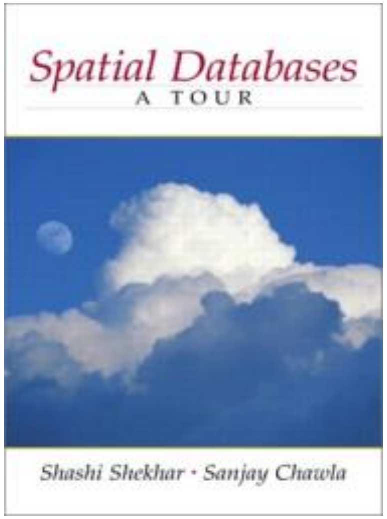
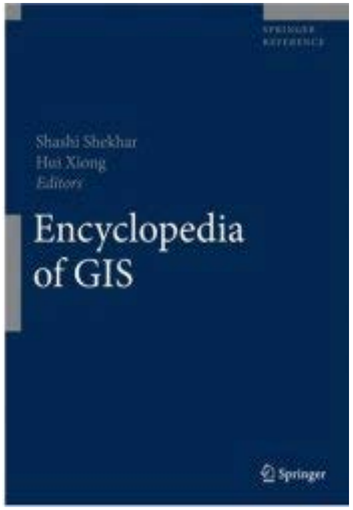
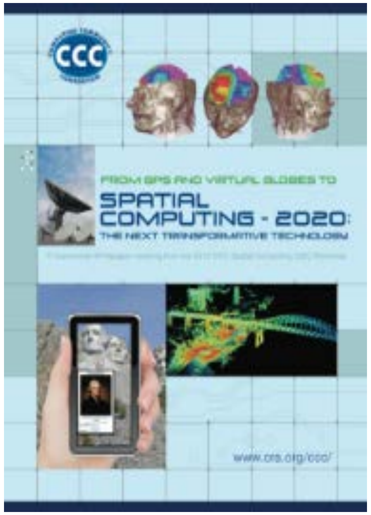
One Size Data Science Does Not Fit All Data: What is Special about Spatial Data Science?

Workshop on Data Science Innovation with NSF Big Data Hubs
Microsoft Research, Seattle, WA. Oct. 29th-30th, 2018

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






CCC
FROM GPS AND VIRTUAL GLOBES TO
SPATIAL COMPUTING - 2020:
THE NEXT TRANSFORMATIVE TECHNOLOGY
www.cra.org/ccc/

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.

University Consortium for
GEOGRAPHIC INFORMATION SCIENCE

Summer 2018

Spatial Computing Examples

Arc ESRI GIS™ X

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

OGC™
Open Geospatial Consortium, Inc.

ORACLE®
SPATIAL

IBM Smarter Planet

OpenStreetMap

Ushahidi

waze
Outsmarting Traffic, Together.

Earth on AWS

Build planetary-scale applications in the cloud with open geospatial data.

Google Earth Engine

MONSANTO®

THE CLIMATE CORPORATION

The Changing World of Spatial Computing

	Last Century	Last Decade
Map User	Well-trained few	Billions
Mappers	Well-trained few	Billions
Software, Hardware	Few layers, e.g., Applications: Arc/GIS, Databases: SQL3/OGIS	Almost all layers
User Expectations & Risks	Modest	Many use-case & Geo-privacy concerns

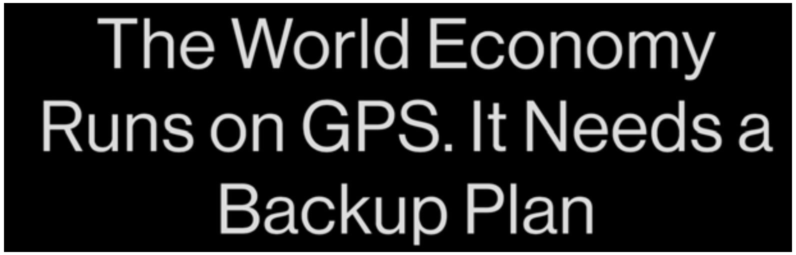
Spatial Computing is a Critical Infrastructure Today!

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

- GPS is the single point of failure for the entire modern economy.
- 50,000 incidents of deliberate (GPS) jamming last two years
  - Against Ubers, Waymo’s self-driving cars, delivery drones from Amazon

Bloomberg Businessweek

July 25, 2018, 4:00 AM CDT

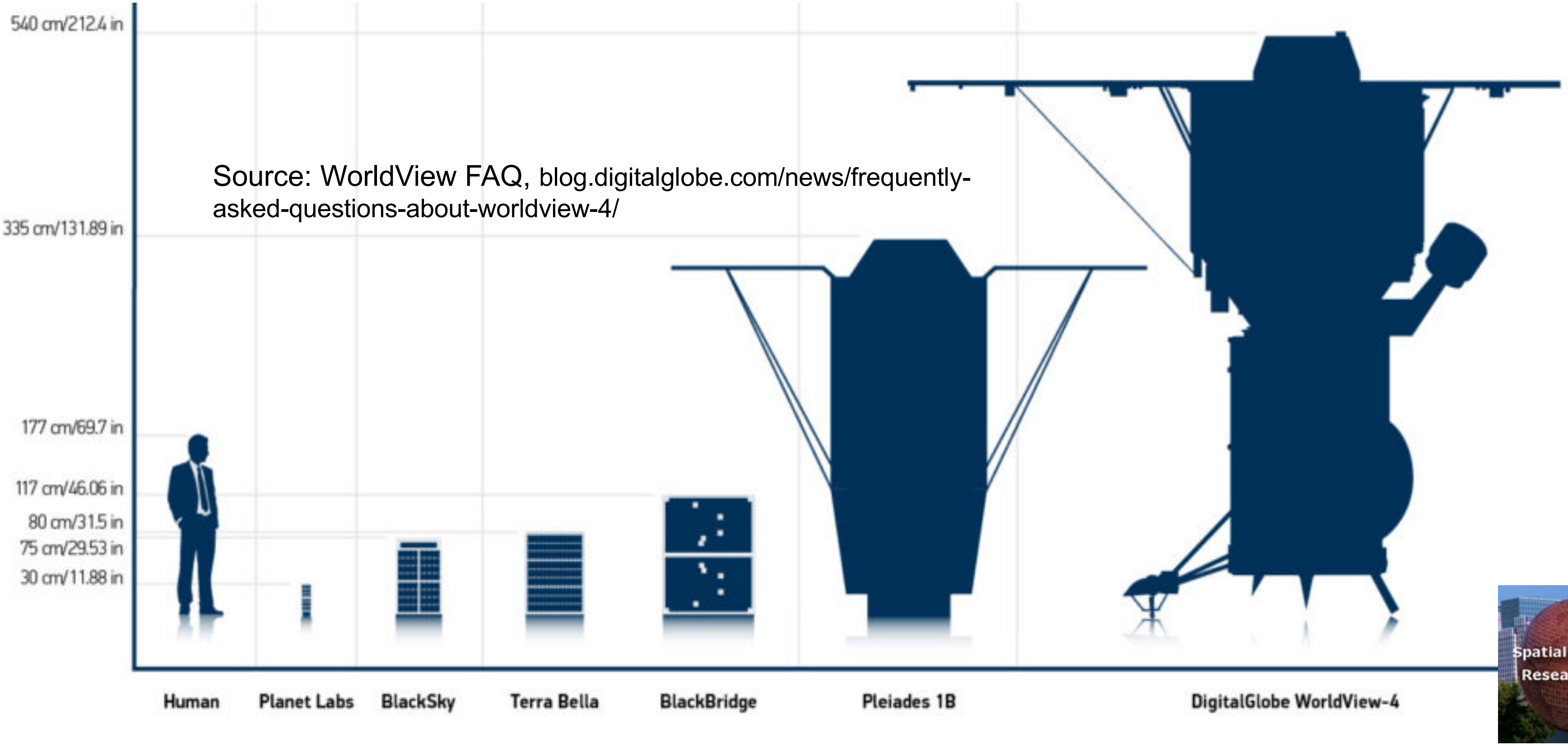


The World Economy Runs on GPS. It Needs a Backup Plan

Source: *https://www.bloomberg.com/news/features/2018-07-25/the-world-economy-runs-on-gps-it-needs-a-backup-plan*

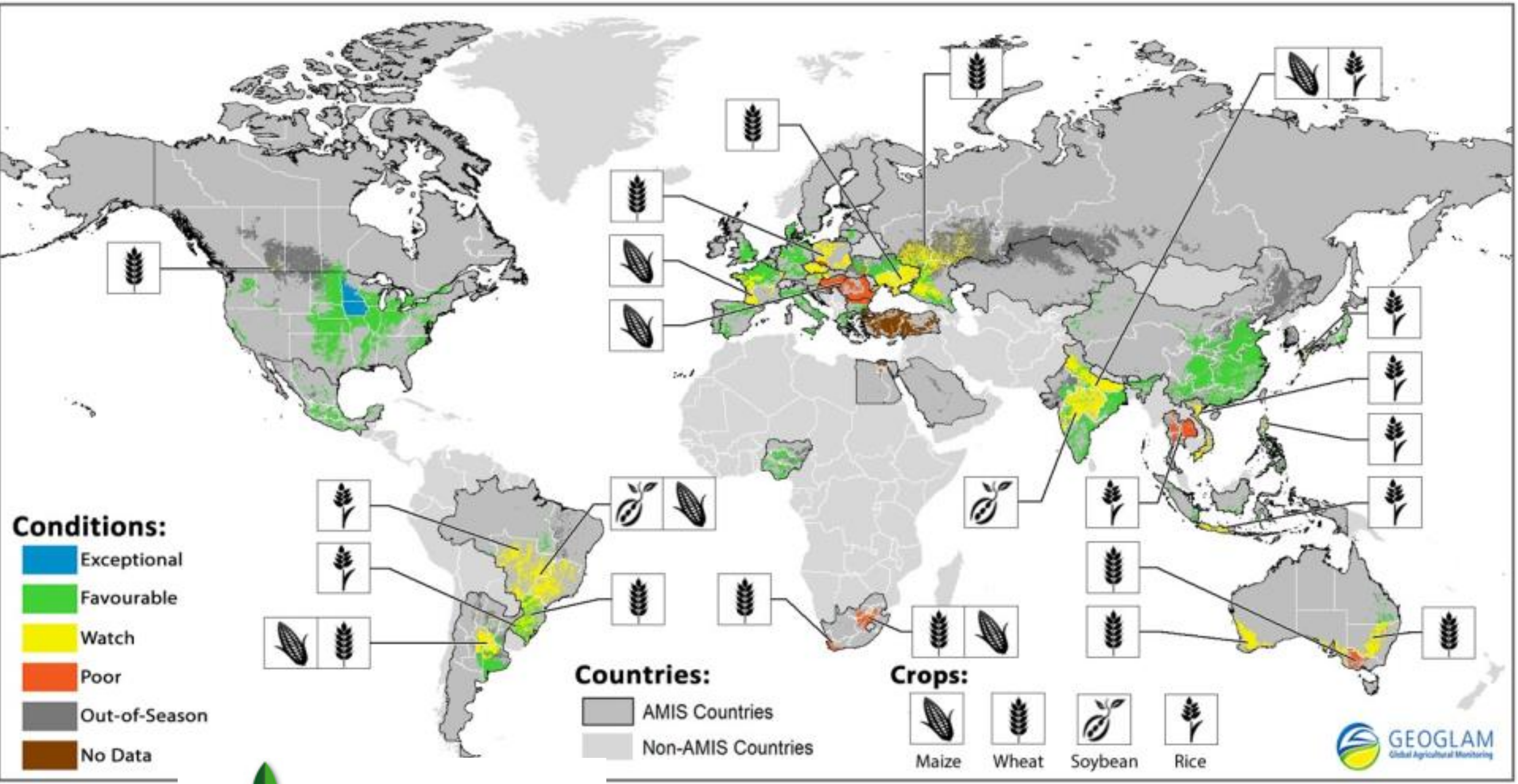
Large Constellations of Small Satellites

- Hi-frequency (e.g., daily or hourly) time-series of imagery of entire earth
  - Monitor illegal fishing, forest fires, crops (2017 DARPA Geospatial Cloud Analytics)
- 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/

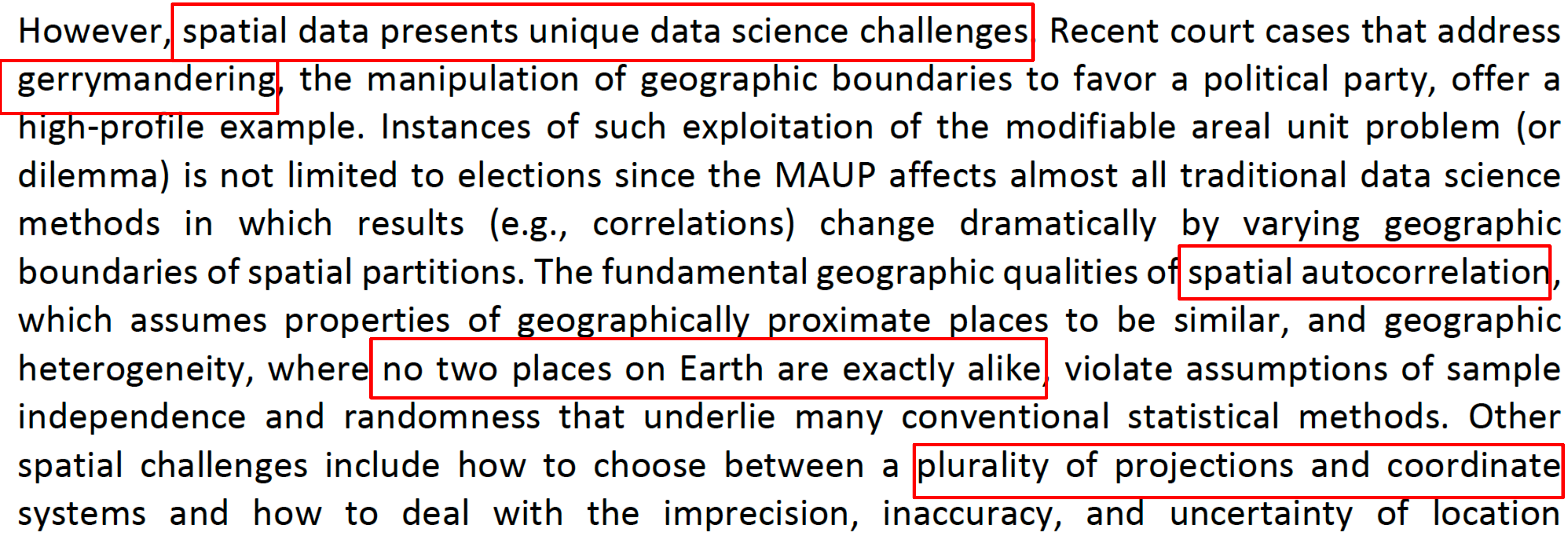
Global Agriculture Monitoring



AMIS Agricultural Market Information System

GEOGLAM Global Agricultural Monitoring

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:

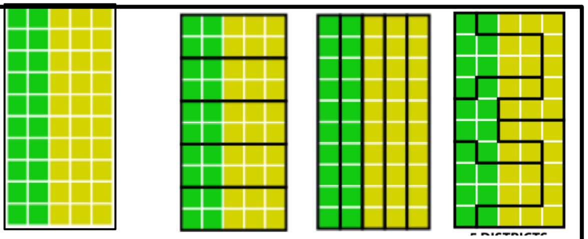
Bringing the Geospatial Perspective to Data Science Degrees and Curricula

University Consortium for
GEOGRAPHIC INFORMATION SCIENCE

Summer 2018

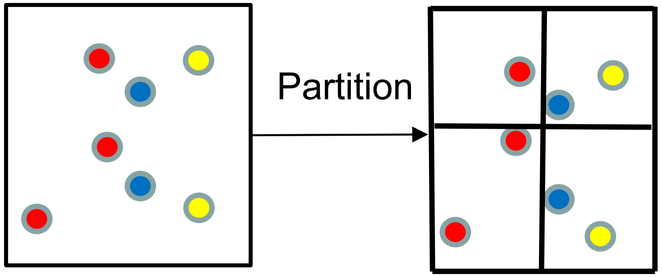
Spatial Partitioning: Gerrymandering

- Space partitioning affects statistical results!
  - Gerrymandering Elections, Correlations
  - Modifiable Areal Unit Problem (MAUP) Dilemma



Election Results 0 - 5 2 - 3 3 - 2

Gini-Index	0.47	0.47	0	0.36

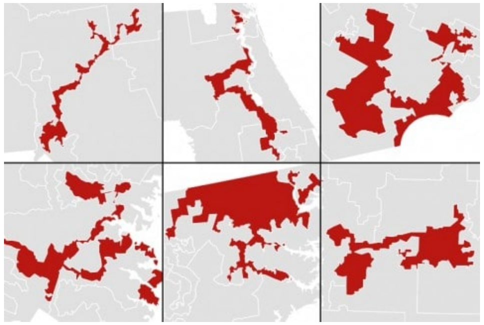


	Partition Based Pearson’s Correlation
🔴 - 🔵	-0.90
🟡 - 🔵	1

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

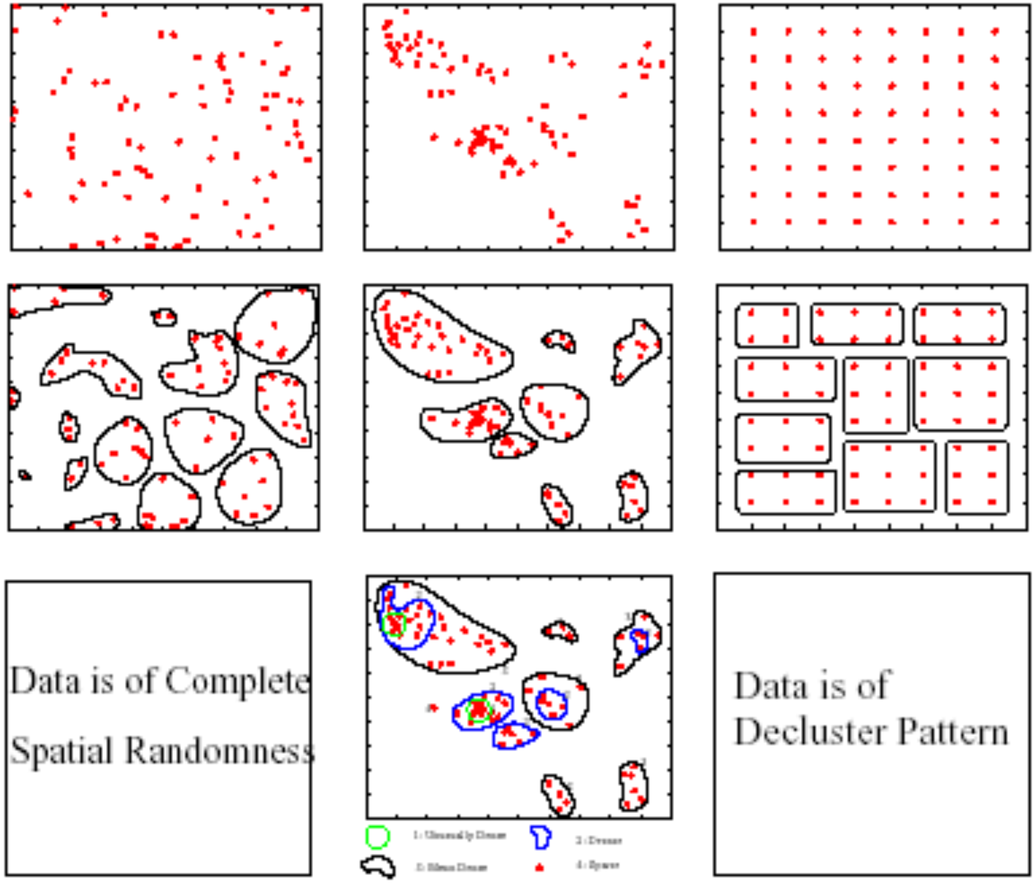
THE WALL STREET JOURNAL.



US Electoral District with Irregular shapes(Source: Washington Post)

Limitation of Traditional Clustering

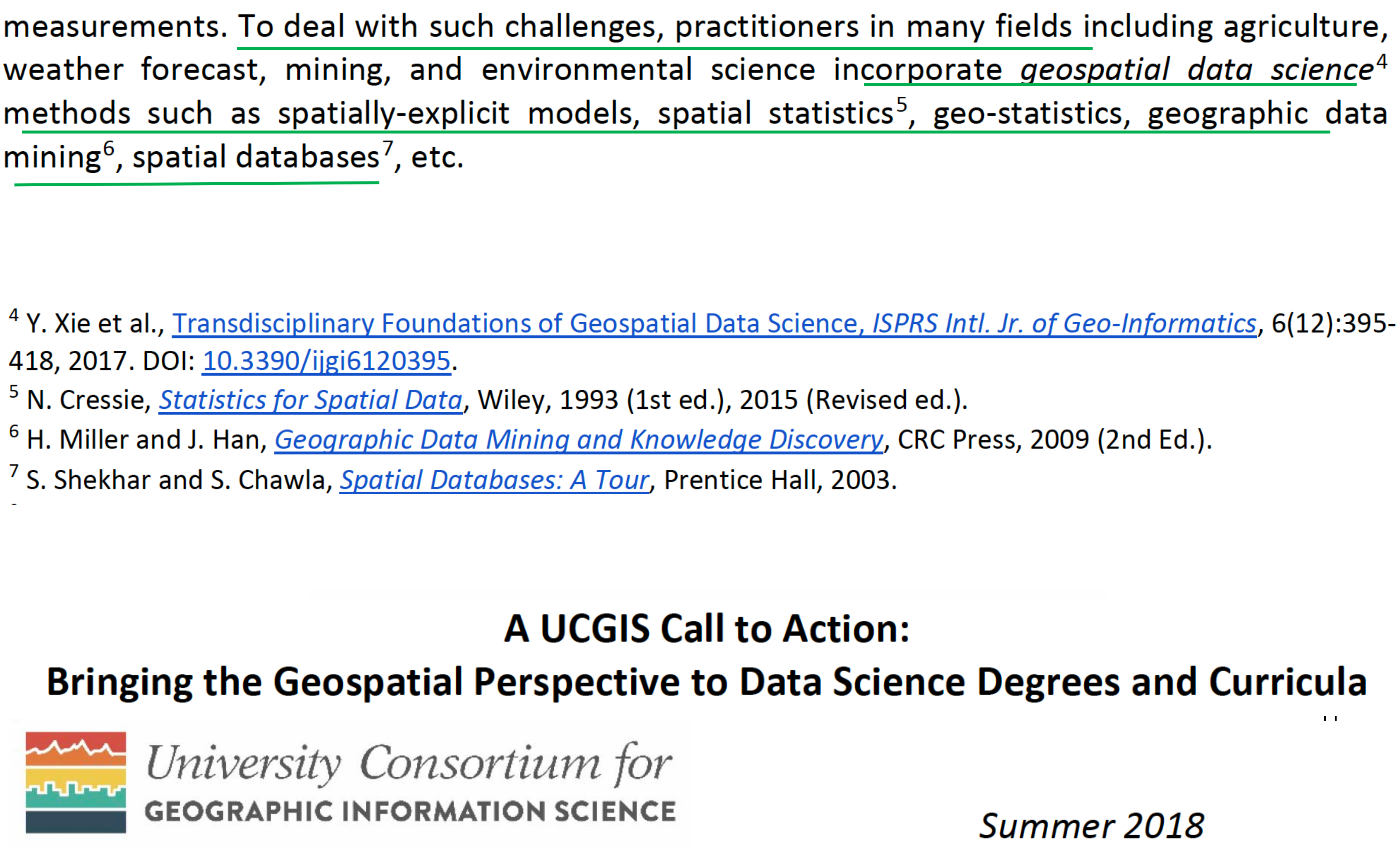
- Challenge: One size does not fit all
  - Prediction error vs. model bias, Cost of false positives, …
- Example. Clustering: Find groups of points


Data is of Complete Spatial Randomness
1: Unusually Dense
2: Dense
3: Mean Dense
4: Sparse
Data is of Decluster Pattern

Traditional Clustering
(K-means always finds clusters)

Spatial Clustering begs to differ!

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.

A UCGIS Call to Action:

Bringing the Geospatial Perspective to Data Science Degrees and Curricula

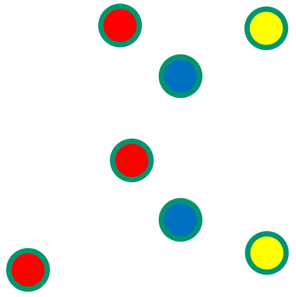
University Consortium for
GEOGRAPHIC INFORMATION SCIENCE

Summer 2018

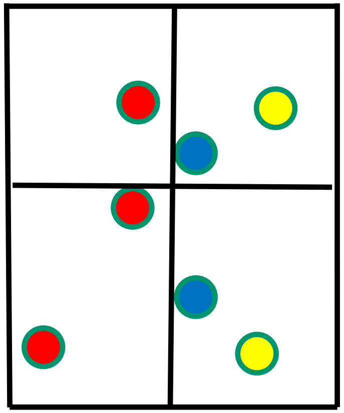
Spatial Interaction Revisited

- Challenge: One size does not fit all

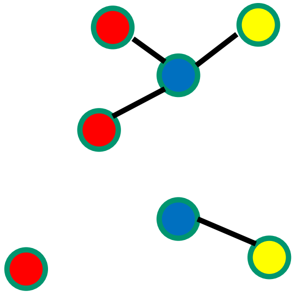
- Ex. Interaction patterns



(a) a map of 3 features



(b) Spatial Partitions

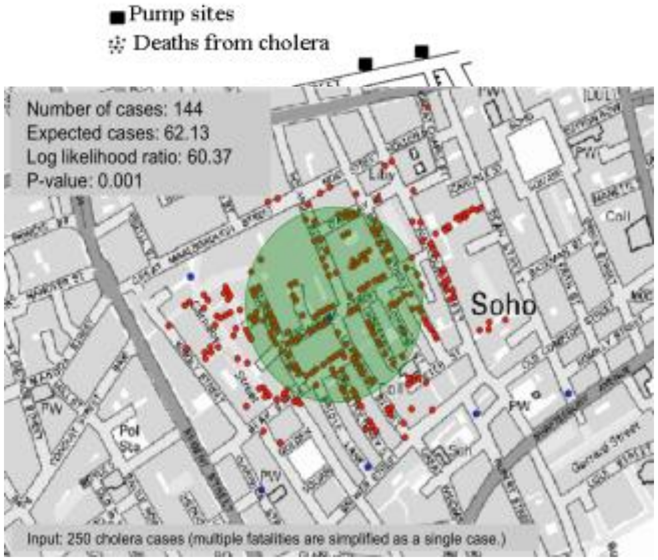


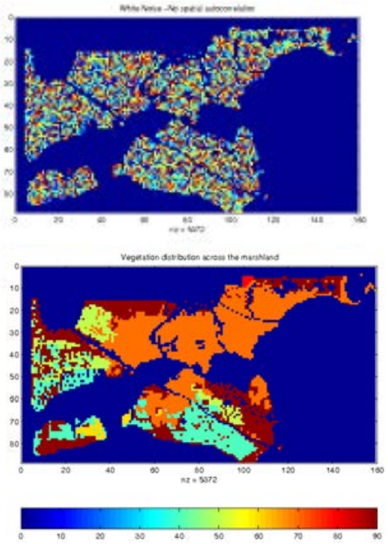
(c) Neighbor graph

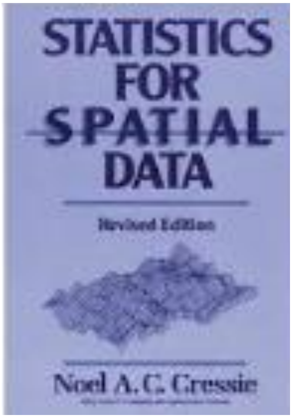
	Pearson's Correlation	Ripley's cross-K	Participation Index
🔴 🔵	-0.90	0.33	0.5
🟡 🔵	1	0.5	1

Sensor Big Data Analysis: Spatial Methods

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

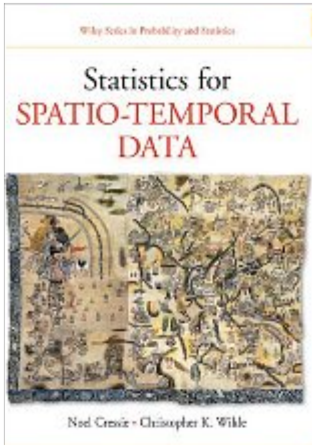

■ Pump sites
Deaths from cholera
Number of cases: 144
Expected cases: 62.13
Log likelihood ratio: 60.37
P-value: 0.001
Soho
Input: 250 cholera cases (multiple fatalities are simplified as a single case.)




STATISTICS FOR SPATIAL DATA

Revised Edition

Noel A.C. Cressie


Wiley Series in Probability and Statistics

Statistics for SPATIO-TEMPORAL DATA

Noel Cressie · Christopher K. Wikle

SaTScan™

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

Spatial Big Data Platforms

Genre	Examples
Relational DBMS, Spatial Library	Oracle, IBM DB2, PostgreSQL, Microsoft 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

Cheap (or free) satellite data on cloud computers

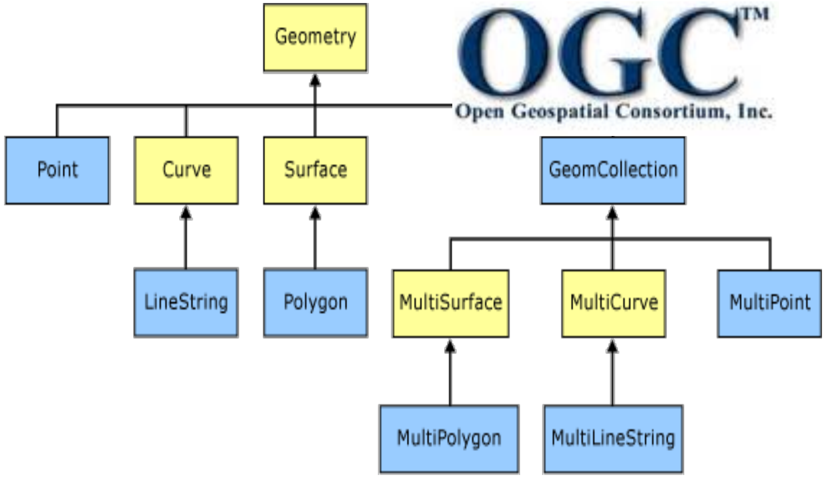
- 2008: USGS gave away 35-year LandSat satellite imagery archive
  - Analog of public availability of GPS signal in late 1980s
- 2017: Many cloud-based Virtual collaboration environment
  - Explosion in machine learning on satelliite imagery to map crops, water, buildings, roads, …

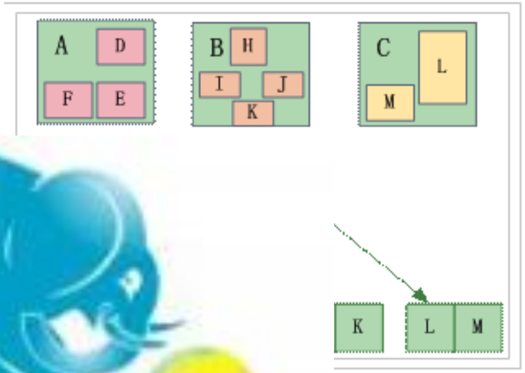
	Google Earth Engines	NEX	AWS Earth
Elevation, Landsat, LOCA, MODIS, NAIP	x	x	x
NOAA	x		x
AVHRR, FIA, GIMMM, GlobCover, NARR, TRIMM, Sentinel-1	x	x	
IARPA, GDELT, MOGREPS, OpenStreetMap, Sentinel-2, SpaceNet (building/road labels for ML)			x
CHIRPS, GeoScience Australia, GSMap, NASS, Oxford Map, PSDI,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

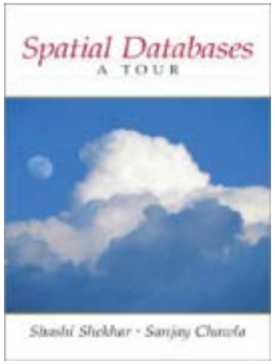
Spatial Big Data Curation

- Meta-data, Schema, DBMS (SQL, Hadoop)
- Challenge: One size does not fit all!

- Ex. Spatial Querying
  - Geo-tag. Checkin, Geo-fence
- Spatial Querying Software
  - OGC Spatial Data Type & Operations
  - Data-structures: B-tree => R-tree
  - Algorithms: Sorting => Geometric
  - Partitioning: random => proximity aware







PostgreSQL

Spatial Hadoop

I can play with your spatial data

Summary : One size data science does not fit all

The World Economy
Runs on GPS.

- Spatial Data are ubiquitous & important
- Current Data Science Tools are inadequate
  - Gerrymandering, Spatial Auto-correlation, …
- **Ask:** BD Hubs & cloud vendors should provide
  - Spatial Data Science Methods
  - Spatial Statistics, Spatial Data Mini