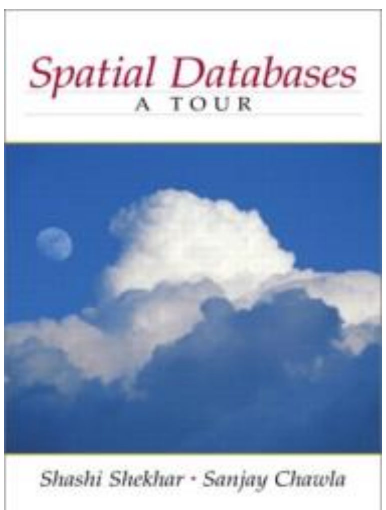
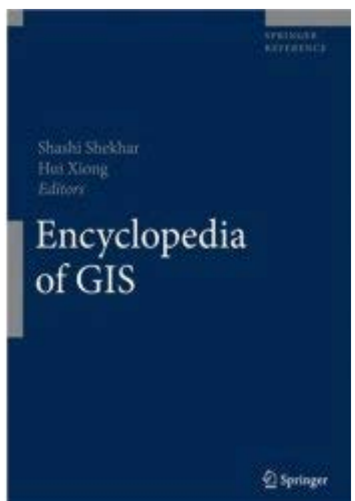
What is Special about **Spatial Data Science and GeoAI?**

Shashi Shekhar
McKnight Distinguished University Professor, University of Minnesota
www.cs.umn.edu/~shekhar, shekhar@umn.edu

Acks: Collaborators, Sponsors (NSF, USDOD NGA, USDOE ARPA-E, USDA NIFA, NIH, …)






University Consortium for GEOGRAPHIC INFORMATION SCIENCE

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

Summer 2018

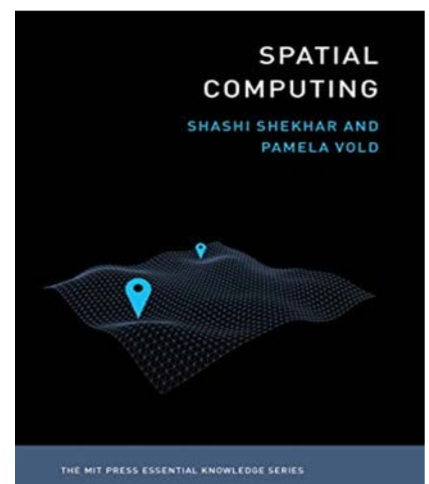
As a long-established information science discipline, the Geographic Information Science & Technology (GIS&T) community has key contributions to make to evolving data science curricula. This statement articulates the University Consortium for Geographic Information Science's (UCGIS) position for the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

Context

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

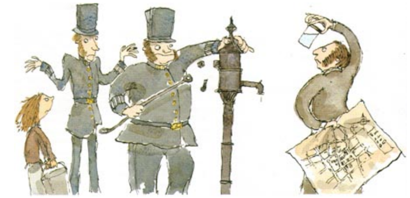
Data that are geographically referenced or contain some type of location markers are both common and of high value (e.g., data subject to state-specific policies, laws and regulations; demographic data from the census; location traces of smartphones and vehicles; remotely sensed imagery from satellites,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

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



A Spatial Data Science Story

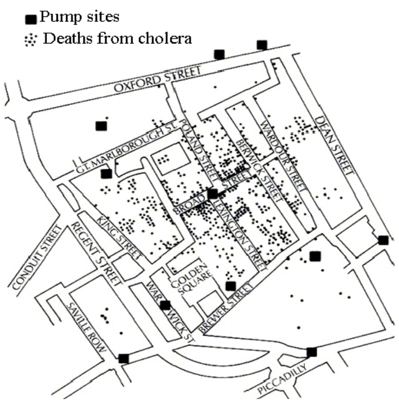
1854: What causes Cholera?

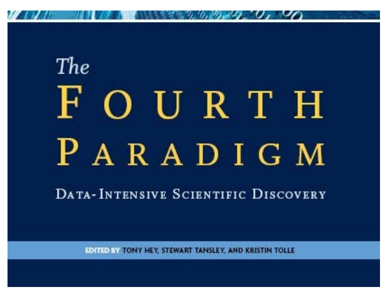


Collect & Curate Data → Discover Patterns, Generate Hypothesis → Test Hypothesis (Experiments) → Develop Theory

? water pump

Remove pump handle





Impact: hygiene, drinking water supply, sewage system, …

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

What has changed? **Spatial Data Revolution**

Spatial	Last Century	Last Decade
Data	Few satellites and sensors	Nano-satellites, Billions of GPS enabled smartphones
Data Access	Need special hardware and network	
Spatial Platforms	ESRI Arc/Info	
Spatial Data Science	Spatial Patterns, e.g., hotspots (SatScan, ESRI Geostatistics)	
Spatial Visualization	Quilt, e.g., MS Terraserver	

Spatial Computing is Ubiquitous 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

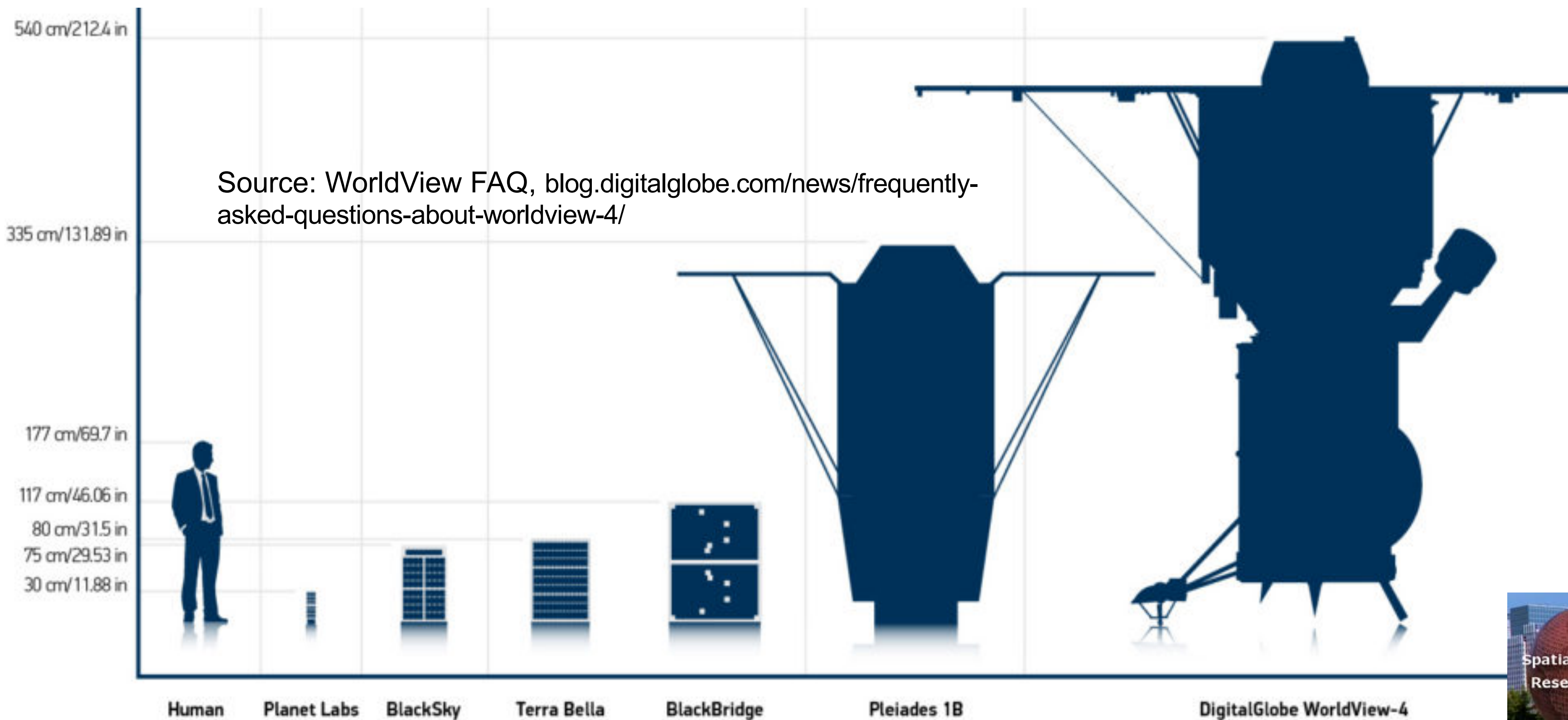
July 25, 2018, 4:00 AM CDT

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*

Large Constellations of Small Satellites

- Hi-frequency (e.g., daily or hourly) time-series of imagery of entire earth
- Small Satellites: video (5-minutes): https://geospatialstream.com/sciencecasts-nasa-embraces-small-satellites/
- Large Constellations
  - 2017: Planet Labs: 100 satellites: daily scan of Earth at 1m resolution in visible band



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

Spatial Computing Research Group

Spatial Data Revolution

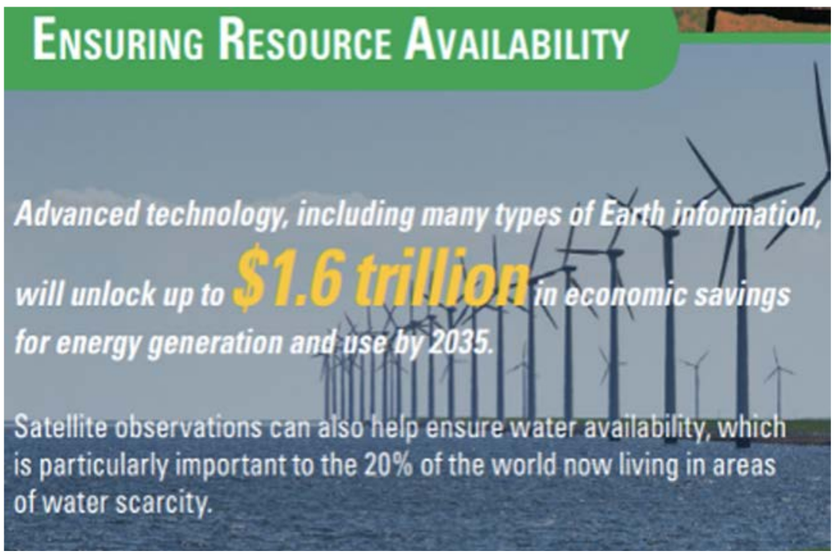
1. **GPS & Location traces**
   - 2 billion GPS receivers today (7 billion by 2022)
   - Reference clock for telecom, banks, …
   - Help understand Spatio-temporal patterns of life

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

Bloomberg Businessweek

July 25, 2018, 4:00 AM CDT

2. **(Nano-)Satellite Imagery**, …



McKinsey Global Institute

The study estimates that the use of personal location data could save consumers worldwide more than $600 billion annually by 2020. Computers determine users' whereabouts by tracking their mobile devices, like cellphones.

The New York Times

Published: May 13, 2011

Source: Y. Xie et al., Transforming Smart Cities With Spatial Computing, Proc. IEEE Intl. Conf. on Smart Cities, 2018.

What has changed? **Spatial Data Access**

Spatial	Last Century	Last Decade
Data	Few satellites and sensors	Nano-satellites, Billions of GPS enabled smartphones
Data Access	Need special hardware and network	Cloud based repositories and analytics (e.g., DARPA GCA)
Spatial Platforms	ESRI Arc/Info	
Spatial Data Science	Spatial Patterns, e.g., hotspots (SatScan, ESRI Geostatistics)	
Spatial Visualization	Quilt, e.g., MS Terraserver	

Easier Access: Cheap (or free) Cloud Repositories

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

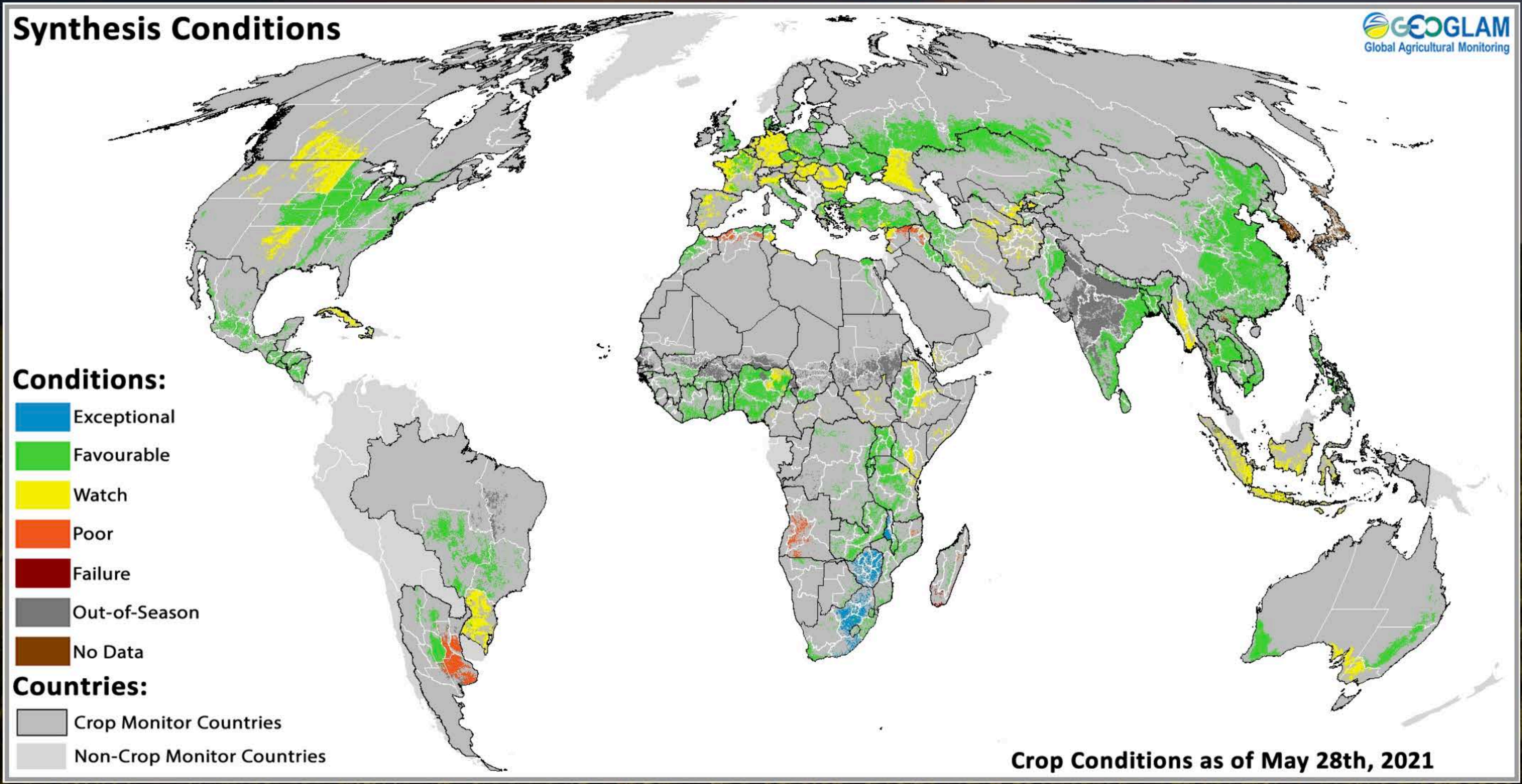
	Google Earth Engines	NEX	AWS Earth
Elevation, Landsat, LOCA, MODIS, NAIP	x	x	x
NOAA	x		x
AVHRR, FIA, GIMMM, GlobCover, NARR, TRIMM, Sentinel-1	x	x	
IARPA, GDELT, MOGREPS, OpenStreetMap, Sentinel-2, SpaceNet (building/road labels for ML)			x
CHIRPS, GeoScience Australia, GSMap, NASS, Oxford Map, PSDI,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

Google Earth Engine

NEX NASA Earth Exchange

Global Agriculture Monitoring

LATEST CONDITIONS PUBLISHED IN OUR REPORTS



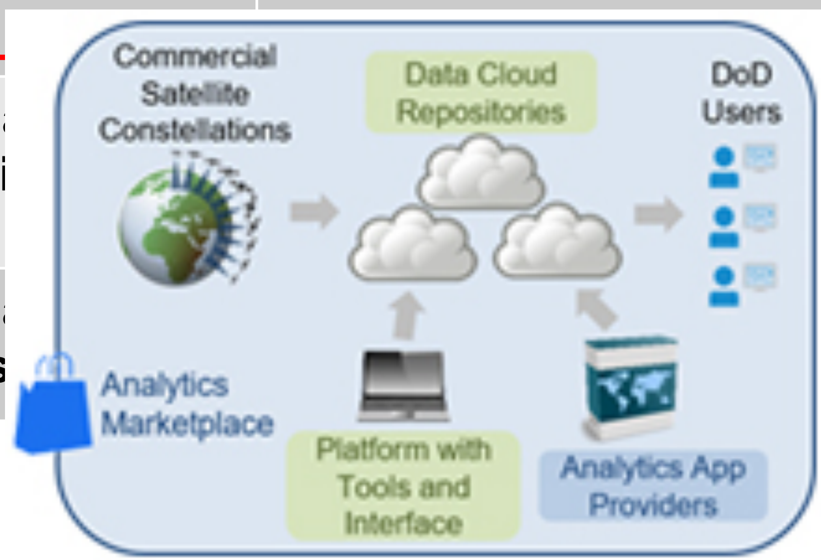
AMIS & EARLY WARNING LATEST SYNTHESIS MAP

AMIS Agricultural Market Information System

GEOGLAM Global Agricultural Monitoring

What has changed? **Spatial Big Data Platforms**

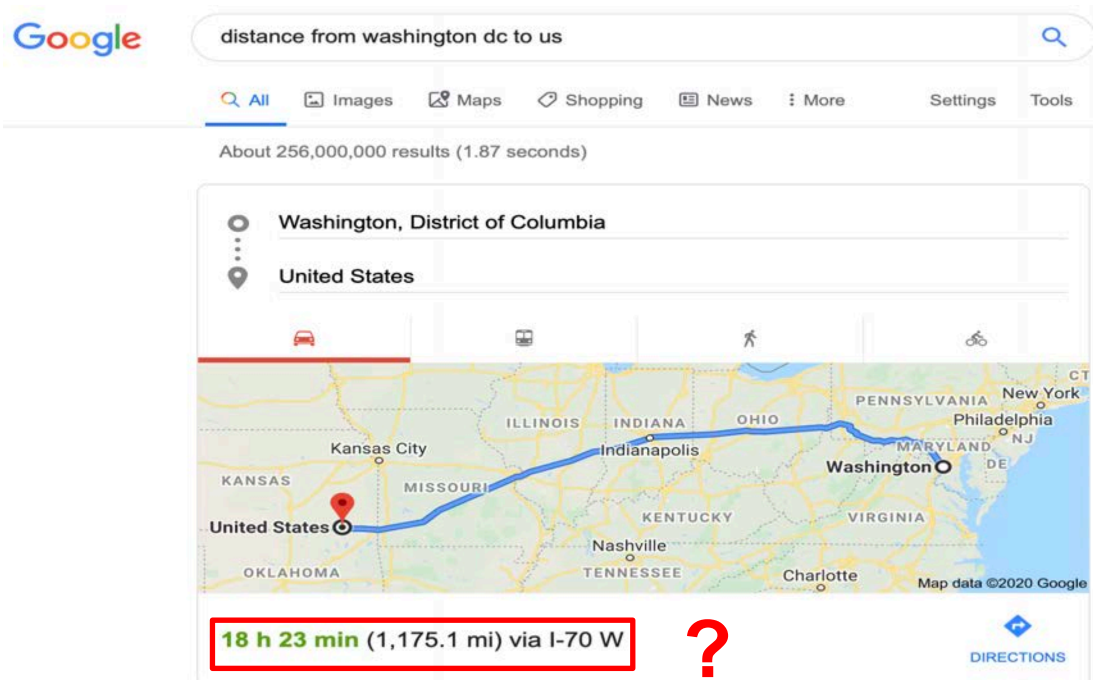
Spatial	Last Century	Last Decade
Data	Few satellites and sensors	Nano-satellites, Billions of GPS enabled smartphones
Data Access	Need special hardware and network	Cloud based repositories, e.g., Earth on AWS
Spatial Platforms	ESRI Arc/Info, SQL3/OGC, e.g., Postgis,	Geospatial Cloud Analytics (Monitor crops, fracking, illegal fishing), ESRI GIS Tools for Hadoop, …
Spa Sci		
Spa Vis		



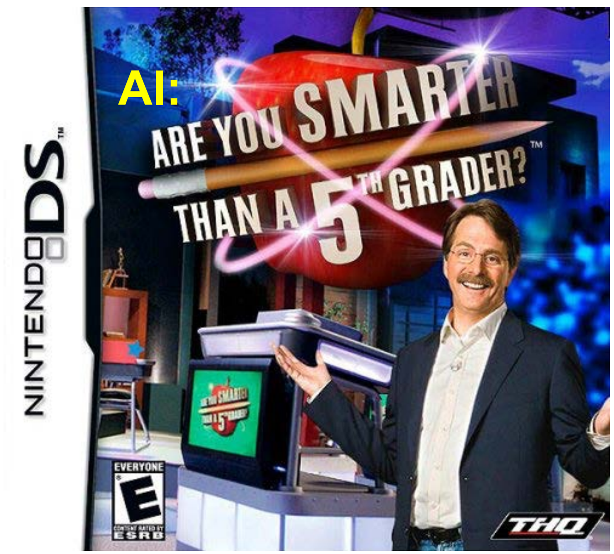
Spatial Data Types >> Points

Q? What is distance between Washington D.C. and U.S.A.?

- Zero (Washington D.C. is inside U.S.A.)
- NSF OKN funded 2 grants on geo-knowledge networks!

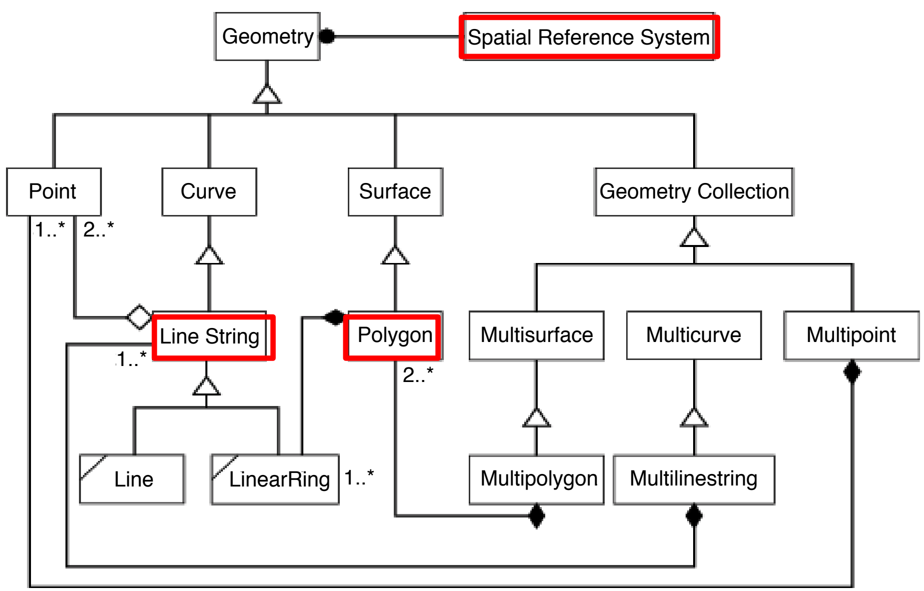


?



Spatial Data Types: OGC Simple Features Standard

- Data types: Point, LineString, Polygon, Collections
- Relationships: Topological, Metric, …
- Helps in feature selection for machine learning
  - Ex. Distance to key geo-features, Neighbor relationship

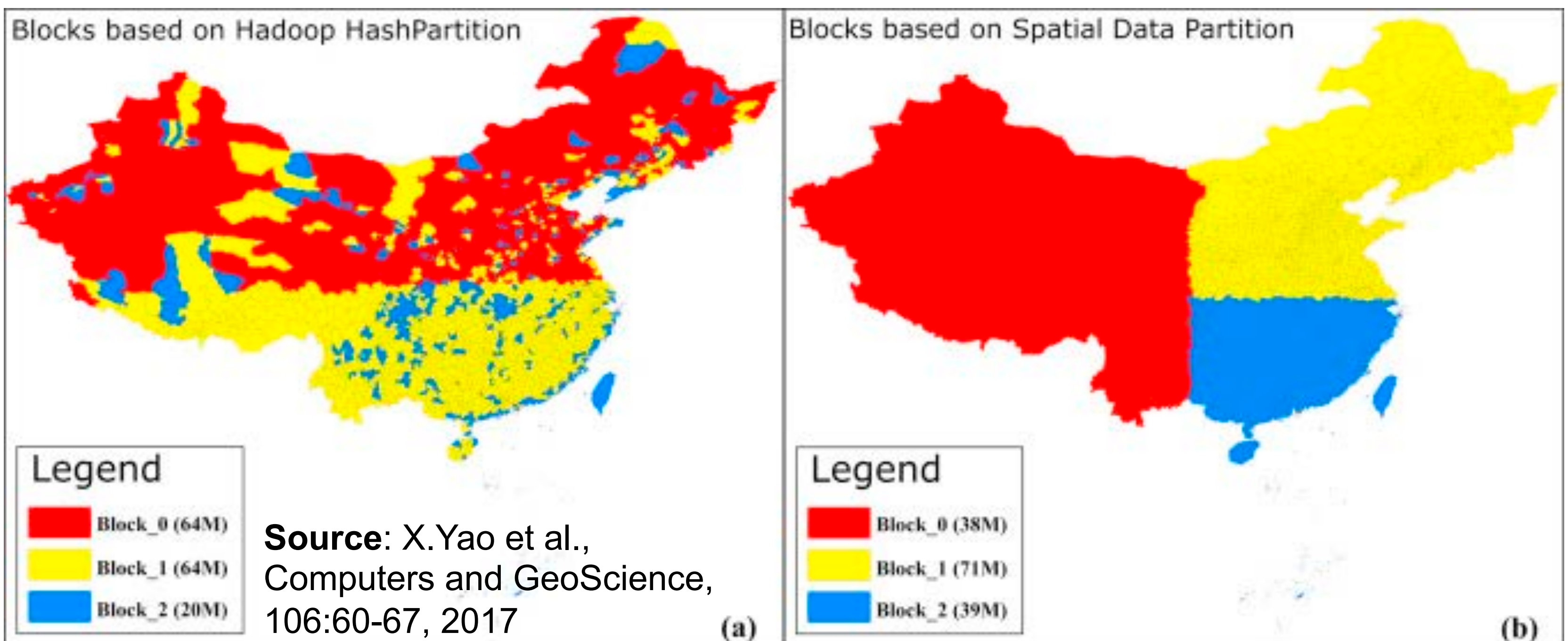


Basic Functions	SpatialReference ()
	Envelop ()
	Export ()
	IsEmpty ()
	IsSimple ()
	Boundary ()
Topological / Set Operators	Equal
	Disjoint
	Intersect
	Touch
	Cross
	Within
	Contains
	Overlap
Spatial Analysis	Distance
	Buffer
	ConvexHull
	Intersection
	Union
	Difference
	DymmDiff

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

Spatial Big Data Platforms

Genre	Examples
Relational DBMS, Spatial Library	Oracle, IBM DB2, PostgreSQL, MS SQL Server OGC Simple Features, …
Parallel DBMS	Teradata, Vertica, Greenplum, DataAllegro, ParAccel
Big Data Platforms	Hadoop, MapReduce, Spark, Hbase, Hive, …
Spatial Big Data Platforms	ESRI GIS Tools for Hadoop, GeoWave, SpatialSpark, GeoSpark, Simba, Hadoop-GIS, SpatialHadoop, ST-Hadoop



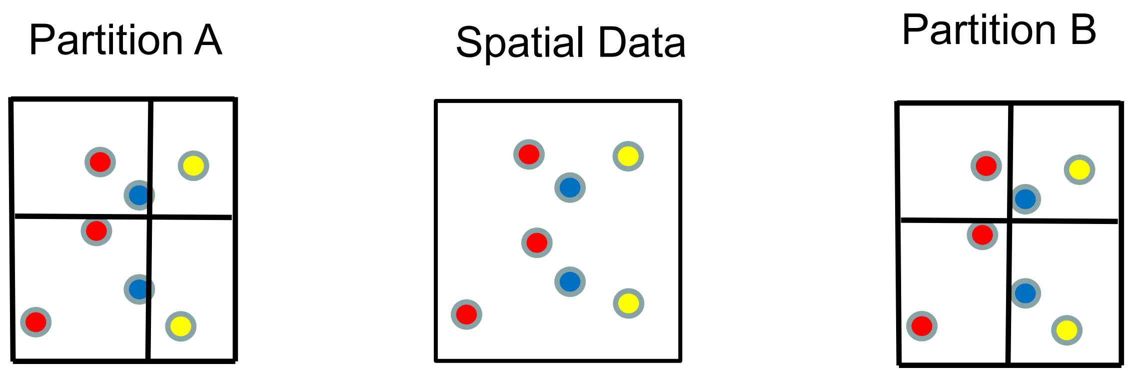
Source: X.Yao et al., Computers and GeoScience, 106:60-67, 2017

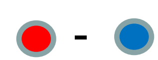
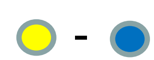
What has changed? **Spatial Data Science**

Spatial	Last Century	Last Decade
Data	Few satellites and sensors	Nano-satellites, Billions of GPS enabled smartphones
Data Access	Need special hardware and network	Cloud based repositories, e.g., Earth on AWS
Spatial Platforms	ESRI Arc/Info	SQL3/OGC, e.g., Postgis, ESRI GIS Tools for Hadoop, Google Earth Engine
Spatial Data Science	Spatial Patterns, e.g., hotspots (SatScan, ESRI Geostatistics)	(a) Spatial Network Patterns, e.g., linear hotspots (b) Spatio-temporal (ST) patterns, e.g., Change time-series
Spatial Visualization	Quilt, e.g., MS Terraserver	

Limitations of Traditional Data Science

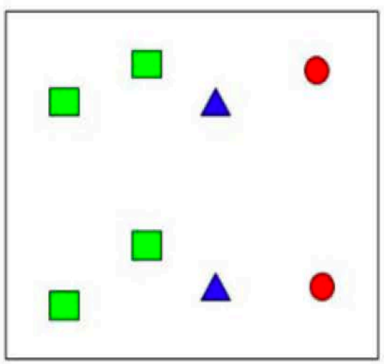
- Traditional methods not robust in face of
  - Spatial continuity
    - Gerrymandering risk: Classical methods not robust
    - Result changes if spatial partitioning changes
  - Auto-correlation, Heterogeneity , Edge-effect, …
  - Noise



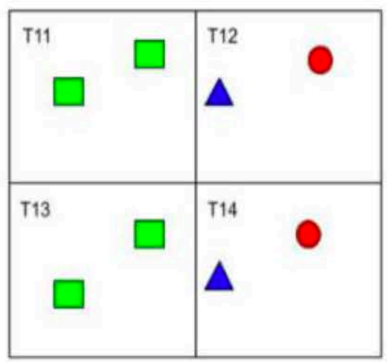
Partition A Based Pearson's Correlation	Pairs	Partition B Based Pearson's Correlation
1		- 0.90
- 0.90		1

Classical Data Mining Methods not robust either!

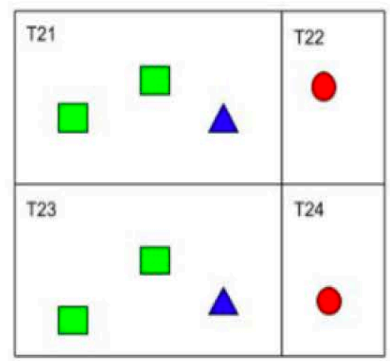
Consider the spatial Data in Figure (a)
Along with 3 alternative partitions in Figures (b), (c) and (d).



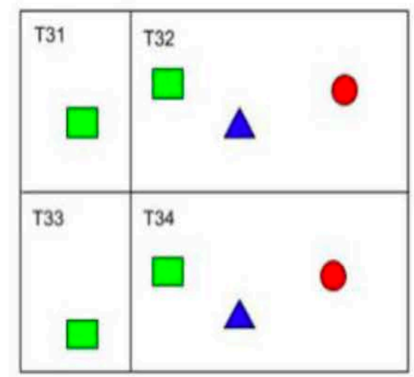
(a) Map of 3 item-types



(b) Spatial Partition P1



(c) Spatial Partition P2

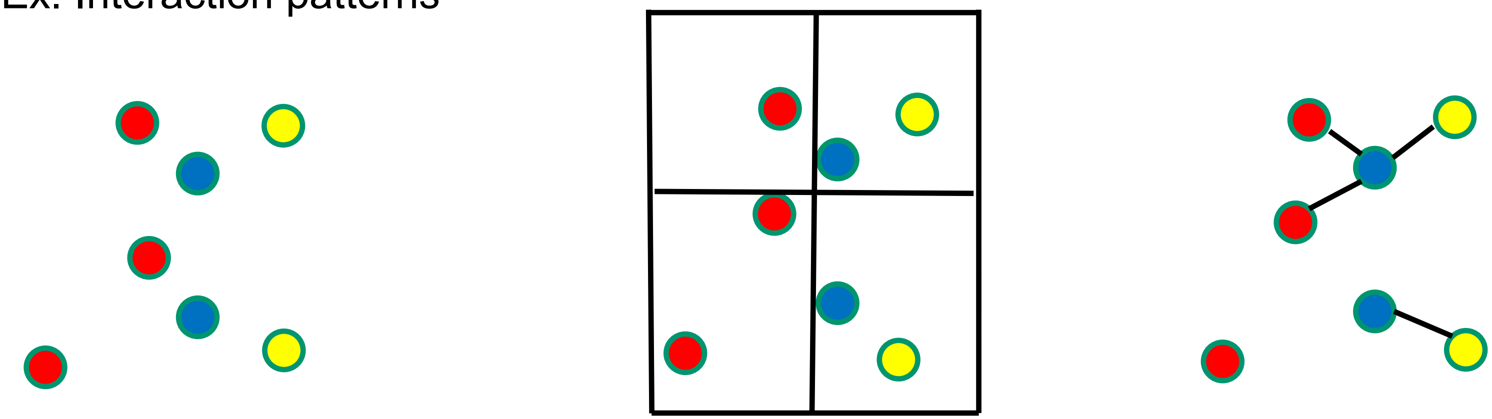


(d) Spatial Partition P3

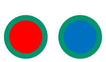
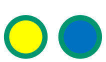
Spatial Partitioning Definition	**P1**	**P2**	**P3**
Transactions	T11, T12, T13, T14	T21, T22, T23, T24	T31, T32, T33, T44
Associations with support >= 0.5	(▲ ●)	(■ ▲)	(■ ▲ ●)

Neighbor Graph Approach

- Challenge: One size does not fit all
- Ex. Interaction patterns



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

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

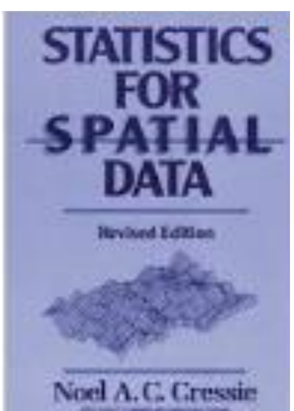
Details: Discovering Spatial Co-location Patterns: A General Approach,
IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

A Metric of Spatial Cross-Correlation

- Ripley's Cross K-Function Definition

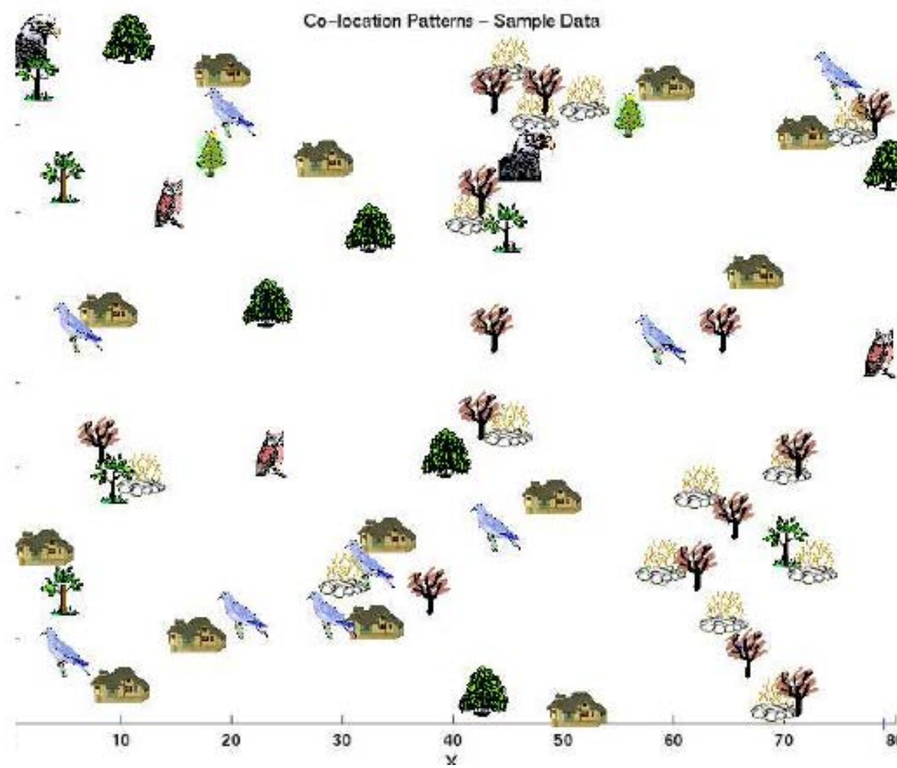
$$K_{ij}(h) = \lambda_j^{-1} E \text{ [number of type } \boldsymbol{j} \text{ event within distance } h \text{ of a randomly chosen type } \boldsymbol{i} \text{ event]}$$

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



Co-locations

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

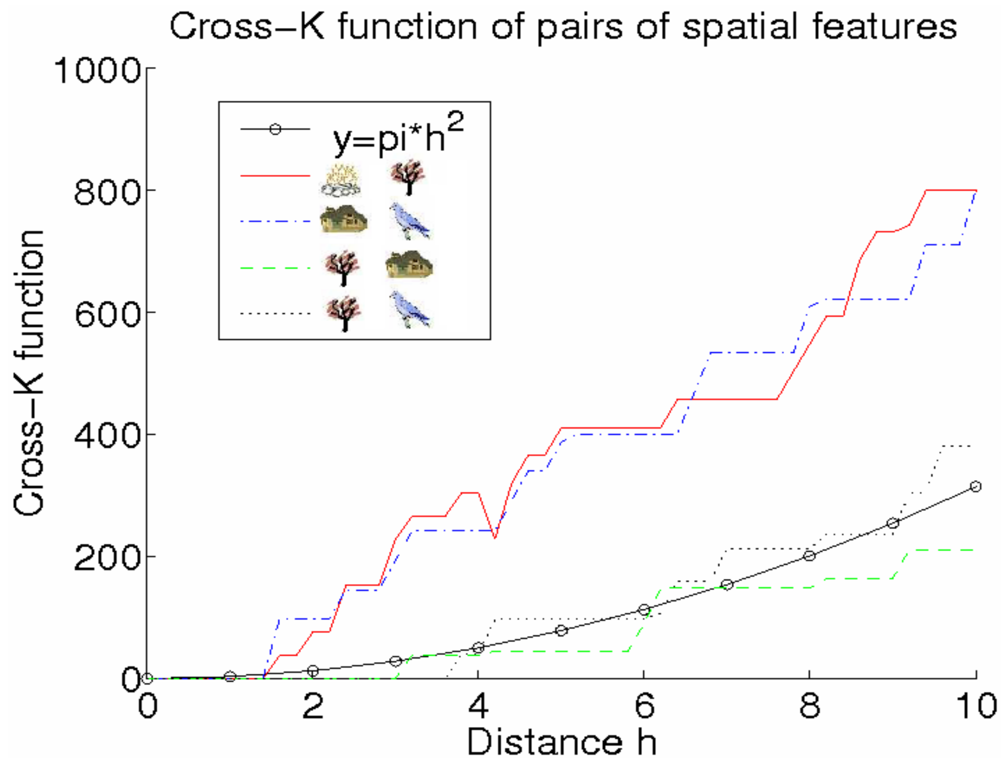


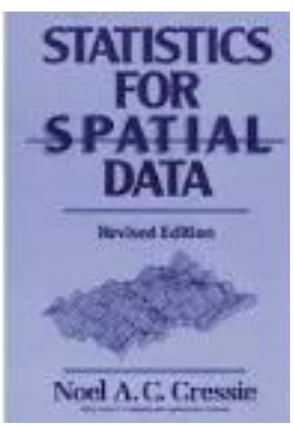
Source: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

Spatial Computing Research Group

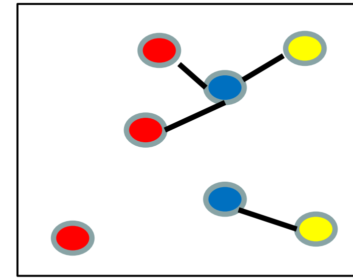
Illustration of Cross-Correlation

- Illustration of Cross K-function for Example Data





Spatial Colocation



Feature set: (, ,)

Feature Subsets:

Participation ratio (pr):

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

pr(,) = ½

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

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

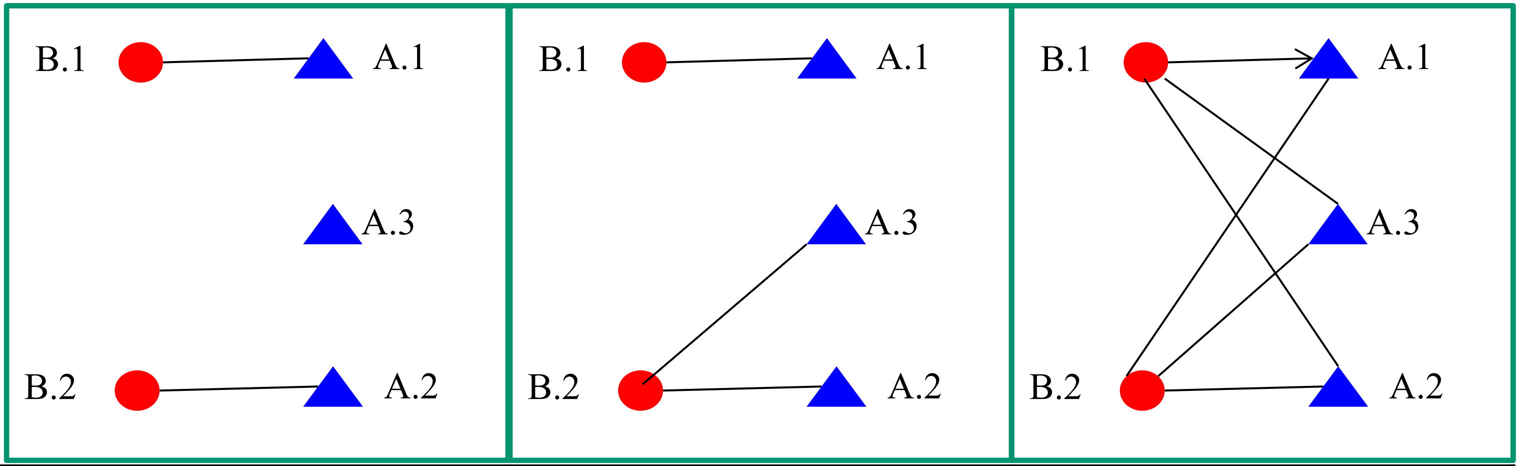
= min (2/3, ½) = ½

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

Participation Index >= Cross-K Function

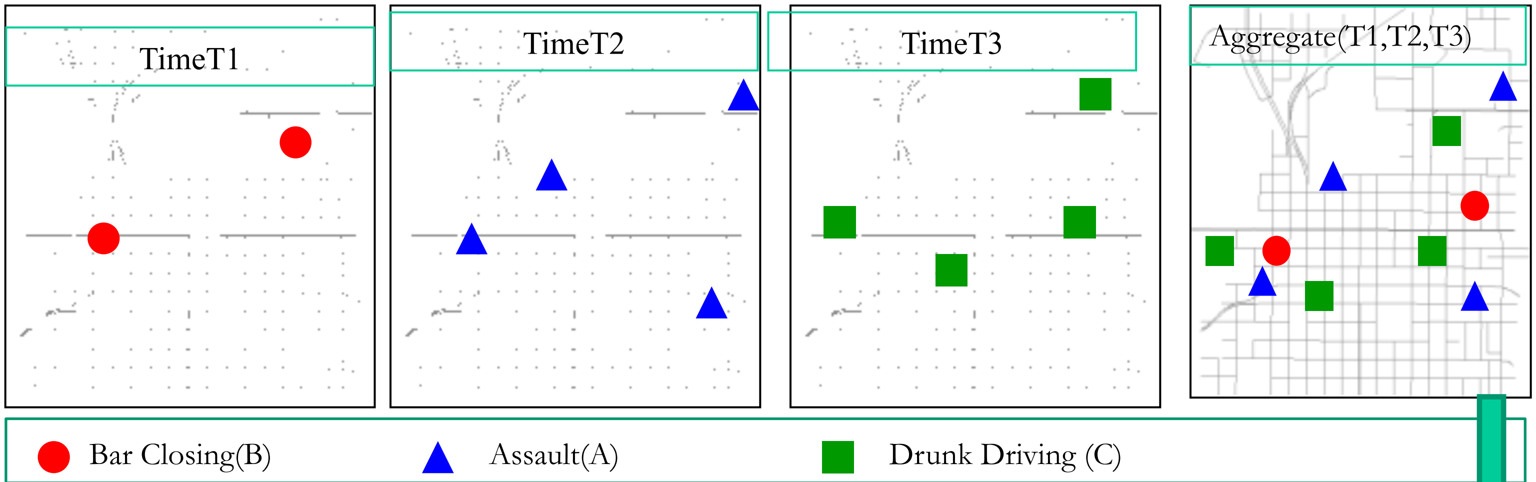


Cross-K (A,B)	2/6 = 0.33	3/6 = 0.5	6/6 = 1
PI (A,B)	2/3 = 0.66	1	1

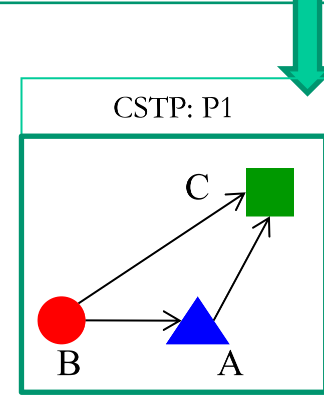
Spatial Colocation: Trends

- 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)



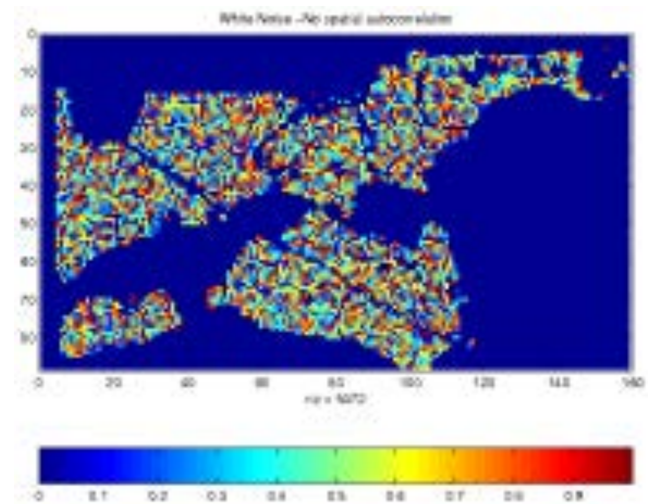
- *Input:* Urban Activity Reports
- *Output: CSTP*
  - *Partially ordered* subsets of ST event types.
  - Located together in space.
  - Occur in *stages* over time.
- Applications: Public Health, Public Safety, …



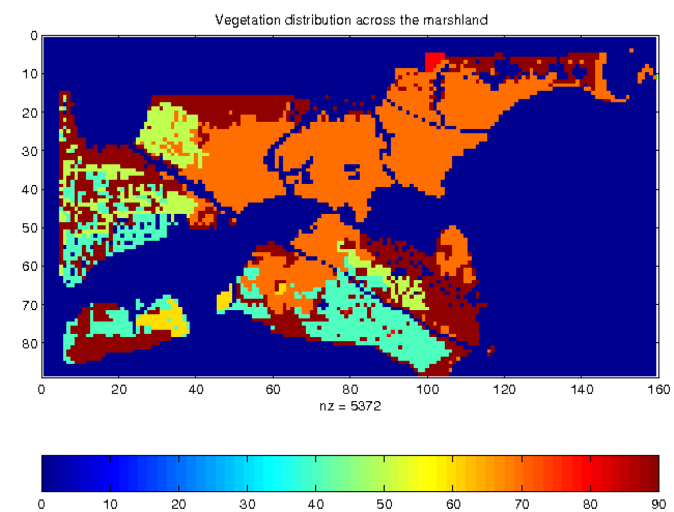
Details: Cascading Spatio-Temporal Pattern Discovery, IEEE Trans. on Know. & Data Eng, 24(11), 2012.

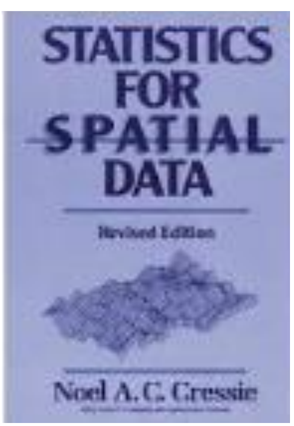
Spatial Auto-correlation and Prediction

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



- Limitation of i.i.d assumption
  - Ignores auto-correlation
  - Salt n Pepper noise





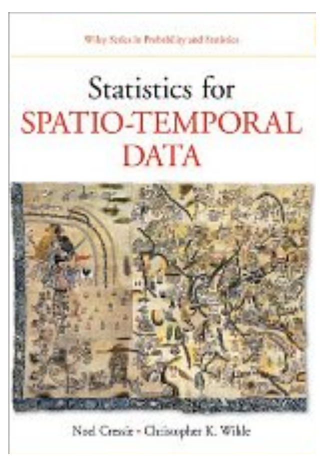
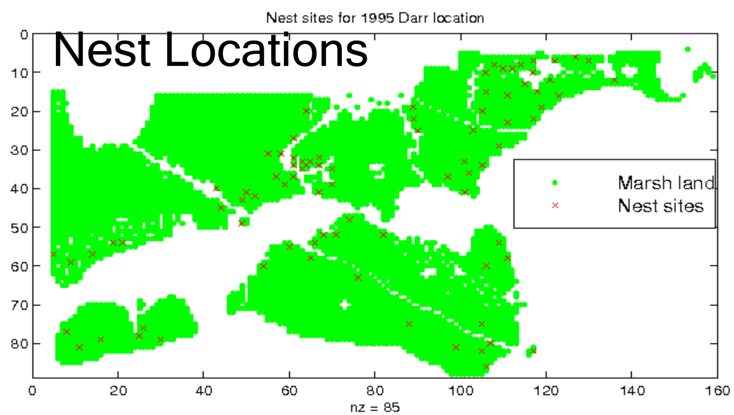
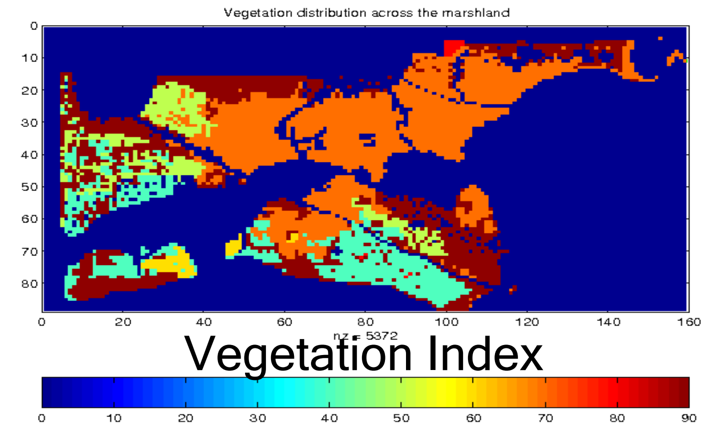
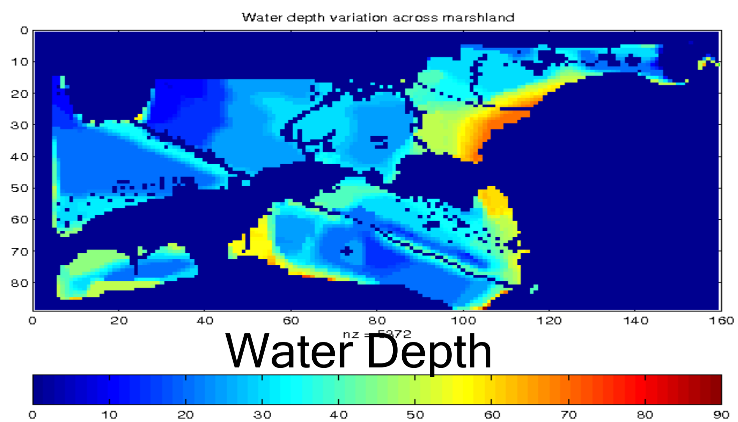
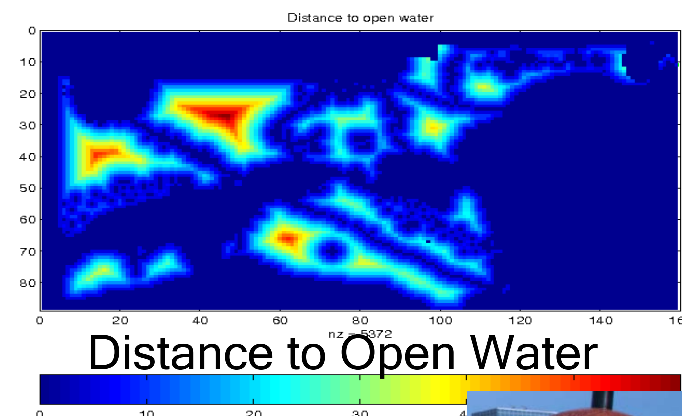


Illustration of Location Prediction Problem









Spatial Computing Research Group

Spatial Auto-Regression & Parameter Estimation

Name	Model
Classical Linear Regression	$\boxed{\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$
Spatial Auto-Regression	$\boxed{\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$

ρ : the spatial auto - regression (auto - correlation) parameter

$\mathbf{W}$: n - by - n neighborhood matrix over spatial framework

- **<u>Maximum Likelihood Estimation</u>**

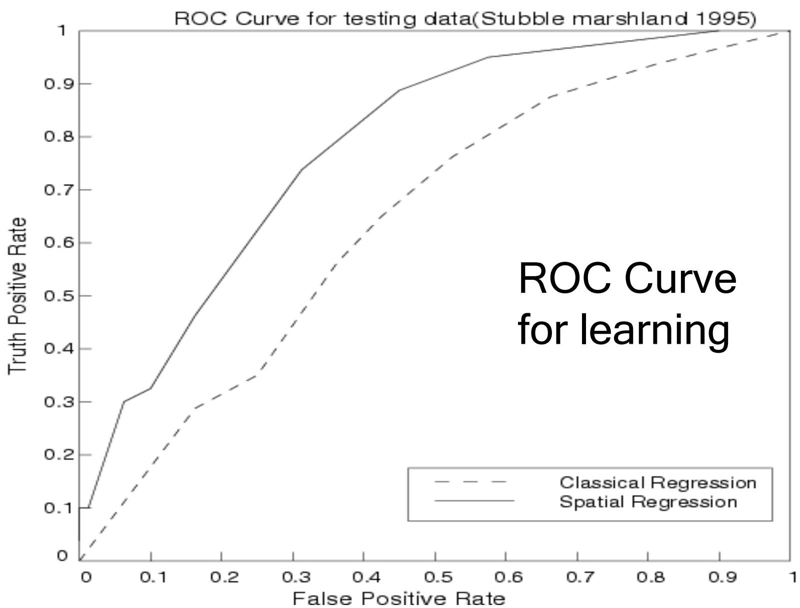
$$\ln(L) = \boxed{\ln\left|\mathbf{I} - \rho\mathbf{W}\right|} - \frac{n\ln(2\pi)}{2} - \frac{n\ln(\sigma^2)}{2} - SSE$$

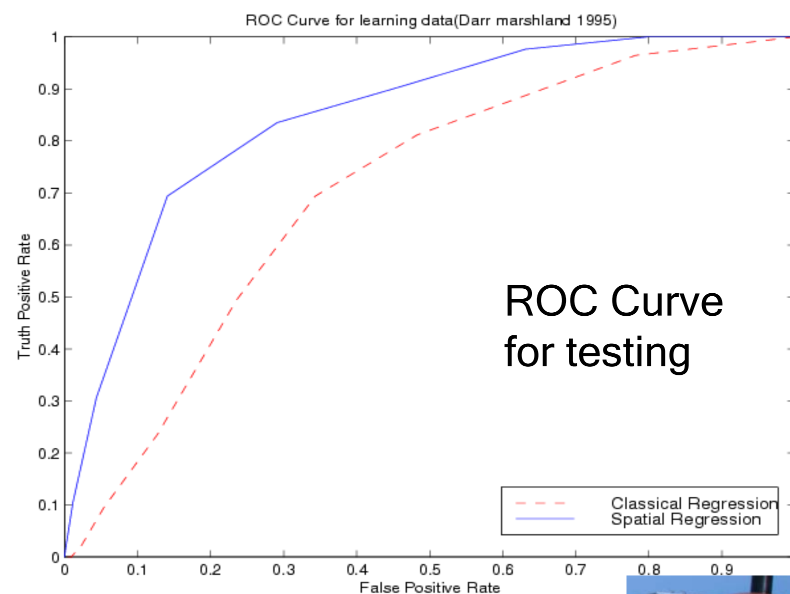
- Computing determinant of large matrix is a hard (open) problem!
  - size(W) is quadratic in number of locations/pixels.
  - Typical raster image has Millions of pixels
  - W is sparse but not banded.

Details: A parallel formulation of the spatial auto-regression model for mining large geo-spatial datasets, SIAM Intl. Workshop on High Perf. and Distr. Data Mining, 2004. (with B. Kazar)

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





Spatial Computing
Research Group

Prediction Error and Bias Trade-off

- Linear Regression (LR)

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

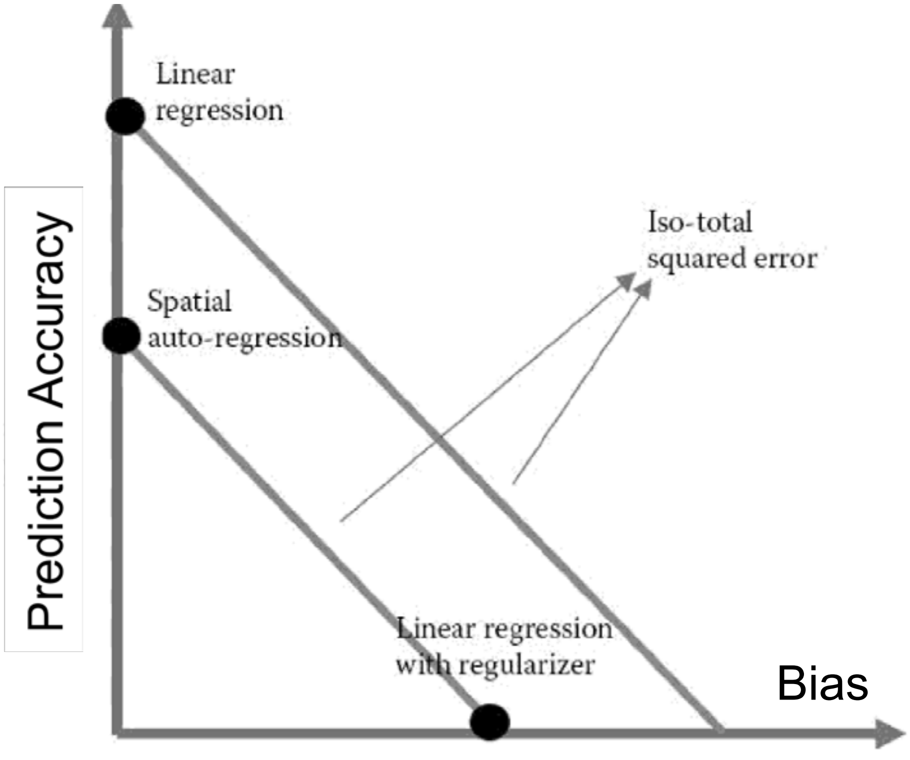
- LR with Auto-correlation Regularizer

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

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

- Spatial Auto-Regression

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

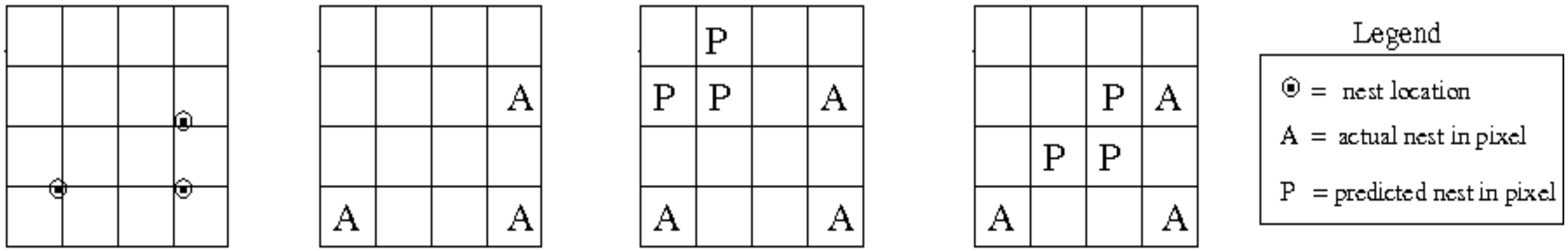


Source: Geospatial Data Science: A Transdisciplinary Approach.
In *Geospatial Data Science Techniques and Applications* (pp. 17-56). CRC Press, 2017
(E. Eftelioglu,R. Ali, X. Tang., Y. Xie, Y., Li and S. Shekhar).

Spatial Computing Research Group

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)



(a) Actual Sites

(b) Pixels with actual sites

(c) Prediction 1

(d) Prediction 2. Spatially more interesting than Prediction 1

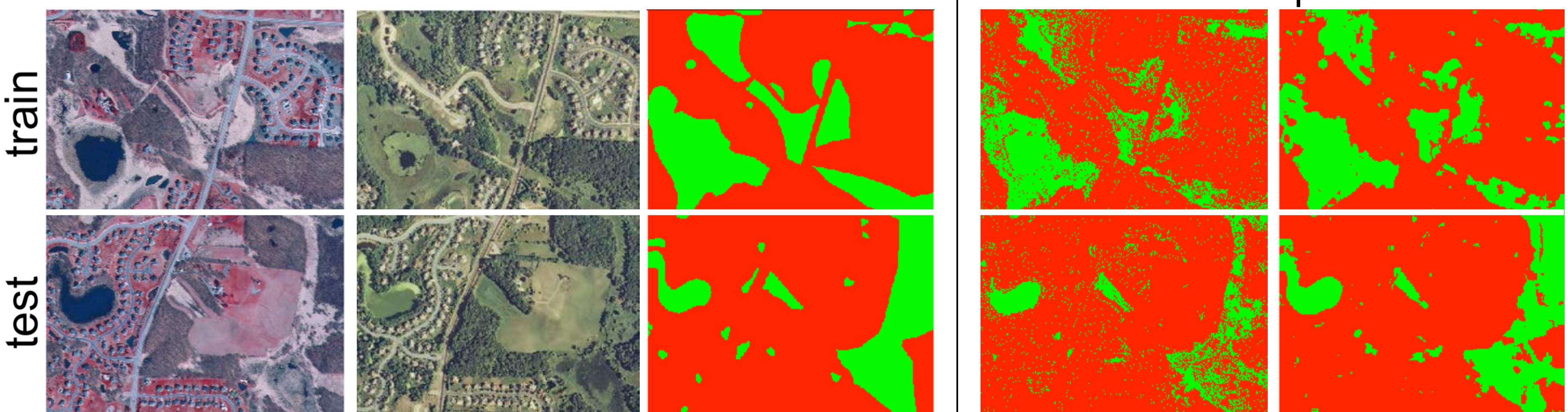
Spatial Computing Research Group

Salt n Pepper Noise

wetland   dry land

Input:

Output:



(a) aerial photo (b) aerial photo (c) true classes (d) DT prediction (Salt n Pepper Noise) (e) SDT prediction

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

DT: decision tree
SDT: spatial decision tree

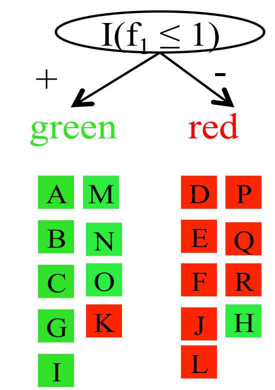
Details: Focal-Test-Based Spatial Decision Tree Learning. IEEE Trans. Knowl. Data Eng. 27(6): 1547-1559, 2015 (summary in Proc. IEEE Intl. Conf. on Data Mining, 2013).(w/ Z. Jiang et al.)

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

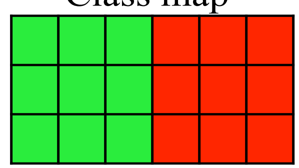
Feature f_1

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

Feature f_2

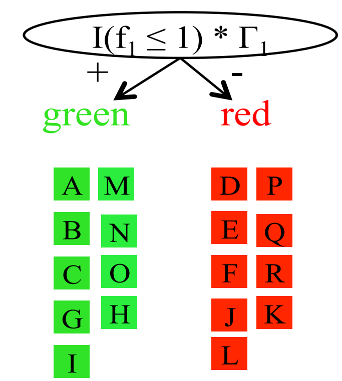
1	1	3	2	2	2
1	1	3	2	2	2
1	1	3	2	2	2

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

Location Prediction Models

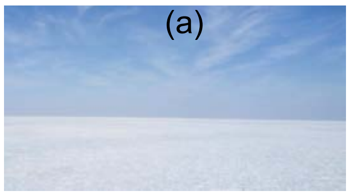
- Traditional Models, e.g., Regression (with Logit or Probit),
  - Linear Regression, Bayes Classifier, …
- Semi-Spatial : auto-correlation regularizer
- Spatial Models
  - Spatial autoregressive model (SAR)
  - Markov random field (MRF) based Bayesian Classifier

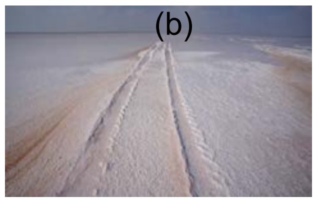
$$\varepsilon = \|y - \beta X\|^2 + \|\beta X - \beta X_{neighbor}\|^2$$

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

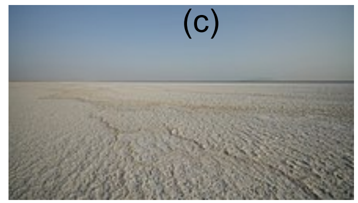
Spatial Variability Challenge: Amorphous Features

Q1. Which images show snow ?

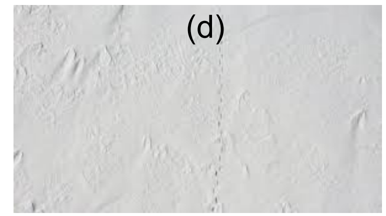




Runn of Kutch, Gujarat, India



Lake Karum, Ethiopia



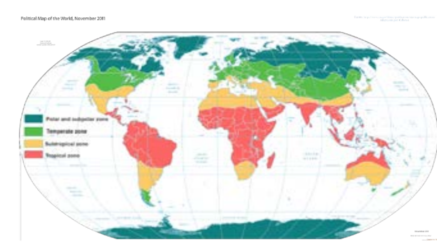
Snow

Snow

Q2. Which geo-challenges are addressed by Convolutional Neural Network (CNN) ?

(a) High Cost of spurious and missed patterns (b) Spatial Auto-correlation

(c) Spatial Heterogeneity (d) Teleconnections

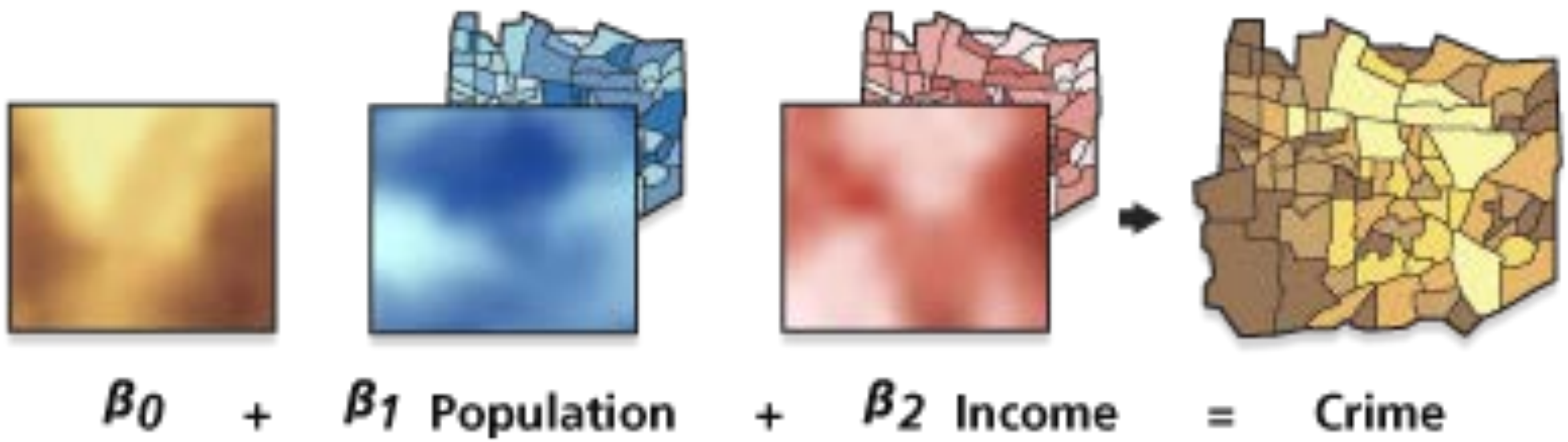


Details: Towards Spatial Variability Aware Deep Neural Networks (SVANN): A Summary of Results, J. Gupta, Y. Xie, and S. Shekhar, DeepSpatial2020 (1st ACM SIGKDD Workshop on Deep Learning for Spatiotemporal Data, Applications, and Systems). Best paper award.

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

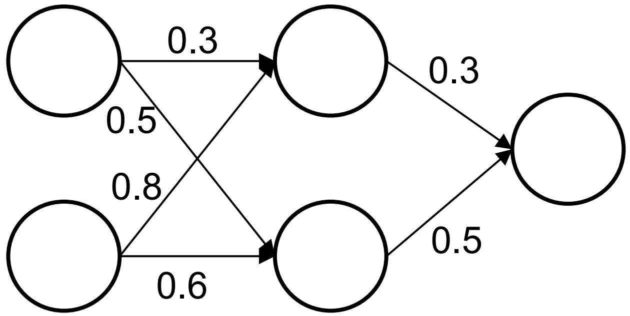


Source: resources.arcgis.com

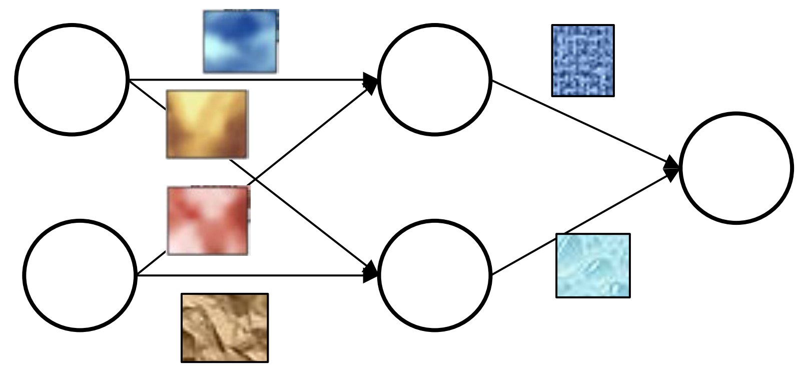
Spatial Computing Research Group

Spatial Variability Aware Neural Networks (SVANN)

A Neural Network (NN)



SVANN



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

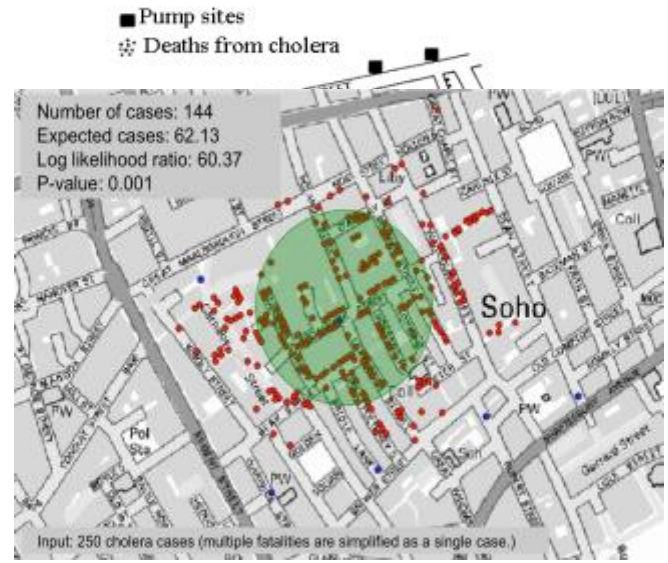
- Evaluation Task:
  - Urban Garden Detection across Hennepin County, MN and Fulton County, GA.
  - SVANN outperformed OSFA by 14.34% on F1-scores.

Details: J. Gupta, Y, Xie and S. Shekhar,
Towards Spatial Variability Aware Deep Neural Networks (SVANN): A Summary of Results,
ACM SIGKDD Workshop on Deep Learning for Spatiotemporal Data, Applications,
and Systems (Deepspatial 2020), 2020. (Best Paper Award). arXiv:2011.08992v1
Full paper accepted for ACM Transaction on Intelligent Systems and Technology.

Dealing with Noise & Spurious Chance Patterns

- Statistics: Deal with Noise
  - Quantify uncertainty, confidence, …
  - Is it (statistically) significant?
  - Is it different from a chance event or rest of dataset?
    - e.g., SaTScan finds circular hot-spots

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



SaTScan™

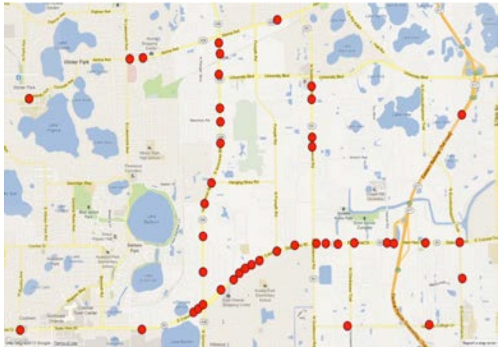
Software for the spatial, temporal, and space-time scan statistics

Spatial Scan Statistics (SatScan)

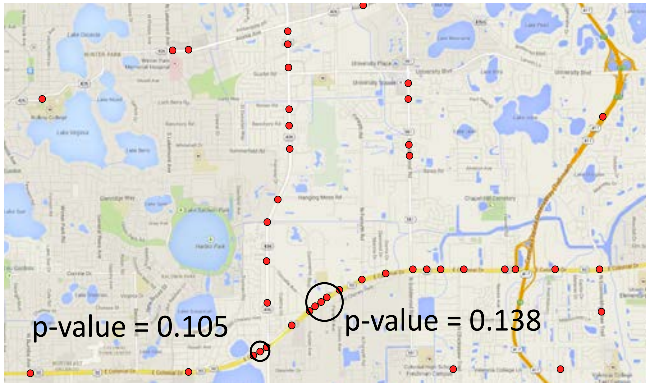
- Goal: Omit chance clusters
- Ideas: Likelihood Ratio, Statistical Significance
- Steps
  - Enumerate candidate zones & choose zone X with highest likelihood ratio (LR)
    - LR(X) = p(H1|data) / p(H0|data)
    - H0: points in zone X show complete spatial randomness (CSR)
    - H1: points in zone X are clustered
  - If LR(Z) >> 1 then test statistical significance
    - Check how often is LR(CSR) > LR(Z) using 1000 Monte Carlo simulations

Beyond SatScan: Spatial Concept/Theory-Aware Hotspots

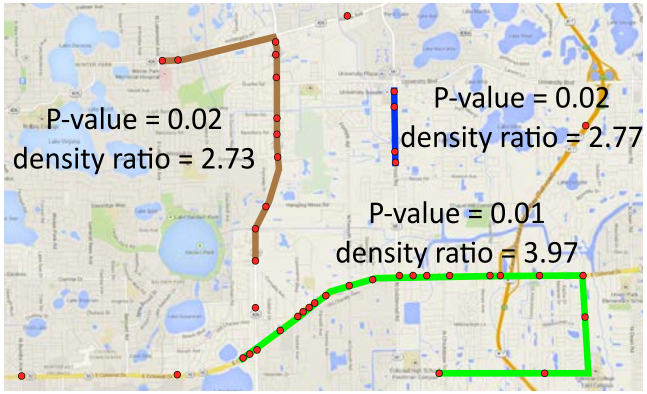
- Geographic features, e.g., rivers, streams, roads, …
  - Hot-spots => Hot Geographic-features, e.g., Linear Hotspots
- Spatial Theories, e.g,, environmental criminology
  - Circles ➔ Doughnut holes



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.)

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.

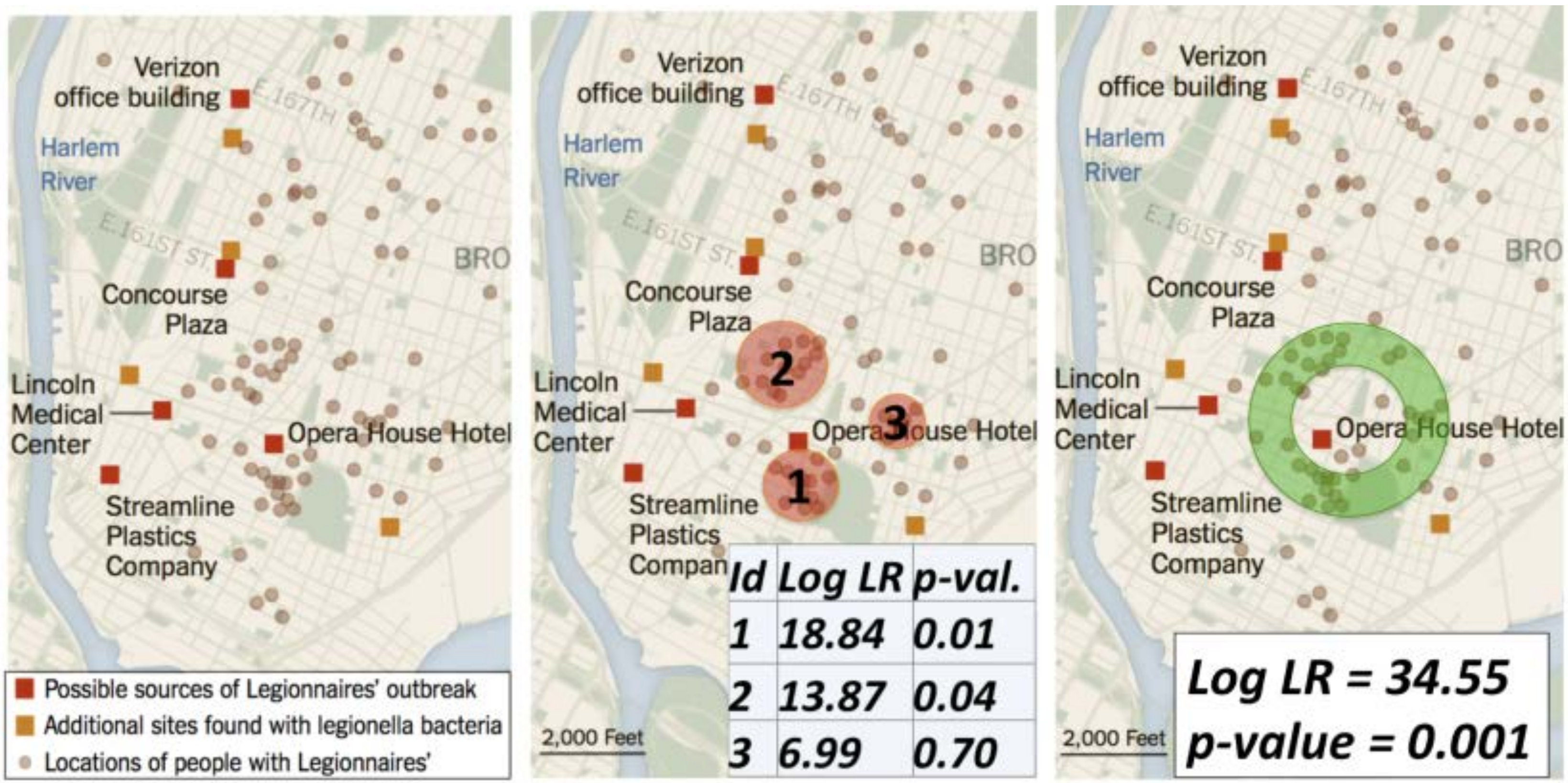
- Possible sources of Legionnaires' outbreak
- Additional sites found with legionella bacteria
- Locations of people with Legionnaires'

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



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

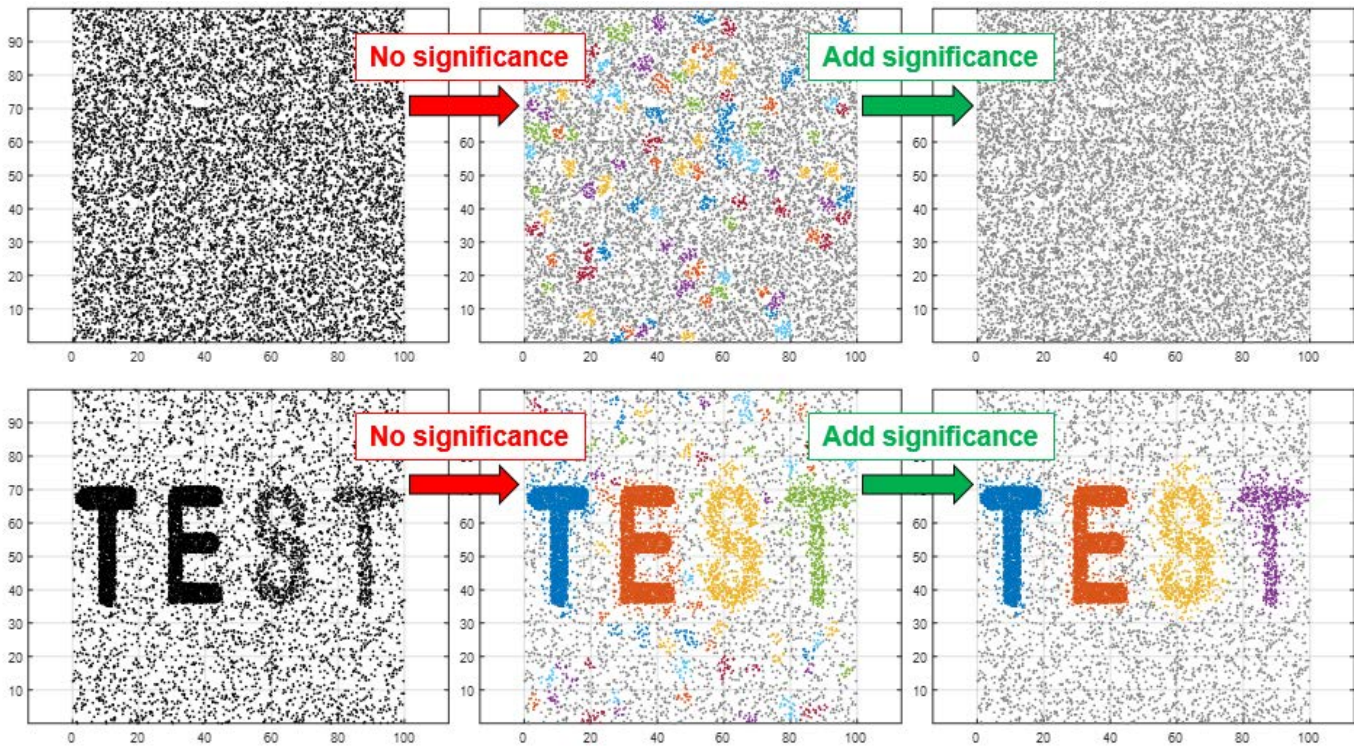
<u>Details: Ring-Shaped Hotspot Detection,</u> IEEE Trans. Know. & Data Eng., 28(12), 2016.
(A Summary in Proc. IEEE ICDM 2014) (w/ E. Eftelioglu et al.)

Robust Clustering (Hotspot Detection)

- **Problem definition**
  - **Inputs:** Collection of event locations, Test statistic; Significance level
  - **Output:** Significant clusters (hotspots)
  - **Constraints:** Avoid chance patterns despite non-trivial noise in data
- **Limitations of Related Work**
  - DBSCAN cannot avoid chance patterns
  - SaTScan cannot detect clusters of arbitrary shapes
- **Contributions**
  - Significance modeling in DBSCAN
  - A fast dual-convergence algorithm

Complete Spatial Random (no significant hotspots)

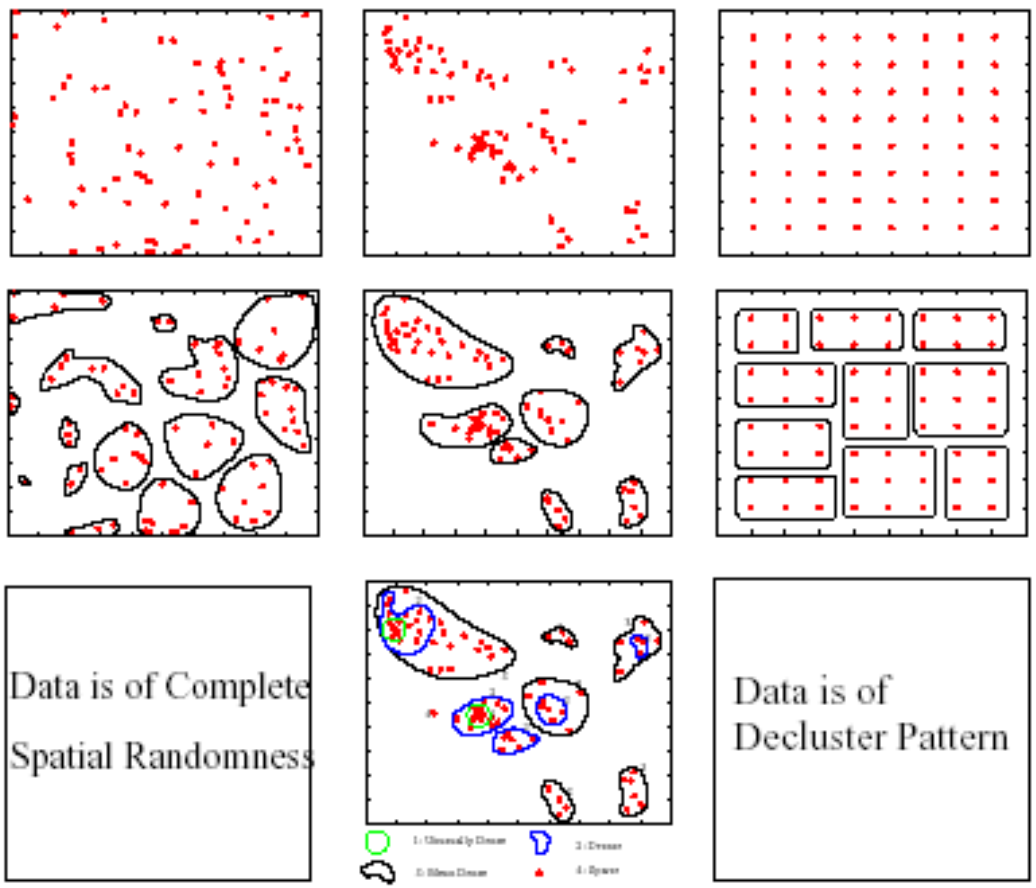
Significant Hotspots with Noise



Details: Significant DBSCAN towards Statistically Robust Clustering, (w/ Yiqun Xie),
In Proc. 16th Intl. Symposium on Spatial and Temporal Databases (SSTD), 2019, ACM. **(Best Paper Award)**

Limitation of Traditional Clustering

- Challenge: One size does not fit all
  - Prediction error vs. model bias, Cost of false positives, …
- Example. Clustering: Find groups of points



Traditional Clustering
(K-means always finds clusters)

Spatial Clustering begs to differ!

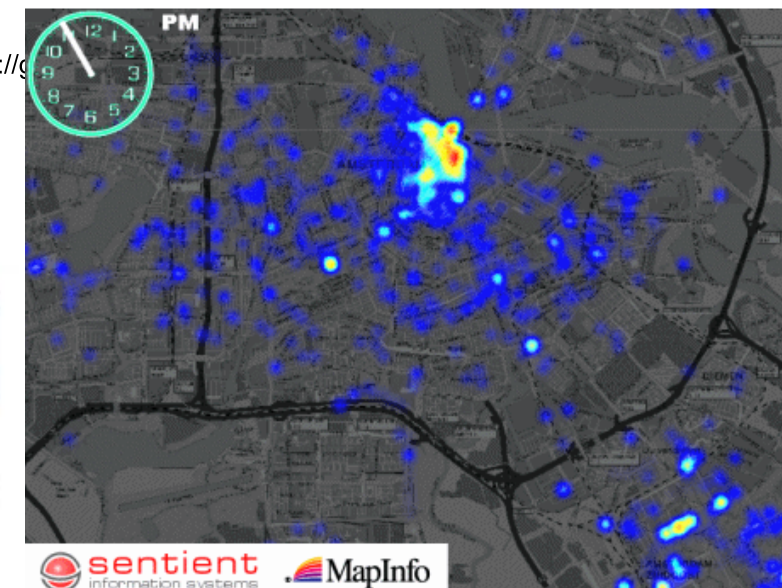
What has changed? **Spatial Data Revolution**

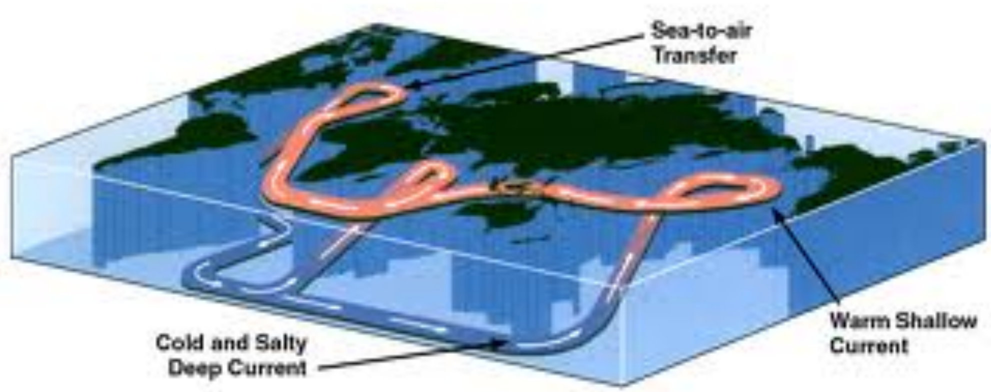
Spatial	Last Century	Last Decade
Data	Few satellites and sensors	Nano-satellites, Billions of GPS enabled smartphones
Data Access	Need special hardware and network	Cloud based repositories, e.g., Earth on AWS
Spatial Platforms	ESRI Arc/Info	SQL3/OGC, e.g., Postgis, ESRI GIS Tools for Hadoop, Google Earth Engine
Spatial Data Science	Spatial Patterns, e.g., hotspots (SatScan, ESRI Geostatistics)	(a) Spatial Network Patterns, e.g., linear hotspots (b) Spatio-temporal (ST) patterns, e.g., Change time-series (Google Timelapse)
Spatial Visualization	Quilt: MS Terraserver Fly through: Google Earth	(a) Space time: Timelapse (b) There Dimensions

Towards Time-Travel and Depth in Virtual Globes

- Virtual globes are snapshots
- How to add time? depth?
  - Ex. Google Earth Engine, NASA NEX
  - Ex. Google Timelapse: 260,000 CPU core-hours for global 29-frame video
- How may one convey provenance, accuracy, age, and data semantics?
- What techniques are needed to integrate and reason about diverse available

http://g html





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

Data that are geographically referenced or contain some type of location markers are both common and of high value (e.g., data subject to state-specific policies, laws and regulations; demographic data from the census; location traces of smartphones and vehicles; remotely sensed imagery from satellites, aircraft and small unmanned aerial vehicles; volunteered geographic information; geographically referenced social media postings). A 2011 McKinsey Global Institute report estimates a value of "about $600 billion annually by 2020" from leveraging personal location data[2] to reduce fuel waste, improve health outcomes, and better match products to consumer needs. Spatial data are critical for societal priorities such as national security, public health & safety, food, energy, water, smart cities, transportation, climate, weather, and the environment. For example, remotely-sensed satellite imagery is used to monitor not only weather and climate but also global crops[3] for early warnings and planning to avoid food shortages.

One Size Data Science Does not Fit All Data!

However, spatial data presents unique data science challenges. Recent court cases that address gerrymandering, the manipulation of geographic boundaries to favor a political party, offer a high-profile example. Instances of such exploitation of the modifiable areal unit problem (or dilemma) is not limited to elections since the MAUP affects almost all traditional data science methods in which results (e.g., correlations) change dramatically by varying geographic boundaries of spatial partitions. The fundamental geographic qualities of spatial autocorrelation, which assumes properties of geographically proximate places to be similar, and geographic heterogeneity, where no two places on Earth are exactly alike, violate assumptions of sample independence and randomness that underlie many conventional statistical methods. Other spatial challenges include how to choose between a plurality of projections and coordinate systems and how to deal with the imprecision, inaccuracy, and uncertainty of location

A UCGIS Call to Action:

Bringing the Geospatial Perspective to Data Science Degrees and Curricula

University Consortium for
GEOGRAPHIC INFORMATION SCIENCE

Summer 2018

Spatial Data Science Tools

measurements. To deal with such challenges, practitioners in many fields including agriculture, weather forecast, mining, and environmental science incorporate *geospatial data science*[4] methods such as spatially-explicit models, spatial statistics[5], geo-statistics, geographic data mining[6], spatial databases[7], etc.

[4] Y. Xie et al., Transdisciplinary Foundations of Geospatial Data Science, *ISPRS Intl. Jr. of Geo-Informatics*, 6(12):395-418, 2017. DOI: 10.3390/ijgi6120395.

[5] N. Cressie, *Statistics for Spatial Data*, Wiley, 1993 (1st ed.), 2015 (Revised ed.).

[6] H. Miller and J. Han, *Geographic Data Mining and Knowledge Discovery*, CRC Press, 2009 (2nd Ed.).

[7] S. Shekhar and S. Chawla, *Spatial Databases: A Tour*, Prentice Hall, 2003.

Summary : One size data science does not fit all

- Spatial Data are ubiquitous & important
- Traditional Data Science Tools are inadequate
  - Gerrymandering, Spatial Auto-correlation, …
- **Ask:**
  - Spatial Data Science Methods
  - Spatial Statistics, Spatial Data Mining, SDBMS, …

References :Surveys, Overviews

- Spatial Computing, MIT Press (Essential Knowledge Series), 2020. (ISBN: 9780262538046).
- 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.

C3. Auto-correlation and Heterogeneity in Prediction

Traditional	Spatial Autocorrelation	Spatial Heterogeneity
Linear Regression	Spatial Auto-Regression	GWR
Bayesian Classifier	Neighborhood Based Bayesian Classifier	
Decision Trees	Spatial Decision Trees	Spatial Ensemble
Neural Networks	Convolutional Neural Networks	SVANN