What is special about mining spatial data?

April 9th, 2018

Quantitative Epidemiology seminar series

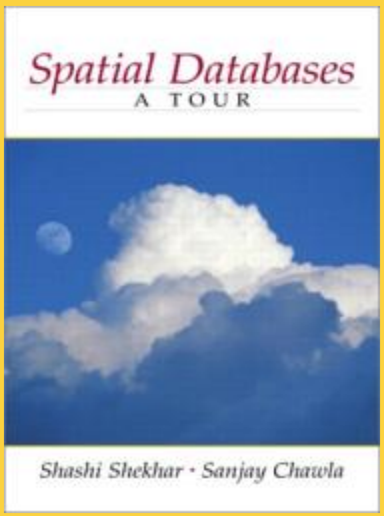
Department of Veterinary Population Medicine, University of Minnesota.

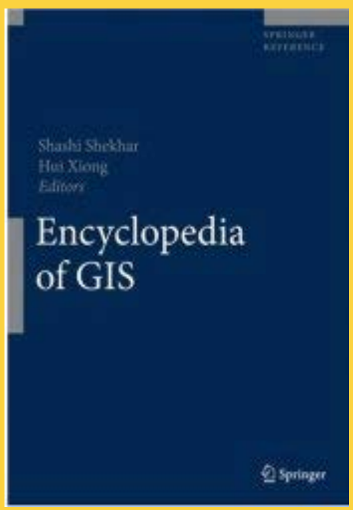
Shashi Shekhar

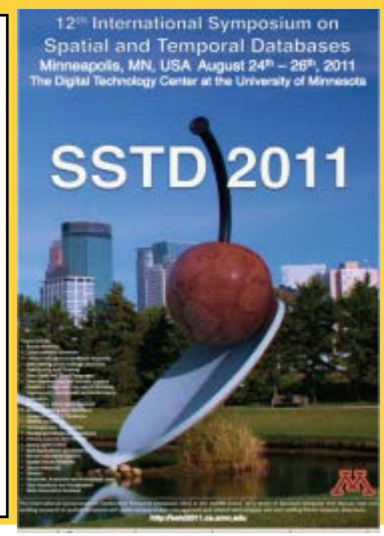
McKnight Distinguished University Professor

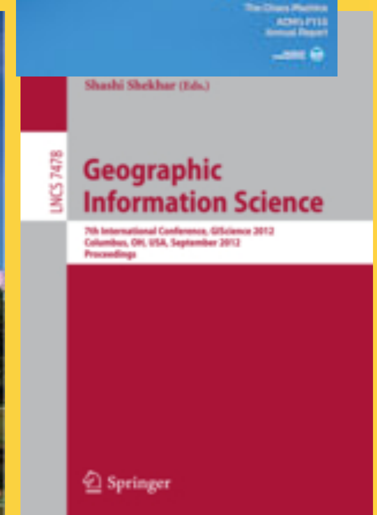
Dept. of Computer Sc. and Eng., University of Minnesota

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









UNIVERSITY OF MINNESOTA

Driven to Discover℠

Acknowledgements

- P.I., Connecting the Smart-City Paradigm with a Sustainable Urban Infrastructure Systems Frame- work to Advance Equity in Communities, National Science Foundation (Award 1737633), $2.5 M, 9/1/2017 - 8/31/2020.
- P.I., III: Small: Investigating Spatial Big Data for Next Generation Routing Services (IIS-1320580), National Science Foundation (NSF), $0.5M, 9/15/2013- 8/31/2018.
- P.I., Identifying and Analyzing Patterns of Evasion (HM0210-13-1-0005), USDOD National Geospatial Intelligence Agency (NGA) $0.6M, 6/10/2013- 9/9/2018.
- Co-P.I., Cloud-Connected Delivery Vehicles: Boosting Fuel Economy Using Physics-Aware Spatio- temporal Data Analysis and Real-Time Powertrain Control, USDOE ARPA-E, $1.78M (1.4M federal), 2/17/2017 - 2/16/2020. (PI: W. Northrop)
- Co-P.I., Increasing Low-Input Turfgrass Adoption Through Breeding, Innovation, and Public Education, Speciality Crop Research Initiative, National Institute for Food and Agriculture (contract 2017-51181-27222), USDA, $5.4 M, 9/1/2017 - 8/31/2021. (with E. Watkins).

Spatial Computing Examples

Arc GIS X ESRI

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

OGC™
Open Geospatial Consortium, Inc.

ORACLE
SPATIAL

IBM Smarter Planet

Google Earth

OpenStreetMap

Ushahidi

waze
Outsmarting Traffic, Together.

Deconstructing Precision Agriculture

#AgInnovates2015

Wednesday, March 4, 2015
Reception | 5:00 to 7:00 pm
House Agriculture Committee Room,
1300 Longworth House Office Building,
Washington, DC

Think Moon landing.

Think Internet.

Think iPhone and Google.

Think bigger.

Come hear U.S. farmers, leading agriculture technology companies, and scientists tell how they work together to fuel U.S. innovation and the economy to solve this global challenge.

The event will exhibit three essential technologies of precision agriculture that originated from a broad spectrum of federally funded science: Guidance Systems and GPS, Data & Mapping with GIS, and Sensors & Robotics.

Moderator

Raj Khosla, Professor of Precision Agriculture at Colorado State Univ.

Farmers

David Hula, of Renwood Farms in Jamestown, Virginia

Rod Weimer, of Fagerberg Produce in Eaton, Colorado

Del Unger, of Del Unger Farms near Carlisle, Indiana

Speakers

Mark Harrington, Vice President of Trimble

Carl J. Williams, Chief of the Quantum Measurement Division at NIST

Bill Raun, Professor at Oklahoma State Univ.

Marvin Stone, Emeritus Professor at Oklahoma State Univ.

J. Alex Thomasson, Professor at Texas A&M Univ.

Dave Gebhardt, Director of Data and Technology at Land O'Lakes/WinField

Shashi Shekhar, Professor at the Univ. of Minnesota

RSVP

http://bit.ly/1CoOYoa

Hosted by the Congressional Soils Caucus

In partnership with

Agricultural Retailers Association
American Society of Plant Biologists
American Physical Society
American Society of Agronomy
Association of Equipment Manufacturers
Coalition for the Advancement of Precision Agriculture
Computing Research Association
CropLife America
Crop Science Society of America
PrecisionAg Institute
Soil Science Society of America
Task Force on American Innovation
Texas A&M AgriLife
Trimble
WinField

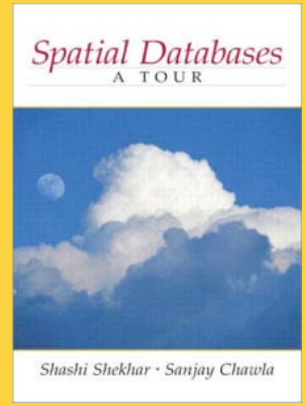
This is about feeding the world.

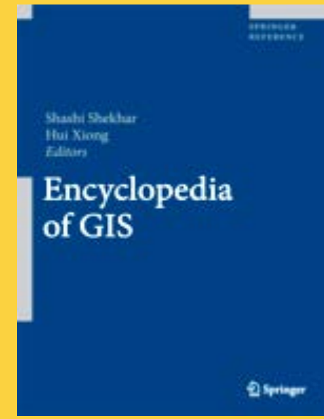
Courses

Csci 5715: Spatial Computing

www.spatial.cs.umn.edu/Courses/Fall17/5715/

- Computing in Navigation, e.g., Google Maps
- Spatial Database Management (SQL3/OGC)
- Spatial Data Mining
- Positioning, e.g., GPS, wi-fi
- Computing in Cartography & Remote Sensing





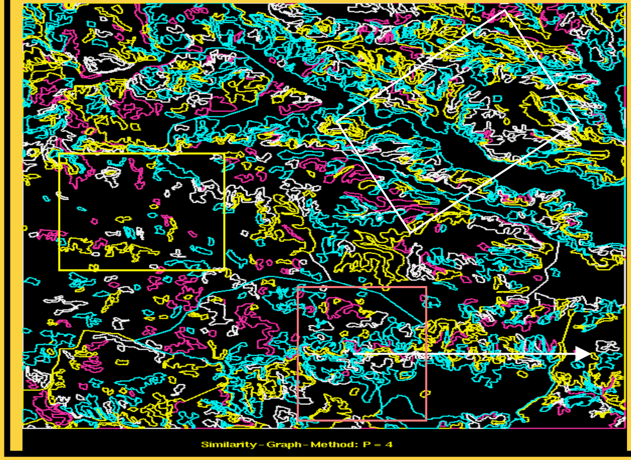
Csci 8715: Spatial Data Science Research

www.spatial.cs.umn.edu/Courses/Spring18/8715/

- Data-driven Sciences: food, energy, water, climate, smart cities, connected cars, spatial thinking, …
- Spatial Data Sciences: data models, query languages, spatial networks, spatial data mining & optimization, …
- Platforms from sensors to cloud
- Trends: spatio-temporal big data, indoors, GPS III, continuous earth observation, accountability, fairness, …

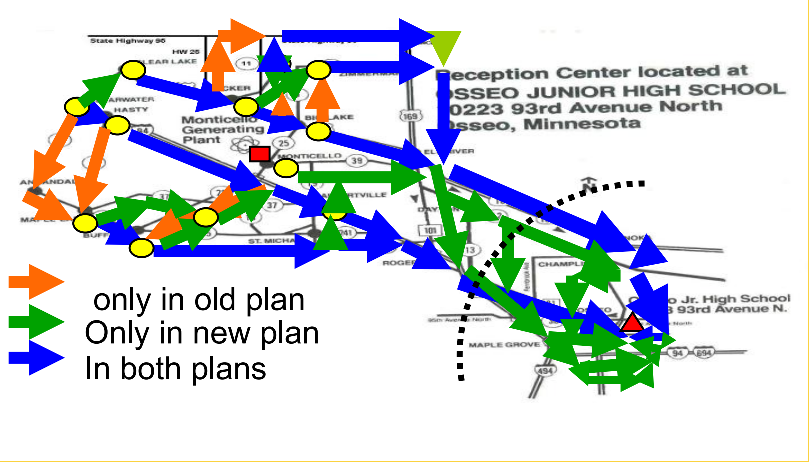
Spatial Databases: Representative Projects

Details: Spatial Databases: Accomplishments and Research Needs, IEEE Transactions on Knowledge and Data Engineering, 11(1), 1999. (and recent update via a technical report)

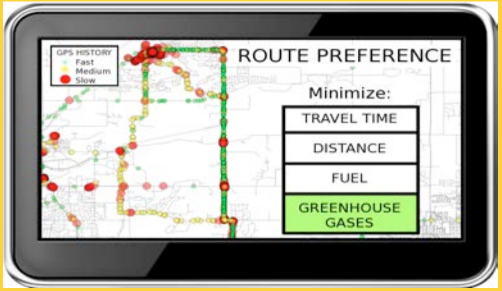


Parallelize Range Queries

Evacutation Route Planning



Eco-Routing



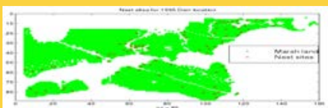
Storing graphs on disk

Spatial Data Mining: Example Projects

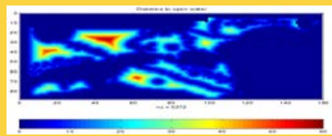
Details: Identifying patterns in spatial information: a survey of methods, Wiley Interdisc. Reviews: Data Mining and Know. Discovery , 1(3):193-214, May/June 2011

Location prediction: nesting sites

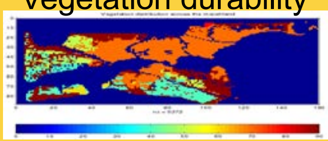
Nest locations



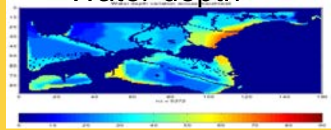
Distance to open water



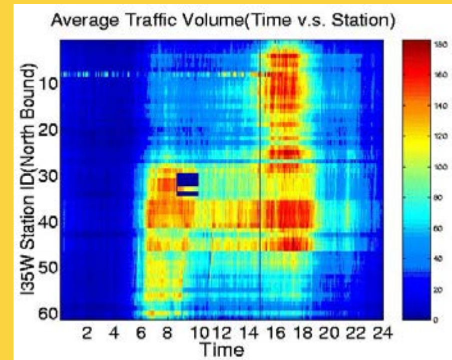
Vegetation durability



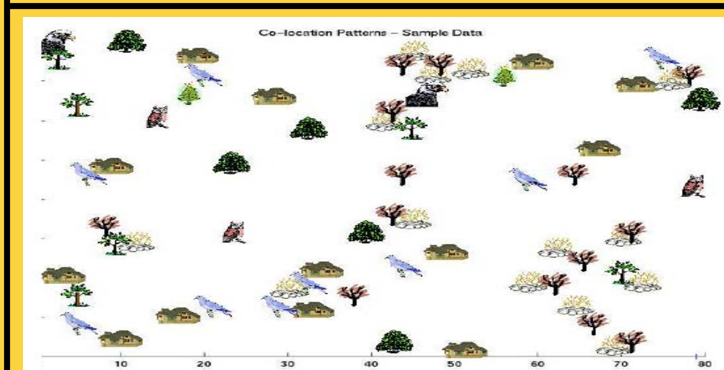
Water depth



Spatial outliers: sensor (#9) on I-35



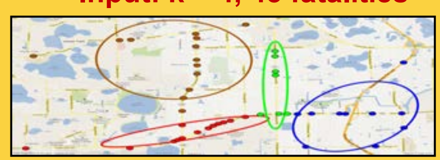
Co-location Patterns



Spatial Network Activity Summarization

Input: k = 4, 43 fatalities

Euclidean Distance



Network Distance

KMR

Outline

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

Why Data Mining?

- <u>Holy Grail</u> - <u>Informed</u> 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
  - …

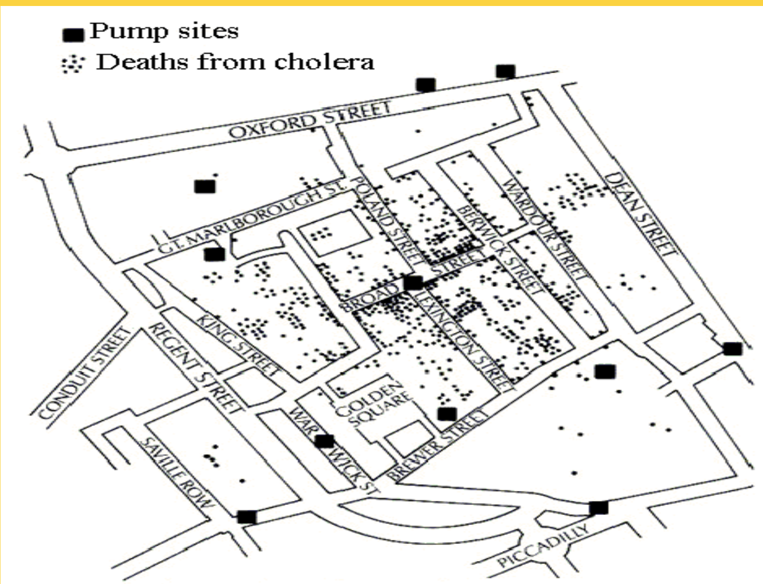
Spatial Data Mining (SDM)

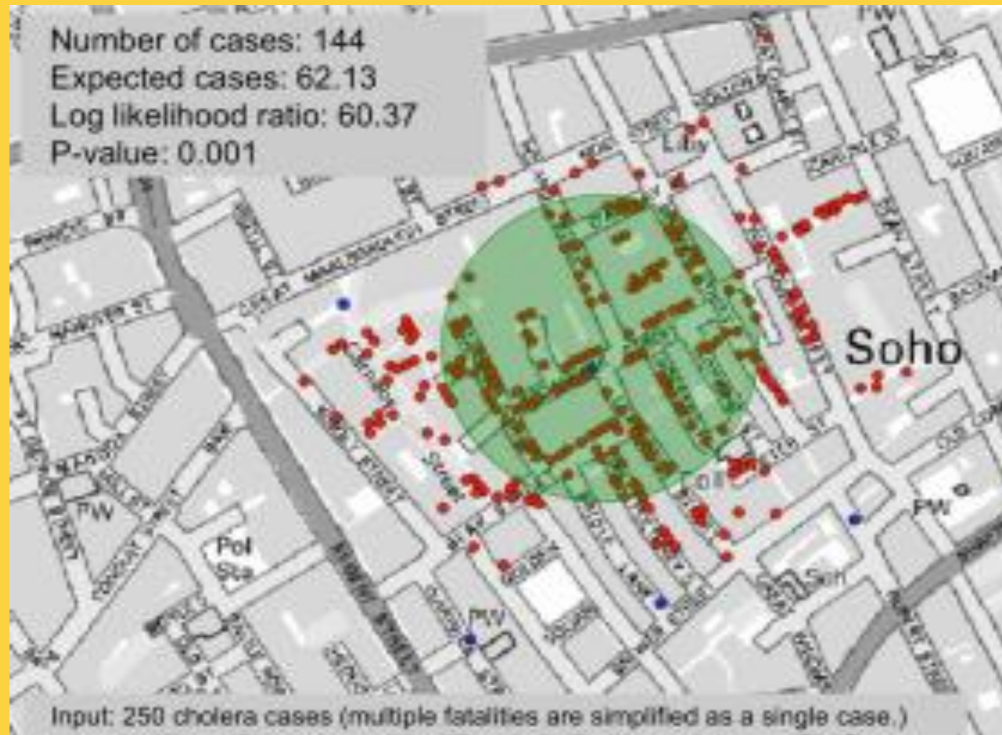
- The process of discovering
  - interesting, useful, non-trivial patterns
    - patterns: non-specialist
    - exception to patterns: specialist
  - from large spatial datasets

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

Pattern Family 1: Hotspots, Spatial Cluster

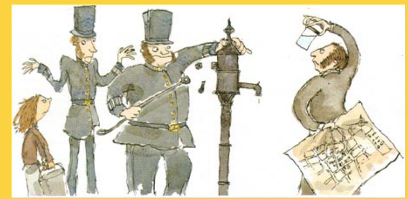
- The 1854 Asiatic Cholera in London
  - Near Broad St. water pump except a brewery



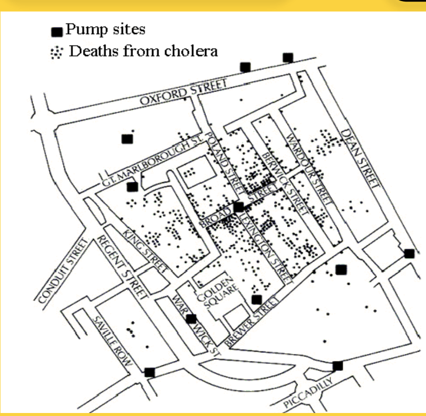


Impact of Snow's Map

<u>1854: What causes Cholera?</u>



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

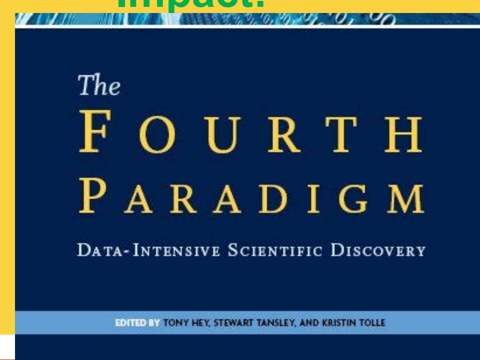


? water pump

Remove pump handle

Germ Theory

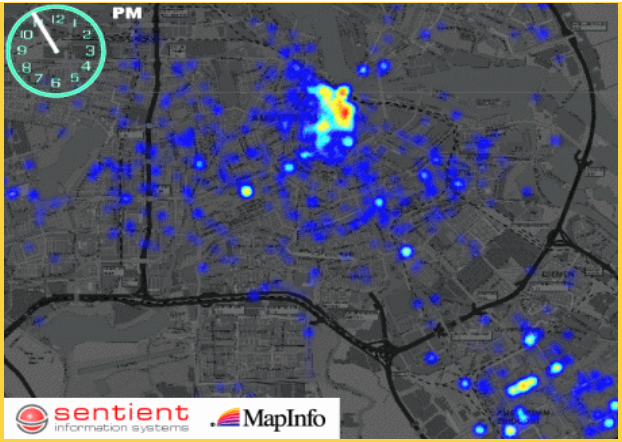
Impact:

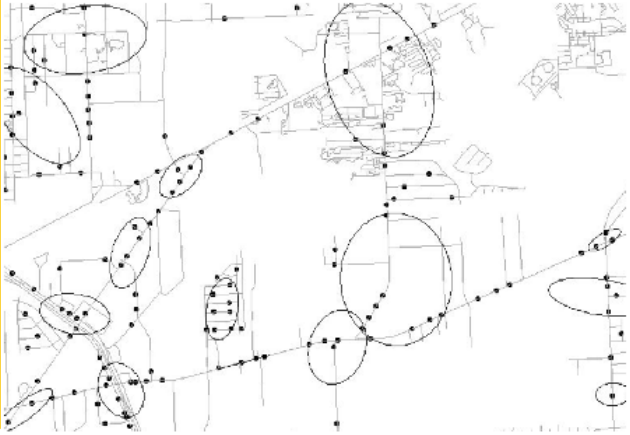


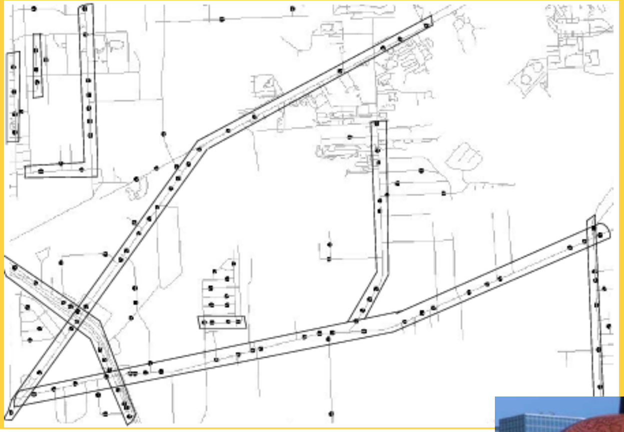
Q? What are Choleras of today?
Q? How may Spatial Data Sc. Help?

Complicated Hotspots

- Complication Dimensions
  - Time
  - Spatial Networks
- Challenges: Trade-off b/w
  - Semantic richness and
  - Scalable algorithms

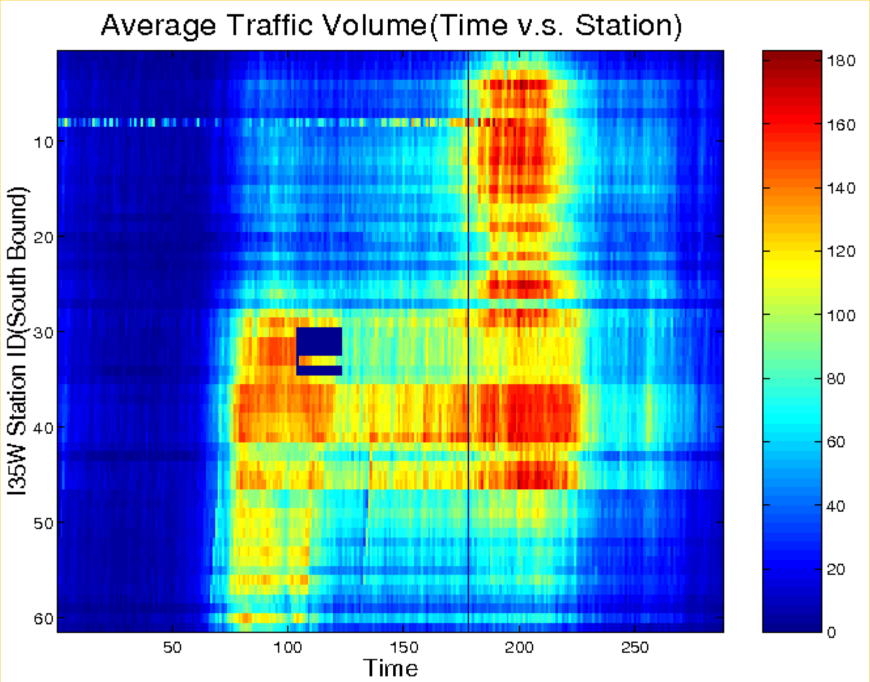






Pattern Family 2: Spatial Outliers

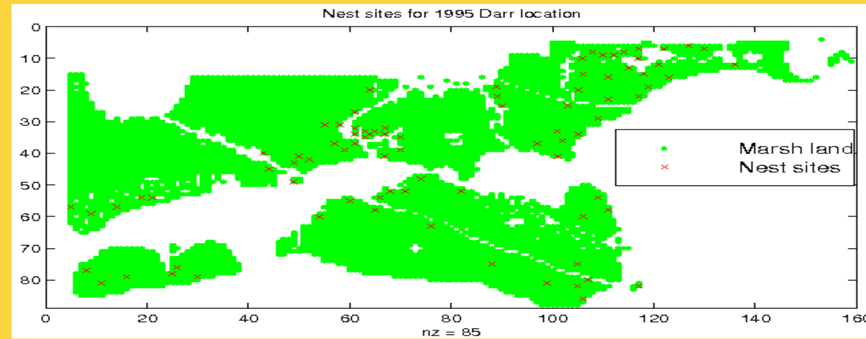
- Spatial Outliers, Anomalies, Discontinuities
  - Traffic Data in Twin Cities
  - Abnormal Sensor Detections
  - Spatial and Temporal Outliers

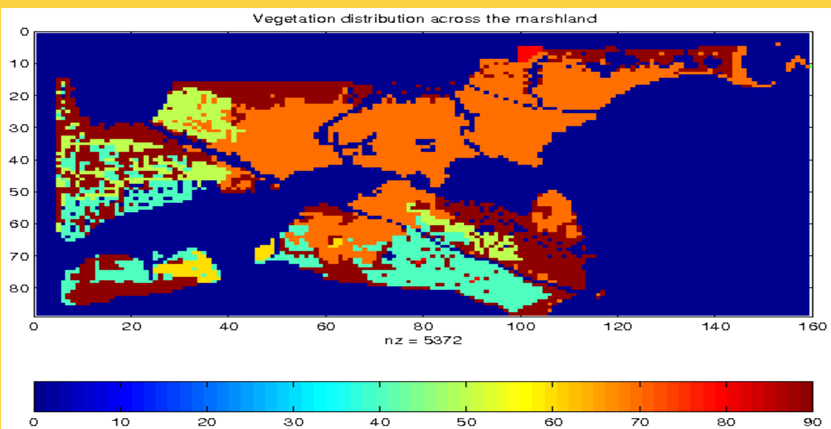


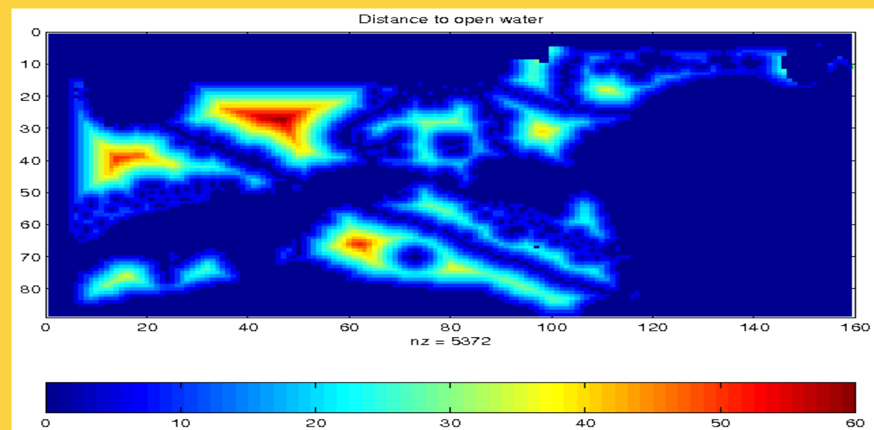
Source: A Unified Approach to Detecting Spatial Outliers, GeoInformatica, 7(2), Springer, June 2003.
(A Summary in Proc. ACM SIGKDD 2001) with C.-T. Lu, P. Zhang.

Pattern Family 3: Predictive Models

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



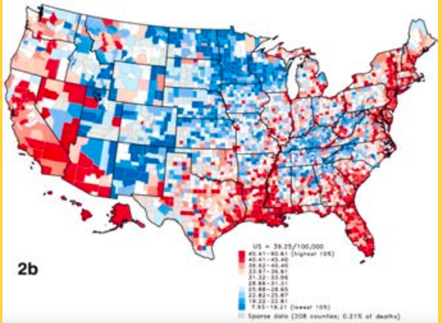




Details: Spatial Contextual Classification and Prediction Models for Mining Geospatial Data, S. Shekhar et al., IEEE Transactions on Multimedia, 4(2):174 - 188. 10.1109/TMM.2002.1017732.

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
- …

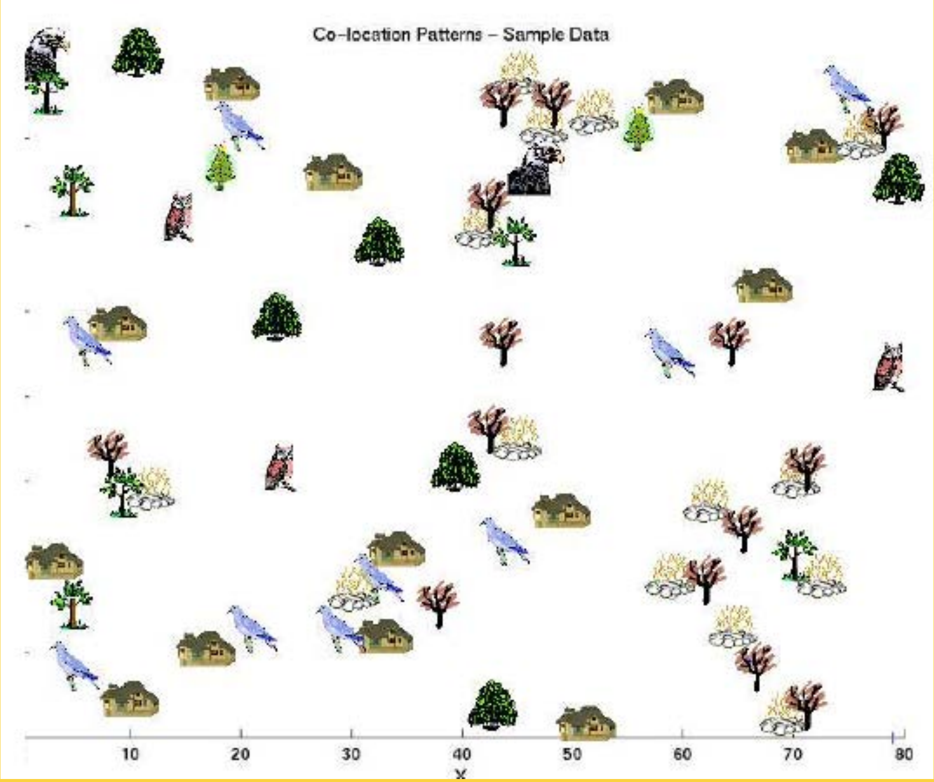
Sources: A. Jemal et al., "Recent Geographic Patterns of Lung Cancer and Mesothelioma Mortality Rates in 49 Shipyard Counties in the U.S., 1970-94", Am J. Ind. Med. 2000, 37(5):512-21.

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



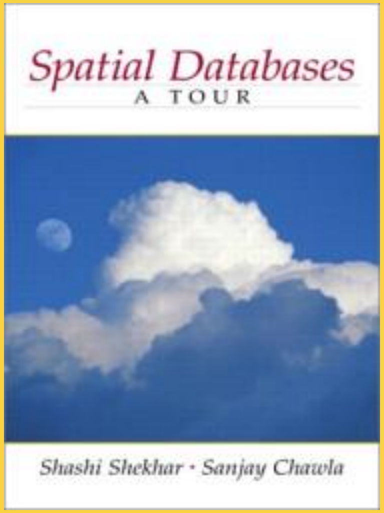
Source: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

What’s NOT Spatial Data Mining (SDM)

- 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



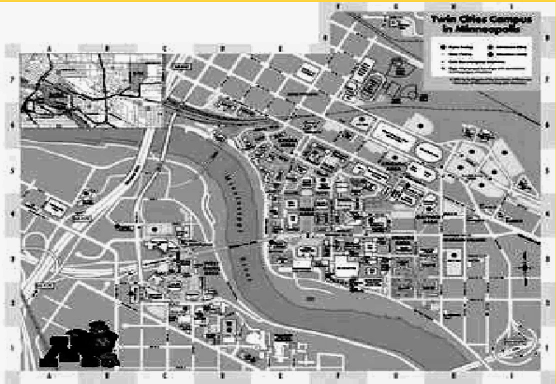
Data-Types: Non-Spatial vs. Spatial

- 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)

Relationships: Non-spatial vs. Spatial

- Non-spatial Relationships
  - Explicitly stored in a database
  - Ex. New Delhi is the capital of India

- Spatial Relationships
  - Implicit, computed on demand
  - Topological: meet, within, overlap, …
  - Directional: North, NE, left, above, behind, …
  - Metric: distance, area, perimeter
  - Focal: slope
  - Zonal: highest point in a country
  - …

OGC Simple Features

- Open GIS Consortium: Simple Feature Types
  - Vector data types: e.g. point, line, polygons
  - Spatial operations :

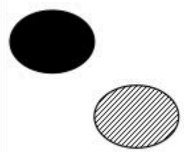
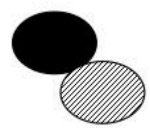
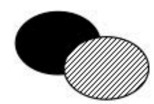
Operator Type	Operator Name
Basic Function	SpatialReference, Envelope, Boundary, Export, IsEmpty, IsSimple
Topological/Set Operations	Equal, Disjoint, Intersect, Touch, Cross, Within, Contains, Overlap
Spatial Analysis	Distance, Buffer, ConvexHull, Intersection, Union, Difference, SymmDiff

Examples of Operations in OGC Model

OGIS - Topological Operations

- Topology: 9-intersections
  - interior
  - boundary
  - exterior

$$
\begin{array}{cccc}
\text{Interior(B)} & \text{Boundary(B)} & \text{Exterior(B)} & \\
\left(\begin{array}{ccc}
(A^{o} \cap B^{o}) & (A^{o} \cap \partial B) & (A^{o} \cap B^{-}) \\
(\partial A \cap B^{o}) & (\partial A \cap \partial B) & (\partial A \cap B^{-}) \\
(A^{-} \cap B^{o}) & (A^{-} \cap \partial B) & (A^{-} \cap B^{-})
\end{array}\right) & & & \begin{array}{l}\text{Interior(A)} \\ \text{Boundary(A)} \\ \text{Exterior(A)}\end{array}
\end{array}
$$

Topological Relationship	 disjoint	 meet	 overlap	 equal
9-intersection model	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Research Needs for Data

- Limitations of OGC Model
  - Direction predicates - e.g. absolute, ego-centric
  - 3D 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

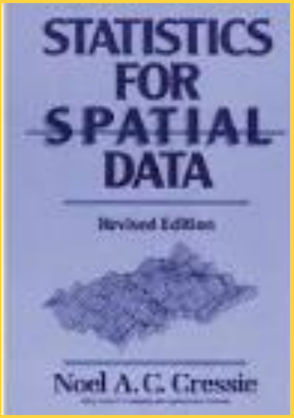
- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
    - Spatial Auto-correlation
    - Heterogeneity
    - Edge Effect
- Spatial Data Mining
- Conclusions

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 Heterogeneity
  - Spatial data samples may not be identically distributed!
  - No two places on Earth are exactly alike!
- …

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

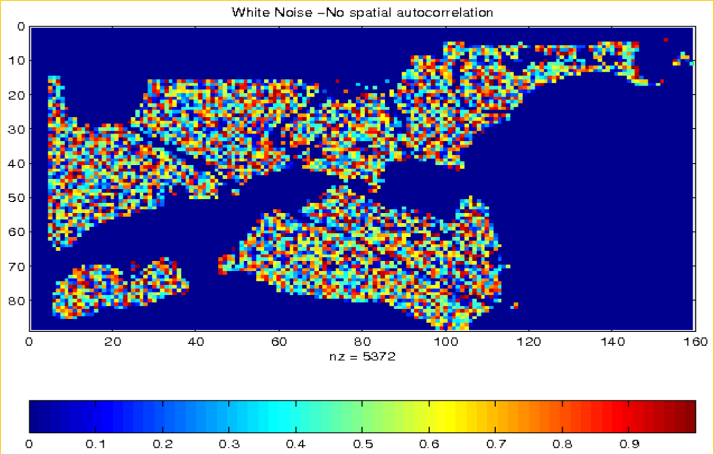


- Geostatistics
  - Continuous phenomena, e.g., rainfall, snow depth, …
  - Methods: Variogram measure how similarity decreases with distance
  - Spatial interpolation, e.g., Kriging

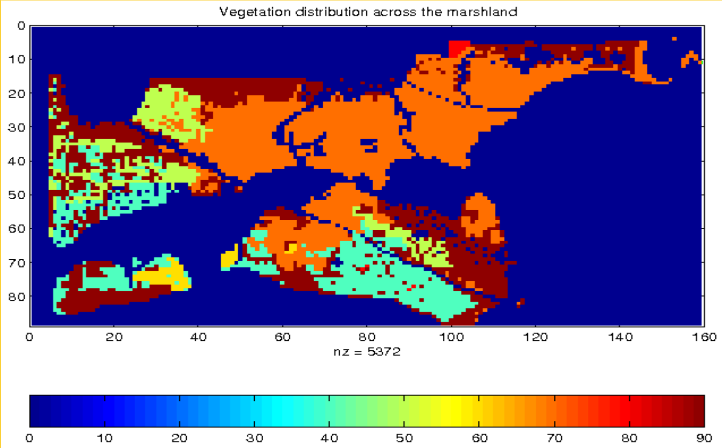
- Lattice-based statistics
  - Polygonal aggregate data, e.g., census, disease rates, pixels in a raster
  - Spatial Gaussian models, Markov Random Fields, Spatial Autoregressive Model

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

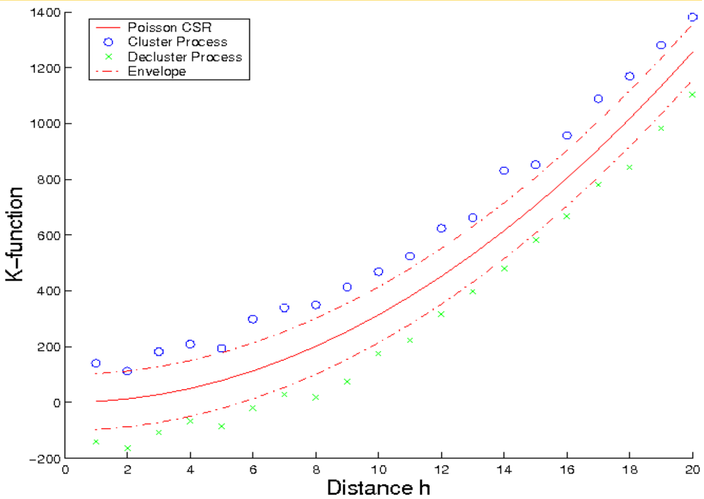
- **Purpose:** Compare a point dataset with a complete spatial random (CSR) data
- **Input:** A set of points

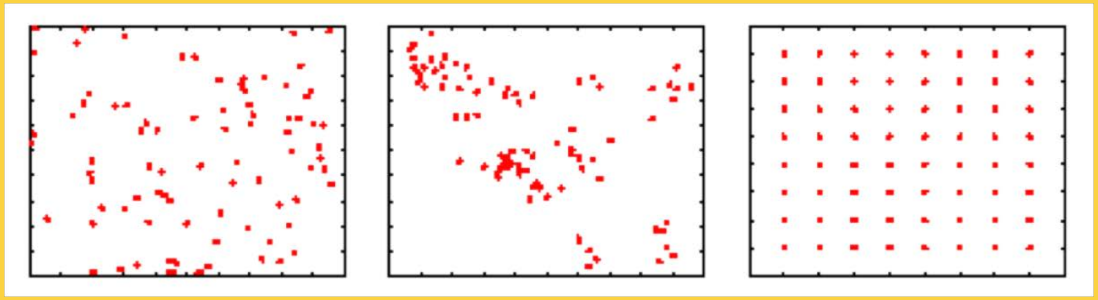
$$K(h, data) = \lambda^{-1} E[\text{number of events within distance } h \text{ of an arbitrary event}]$$

  - 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

Cross-Correlation

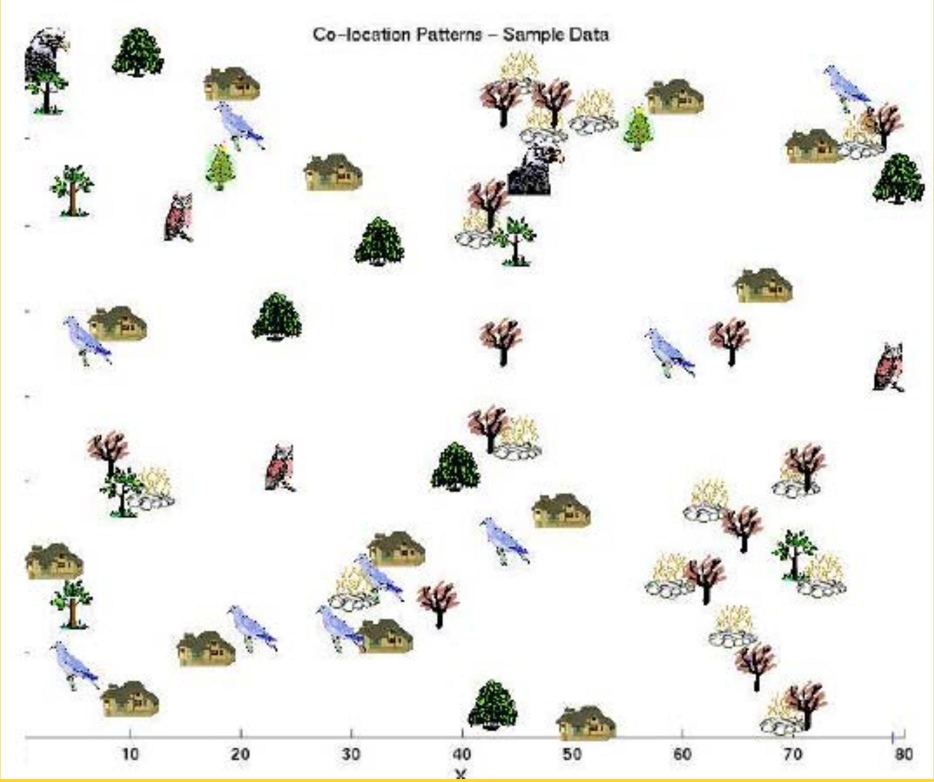
- Cross K-Function Definition

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

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

Recall Pattern Family 4: Co-locations

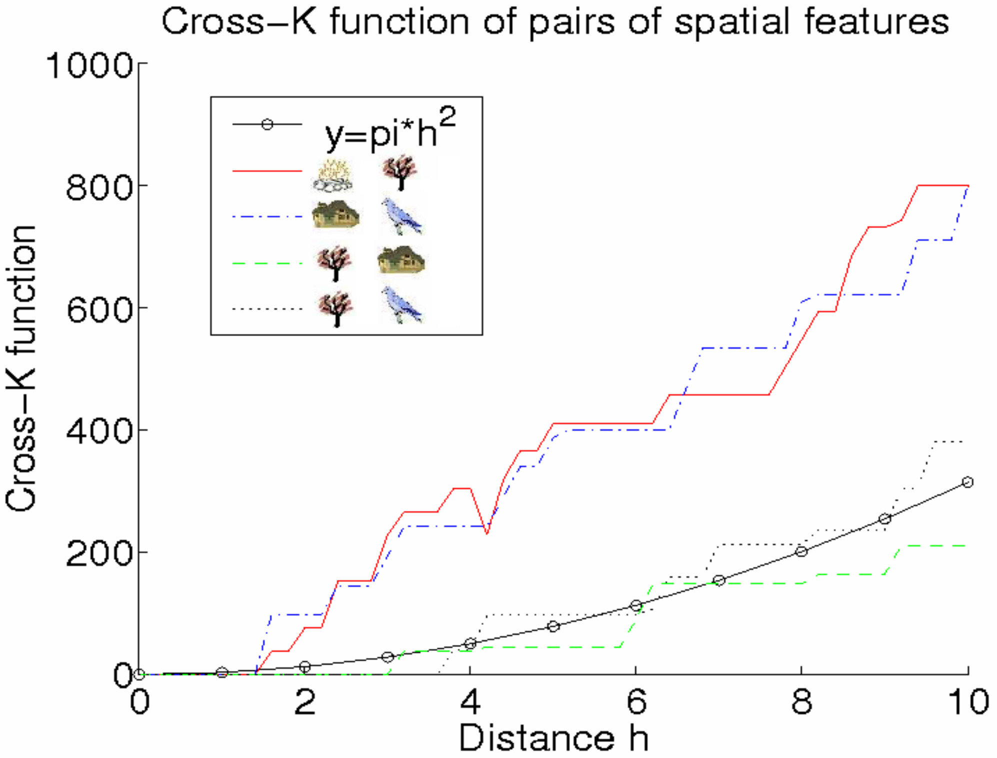
- Given: A collection of different types of spatial events
- Find: Co-located subsets of event types



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

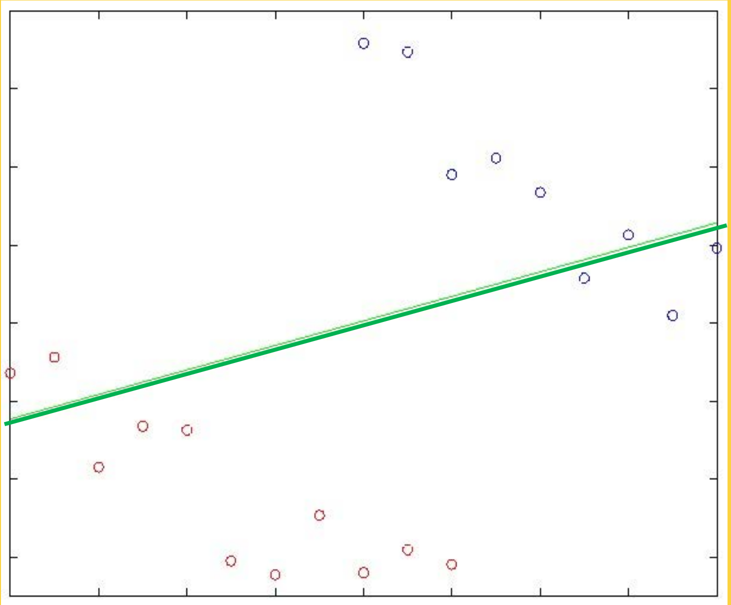
- Illustration of Cross K-function for Example Data

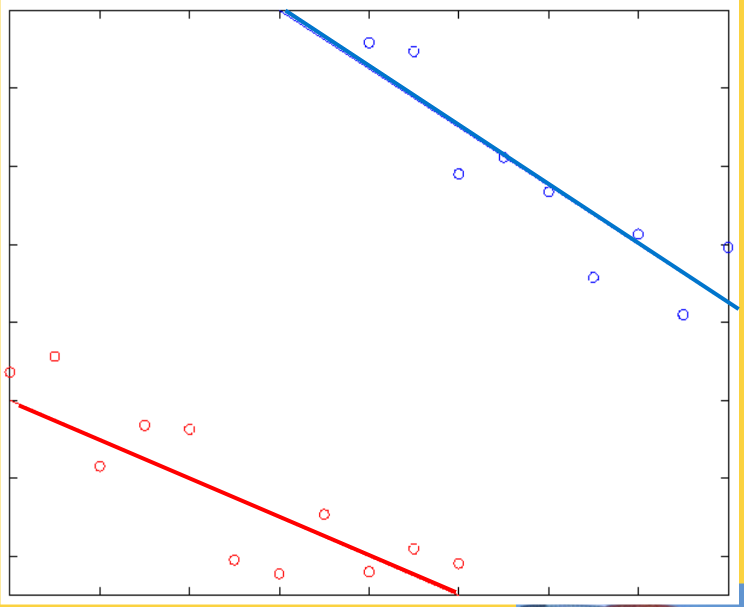


Cross-K Function for Example Data

Spatial Heterogeneity

- “Second law of geography” [M. Goodchild, UCGIS 2003]
- Global model might be inconsistent with regional models
  - Spatial Simpson’s Paradox
- May improve the effectiveness of SDM, show support regions of a pattern





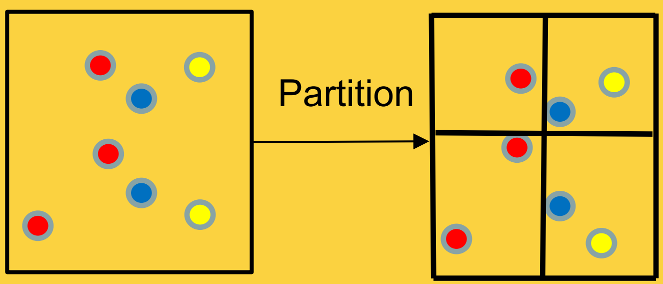
Spatial Heterogeneity: Gerrymandering

Gerrymandering, a Tradition as Old as the Republic, Faces a Reckoning

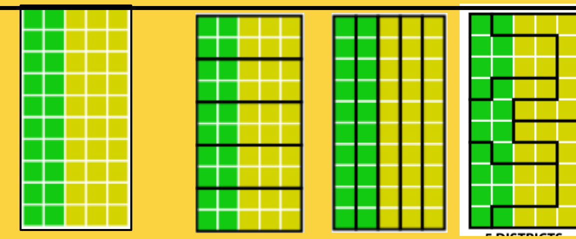
Supreme Court to hear arguments on whether contorted voting maps drawn by both parties to cement power have finally gone too far

THE WALL STREET JOURNAL.

- Space partitioning affects statistical results!
  - Gerrymandering Elections
  - Gini-Index, Entropy
  - Associations & correlations
  - Modifiable Areal Unit Problem (MAUP)



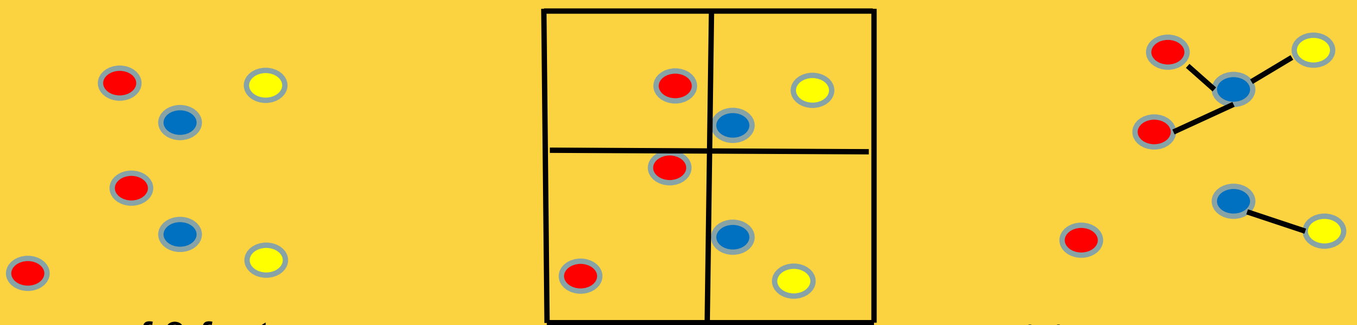
	Partition Based Pearson's Correlation
-	
-	



Election Results		0 - 5	2 - 3	3 - 2
Gini-Index	0.47	0.47	0	0.36
Entropy	0.97	0.97	0	0.77

Neighbor Relationship vs. Space Partitioning

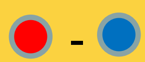
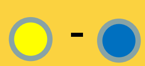
- Neighbor relationship graph
  - Honors continuity of geographic space methods
  - Partitions miss spatial interactions



(a) a map of 3 features

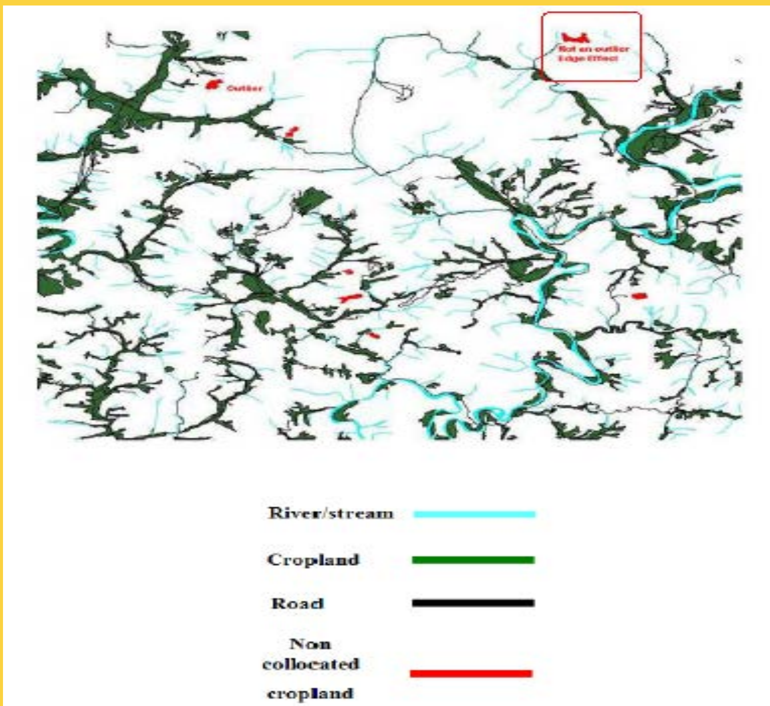
(b) Spatial Partitions

(c) Neighbor graph

	Pearson's Correlation (Partition based)	Ripley's cross-K
	-0.90	
	1	

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.

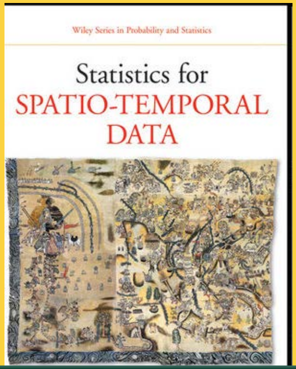
Research Challenges of Spatial Statistics

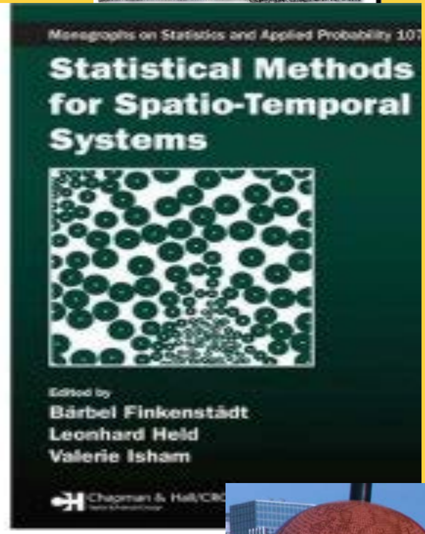
- State-of-the-art of Spatial Statistics

		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

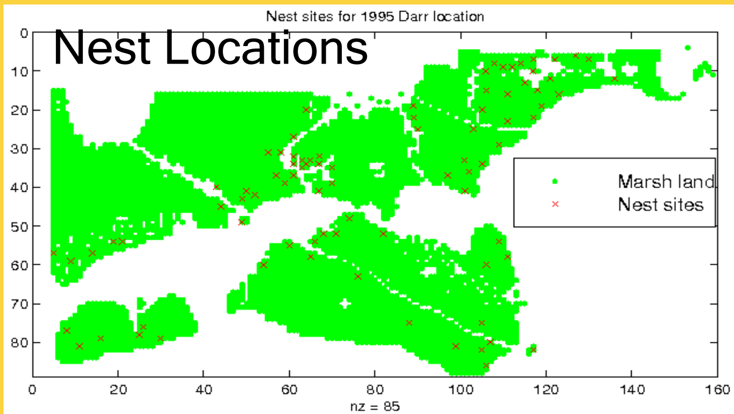


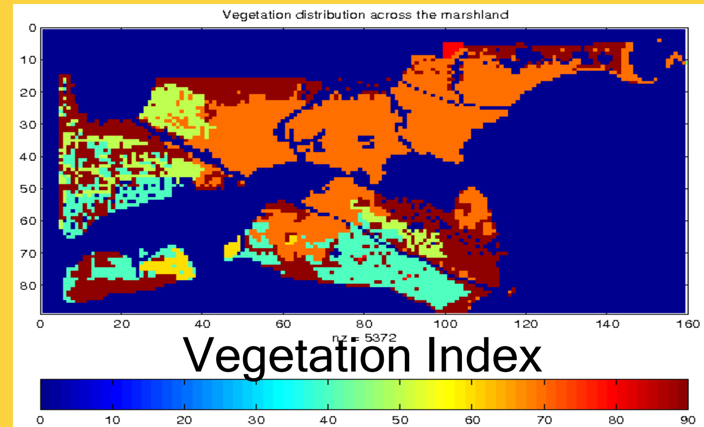


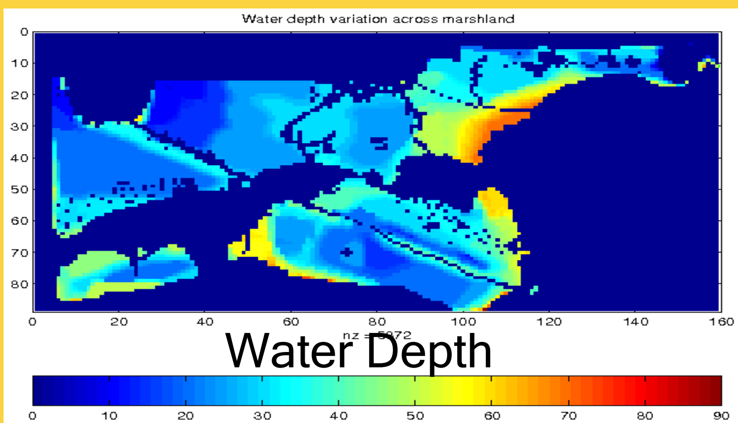
Outline

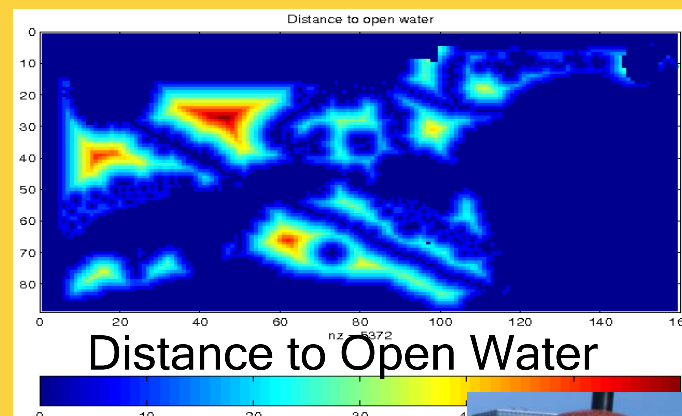
- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
  - Location Prediction
  - Hotspots
  - Spatial Outliers
  - Colocations
- Conclusions

Illustration of Location Prediction Problem

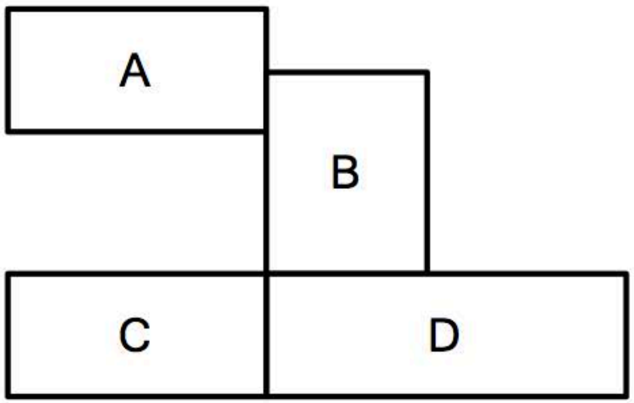








Neighbor Relationship: W Matrix



(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

Location Prediction Models

- Traditional Models, e.g., Regression (with Logit or Probit),
  - Bayes Classifier, …
- Spatial Models
  - Spatial autoregressive model (SAR)
  - Markov random field (MRF) based Bayesian Classifier

Classical	Spatial
$y = X\beta + \varepsilon$	$y = \rho W y + X\beta + \varepsilon$
$\Pr(C_i \mid X) = \dfrac{\Pr(X \mid C_i)\Pr(C_i)}{\Pr(X)}$	$\Pr(c_i \mid X, C_N) = \dfrac{\Pr(C_i)\Pr(X, C_N \mid c_i)}{\Pr(X, C_N)}$

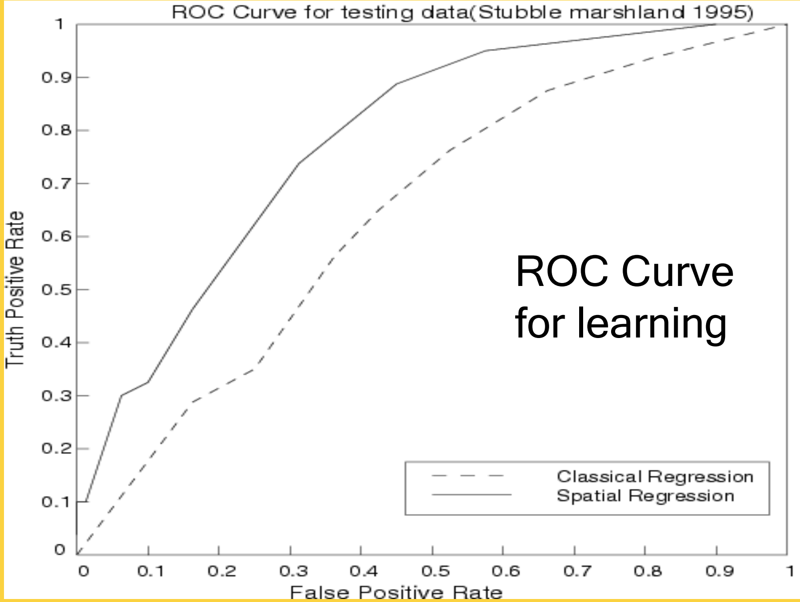
Location Prediction Models

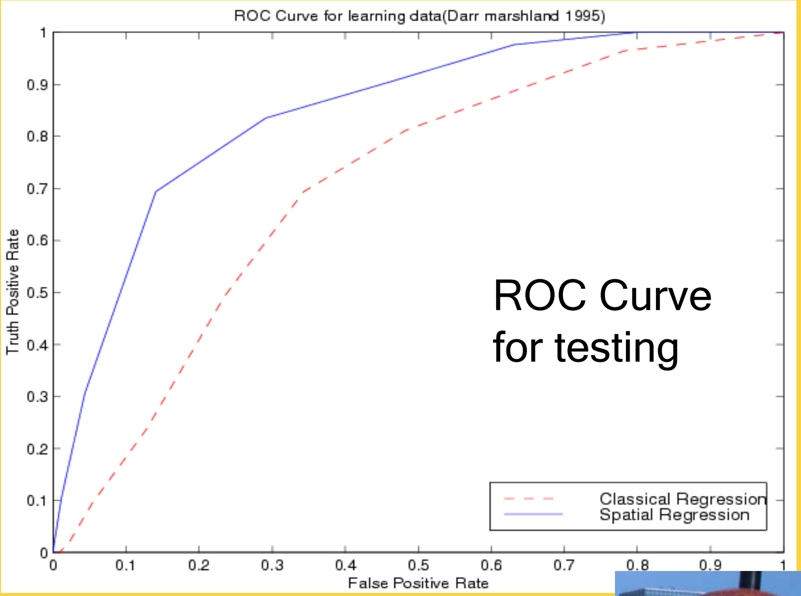
- Traditional Models, e.g., Regression (with Logit or Probit),
  - Linear Regression, Bayes Classifier, …
- Semi-Spatial : auto-correlation regularizer $\varepsilon = \|y - \beta X\|^2 + \|\beta X - \beta X_{neighbor}\|^2$
- Spatial Models
  - Spatial autoregressive model (SAR)
  - Markov random field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

Traditional	Spatial
$y = X\beta + \varepsilon$	$y = \rho W y + X\beta + \varepsilon$
$\Pr(C_i \mid X) = \frac{\Pr(X \mid C_i)\Pr(C_i)}{\Pr(X)}$	$\Pr(c_i \mid X, C_N) = \frac{\Pr(C_i)\Pr(X, C_N \mid c_i)}{\Pr(X, C_N)}$
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

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

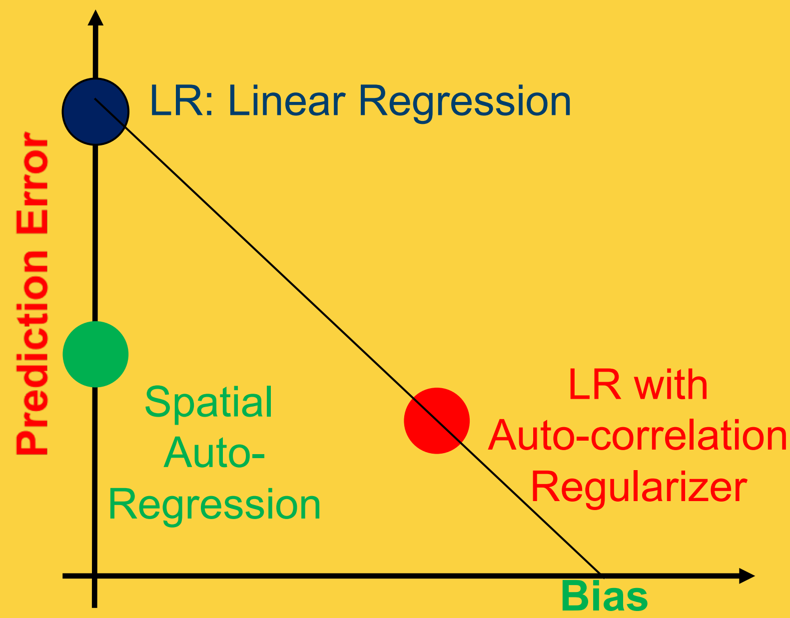
- LR with Auto-correlation Regularizer
  - Least squares estimator

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

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

- 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 Decision Tree

Input:

(a) aerial photo (b) aerial photo (c) true classes

Output:

DT: decision tree
SDT: spatial decision tree

Training samples: upper half
Test samples: lower half
Spatial neighborhood: maximum 11 pixels by11 pixels

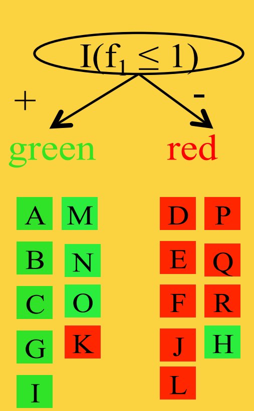
Details: Focal-Test-Based Spatial Decision Tree Learning. IEEE Trans. Knowl. Data Eng. 27(6): 1547-1559, 2015 (summary in Proc. IEEE Intl. Conf. on Data Mining, 2013).(w/ Z. Jiang et al.)

Spatial Decision Tree

Inputs: table of records

ID	f_1	f_2	Γ_1	class
A	1	1	1	green
B	1	1	0.3	green
C	1	3	0.3	green
G	1	1	0.3	green
I	1	3	0	green
K	1	2	-1	red
M	1	1	1	green
N	1	1	0.3	green
O	1	3	0.3	green
D	3	2	0.3	red
E	3	2	0.3	red
F	3	2	1	red
H	3	1	-1	green
J	3	2	0	red
L	3	2	0.3	red
P	3	2	0.3	red
Q	3	2	0.3	red
R	3	2	1	red

feature test	information gain
$f_1 \leq 1$	0.50
$f_2 \leq 1$	0.46
$f_2 \leq 2$	0.19



Predicted map

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R

Inputs:

- feature maps, class map
- Rook neighborhood

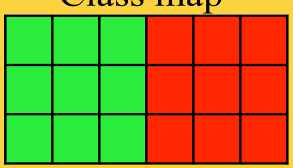
Feature f_1

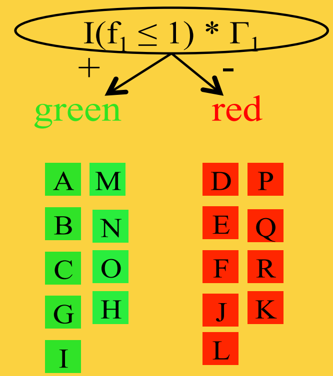
1	1	1	3	3	3
1	3	1	3	1	3
1	1	1	3	3	3

Feature f_2

1	1	3	2	2	2
1	1	3	2	2	2
1	1	3	2	2	2

Class map





Focal function Γ_1

1	.3	.3	.3	.3	1
.3	-1	0	0	-1	.3
1	.3	.3	.3	.3	1

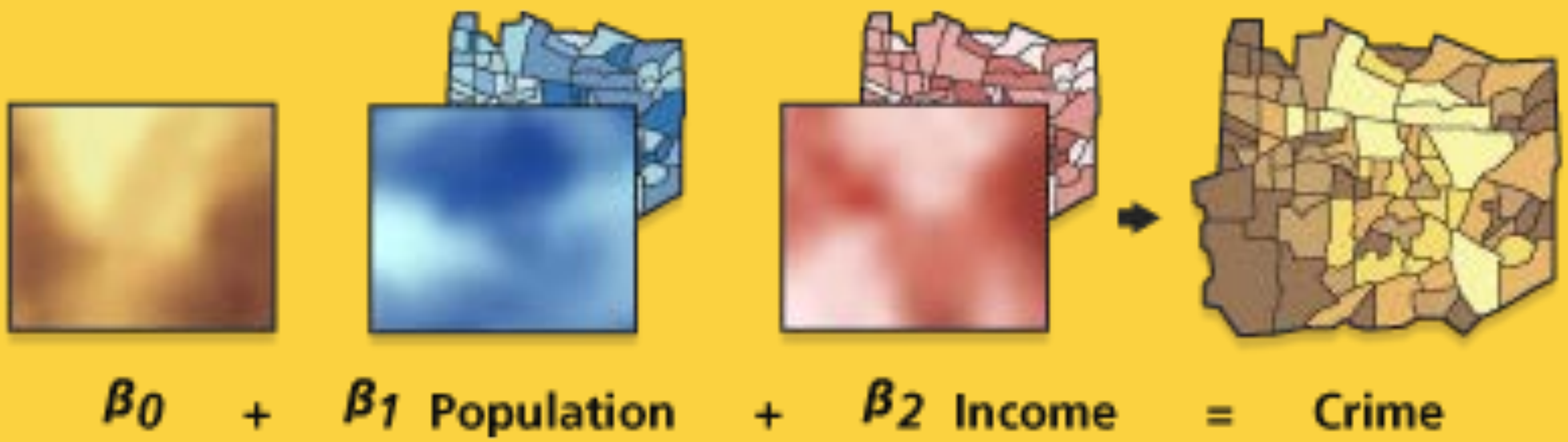
Predicted map

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R

Modeling Spatial Heterogeneity: GWR

- Geographically Weighted Regression (GWR)
  - Goal: Model spatially varying relationships
  - Example: $y = X\beta' + \varepsilon'$

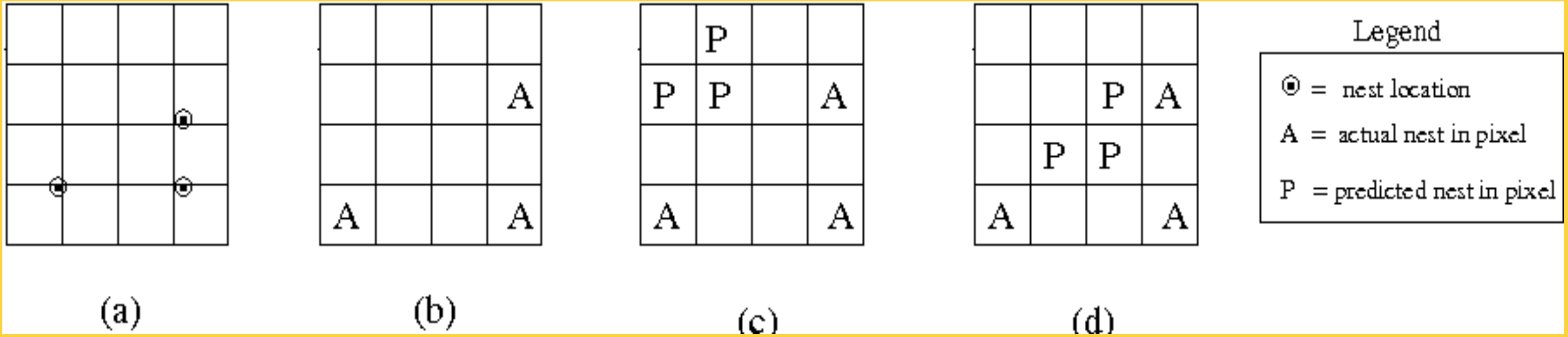
    Where β' and ε' are location dependent



Source: resources.arcgis.com

Research Needs for Location 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

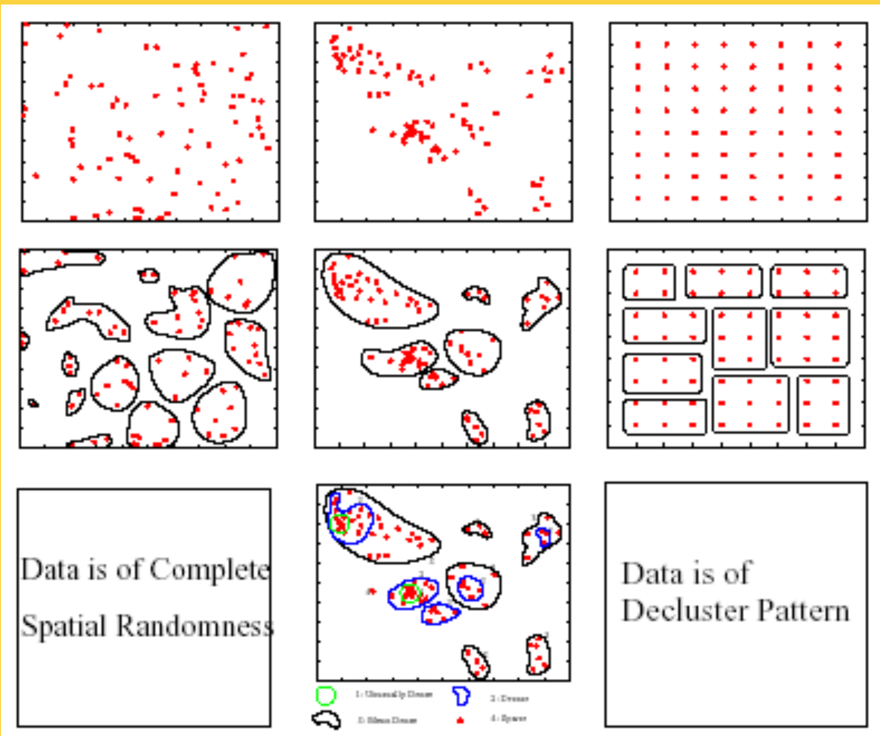
Prediction 2.
Spatially more interesting than Prediction 1

Outline

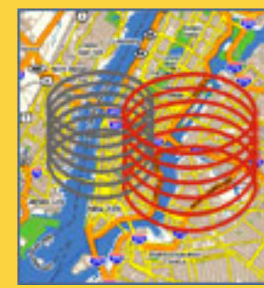
- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
  - Location Prediction
  - Hotspots
  - Spatial Outliers
  - Colocations
- Conclusions

Limitations of K-Means

- K-Means does test Statistical Significance
  - Finds chance clusters in complete spatial randomness (CSR)



Classical Clustering



Spatial Clustering

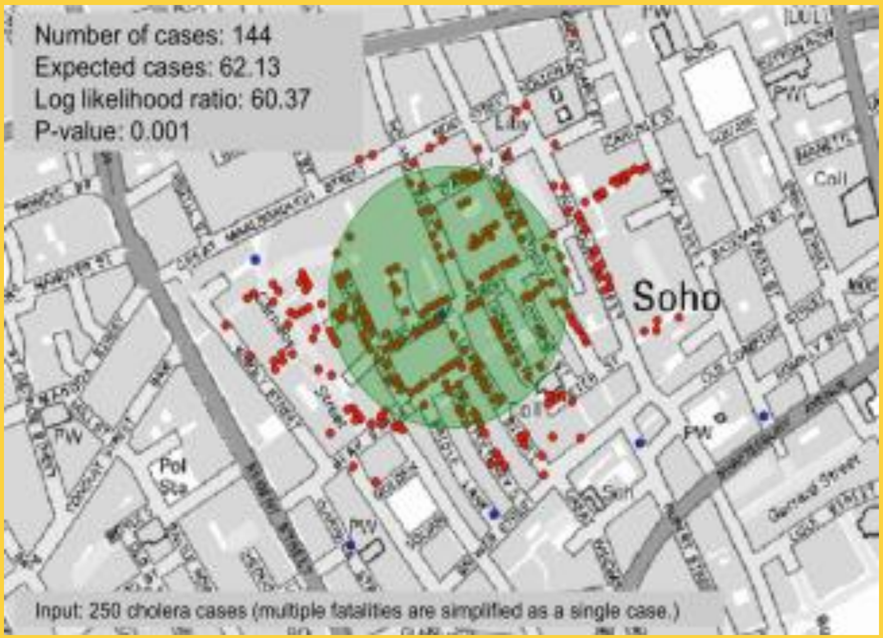
SaTScan™
Software for the spatial, temporal, and space-time scan statistics

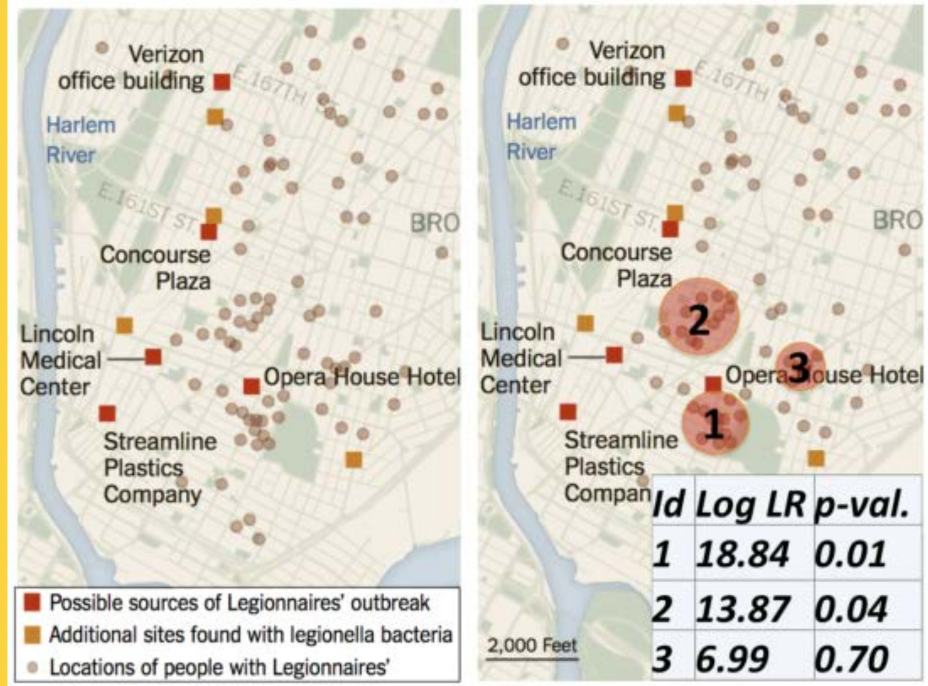
Spatial Scan Statistics (SatScan)

- 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 Example

1854 London Cholera, p-value = 0.001
Output: A hotspot!





Id	Log LR	p-val.
1	18.84	0.01
2	13.87	0.04
3	6.99	0.70

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

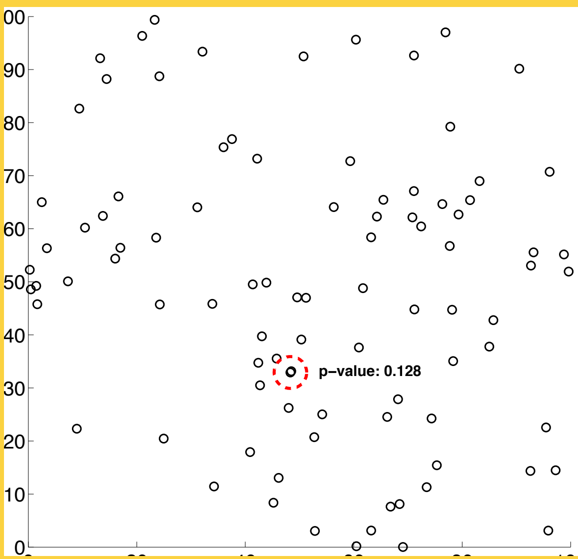
<u>Source: Ring-Shaped Hotspot Detection,</u> IEEE Trans. Know. & Data Eng., 28(12), 2016.
(A Summary in Proc. IEEE ICDM 2014) (w/ E. Eftelioglu et al.)

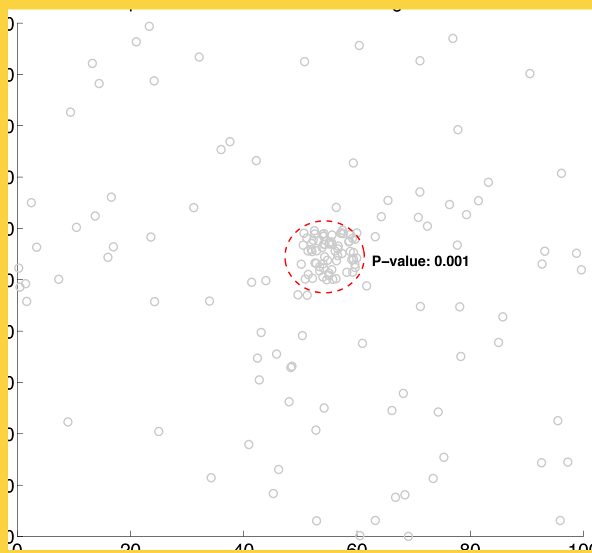
SatScan Examples

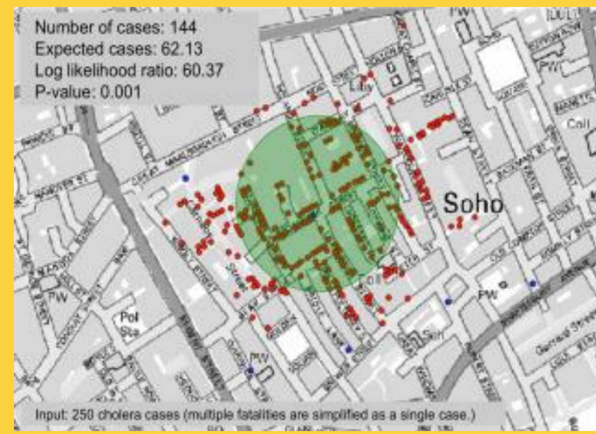
Complete Spatial Randomness
Output: No hotspots !
Highest LR circle p-value = 0.128

Data with a hotspot
Output: A hotspot!
p-value = 0.001

1854 London Cholera
Output: A hotspot!
p-value = 0.001

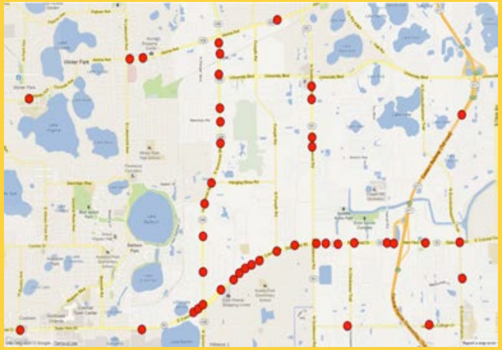




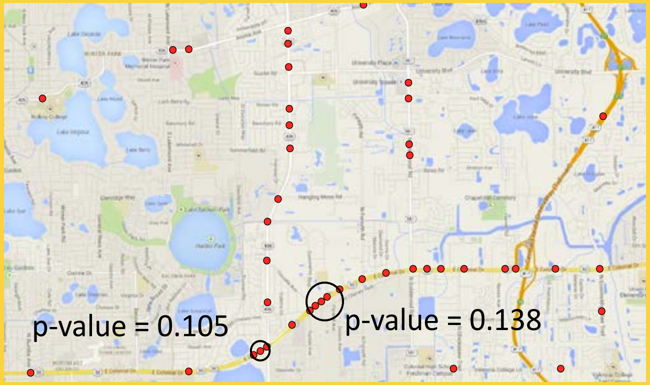


Spatial-Concept/Theory-Aware Clusters

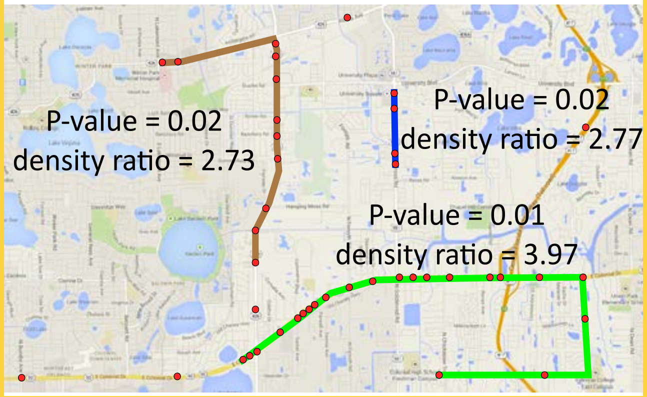
- Geographic features, e.g., rivers, streams, roads, …
  - Hot-spots => Hot Geographic-features, e.g., Linear Hotspots
- Spatial Theories, e.g,, environmental criminology
  - Circles ➔ Doughnut holes



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.

N.Y. / REGION

34 COMMENTS

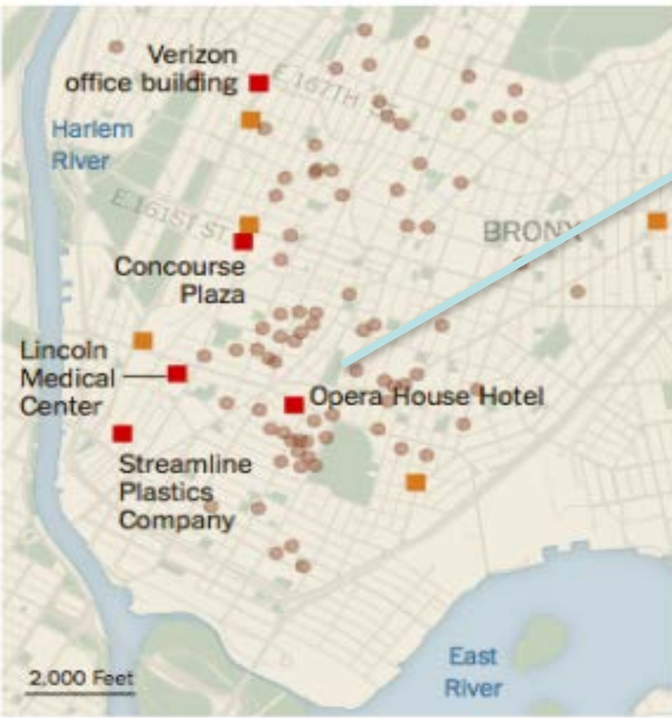
Hotel That Enlivened the Bronx Is Now a 'Hot Spot' for Legionnaires'

By WINNIE HU and NOAH REMNICK AUG. 10, 2015

Contaminated Cooling Towers

Five buildings have been identified as the potential source of the Legionnaires' disease outbreak in the South Bronx.

- Possible sources of Legionnaires' outbreak
- Additional sites found with legionella bacteria
- Locations of people with Legionnaires'



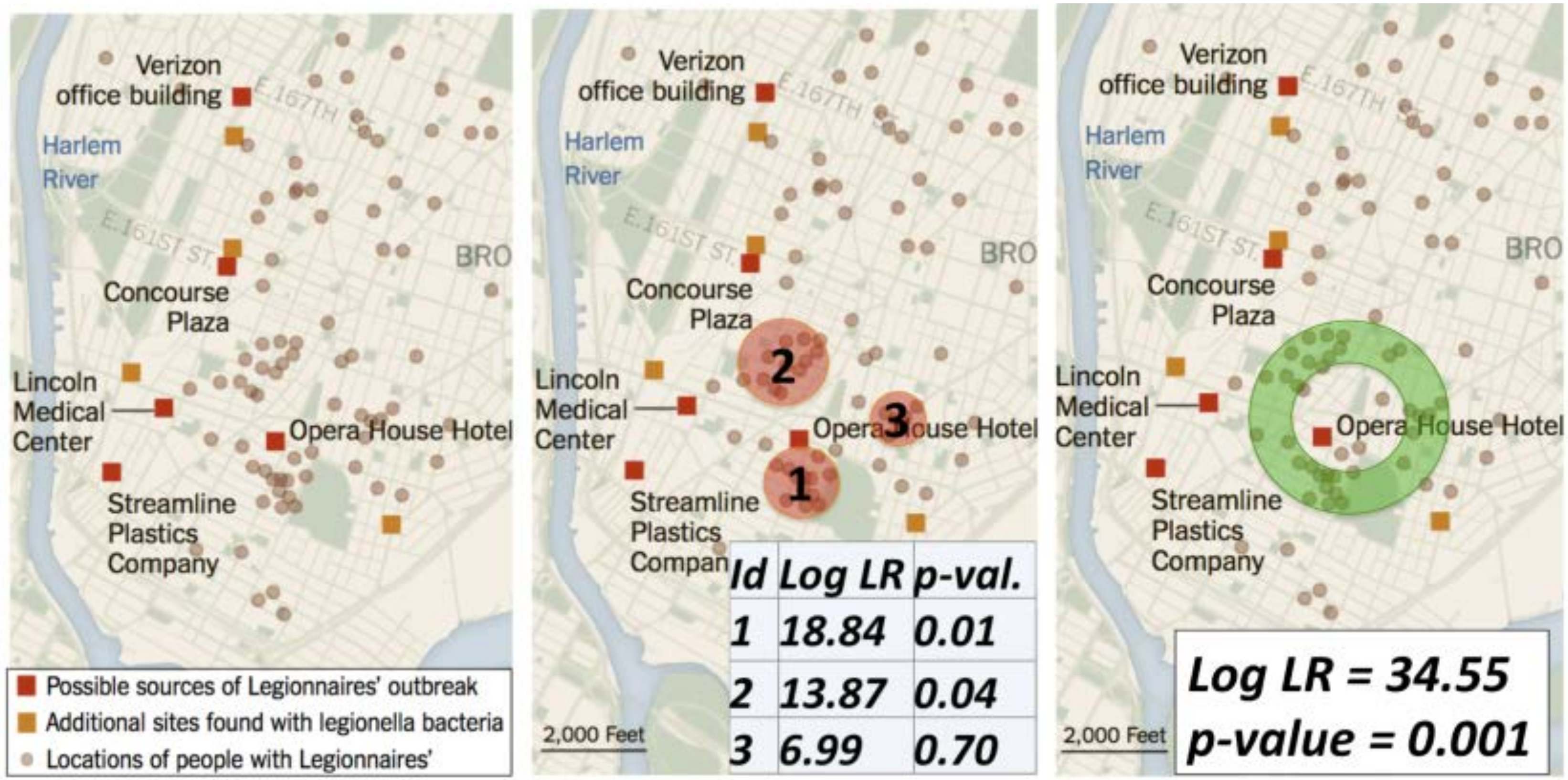
Source: New York Mayor's Office

By The New York Times

The Opera House Hotel is at the center of the outbreak. Edwin J. Torres for The New York Times

Legionnaires' Disease Outbreak in New York

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

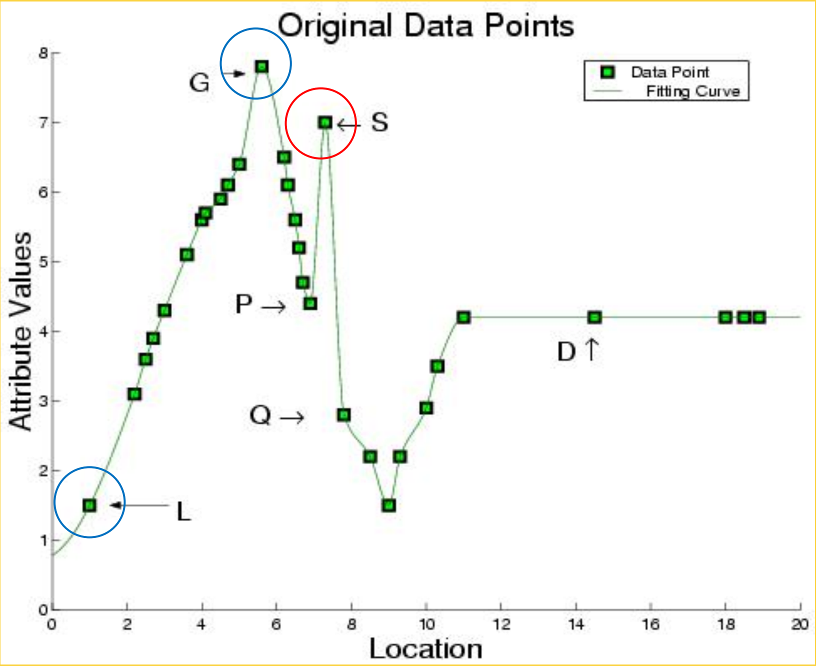


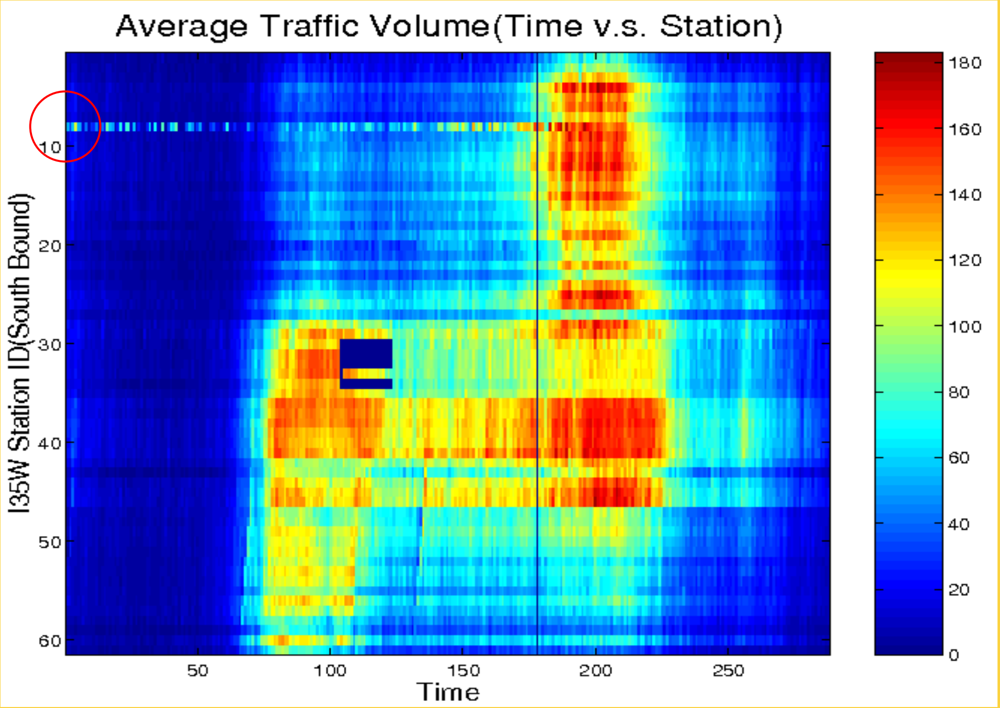
(a) Legionnaire's in New York (2015) (b) Output of SaTScan (c) Output of RHD

Outline

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

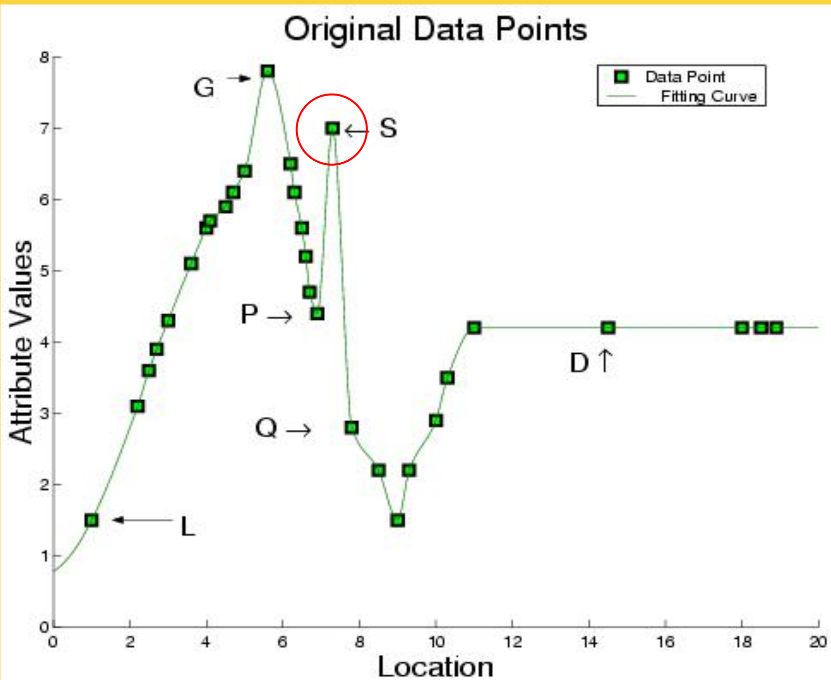
Outliers: Global (G) vs. Spatial (S)

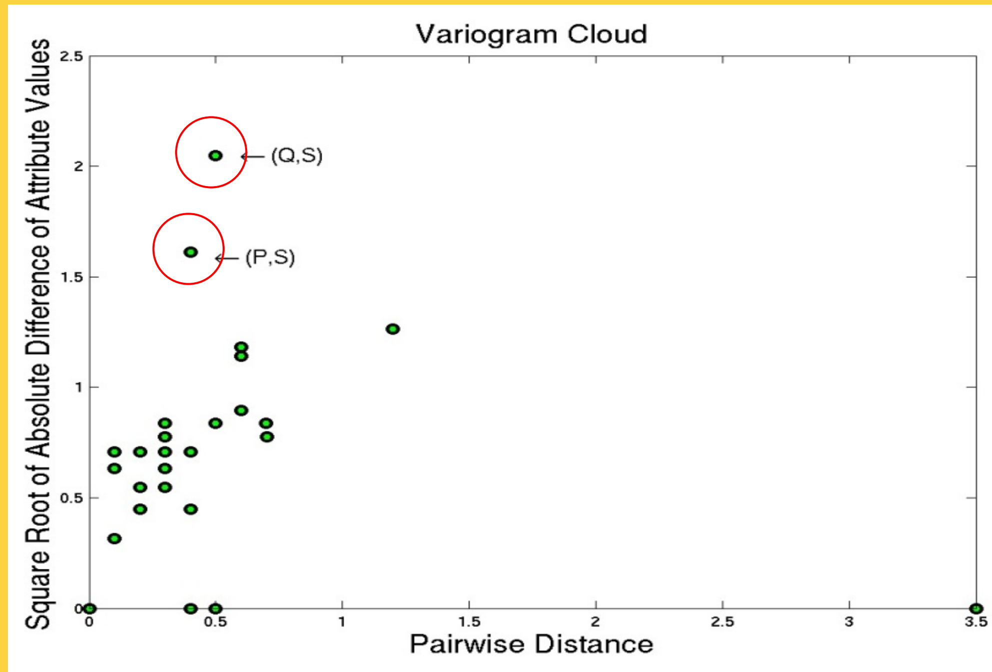




Outlier Detection Tests: Variogram Cloud

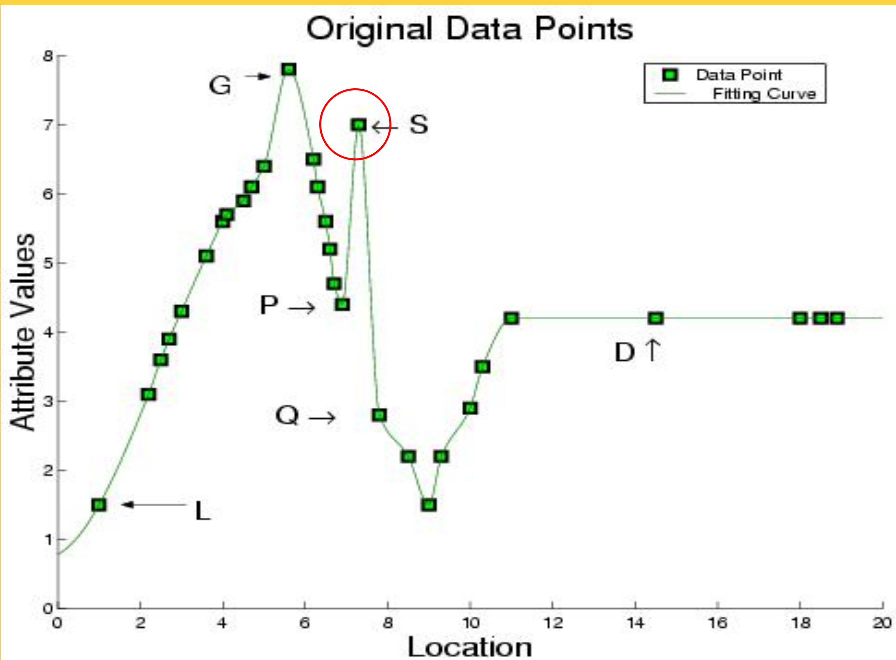
- Graphical Test: Variogram Cloud

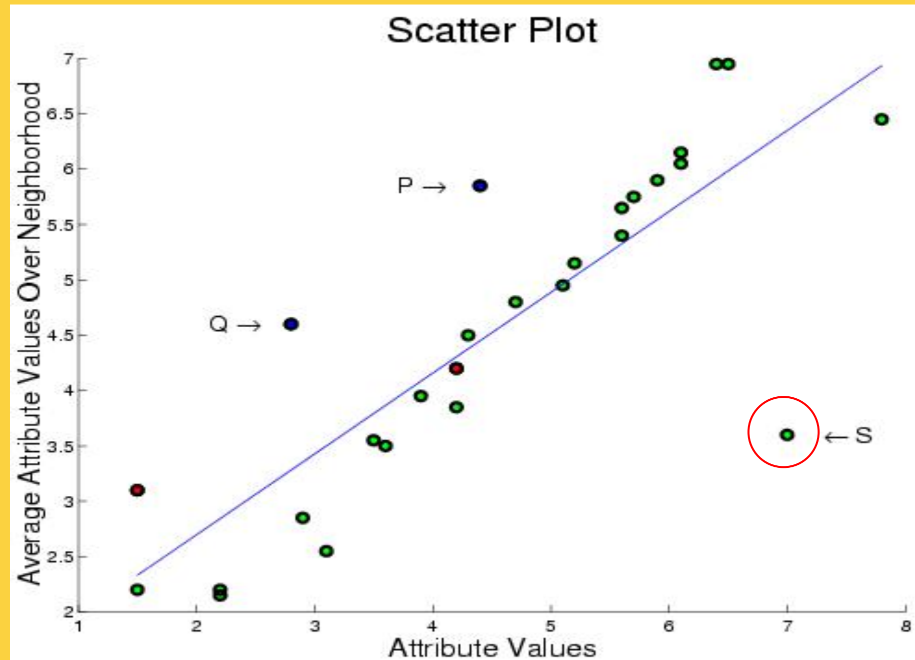




Outlier Detection – Scatterplot

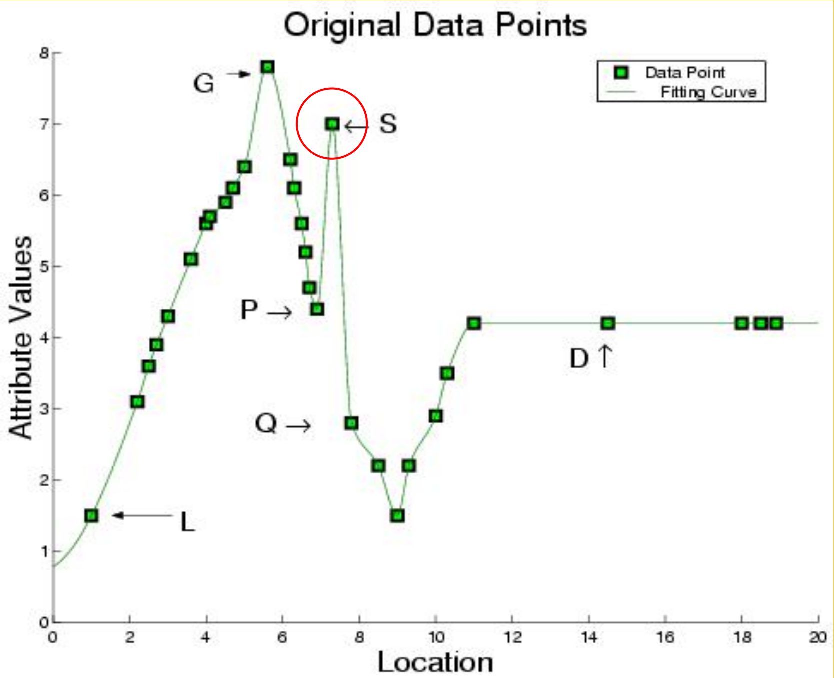
- Quantitative Tests: Scatter Plot

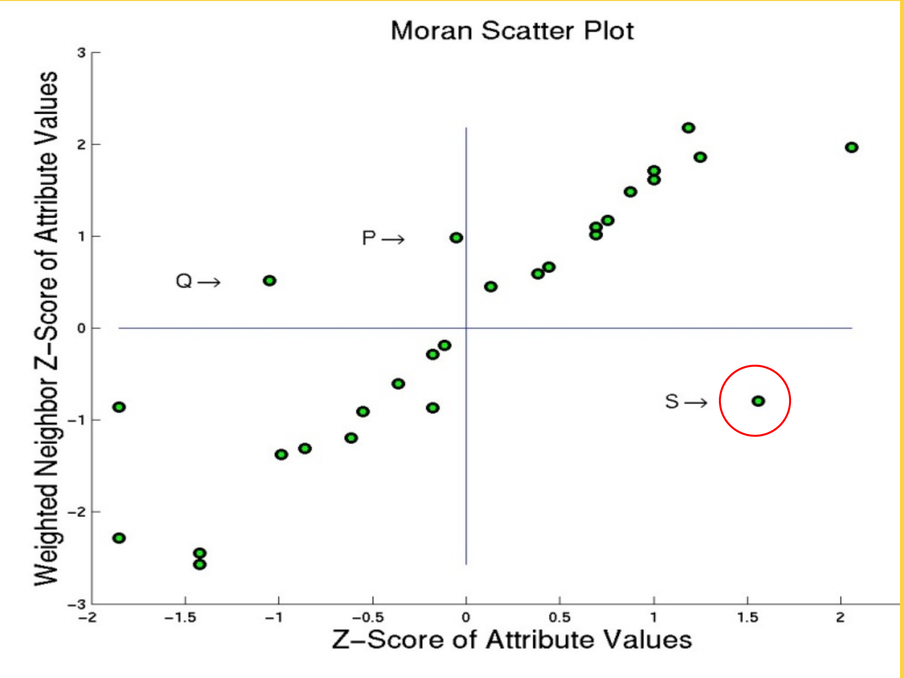




Outlier Detection Test: Moran Scatterplot

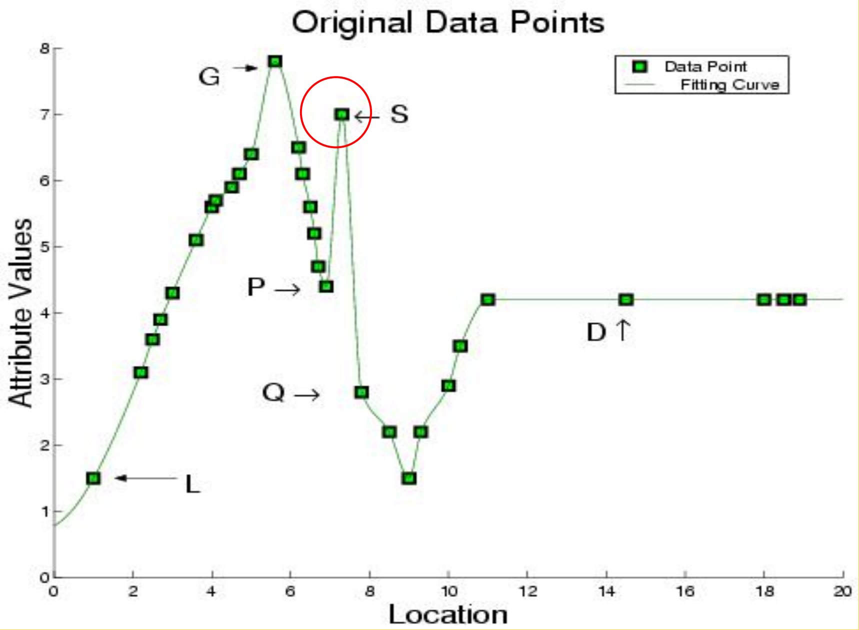
- Graphical Test: Moran Scatter Plot

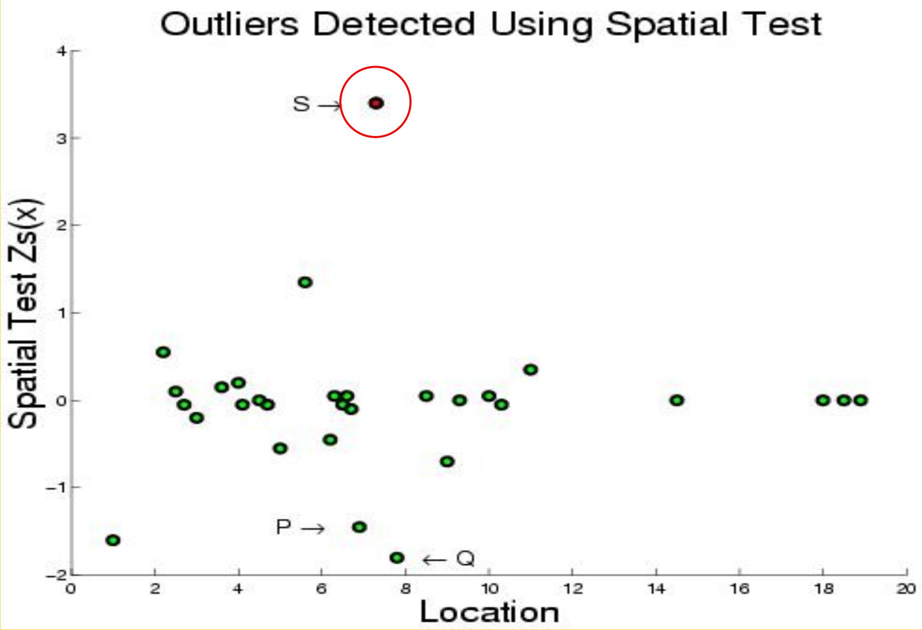




Outlier Detection Tests: Spatial Z-test

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

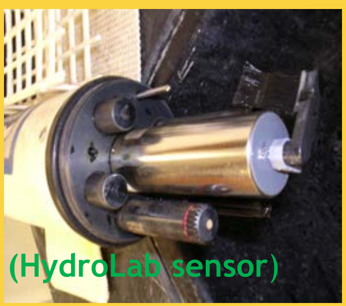


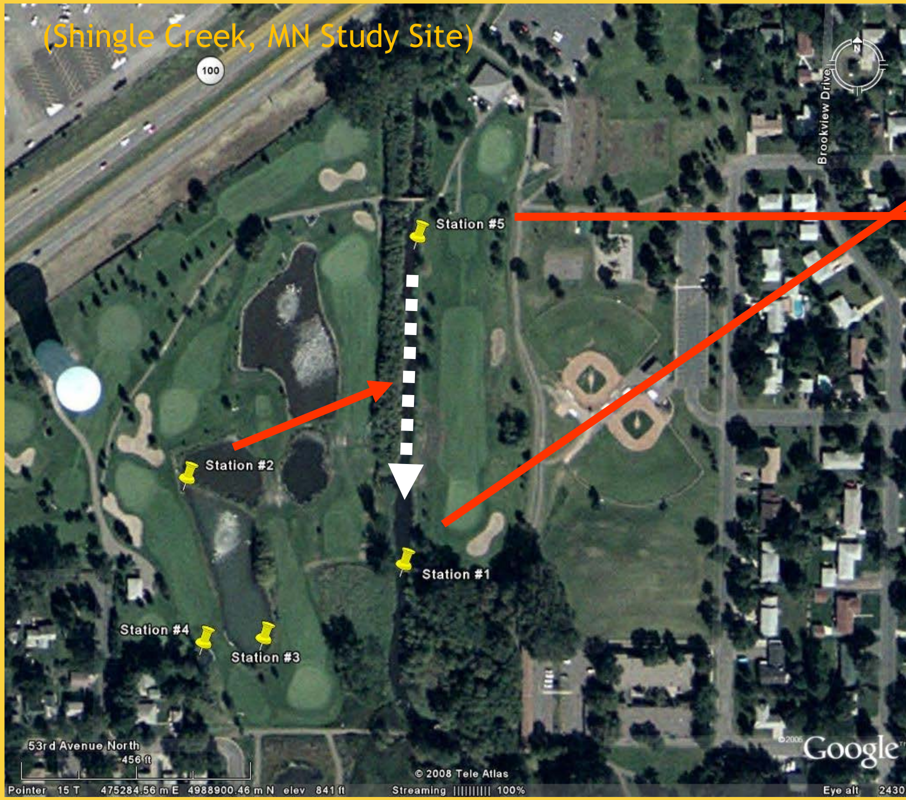


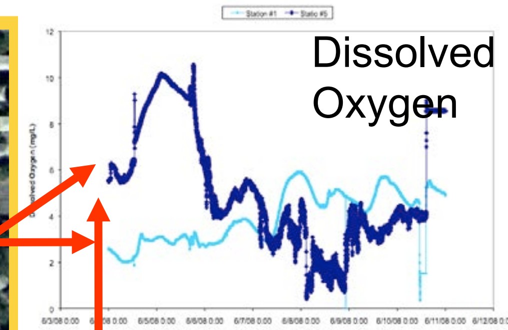
Flow Anomalies

Example Forensics: When and where do contaminants enter Shingle Creek?

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

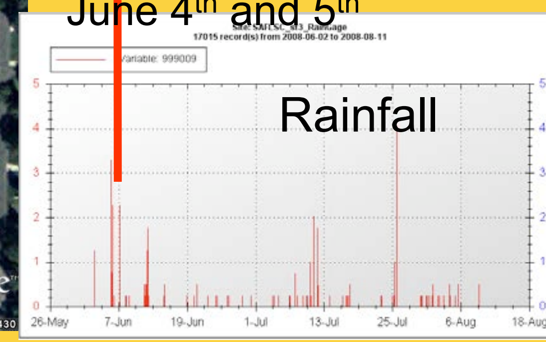






6/4/08 13:06 - 6/5/08 19:34

After heavy rains on June 4th and 5th



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

Learning Objectives

- After this segment, students will be able to
  - Contrast colocations and associations
  - Describe colocation interest measures

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

Association Rules Limitations

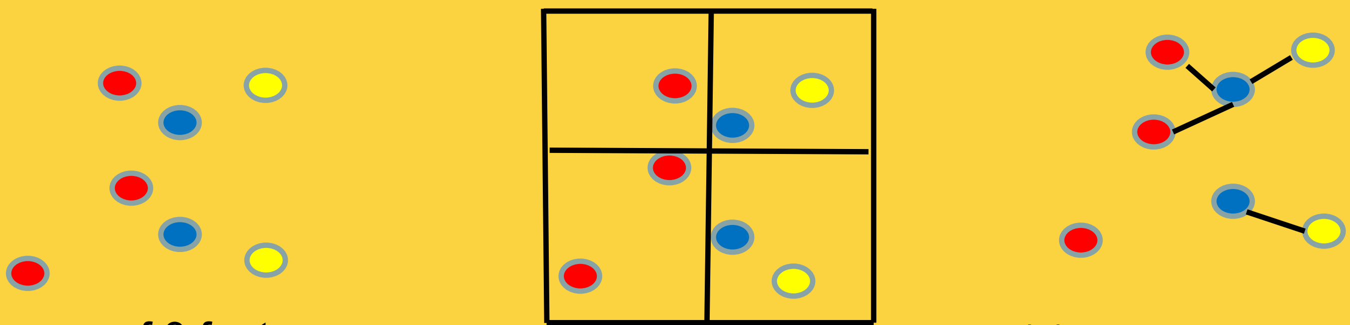
- **Transaction is a core concept!**
  - Support is defined using transactions
  - Apriori algorithm uses transaction based Support for pruning

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

- However, spatial data is embedded in continuous space
  - Transactionizing continuous space is non-trivial !
  - Recall Gerrymandering (Modifiable Areal Unit Problem)

Association Rules and Gerrymandering (MAUP)

- Support is sensitive to spatial partitioning
  - Association Rules may miss spatial interactions
- However, Ripley's K are computationally expensive



(a) a map of 3 features

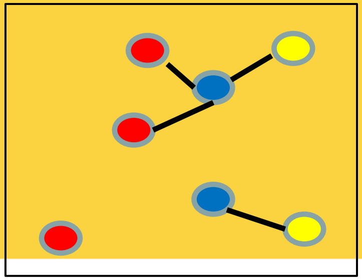
(b) Spatial Partitions

(c) Neighbor graph

	Pearson's Correlation	Support	Ripley's cross-K
-	-0.90	0	0.33
-	1	0.5	0.5

Spatial Colocation

Details: 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.).



Feature set: (, ,)

Feature Subsets:

Participation ratio (pr):

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

Participation index ( ) = **pi**( )

= min{ **pr**(,), **pr**(,) }

= min (2/3, ½) = ½

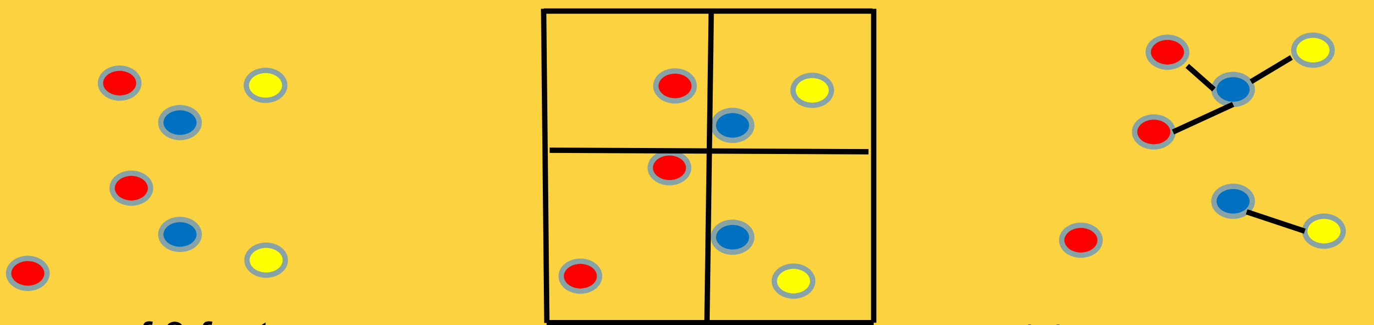
Participation Index Properties:

(1) <u>Computational</u>: Non-monotonically decreasing like support measure

(2) <u>Statistical</u>: Upper bound on Ripley's Cross-K function

Neighbor Relationship vs. Space Partitioning

- Neighbor relationship graph
  - Honors continuity of geographic space methods
  - Partitions miss spatial interactions



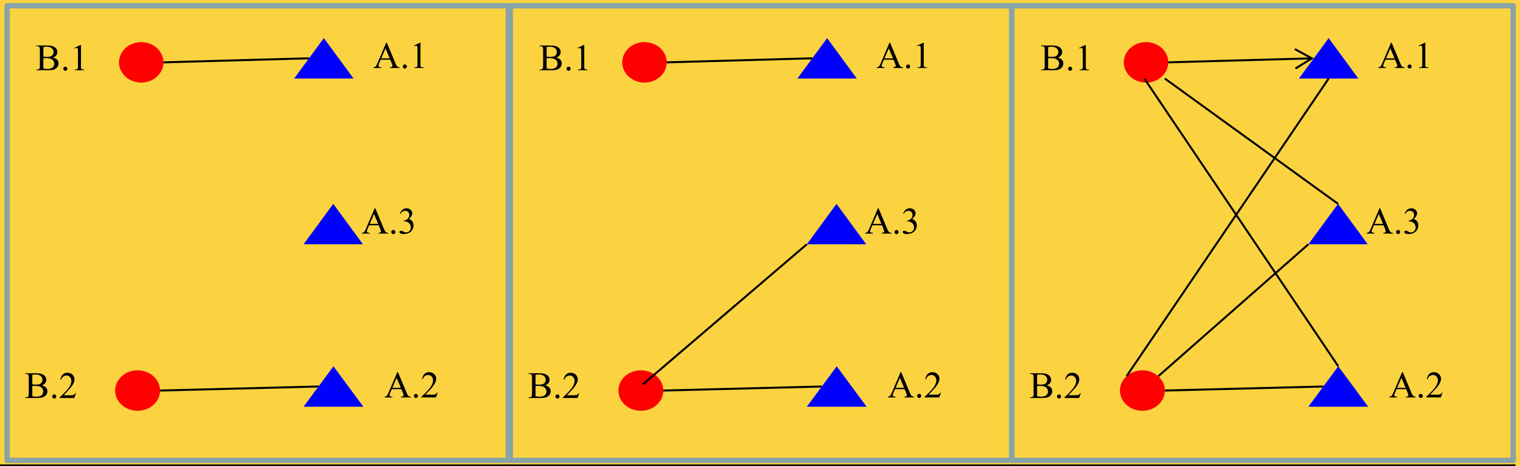
(a) a map of 3 features

(b) Spatial Partitions

(c) Neighbor graph

	Pearson's Correlation	Support	Ripley's cross-K	Participation Index (colocation)
-	-0.90	0	0.33	0.5
-	1	0.5	0.5	1

Participation Index >= Cross-K Function

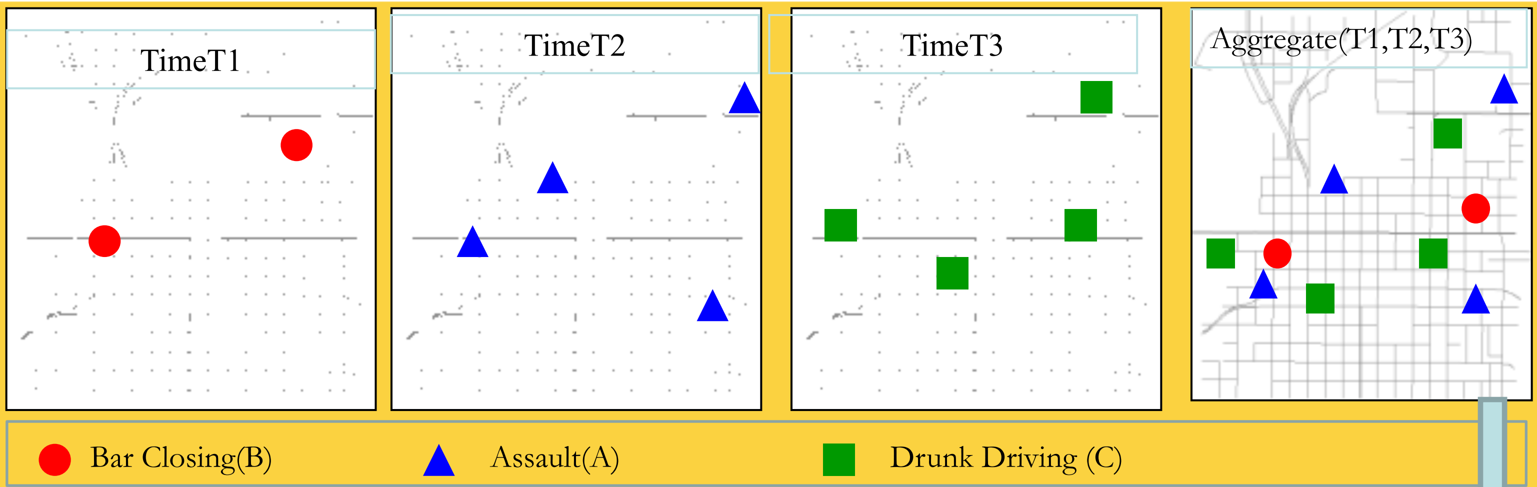


Cross-K (A,B)	2/6 = 0.33	3/6 = 0.5	6/6 = 1
PI (A,B)	2/3 = 0.66	1	1

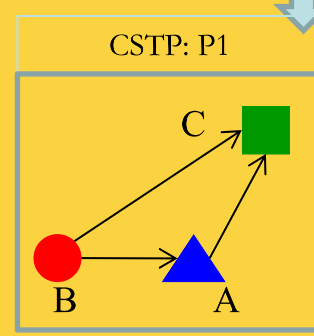
Spatial Colocation: Trends

- 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)

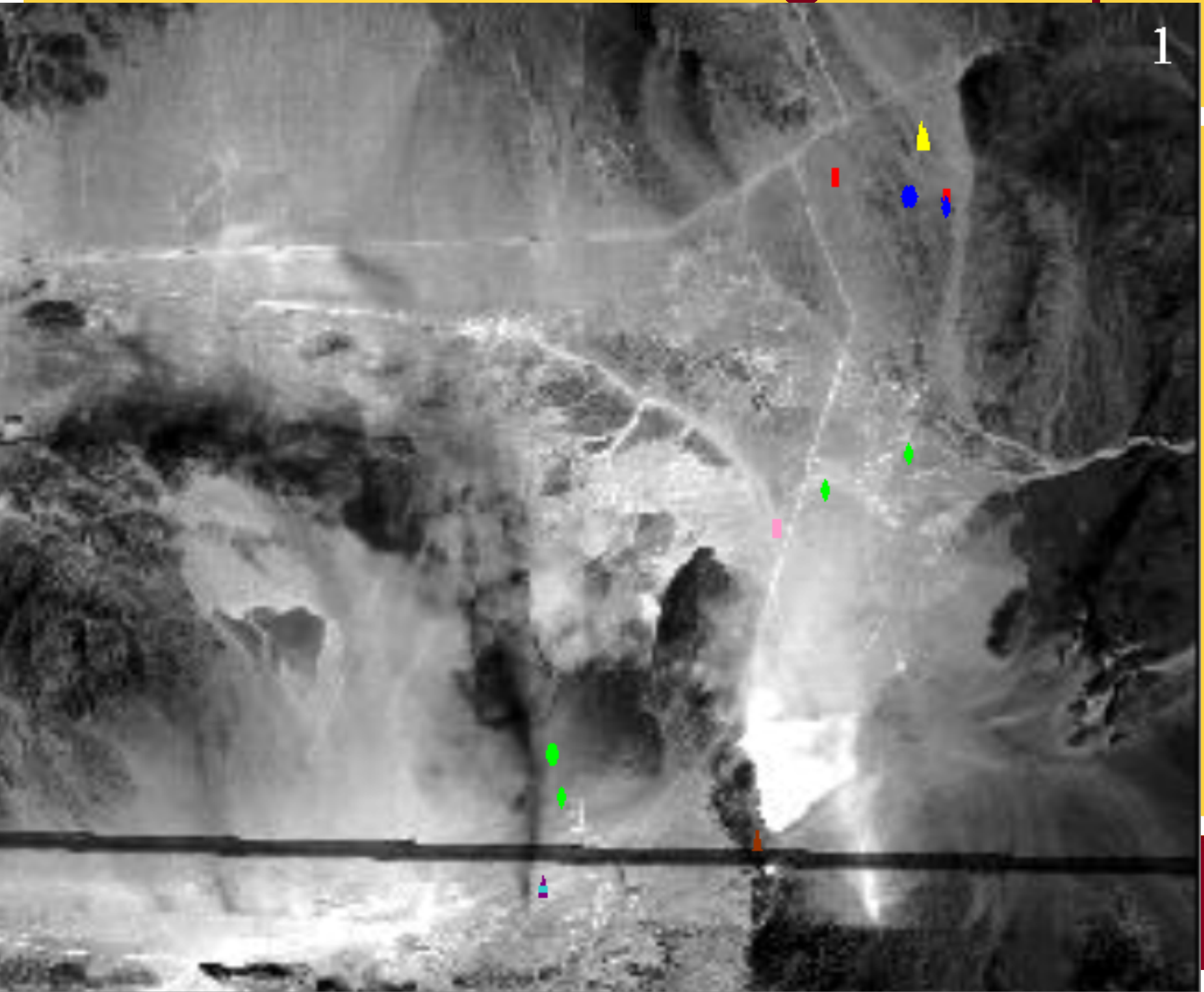


- ❑ *Input:* Urban Activity Reports
- ❑ *Output: CSTP*
  - ❑ *Partially ordered* subsets of ST event types.
  - ❑ Located together in space.
  - ❑ Occur in *stages* over time.
- ❑ Applications: Public Health, Public Safety, …



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

MDCOP Motivating Example :



Manpack stinger
(2 Objects)

M1A1_tank
(3 Objects)

M2_IFV
(3 Objects)

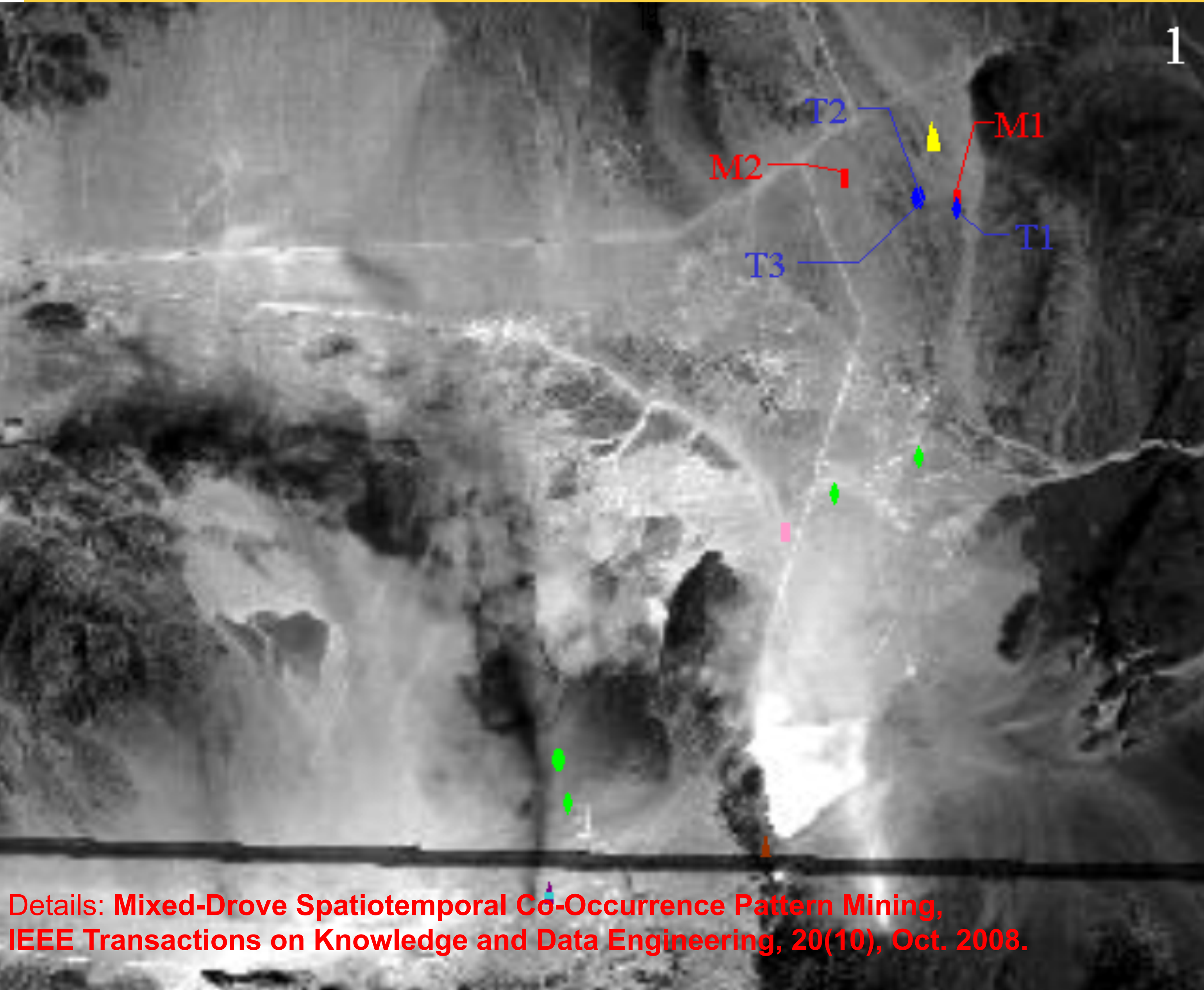
Field_Marker
(6 Objects)

T80_tank
(2 Objects)

BRDM_AT5
(enemy) (1 Object)

BMP1
(1 Object)

MDCOP Motivating Example : Output



- Manpack stinger (2 Objects)

- M1A1_tank (3 Objects)

- M2_IFV (3 Objects)

- Field_Marker (6 Objects)
- T80_tank (2 Objects)

- BRDM_AT5 (enemy) (1 Object)

- BMP1 (1 Object)

Details: **Mixed-Drove Spatiotemporal Co-Occurrence Pattern Mining, IEEE Transactions on Knowledge and Data Engineering, 20(10), Oct. 2008.**

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	
	Clusters	Hot-spots	Flock pattern Moving Clusters
	Outlier	Spatial outlier	Spatio-Temporal outlier
	Prediction	Location prediction	Future Location prediction

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.

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.)
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.)

NSF INFEWS Data Science Workshop

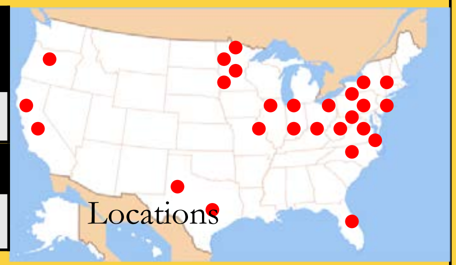
(@ USDA NIFA, Oct. 5th-6th, 2015; Shekhar, Mulla, & Schmoldt; www.spatial.cs.umn.edu/few)

Goals:
- Design compelling visions, Identify gaps
- Develop a research agenda

55 Participants (Data-driven FEW & Data Sciences)

Food	Energy	Water	Data Sc.
14	10	11	20

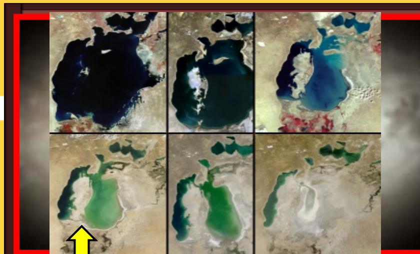
Gov.	Aca.	Industry
26	24	5



Finding 1: Data & Data Science are crucial!
- Understand problems, connections, impacts
- Monitor FEW resources, and trends to detect risks
- Support decision and policy making
- Communicate with public and stakeholders

Finding 2: However, there are show-stopper gaps.

1. Data Gaps: No global water & energy census, Heterogeneous data formats & collection protocols
2. Data Science (DS) Gaps: Current DS methods are inadequate for spatio-temporal-network FEW data.



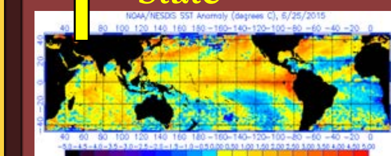
Aral Sea Shrinkage (1978-2014)
Due to Cotton Farms

Alerts

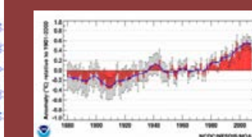
Nexus Dashboard

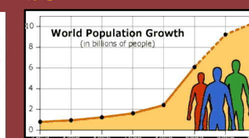
State

Trends



Sea-Surface Temperature Anomaly


Global Temperature



Global Population

Potentially Transformative Research Agenda:
- National FEW Nexus Observatory & Dashboard for chokepoint monitoring, alerts, warnings
- Novel Physics-aware Data Science for mining nexus patterns in multi-scale spatio-temporal-network data despite non-stationarity, auto-correlation, uncertainty, etc.
- Scalable tools for consensus Geo-design via participative planning with nexus observations and policy projections
- An INFEWS data science community to address crucial gaps, and shape next-generation Data Science

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Global Population