What is special about mining spatial data?

Sept. 13th, 2022

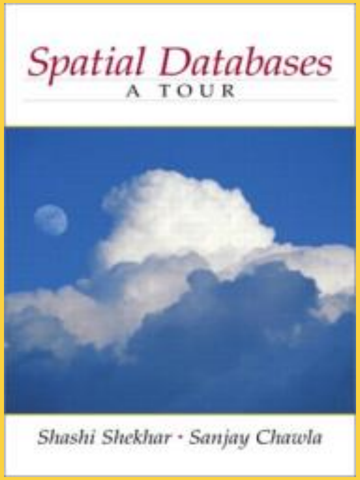
Monthly Seminar of the HDR Institute: HARP- Harnessing Data and Model Revolution in the Polar Regions

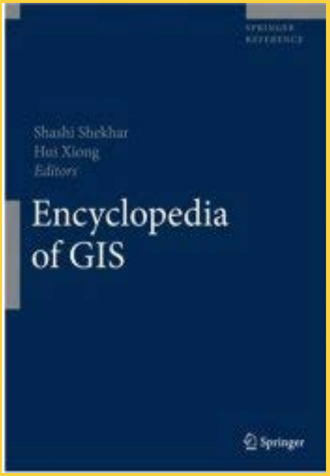
Shashi Shekhar

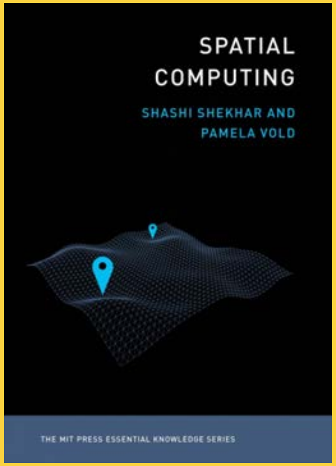
McKnight Distinguished University Professor

Dept. of Computer Sc. and Eng., University of Minnesota

www.cs.umn.edu/~shekhar : shekhar@umn.edu







University of Minnesota

Driven to Discover℠

Spatial Revolution

ESRI ArcGIS X

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

OGC™
Open Geospatial Consortium, Inc.

ORACLE
SPATIAL

IBM Smarter Planet

Google Maps

OpenStreetMap

Ushahidi

waze

DiDi
GRABTAXI
lyft
OLA

Uber Eats
GRUBHUB
instacart
DOORDASH

Google Earth Engine

MONSANTO
THE CLIMATE CORPORATION

Spatial is a Critical Infrastructure Today!

- 2 billion GPS receivers in use, will hit 7 billion by 2022.
- Besides location, it reference time for critical infrastructure
  - Telecommunications industry, Banks, Airlines...
- 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

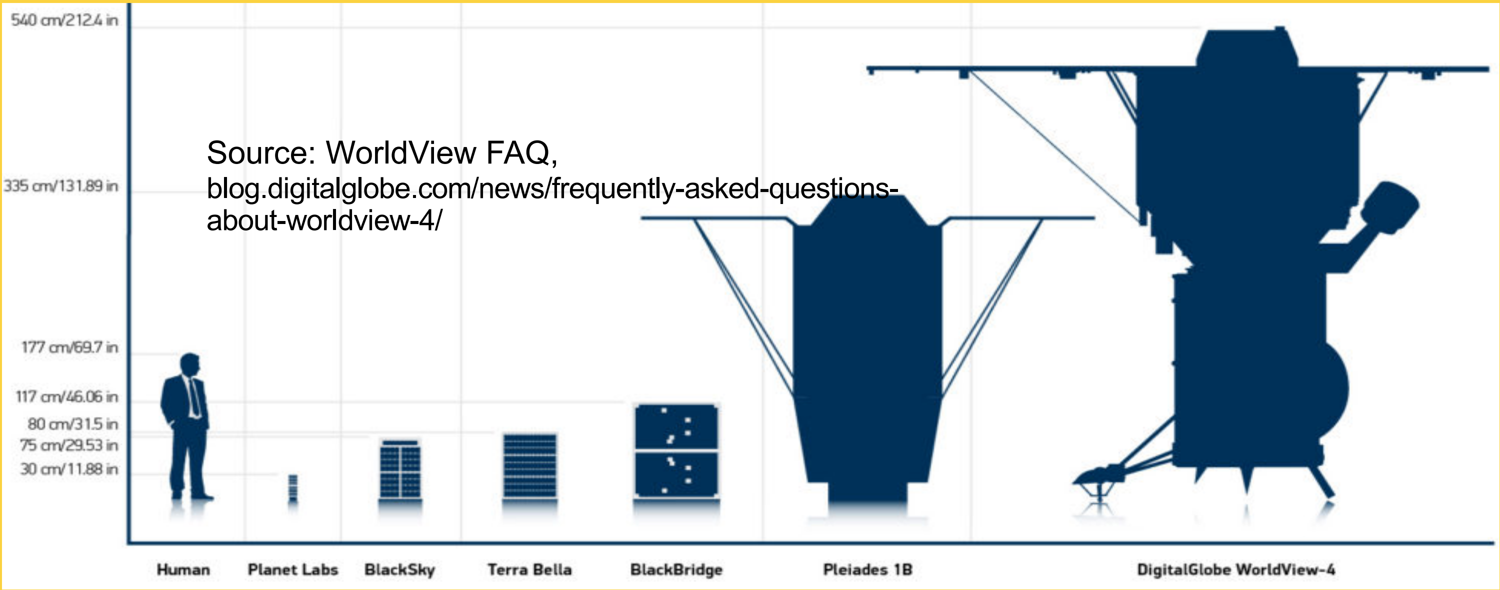
July 25, 2018, 4:00 AM CDT

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*

Growth of Spatial Data

- 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: 200+ satellites: daily scan of Earth at 1m resolution



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

Easier Access to Spatial Data

- 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, …

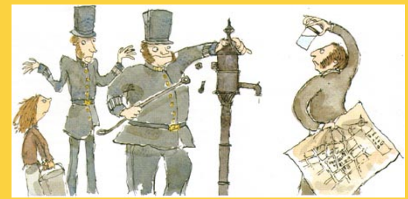
	Google Earth Engines	NEX	AWS Earth
Elevation, Landsat, LOCA, MODIS, NAIP	x	x	x
NOAA	x		x
AVHRR, FIA, GIMMM, GlobCover, NARR, TRIMM, Sentinel-1	x	x	
IARPA, GDELT, MOGREPS, OpenStreetMap, Sentinel-2, SpaceNet (building/road labels for ML)			x
CHIRPS, GeoScience Australia, GSMap, NASS, Oxford Map, PSDI,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

Outline

- Motivation
  - Use cases
  - Pattern families
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
- Conclusions

A Spatial Data Mining Story

<u>1854: How does Cholera spread?</u>

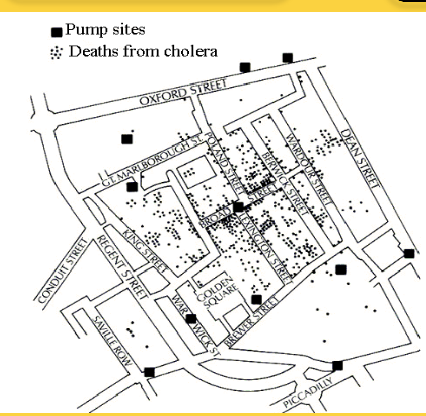


Collect & Curate Data ⇨ Discover Patterns, Generate Hypothesis ⇨ Test Hypothesis (Experiments) ⇨ Develop Theory

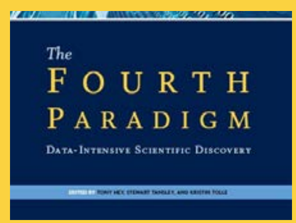
? water pump

Remove pump handle

Germ Theory



Impact:
Hygiene
Drinking water supply,
Sewage system,
…



Why Data Mining?

- Holy Grail - Informed Decision Making
- Sensors & Databases increased rate of Data Collection
  - Transactions, Web logs, GPS-track, Remote sensing, …
- Challenges:
  - Volume (data) >> number of human analysts
  - Some automation needed
- Approaches
  - Database Querying, e.g., SQL3/OGIS
  - Data Mining for Patterns
  - …

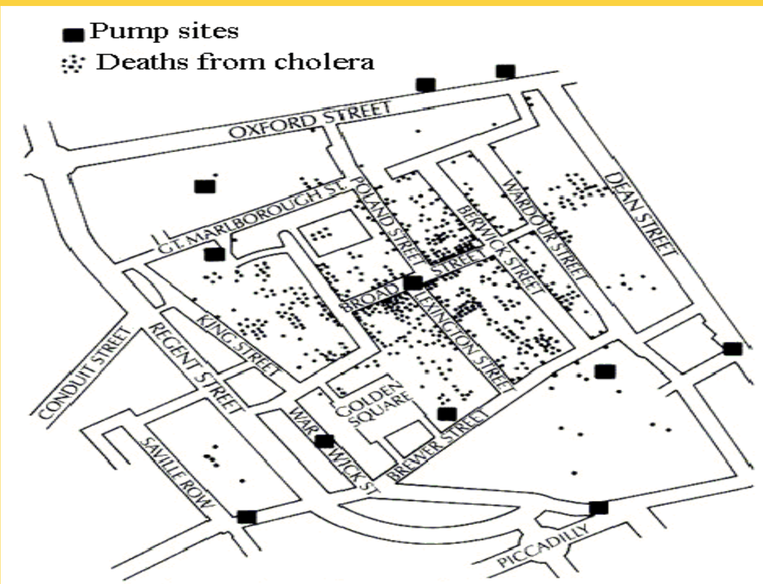
Spatial Data Mining (SDM)

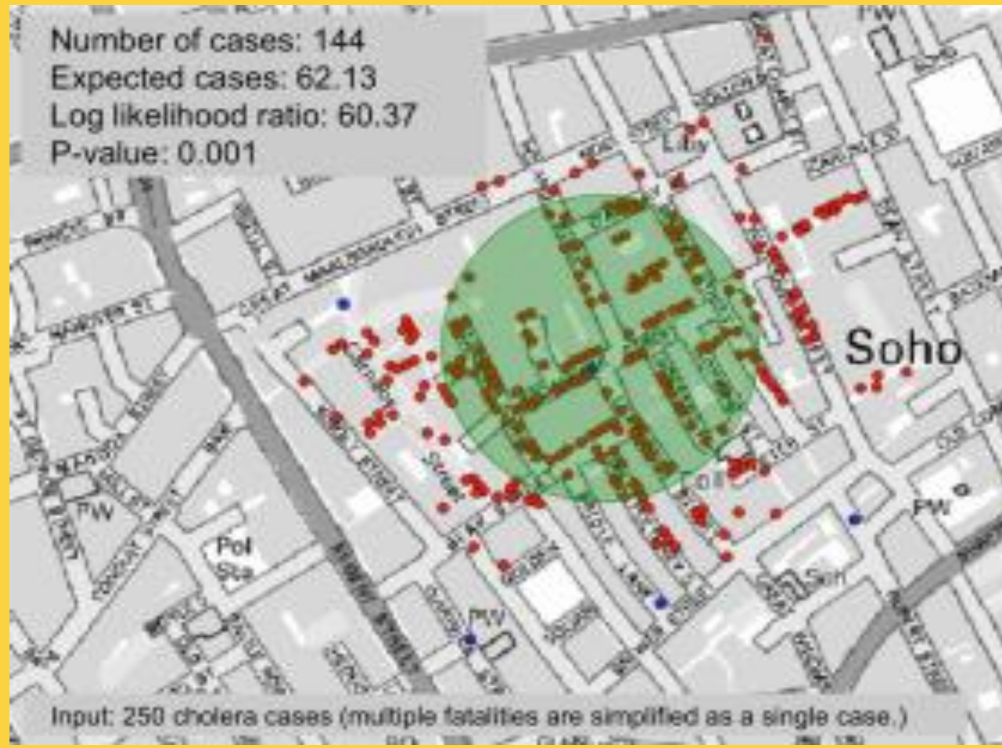
- The process of discovering
  - interesting, useful, non-trivial patterns
    - patterns: non-specialist
    - exception to patterns: specialist
  - from large spatial datasets

- Spatial pattern families
  - Hotspots, Spatial clusters
  - Spatial outlier, discontinuities
  - Co-locations, co-occurrences
  - Location prediction models
  - …

Pattern Family 1: Hotspots, Spatial Cluster

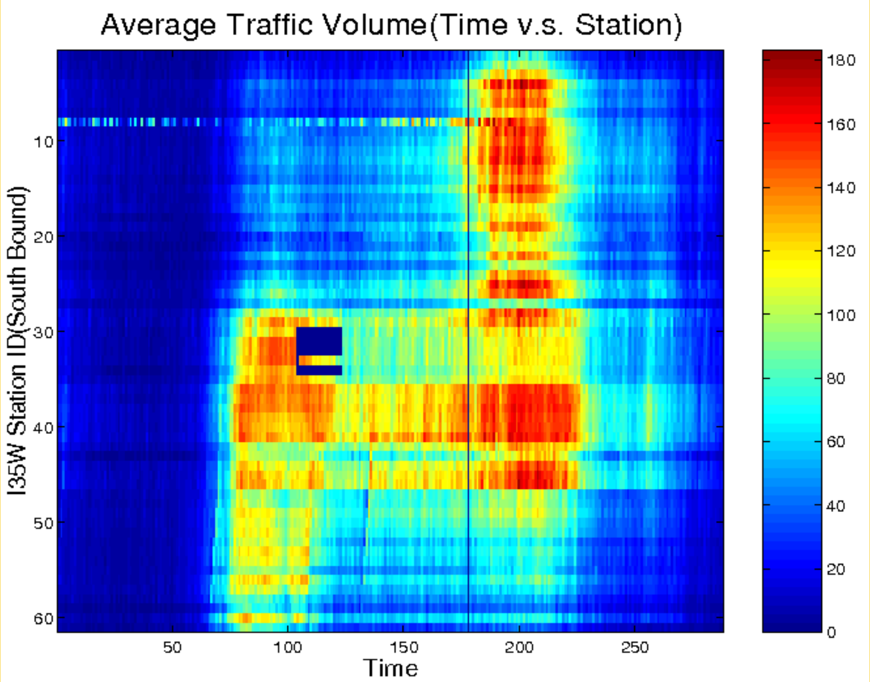
- The 1854 Asiatic Cholera in London
  - Near Broad St. water pump except a brewery





Pattern Family 2: Spatial Outliers

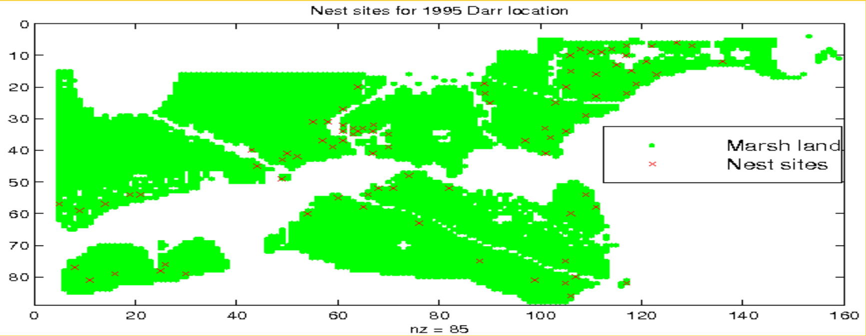
- Spatial Outliers, Anomalies, Discontinuities
  - Traffic Data in Twin Cities
  - Abnormal Sensor Detections
  - Spatial and Temporal Outliers

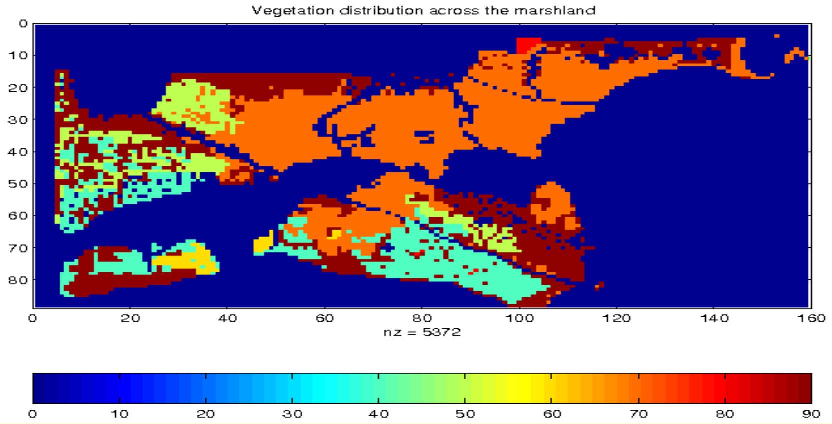


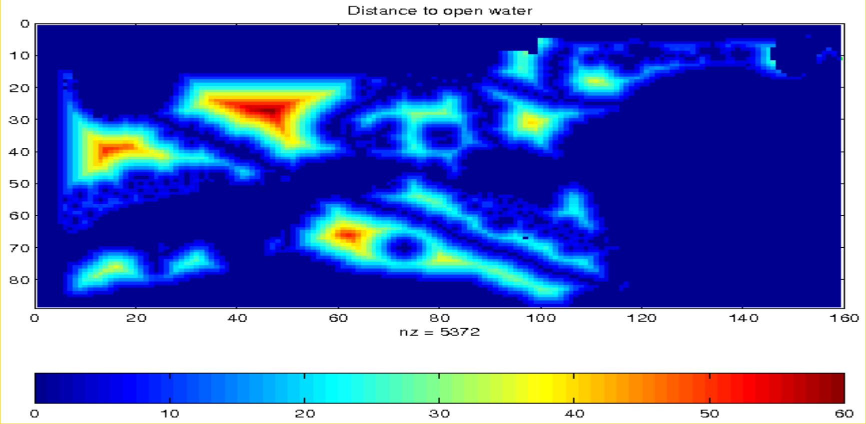
Source: A Unified Approach to Detecting Spatial Outliers, GeoInformatica, 7(2), Springer, June 2003. (A Summary in Proc. ACM SIGKDD 2001) with C.-T. Lu, P. Zhang.

Pattern Family 3: Spatial Prediction

- Location Prediction:
  - Predict Bird Habitat Prediction
  - Using environmental variables



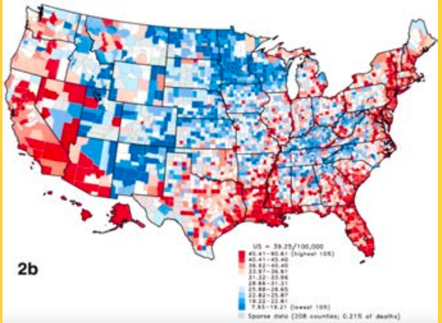




Details: Spatial Contextual Classification and Prediction Models for Mining Geospatial Data, S. Shekhar et al., IEEE Transactions on Multimedia, 4(2):174 - 188. 10.1109/TMM.2002.1017732.

Pattern Family 4: Colocation Example

- Cholera death, Broad Street water pump (1854, London)
- Higher Lung-cancer mortality (white males, 1950-69), WW2 ship building (Asbestos)



- Food deserts, increased rate of obesity & cancer
- …

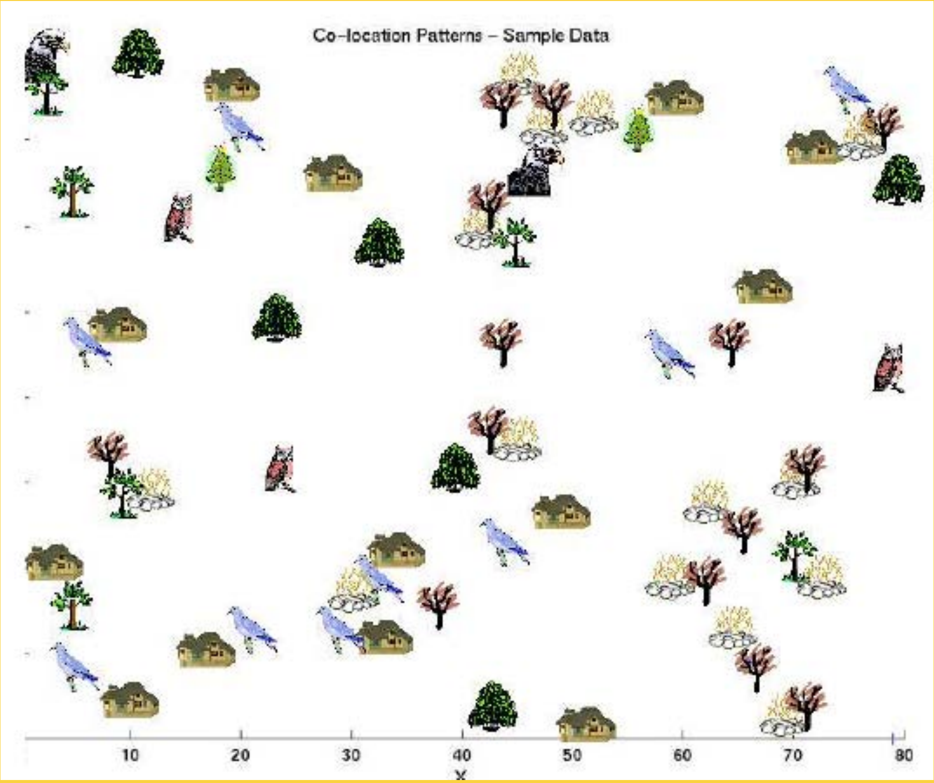
Sources: A. Jemal et al., “Recent Geographic Patterns of Lung Cancer and Mesothelioma Mortality Rates in 49 Shipyard Counties in the U.S., 1970-94”, Am J. Ind. Med. 2000, 37(5):512-21.

E. Paskett, Place as a rick factor: how Geography shapes where cancer strikes, Elektra Paskett, www.nyp.org/cancer/cancerprevention/cancer-prevention-articles/029-how-geography-shapes-where-cancer-strikes;

B. Tedeschi, Breaking the cycle of despair: One woman's battle for the health of Appalachia, June 20, 2016. https://www.statnews.com/2016/06/20/breaking-cycle-despair-one-womans-battle-health-appalachia/

Family 4: Co-locations/Co-occurrence

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



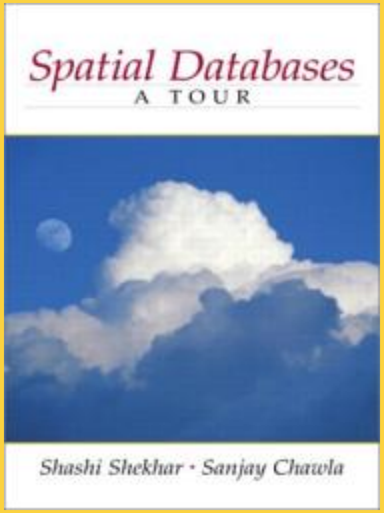
Source: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

Scope: What's NOT Spatial Data Mining?

- Simple Querying of Spatial Data
  - Find neighbors of Canada, or shortest path from Boston to Houston
- Testing a hypothesis via a primary data analysis
  - Ex. Is cancer rate inside Hinkley, CA higher than outside ?
  - SDM: Which places have significantly higher cancer rates?
- Uninteresting, obvious or well-known patterns
  - Ex. (Warmer winter in St. Paul, MN) => (warmer winter in Minneapolis, MN)
  - SDM: (Pacific warming, e.g. El Nino) => (warmer winter in Minneapolis, MN)
- Non-spatial data or pattern
  - Ex. Diaper and beer sales are correlated
  - SDM: Diaper and beer sales are correlated in blue-collar areas (weekday evening)

Outline

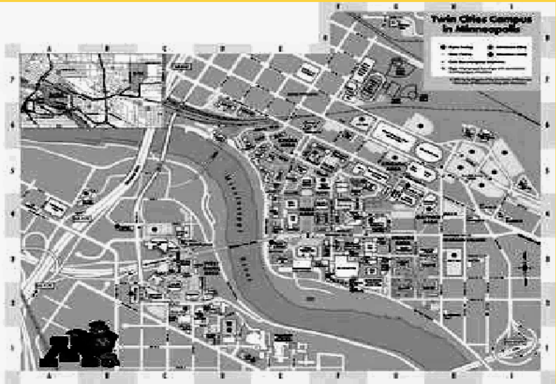
- Motivation
- Spatial Data
  - Spatial Data Types & Relationships
  - OGIS Simple Feature Types
- Spatial Statistical Foundations
- Spatial Data Mining
- Conclusions



Data-Types: Non-Spatial vs. Spatial

- Non-spatial
  - Numbers, text-string, …
  - e.g., city name, population
- Spatial (Geographically referenced)
  - Location, e.g., longitude, latitude, elevation
  - Neighborhood and extent
- Spatial Data-types
  - Raster: gridded space
  - Vector: point, line, polygon, …
  - Graph: node, edge, path

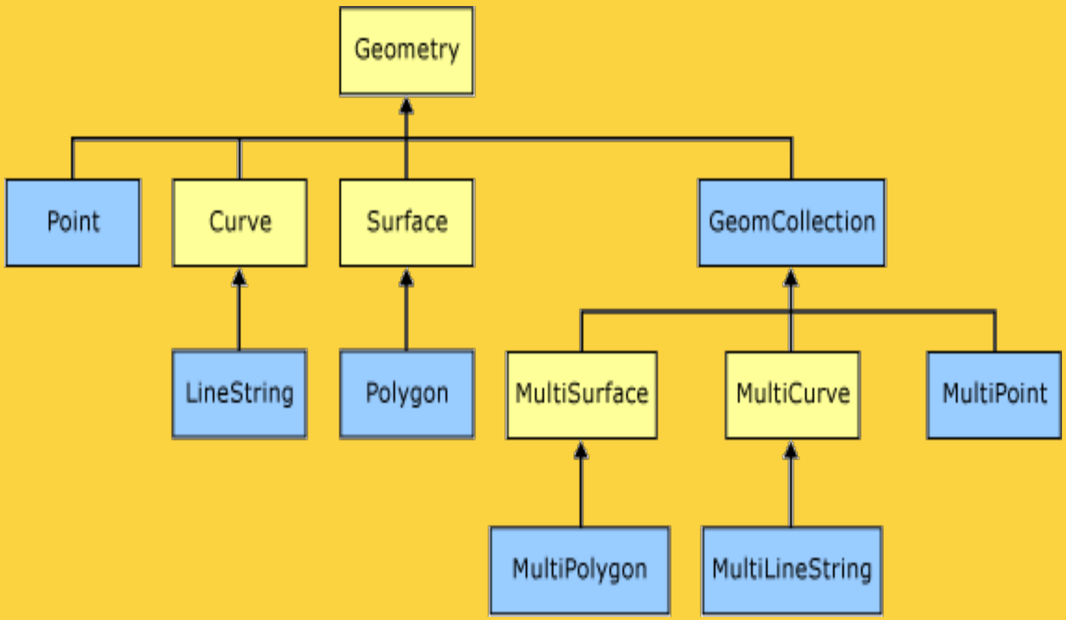
Raster (Courtesy: UMN)



Vector (Courtesy: MapQuest)

OGC Simple Features Standard

OGC Open Geospatial Consortium, Inc.



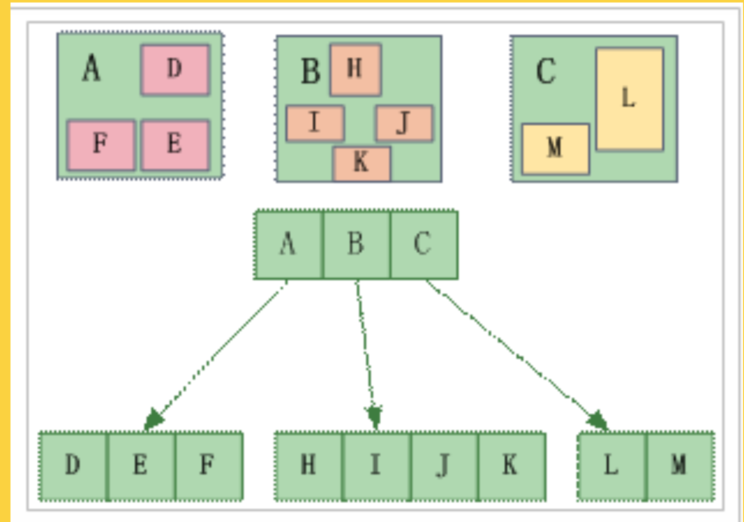
Spatial Analysis	Distance
	Buffer
	ConvexHull
	Intersection
	Union
	Difference
	DymmDiff

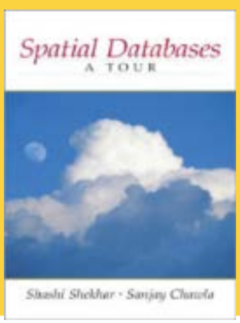
Basic Functions	SpatialReference ()
	Envelop ()
	Export ()
	IsEmpty ()
	IsSimple ()
	Boundary ()
Topological / Set Operators	Equal
	Disjoint
	Intersect
	Touch
	Cross
	Within
	Contains
	Overlap

[**Details:** Spatial Databases: Accomplishments and Research Needs, S. Shekhar et al., IEEE Trans. on Knowledge and Data Eng., 11(1), Jan.-Feb. 1999.

Spatial Database Management Systems

- Meta-data, Schema, DBMS (SQL, Hadoop)
- Challenge: One size does not fit all!
- Ex. Spatial Querying
  - Geo-tag. Checkin, Geo-fence
- Spatial Querying Software
  - OGC Spatial Data Type & Operations
  - Data-structures: B-tree => R-tree
  - Algorithms: Sorting => Geometric
  - Partitioning: random => proximity aware





ORACLE SPATIAL

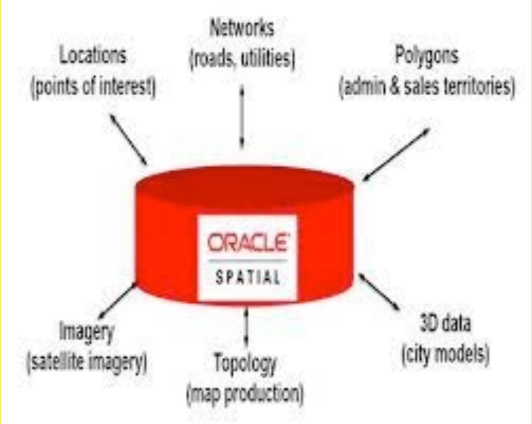
PostgreSQL

Spatial Hadoop
I can play with your spatial data

Research Needs for Data

- Limitations of OGC Simple Features
  - Direction predicates - e.g. absolute, ego-centric
  - Terrains and visibility, Network analysis, Raster operations
  - Spatio-temporal

- Needs for New Standards & Research
  - Modeling richer spatial properties listed above
  - Spatio-temporal data, e.g., moving objects



Outline

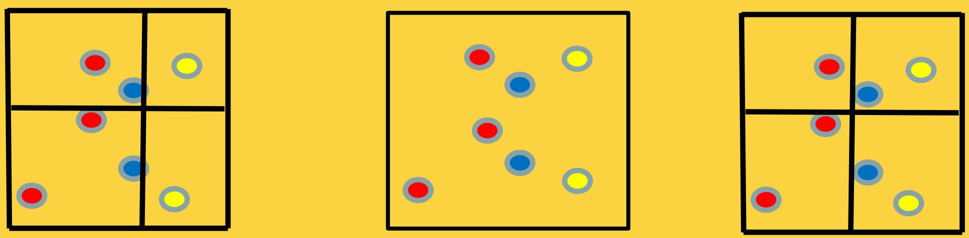
- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
  - Spatial Auto-correlation
  - Heterogeneity
  - Edge Effect
- Spatial Data Mining
- Conclusions

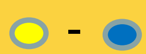
Limitations of Traditional Statistics

- Classical Statistics
    - Data samples: independent and identically distributed (i.i.d)
    - Simplifies mathematics underlying statistical methods, e.g., Linear Regression
- Spatial data samples are not independent
    - Spatial Autocorrelation metrics
        - distance-based (e.g., K-function), neighbor-based (e.g., Moran's I)
    - Spatial Cross-Correlation metrics
- Spatial Variability and Heterogeneity
    - Spatial data samples may not be identically distributed!
    - No two places on Earth are exactly alike!
- …

Challenge: Modifiable Areal Unit Problem (MAUP)

- Result changes if spatial partitioning changes (similar to Gerrymandering)
  - Neighbor Graph Based Measures are more robust

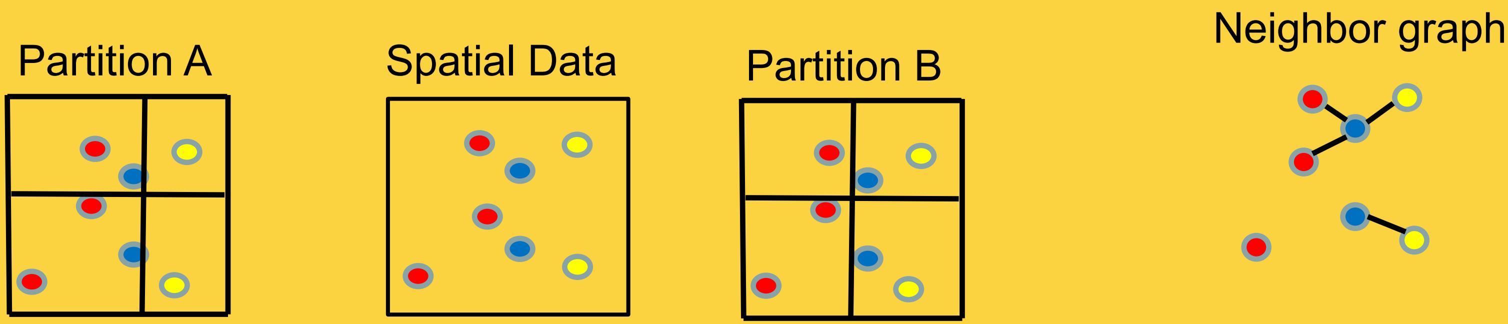


Partition A Based Pearson's Correlation	Pairs	Partition B Based Pearson's Correlation
1		- 0.90
- 0.90		1

Details: [Data Science for Earth: The Earth Day Report](#) , E. Eftelioglu, S. Shekhar, J. Hudson, L. Joppa, C. Baru, V. Janeja, et al. ACM SIGKDD Explorations Newsletter, 22(1), May 2020.

Challenge: Modifiable Areal Unit Problem (MAUP)

- Result changes if spatial partitioning changes (similar to Gerrymandering)
  - Neighbor Graph Based Measures are more robust

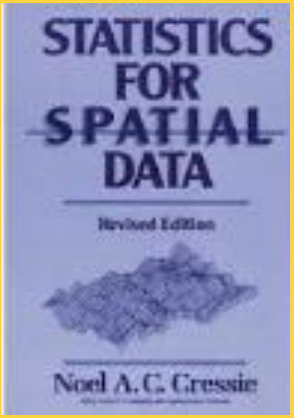


Partition A Based Pearson's Correlation	Pairs	Partition B Based Pearson's Correlation	Ripley's Cross-K	Participation Index
1		- 0.90	0.33	0.66
- 0.90		1	0.5	1

Details: [Data Science for Earth: The Earth Day Report](), E. Eftelioglu, S. Shekhar, J. Hudson, L. Joppa, C. Baru, V. Janeja, et al. ACM SIGKDD Explorations Newsletter, 22(1), May 2020.

Spatial Statistics: An Overview

- Point process
  - Discrete points, e.g., locations of trees, accidents, crimes, …
  - Complete spatial randomness (CSR): Poisson process in space
  - K-function: test of CSR

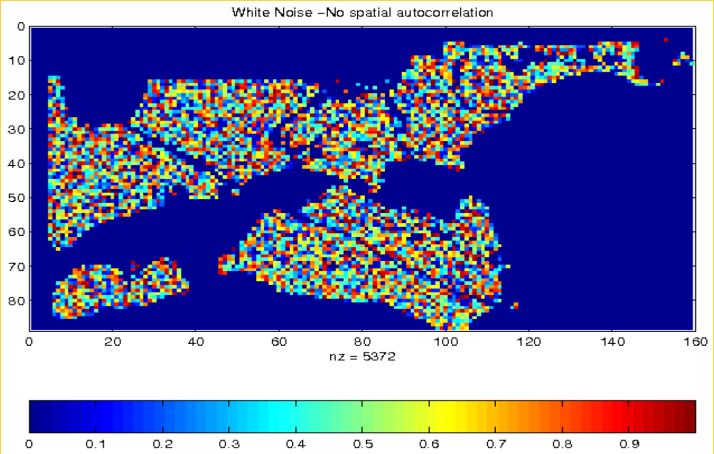


- Geostatistics
  - Continuous phenomena, e.g., rainfall, snow depth, …
  - Methods: Variogram measure how similarity decreases with distance
  - Spatial interpolation, e.g., Kriging

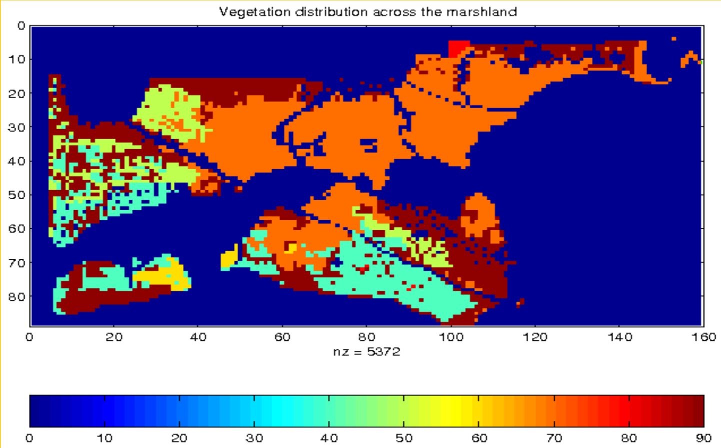
- Lattice-based statistics
  - Polygonal aggregate data, e.g., census, disease rates, pixels in a raster
  - Spatial Gaussian models, Markov Random Fields, Spatial Autoregressive Model

Spatial Autocorrelation (SA)

- First Law of Geography
  - All things are related, but nearby things are more related than distant things. [Tobler70]
- Spatial autocorrelation
  - Traditional i.i.d. assumption is not valid
  - Measures: K-function, Moran's I, Variogram, …



Independent, Identically Distributed pixel property



Vegetation Durability with SA

Spatial Autocorrelation: K-Function

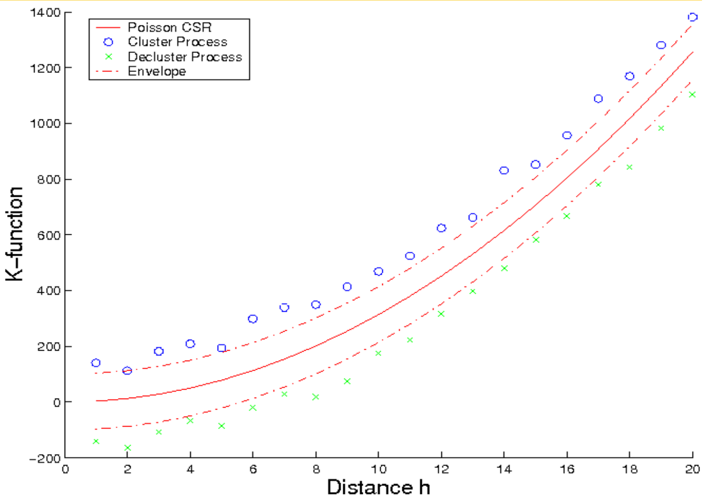
- **Purpose:** Compare a point dataset with a complete spatial random (CSR) data
- **Input:** A set of points

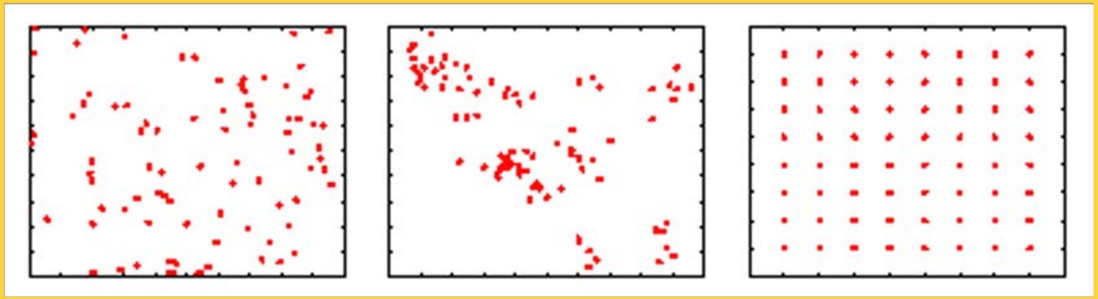
$K(h, data) \;=\; \lambda^{-1} E$ [number of events within distance *h* of an arbitrary event]

  - 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



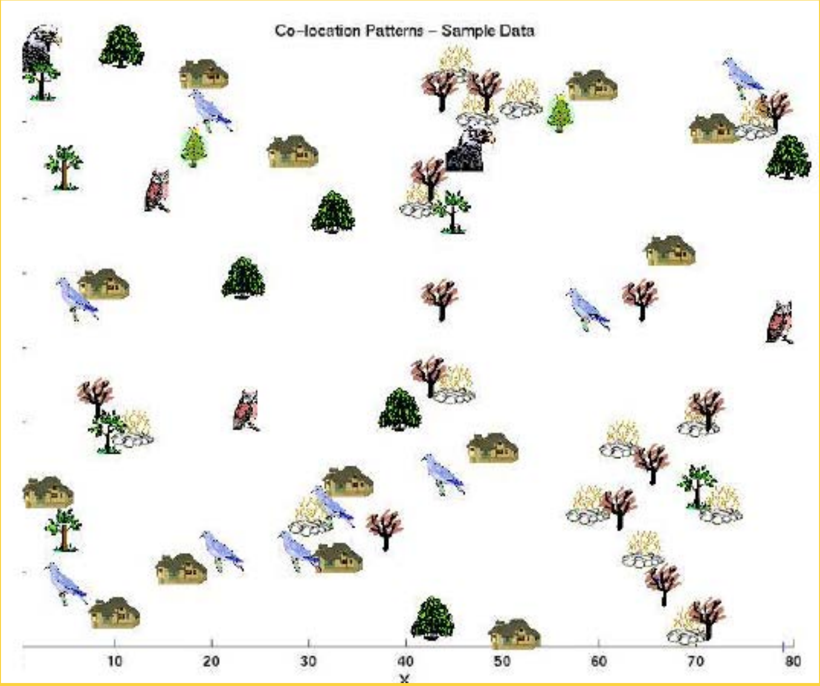


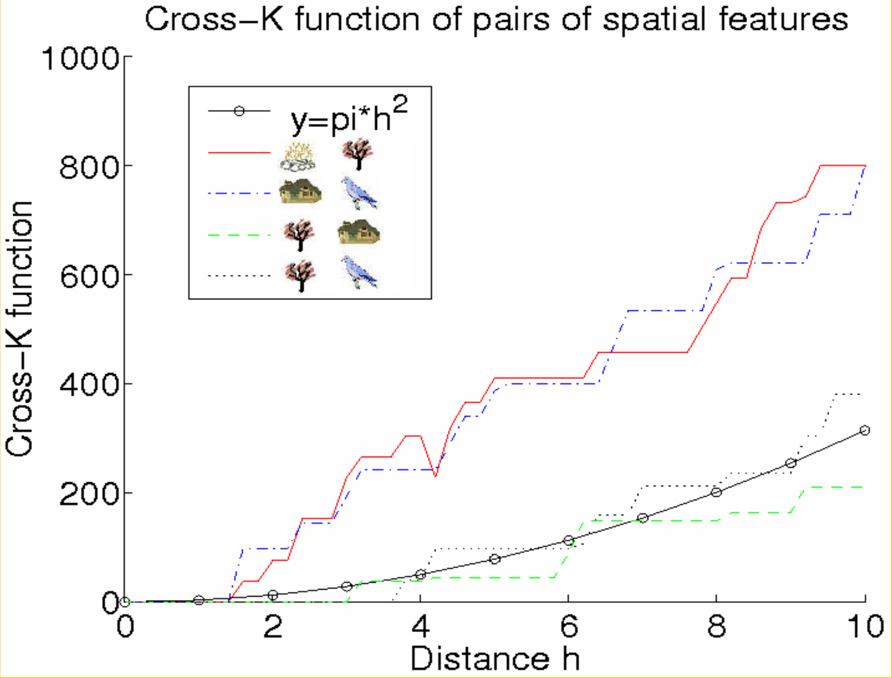
CSR Clustered De-clustered

Spatial Cross-Correlation

- Cross K-Function Definition

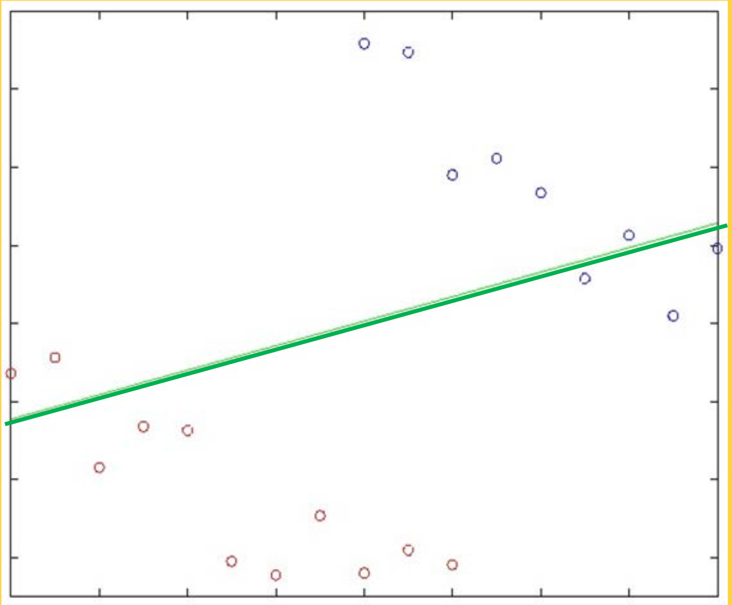
$$K_{ij}(h) = \lambda_j^{-1} E[\text{number of type } j \text{ event within distance } h \text{ of a type } i \text{ event}]$$

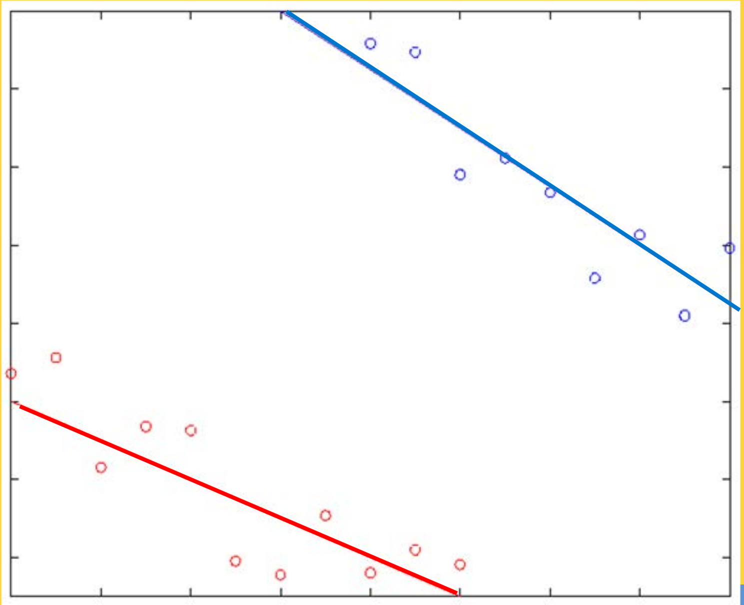




Spatial Variability and Heterogeneity

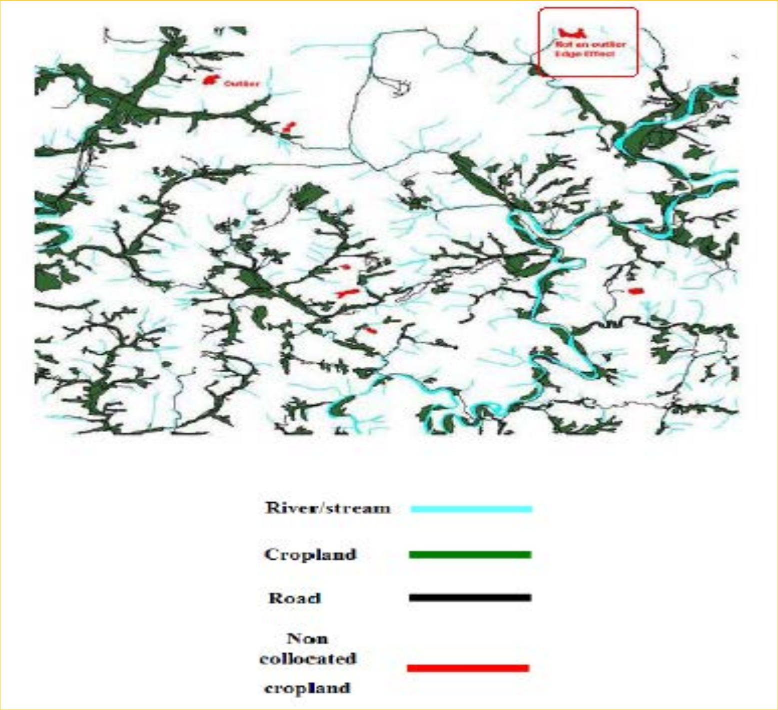
- "Second law of geography" [M. Goodchild, UCGIS 2003]
- Global model might be inconsistent with regional models
  - Spatial Simpson's Paradox (linked to MAUP)
- May improve the effectiveness of SDM, show support regions of a pattern





Edge Effect

- Cropland on edges may not be classified as outliers
- No concept of spatial edges in classical data mining



Korea Dataset,
Courtesy: Architecture Technology Corp.

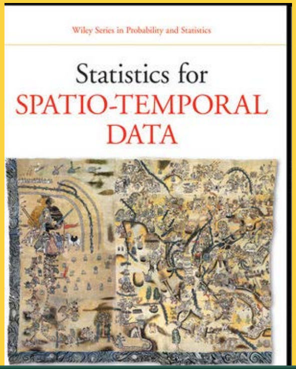
Research Needs

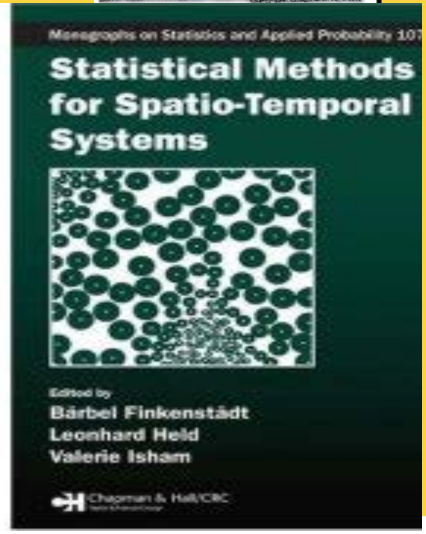
- State-of-the-art of Spatial Statistics

		Point Process	Lattice	Geostatistics
raster			√	√
Vector	Point	√	√	√
	Line			√
	Polygon		√	√
graph				

Data Types and Statistical Models

- Research Needs
  - Correlating extended features, road, rivers, cropland
  - Spatio-temporal statistics
  - Spatial graphs, e.g., reports with street address

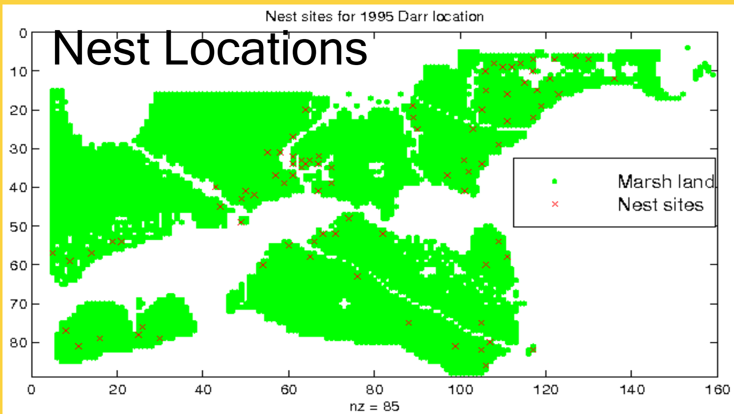


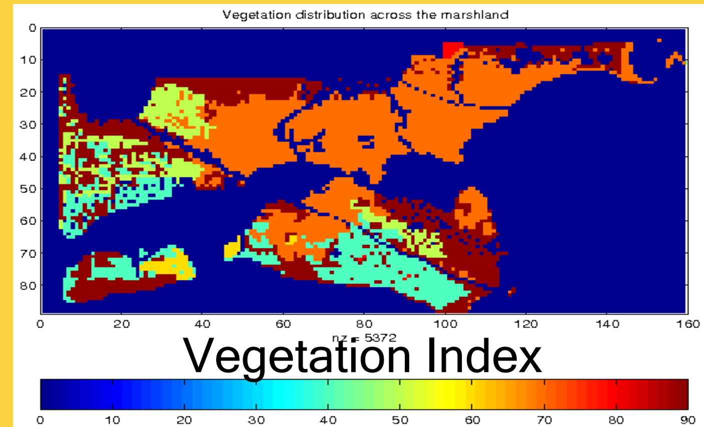


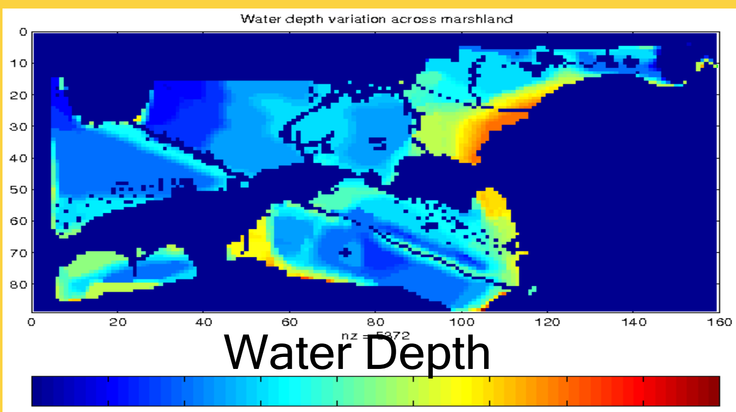
Outline

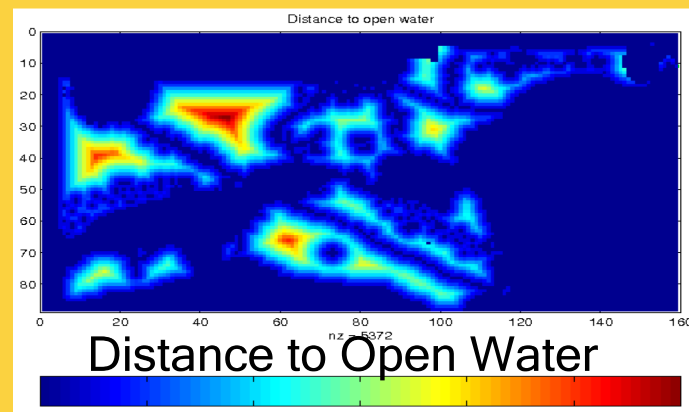
- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
  - Location Prediction
  - Hotspots
  - Spatial Outliers
  - Colocations
- Conclusions

Illustration of Spatial Prediction Problem

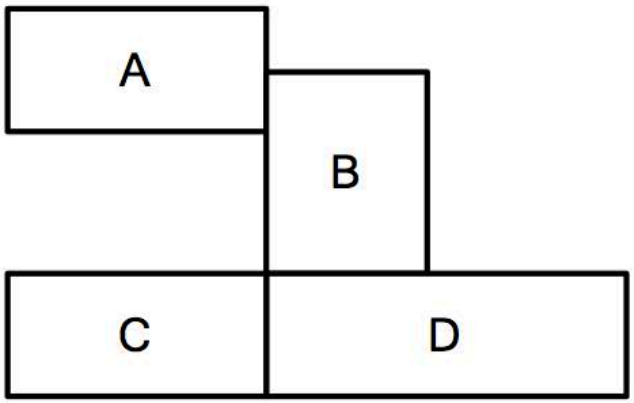








Neighbor Relationship: W Matrix



(a) Map

$$
\begin{array}{c|cccc|}
 & A & B & C & D \\
\hline
A & 0 & 1 & 0 & 0 \\
B & 1 & 0 & 1 & 1 \\
C & 0 & 1 & 0 & 1 \\
D & 0 & 1 & 1 & 0
\end{array}
$$

(b) Boolean W

$$
\begin{array}{c|cccc|}
 & A & B & C & D \\
\hline
A & 0 & 1 & 0 & 0 \\
B & 0.3 & 0 & 0.3 & 0.3 \\
C & 0 & 0.5 & 0 & 0.5 \\
D & 0 & 0.5 & 0.5 & 0
\end{array}
$$

(c) Row-normalized W

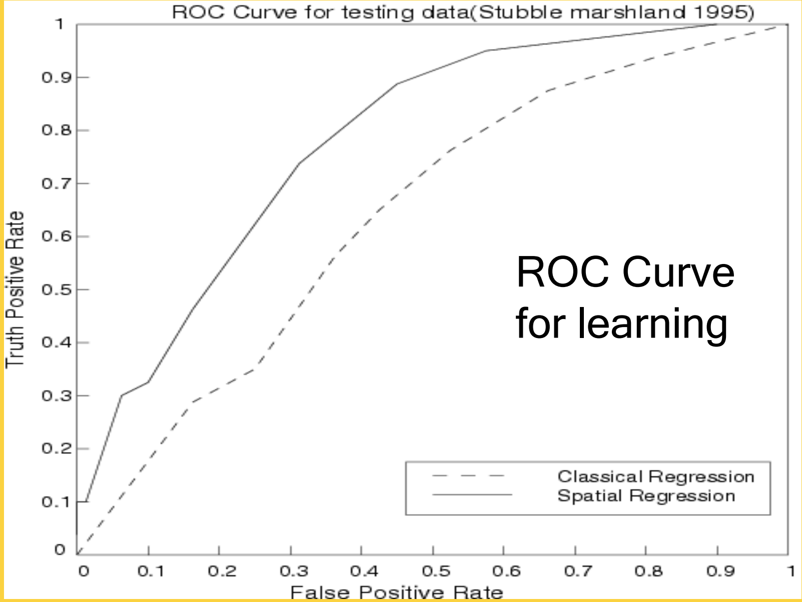
Spatial Prediction Models

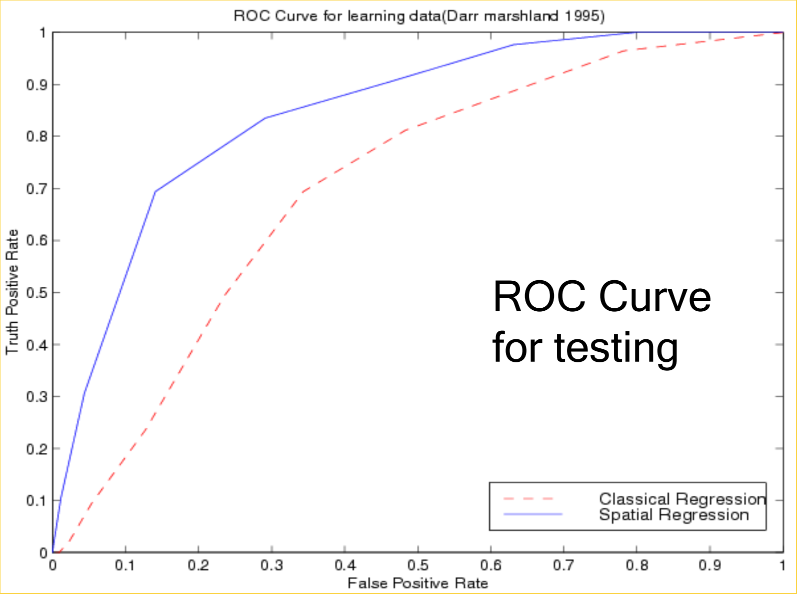
- Traditional Models, e.g., Regression (with Logit or Probit),
  - Linear Regression, Bayes Classifier, …
- Semi-Spatial : auto-correlation regularizer $\varepsilon = \|y - \beta X\|^2 + \|\beta X - \beta X_{neighbor}\|^2$
- Spatial Models
  - Spatial autoregressive model (SAR)
  - Markov random field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

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

Comparing Traditional and Spatial Models

- Dataset: Bird Nest prediction
- Linear Regression
  - Lower prediction accuracy, coefficient of determination,
  - Residual error with spatial auto-correlation
- Spatial Auto-regression outperformed linear regression





Prediction Error and Bias Trade-off

- Linear Regression (LR): Least Squares estimator

$$y = X\beta + \varepsilon$$

- LR with Auto-correlation Regularizer
  - Least squares estimator

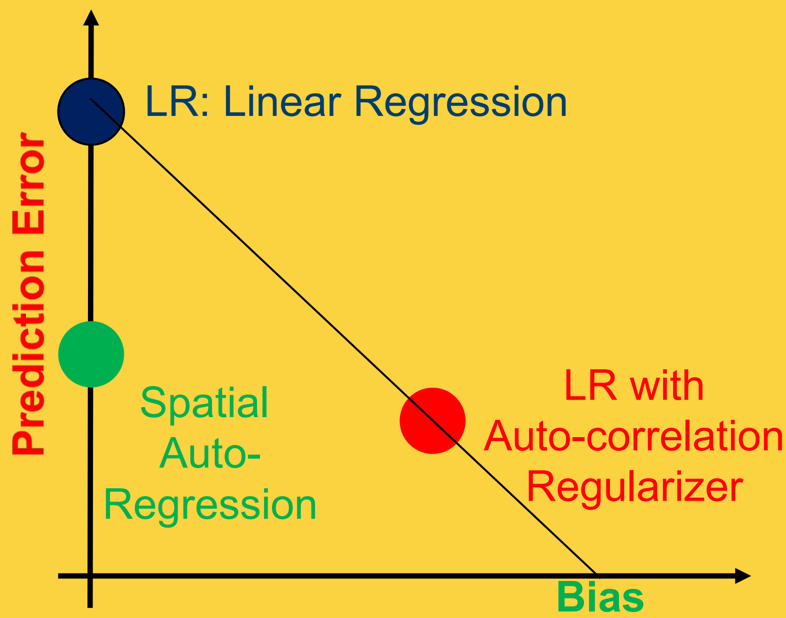
$$y = X\beta + \varepsilon$$

$$\varepsilon = \|y - \beta X\|^2 + \|\beta X - \beta X_{neighbor}\|^2$$

$$\varepsilon = \|y - \beta X\|^2 + \|y - \beta X_{neighbor}\|^2$$

- Spatial Auto-Regression:
  - Maximum Likelihood Estimator

$$y = \rho W y + X\beta + \varepsilon$$



Source: Geospatial Data Science: A Transdisciplinary Approach. In *Geospatial Data Science Techniques and Applications* (pp. 17-56). CRC Press, 2017 (E. Eftelioglu,R. Ali, X. Tang., Y. Xie, Y., Li and S. Shekhar).

Spatial Heterogeneity

- Knowledge of location can improve land-cover and object recognition (Ex. Snow vs. salt)

Salt Marsh (Runn of Kutch, India)

Snow

Snow

- Coarse Satellite Imagery (e.g., 30m pixels)
  - Better for mono-crop farms than mixed-crop plots

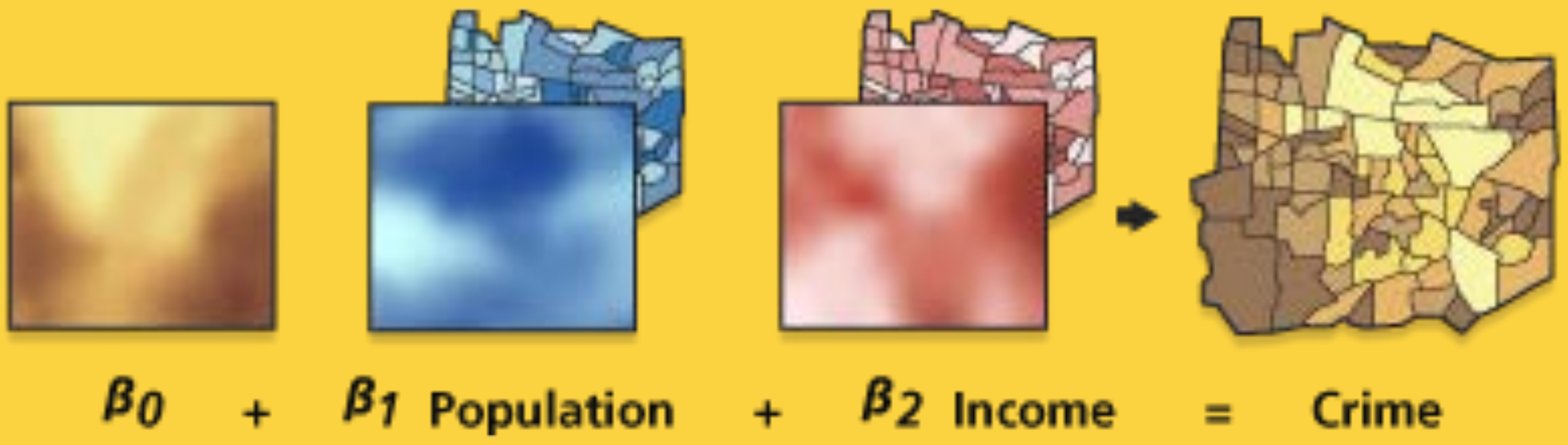
However, Convolutional Neural Networks does not model geographic heterogeneity.

Q? Which of these problem may be addressed by “attention” in DNN ?

Geographically Weighted Regression (GWR)

- Goal: Model spatially varying relationships
- Example: $y = X\beta' + \varepsilon'$

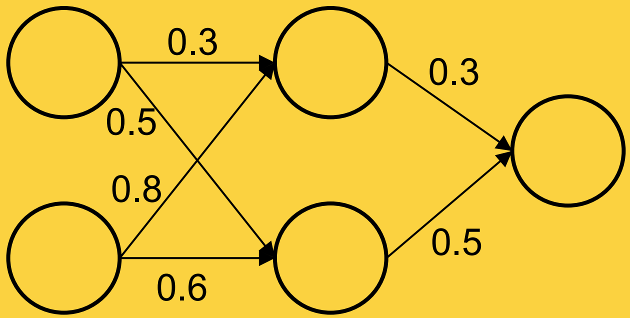
  Where β' and ε' are location dependent



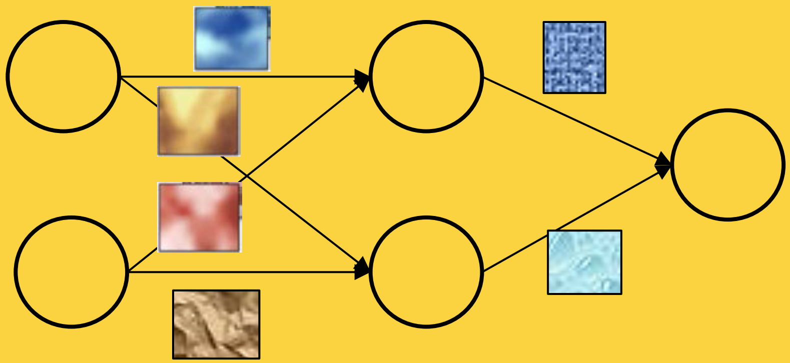
Source: resources.arcgis.com

Spatial Variability Aware Neural Networks (SVANN)

- Each NN parameter is a map i.e., a function of location
  - Similar to Geographically Weighted Regression



A Neural Network (NN)



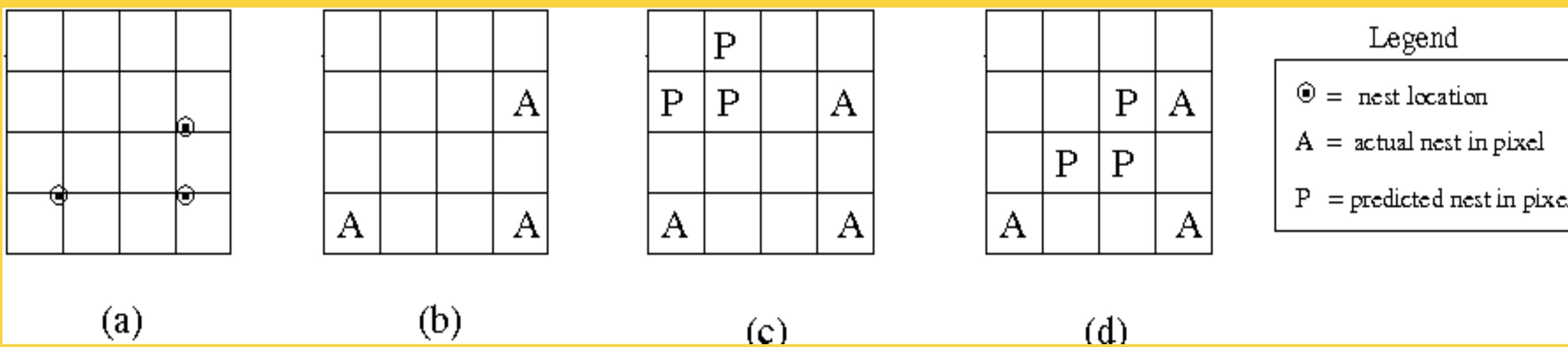
SVANN

- Evaluation:
  - Urban Garden Detection across Hennepin County, MN and Fulton County, GA.
  - SVANN outperformed OSFA by 14.34% on F1-scores.

Details: Towards Spatial Variability Aware Deep Neural Networks (SVANN), ACM Trans. on Intelligent Systems and Tech, 12(6):1-21, 2021. (A Summary in ACM SIGKDD DeepSpaial, 2020. (Best Paper Award)

Research Needs in Spatial Prediction

- Spatial Auto-Regression
  - Estimate W
  - Scaling issue $\rho W y$ vs. $X\beta$
- Spatial interest measure
  - e.g., distance(actual, predicted)



Actual Sites

Pixels with actual sites

Prediction 1

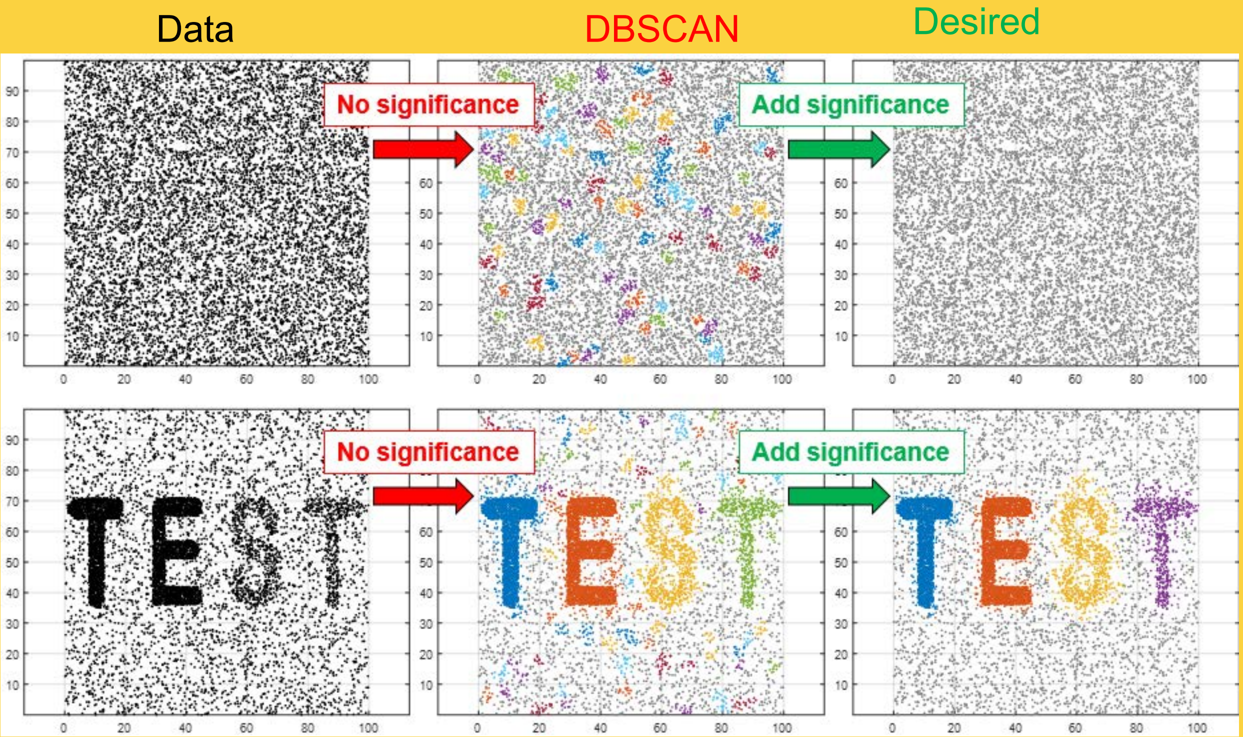
Prediction 2. Spatially more interesting than Prediction 1

Outline

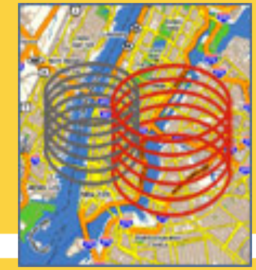
- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
    - Location Prediction
    - Hotspots
    - Spatial Outliers
    - Colocations
- Conclusions

Limitations of Classical Clustering Methods

- Easily fooled by noise



Spatial Scan Statistics



SaTScan™

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

- Goal: Omit chance clusters

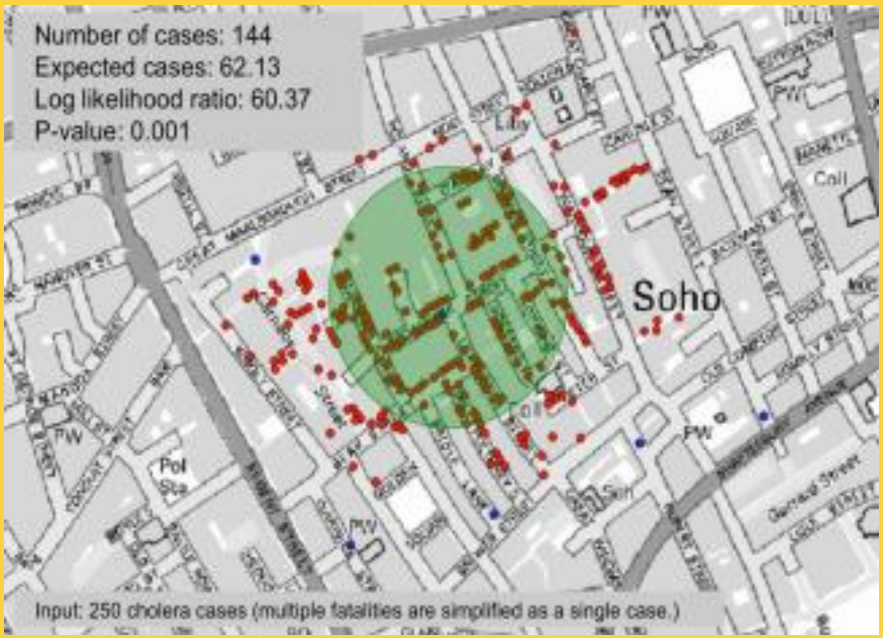
- Ideas: Likelihood Ratio, Statistical Significance

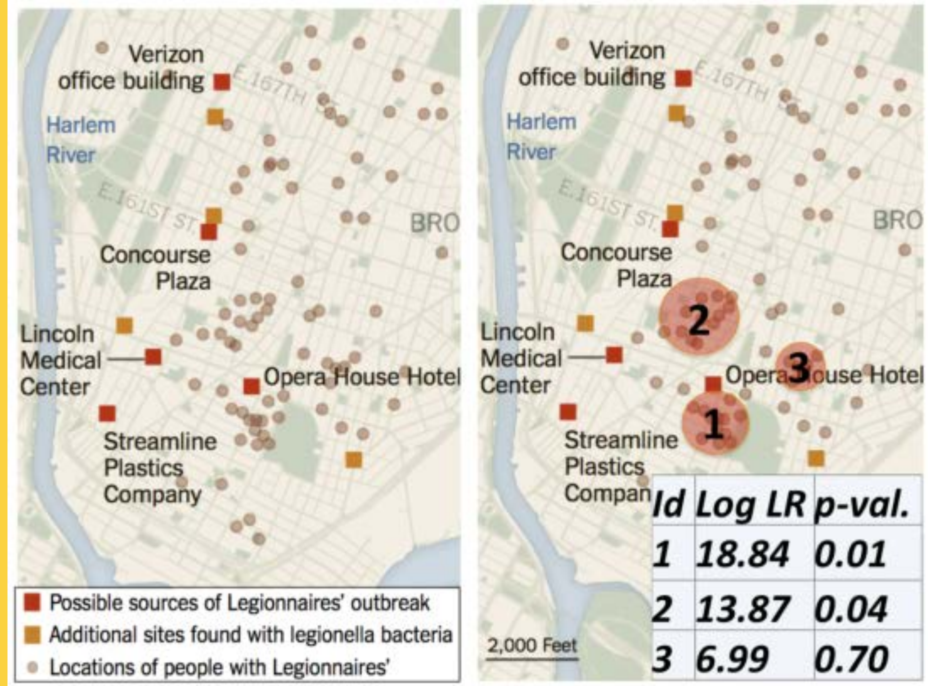
- Steps
  - Enumerate candidate zones & choose zone X with highest likelihood ratio (LR)
    - LR(X) = p(H1|data) / p(H0|data)
    - H0: points in zone X show complete spatial randomness (CSR)
    - H1: points in zone X are clustered

  - If LR(Z) >> 1 then test statistical significance
    - Check how often is LR(CSR) > LR(Z) using 1000 Monte Carlo simulations

SaTScan Examples

1854 London Cholera, p-value = 0.001



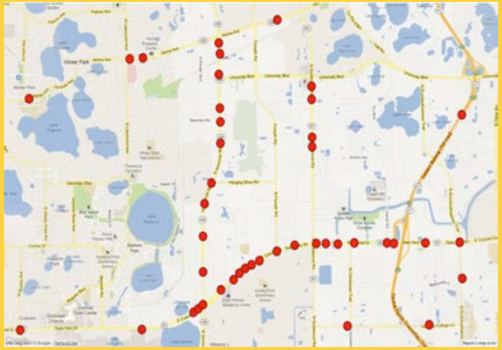


(a) Legionnaire's in New York (2015) (b) Output of SaTScan

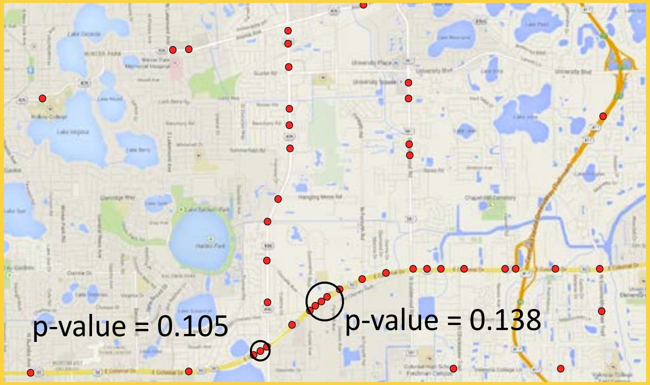
<u>Source: Ring-Shaped Hotspot Detection,</u> IEEE Trans. Know. & Data Eng., 28(12), 2016.
(A Summary in Proc. IEEE ICDM 2014) (w/ E. Eftelioglu et al.)

Non-circular Hotspots

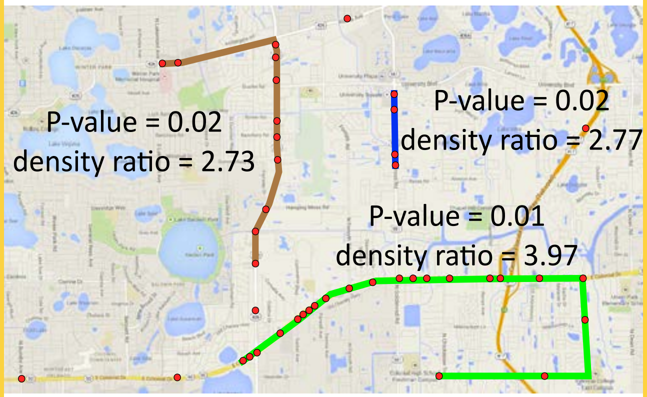
- Geographic features, e.g., rivers, streams, roads, …
  - Hot-spots => Hot Geographic-features, e.g., Linear Hotspots
- Spatial Theories, e.g,, environmental criminology
  - Circles ➔ Doughnut holes



Pedestrian fatalities
Orlando, FL



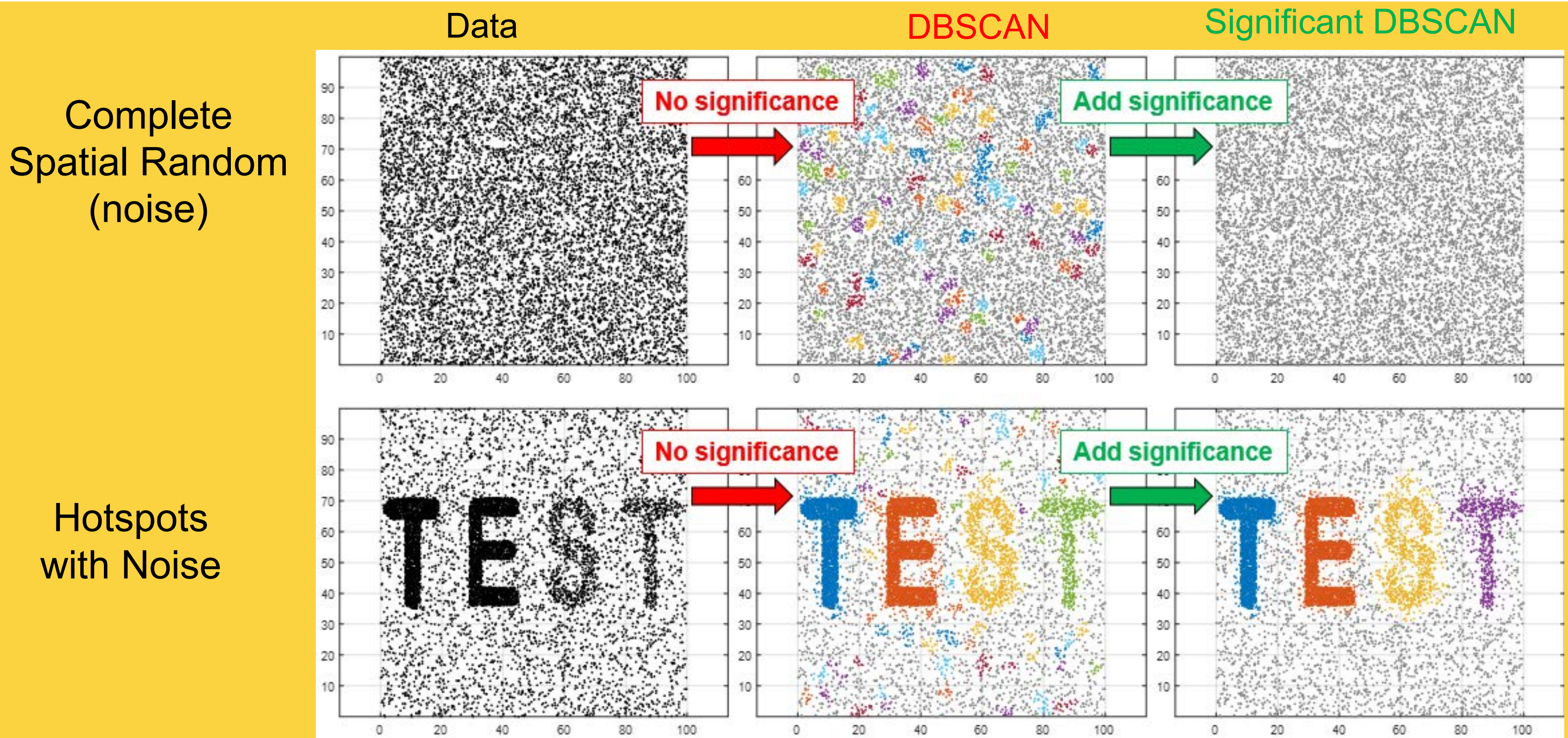
Circular hotspots
by SatScan



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.

Hotspots with Flexible Shapes

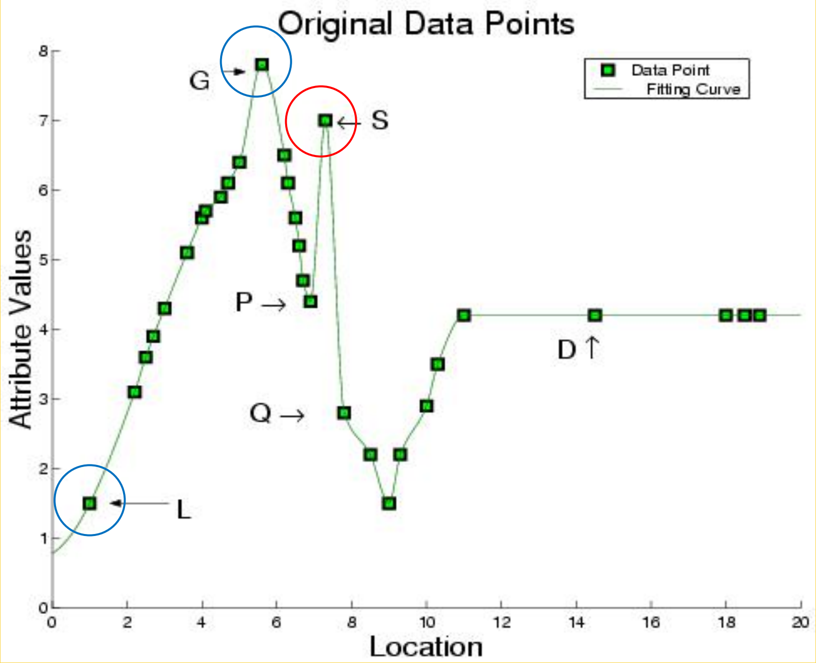


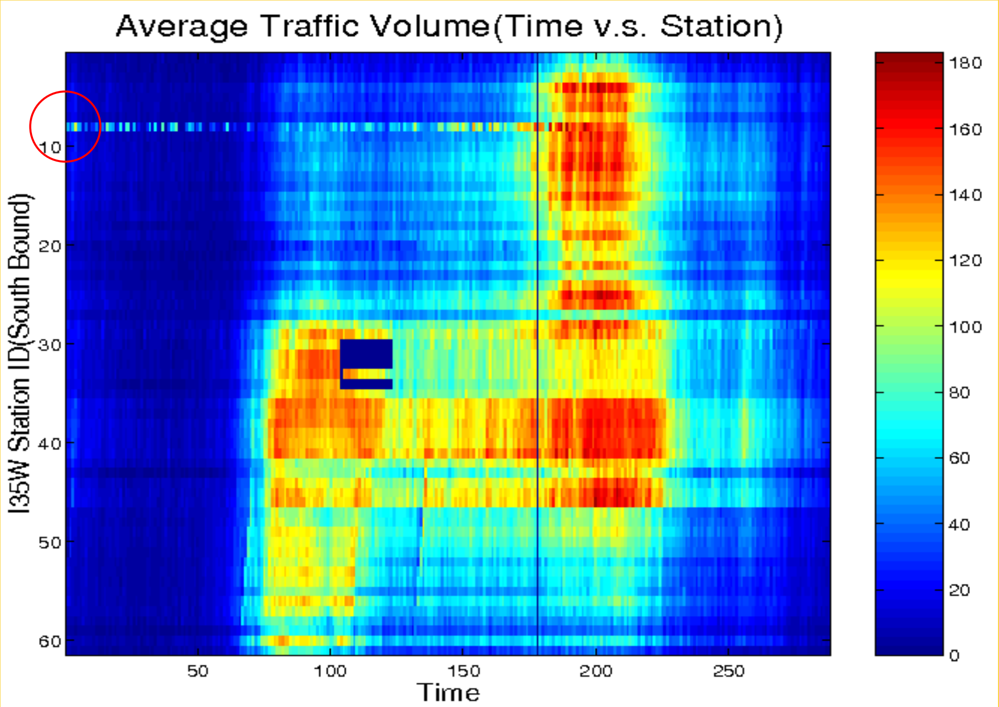
Details: Significant DBSCAN towards Statistically Robust Clustering, [ACM Trans. on Intelligent Systems and Tech](), 12(5):1-26, Oct. 2021. (A summary in 16th Intl. Symp. on Spatial and Temporal Databases, 2019. **(Best Paper Award)**

Outline

- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
    - Location Prediction
    - Hotspots
    - Spatial Outliers
    - Colocations
- Conclusions

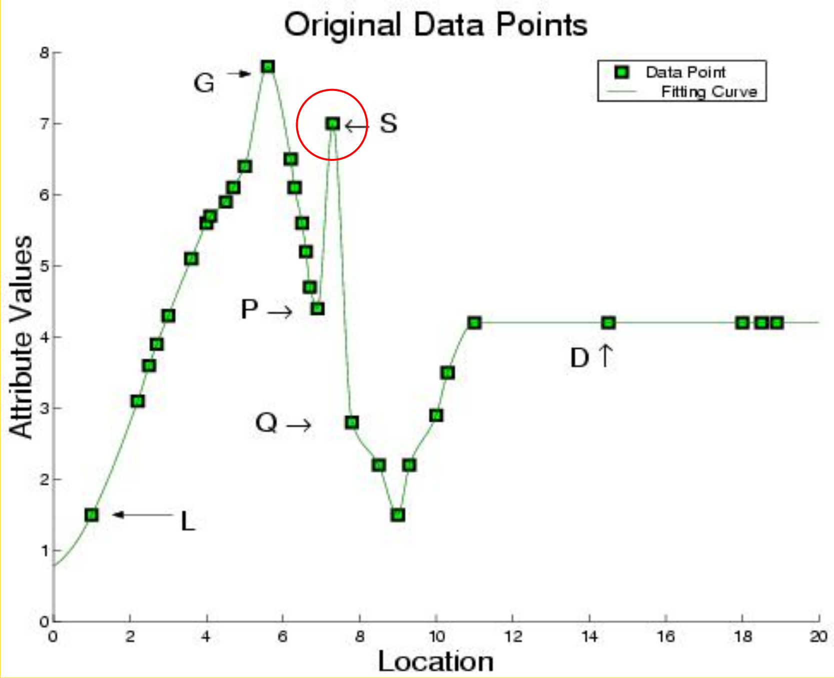
Outliers: Global (G) vs. Spatial (S)

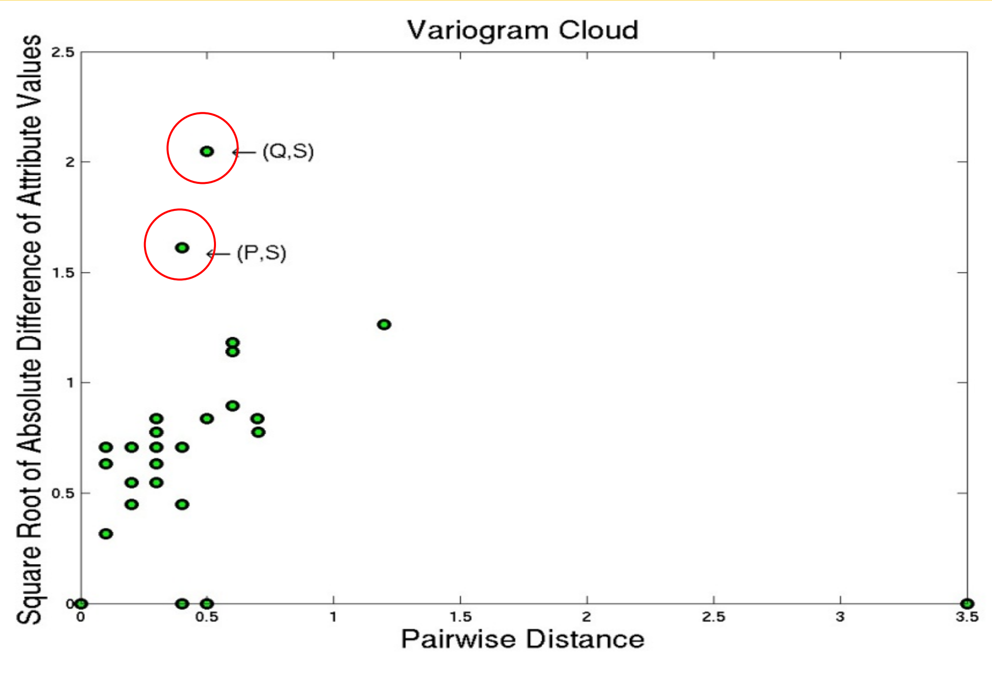




Outlier Detection Tests: Variogram Cloud

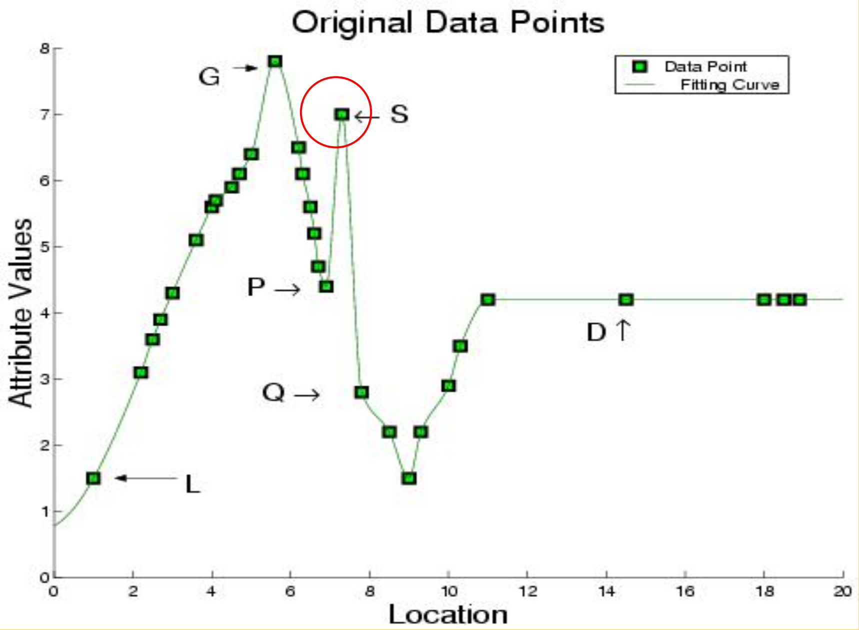
- Graphical Test: Variogram Cloud

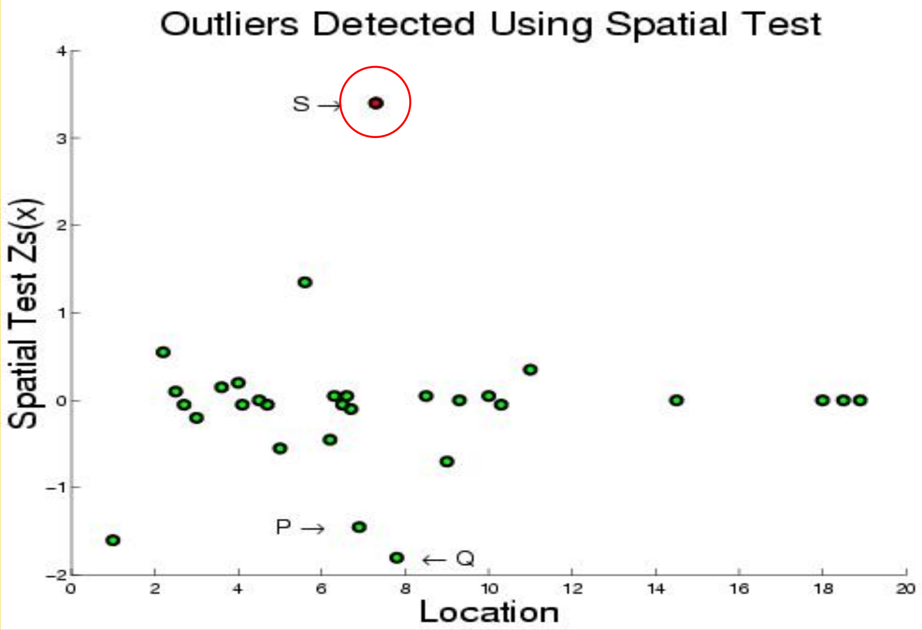




Outlier Detection Tests: Spatial Z-test

- Quantitative Tests: Spatial Z-test
  - Algorithmic Structure: Spatial Join on neighbor relation

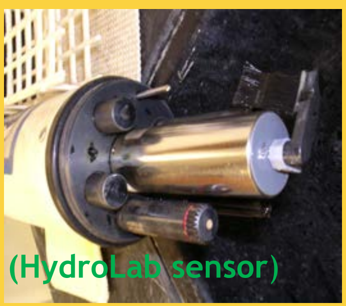


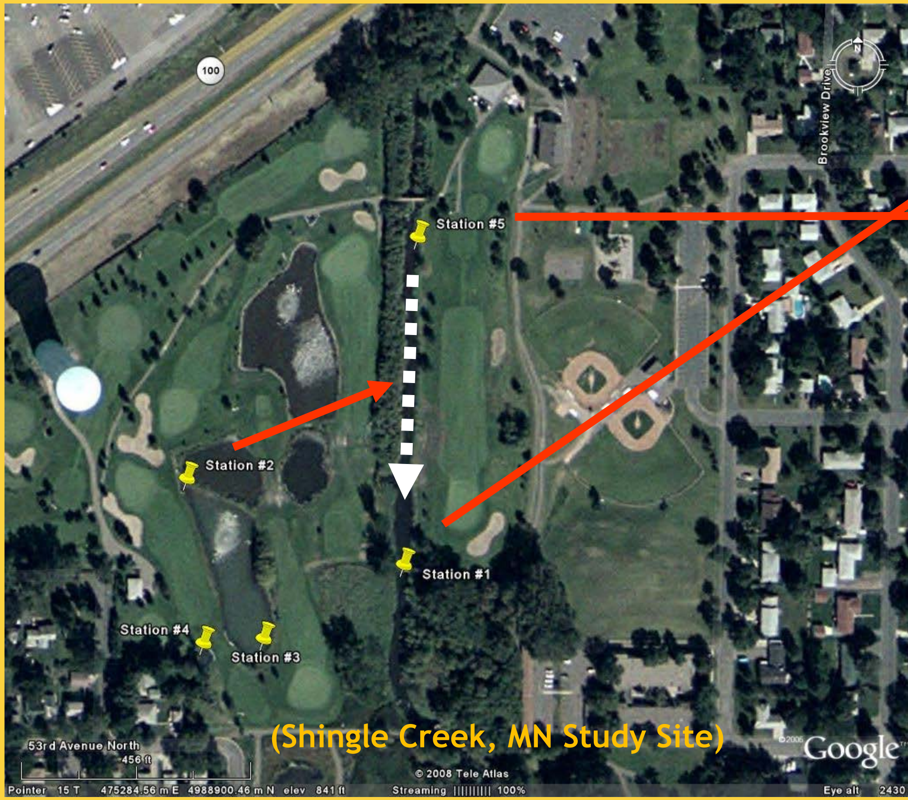


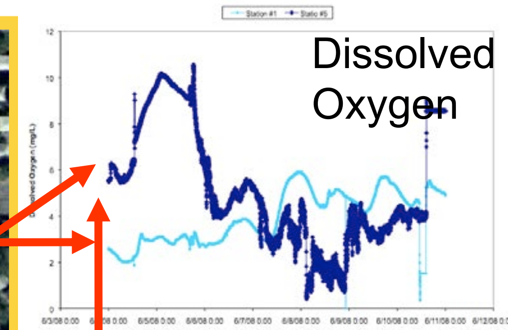
Flow Anomalies

Example **Forensics: When and where do contaminants enter a Creek?**

www.sfgate.com/cgi-bin/news/oilspill/busan

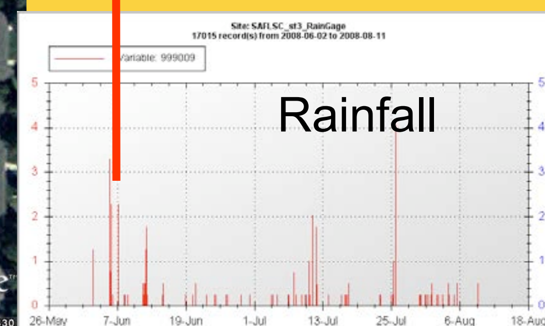






6/4/08 13:06 - 6/5/08 19:34

After heavy rain (June 4 & 5)



Details: Discovering Flow Anomalies: A SWEET Approach, IEEE Intl. Conf. on Data Mining, 2008 (w/J. Kang et al.).

Spatial Outlier Detection: Computation

- Separate two phases
  - Model Building
  - Testing: test a node (or a set of nodes)

- Computation Structure of Model Building
  - Key insights:
    - Spatial self join using N(x) relationship
    - Algebraic aggregate function computed in one scan of spatial join

Trends in Spatial Outlier Detection

- Multiple spatial outlier detection
  - Eliminating the influence of neighboring outliers
- Multi-attribute spatial outlier detection
  - Use multiple attributes as features
- Spatio-temporal anomalies
  - Anomalous trajectories, patterns of life
- Scale up for large data

Outline

- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
    - Location Prediction
    - Hotspots
    - Spatial Outliers
    - Colocations
- Conclusions

Background: Association Rules

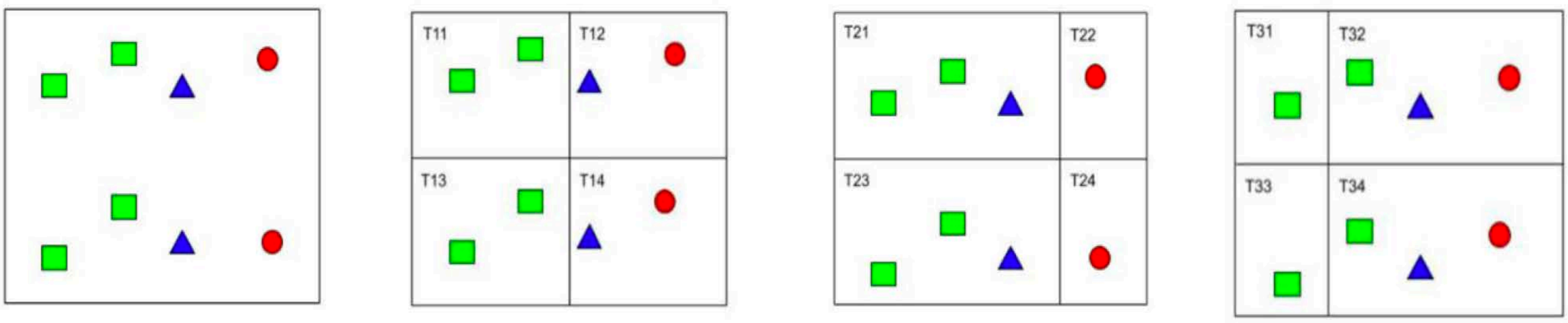
- Association rule e.g. (Diaper in T => Beer in T)

Transaction	Items Bought
1	{socks, , milk, , beef, egg, …}
2	{pillow, , toothbrush, ice-cream, muffin, …}
3	{ , , pacifier, formula, blanket, …}
…	…
n	{battery, juice, beef, egg, chicken, …}

  - Support: probability (Diaper and Beer in T) = 2/5
  - Confidence: probability (Beer in T \| Diaper in T) = 2/2

- Apriori Algorithm
  - Support based pruning using monotonicity
  - Computationally efficient, scales to larger dataset than correlation coefficient

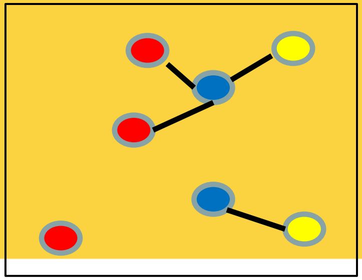
Limitations of Association Rules



(a) Map of 3 item-types (b) Spatial Partition P1 (c) Spatial Partition P2 (d) Spatial Partition P3

Spatial Partitioning	**P1**	**P2**	**P3**
Transactions	T11, T12, T13, T14	T21, T22, T23, T24	T31, T32, T33, T44
Associations with support >= 0.5	()	()	()

Spatial Colocation



Feature set: (, ,)

Feature Subsets:

Participation ratio (pr):

pr(,) = fraction of instances neighboring feature { } = 2/3

pr(,) = ½

Participation index (A,B.) = **pi**(A,B.)

= min{ **pr**(A. , (A, **pr**(B B).) }

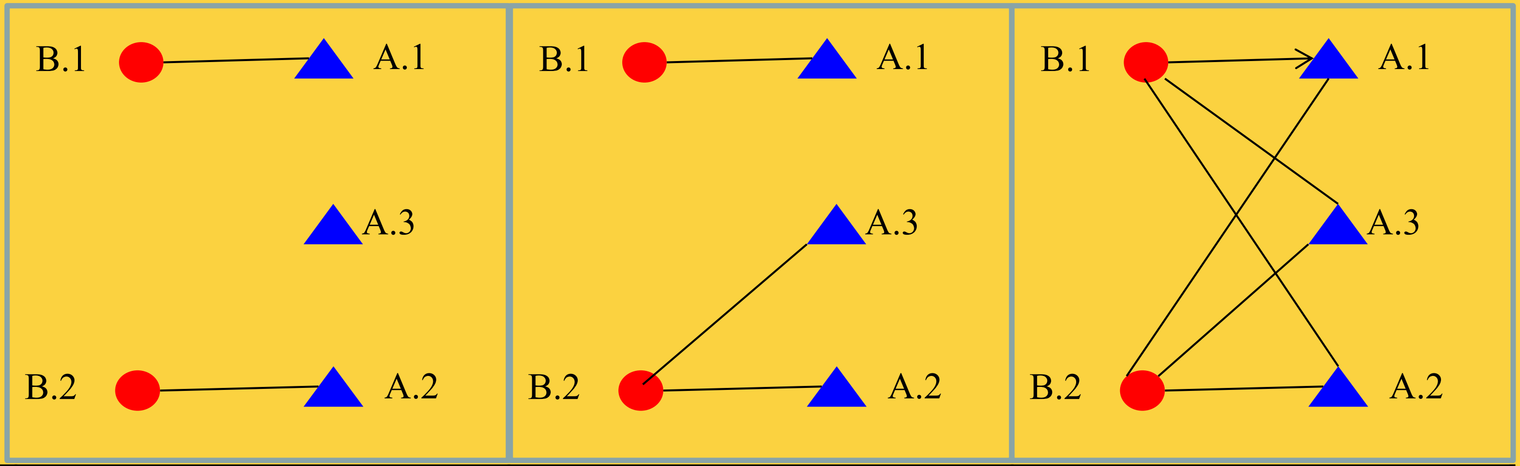
= min (2/3,) =

Participation Index Properties:

(1) Computational: Non-monotonically decreasing like support measure

(2) Statistical: Upper bound on Ripley's Cross-K function

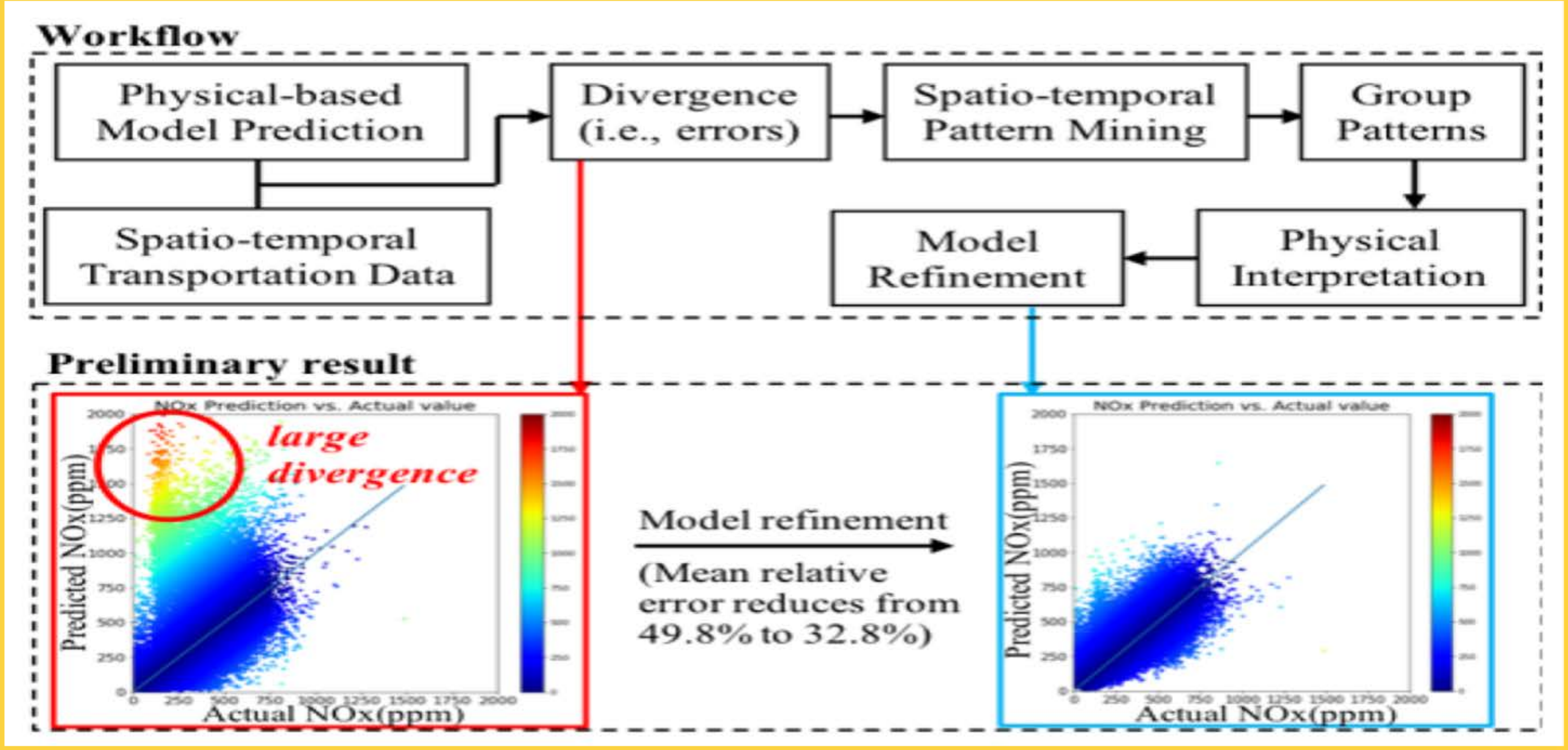
Participation Index >= Cross-K Function



Cross-K (A,B)	2/6 = 0.33	3/6 = 0.5	6/6 = 1
PI (A,B)	2/3 = 0.66	1	1

Co-occurrence Patterns to Refine a Physical Model

Details: R. Ali, V. Gunturi, A. Kotz, E. Eftelioglu, S. Shekhar, and W. Northrop "*Discovering Non-compliant Window Co-Occurrence Patterns.*" GeoInformatica, 21(4), 829-866 (2017).



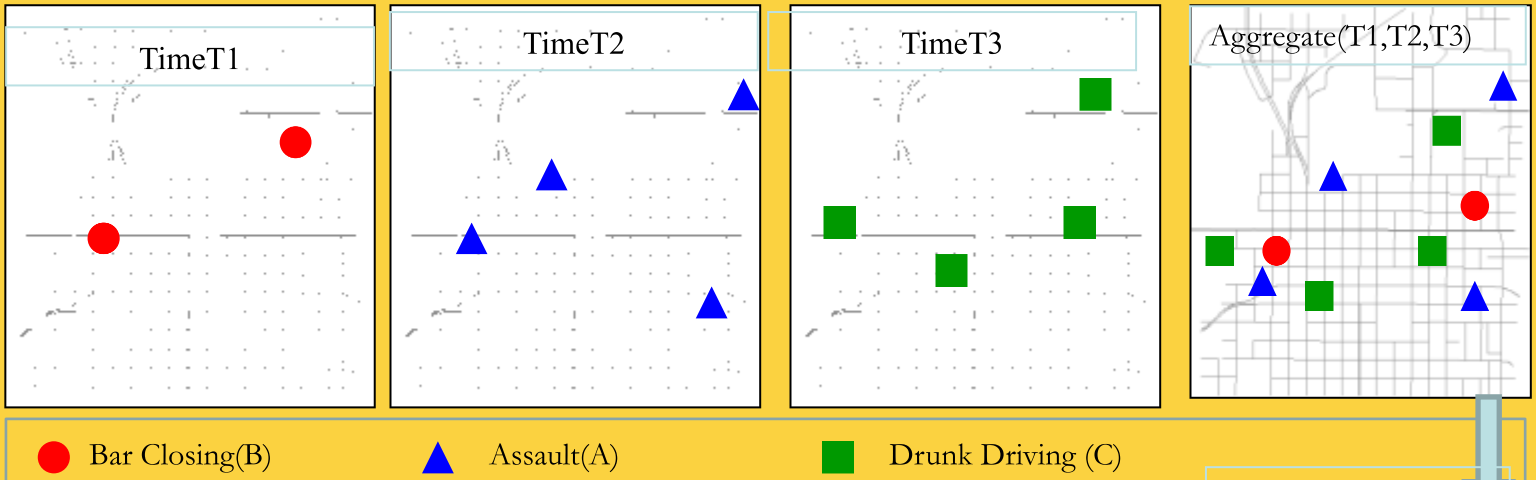
Spatial Colocation: Trends

- Algorithms
  - Join-based algorithms
    - One spatial join per candidate colocation
  - Join-less algorithms

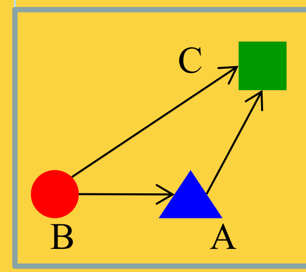
- Statistical Significance
  - ?Chance-patterns

- Spatio-temporal
  - Which events co-occur in space and time?
    - (bar-closing, minor offenses, drunk-driving citations)
  - Which types of objects move together?

Cascading spatio-temporal pattern (CSTP)



CSTP: P1



- *Input:* Urban Activity Reports
- *Output: CSTP*
  - *Partially ordered* subsets of ST event types.
  - Located together in space + Occur in *stages* over time.
- Applications: Public Health, Public Safety, …

Details: Cascading Spatio-Temporal Pattern Discovery, IEEE Trans. on Know. & Data Eng, 24(11), 2012.

Outline

- Motivation
  - Use cases
  - Pattern families
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
- Conclusions

Summary

What's Special About Mining Spatial Data ?

		Spatial DM	Spatio-Temporal DM
Input Data		Often implicit relationships, complex types	Another dimension – Time. Implicit relationships changing over time
Statistical Foundation		Spatial autocorrelation	Spatial autocorrelation and Temporal correlation
Output	Association	Colocation	Frequent Patterns of Change
	Clusters	Hot-spots	Flock pattern Moving Clusters
	Outlier	Spatial outlier	Change Detection
	Prediction	Location prediction	Future Location prediction

References :Surveys, Overviews

- Spatial Computing, The MIT Press Essential Knowledge series, Feb. 2020.
- Spatial Computing (html , short video , tweet), Communications of the ACM, 59(1):72-81, Jan. 2016.
- AM-97 - An Introduction to Spatial Data Mining , The Geographic Information Science & Technology Body of Knowledge, 2020, J. Wilson (Ed.). DOI:10.22224/gistbok/2020.4.5. (Also UMN CS technical report 18-013, 2018).
- 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.

References: Details

Colocations	• Discovering colocation patterns from spatial data sets: a general approach, *IEEE Trans. on Know. and Data Eng*., 16(12), 2004 (w/ Y. Huang et al.). • A join-less approach for mining spatial colocation patterns, IEEE Trans. on Know. and Data Eng.,18 (10), 2006. (w/ J. Yoo). • Cascading Spatio-Temporal Pattern Discovery. IEEE Trans. Knowl. Data Eng. 24(11): 1977-1992, 2012 (w/ P. Mohan et al.).
Spatial Outliers	• Detecting graph-based spatial outliers: algorithms and applications (a summary of results), Proc.: ACM Intl. Conf. on Knowledge Discovery & Data Mining, 2001 (with Q. Lu et al.) • A unified approach to detecting spatial outliers, Springer GeoInformatica, 7 (2), 2003. (w/ C. Lu, et al.) • Discovering Flow Anomalies: A SWEET Approach, IEEE Intl. Conf. on Data Mining, 2008 (w/ J. Kang).
Hot Spots	• Discovering personally meaningful places: An interactive clustering approach, ACM Trans. on Info. Systems (TOIS) 25 (3), 2007. (with C. Zhou et al.) • A K-Main Routes Approach to Spatial Network Activity Summarization, IEEE Trans on Know. & Data Eng., 26(6), 2014. (with D. Oliver et al.) • Significant Linear Hotspot Discovery, IEEE Trans. Big Data 3(2): 140-153, 2017, (w/ X.Tang et al.) • Statistically-Robust Clustering Techniques for Mapping Spatial Hotspots: A Survey, ACM Computing Surveys, 55(2):1-38, March 2023, https://doi.org/10.1145/3487893
Location Prediction	• Spatial contextual classification and prediction models for mining geospatial data, IEEE Transactions on Multimedia, 4 (2), 2002. (with P. Schrater et al.) • Focal-Test-Based Spatial Decision Tree Learning. IEEE Trans. Knowl. Data Eng. 27(6): 1547-1559, 2015 (summary in Proc. IEEE Intl. Conf. on Data Mining, 2013) (w/ Z. Jiang et al.).
Change Detection	• Spatiotemporal change footprint pattern discovery: an inter-disciplinary survey. Wiley Interdisc. Rew.: Data Mining and Know. Discovery 4(1), 2014. (with X. Zhou et al.) https://doi.org/10.1002/widm.1113