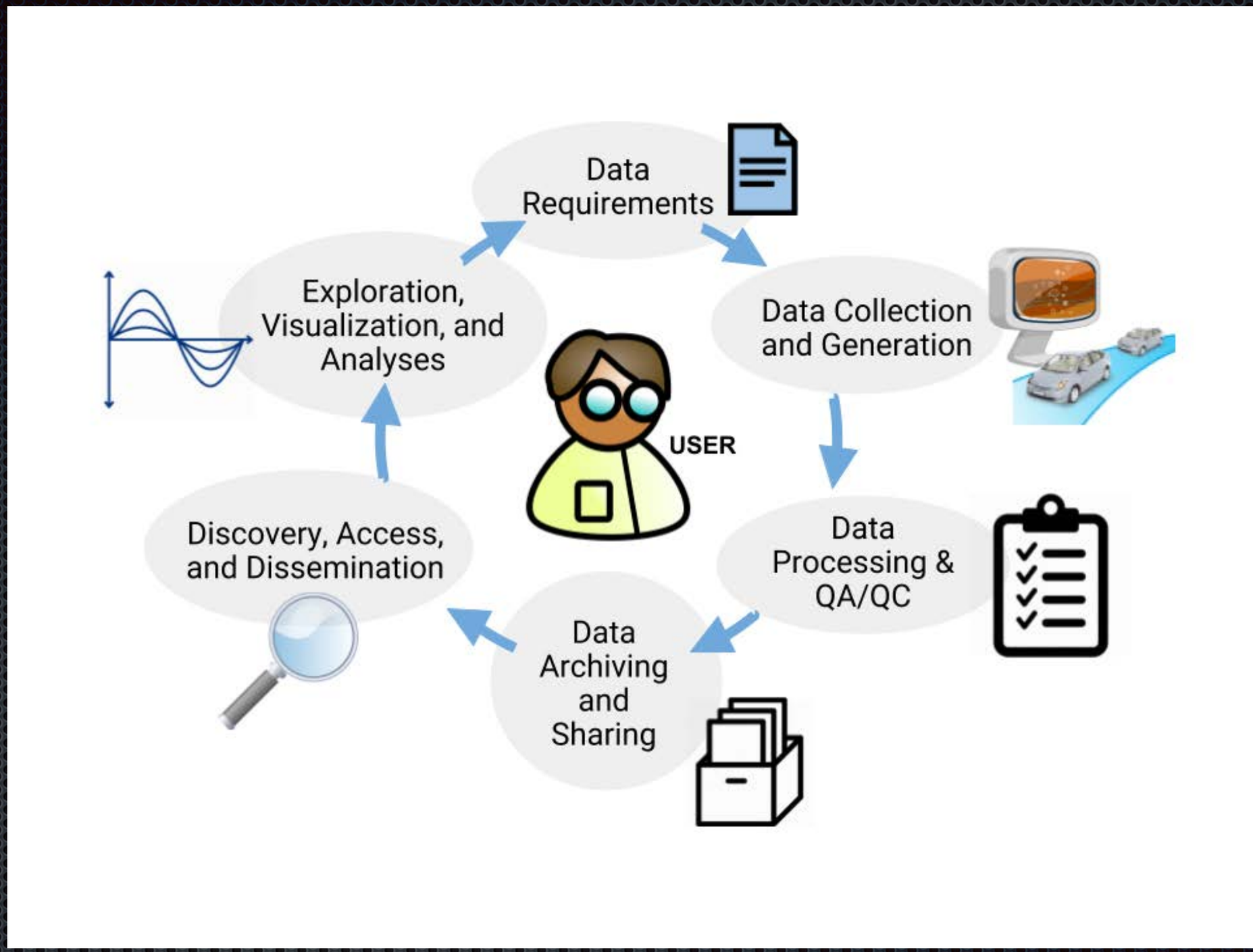
KNOW YOUR DATA

Jane Macfarlane
UCB / LBNL
May 2018

Data life cycle management



Unfortunately, we tend to work in silos…

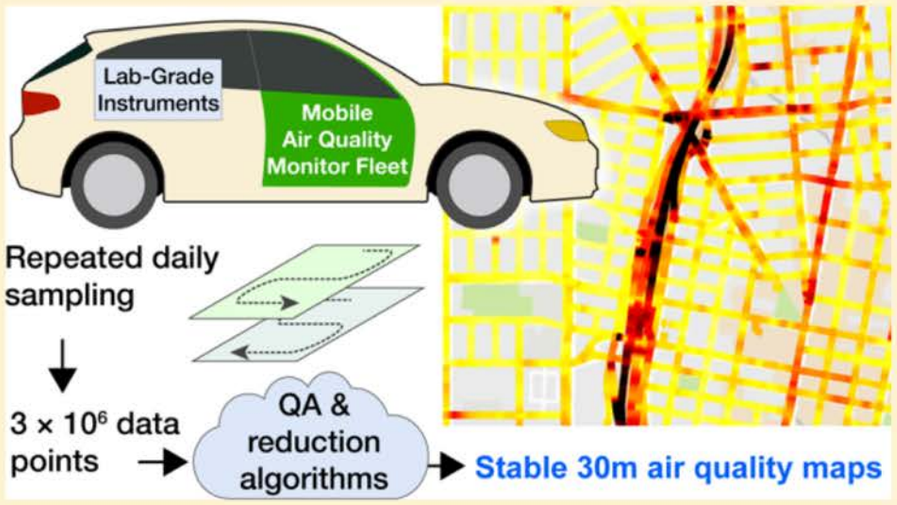
High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data

Joshua S. Apte,[*,†] Kyle P. Messier,[†,‡] Shahzad Gani,[†] Michael Brauer,[§] Thomas W. Kirchstetter,[∥] Melissa M. Lunden,[⊥] Julian D. Marshall,[#] Christopher J. Portier,[‡] Roel C.H. Vermeulen,[▽] and Steven P. Hamburg[‡]

[†]Department of Civil, Architectural and Environmental Engineering, University of Texas at Austin, Austin, Texas 78712 United States
[‡]Environmental Defense Fund, New York, New York 10010 United States
[§]School of Population and Public Health, University of British Columbia, Vancouver V6T 1Z3 Canada
[∥]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, California 94720 United States
[⊥]Aclima, Inc., 10 Lombard St., San Francisco, California 94111 United States
[#]Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington 98195 United States
[▽]Institute for Risk Assessment Science, Utrecht University, Utrecht 3584 CM Netherlands

Supporting Information

ABSTRACT: Air pollution affects billions of people worldwide, yet ambient pollution measurements are limited for much of the world. Urban air pollution concentrations vary sharply over short distances (≪1 km) owing to unevenly distributed emission sources, dilution, and physicochemical transformations. Accordingly, even where present, conventional fixed-site pollution monitoring methods lack the spatial resolution needed to characterize heterogeneous human exposures and localized pollution hotspots. Here, we demonstrate a measurement approach to reveal urban air pollution patterns at 4−5 orders of magnitude greater spatial precision than possible with current central-site ambient monitoring. We equipped Google Street View vehicles with a fast-response pollution measurement platform and repeatedly sampled every street in a 30-km^2 area of Oakland, CA, developing the largest urban air quality data set of its type. Resulting maps of annual daytime NO, NO_2, and black carbon at 30 m-scale reveal stable, persistent pollution patterns with surprisingly sharp small-scale variability attributable to local sources, up to 5−8× within individual city blocks. Since local variation in air quality profoundly impacts public health and environmental equity, our results have important implications for how air pollution is measured and managed. If validated elsewhere, this readily scalable measurement approach could address major air quality data gaps worldwide.



Types of data in the wild.....

- Street level imagery
- LiDAR
- Photogrammetry
- Crowd sourced imagery
- Social media
- Device location and mobility information

Figure courtesy of HERE Research

urban canyon not just a gps problem...

us_zoom2neighborhood160milpts_dark3

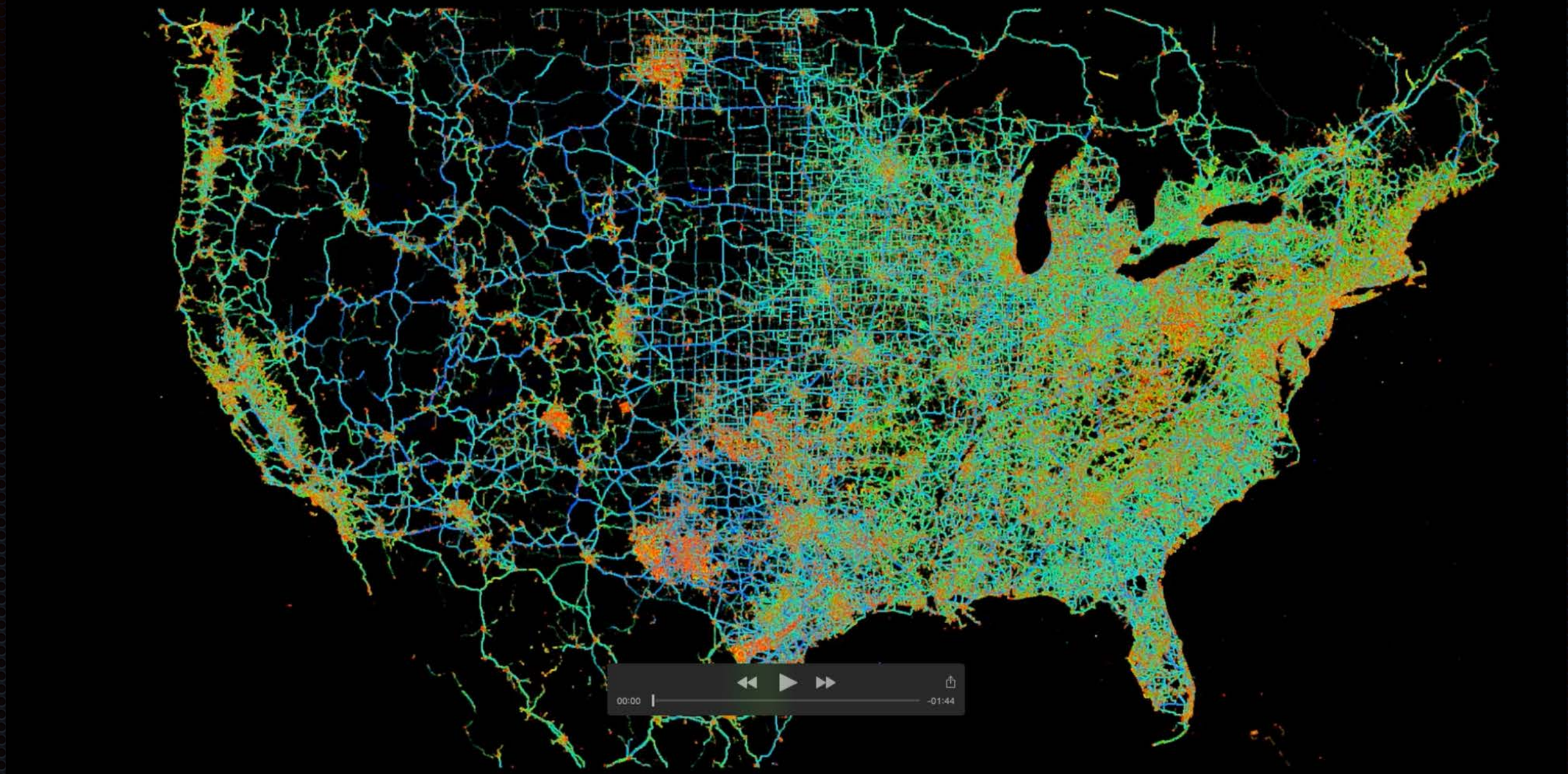


Figure courtesy of HERE Research

Figure courtesy of HERE Research

Good data....

Figure courtesy of HERE Research

Bad data….

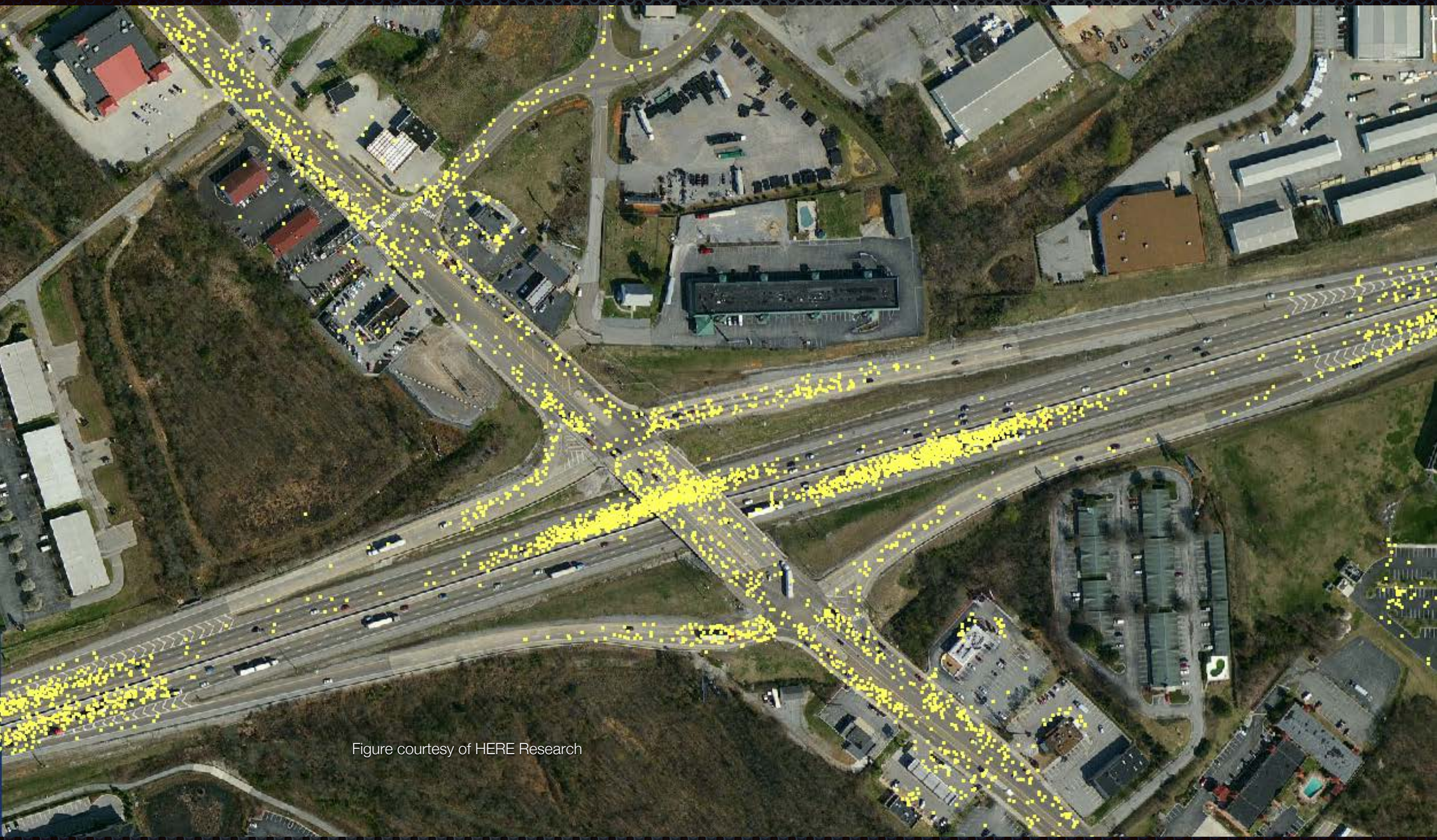


Figure courtesy of HERE Research

Off the charts bad…

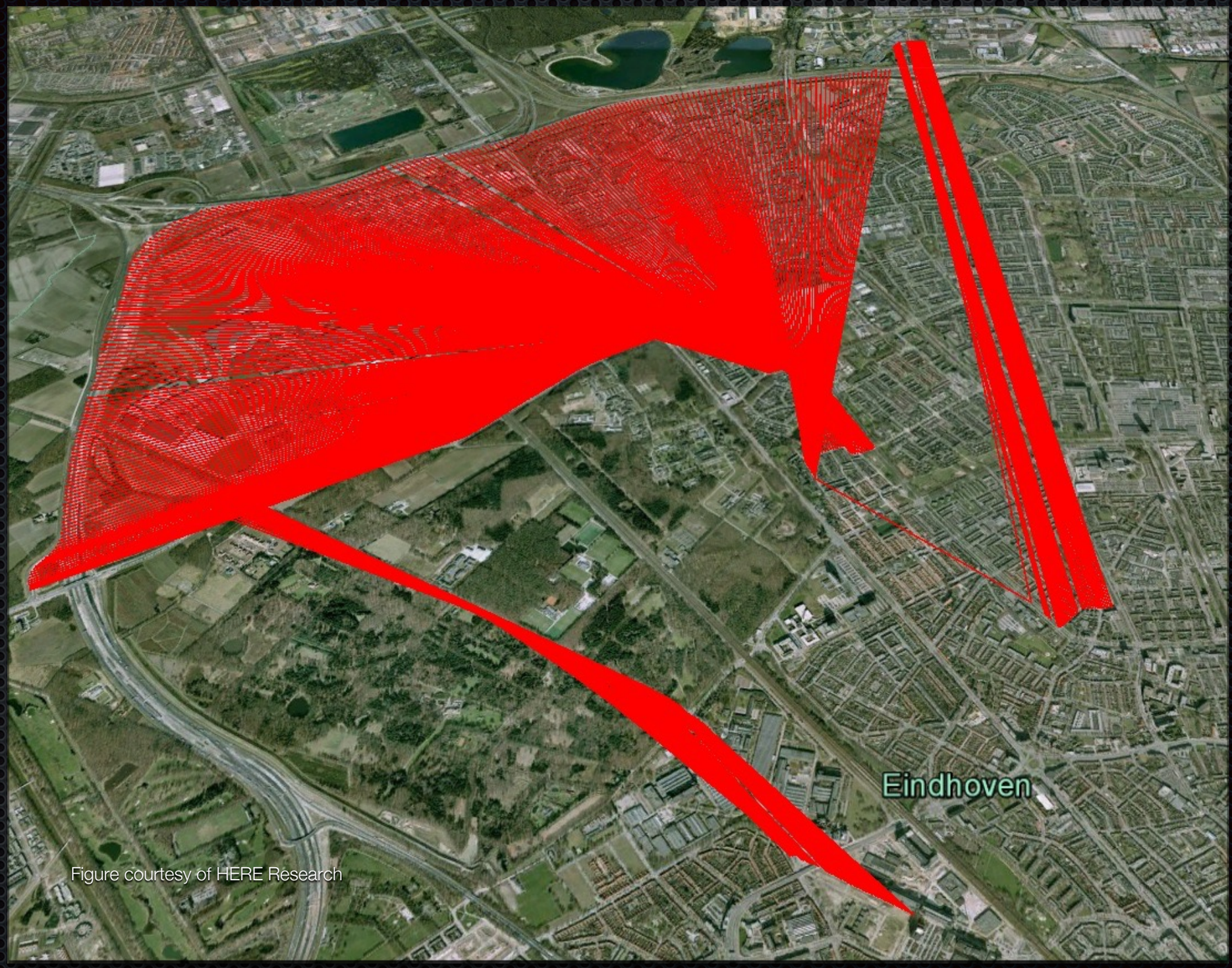


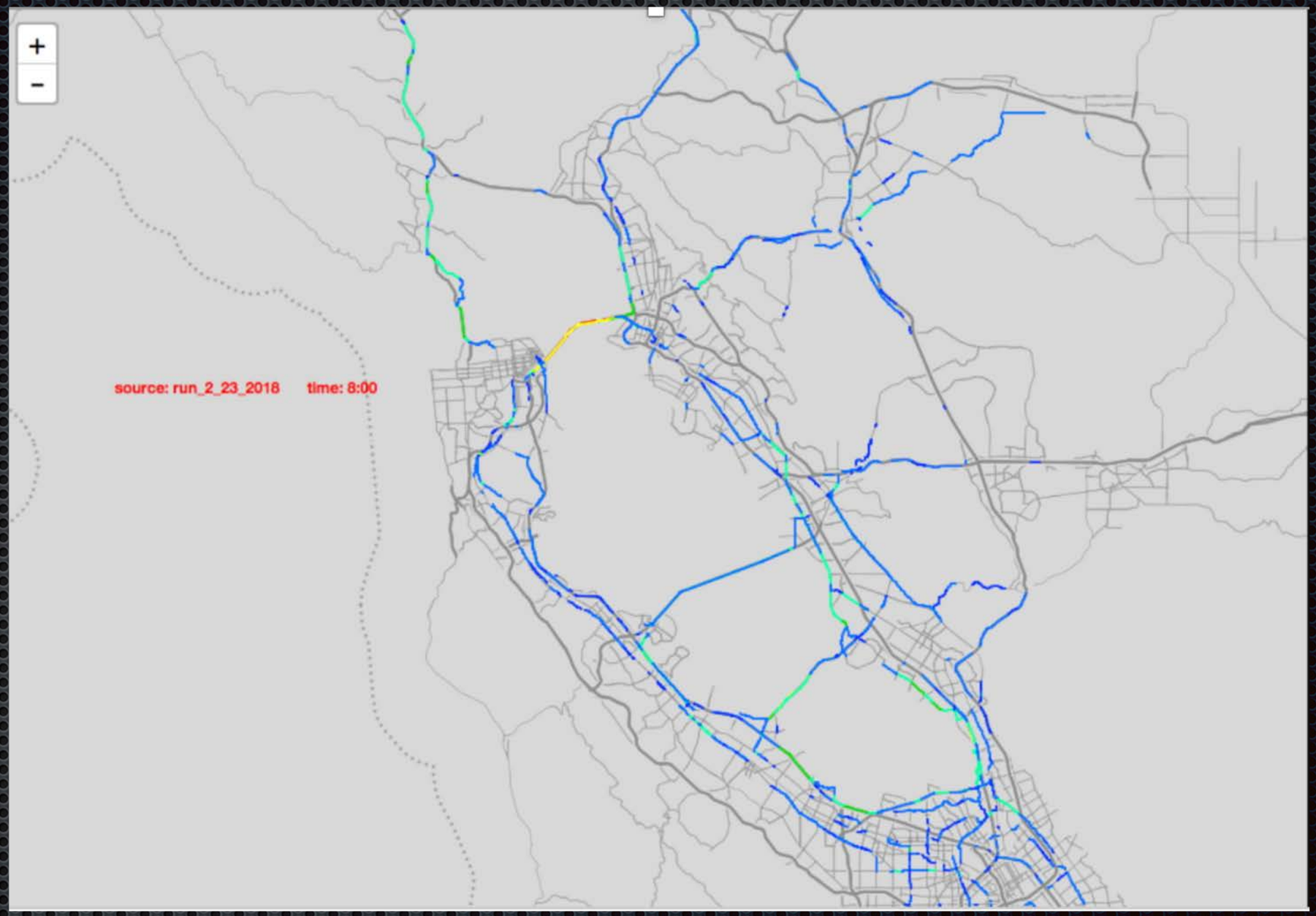
Figure courtesy of HERE Research

Lessons

- Find every occasion to gather labeled data sets
- Know your data
- ML is not new, we just have more data and compute
- Big players are gathering labeled data sets and situational data (e.g. automated vehicles)
- ML cannot be used as a black box
- Small data sets => don’t expect good results
- Know your data :-)
- Fusion of data will likely improve your results
- Educate your customer - balance of True Positive….

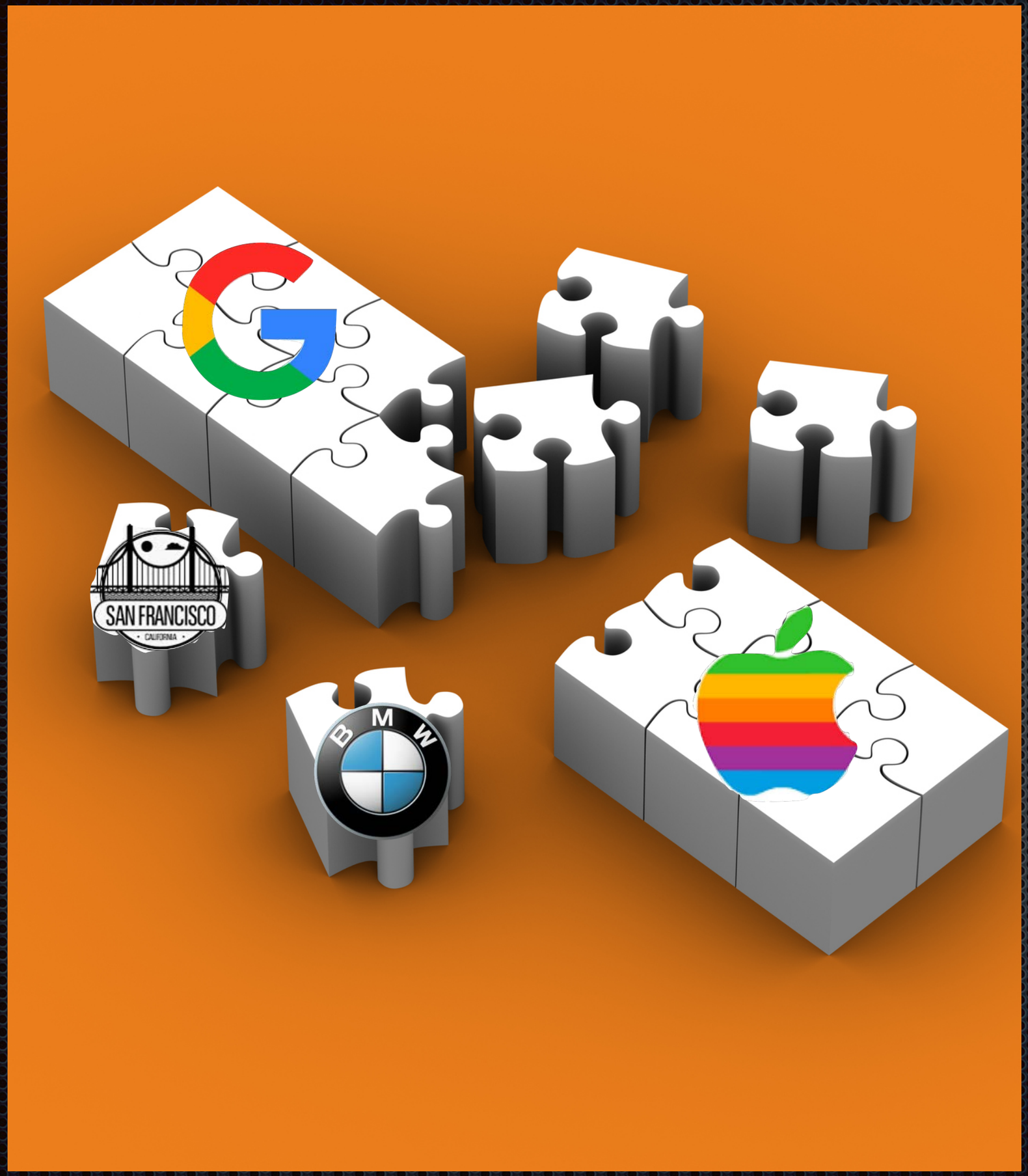
SuperComputing can get us there….

Mobiliti



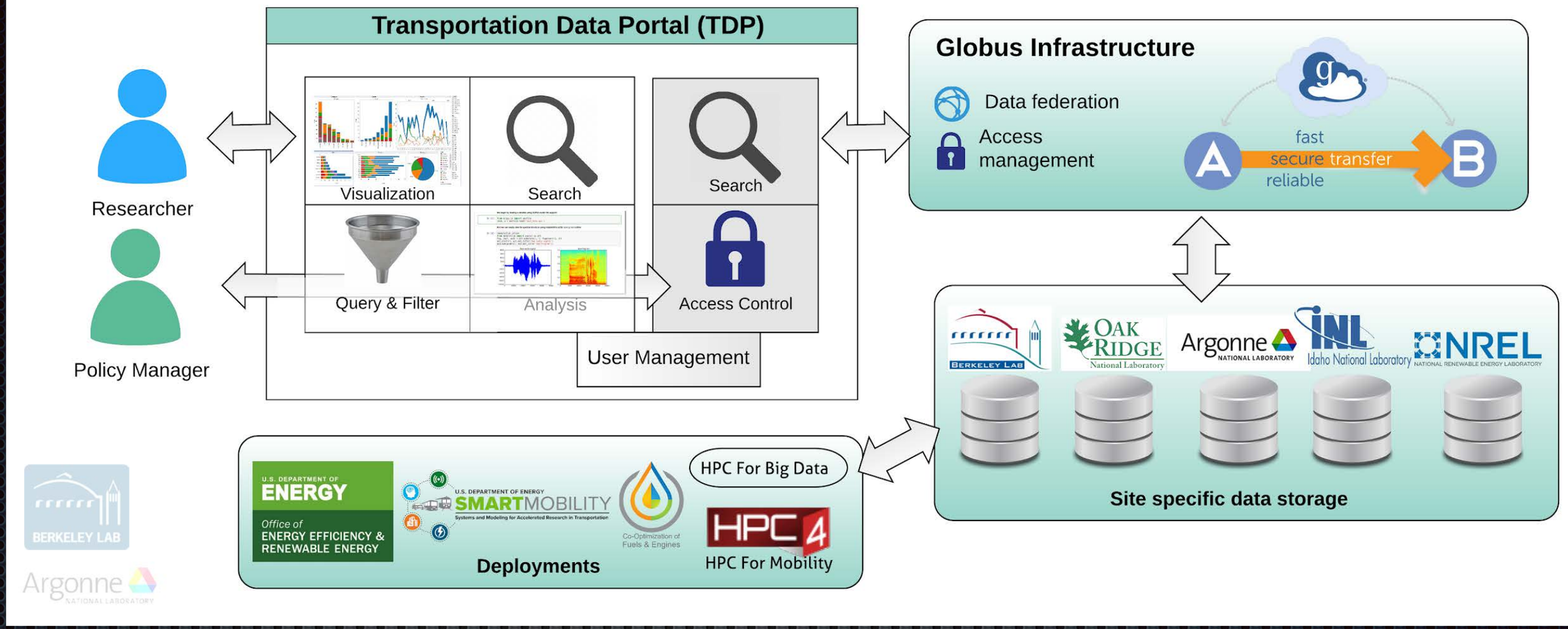
Urban scale simulations – 7.3M vehicles, 2M link network

Need to work from the top down not from the bottom up



We need to break down the silos – one persons data exhaust is another persons gold

KITT – an ecosystem for knowledge exchange



Domain specific semantic knowledge representation connected to site owned data repositories with search, controlled access and exchange, and first level analytics/visualization

Remember…

"On the Internet, nobody knows you're a dog."

Peter Steiner
as published in The New Yorker 1993

Public

Academia

Private

Corporate
Responsibility
To Participate