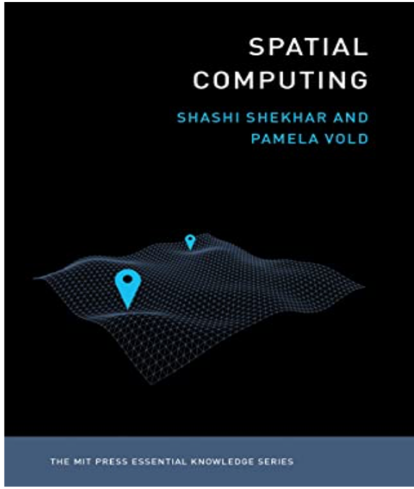
Panel on Future of AI for Spatio-Temporal Data Science

1st ACM SIGKDD Workshop on Deep Learning for Spatiotemporal Data, Applications, and Systems (2020 DeepSpatial)

What's Special About Spatial?


SPATIAL COMPUTING
SHASHI SHEKHAR AND PAMELA VOLD
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

Shashi Shekhar

Member, Computing Research Association Board

McKnight Distinguished University Professor, Univ. of Minnesota

www.cs.umn.edu/~shekhar

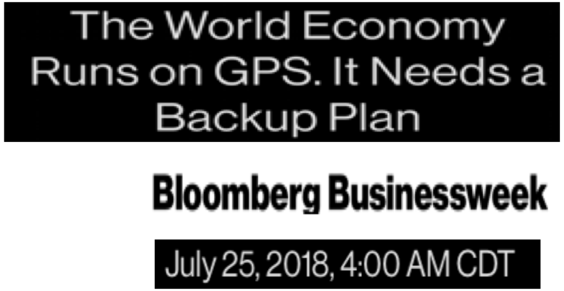
Acks.: NSF, USDOD-NGA, USDOE-ARPA-E, USDA

UNIVERSITY OF MINNESOTA

Driven to Discover®

Spatial Revolution

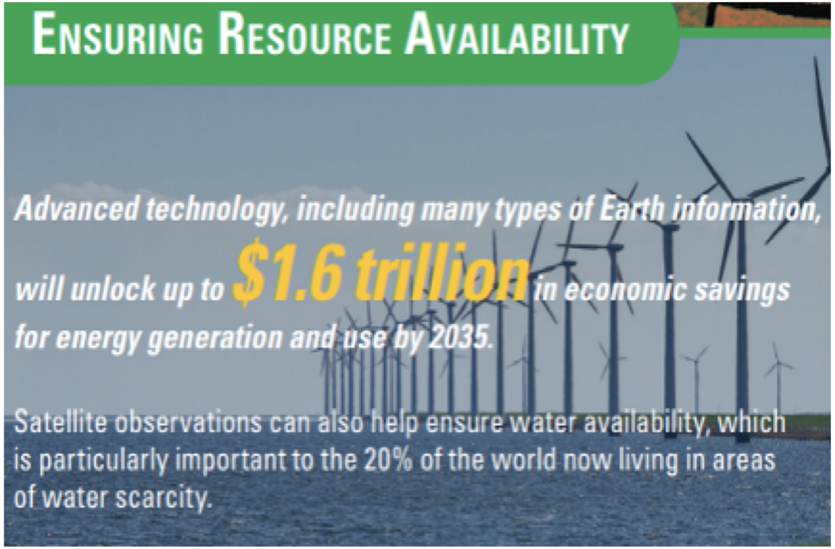
- GPS & Location traces
  - 2 billion GPS receivers today (7 billion by 2022)
  - Reference clock for telecom, banks, …
  - Help understand Spatio-temporal patterns of life
- (Nano-)Satellite Imagery, …



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

Bloomberg Businessweek

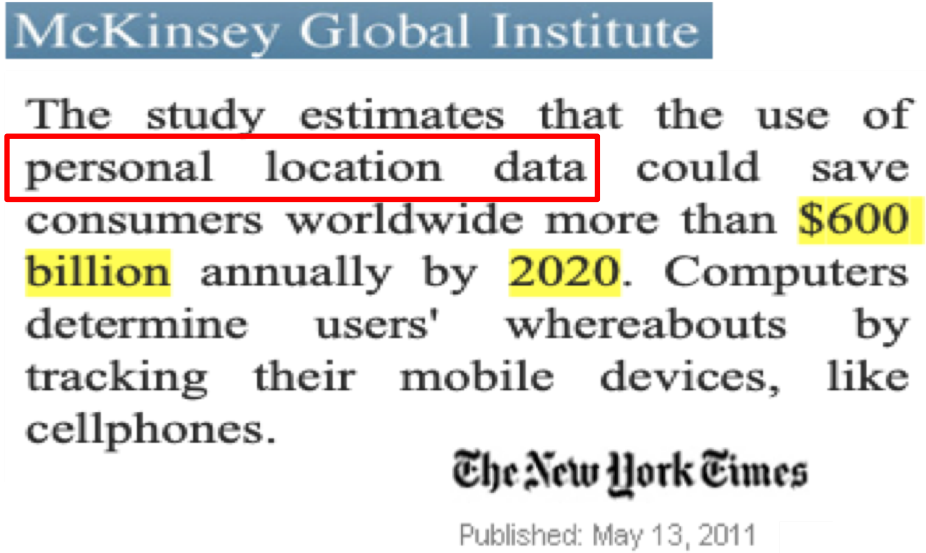
July 25, 2018, 4:00 AM CDT



ENSURING RESOURCE AVAILABILITY

Advanced technology, including many types of Earth information, will unlock up to $1.6 trillion in economic savings for energy generation and use by 2035.

Satellite observations can also help ensure water availability, which is particularly important to the 20% of the world now living in areas of water scarcity.



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

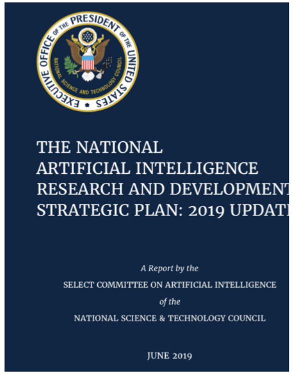
Source: Y. Xie et al., Transforming Smart Cities With Spatial Computing, Proc. IEEE Intl. Conf. on Smart Cities, 2018.

AI promise for Spatial Problems

- From Satellite Imagery: Classify Land-cover, Map buildings
- Inverse Problems: Find geo-location given picture(s)
- Content based Querying
  - 2009 Haiti Earthquake: Map building damage
  - 2017: DARPA Geospatial Cloud Analytics
    - food shortages, fracking, illegal fishing vessels
  - 2020: DARPA
    - Spatiotemporal: Construction & classify Stage
    - Map Underground: (Subterranean Challenge)
- But many hurdles (Machine is still learning!)

.

Sources: 1. DARPA(https://www.darpa.mil/attachments/DARPA-2019-framework.pdf)
2. J. Aardt et al., Geospatial Disaster Response during the Haiti Earthquake: A Case Study Spanning Airborne Deployment, Data Collection, Transfer, Processing, and Dissemination, Photogrammetric Eng. & Remote Sensing, 77(9):943-952, Sept. 2011.



THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE

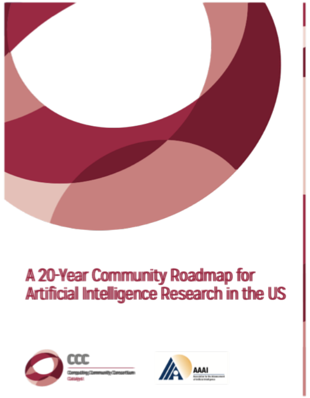
A Report by the

SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE

of the

NATIONAL SCIENCE & TECHNOLOGY COUNCIL

JUNE 2019



A 20-Year Community Roadmap for Artificial Intelligence Research in the US

Advancing data-focused methodologies for knowledge discovery

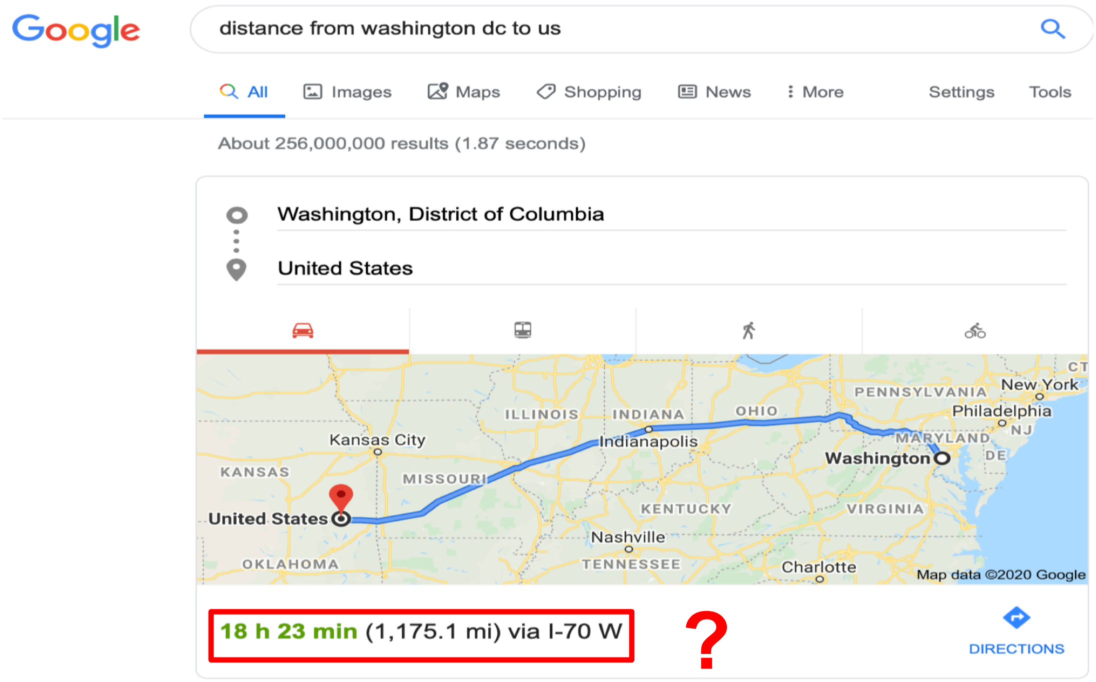
As discussed in the 2016 *Federal Big Data Research and Development Strategic Plan*,[34] many fundamental new tools and technologies are needed to achieve intelligent data understanding and knowledge discovery. Further progress is needed in the development of more advanced machine learning algorithms that can identify all the useful information hidden in big data. Many open research questions revolve around the creation and use of data, including its veracity and appropriateness for AI system training. The veracity of data is particularly challenging when dealing with vast amounts of data, making it difficult for humans to assess and extract knowledge from it. While much research has dealt with veracity through data quality assurance methods to perform data cleaning and knowledge discovery, further study is needed to improve the efficiency of data cleaning techniques, to create methods for discovering inconsistencies and anomalies in the data, and to develop approaches for incorporating human feedback. Researchers need to explore new methods to enable data and associated metadata to be mined simultaneously.

Many AI applications are interdisciplinary in nature and make use of heterogeneous data. Further investigation of multimodality machine learning is needed to enable knowledge discovery from a wide variety of different types of data (e.g., discrete, continuous, text, spatial, temporal, spatio-temporal, graphs). AI investigators must determine the amount of data needed for training and to properly address large-scale versus long-tail data needs. They must also determine how to identify and process rare events beyond purely statistical approaches; to work with knowledge sources (i.e., any type of information that explains the world, such as knowledge of the law of gravity or of social norms) as well as data sources, integrating models and ontologies in the learning process; and to obtain effective learning performance with little data when big data sources may not be available.

Challenge 1: 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!



distance from washington dc to us

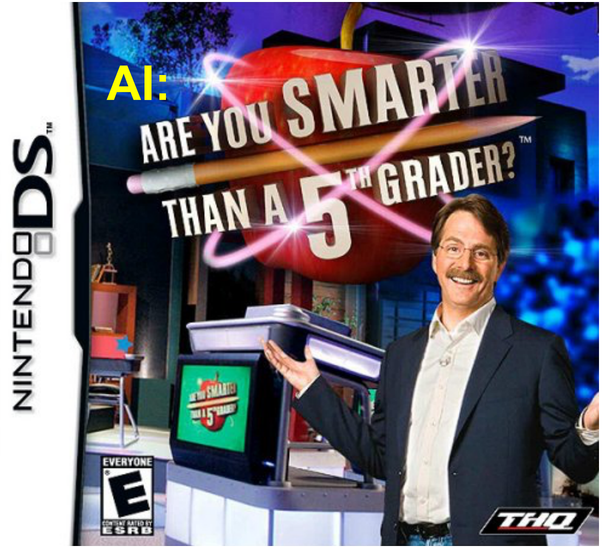
About 256,000,000 results (1.87 seconds)

Washington, District of Columbia

United States

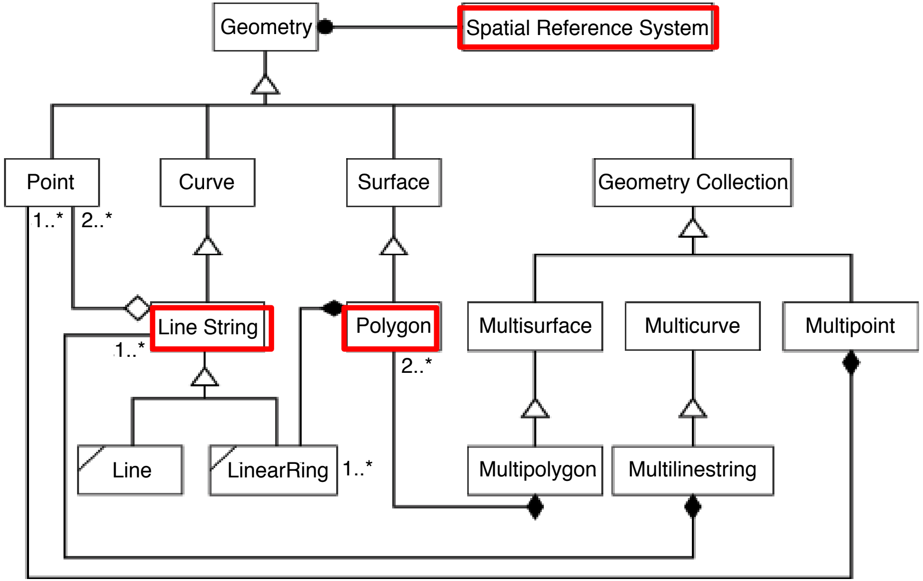
18 h 23 min (1,175.1 mi) via I-70 W

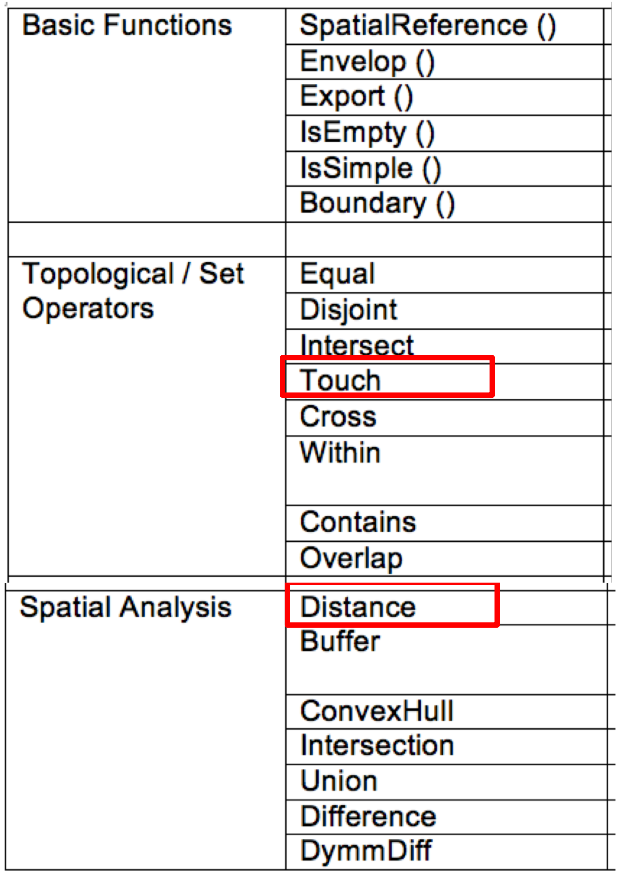
?



Spatial Data Types: OGC Simple Features Standard

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



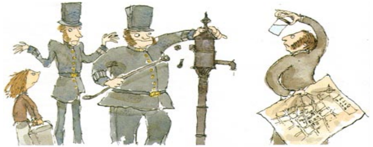


Basic Functions	SpatialReference ()
	Envelop ()
	Export ()
	IsEmpty ()
	IsSimple ()
	Boundary ()
Topological / Set Operators	Equal
	Disjoint
	Intersect
	Touch
	Cross
	Within
	Contains
	Overlap
Spatial Analysis	Distance
	Buffer
	ConvexHull
	Intersection
	Union
	Difference
	DymmDiff

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

Challenge 2: Spatial Data Science >> Patterns

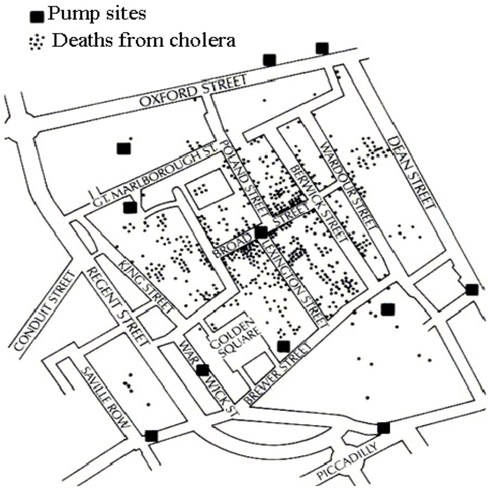
1854: What causes Cholera?



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

? water pump

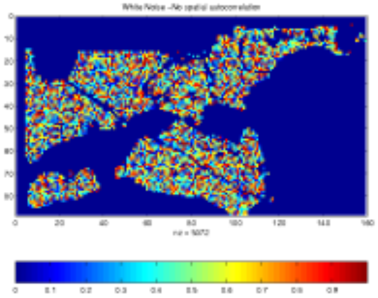
Remove pump handle

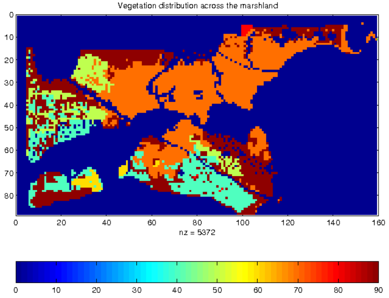


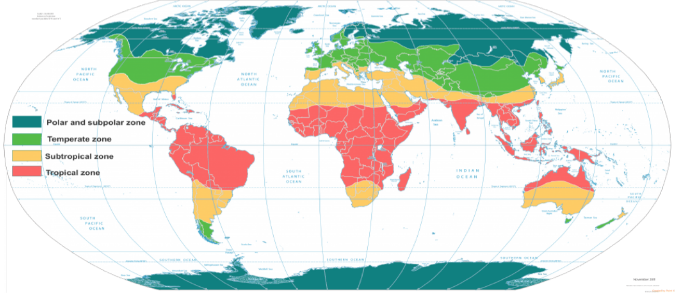
Societal Impact:
Sewage system,
Drinking water supply,
Lower urban density (parks),
…
…

Challenge 3: Spatial Patterns >> Traditional Machine Learning

- **A. High cost of missed or spurious patterns**
  - Pr.[Self-driving car fail to detect a red traffic light] > 0
  - Loss of life, stigmatization, economic loss
- **B. Gerrymandering: Spatial partitioning choice alters results**
- **C. Spatial data violates ubiquitous assumptions**
  - Data samples: independent and identically distributed (i.i.d)
  - Nearby spatial data samples are not independent
  - No two places on Earth are exactly alike!



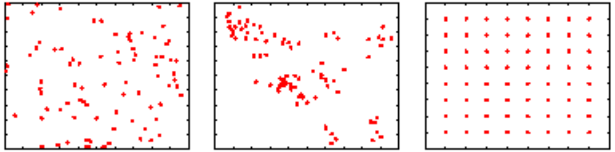


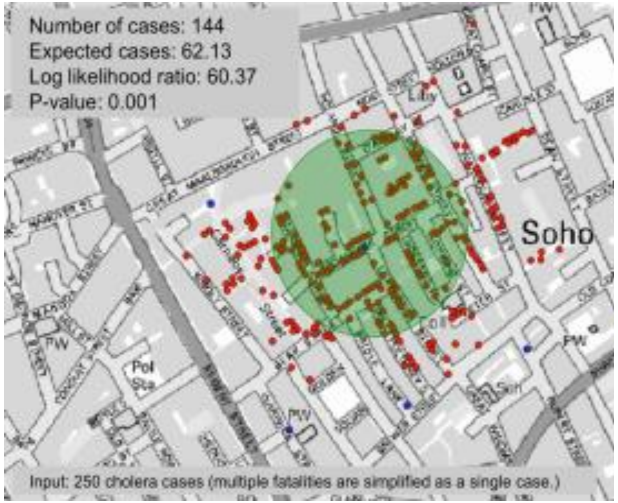


Source: Spatiotemporal Data Mining: A Computational Perspective , ISPRS Intl. Jr. Geo-Info., 4(4):2306-2338, 2015

3A. An Approach to Reduce Spurious Patterns

- SatScan (National Cancer Institute)
  - Compare with complete spatial random
  - Monte Carlo simulation

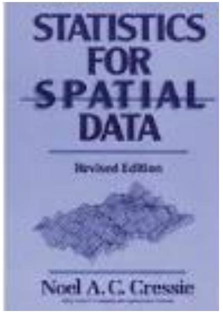


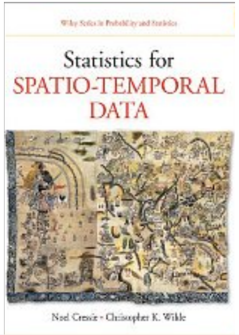


Number of cases: 144
Expected cases: 62.13
Log likelihood ratio: 60.37
P-value: 0.001

Input: 250 cholera cases (multiple fatalities are simplified as a single case.)

- Spatial Statistics
  - Quantify uncertainty, confidence, …
  - Model Auto-correlation, heterogeneity, …





Statistics for SPATIO-TEMPORAL DATA
Noel Cressie · Christopher K. Wikle

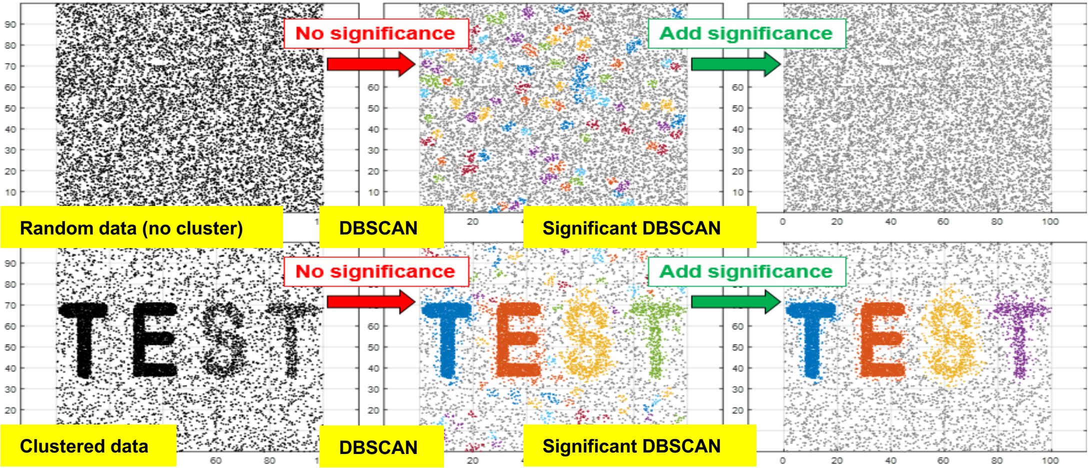
SaTScan™

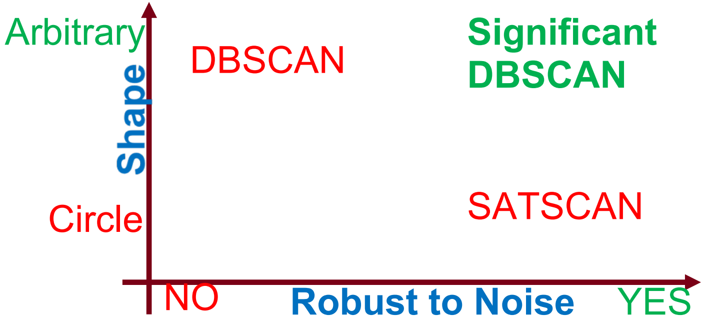
Software for the spatial, temporal, and space-time scan statistics

Source: https://www.satscan.org

3A. Reducing Spurious Arbitrary Shape Hotspots

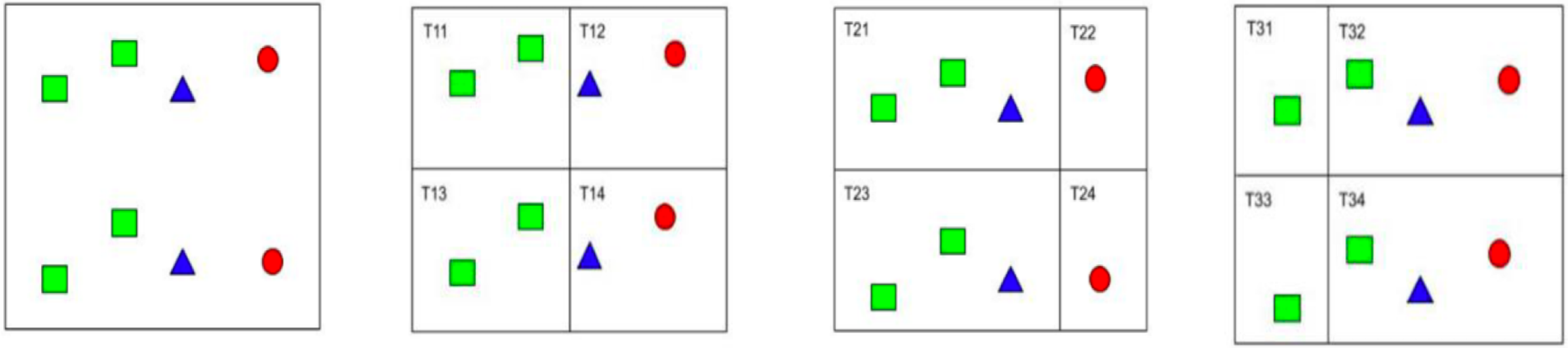
- Q? How to detect Statistically Significant Arbitrary Shape Hotspots?
- **Significant DBSCAN** [SSTD 2019]
  - Significance modeling in DBSCAN + A fast dual-convergence algorithm





Details: Significant DBSCAN towards Statistically Robust Clustering, Y. Xie & S. Shekhar, Proc. 16th Intl. Symp. on Spatial and Temporal Databases (SSTD '19), 2019, ACM **(Best Paper).**

3B. Gerrymandering Risk in Traditional Association Rules

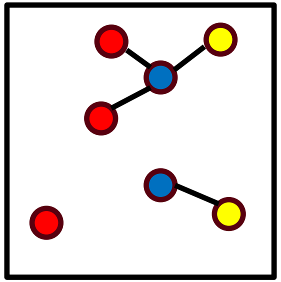


(a) Map of 3 item-types (b) Spatial Partition P1 (c) Spatial Partition P2 (d) Spatial Partition P3

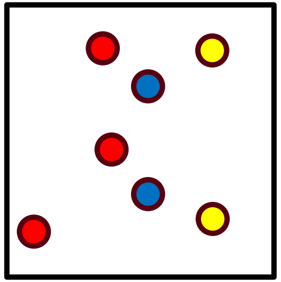
Partitioning	P1	P2	P3
Transactions	T11, T12, T13, T14	T21, T22, T23, T24	T31, T32, T33, T44
Associations with support >= 0.5	(▲ 🔴)	(🟩 ▲)	(🟩 ▲ 🔴)

Details: Data Science for Earth: The Earth Day Report , E. Eftelioglu, et al., ACM SIGKDD Explorations Newsletter, 22(1), May 2020.

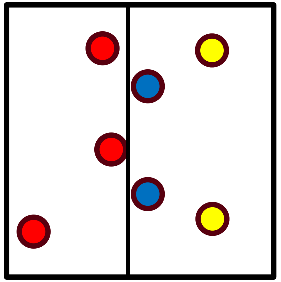
3B. Neighbor Graph to Reduce Gerrymandering Risks



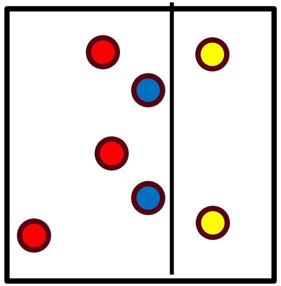
(d) Neighbor graph



(a) a map



(b) Partition A



(c) Partition C

Participation Index	Ripley’s Cross-K	Pattern	Pearson Correlation	Pearson Correlation
0.5	0.33	🔴🔵	(-) 0.9	1
1	1	🟡🔵	1	(-) 0.9

3C. Modeling Auto-correlation in Prediction Models

- Traditional Models
  - Linear Regression (e.g., Logit), Bayes Classifier, Neural Networks, Decision Trees
- Semi-Spatial : auto-correlation in regularizer
- Spatial Models
  - W = neighbor matrix (row-normalized)
  - Spatial autoregressive model (SAR), …

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

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

Prediction Error and Bias Trade-off

- Linear Regression (LR): Least Squares estimator

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

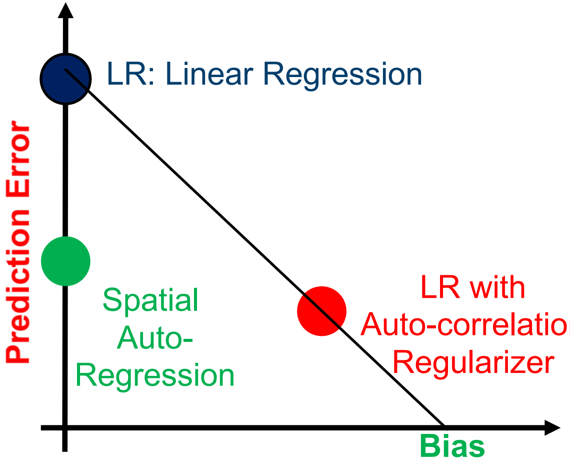
- LR with Auto-correlation Regularizor
  - Least squares estimator

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

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

- Spatial Auto-Regression:
  - Maximum Likelihood Estimator

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



Source: Y. Xie et al., Transdisciplinary Foundations of Geospatial Data Science (html , pdf)
ISPRS Intl. Jr. of Geo-Informatics, 6(12), 2017. doi:10.3390/ijgi6120395.

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>**

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

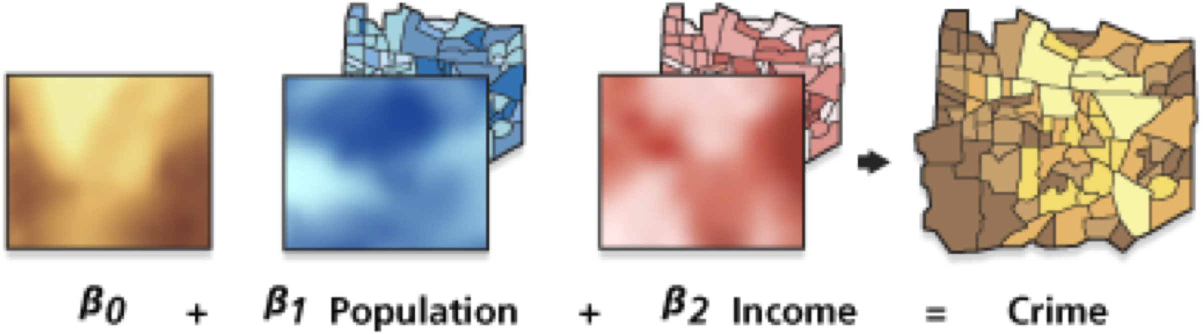
- Computing determinant of large matrix is a hard (open) problem!
  - size(W) is quadratic in number of locations/pixels.
  - Typical raster image has Millions of pixels
  - W is sparse but not banded.

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)

3C. Modeling Spatial Heterogeneity: GWR, SVANN

- Geographically Weighted Regression (GWR) $y = X\beta' + \varepsilon'$

  Where β' and ε' are location dependent



Source:
resources.arcgis.com

- Spatial Variability Aware Neural Networks (SVANN)
  - Each Neural network parameter is a map (location dependent) not scalar

Details: Towards Spatial Variability Aware Deep Neural Networks (SVANN): A Summary of Results, J. Gupta et al., Proc.1st ACM SIGKDD Workshop on Deep Learning for Spatiotemporal Data, Applications, and Systems , 2020 (Best Paper).

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

Summary

The World Economy Runs on GPS.

- Spatial Data has already transformed our society
  - It is only a beginning!
  - It promises astonishing opportunities in coming decade

- AI (e.g., ML) has promise but faces major challenges
  - – Rich Data Types, e.g., lineStrings, polygons, …
  - – High cost of errors, Gerrymandering, Spatial Auto-correlation, .

- **Ask**
  - – Research Sponsors: Nurture approaches to overcome challenge
  - – Acdemics: Include Spatial topics in courses and curricula