Spatial Big Data Challenges

ACM SIG-SPATIAL Workshop on Analytics for Big Geospatial Data

November 6th, 2012.

!

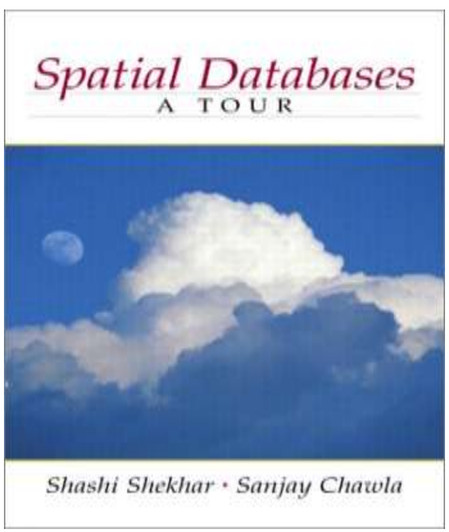
Shashi Shekhar

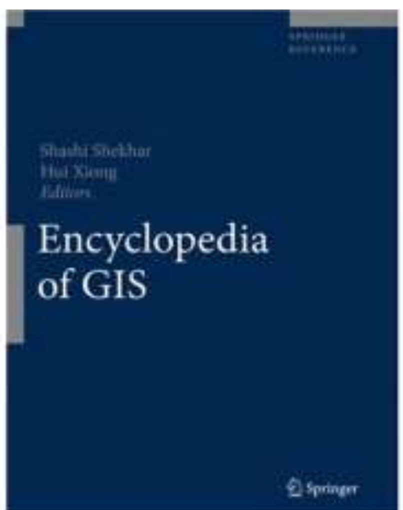
McKnight Distinguished University Professor

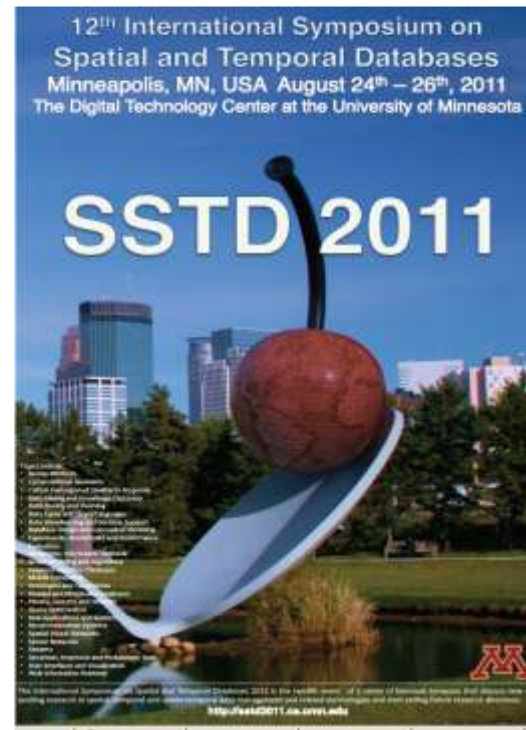
Department of Computer Science and Engineering

University of Minnesota

www.cs.umn.edu/~shekhar







CCC Visioning Workshop: Making a Case for Spatial Computing 2020

http://cra.org/ccc/spatial_computing.php

Computing Community Consortium

We support the computing research community in creating compelling research visions and the mechanisms to realize these visions.

HOME | ABOUT | YOUR VISION | ACTIVITIES | RESOURCES | CONTACT | GO

Funded Visioning Activities

Disaster Management | SEES IT | HealthIT | Interactive Tech | Architecture | XLayer | Robotics | Learning Tech | Open Source | Cyber Physical Systems | Global Development | Theoretical CS | Big Data Computing | NetSE | Spatial Computing

From GPS and Virtual Globes to Spatial Computing-2020

About the workshop

This workshop outlines an effort to develop and promote a unified agenda for Spatial Computing research and development across US agencies, industries, and universities. See the original workshop proposal **here**.

Spatial Computing

Spatial Computing is a set of ideas and technologies that will transform our lives by understanding the physical world, knowing and communicating our relation to places in that world, and navigating through those places.

The transformational potential of Spatial Computing is already evident. From Virtual Globes such as Google Maps and Microsoft Bing Maps to consumer GPS devices, our society has benefitted immensely from spatial technology. We've reached the point where a hiker in Yellowstone, a schoolgirl in DC, a biker in Minneapolis, and a taxi driver in Manhattan know precisely where they are, nearby points of interest, and how to reach their destinations. Large

Logistics

Date: Sept. 10th-11th, 2012
Location: Keck Center
Hotel: Liaison Hotel

Steering Committee

Erwin Gianchandani

Hank Korth

Organizing Committee

Peggy Agouris, George Mason University

Walid Aref, Purdue University

Michael F. Goodchild, University of California - Santa Barbara

Spatial Computing Has Already Transformed Our Lives!

Spatial Computing

- Spatial
  - Space and Time
  - Physical Spaces:
    - Geo, Astronomy, Indoors, Human Body, …
  - Virtual Spaces
    - Localize video, image, document, IP address, …

- Computing
  - Theory, AI, Analytics, …
  - Hardware, Networks, Software, Databases, …
  - Visualization, Augmented Reality
  - Collaboration, CHI,
  - Location Based Services
  - Mobile Computing,
  - Privacy, Data Quality, Uncertainty,
  - …

It is widely used by Government

May 18, 2011

Geospatial Information and Geographic Information Systems (GIS): An Overview for Congress

Congressional Research Service

Table 1. Members of the Federal Geographic Data Committee (FGDC)

Dept. of Agriculture	Environmental Protection Agency
Dept. of Commerce	Federal Emergency Management Agency
Dept. of Defense	General Services Administration
Dept. of Energy	Library of Congress
Dept. of Health and Human Services	National Aeronautics and Space Administration
Dept. of Housing and Urban Development	National Archives and Records Administration
Dept. of the Interior (Chair)	National Science Foundation
Dept. of Justice	Tennessee Valley Authority
Dept. of State	
Dept. of Transportation	Office of Management and Budget (Co-Chair)

It is only a start! Bigger Opportunities ahead!

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 study cites smartphone location services including Foursquare and Loopt, for locating friends, and ones for finding nearby stores and restaurants.

But the biggest single consumer benefit, the study says, is going to come from time and fuel savings from location-based services — tapping into real-time traffic and weather data — that help drivers avoid congestion and suggest alternative routes. The location tracking, McKinsey says, will work either from drivers' mobile phones or GPS systems in cars.

The New York Times

New Ways to Exploit Raw Data May Bring Surge of Innovation, a Study Says

Published: May 13, 2011

Agenda

Time	Day 1 - Activity
830- 9	Opening Remarks, Current Initiatives
9 – 11	**Push Panel: SC Platform Trends, Disruptive Technologies**
	Breakouts on new SC research opportunities from platform trends
1330	Breakout Report Back
1400	**Pull Panel: National Priorities, Societal Applications**
1600	Identify Cross-cutting Characteristics
1600	*Breakout: on new SC research opportunities from application trends*
1700	Report back

Graphics & Vision: John Keyser, TAMU
Interaction Devices: Steven Feiner, Columbia U
LiDAR : Avideh Zakhor, UCB
GPS Modernization: Mark Abrams, NRO
Cell Phones: Ramon Caceres, AT&T
Indoor Localization: Greg Welch, UNC
Internet Localization: Rajesh Gupta, UCSD
Cloud Computing: Divyakant Agarwal, UCSB

Chair: OSTP: Dr. Henry Kelley
US-DoD: Patterns of Life: Eric Vessey
US-DoD: GEOINT: Todd Johanesen
NIH/NIEHS: Exposomics: Michelle Heacock
NASA: Climate Change: John L Schnase
DHS: Disaster Resilience: Nabil Adam
NSF: EarthCube: Clifford Jacobs
DOT: Intellidrive, NextGen : Walton Fehr
DOE: Eco-routing, Bio-fuels: Alicia Lindauer

Time	Day 2 - Activity
9am	Present 1st Draft
11am-12noon	*Breakout: Refine draft based on peer review*
12noon	Present Revised Draft
145pm	Wrap Up, Assignments

Trends to Challenge-Themes

Trends	Challenges	Themes
Indoor Localization Internet Localization	We long for Location-based services everywhere (e.g., indoors)	Cross-cutting
Cloud Computing: Cell Phones Graphics & Vision	Every platform is location-aware	From Sensors to Cloud
	Everyone is a Map-maker	From Fusion to Synergetics
	Everyone uses location-based service	Spatial Cognition First
Interaction Devices LiDAR, sensors	ST Map layers are diversifying fast & old analytics techniques are too weak	From Fusion to Synergetics
GPS Modernization	We are worried about GPS-spoofing, stalking, and loss of privacy!	Cross-Cutting

Agency: Theme	Link	Challenge
NIH/NIEHS: Exposomics	Phone sensor	We long for Location-based services everywhere (e.g., indoors, outdoors, …)
DHS: Disaster Resilience	Geo-targetting	Every platform is location-aware
NSF: EarthCube: DoD/NSA:	Geo. Sc. Community	Everyone is a (collaborative ST) Map-maker
		Everyone uses location-based service
Patterns of Life DoD/NGA: GEOINT NASA: Climate Change	ST pattern mining	ST Map layers are diversifying fast & old analytics techniques are too weak
DOE: Eco-routing, Bio-fuels DOT: Intellidrive, NextGen	Secure smart cars, Private eco-routing	We are worried about GPS-spoofing, stalking, and loss of privacy!

Four Breakout Groups

- SC Sciences : From Fusion to Synergetics
  - Theory, AI, G.I.Science, Analytics, …
- SC Systems : From Sensors to Clouds
  - Hardware, Networks, Software, Database, …
- SC Services : Spatial Cognition First
  - Visualization, Augmented Reality
  - Collaboration
  - Location Based Services …
- Cross-Cutting
  - Mobile Computing,
  - Privacy, Security, Trust
  - Data Quality, Uncertainty, …

Cross-Cutting Breakout Group (C1000)

Budhendra Bhaduri	ORNL
Daniel Z. Sui	Ohio State University
Lea Shanley	Wilson Center
Michael Goodchild	UC Santa Barbara
Ouri E. Wolfson	Univ. of Illinois at Chicago
Paul Torrens	University of Maryland
Ramon Caceres	AT&T Research
Shaowen Wang	University of Illinois at UC
Xuan Liu	IBM
May Yuan	University of Oklahoma
Dev Oliver	University of Minnesota

Science Breakout Group (206)

Benjamin Kuipers	University of Michigan
Jie Gao	Stony Brook University
Jim Shine	Army Research
Mike Worboys	University of Maine
Norman Sadeh	CMU
Sara Graves	UA Huntsville
Stephen Hirtle	University of Pittsburgh
Vipin Kumar	University of Minnesota
Craig A. Knoblock	Information Sciences Institute
Raju Vatsavai	ORNL

Services Breakout Group (208)

Cecilia Aragon	University of Washington
Chuck Hansen	University of Utah
Dinesh Manocha	University of North Carolina
Greg Welch	University of North Carolina
John Keyser	Texas A&M University
Lee Allison	Arizona Geological Survey
Steven Feiner	Columbia University
Tom Erickson	IBM
Peggy Agouris	George Mason University
Dan Keefe	University of Minnesota

Systems Breakout Group (C700)

Avideh Zakhor	UC Berkeley
Chang-Tien Lu	Virginia Tech
Divyakant Agrawal	UC Santa Barbara
Edward M. Mikhail	Purdue
Jagan Sankaranarayanan	NEC Labs
Mohamed Ali	Microsoft
Rajesh Gupta	UC San Diego
Siva Ravada	Oracle
Vijay Atluri	NSF
Walid G. Aref	Purdue
Michael R. Evans	UMN

Breakout Goals

Day1 AM – Questions to address:

1. What role will Spatial Computing play in our lives in 2020?
2. What are most compelling transformative opportunities ?

Day 1 PM - Quadcharts (1 per questions)

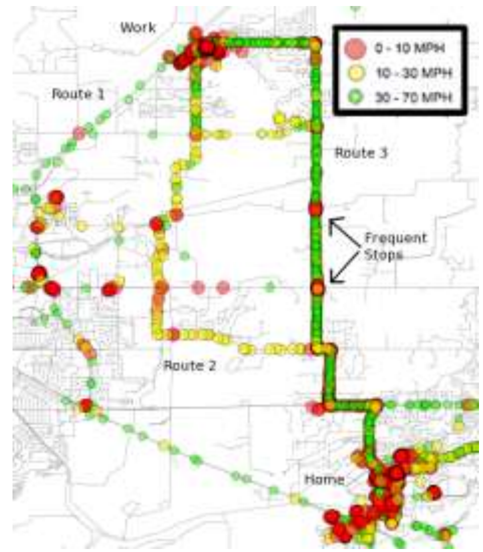
Example on next slide.

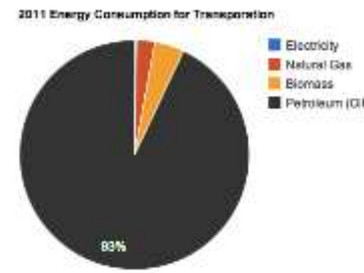
Day 2 AM - Paragraphs

Sample Quad-Chart: Eco-Routing Using Spatial Big Data

OBJECTIVES

- Next-generation Routing services to minimize fuel or GHG emissions instead of distance or travel-time
- Expolit Spatial Big Data, e.g., gps-traces and temporally detailed roadmaps, to identify fuel-saving opportunities
- Novel representation, algorithms, and architecture for SBD and problems violating Dynamic Programming assumption





SPATIAL COMPUTING CHALLENGES

- Change in frame of reference from a snapshot perspective to the perspective of the individual traveling through a transportation network
- Diversity of SBD significantly increases computational cost because it magnifies the impact of the partial nature and ambiguity of traditional routing query speciation
- Route ranking changes over time violating dynamic programming assumptions underlying routing algorthms.
- Spatial Big Data volume, velocity and variety exceed capacity of current spatial computing systems

TRANSFORMATIVE POTENTIAL

- Significantly reduce US consumption of petroleum, the dominant source of energy for transportation
- Reduce the gap between domestic petroleum consumption and production
- Reduce greenhouse gas (GHG) emissions
- A 2011 McKinsey Global Institute report estimates savings of “about $600 billion annually by 2020 via vehicles avoiding congestion and reducing idling

From Workshop to Report Outline

Recent Grand Challenges Discussions

- 2003 NRC report (Geospatial Future)
- 2009 NSF Workshop (P. Agouris)
- 2010 AAG Panel

Proposal

1. Introduction
2. …

Workshop Activities

Opening Remarks, Current Initiatives

Push Panel: SC Platform Trends, Disruptive Technologies

Breakouts on new SC research opportunities from platform trends

Lunch, Breakout Report Back

Pull Panel: National Priorities, Societal Applications of Spatial Computing (SC)

Break

Identify cross-cutting SC application characteristics

Breakout: on new SC research opportunities from application trends

Report back

Synthesis,

Report Outline, Writing Assignments

Report Outline

1. Introduction

1.0 Why Spatial Computing?

1.1 Challenges

1.2 Opportunities

2. Research Directions

2.1 SC Sciences: From Fusion to Synergetics

2.2 SC Systems: From Sensors to Cloud

2.3 SC Services: Spatial Cognition First

2.4 Cross-Cutting

3. Geo-Privacy Policy Opportunities

4. Closing

5. About This Document

A. List of Contributors

B: Emerging Application Attributes

B.1 Example Applications

B.2 Attributes of Applications

C. Emerging Platform Trends

Report – Sample Research Directions

Report Outline

1. Introduction
1.0 Why Spatial Computing?
1.1 Challenges
1.2 Opportunities
2. Research Directions
2.1 SC Sciences: From Fusion to Synergetics
2.2 SC Systems: From Sensors to Cloud
2.3 SC Services: Spatial Cognition First
2.4 Cross-Cutting
3. Geo-Privacy Policy Opportunities
4. Closing
5. About This Document
A. List of Contributors
B: Emerging Application Attributes
B.1 Example Applications
B.2 Attributes of Applications
C. Emerging Platform Trends

2.1.1 Manipulating Qualitative Spatio-Temporal Data
2.1.2 (Spatio-temporal) Prediction
2.1.3 Synthesizing Multiple Projects of Past & Future
2.1.4 Collection, Fusion, Curation of Sensed Data
2.1.5 Spatial Computing Standards

2.2.1 Computational Issues in Spatial Big Data
2.2.2 Spatial Computing Infrastructure
2.2.3 Augmented Reality
2.2.4 Device to Device Spatial Computing

2.3.1 Human Spatial-Computing Interaction
2.3.2 Spatial Cognitive Assistance
2.3.3 Context-aware Spatial Computing
2.3.4 SC Assisted Human-Human Interactions
2.3.5 Spatial Cognition and Spatial Abilities

2.4.1 Ubiquitous Computing
2.4.2 Persistent Sensing & Monitoring
2.4.3 Trustworthy SC Systems, e.g., Transportation
2.4.4 Geo-Privacy

Report –Section 3: Sample Principles & Policy Possibilities

CCC Council: Review Nov. 2nd, 2012. – Nov. 16th, 2012.

Next: Choose message for policy makers (Need your help!)

- Ex.: spatial economy: location-based-commerce, mobile commerce

Report Outline

1. Introduction
1.0 Why Spatial Computing?
1.1 Challenges
1.2 Opportunities
2. Research Directions
2.1 SC Sciences: From Fusion to Synergetics
2.2 SC Systems: From Sensors to Cloud
2.3 SC Services: Spatial Cognition First
2.4 Cross-Cutting
3. Geo-Privacy Policy Opportunities
4. Closing
5. About This Document
A. List of Contributors
B: Emerging Application Attributes
B.1 Example Applications
B.2 Attributes of Applications
C. Emerging Platform Trends

1. Emergencies are different! E911
2. Differential Privacy: E911 → PLAN, CMAS
3. Send Apps to Data, not vice versa (e.g., Eco-routing)
4. (Transparent) Transactions for location traces
5. Responsible Entities for location traces
    1. Credit-bureau/Census
    2. HIPPA++ for responsible parties

3.1 What can policy makers do?
3.2 Urgency
3.3 Benefits and Costs

3.2 Urgency: Tech Giants scramble to get upto speed (NYTimes, Oct. 22, 2012)

3.3 Cusp of an economic revolution leveraging Emerging spatial data. Additional benefits in Energy independence, disaster resiliency, env. health

Outline

- Motivation
- What is Spatial Big Data (SBD)?
  - Definitions
  - Examples & Use Cases
- SBD Infrastructure
- SBD Analytics
- Conclusions

Spatial Big Data Definitions

- Spatial datasets exceeding capacity of current computing systems
  - To manage, process, or analyze the data with reasonable effort
  - Due to Volume, Velocity, Variety, …

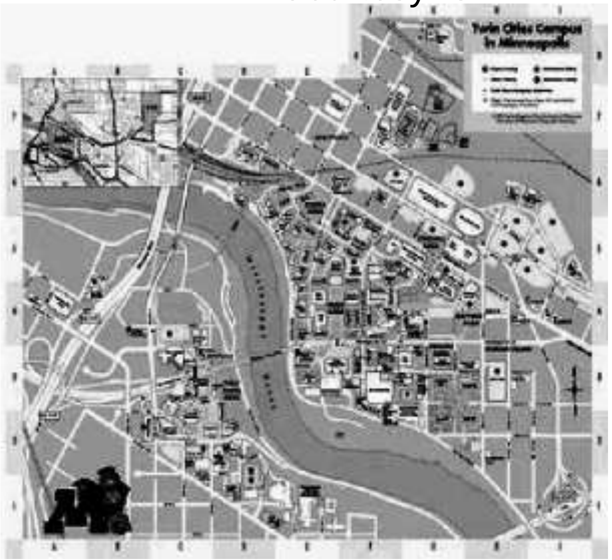
- SBD History
  - Data-intensive Computing: Cloud Computing, Map-Reduce, Pregel
  - Middleware
  - Big-Data including data mining, machine learning, …

Traditional Spatial Data

- Spatial attribute:
  - Neighborhood and extent
  - Geo-Reference: longitude, latitude, elevation
- Spatial data genre
  - **Raster**: geo-images e.g., Google Earth
  - **Vector:** point, line, polygons
  - **Graph**, e.g., roadmap: node, edge, path

Raster Data for UMN Campus
Courtesy: UMN

Graph Data for UMN Campus
Courtesy: Bing



Vector Data for UMN Campus
Courtesy: MapQuest

Raster SBD

The New York Times : January 10, 2010

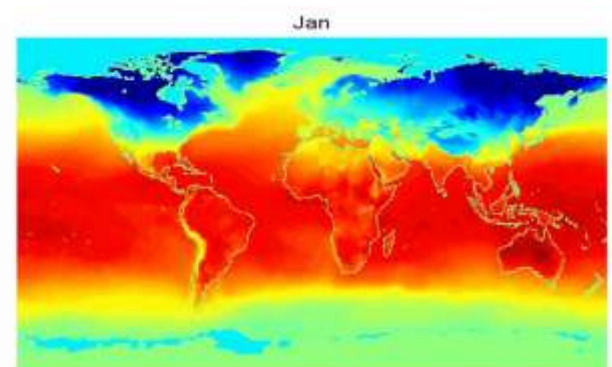
Military Is Awash in Data From Drones

adding 2,500 analysts to help handle the growing volume of data.

With a new $500 million computer system

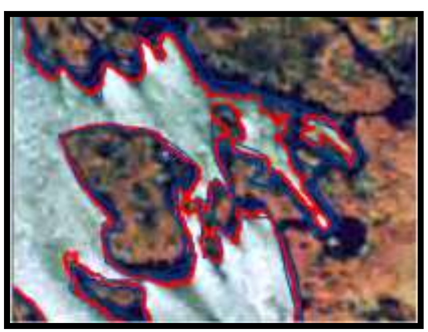
- Data Sets >> Google Earth
  - Geo-videos from UAVs, security cameras
  - Satellite Imagery (periodic scan), LiDAR, …
  - Climate simulation outputs for next century
- Example use cases
  - Patterns of Life
  - Change detection, Feature extraction, Urban terrain

LiDAR & Urban Terrain

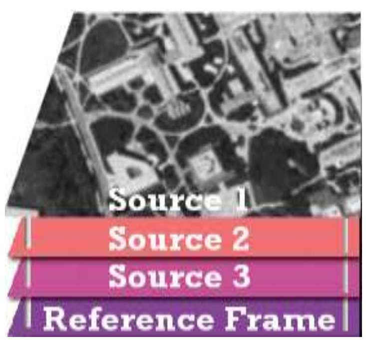


Average Monthly Temperature

(Courtsey: Prof. V. Kumar)



Feature Extraction

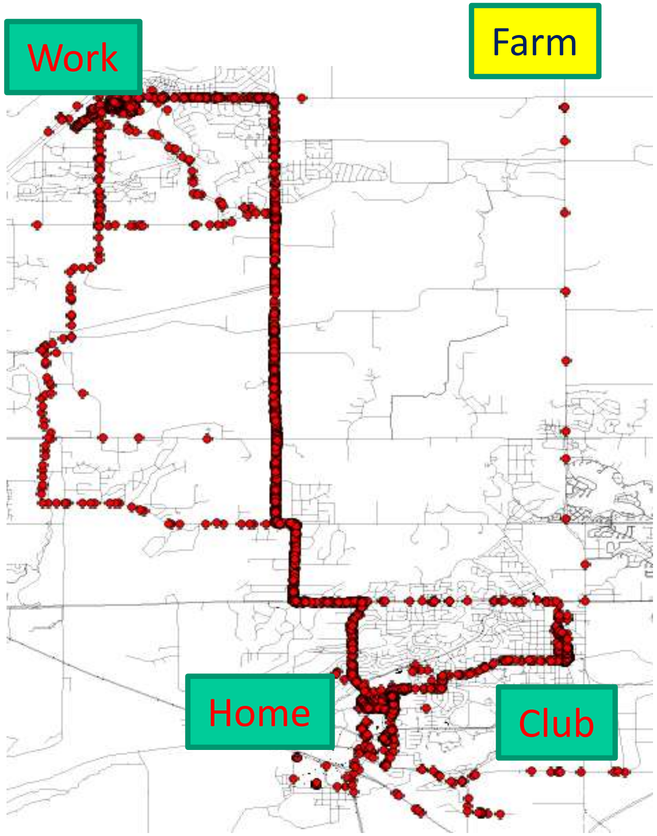


Change Detection

Use Case: Patterns of Life

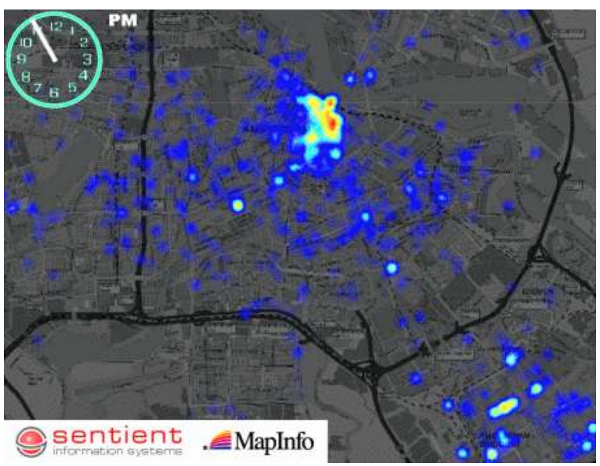
- Weekday GPS track for 3 months
  - Patterns of life
  - Usual places and visits
  - Rare places, Rare visits

	Morning 7am – 12am	Afternoon 12noon – 5pm	Evening 5pm – 12pm	Midnight 12midnight – 7pm	Total
Home	10	2	15	29	54
Work	19	20	10	1	50
Club	4	5	4		15
Farm			1		1
Total	30	30	30	30	120



Vector SBD from Geo-Social Media

- Vector data sub-genre
  - Point: location of a tweet, Ushahidi report, checkin, …
  - Line-strings, Polygons: roads in openStreetMap
- Use cases: Persistent Surveillance
  - Outbreaks of disease, Disaster, Unrest, Crime, …
  - Hot-spots, emerging hot-spots
  - Spatial Correlations: co-location, teleconnection



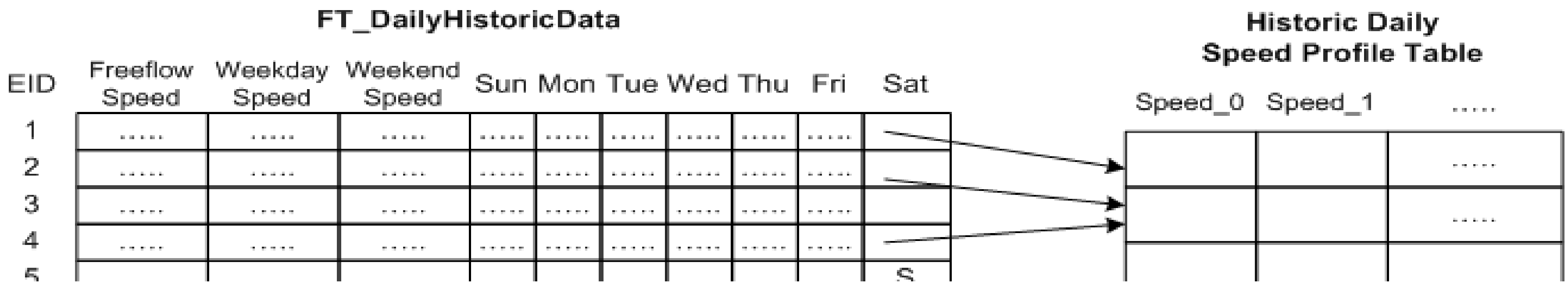
Persistent Surveillance at American Red Cross

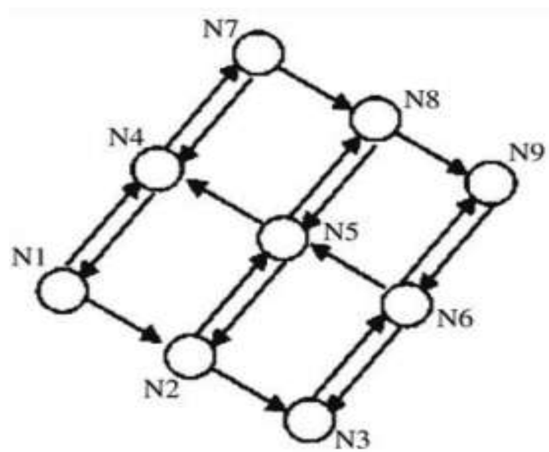
- Even before cable news outlets began reporting the tornadoes that ripped through Texas on Tuesday, a **map** of the state began blinking red on a screen in the Red Cross' new social media monitoring center, alerting weather watchers that something was happening in the hard-hit area. (AP, April 16th, 2012)

American
Red Cross

Graphs SBDs: Temporally Detailed

- Spatial Graphs, e.g., Roadmaps, Electric grid, Supply Chains, …
  - Temporally detailed roadmaps [Navteq]
- Use cases: Best start time, Best route at different start-times





Nodes

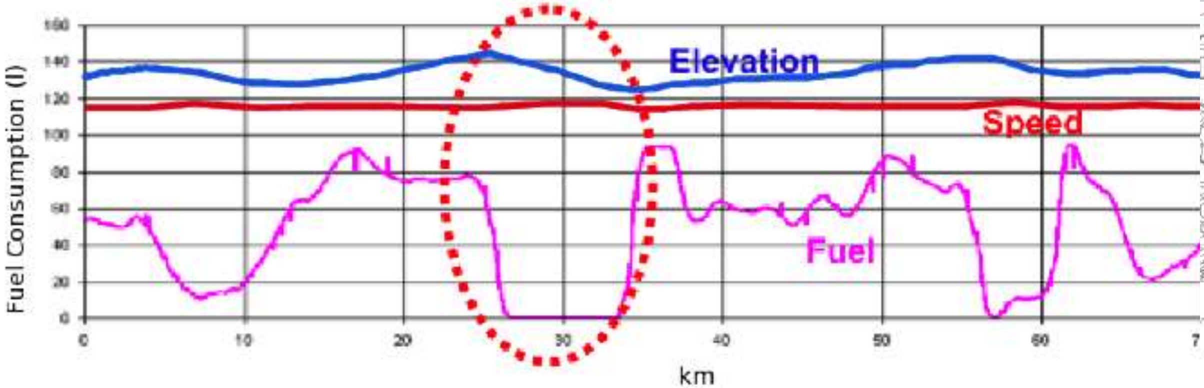
NID
N1
N2
N3
N4
N5
N6
N7
N8
N9

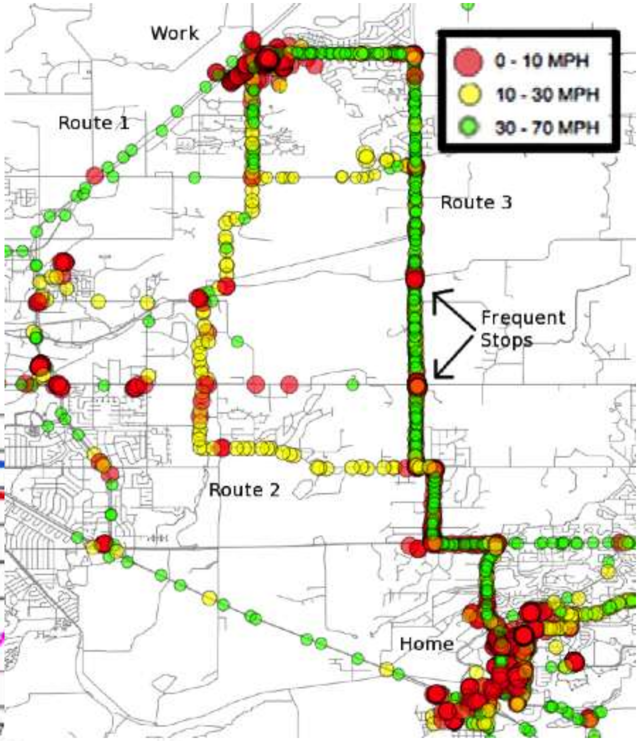
Edges

EID	From	To	Speed	Distance
E1	N1	N2	35mph	0.075mi
E2	N1	N4	30mph	0.075mi
E3	N2	N3	35mph	0.078mi
E4	N2	N5	30mph	0.078mi
E5	N3	N6	30mph	0.077mi
E6	N4	N1	30mph	0.075mi
E7	N4	N7	30mph	0.078mi
E8	N5	N2	30mph	0.078mi
...	...	...	...	...

Emerging SBD: Mobile Device2Device

- Mobile Device
  - Cell-phones, cars, trucks, airplanes, …
  - RFID-tags, bar-codes, GPS-collars, …
- Trajectory & Measurements sub-genre
  - Receiver: GPS tracks, …
  - System: Cameras, RFID readers, …
- Use cases:
  - Tracking, Tracing,
    - Improve service, deter theft …
  - Geo-fencing, Identify nearby friends
  - Pattersns of Life
  - Eco-routing





36

Emergin Use-Case: Eco-Routing

- Minimize fuel consumption and GPG emission
  - rather than proxies, e.g. distance, travel-time
  - avoid congestion, idling at red-lights, turns and elevation changes, etc.

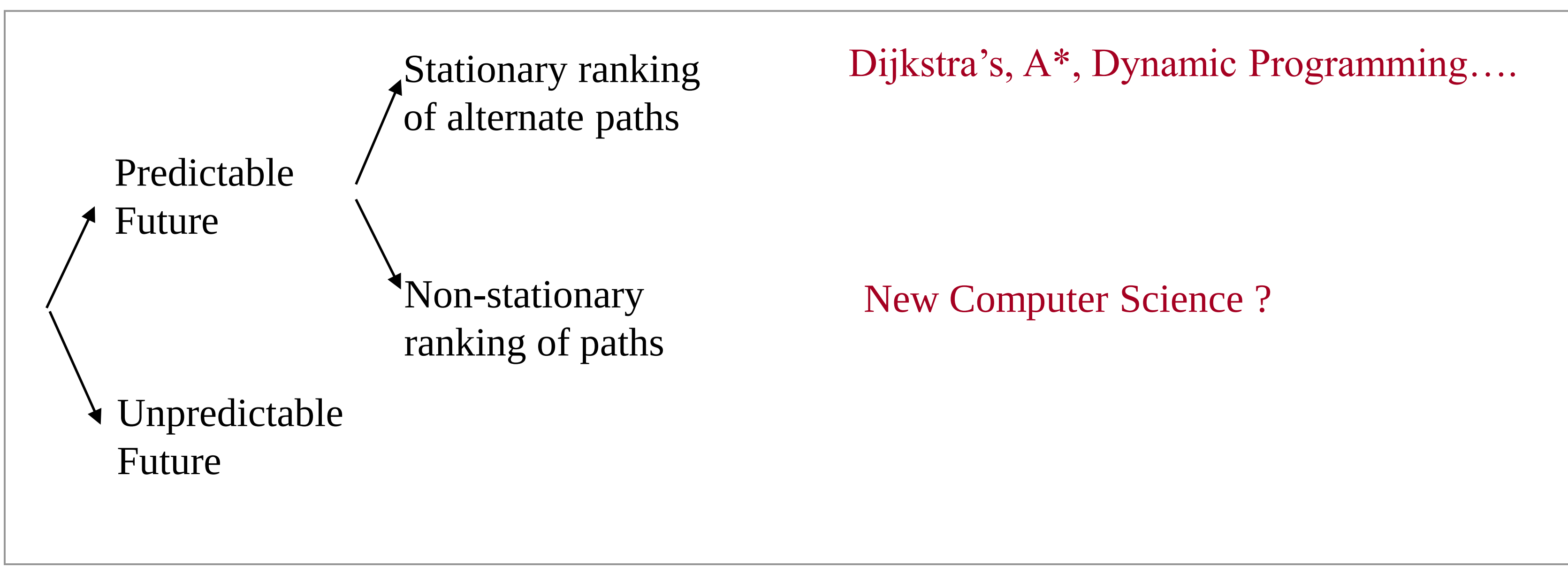
The New York Times

U.P.S. Embraces High-Tech Delivery Methods (July 12, 2007)

By "The research at U.P.S. is paying off. ....— saving roughly three million gallons of fuel in good part by mapping routes that minimize left turns."

Eco-Routing Questions

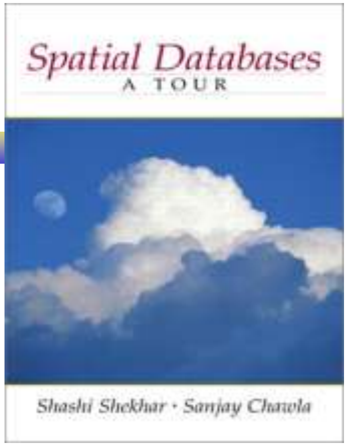
- What are expected fuel saving from use of GPS devices with static roadmaps?
- What is the value-added by historical traffic and congestion information?
- How much additional value is added by real-time traffic information?
- What are the impacts of following on fuel savings and green house emissions?
  - traffic management systems (e.g. traffic light timing policies),
  - vehicles (e.g. weight, engine size, energy-source),
  - driver behavior (e.g. gentle acceleration/braking), environment (e.g. weather)
- What is computational structure of the Eco-Routing problem?

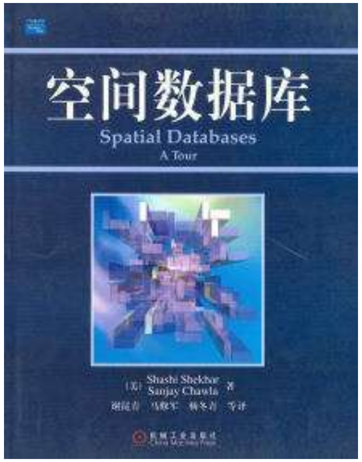


Relational to Spatial DBMS to SBD Management

- 1980s: Relational DBMS
  - Relational Algebra, B+Tree index
  - Query Processing, e.g. sort-merge equi-join algorithms, …
- Spatial customer (e.g. NASA, USPS) faced challenges
  - Semantic Gap
    - Verbose description for distance, direction, overlap
    - Shortest path is Transitive closure
  - Performance challenge due to linearity assumption
    - Are Sorting & B+ tree appropriate for geographic data?
- New ideas emerged in 1990s
  - Spatial data types and operations (e.g. OGIS Simple Features)
  - R-tree, Spatial-Join-Index, space partitioning, …

- SBD may require new thinking for
  - Temporally detailed roadmaps
  - Eco-routing queries
  - Privacy vs. Utility Trade-off





Outline

- Motivation
- SBD Definitions & Examples
- SBD Analytics
  - Spatial Data Mining
  - SDM Limitations & SBD Opportunities
- SBD Infrastructure
- Conclusions

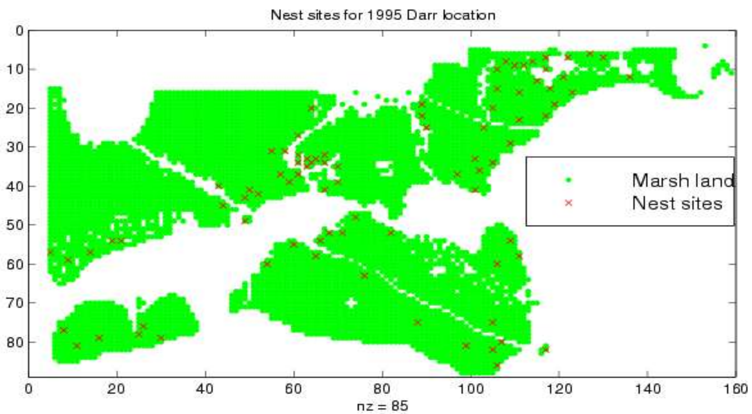
Data Mining to Spatial Data Mining to SBD Analytics

November 14, 2004 The New York Times
What Wal-Mart Knows About Customers' Habits

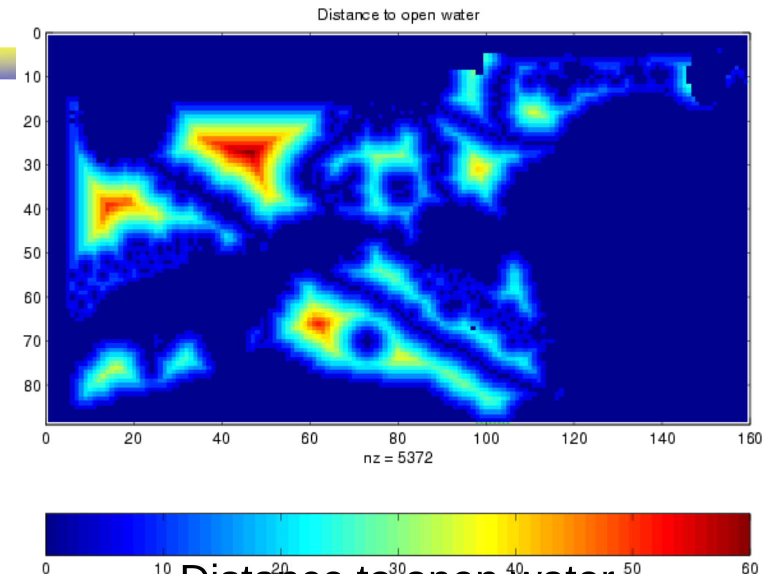
- 1990s: Data Mining
  - Scale up traditional models (e.g., Regression) to large relational databases (460 Tbytes)
  - New pattern families: Associations : Which items are bought together? (Ex. Diaper, beer)
- Spatial customers
  - Walmart: Which items are bought just before/after events, e.g. hurricanes?
    - Where is a pattern (e.g., (diaper-beer) prevalent?
  - Global climate change: tele-connections

- But faced challenges
  - Independent Identical Distribution assumption not reasonable for spatial data
  - Transactions, i.e. disjoint partitioning of data, not natural for continuous space
- This led to Spatial Data Mining (last decade)
- SBD raise new questions
  - May SBD address open questions, e.g. estimate spatial neighborhood (e.g., W matrix)?
  - Does SBD facilitate better spatial models, e.g., place based ensembles beyond GWR?
  - (When) Does bigger spatial data lead to simpler models, e.g. database as a model ?
  - On-line Spatio-temporal Data Analytics

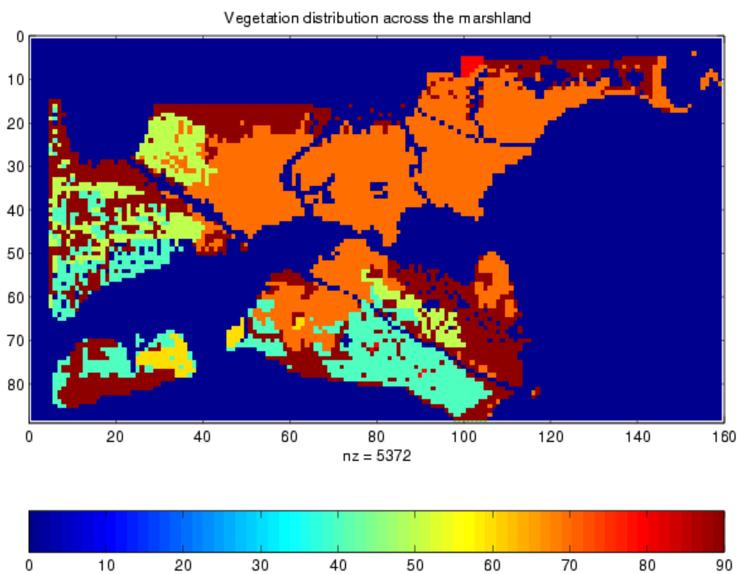
Spatial Data Mining Example 1: Spatial Prediction



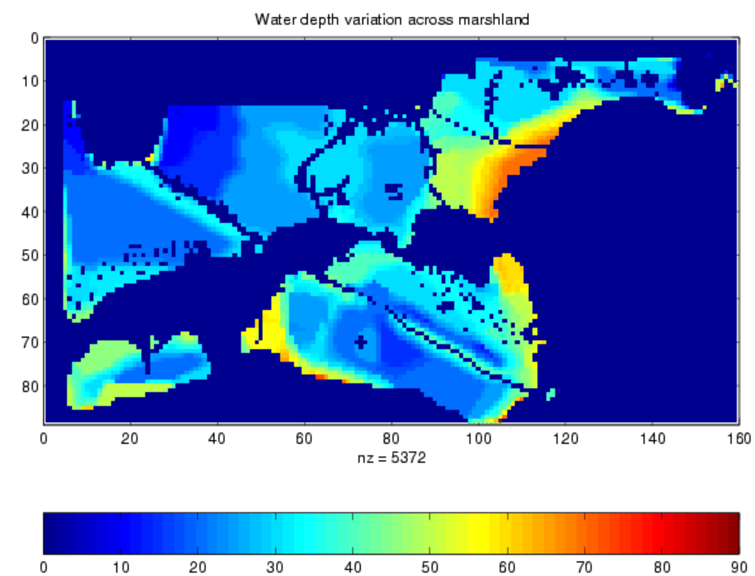
Nest locations



Distance to open water



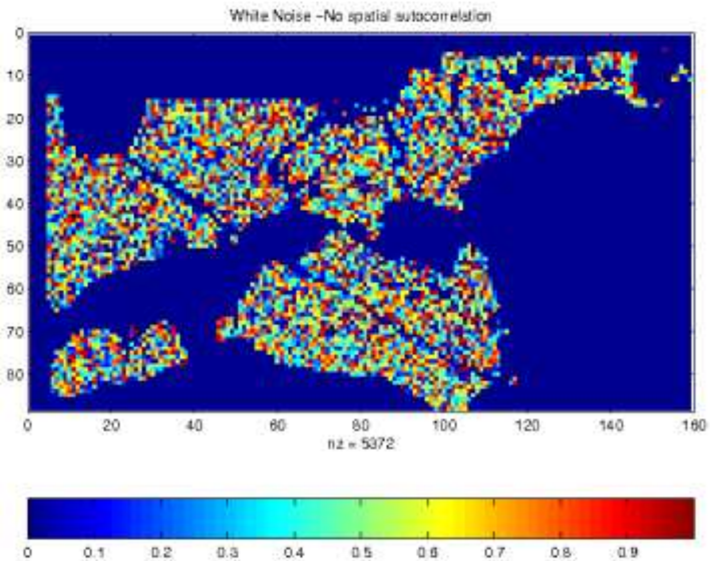
Vegetation durability



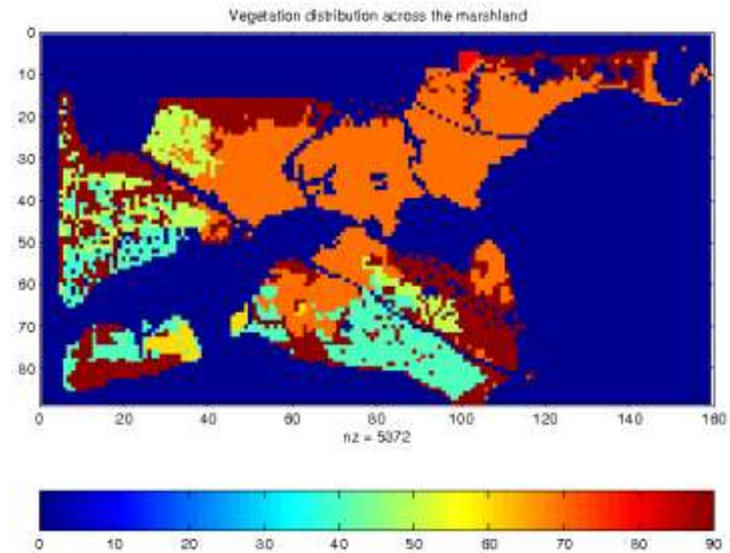
Water depth

Spatial Autocorrelation (SA)

- First Law of Geography
  - “All things are related, but nearby things are more related than distant things. [Tobler, 1970]”



Pixel property with independent identical Distribution (i.i.d)



Vegetation Durability with SA

- Autocorrelation
  - Traditional i.i.d. assumption is not valid
  - Measures: K-function, Moran’s I, Variogram, …

Parameter Estimation for Spatial Auto-regression

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

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

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

- **Maximum Likelihood Estimation**
- Computationally Expensive
  - Determinant of a large matrix
- Iterative Computation
  - Golden Section Search for ρ

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

SBD Opportunity 1: Estimate Spatial Neighbor Relationship

- SDM Limitation 1: Neighbor relationship is End-users' burden !
  - Colocation mining, hotspot detection, spatial outlier detection, …
  - Example: W matrix in spatial auto-regression
  - Reason: W quadratic in number of location
  - Reliable estimation of W needs very large number data samples
- SBD Opportunity 1: Post-Markov Assumption
  - SBD may be large enough to provide reliable estimate of W
  - This will relieve user burden and may improve model accuracy
  - One may not have assume
    - Limited interaction length, e.g. Markov assumption
    - Spatially invariant neighbor relationships, e.g., 8-neighbor
    - Tele-connections are derived from short-distance relationships

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

SBD Opportunity 2: Place Based Ensemble of Models

- SDM Limitation 2: Modeling of Spatial Heterogeneity is rare
  - Spatial Heterogeneity: No two places on Earth are identical
  - Yet, Astro-Physics tradition focused on place-independent models
  - Was it due to paucity of data ?
  - Exception: Geographically Weighted Regression or GWR [Fortheringham et al.]
  - GWR provides an ensemble of linear regression models, one per place of interest

- Opportunity 2: SBD may support Place based ensemble of models beyond GWR
  - Example: Place based ensemble of Decision Trees for Land-cover Classification
  - Example: Place based ensemble of Spatial Auto-Regression Models
  - Computational Challenge:
    - Naïve approach may run a learning algorithm for each place.
    - Is it possible to reduce computation cost by exploiting spatial auto-correlation ?

Outline

- Motivation
- SBD Definition and Examples
- SBD Analytics
- SBD Infrastructure
  - Parallelizing Spatial Computations
  - Implications for Cloud Platforms
- Conclusions

Parallelizing Spatial Big Data on Cloud Computing

- Case 1: Compute Spatial-Autocorrelation Simpler to Parallelize
  - Map-reduce is okay
  - Should it provide spatial de-clustering services?
  - Can query-compiler generate map-reduce parallel code?

- Case 2: Harder : Parallelize Range Query on Polygon Maps
  - Need dynamic load balancing beyond map-reduce
  - MPI or OpenMP is better!

- Case 3: Estimate Spatial Auto-Regression Parameters, Routing
  - Map-reduce is inefficient for iterative computations due to expensive “reduce”!
  - MPI, OpenMP, Pregel or Spatial Hadoop is essential!
  - Ex. Golden section search, Determinant of large matrix
  - Ex. Eco-routing algorithms, Evacuation route planning

Ex. 3: Hardest to Parallelize

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

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

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

- **Maximum Likelihood Estimation** $\ln(L) = \boxed{\ln\left|\mathbf{I} - \rho\mathbf{W}\right|} - \frac{n\ln(2\pi)}{2} - \frac{n\ln(\sigma^2)}{2} - SSE$
- Need cloud computing to scale up to large spatial dataset.
- However,
  - Map reduce is too slow for iterative computations!
  - computing determinant of large matrix is an open problem!

Spatial Big Data (SBD)

- SBD Definitions
- SBD Applications
- SBD Analytics
- SBD Infrastructure
- Conclusions

Summary

- SBD are important to society
  - Ex. Eco-routing, Public Safety & Security, Understanding Climate Change
- SBD exceed capacity of current computing systems
- DBMS Oppotunities
  - Eco-Routing: Lagrangian frame, Non-Stationary Ranking
  - Privacy vs. Utility Trade-offs
- Data Analytics Opportunities
  - Post Markov Assumption – Estimate Neighbor Relationship from SBD
  - Place based Ensemble Models to address spatial heterogeneity
  - Bigger the spatial data, simpler may be the spatial models
  - Online Spatial Data Analytics
- Platform Opportunities
  - Map-reduce – expensive reduce not suitable for iterative computations
  - Load balancing is harder for maps with polygons and line-strings
  - Spatial Hadoop ?