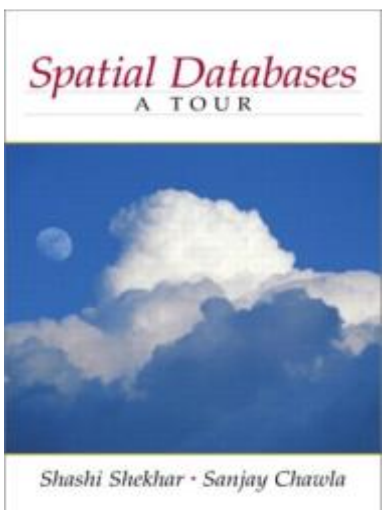
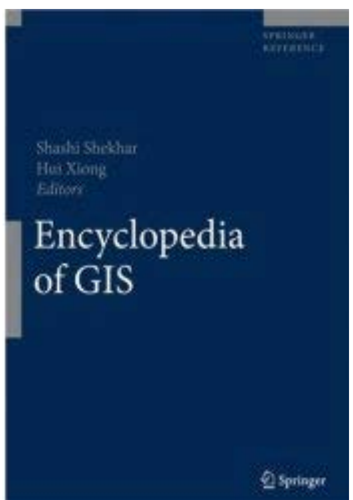
One Size Data Science Does Not Fit All Data:
What is Special about Spatial Data Science?

NSF ERC Planning [Workshop on Reimagining Road Infrastructure]
Oct.3rd-4th 2019, Alexandria, VA

Shashi Shekhar
Member, CRA Board & Midwest Big Data Hub Board
Former President, University Consortium for GIS
McKnight Distinguished University Professor, University of Minnesota
www.cs.umn.edu/~shekhar, shekhar@umn.edu





University Consortium for GEOGRAPHIC INFORMATION SCIENCE

A UCGIS Call to Action:
Bringing the Geospatial Perspective to Data Science Degrees and Curricula

Summer 2018

As a long-established information science discipline, the Geographic Information Science & Technology (GIS&T) community has key contributions to make to evolving data science curricula. This statement articulates the University Consortium for Geographic Information Science's (UCGIS) position for the academic GIS&T community and provides recommendations and action items for the benefit of both internal and external audiences. On May 22-24, 2018, UCGIS held its annual Symposium under the theme of *Frontiers of Geospatial Data Science*, coordinated this year with the AutoCarto conference of the Cartography and Geographic Information Society (CaGIS). Drawing from discussions at that event, together with many months of internal exchanges, UCGIS offers these statements for the benefit of its member organizations as well as the broader geospatial community. The goals of this white paper and its recommendations are to 1) describe and clarify the value of incorporating geospatial knowledge, skills, and data for students, employees, and employers within the emerging field of data science; 2) highlight potential pathways and opportunities for academic geospatial scientists to establish connections with data science programs and personnel on their university campuses; and 3) initiate a national dialogue about the synergistic benefits of mutually enriching data science and geospatial science curricula.

Context

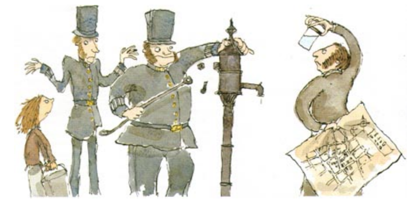
Virtually every sector of industry, business, government, and science is awash in data of great volume, variety, and velocity. In light of calls for fairness, accountability, transparency, and reproducibility, data accuracy and authority are also highly relevant. As an interdisciplinary field, there are high expectations for the capabilities of data science[1] to address myriad demands for innovative breakthroughs. "Data Scientist" has become an in-demand job title, though the nature of the positions varies widely. The most common skill sets required are analytical and quantitative in nature: to be able to manage and help others interpret large and diverse data sets.

Data that are geographically referenced or contain some type of location markers are both common and of high value (e.g., data subject to state-specific policies, laws and regulations; demographic data from the census; location traces of smartphones and vehicles; remotely sensed imagery from satellites, aircraft and small unmanned aerial vehicles; volunteered geographic information; geographically referenced social media postings). A 2011 McKinsey

[1] F. Berman et al., Realizing the Potential of Data Science, Communications of the ACM, 61(4):67-72, April 2018. DOI: 10.1145/3188721.

A Spatial Data Science Story

1854: What causes Cholera?

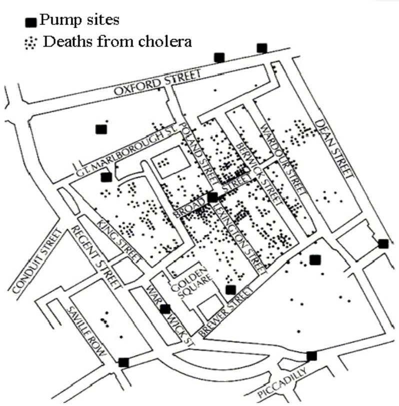


Collect & Curate Data → Discover Patterns, Generate Hypothesis → Test Hypothesis (Experiments) → Develop Theory

? water pump

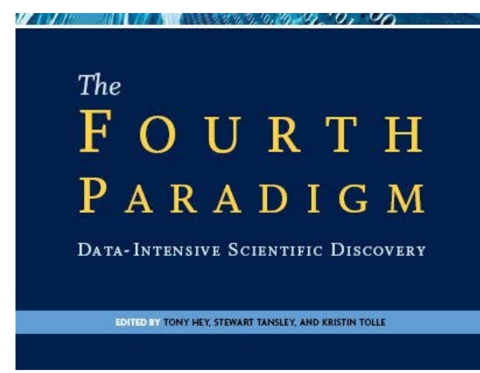
Remove pump handle

Germ Theory



Impact:
sewage system,
drinking water supply
…

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



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*

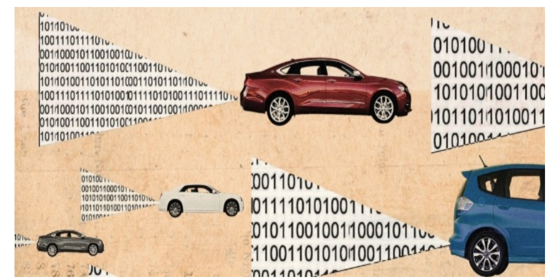
Spatial Big Data has Big Value

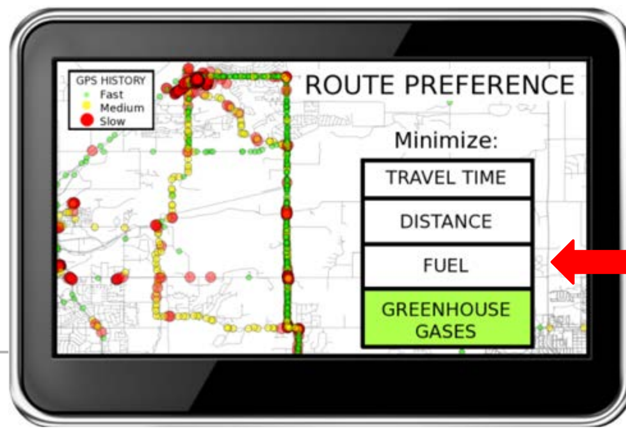
The New York Times

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.





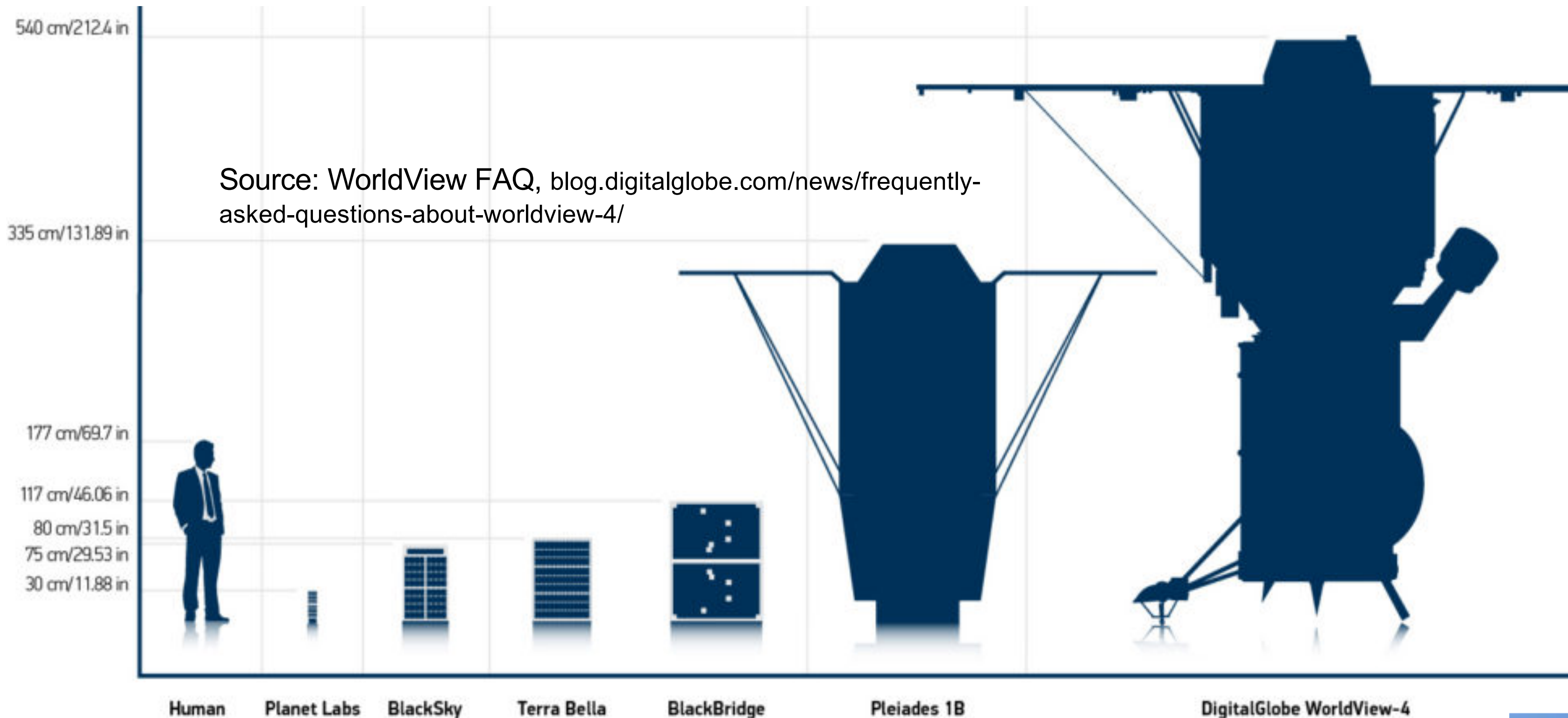
The New York Times

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**."*

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/

Spatial Computing Research Group

Easier Access: Cheap (or free) Cloud Repositories

- 2008: USGS gave away 35-year LandSat satellite imagery archive
  - Analog of public availability of GPS signal in late 1980s
- 2017: Cloud-based repositories of geospatial data
  - Explosion in machine learning on satellite imagery to map crops, water, buildings, roads,

	Google Earth Engine	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	

Google Earth Engine

NEX NASA Earth Exchange

Ground Truth Collection: Volunteered Geographic Information

- Context: Labeled data crucial for Machine Learning
- Last century: Ground Truth official, expensive, sparse
- Recent: Augment with Citizen Science: Zooinverse, GalaxyZoo, …
  - Limited in support for spatial data science
- Volunteered Geographic Information (VGI)
  - Undirected: Flickr, eBird, …
  - Directed: Ushahidi, GIS Corps, Open Street Map (OSM) …
  - OSM: Roadmaps for many country, e.g., Haiti Earth Quake (2009)

GIS Corp.

Ushahidi

Spatial Big Data, e.g., GPS trace

Ground Truth, e.g., VGI

Spatial Statistics, Spatial Data Mining

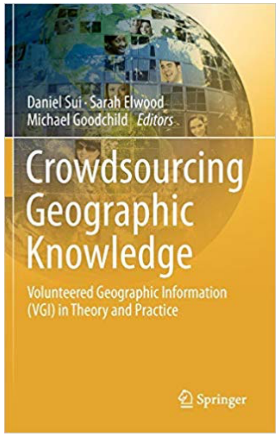
frontiers in Neuroinformatics

METHODS
published: 08 May 2019
doi: 10.3389/fninf.2019.00029

Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging

Anisha Keshavan[1,2,3,4], Jason D. Yeatman[3,4] and Ariel Rokem[1,2]*

[1] eScience Institute, University of Washington, Seattle, WA, United States, [2] Institute for Neuroengineering, University of Washington, Seattle, WA, United States, [3] Institute for Learning and Brain Sciences, University of Washington, Seattle, WA, United States, [4] Department of Speech and Hearing, University of Washington, Seattle, WA, United States



Spatial Big Data is transforming our Society!

ArcGIS X ESRI

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

OGC™
Open Geospatial Consortium, Inc.

ORACLE SPATIAL

IBM Smarter Planet

vs

waze
Outsmarting Traffic, Together.

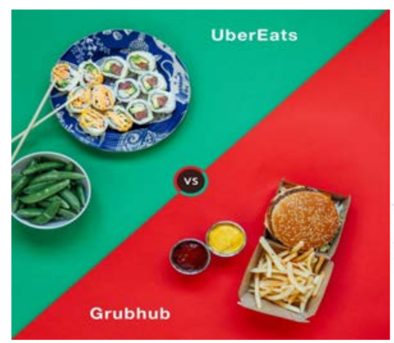
Google Earth

OpenStreetMap

Ushahidi

Leading Market Players

LimeBike Lime
SPIN Spin
Bird



Google Earth Engine

MONSANTO

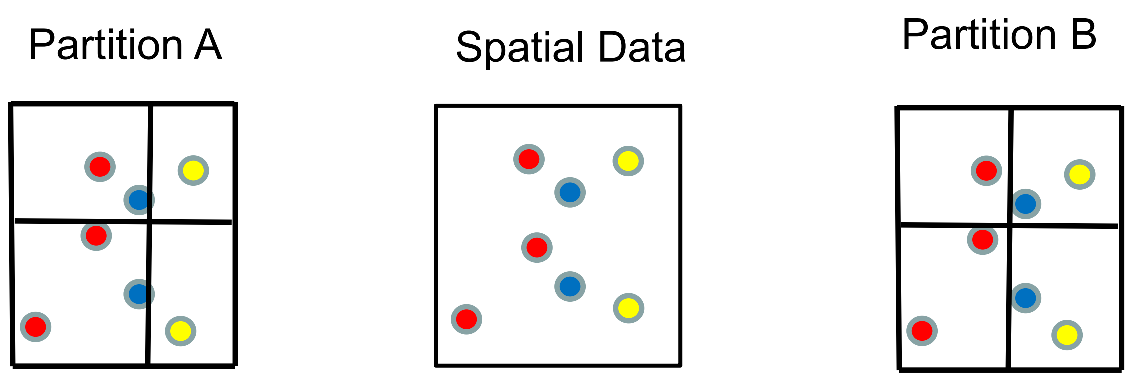
THE CLIMATE CORPORATION

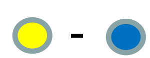
A few Questions in Transportation Domain

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?	Object Detection
Public Safety	Which transportation mode is a GPS trace in? Which transit routes are taken by criminals?	Prediction
Vehicle engineers	Which locations have high NOx emission? What is co-located there?	Hotspot, Co-location

Limitations of Traditional Data Science

- Traditional methods not robust in face of
  - Spatial continuity
    - Gerrymandering risk: Classical methods not robust
    - Result changes if spatial partitioning changes
  - Auto-correlation, Heterogeneity , Edge-effect, …
  - Noise

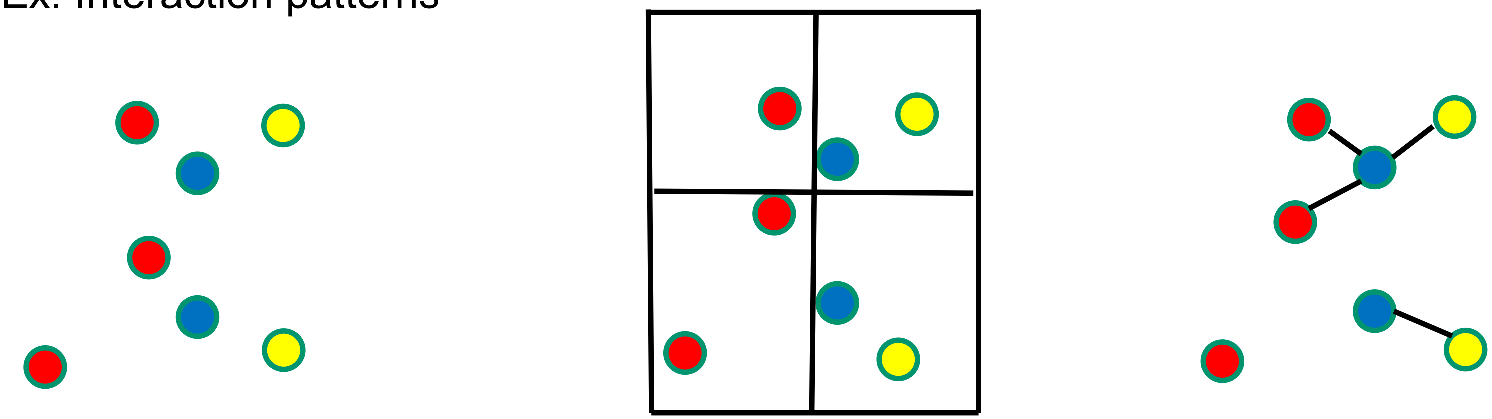


Partition A Based Pearson's Correlation	Pairs	Partition B Based Pearson's Correlation
1		- 0.90
- 0.90		1

Neighbor Graph Approach

- Challenge: One size does not fit all

- Ex. Interaction patterns



(a) a map of 3 features (b) Spatial Partitions (c) Neighbor graph

	Pearson's Correlation	Ripley's cross-K	Participation Index
	-0.90	0.33	0.5
	1	0.5	1

Details: Discovering Spatial Co-location Patterns: A General Approach,
IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

Spatial Autocorrelation: K-Function

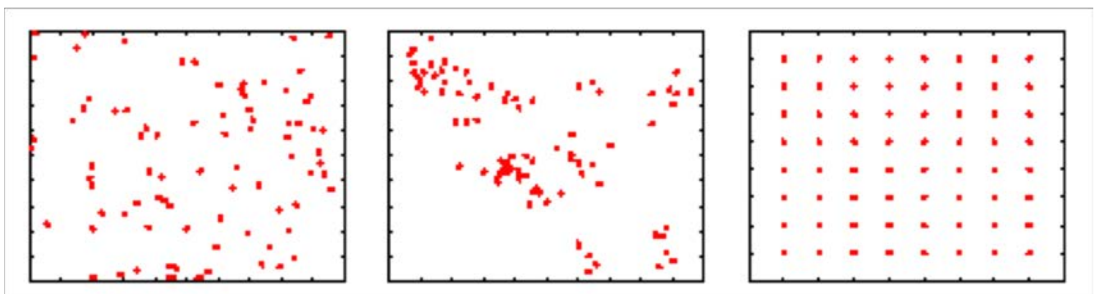
- **Purpose:** Compare a point dataset with a complete spatial random (CSR) data
- **Input:** A set of points

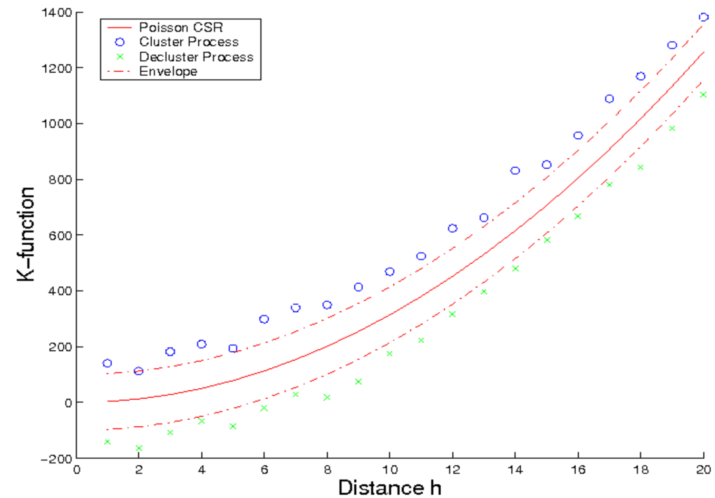
$K(h, data) = \lambda^{-1} E$ [number of events within distance *h* of an arbitrary event]

  - where λ is intensity of event
- **Interpretation:** Compare k(h, data) with *K(h,* CSR)
  - *K(h, data)* = *k(h, CSR):* Points are CSR

    > means Points are clustered

    < means Points are de-clustered

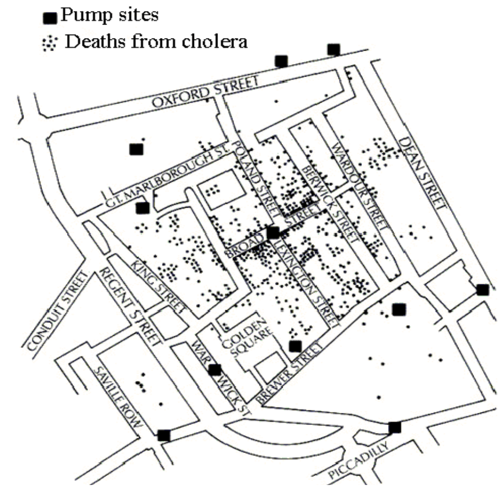




CSR Clustered De-clustered

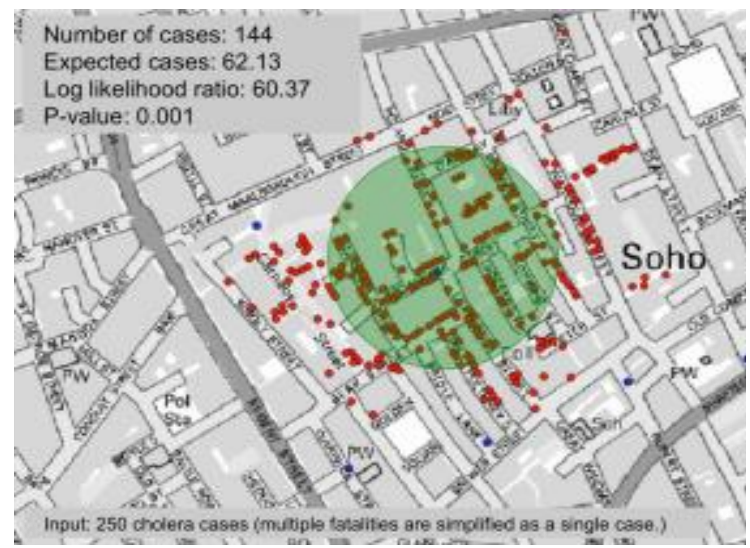
Spatial Computing Research Group

Defining Spatial Data Science



- The process of discovering
  - interesting, useful, non-trivial patterns
    - patterns: non-specialist
    - exception to patterns: specialist
  - from large spatial datasets
    - Spatial Big Data
    - Volunteered Geographic Information

- 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

Transdisciplinary Foundations of Geospatial Data Science. *ISPRS International Journal of Geo-Information*, *6*(12), p.395, 2017.

Identifying patterns in spatial information: A survey of methods. *Wiley Interdisci. Reviews: Data Mining and Knowl. Discovery*, *1*(3):193-214, 2011

A. Hotspots, Spatial clusters

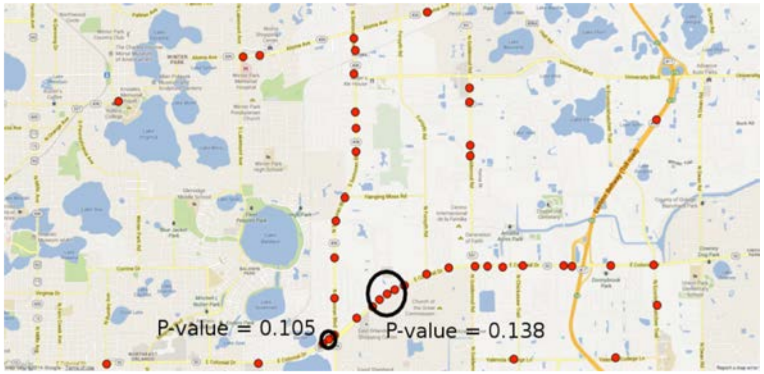
- **Question:** Which corridors are accident-prone?
- **Data:**
  - 43 Pedestrian fatalities in Orlando, FL (2000-9)
  - USDOT Fatality Analysis Reporting System

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

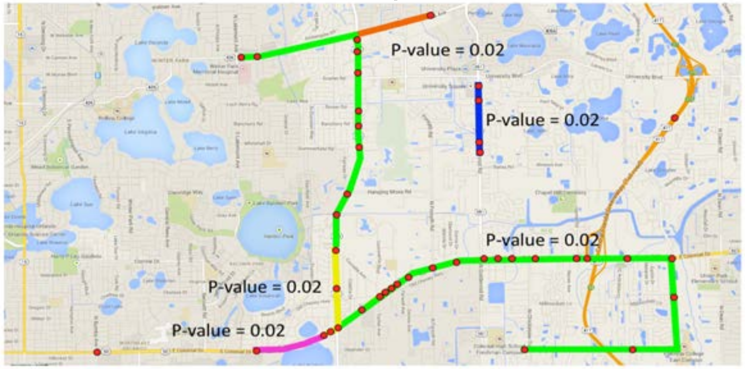
- **Patterns:**
  - Circular results from SaTScan
  - Linear hotspots
- **Interpretation:**

Unsafe pedestrian walkway

SaTScan Result



Linear hotspots



Details: Significant Linear Hotspot Discovery, IEEE Transactions on Big Data, 3(2):140-153, 2017.
(Summary in Proc. Geographic Info. Sc., Springer LNCS 8728, pp. 284-300, 2014.)

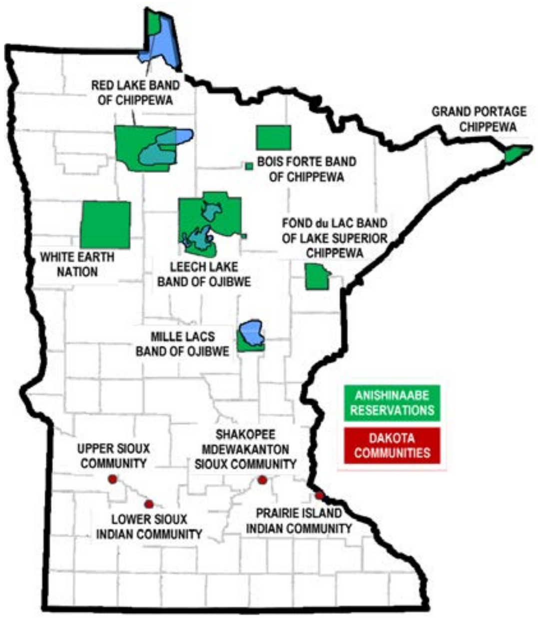
Minnesota Examples

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



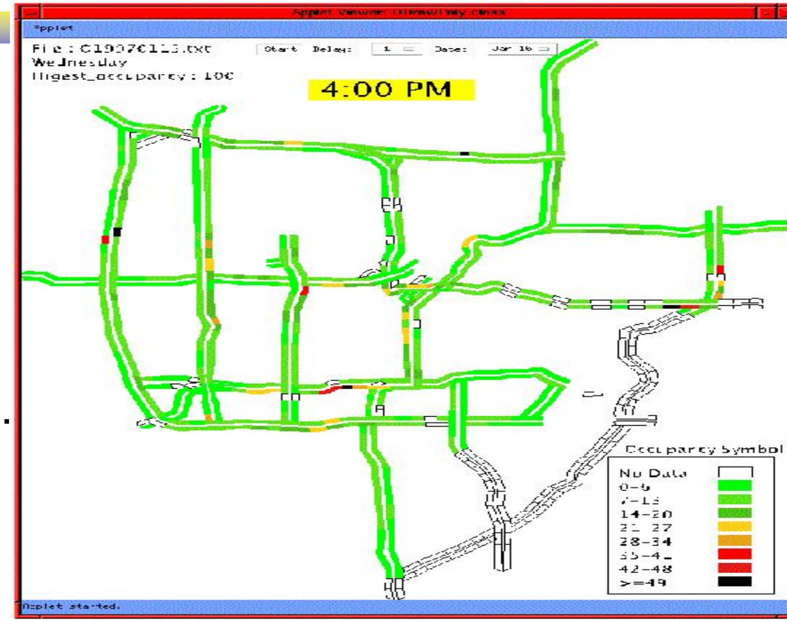
https://www.researchgate.net/figure/Location-of-reservations-in-Minnesota-Source-Indian-Affairs-Council-of-State-of_fig3_328759103

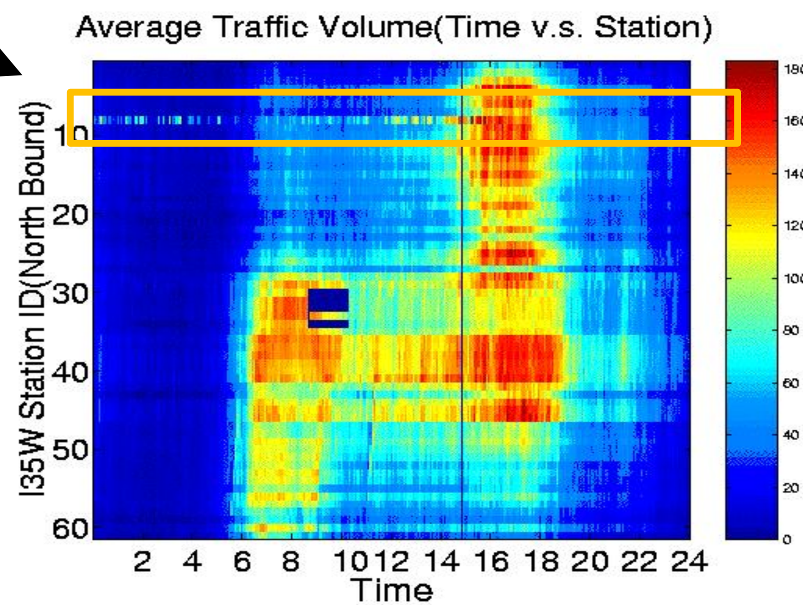
http://www.startribune.com/report-shows-that-pedestrian-safety-is-a-major-concern-on-minnesota-s-american-indian-reservations/505941632/

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

B. Spatial outlier, Discontinuities

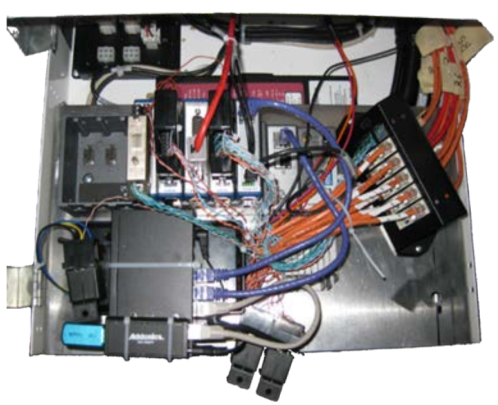
- **Question:** Which loop detector stations are very different from their neighbors?
- **Data:**
  - 900 stations (with 1 to 4 loop detectors each).
- **Pattern:**
  - Spatial outlier at Station 9.
- **Interpretation:**
  - Hypothesis: faulty loop detector?
  - Action: Test station 8 detectors





C. 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

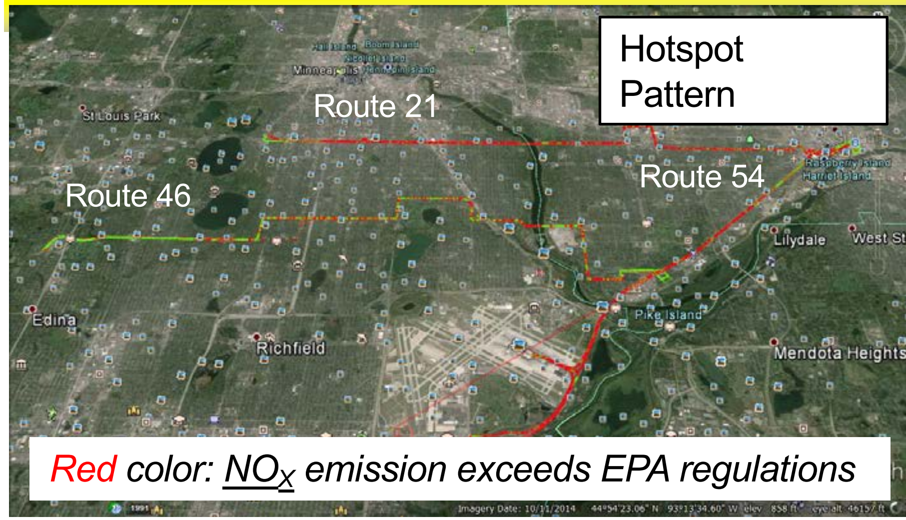


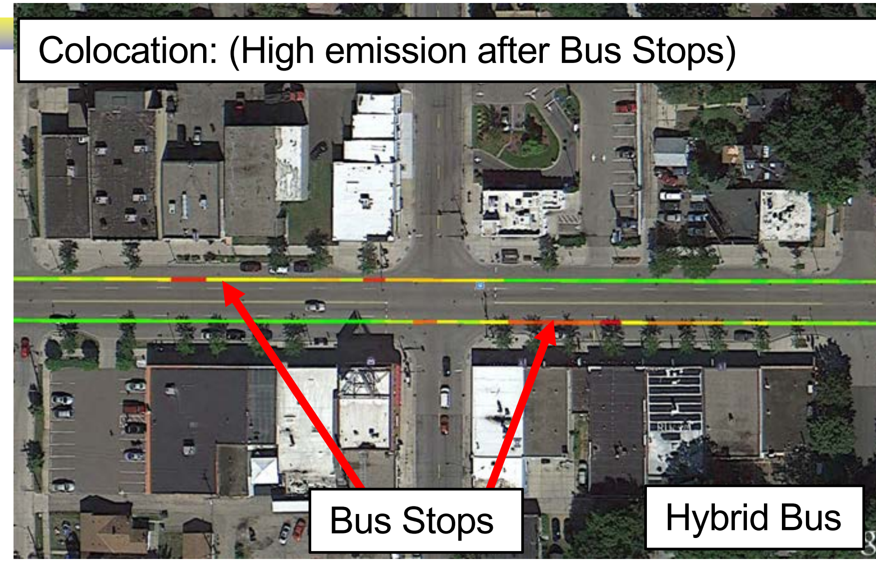
Variables sampled every second:

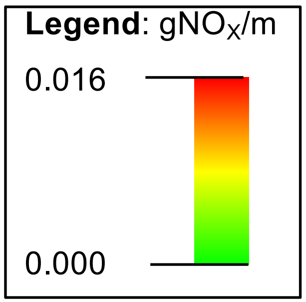
- GPS location
- Speed
- Vehicle Load
- Engine and Heater Fuel Flow
- Exhaust Temp and Mass Flow
- Intake Temp And Mass Flow
- Engine Torque and RPM
- Engine Coolant Temp
- Odometer
- **NOx emission**
-
-
- ….measurements on 200+ variables

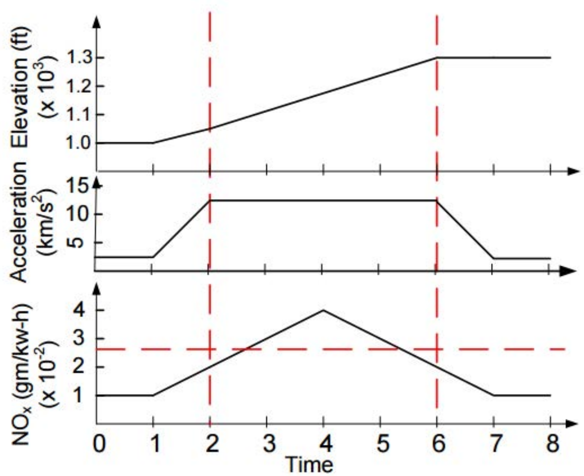
Details: “*Discovering Non-compliant Window Co-Occurrence Patterns.*” (R. Ali et al.) GeoInformatica, 21(4): 829-866, Springer, 2017

C. Emission Hotspots, Co-locations

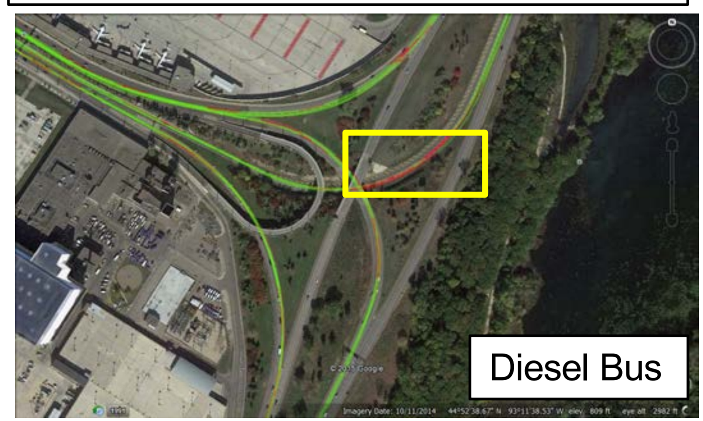








Colocation:(High emission, uphill ramp)

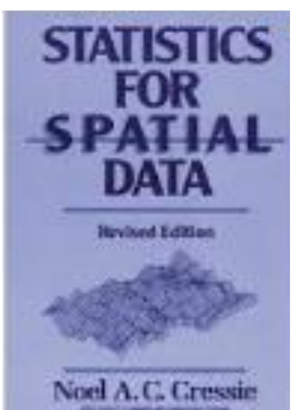


A Metric of Spatial Cross-Correlation

- Ripley's Cross K-Function Definition

$$K_{ij}(h) = \lambda_j^{-1} E\ [\text{number of type } \boldsymbol{j} \text{ event within distance } h \text{ of a randomly chosen type } \boldsymbol{i} \text{ event}]$$

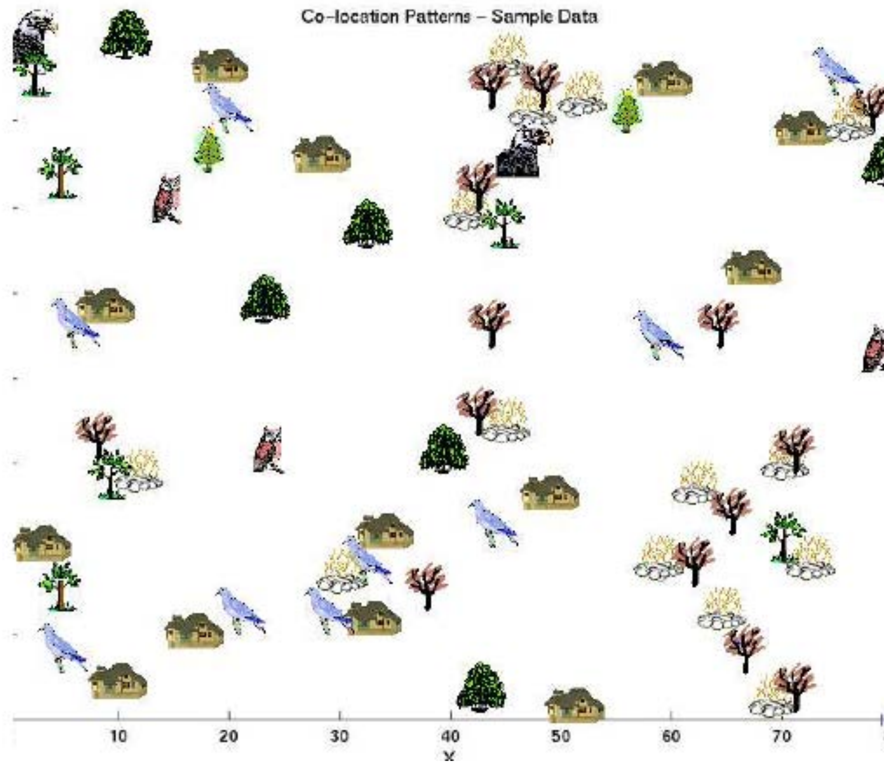
  - Cross K-function of some pair of spatial feature types
  - Example
    - Which pairs are frequently co-located
    - Statistical significance



Spatial Computing Research Group

Co-locations

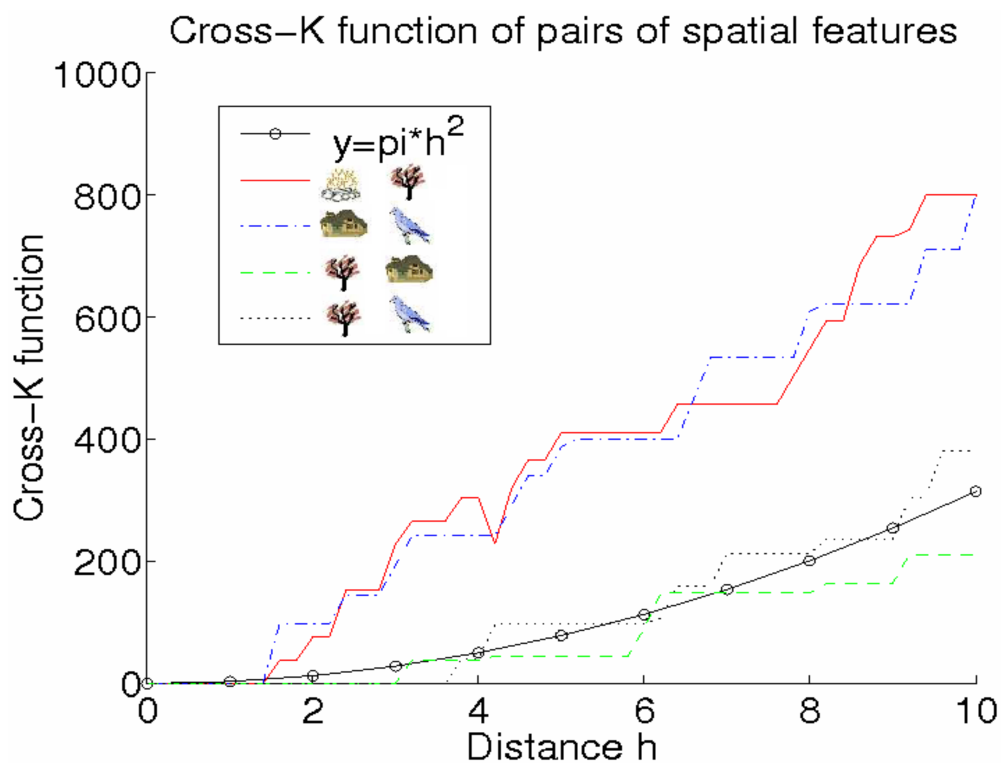
- Given: A collection of different types of spatial events
- Find: Co-located subsets of event types

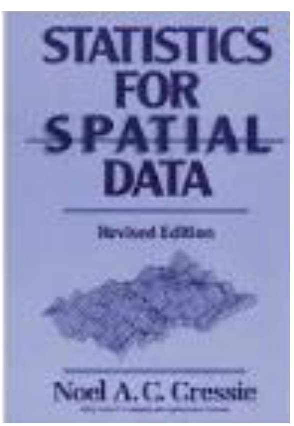


Source: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

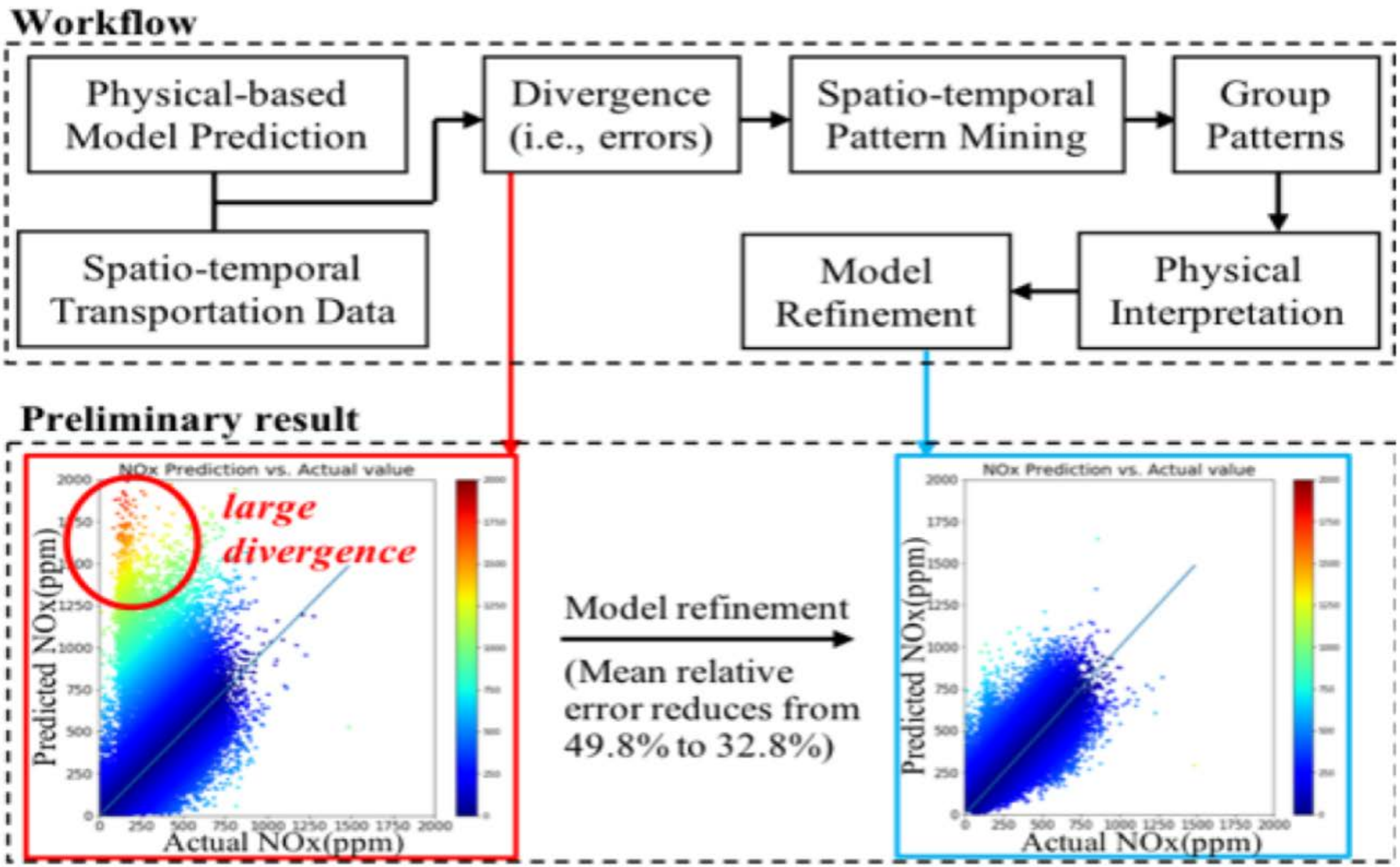
Illustration of Cross-Correlation

- Illustration of Cross K-function for Example Data





Co-occurrence Patterns to Refine (NOx) Model



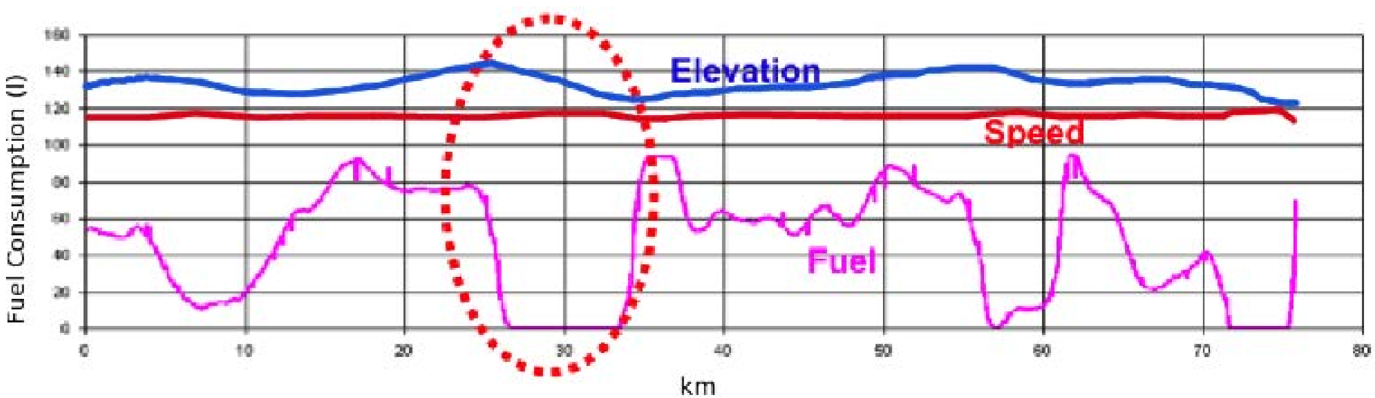
Discovering Co-occurrence Patterns of Model Errors

- **Question:** What OBD variable values co-occurs with high error in an NOx model ?
- **OBD =** On Board Diagnostics Data from Diesel Buses (MetroTransit)

Pattern Group	Example Patterns		
Low Vehicle Speed Condition	1. Wheelspeed: $w_0\ w_0\ w_0\ w_0\ w_0$ RailMPa: $r_1\ r_1\ r_1\ r_1\ r_1$ IntakeT: $I_6\ I_6\ I_6\ I_6\ I_6$	2. Wheelspeed: $w_0\ w_0\ w_0\ w_0\ w_0$ IntakeT: $I_6\ I_6\ I_6\ I_6\ I_6$ Fuelconskgph: $f_1\ f_1\ f_1\ f_1\ f_1$	3. Wheelspeed: $w_1\ w_0\ w_0\ w_0\ w_0$ Enginespeed: $s_1\ s_1\ s_2\ s_3\ s_3$ Enginepower: $p_5\ p_5\ p_5\ p_5\ p_5$
Low EGR Condition	4. Acceleration: $a_6\ a_6\ a_6\ a_6\ a_6$ EGRkgph: $g_0\ g_0\ g_0\ g_0\ g_0$	5. Bkpwr: $B_4\ B_4\ B_4\ B_4\ B_4$ EGRkgph: $g_0\ g_0\ g_0\ g_0\ g_0$	
Transient Condition	6. Wheelspeed: $w_7\ w_7\ w_7\ w_7\ w_7$ Bkpwr: $B_5\ B_4\ B_4\ B_4\ B_4$ Fuelconskgph: $f_1\ f_1\ f_0\ f_0\ f_0$	7. Acceleration: $a_6\ a_6\ a_6\ a_6\ a_5$ RailMPa: $r_4\ r_4\ r_4\ r_4\ r_4$	

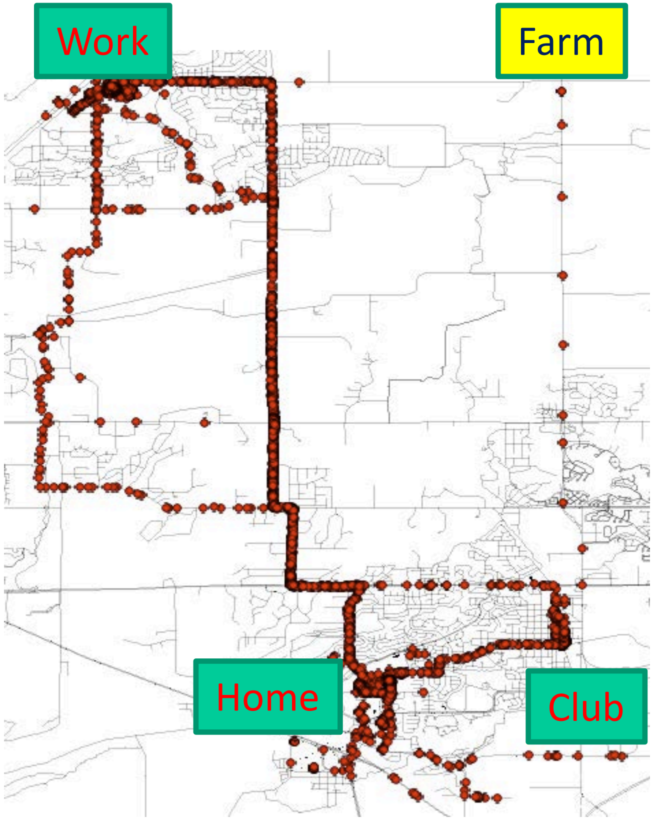
Legend

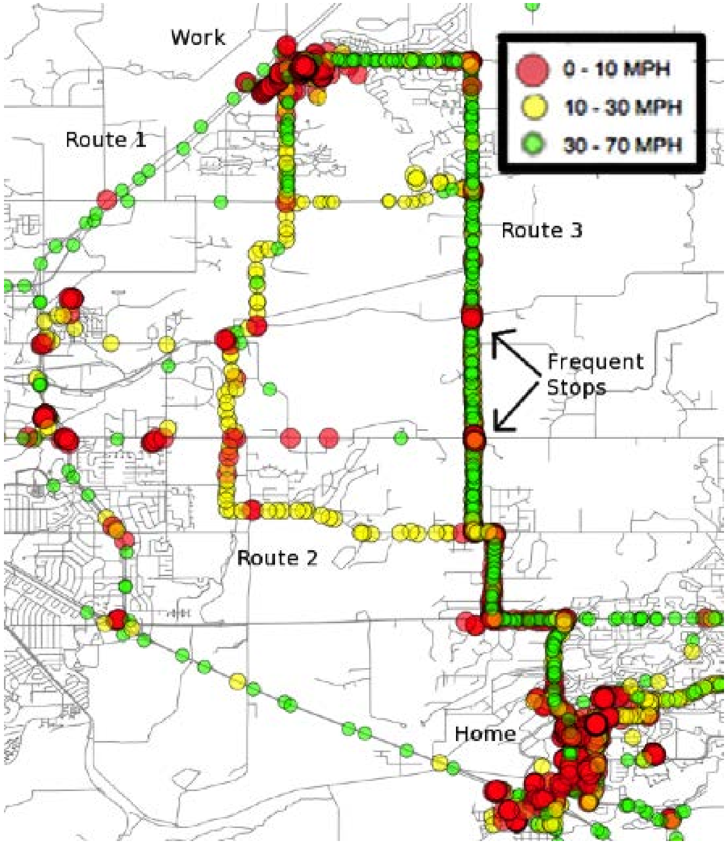
Subscript	Scale of the values
0, 1	Very low value
2, 3	Low value
4, 5	Medium value
6, 7	High value



D. Classification: GPS trace → Transportation Modes

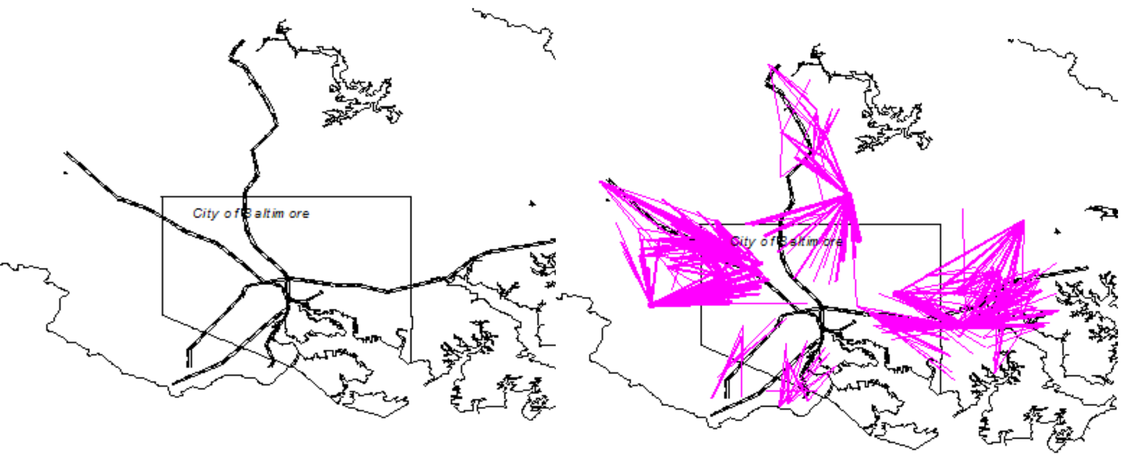
- Weekday GPS track for 3 months
  - Patterns of life, usual places (e.g., home, work, turf/tribe), commute routes
  - Predict Transport Modes, e.g., car, bicycle, walking, … (e.g., Travel Diary App)
  - Q? Guess transport modes for yellow and green commute routes?
  - Hint: see speed



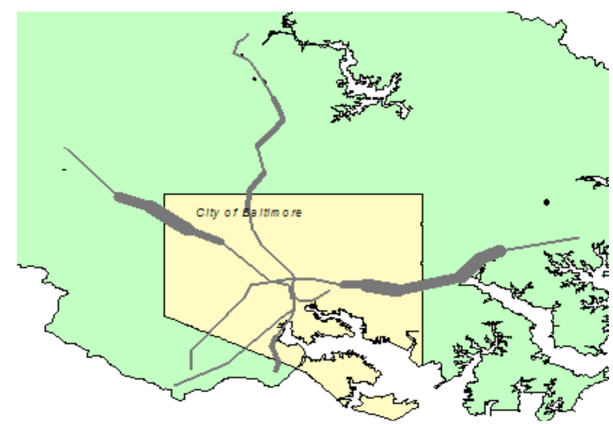


D. Prediction of Routes

- **Q:?** Which transit routes are used frequently by criminals ?



Input: Train network &
Lines *connecting crime location & criminal's residence*

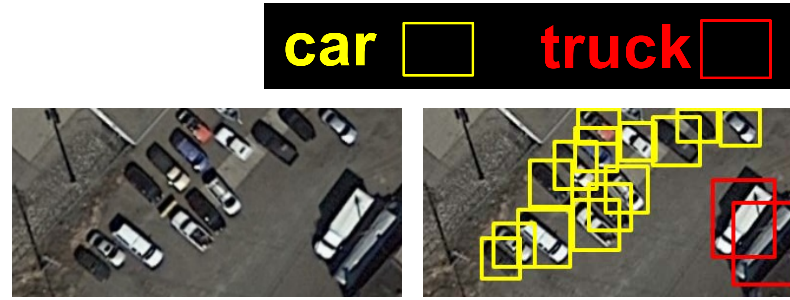


Output: Journey- to-Crime
(thick lines = common routes)

Journey-to-Crime **Prediction via the CrimeStat software**

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, …
- **Approach:**
  - Convolutional Neural Networks
  - You Only Look Once (YOLO) architecture



Input training image **Input training MOBRs**

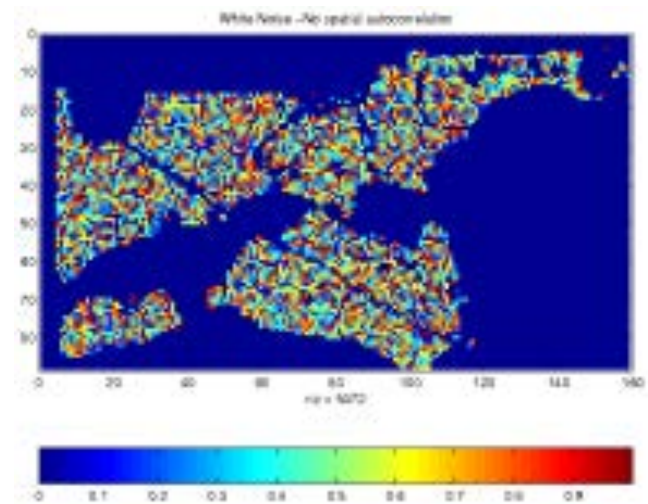
Test image **Output MBRs**

YOLO (baseline) **Proposed method**

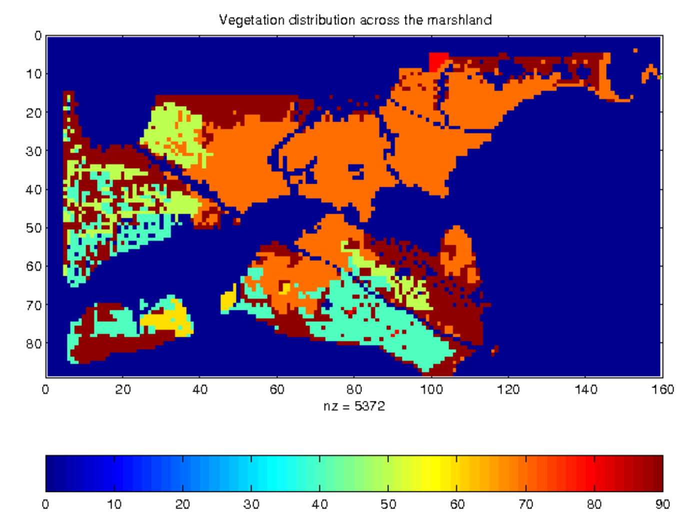
An unsupervised augmentation framework for deep learning based geospatial object detection: A summary of results, Proc. ACM SIGSPATIAL Intl. Conf. on Adv. in GIS (pp. 349-358). ACM, 2018 (w/ Y. Xie et al.)

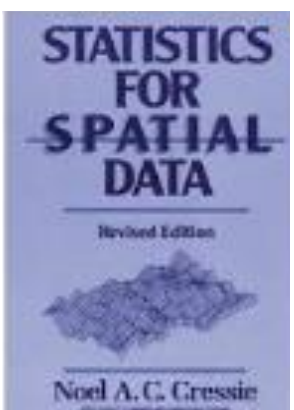
Spatial Auto-correlation

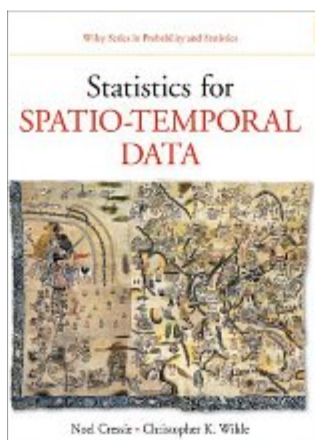
- Spatial Statistics, Spatial Data Mining
  - Honor spatial continuity
  - Auto-correlation
  - Heterogeneity
  - Edge-effect, …



- Limitation of i.i.d assumption
  - Ignores auto-correlation
  - Salt n Pepper noise (next slide)







Spatial Auto-correlation in Prediction Models

- Traditional Models, e.g., Regression (with Logit or Probit),
  - Linear Regression, Bayes Classifier, …
- Semi-Spatial : auto-correlation regularizer $\varepsilon = \|y - \beta X\|^2 + \|\beta X - \beta X_{neighbor}\|^2$
- Spatial Models
  - Spatial autoregressive model (SAR)
  - Markov random field (MRF) based Bayesian Classifier

Traditional	**Spatial**
$y = X\beta + \varepsilon$	$y = \rho W y + X\beta + \varepsilon$
$\Pr(C_i \mid X) = \frac{\Pr(X \mid C_i)\Pr(C_i)}{\Pr(X)}$	$\Pr(c_i \mid X, C_N) = \frac{\Pr(C_i)\Pr(X, C_N \mid c_i)}{\Pr(X, C_N)}$
Neural Networks	Convolutional Neural Networks
Decision Trees	Spatial Decision Trees

Open Problems in Spatial Data Science

- Spatial Statistics mature for low-dimensional Isotropic spaces
- Not mature for Anisotropic spaces (e.g., road networks), Spatio-temporal phenomena
- Open Questions
  - How to quantify and efficiently mine interesting patterns on road networks?
    - Spatio-temporal hotspots, Linear Hotspots on non-shortest paths
    - Co-occurrences of spatial network events, Prediction of their properties
    - Change Detection in spatial network patterns, e.g., displacement
    - Multi-scale space/time
  - Other Questions
    - How to increase Ground Truth data? e.g., citizen science
    - Fairness (e.g., pothole reports by smartphone apps)
    - Accountability (e.g., cost of spurious hotspots)
    - Transparency, e.g., interpretation using transportation concepts & theories
    - Ethics (e.g., geo-privacy, data ownership, gerrymandering)

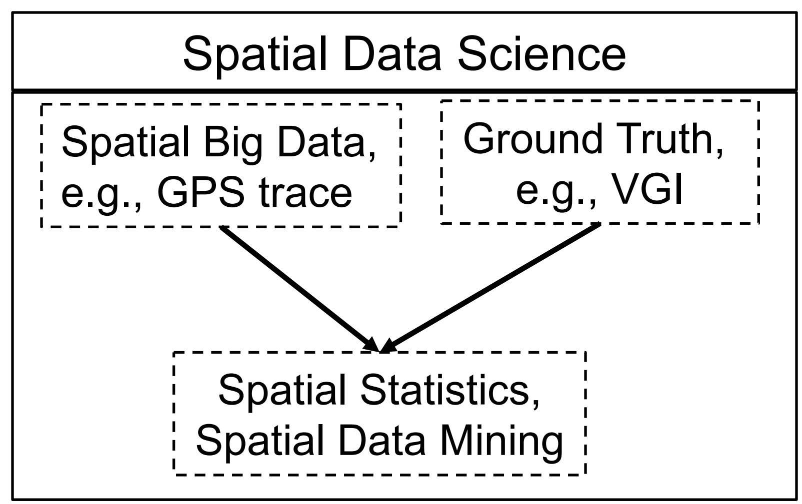
Transdisciplinary Foundations of Geospatial Data Science. *ISPRS International Journal of Geo-Information*, *6*(12), p.395, 2017.

Identifying patterns in spatial information: A survey of methods. *Wiley Interdisci. Reviews: Data Mining and Knowl. Discovery*, *1*(3):193-214, 2011

Summary : One size data science does not fit all

- Spatial Data are ubiquitous & important

- Traditional Data Science Tools are inadequate
  - Gerrymandering, Spatial Auto-correlation, …
- Spatial Data Science
  - Spatial Big Data
  - Ground Truth (e.g., official or VGI)
  - Spatial Statistics/Data Mining
    - Mature in isotropic space
    - Not for road maps, spatio-temporal phenomena



References :Surveys, Overviews

- Spatial Computing (html , short video , tweet), Communications of the ACM, 59(1):72-81, January, 2016.
- Transdisciplinary Foundations of Geospatial Data Science (html , pdf), ISPRS Intl. Jr. of Geo-Informatics, 6(12):395-429, 2017. (doi:10.3390/ijgi6120395)
- Spatiotemporal Data Mining: A Computational Perspective , ISPRS Intl. Jr. on Geo-Information, 4(4):2306-2338, 2015 (DOI: 10.3390/ijgi4042306).
- Identifying patterns in spatial information: a survey of methods (pdf), Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3):193-214, May/June 2011. (DOI: 10.1002/widm.25).
- Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, IEEE Transactions on Knowledge and Dat Mining, 29(10):2318-2331, June 2017. (DOI: 10.1109/TKDE.2017.2720168).
- Parallel Processing over Spatial-Temporal Datasets from Geo, Bio, Climate and Social Science Communities: A Research Roadmap. IEEE BigData Congress 2017: 232-250.
- Spatial Databases: Accomplishments and Research Needs, IEEE Transactions on Knowledge and Data Engineering, 11(1):45-55, 1999.