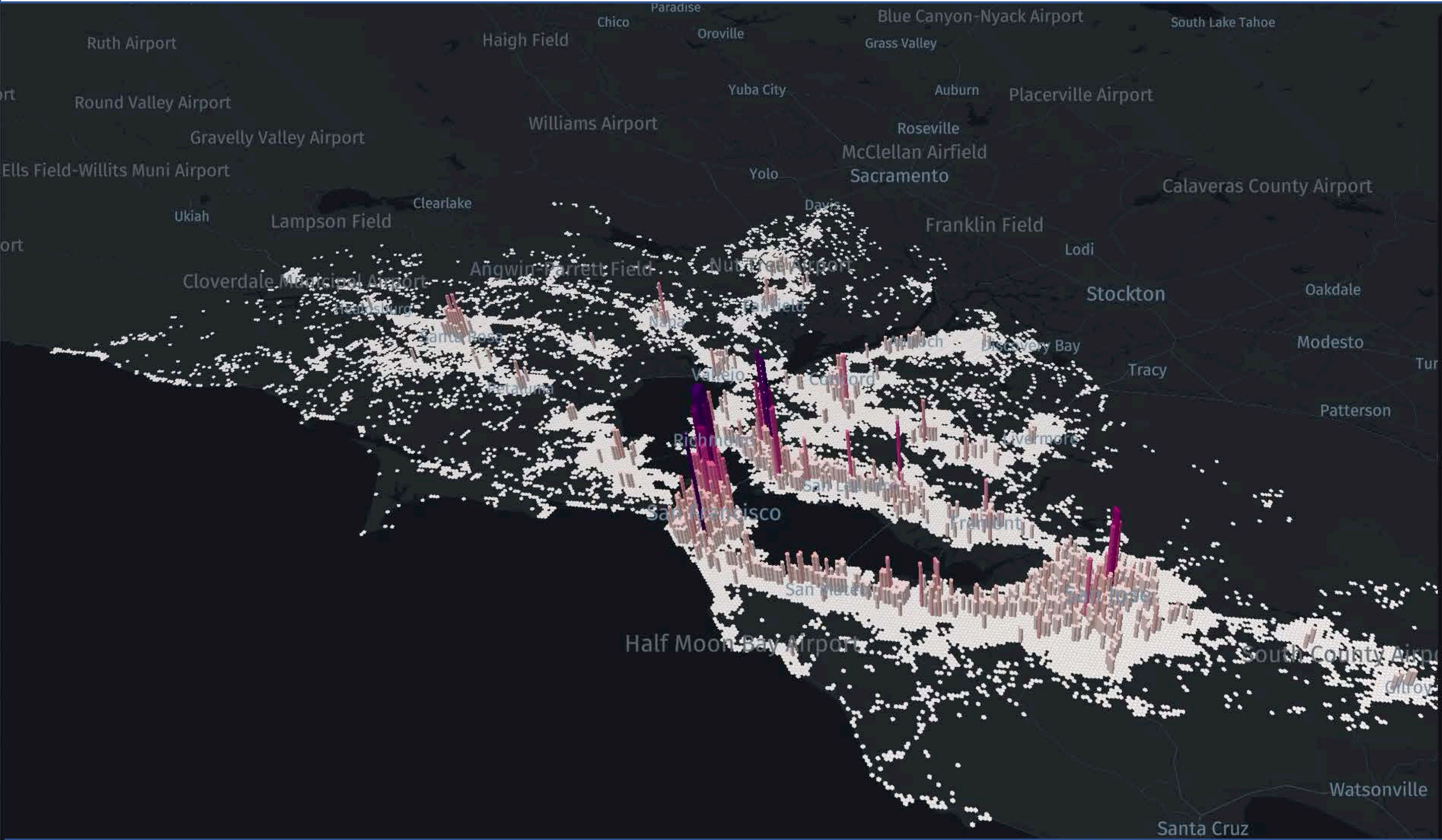
- Simulation is parallelized by splitting links across multiple computer nodes/processes/threads to logical processors (LPs)
- Vehicles traverse between LPs and must be communicated between ranks
- Leverages GASNet-Ex and Devastator (PDES engine)
- Avoids synchronization through optimistic asynchronous execution

Conservative (window-based) PDES:
Requires every rank to be synchronized to a global time step
Global time step determined by fastest possible agent interaction



Geospatial Partitioning of the Network to support Distributed Memory Computation

Optimistic PDES:
Allows ranks to execute without synchronization and enforces causality by rolling back mis-speculatively executed events. Reduce simulation overheads by multi-objective partitioning of actors based on loads and interaction.



Bay Area Large-Scale Traffic Simulation

Origin locations of the 21.6 million trips taken on the Bay Area road network during an average weekday.

Berkeley Labs is simulating the vehicle flow rates and resulting congestion on each of the 2.2 million road links in the system.

Legend

The greater the height of the hexbin, the larger the number of origins points.

0K+ 4K+ 9K+ 13K+ 18K+ 22K+ 27K+ 31K+ 36K+

Map View

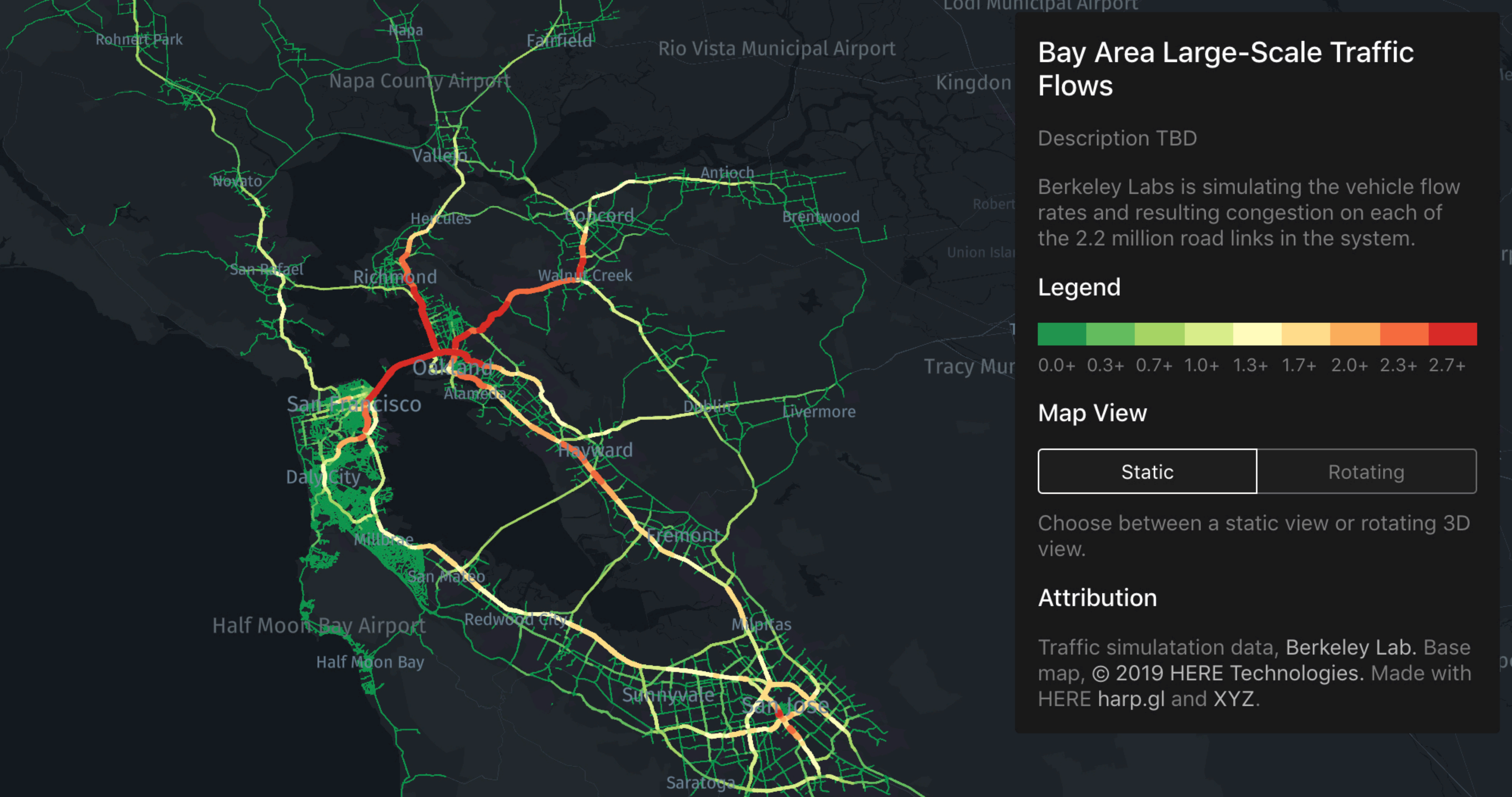
Static | Rotating

Choose between a static view or rotating 3D view.

Attribution

Traffic simulatation data, Berkeley Lab. Base map, © 2019 HERE Technologies. Made with HERE harp.gl and XYZ.

Yojimbo



Demand 22M ODS

Network Size 2M links, 1 M nodes

Active Control?

Emerging Bottom Up Solutions

Active Control

Unregulated impacts the quality of life of the city

Control Systems View

Infrastructure Control

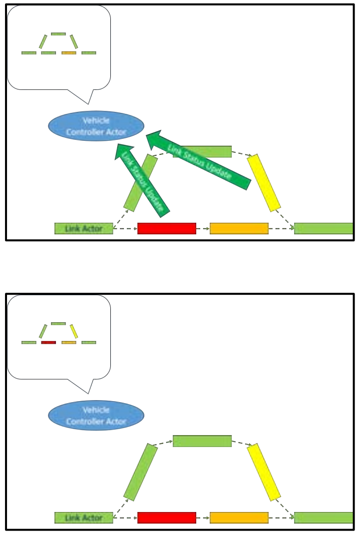
Individual Control

Knowledge Based Routing

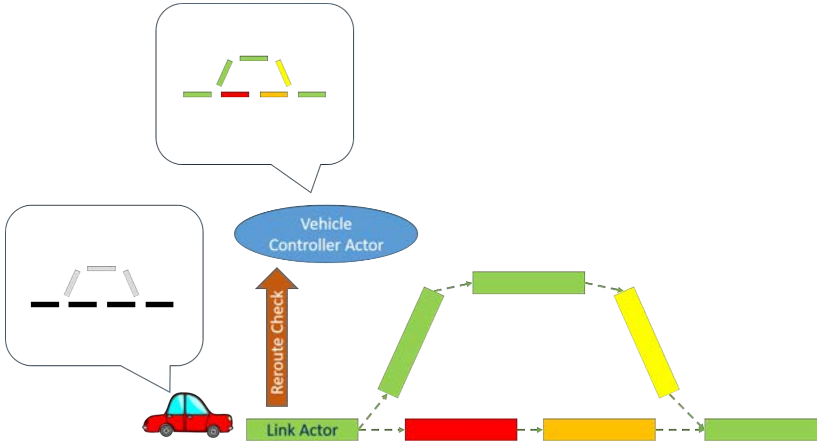
Dynamic Routing

Connected (Automated?) Routing

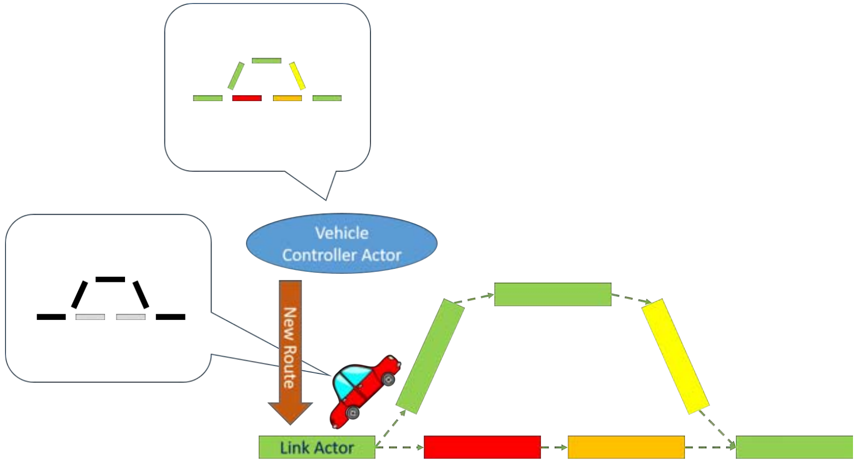
Estimating Impact of Dynamic Rerouting



Link status updates



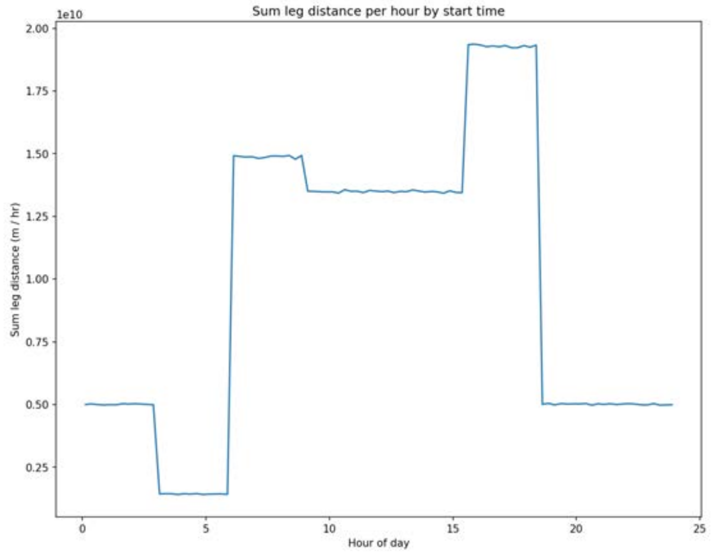
Vehicle checks with controller before traversing congested links



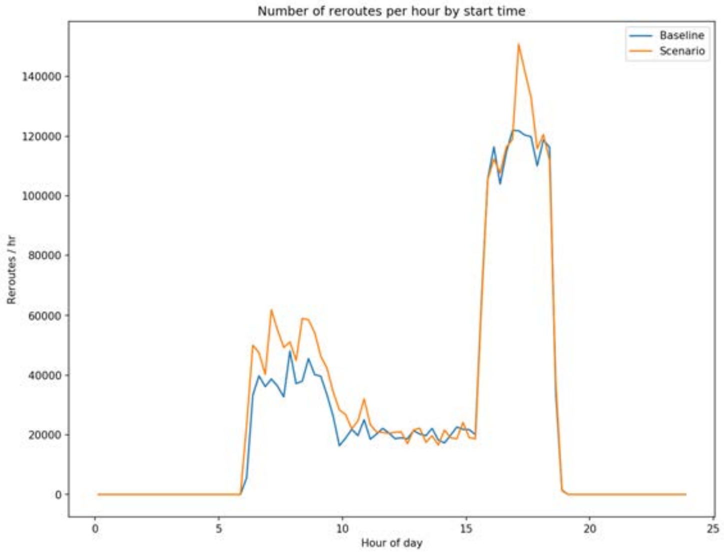
New route avoids congested links

- Through the addition of vehicle controller actors, we can enable a parameterized fraction of vehicles to dynamically reroute around congestion
- This is a key capability to allow the prediction of emergent behavior in response to unexpected changes in the road network or demand
- Example: optimize government response to major traffic incidents or evacuation scenarios

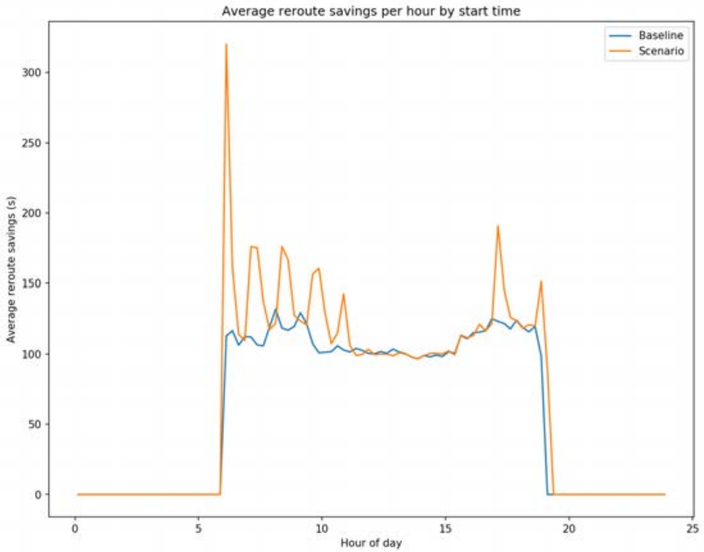
Temporal Distribution of Vehicle Rerouting



System-wide traffic demand



Reroutes per hour

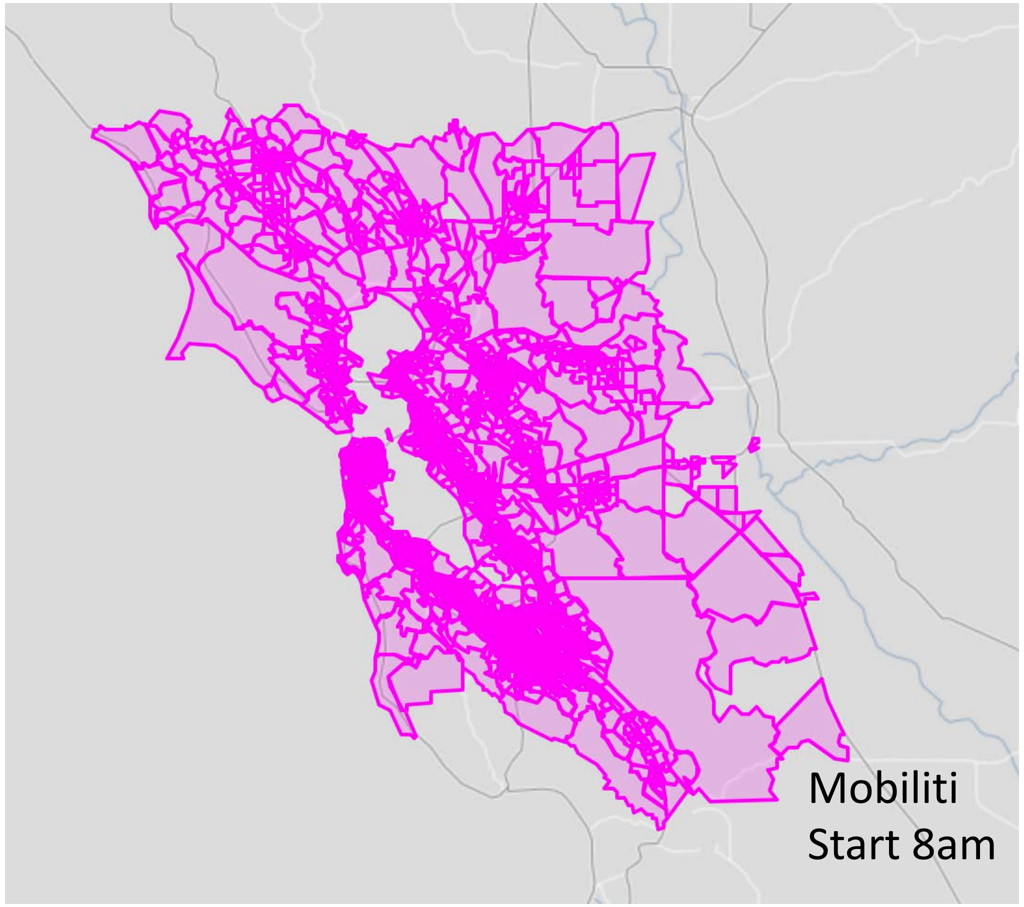


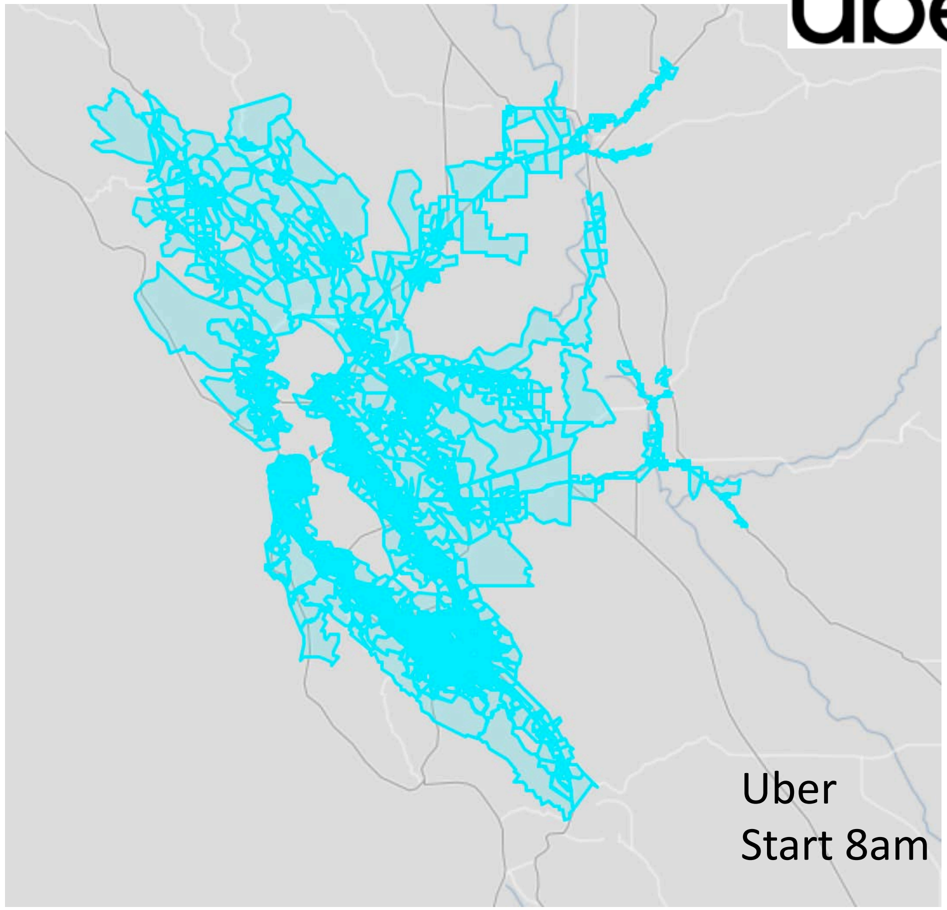
Average time saved per reroute

- We observe the time dependence of vehicle rerouting depends on the demand profile on the traffic system as well as timing of the incident scenario
- Reroutes peak during morning and evening peaks, as well as during the induced congestion due to incident scenario
- Value of rerouting (average time saved) dramatically increases during a major road network incident

Validation

Simulation Validated with Relevant Real World Data





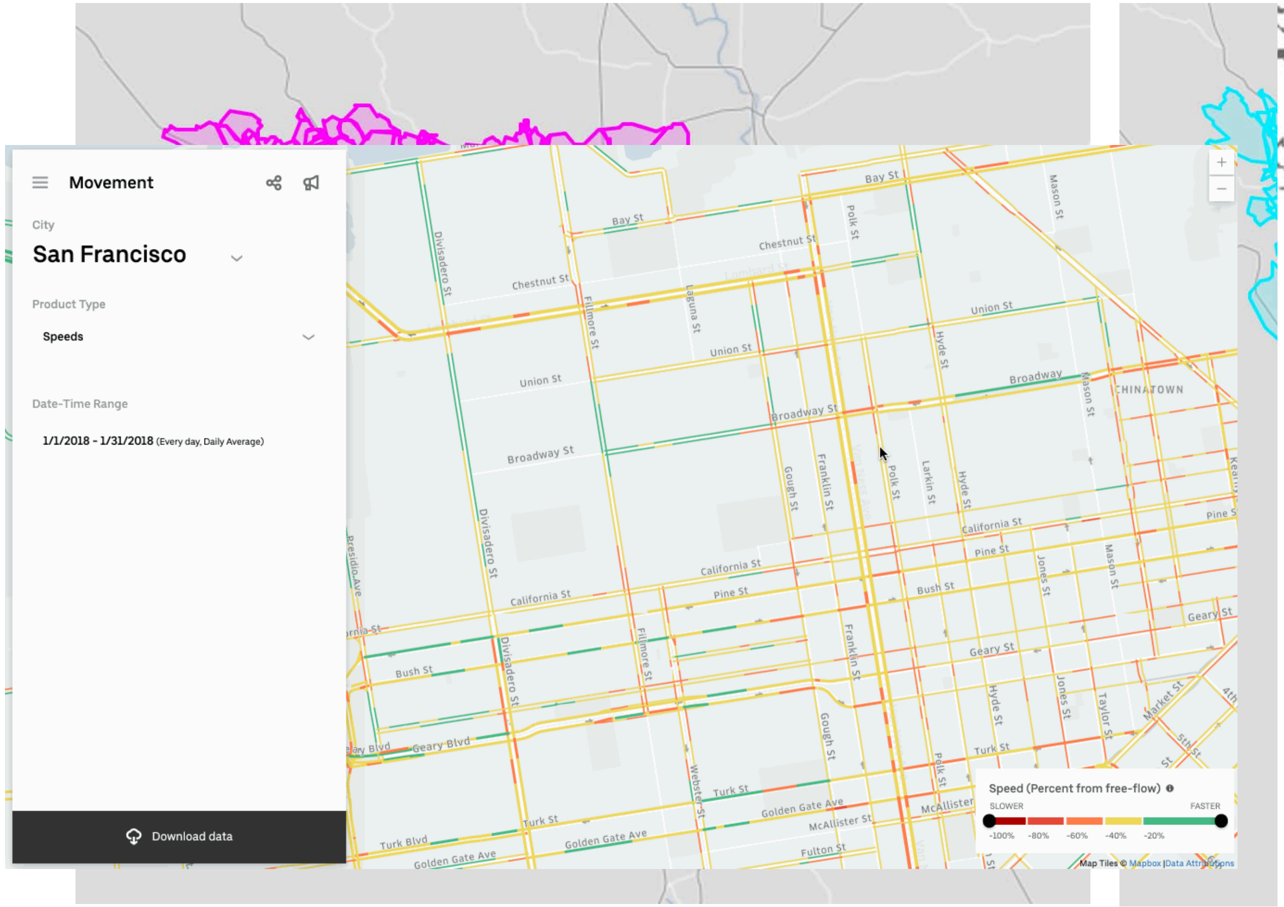
Uber
Movement
Travel
Times

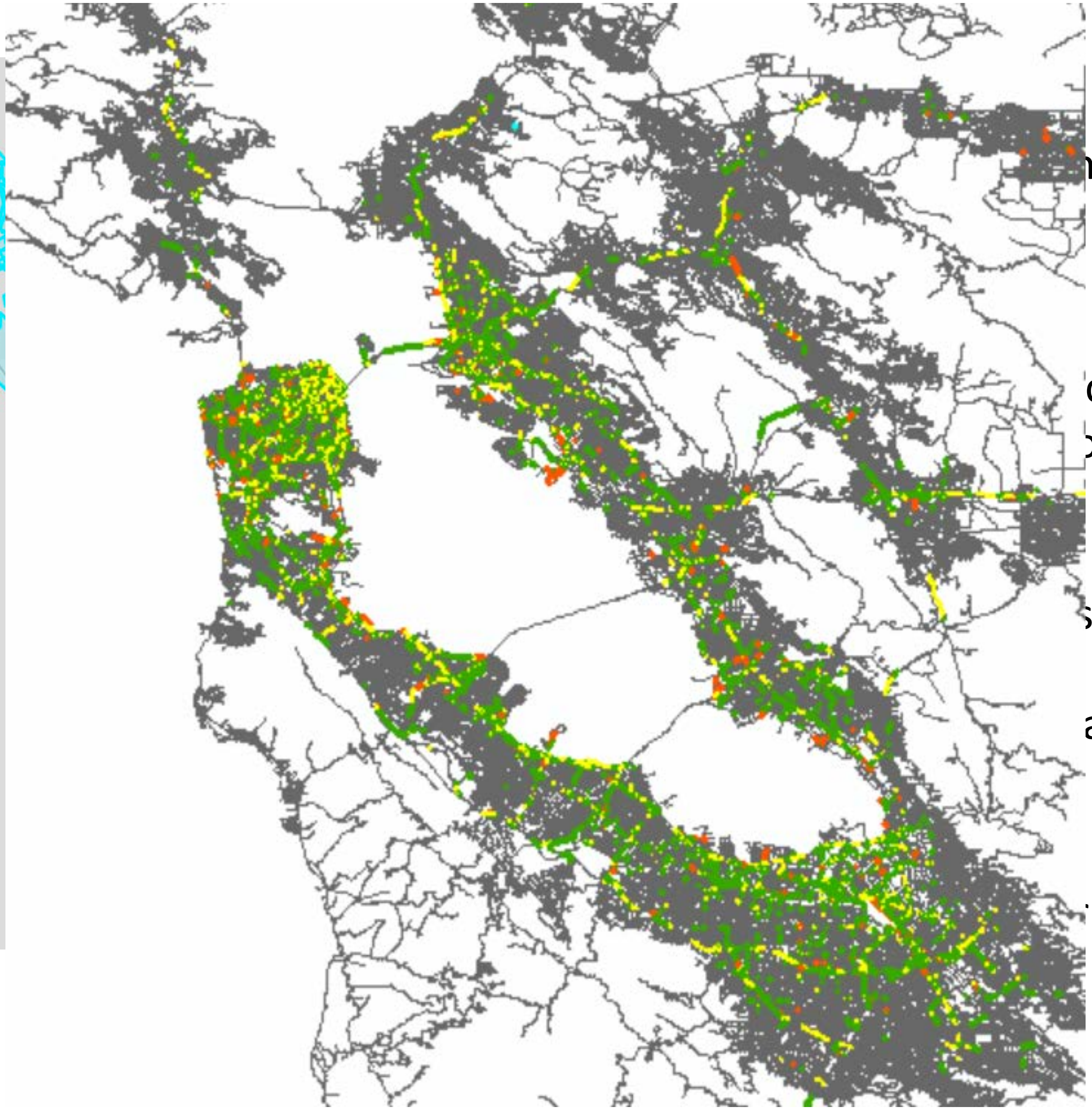
San Francisco
Planning Group
Demand
Model
22M OD’s

Problem Space
719M

8 M overlap

Simulation Validation





nt

co
oup

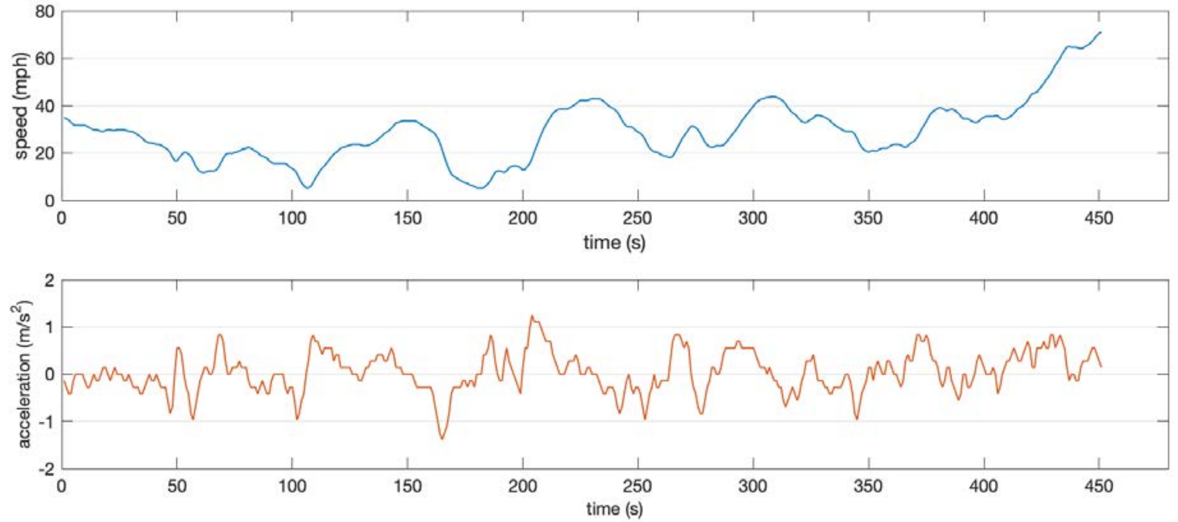
ace

lap

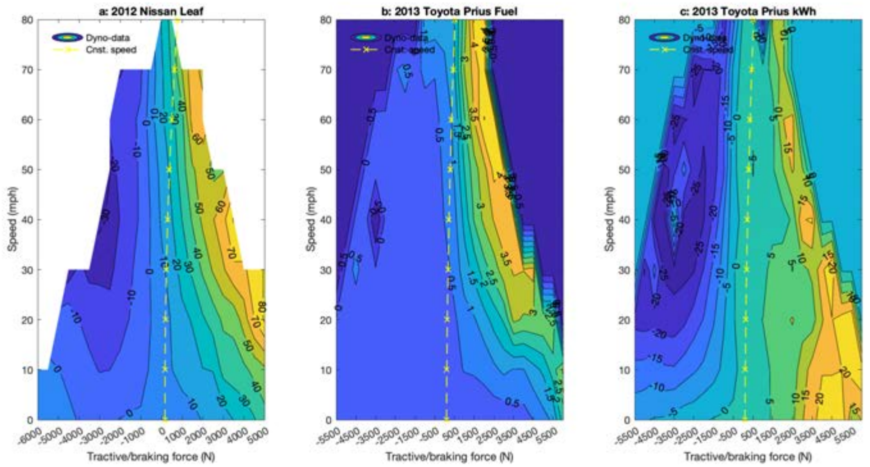
Energy Budget?

Energy Consumption Estimates from Real-World Devices

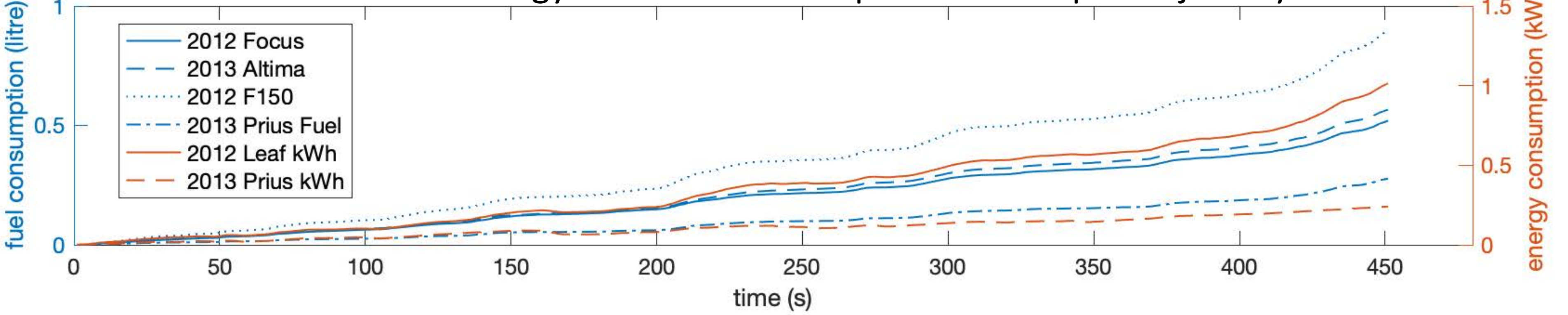
Sample Trajectory in Congestion



ML Derived Fuel and Energy Consumption Rates for Plug-In Hybrid Vehicles from ANL D3 Datasets



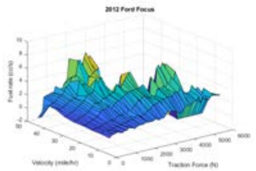
Accumulative Energy and Fuel Consumption for Sample Trajectory



Prediction of Instant Fuel Consumption Rate

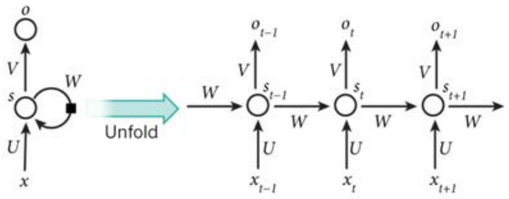
- Data-driven approach to combine the D3 dynamometer datasets with real-world speed trajectories from HERE probe data
- Long short-term memory network (LSTM) provides the prediction capability

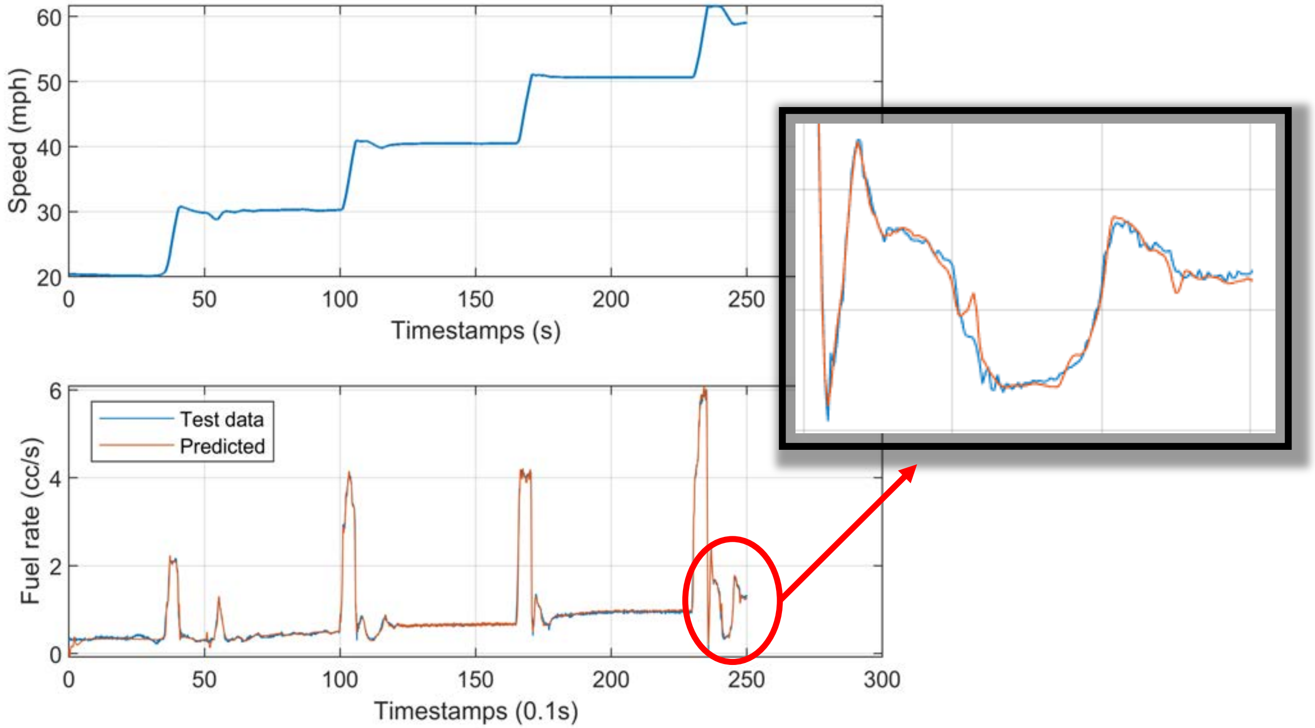
$$c(t) \Leftarrow Fuel\_map(F_T(t), V(t))$$



$$\begin{aligned} c(t) \Leftarrow \boldsymbol{F}[&F_T(t), F_T(t-1), \cdots, F_T(t-N),\\ &V(t), V(t-1), \cdots, V(t-N),\\ &RPM(t), RPM(t-1), \cdots, RPM(t-N),\\ &T(t), T(t-1), \cdots, T(t-N),\\ &\vdots\\ &c(t-1), c(t-2), \cdots, c(t-N-1)] \end{aligned}$$

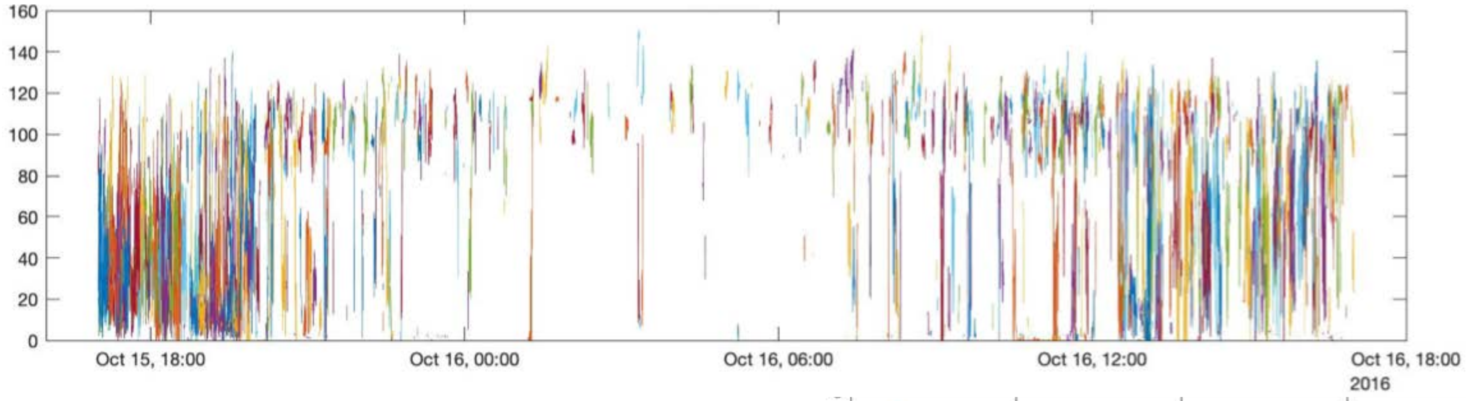
F: Recurrent Neural Networks (RNN)



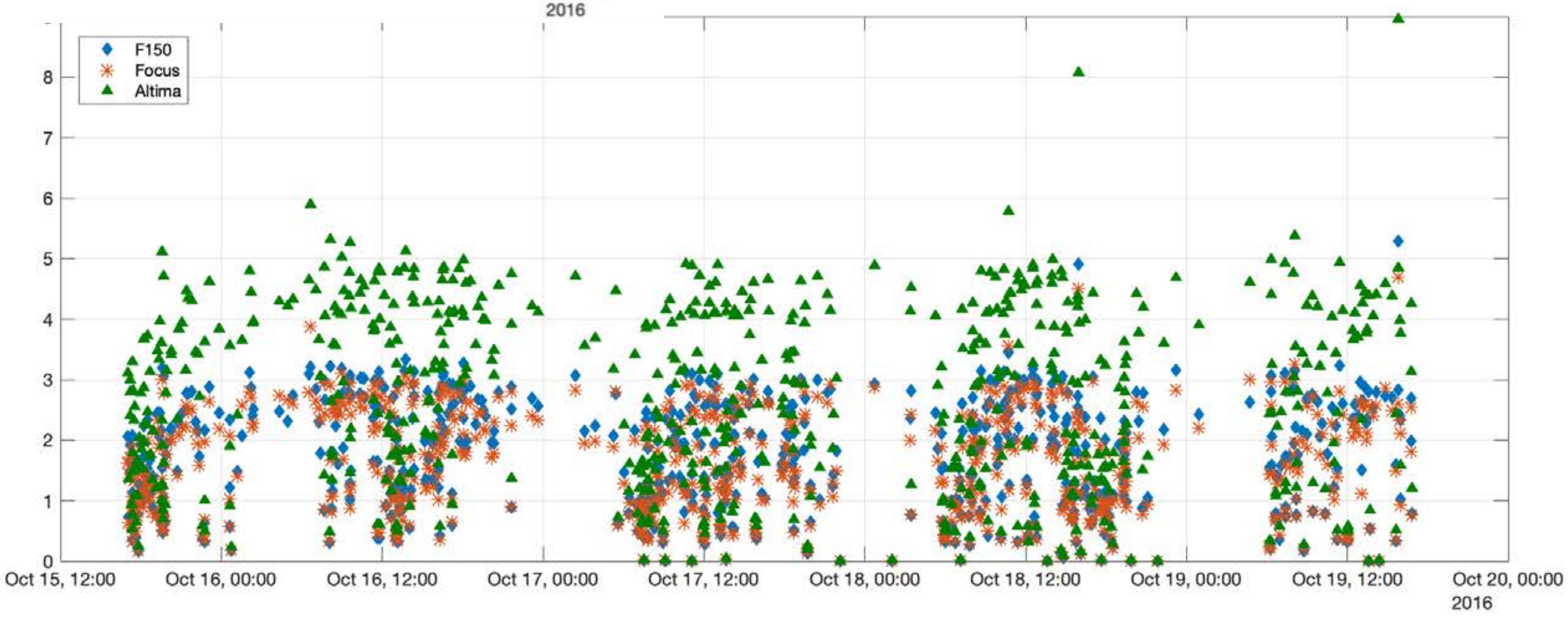


ML Models for Energy Consumption sample I210 trajectories

Trajectories from Mobile Devices on I210



Energy Consumption Sample for Three Vehicles Types



Grid

- **Develop *intelligent, scalable and computationally efficient* solutions for coupled grid-transportation co-optimization**
    - Reduction and decomposition;
    - Bi-level optimization Stackelberg game;
    - Massively-parallelization;
    - Hybrid solution space method: evolutionary + gradient methods;

I.

P

$$\min_{X_1,X_2} f(X_1, X_2, \cdots)$$

Subject to:

$$g(X_1, X_2, \cdots) \le b$$
$$h(X_1, X_2, \cdots) = c$$

$\longrightarrow$

p₁

$$\min_{X_1} f(X_1, \hat{X}_2, \cdots)$$

Subject to:

$$g(X_1, \hat{X}_2, \cdots) \le b$$
$$h(X_1, \hat{X}_2, \cdots) = c$$

$+$

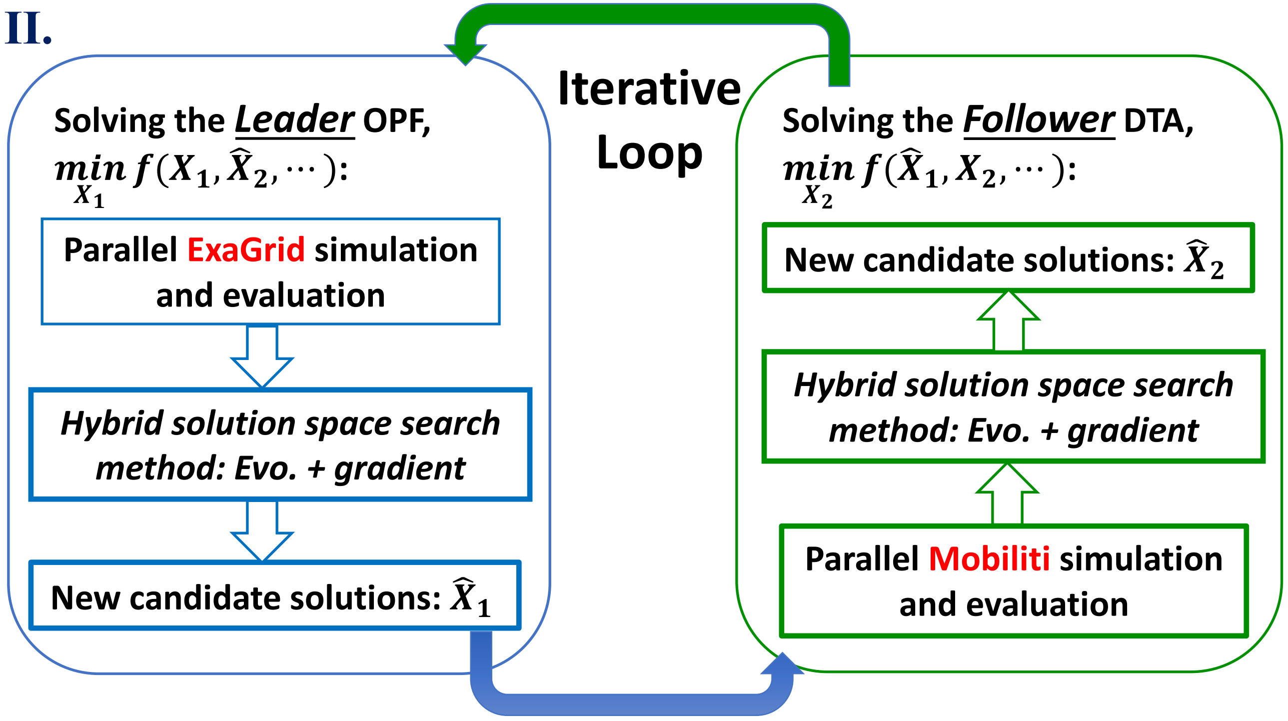
p₂

$$\min_{X_2} f(\hat{X}_1, X_2, \cdots)$$

Subject to:

$$g(\hat{X}_1, X_2, \cdots) \le b$$
$$h(\hat{X}_1, X_2, \cdots) = c$$

II.



High-performance Computing (HPC)

Thank you!

Cy Chan, Bin Wang, John Bachan
LBNL

Prasanna Balaprakash, Tanwi Mallick, Eric Rask
ANL