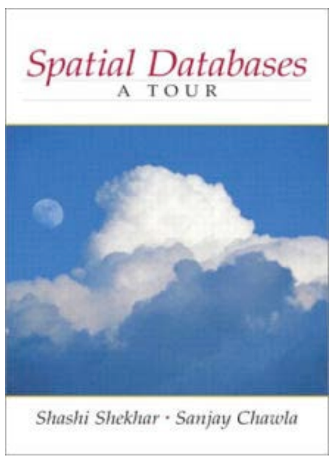
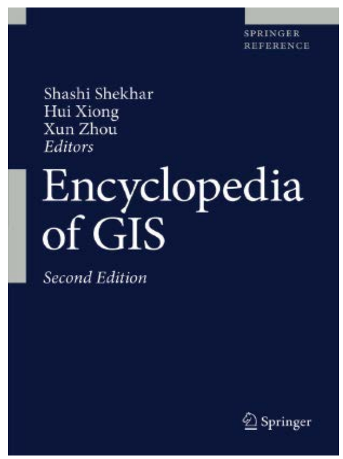
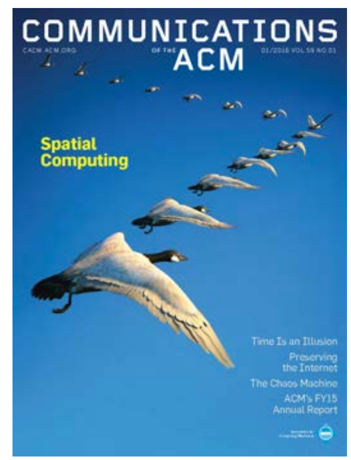
Spatial Data Science and Transportation

Shashi Shekhar

CTS Scholar & McKnight Distinguished University Professor
Dept. of Computer Sc. and Eng., University of Minnesota
shekhar@umn.edu





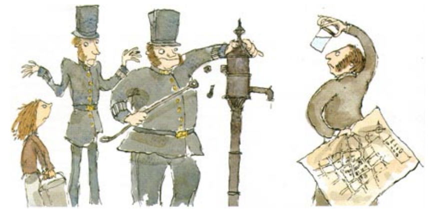


Acknowledgement: Slides prepared by Xun Tang, Yan Li. This material is based upon work supported by the National Science Foundation, the USDOD, the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, the NIH, and the UMN Center for Transportation Studies.

University of Minnesota
Driven to Discover℠

A Spatial Data Science Story

1854: What causes Cholera?

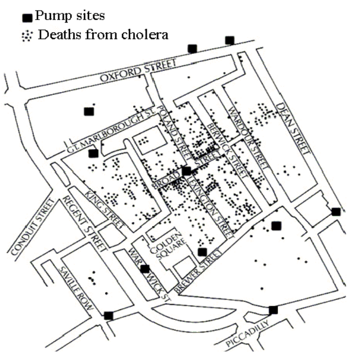


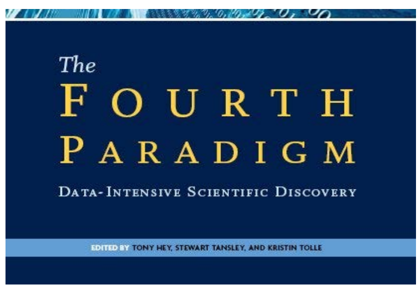
Collect & Curate Data → Discover Patterns, Generate Hypothesis → Test Hypothesis (Experiments) → Develop Theory

? water pump

Remove pump handle

Germ Theory





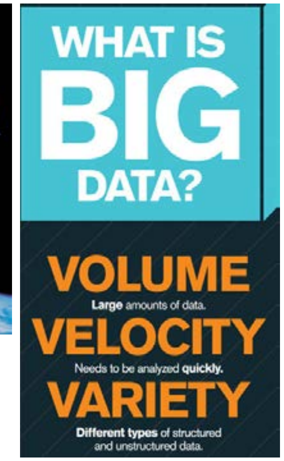
Impact:

sewage system, drinking water supply …

Q? What are Choleras of today?
Q? How may Spatial Data Mining Help?

Details: (1) Spatial computing. (S. Shekhar et al.) *Communications of the ACM*, *59*(1):72-81, 2016.
(2) Transforming Smart Cities with Spatial Computing (Y. Xie et al.) . Proc. IEEE Intl. Smart Cities Conference, 2018.

What is new since Snow's map? **Spatial Big Data**



- 1980s : USDOD opens GPS for civilian use
  - 1990s: use in Intelligent Transportation Systems
- Today: 2 billion GPS receivers in use (7 billion by 2022).
  - Many share location every second
  - Generating a large volume of location traces

- GPS also provides reference time for many infrastructure
  - Airlines, Telecommunications, Banks
- GPS is the single point of failure for the entire modern economy.
- 50,000 incidents of deliberate (GPS) jamming last two years
  - Against Ubers, Waymo's self-driving cars, delivery drones from Amazon

Bloomberg Businessweek

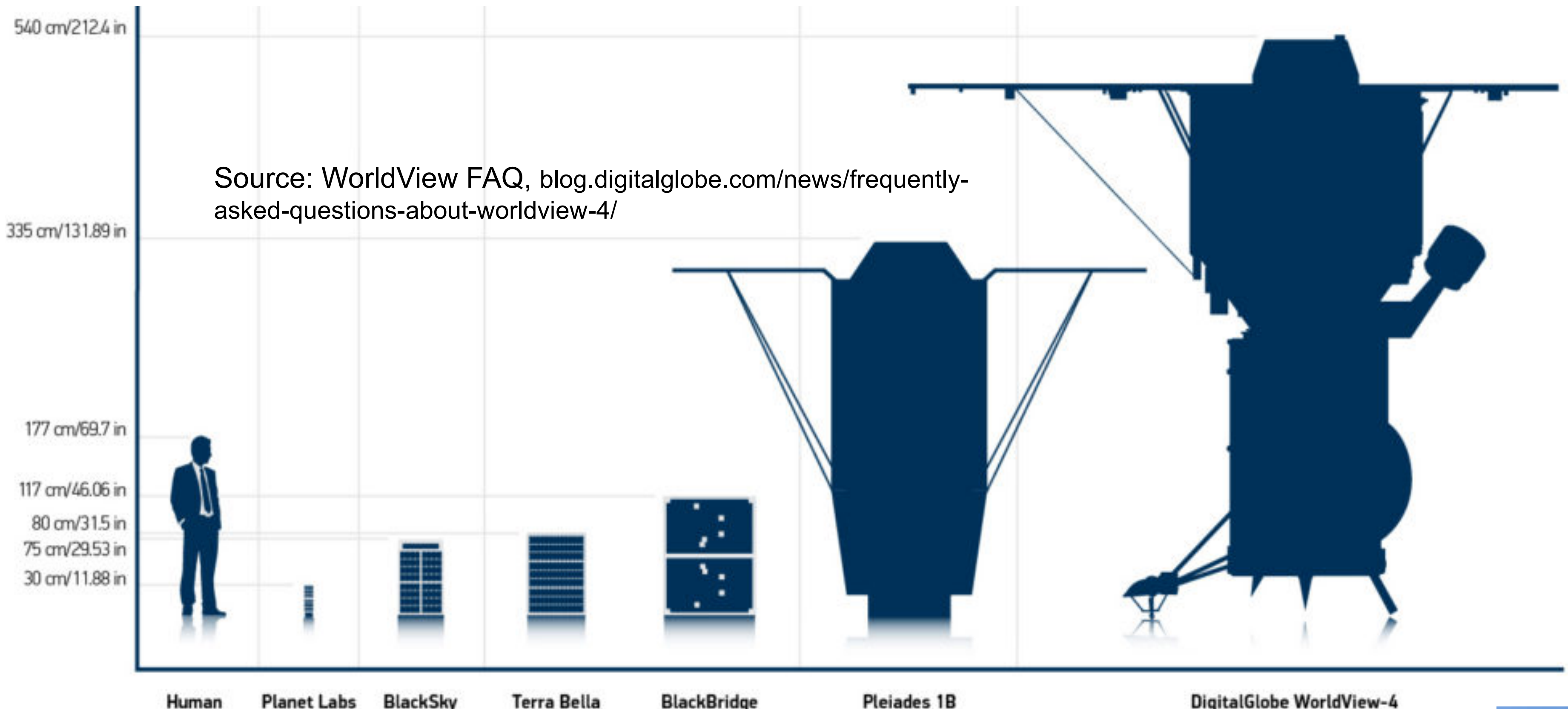
July 25, 2018, 4:00 AM CDT

The World Economy Runs on GPS. It Needs a Backup Plan

Source: *https://www.bloomberg.com/news/features/2018-07-25/the-world-economy-runs-on-gps-it-needs-a-backup-plan*

Large Constellations of Small Satellites

- Hi-frequency (e.g., daily or hourly) time-series of imagery of entire earth
  – Monitor illegal fishing, forest fires, crops (2017 DARPA Geospatial Cloud Analytics)
- Large Constellations
  – 2017: Planet Labs: 100 satellites: daily scan of Earth at 1m resolution in visible band



Source: WorldView FAQ, blog.digitalglobe.com/news/frequently-asked-questions-about-worldview-4/

Cheap (or free) satellite data on cloud computers

- 2008: USGS gave away 35-year LandSat satellite imagery archive
  - Analog of public availability of GPS signal in late 1980s
- 2017: Many cloud-based Virtual collaboration environment
  - Explosion in machine learning on satelliite imagery to map crops, water, buildings, roads, …

	Google Earth Engines	NEX	AWS Earth
Elevation, Landsat, LOCA, MODIS, NAIP	x	x	x
NOAA	x		x
AVHRR, FIA, GIMMM, GlobCover, NARR, TRIMM, Sentinel-1	x	x	
IARPA, GDELT, MOGREPS, OpenStreetMap, Sentinel-2, SpaceNet (building/road labels for ML)			x
CHIRPS, GeoScience Australia, GSMap, NASS, Oxford Map, PSDI, WHRC, WorldClim, WorldPop, WWF,	x		
BCCA, FLUXNET		x	

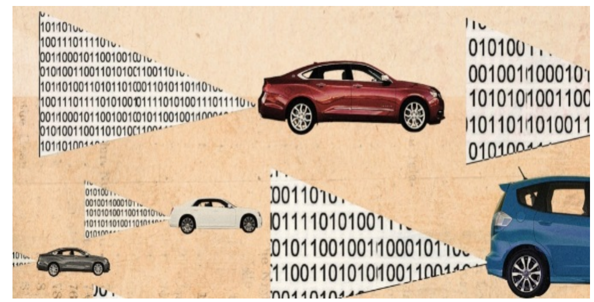
Spatial Big Data has Big Value

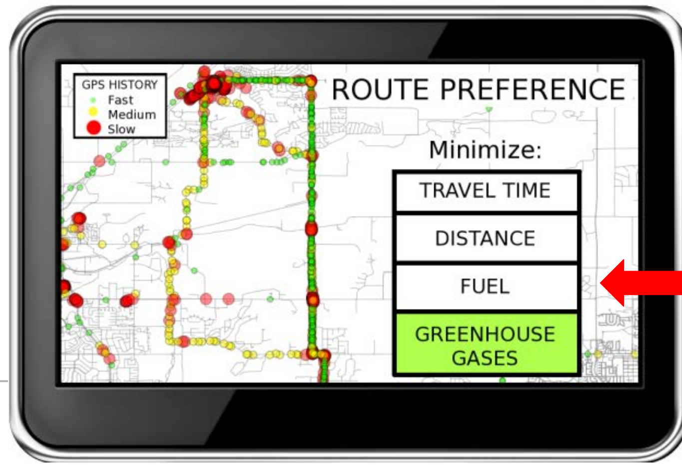
The New York Times

New Ways to Exploit Raw Data May Bring Surge of Innovation, a Study Says (May 13, 2011)

The study estimates that the use of personal location data could save consumers worldwide more than $600 billion annually by 2020. Computers determine users' whereabouts by tracking their mobile devices, like cellphones. The study cites smartphone location services including Foursquare and Loopt, for locating friends, and ones for finding nearby stores and restaurants.

But the biggest single consumer benefit, the study says, is going to come from time and fuel savings from location-based services — tapping into real-time traffic and weather data — that help drivers avoid congestion and suggest alternative routes. The location tracking, McKinsey says, will work either from drivers' mobile phones or GPS systems in cars.





The New York Times

U.P.S. Embraces High-Tech Delivery Methods (July 12, 2007)

*By "The research at U.P.S. is paying off. ……..— saving roughly **three million gallons of fuel** in good part by mapping routes that **minimize left turns**."*

Spatial Big Data is transforming our Society!

ESRI ArcGIS™ X

SaTScan™
Software for the spatial, temporal, and space-time scan statistics

OGC™
Open Geospatial Consortium, Inc.

ORACLE SPATIAL

IBM Smarter Planet

vs

OpenStreetMap

Ushahidi

waze
Outsmarting Traffic, Together.

Google Earth Engine

MONSANTO

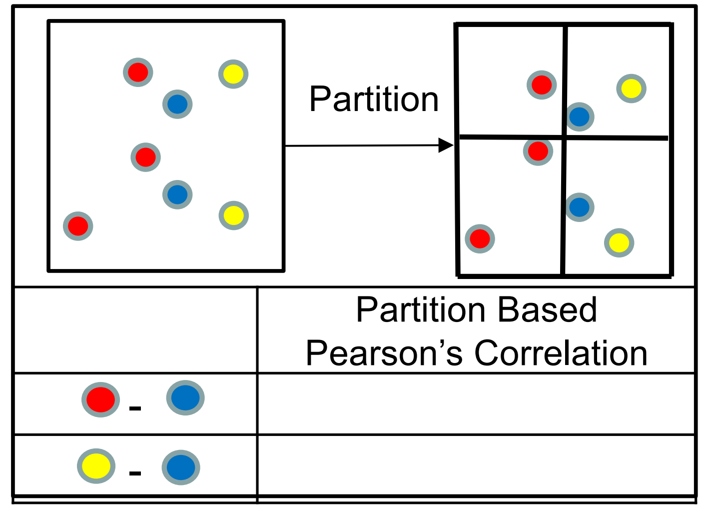
THE CLIMATE CORPORATION

A few Questions in Transportation Domain

Role	Questions	Pattern Family
Traveler, Commuter	What will be the travel time on a route?	Prediction
Transportation manager	Which corridors are accident-prone?	Hotspot
	Where and when are traffic flow anomalies?	Spatial Outlier
Traffic engineering	Which loop detector stations are very different from their neighbors?	Spatial Outlier
	Where are the congestion (in time and space)?	Hotspot
Planner and researchers	What will be travel demand in future?	Prediction
	How many trucks are there in a parking lot?	Prediction
	What road types are co-located? Where are they?	Co-location
Vehicle engineers	Which locations have high NOx emission? What is co-located there?	Hotspot, Co-location

Spatial Data Mining

- Challenge:
  - (Data Volume) >> (Number of Human Analysts)
  - Need automated methods
  - Need tools to amplify human capabilities



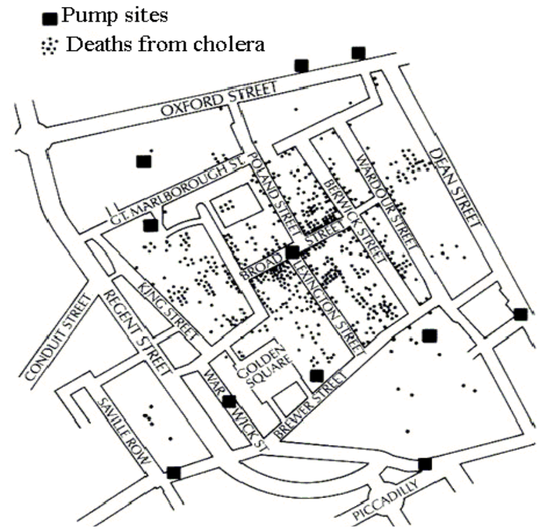
- Spatial Data are ubiquitous & important
- Current Data Science Tools are inadequate
  - Gerrymandering, Spatial Auto-correlation, …
- Practitioners in fields including:
  - Transportation, agriculture, weather, environment, …

Details: *A UCGIS Call to Action: Bringing the Geospatial Perspective to Data Science Degrees and Curricula*.
https://www.ucgis.org/index.php?option=com_dailyplanetblog

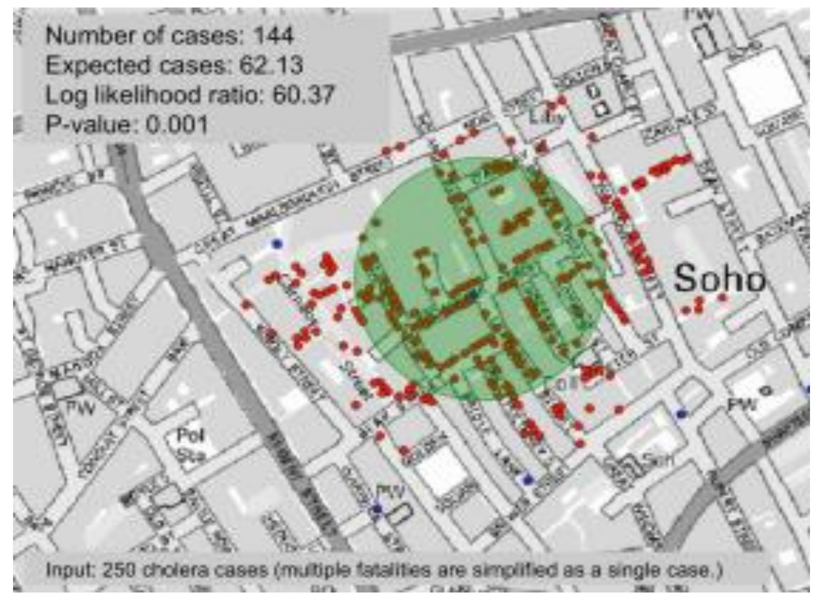
University Consortium for
GEOGRAPHIC INFORMATION SCIENCE

Defining Spatial Data Mining

- The process of discovering
  - interesting, useful, non-trivial patterns
    - patterns: non-specialist
    - exception to patterns: specialist
  - from large spatial datasets



- Spatial pattern families
  - A. Hotspots, Spatial clusters
  - B. Spatial outlier, discontinuities
  - C. Co-locations, co-occurrences
  - D. Spatial classification, prediction
  - E. Object detection
  - F. …



SaTScan Result

Xie, Y., Eftelioglu, E., Ali, R.Y., Tang, X., Li, Y., Doshi, R. and Shekhar, S., 2017. Transdisciplinary Foundations of Geospatial Data Science. *ISPRS International Journal of Geo-Information*, *6*(12), p.395.

Shekhar, S., Evans, M.R., Kang, J.M. and Mohan, P., 2011. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), pp.193-214.

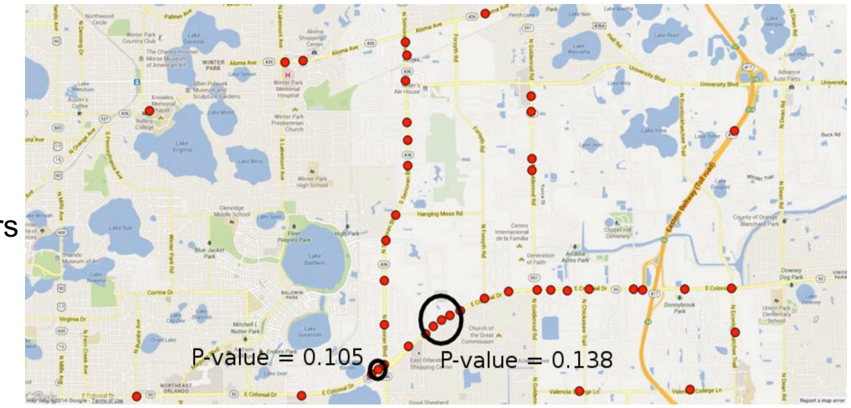
A. Hotspots, Spatial clusters

- **Question:** Which corridors are accident-prone?
- **Data:**
  - 43 Pedestrian fatalities in Orlando, FL (2000-9)
  - USDOT Fatality Analysis Reporting System

https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

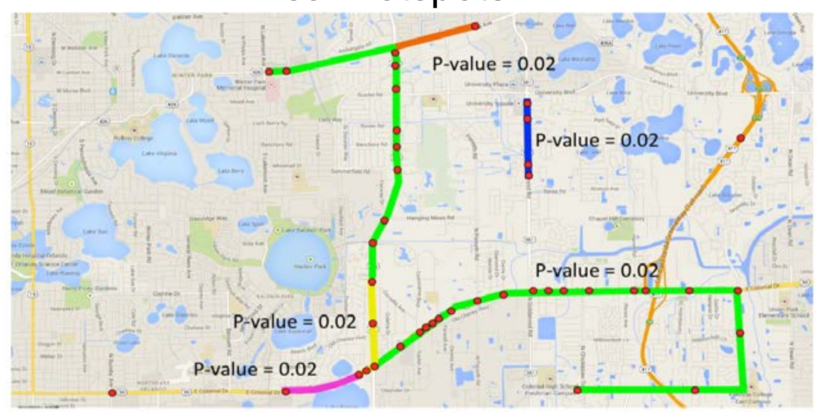
- **Patterns:**
  - Circular results from SaTScan
  - Linear hotspots
- **Interpretation:**

SaTScan Result



Unsafe pedestrian walkway

Linear hotspots



Details: Significant Linear Hotspot Discovery (X. Tang et al.). *IEEE Transactions on Big Data*, *3*(2), pp.140-153, 2017.

UNIVERSITY OF MINNESOTA
Driven to Discover®

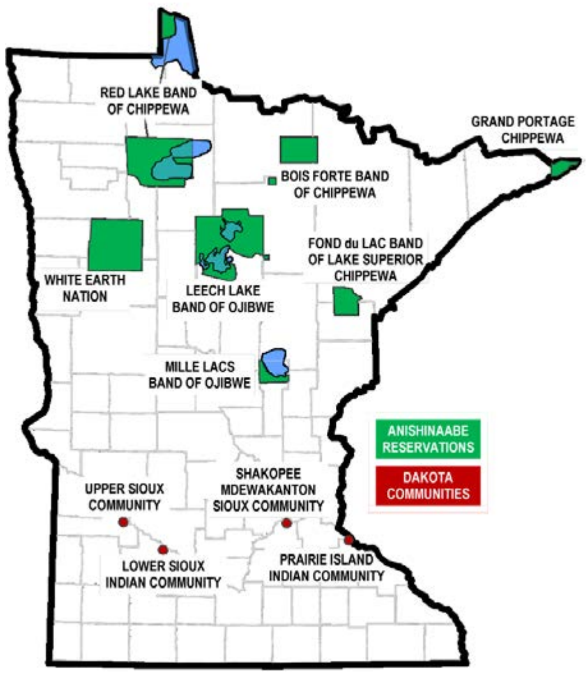
Minnesota Examples

LOCAL

Report shows that pedestrian safety is a major concern on Minnesota's American Indian reservations

More residents get around on foot, often on well-traveled roads

By Kelly Smith | FEBRUARY 18, 2019 — 5:25PM



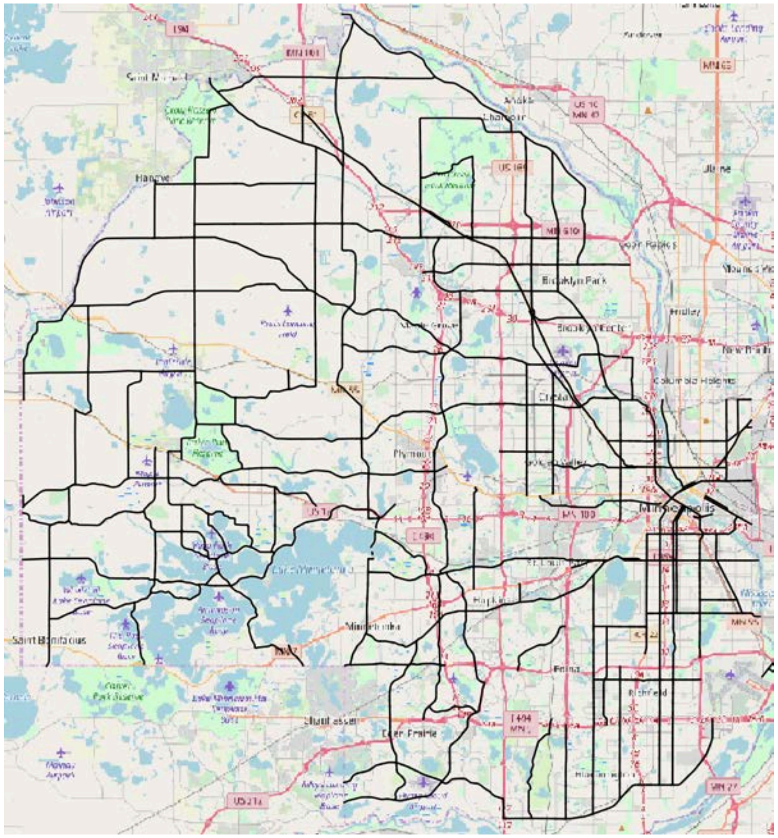
https://www.researchgate.net/figure/Location-of-reservations-in-Minnesota-Source-Indian-Affairs-Council-of-State-of_fig3_328759103

http://www.startribune.com/report-shows-that-pedestrian-safety-is-a-major-concern-on-minnesota-s-american-indian-reservations/505941632/

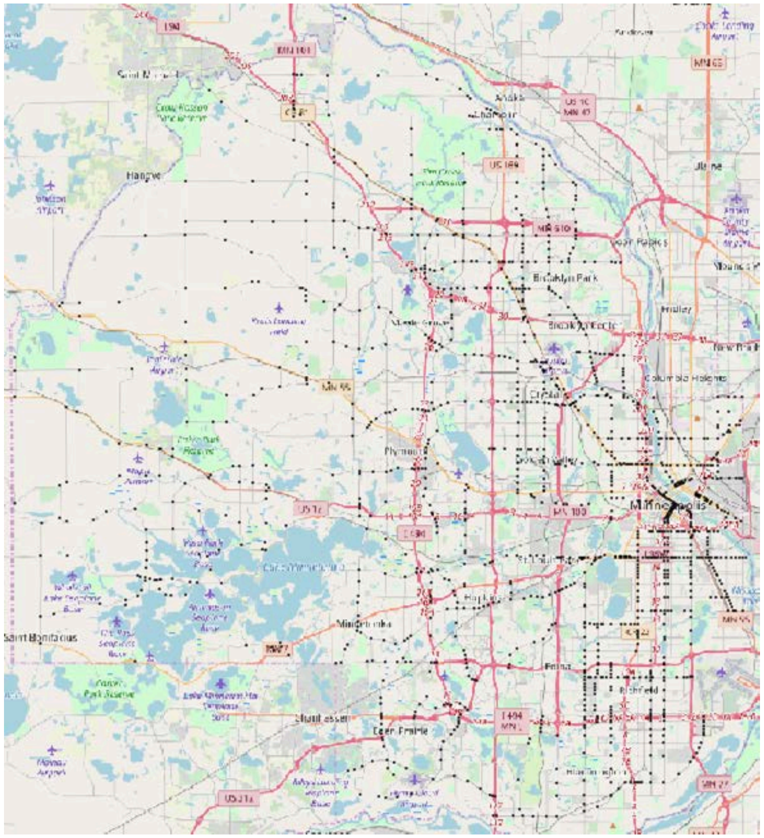
https://www.completecommunitiesde.org/planning/complete-streets/winter-maintenance-2/

A. Hotspots, Spatial clusters:
Case Study on Hennepin County Crashes

- **Question:** Which corridors are accident-prone?
- **Data:**
  - 1345 crashes on Hennepin County road intersections (2010 - 2015)
  - Source: Hennepin County Public Works



Major road network



Crashes (black dots)

Data Source: https://www.hennepin.us/business/work-with-henn-co/transportation-planning-design

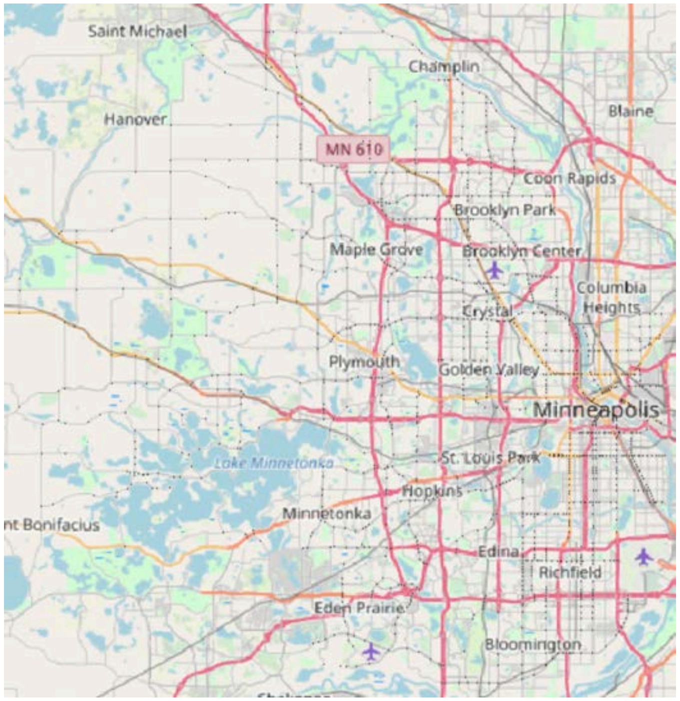
A. Hotspots, Spatial clusters:
Case Study on Hennepin County Crashes

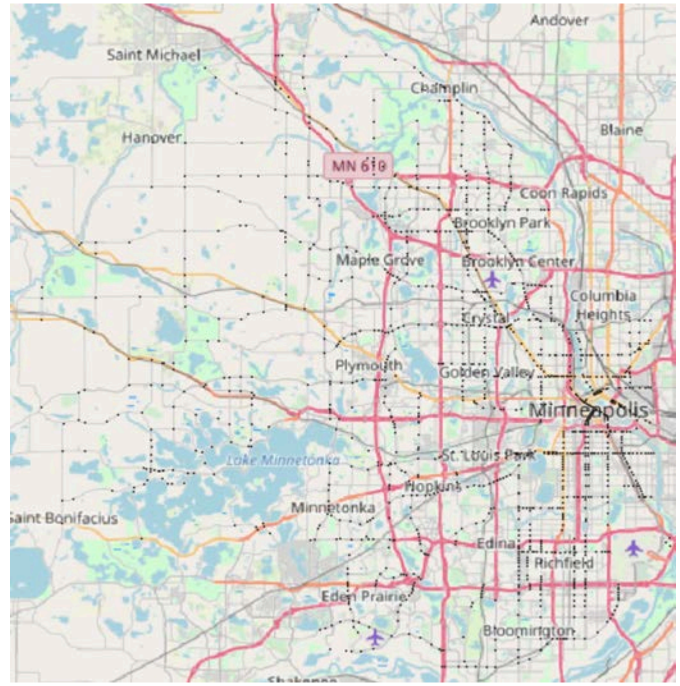
- **Data:**
  - 1345 crashes on Hennepin County major intersections (2010-2015)
  - Source: Hennepin County PWD
- **Patterns:**
  - Linear hotspots (p-value = 0.05)
    - Minimum length: 500 meters
    - No turns over 45 degrees in the path (constrained on single street)
- **Interpretation:**
  - Intersections to corridors
  - Feasibility study
- **Next:**
  - Include other roads
  - Consider traffic volume

Data Source: https://www.hennepin.us/business/work-with-henn-co/transportation-planning-design

Dot sizes fool human eye but not algorithms

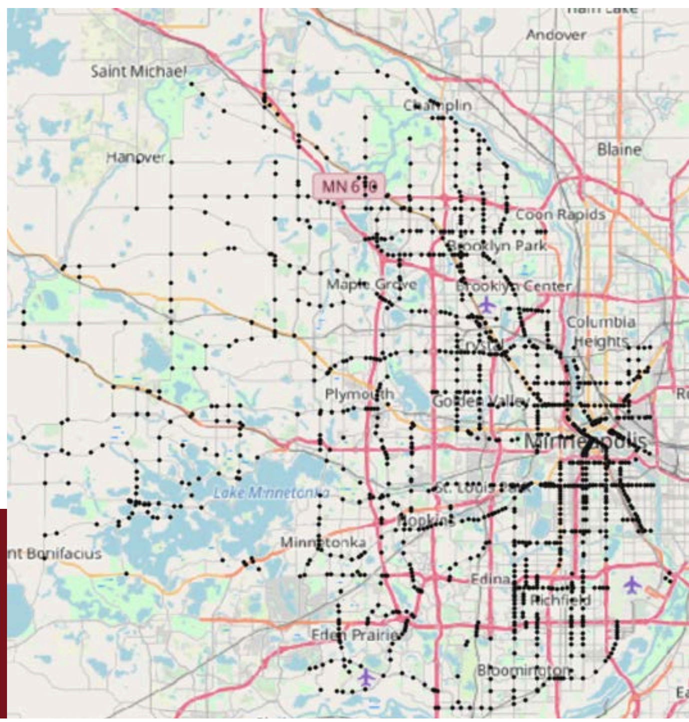
dot size = 0.25

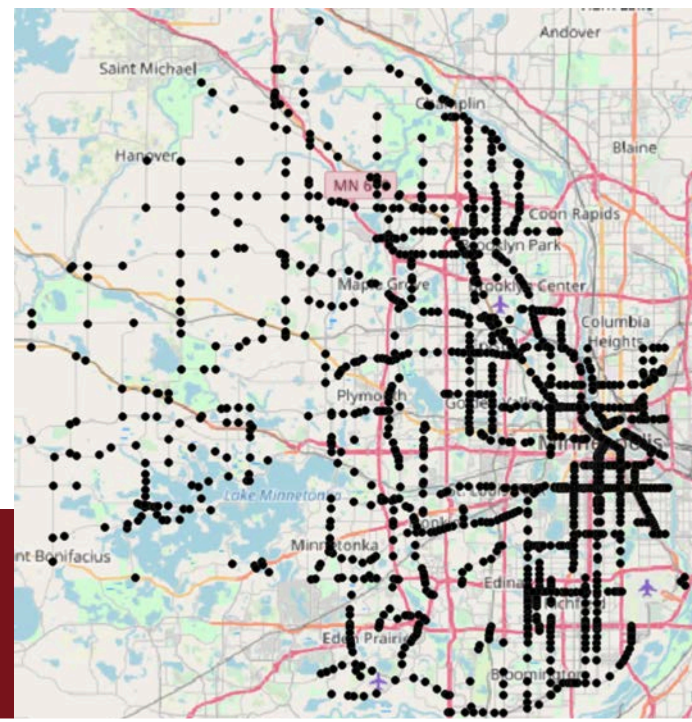




dot size = 0.5

dot size = 1





dot size = 2

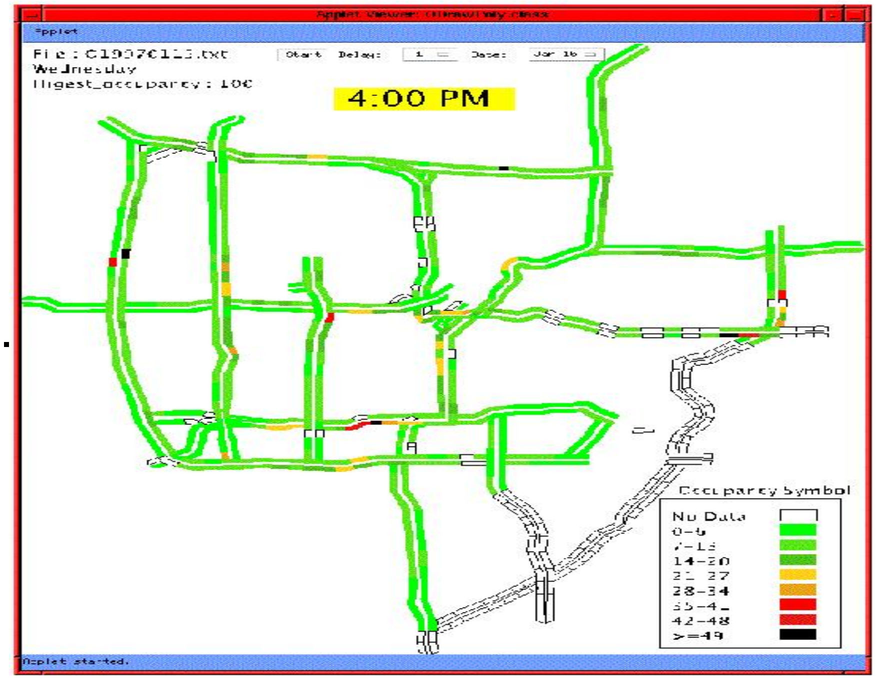
B. Spatial outlier, Discontinuities

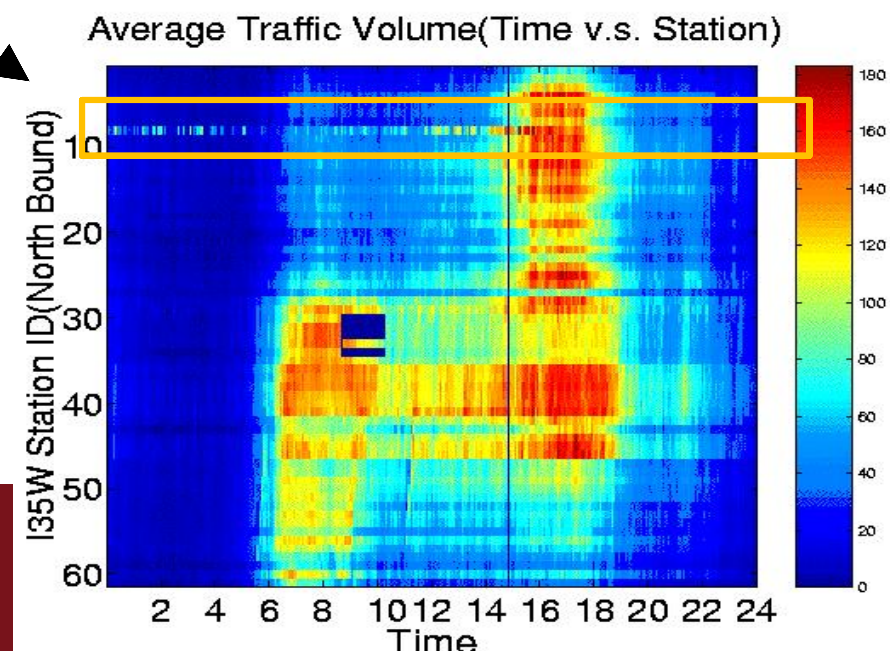
- **Question:** Which loop detector stations are very different from their neighbors?

- **Data:**
  - 900 stations (with 1 to 4 loop detectors each).

- **Pattern:**
  - Spatial outlier at Station 9.

- **Interpretation:**
  - Hypothesis: faulty loop detector?
  - Action: Test station 8 detectors





Details: *A unified approach to detecting spatial outliers*. GeoInformatica, (S. Shekhar et al.), 7(2), Springer, 2003 (Summary in ACM SIGKDD '01).

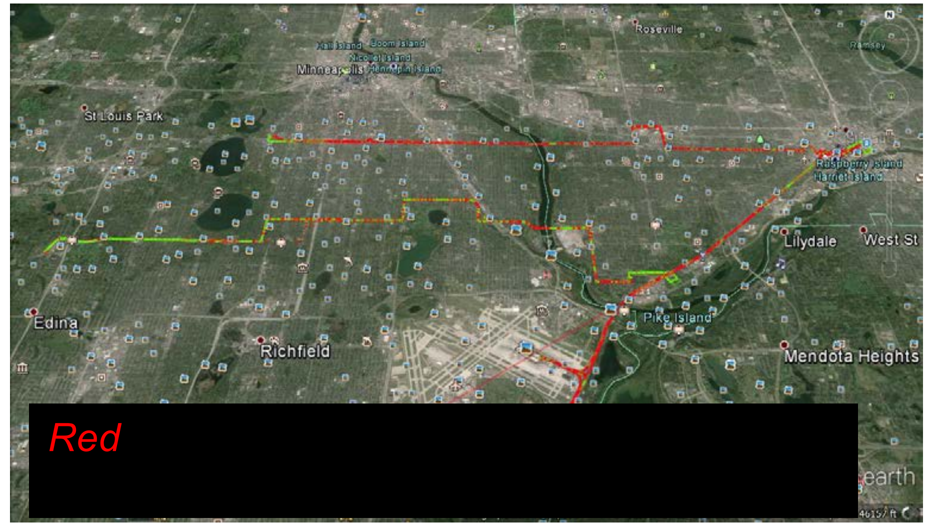
Discovering Sub-time-series Co-occurrence Patterns of Non-compliance

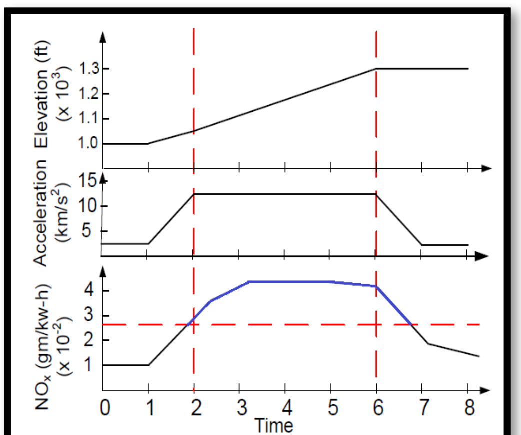
Given:

- A set of multivariate event trajectories and a set of non-compliant windows
- A cross-k function threshold ε
- A time lag δ
- A minimum support threshold *minsupp*

Find:

- Co-occurrence patterns whose cross-K function at distance δ exceeds ε and whose support exceed minsupp





ID	Co-occurrence Pattern C	$\hat{K}_{C,W_N}(2)$
1	Wheel speed: $\{w_0\ w_0\ w_0\ w_1\ w_2\}$	21.57
2	Engine RPM: $\{s_1\ s_2\ s_3\ s_3\ s_3\}$ Engine power: $\{r_5 r_5 r_5 r_5 r_5\}$ Wheel speed: $\{w_0 w_0 w_0 w_0 w_0\}$ Acceleration: $\{a_{16}\ a_{16}\ a_{17}\ a_{17}\ a_{17}\}$	16.28
3	Engine RPM: $\{s_1\ s_1\ s_2\ s_3\ s_3\}$ Engine power: $\{r_5 r_5 r_5 r_5 r_5\}$ Wheel speed: $\{w_1\ w_0 w_0 w_0 w_0\}$	17.15

University of Minnesota
Driven to Discover

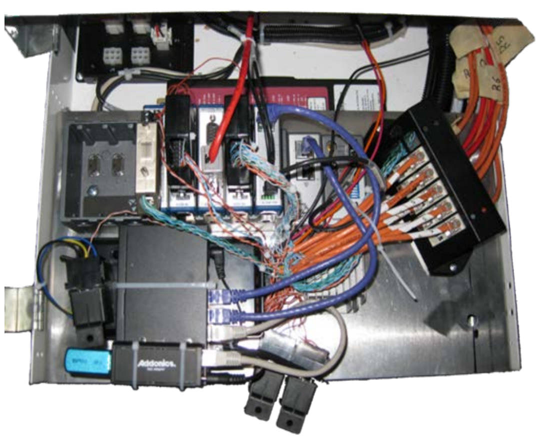
Reem Y. Ali, Venkata M.V. Gunturi, Andrew J. Kotz, Emre Eftelioglu, Shashi Shekhar, and William F. Northrop "*Discovering Non-compliant Window Co-Occurrence Patterns.*" GeoInformatica, 21(4), 829-866 (2017), Springer.

C. Hotspots, Co-locations, Co-occurrences

- **Question:** Where are high transit-NOx emissions? What is co-located there?

- **Data:**
  - On Board Diagnostics Data from Metro-Transit Buses

Metro Transit

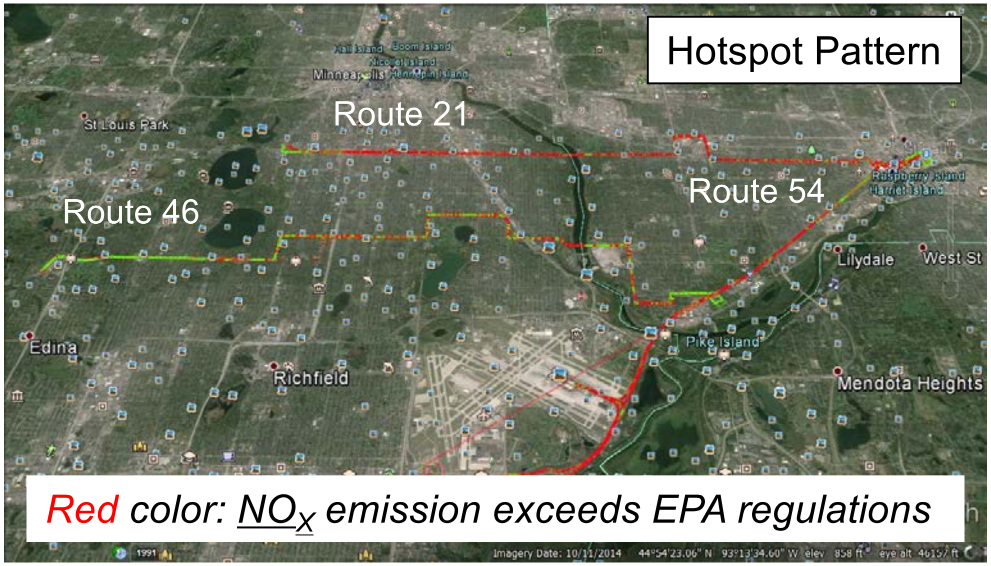


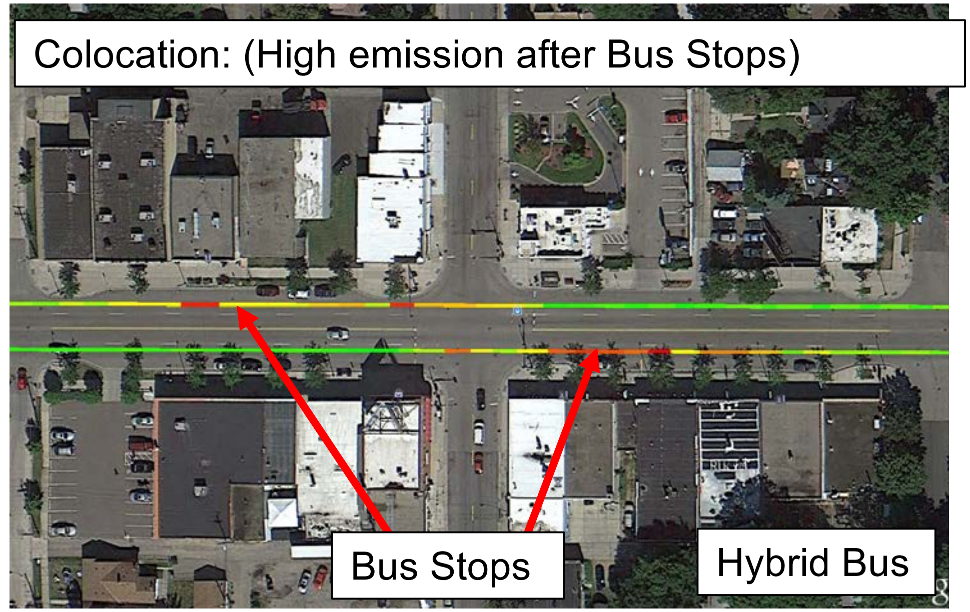
Variables sampled every second:

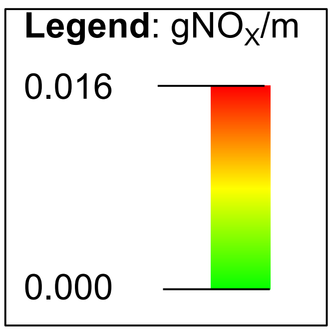
- GPS location
- Speed
- Vehicle Load
- Engine and Heater Fuel Flow
- Exhaust Temp and Mass Flow
- Intake Temp And Mass Flow
- Engine Torque and RPM
- Engine Coolant Temp
- Odometer
- **NOx emission**
-
-
- ….measurements on 200+ variables

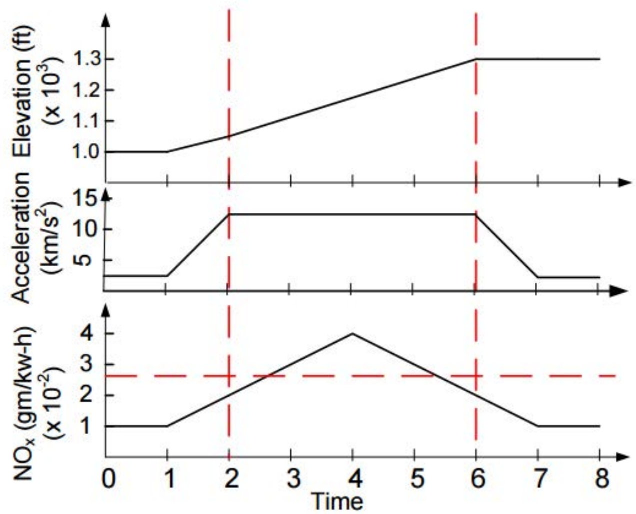
Details: "*Discovering Non-compliant Window Co-Occurrence Patterns.*" (R. Ali et al.) GeoInformatica, 21(4): 829-866, Springer, 2017

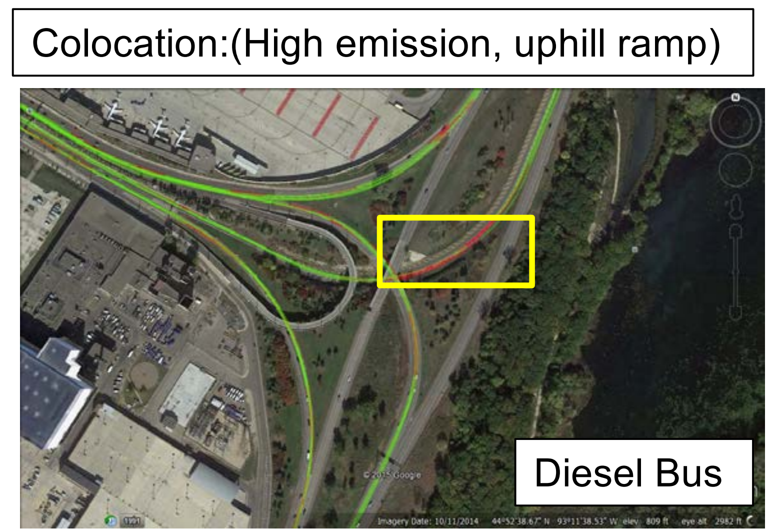
C. Emission Hotspots, Co-locations











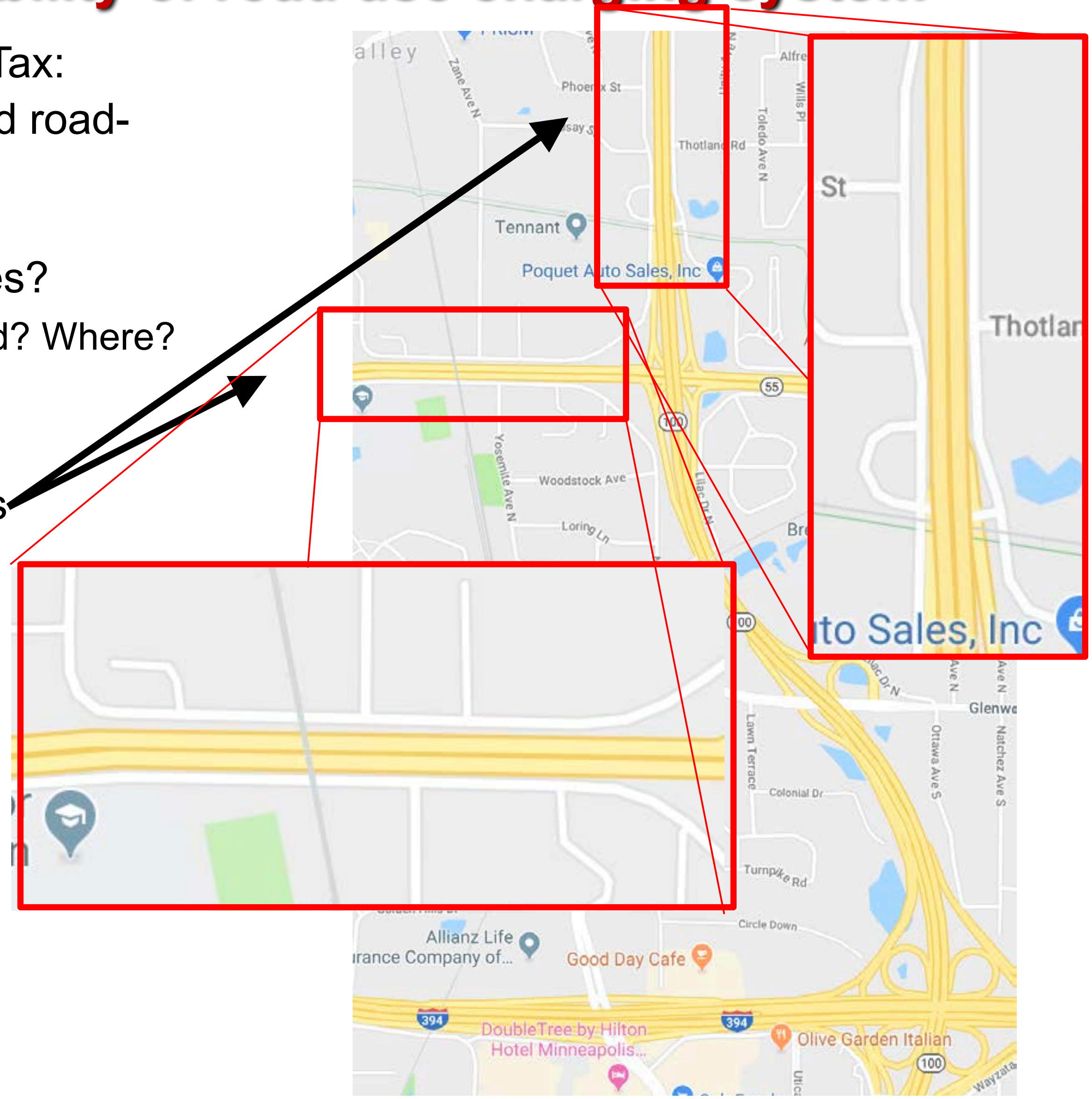
Details: "Discovering non-compliant window co-occurrence patterns: A summary of results." R. Ali et al., Proc. Intl. Symp. on Spatial and Temporal Databases,,pp. 391-410. Springer, 2015.

UNIVERSITY OF MINNESOTA
Driven to Discover®

C. Co-locations, Co-occurrences

Case Study: Test feasibility of road use charging system

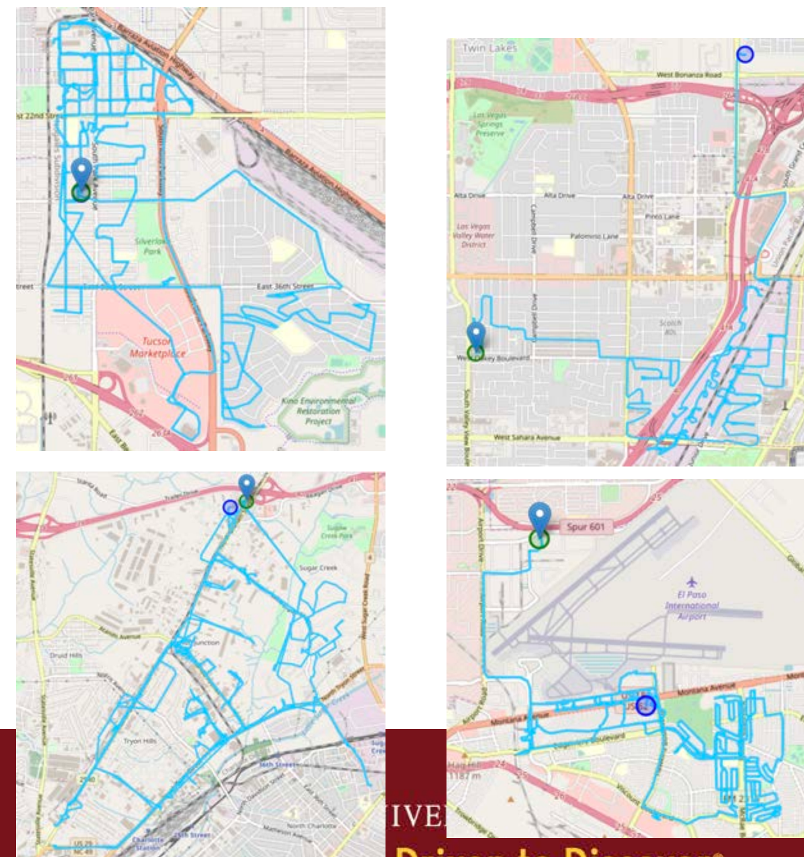
- **Use Case:** Impact of EV on Gas Tax:
- Test technology for road-type based road-usage based charging.
- **Q?** Can GPS distinguish road-types?
- Which road types are closely co-located? Where?
- **Input:** Road map with road-types
- **Pattern:** Co-location of road-types



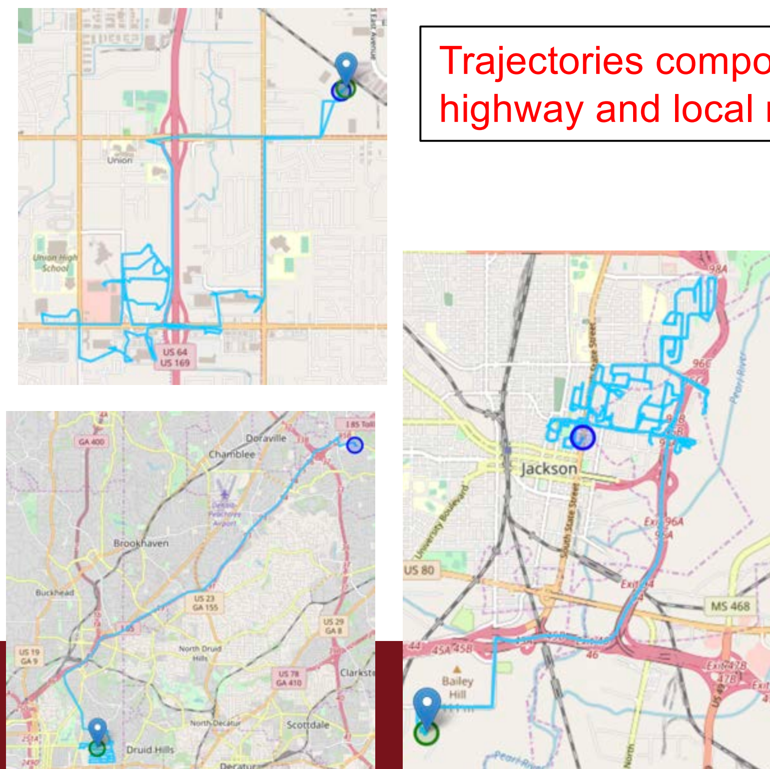
D. Spatial Classification, Prediction

- **Question:** Are there natural groups for UPS delivery trajectories?
- **Data:** A set of historical trajectories with on-board diagnostic data from UPS trucks.
- **Pattern:** Clusters of trajectories with similar spatial properties.
- **Interpretation:** Delivery zones are small, but the distance between each delivery zone and UPS depots is different.

Trajectories composed of only local road trips



Trajectories composed of highway and local road trips



Li, Y., Shekhar, S., Wang, P. and Northrop, W., 2018, November. Physics-guided energy-efficient path selection: a summary of results. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 99-108). ACM.

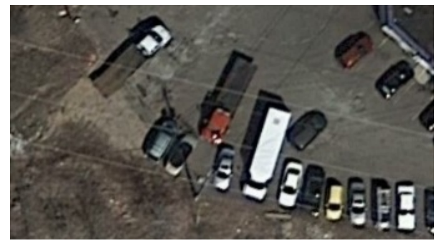
E. Geospatial Object Detection

- **Q:?** How many trucks are there in a lot? City?
- **Ex.:** Estimate truck supply in a city (CH Robinson)
- **Data:**
  - Aerial imagery (3 inch pixels)
    - Hennepin & Ramsey counties
  - NAIP Imagery (1 meter pixels, 2017)
    - MA Buildings Dataset. https://www.cs.toronto.edu/~vmnih/data/
- **Pattern:**
  - Detected geospatial objects
    - Cars, trucks,
    - Houses, …

car truck

Input training image

Input training MOBRs



Test image

Output MBRs

YOLO (baseline)

Proposed method

Xie, Y., Bhojwani, R., Shekhar, S. and Knight, J., 2018. An unsupervised augmentation framework for deep learning based geospatial object detection: a summary of results. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 349-358). ACM.

Data Science Education - Nationwide

DOI:10.1145/3188721

Data science promises new insights, helping transform information into knowledge that can drive science and industry.

BY FRANCINE BERMAN, ROB RUTENBAR, BRENT HAILPERN, HENRIK CHRISTENSEN, SUSAN DAVIDSON, DEBORAH ESTRIN, MICHAEL FRANKLIN, MARGARET MARTONOSI, PADMA RAGHAVAN, VICTORIA STODDEN, AND ALEXANDER S. SZALAY

Realizing the Potential of Data Science

Berman F. et al.,
Realizing the Potential of Data Science, Communications of the ACM,
April 2018, Vol. 61 No. 4, pp. 67-72,
10.1145/3188721

Teaching Data Science: Many Flowers Blooming

- **University of California, Berkeley**:
  - Recently established division of data science (same level as college and school)
  - Opened Introductory, foundational, and advanced courses.
  - <u>Undergraduate</u> program in Data Science

- **University of Michigan, Ann Arbor**:
  - <u>Undergraduate</u> program in Data Science

- **Columbia University**:
  - <u>Master</u> of Data Science offered by Data Science Institute

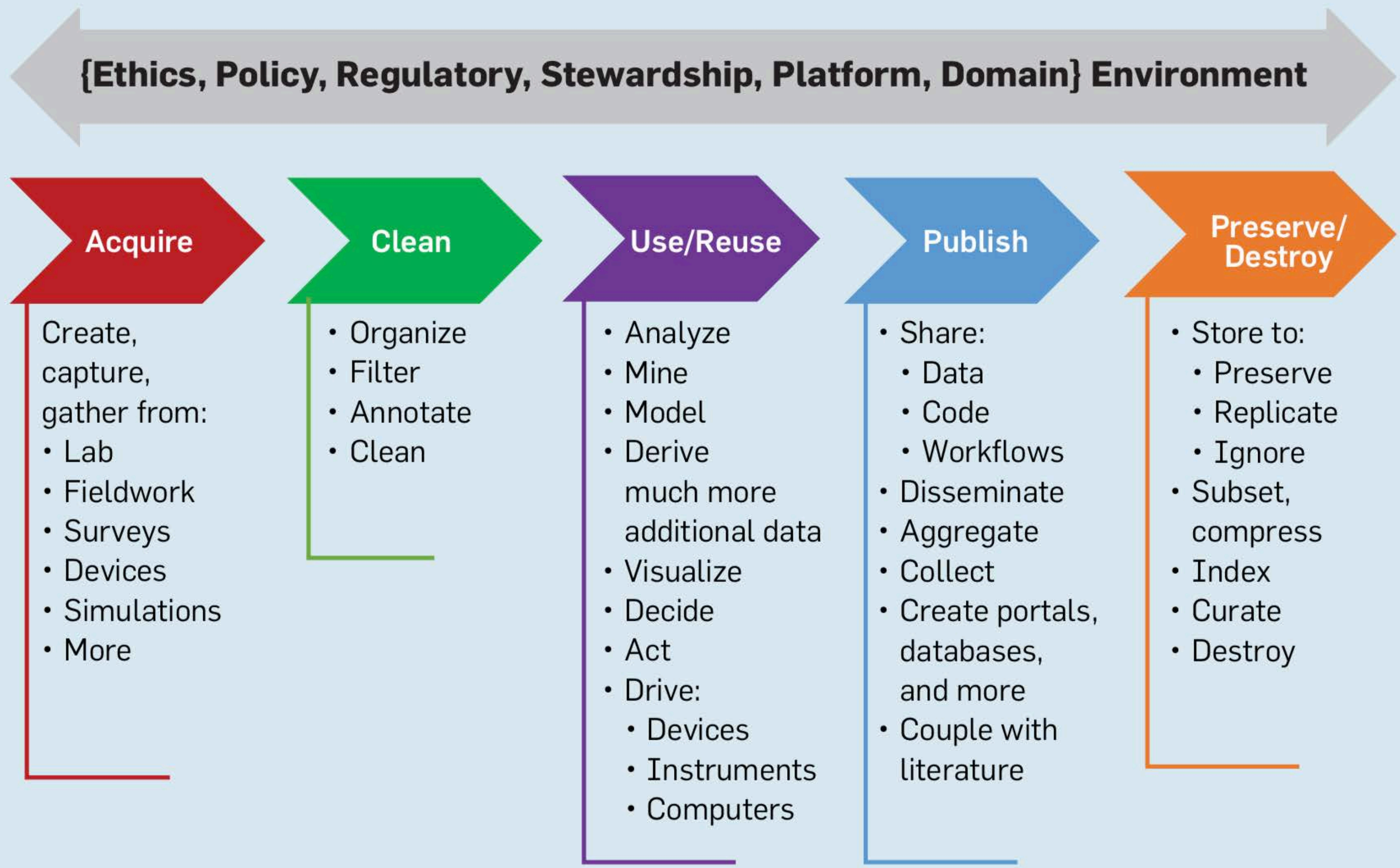
- **University of Illinois, Urbana-Champaign**:
  - <u>Master</u> of Computer Science in Data Science offered as an online professional course

- **University of Chicago**:
  - <u>Master</u> of Science in Computational Analysis and Public Policy program

Berman F. et al.,
Realizing the Potential of Data Science, Communications of the ACM,
April 2018, Vol. 61 No. 4, pp. 67-72, 10.1145/3188721

Data Life Cycle

The data life cycle and surrounding data ecosystem from the *Realizing the Potential of Data Science Report.*[2]



Berman F. et al.,
Realizing the Potential of Data Science, Communications of the ACM,
April 2018, Vol. 61 No. 4, pp. 67-72, 10.1145/3188721

University of Minnesota
Driven to Discover®

Data Science Skills

The data life cycle

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Acquire → **Clean** → **Judicious use/reuse** → **Publish** → **Preserve & Destroy**

Acquire
- Survey
- Sensor
- Citizen Science

Clean
- Filter
- Annotation

Judicious use/reuse
- Coding
- Querying
- Machine learning
- Data mining
- Statistics
- Optimization
- Visualization
- Spatial data analysis
- Interpretation
- Decision Making

Publish
- Portal
- Share

Preserve & Destroy
- Curation
- Indexing

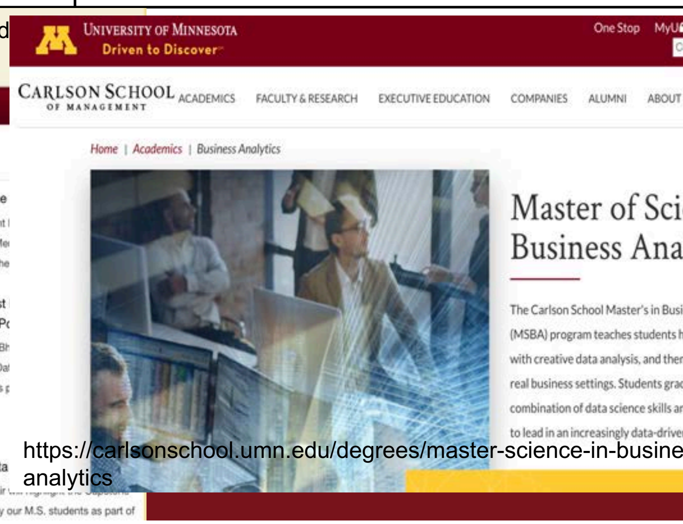
Data Science Tools

Skills	Tools
Coding	• Python • Matlab
Querying	• SQL • Hive
Machine learning	• Scikit-learn • Tensorflow • Mllib for Spark
Data mining	• Rapid miner • Oracle data mining • Weka
Statistics	• R • SAS
Optimization	• Cplex • GAMS • GUrobi
Spatial data analysis	• ArcGIS • QGIS • SaTScan

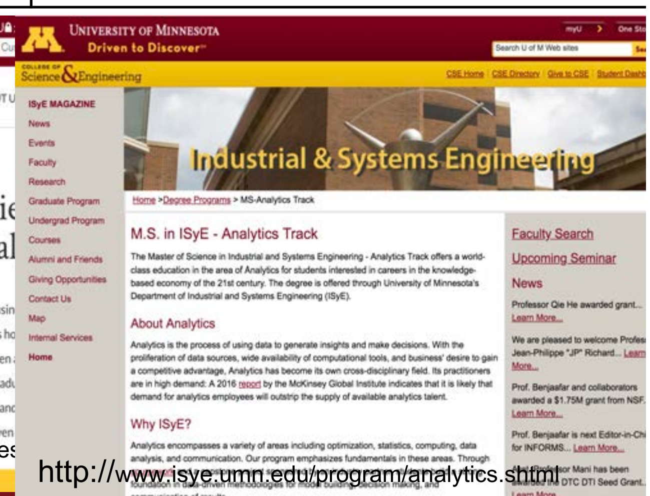
Education in Data Science - UMN

Name of Degrees		Focused skills	Name of Schools
Bachelor	Coming soon		College of Science & Engineering College of Liberal Arts School of Public Health
Certificate (12 credits)	Post-Baccalaureate Certificate in Data Science	Coding, Querying, Machine learning, Data mining	
Master (31 credits)	Master's of Science in Data Science		
	Master of Science in Business Analytics	Interpretation, Decision making	Carlson School of Management
	M.S. in Industrial and Systems Engineering - Analytics Track	Optimization, Decision making	College of Science and Engineering • Department of Industrial and Systems Engineering (ISyE)

https://www.cs.umn.edu/ad duate/ms_data_science



https://carlsonschool.umn.edu/degrees/master-science-in-busines analytics



http://www.isye.umn.edu/program/analytics.shtml

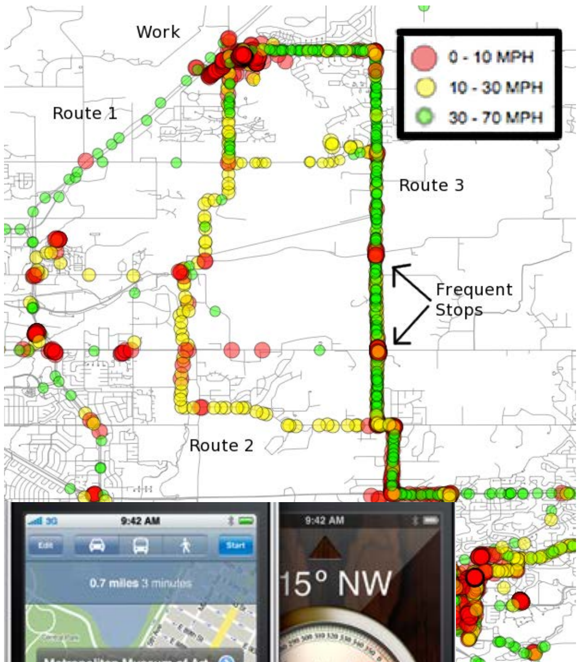
References

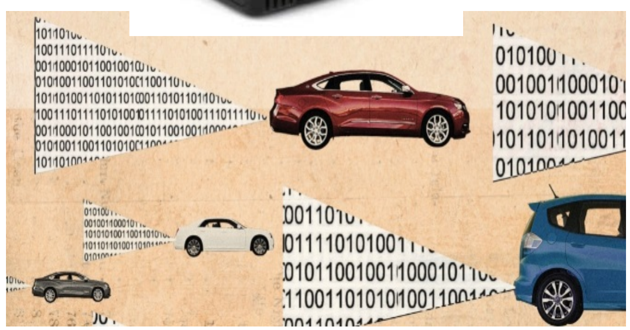
- Shekhar, S., Feiner, S.K. and Aref, W.G., 2016. Spatial computing. *Commun. ACM*, *59*(1), pp.72-81.
- Yiqun Xie, Jayant Gupta, Yan Li and Shashi Shekhar. Transforming Smart Cities with Spatial Computing. Accepted at: IEEE International Smart Cities Conference (ISC2 2018), Kansas City, MO, Sep. 2018.
- Xie, Y., Eftelioglu, E., Ali, R.Y., Tang, X., Li, Y., Doshi, R. and Shekhar, S., 2017. Transdisciplinary Foundations of Geospatial Data Science. *ISPRS International Journal of Geo-Information*, *6*(12), p.395.
- Shekhar, S., Evans, M.R., Kang, J.M. and Mohan, P., 2011. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), pp.193-214.
- Tang, X., Eftelioglu, E., Oliver, D. and Shekhar, S., 2017. Significant Linear Hotspot Discovery. *IEEE Transactions on Big Data*, *3*(2), pp.140-153.
- Oliver, D., Shekhar, S., Zhou, X., Eftelioglu, E., Evans, M.R., Zhuang, Q., Kang, J.M., Laubscher, R. and Farah, C., 2014, September. Significant route discovery: A summary of results. In *International Conference on Geographic Information Science* (pp. 284-300). Springer, Cham.
- S. Shekhar, C.T. Lu, and P. Zhang. *A unified approach to detecting spatial outliers*. GeoInformatica, 7(2), 2003 (Earlier version appeared in SIGKDD '01). Springer.
- Reem Y. Ali, Venkata M.V. Gunturi, Andrew J. Kotz, Emre Eftelioglu, Shashi Shekhar, and William F. Northrop "*Discovering Non-compliant Window Co-Occurrence Patterns.*" GeoInformatica, 21(4), 829-866 (2017), Springer.
- Li, Y., Shekhar, S., Wang, P. and Northrop, W., 2018, November. Physics-guided energy-efficient path selection: a summary of results. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 99-108). ACM.
- Xie, Y., Bhojwani, R., Shekhar, S. and Knight, J., 2018. An unsupervised augmentation framework for deep learning based geospatial object detection: a summary of results. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 349-358).
- Berman F. et al., *Realizing the Potential of Data Science, Communications of the ACM*, April 2018, Vol. 61 No. 4, pp. 67-72, 10.1145/3188721

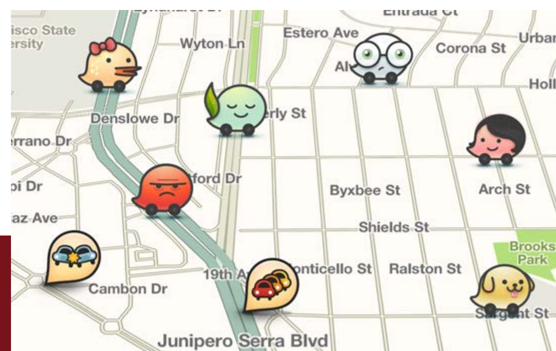
Spatial Big Data driven Eco-Routing

Spatially oriented datasets exceeding capacity of current routing systems

- Due to Volume, Velocity (Update-rate) and, Variety







Waze.com