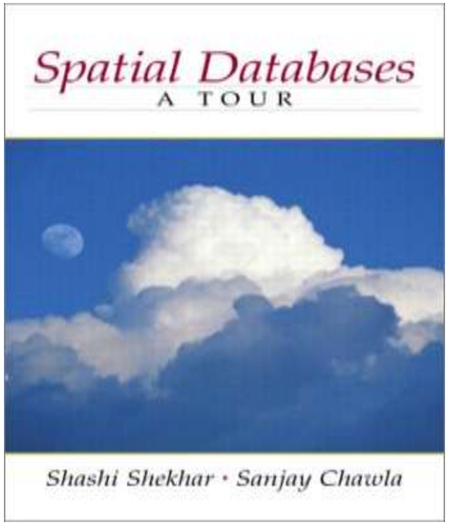
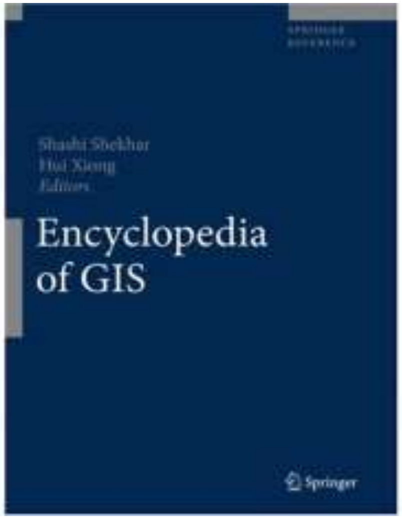
Spatial Big Data Challenges

ARO/NSF Workshop on Big Data at Large: Applications and Algorithms (Durham, NC)
June 14th, 2012.

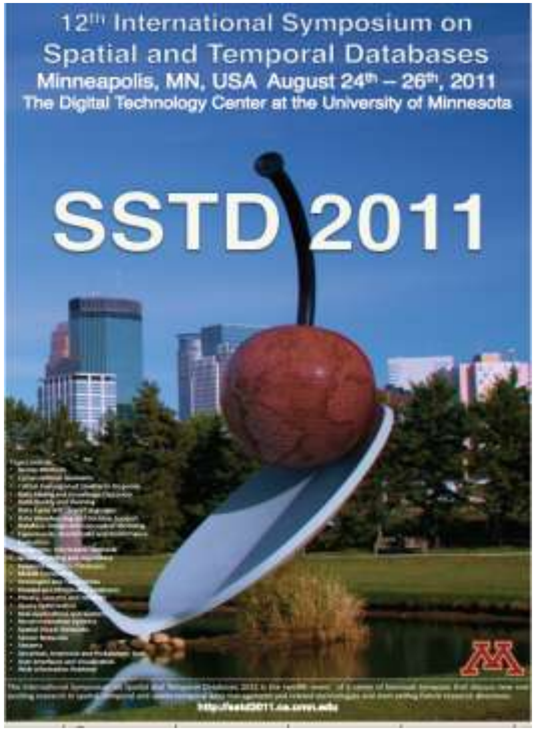
Congratulations Army on 237th Anniversary!

Shashi Shekhar
McKnight Distinguished University Professor
Department of Computer Science and Engineering
University of Minnesota
www.cs.umn.edu/~shekhar





GeoInformatica
Editors-in-Chief: Patrick Bergougnoux
John R. Herring
Shashi Shekhar
Volume 9, Number 2, June 2005
Springer


12th International Symposium on
Spatial and Temporal Databases
Minneapolis, MN, USA August 24th – 26th, 2011
The Digital Technology Center at the University of Minnesota
SSTD 2011

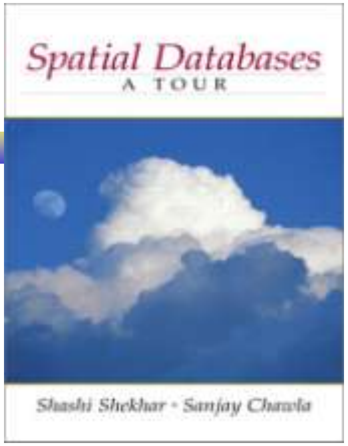
Spatial Big Data (SBD) - Summary

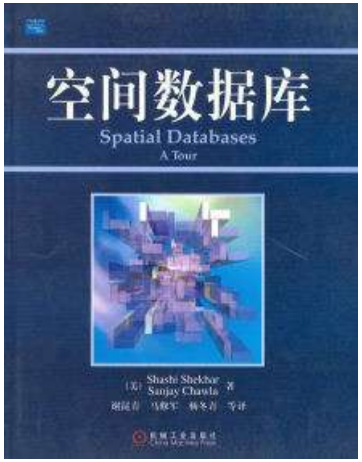
- SBD are important to society
  - Ex. Eco-routing, Public Safety & Security, Understanding Climate Change
- SBD exceed capacity of current computing systems
- DBMS Challenges
  - Eco-Routing: Lagrangian frame, Non-Stationary Ranking
  - Privacy vs. Utility Trade-offs
- Data Analytics Opportunities
  - 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 Challenges
  - Map-reduce – expensive reduce not suitable for iterative computations
  - Load balancing is harder for maps with polygons and line-strings
  - Spatial Hadoop ?

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





Шаши Шекхар, Санжей Чаула
Основы пространственных
баз данных
учебное пособие
Геоинформационные системы • САПР • Мультимедиа

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

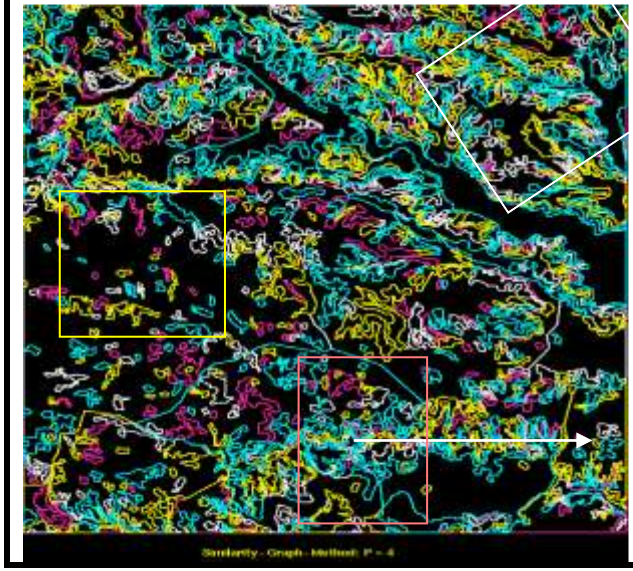
Parallelizing Spatial Big Data on Cloud Computing

- 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!

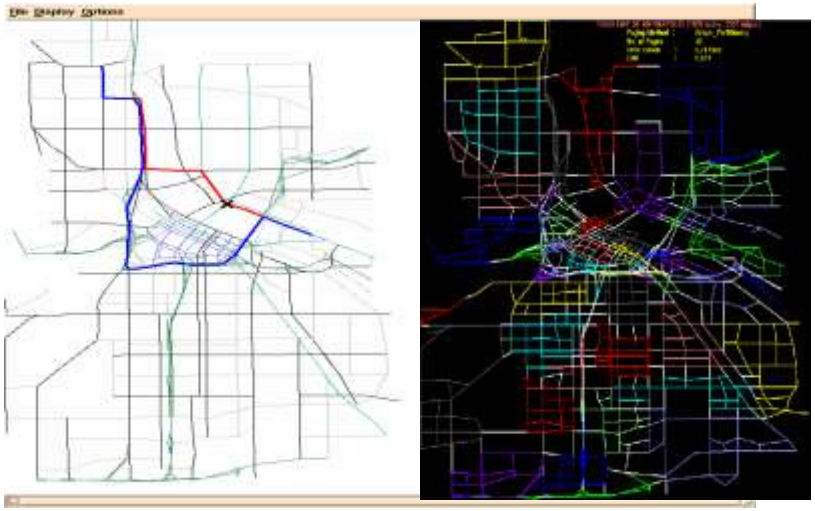
- Case 3: Estimate Spatial Auto-Regression Parameters, Routing
  - Map-reduce is inefficient for iterative computations due to expensive "reduce"!
  - Ex. Golden section search, Determinant of large matrix
  - Ex. Eco-routing algorithms, Evacuation route planning
  - Option 1: Develop non-iterative formulations of spatial problems
  - Option 2: Alternative Platform: MPI, OpenMP, Pregel or Spatial Hadoop

Spatial Databases: Representative Projects

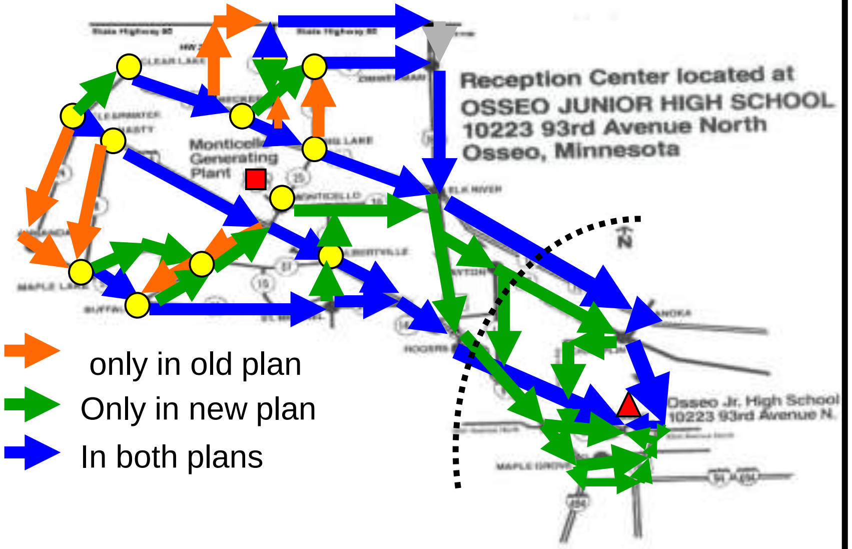


Parallelize Range Queries

Shortest Paths   Storing graphs in disk blocks



Evacutation Route Planning



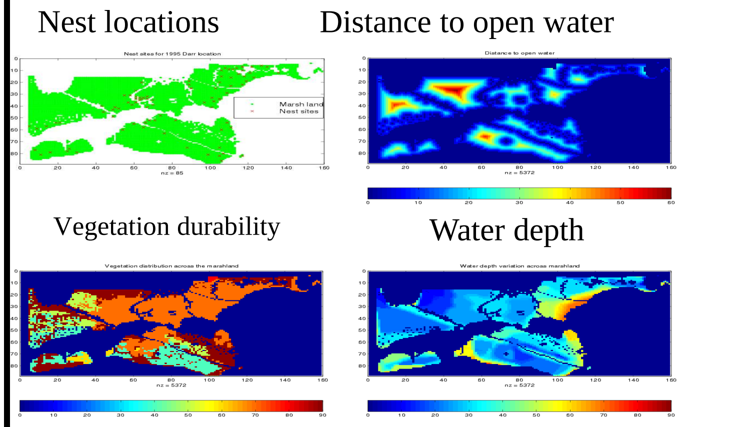
➡ only in old plan

➡ Only in new plan

➡ In both plans

Spatial Data Mining : Representative Projects

Location prediction: nesting sites



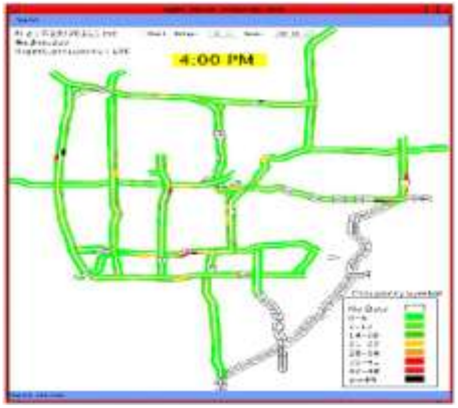
Nest locations

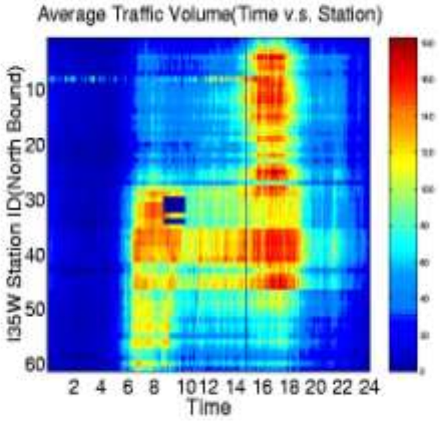
Distance to open water

Vegetation durability

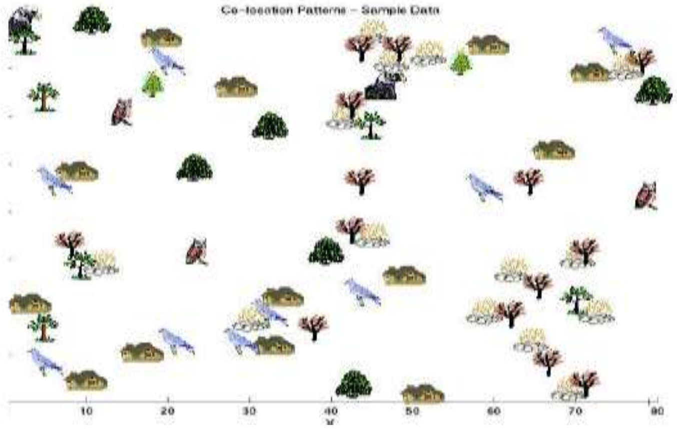
Water depth

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

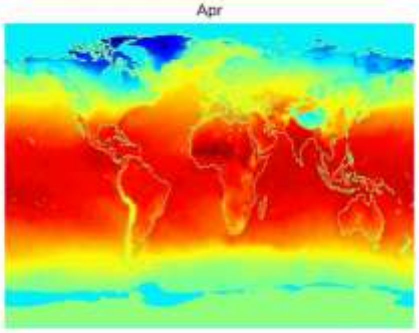




Co-location Patterns



Tele connections



W
W
W
R
R
R
R
R
D
D
D
El-Niño

Spatial Computing: Recent Trends

Microsoft® Virtual Earth™

ESRI ArcGIS X™

IBM Smarter Planet

Google Earth

acm GIS

SIG SPATIAL

Spatial Statistics

BRIAN D. RIPLEY

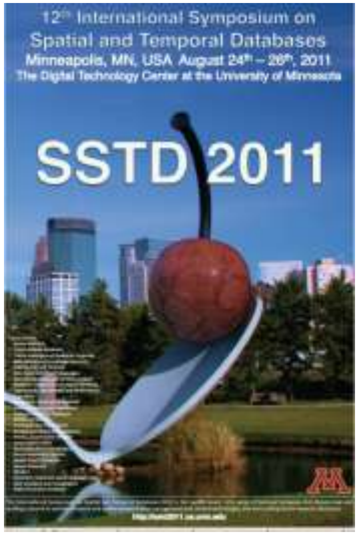
WILEY SERIES IN PROBABILITY AND STATISTICS

waze way to go

foursquare

ORACLE®

SPATIAL



12th International Symposium on Spatial and Temporal Databases

Minneapolis, MN, USA August 24th – 26th, 2011

The Digital Technology Center at the University of Minnesota

SSTD 2011

GeoInformatica

Editors-in-Chief: Patrick Bergougnoux, Shashi Shekhar, [illegible]

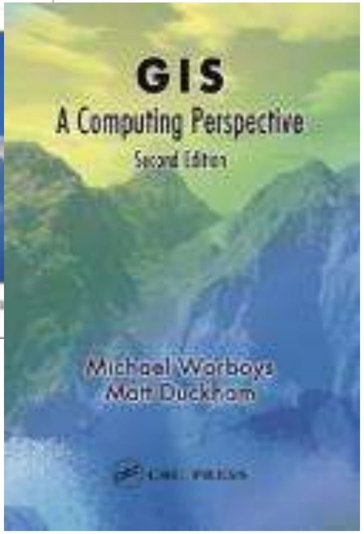
Volume 14, Number 1, January 2010

Springer

Spatial Databases

A TOUR

Shashi Shekhar · Sanjay Cha



GIS

A Computing Perspective

Second Edition

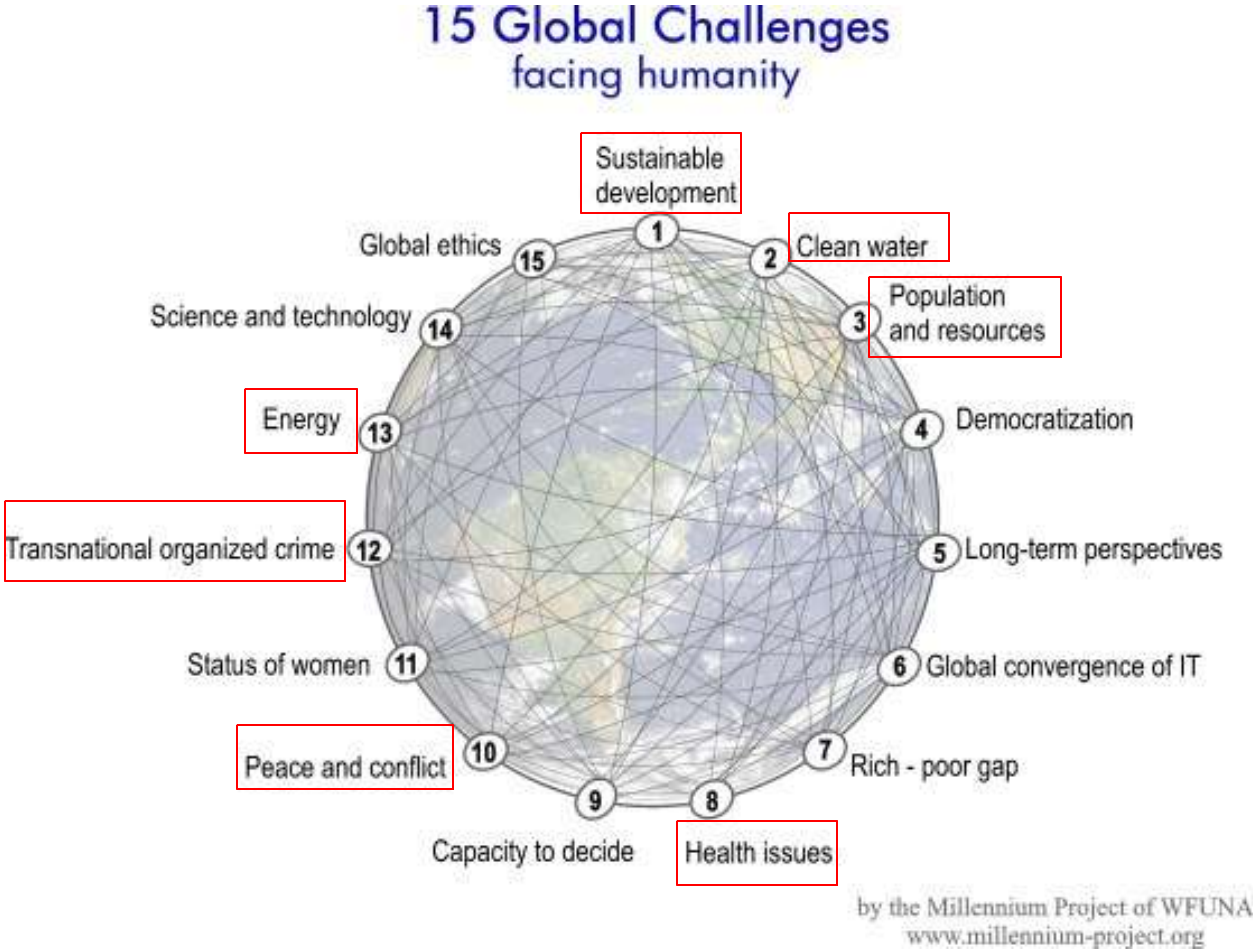
Michael Worboys

Matt Duckham

CRC PRESS

Motivation for Spatial Big Data (SBD)

- Societal:
  - Google Earth, Google Maps, Navigation, location-based service
  - Global Challenges facing humanity – many are geo-spatial!
  - Many may benefit from Big Spatial Data



SBD Motivation

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

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

Published: May 13, 2011

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

Outline

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

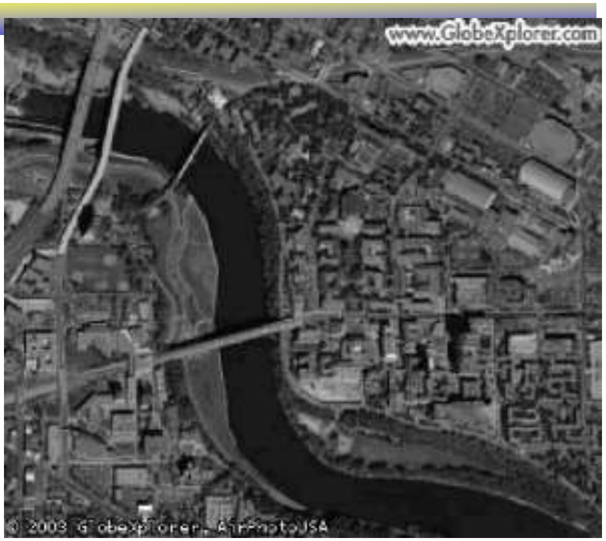
Spatial Big Data Definitions

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

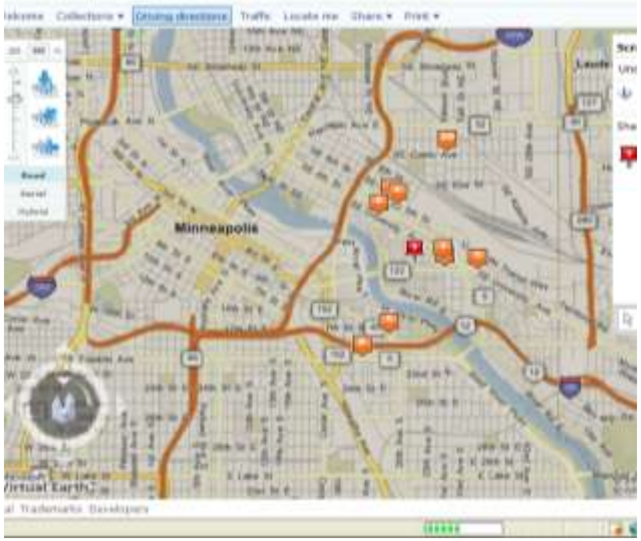
- SBD History
    - Data-intensive Computing: Cloud Computing, Map-Reduce, Pregel
    - Middleware
    - Big-Data including data mining, machine learning, …

Traditional Spatial Data

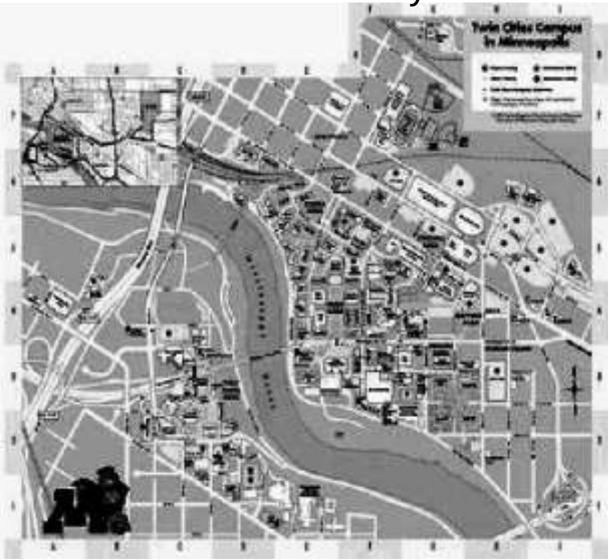
- Spatial attribute:
  - Neighborhood and extent
  - Geo-Reference: longitude, latitude, elevation
- Spatial data genre
  - **Raster**: geo-images e.g., Google Earth
  - **Vector:** point, line, polygons
  - **Graph**, e.g., roadmap: node, edge, path



Raster Data for UMN Campus
Courtesy: UMN



Graph Data for UMN Campus
Courtesy: Bing



Vector Data for UMN Campus
Courtesy: MapQuest

Raster SBD

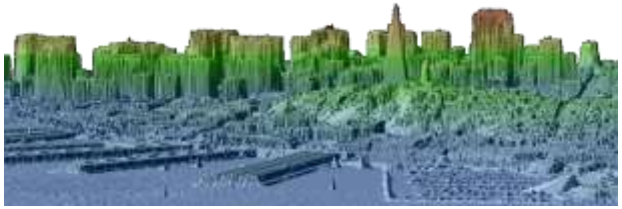
The New York Times d: January 10, 2010

Military Is Awash in Data From Drones

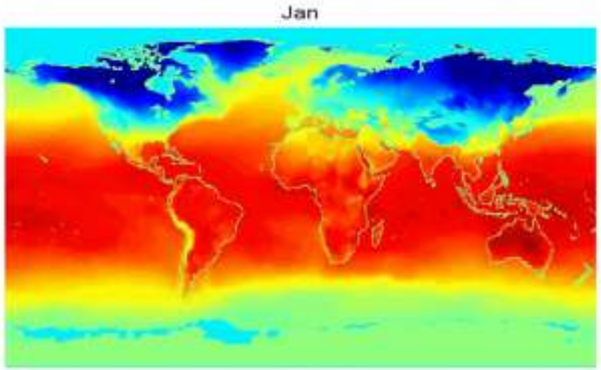
adding 2,500 analysts to help handle the growing volume of data.

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

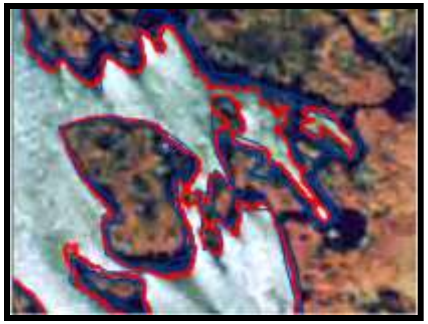


LiDAR & Urban Terrain

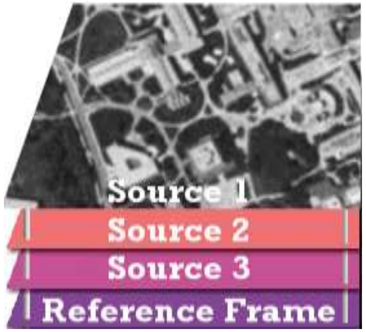


Average Monthly Temperature

(Courtsey: Prof. V. Kumar)



Feature Extraction

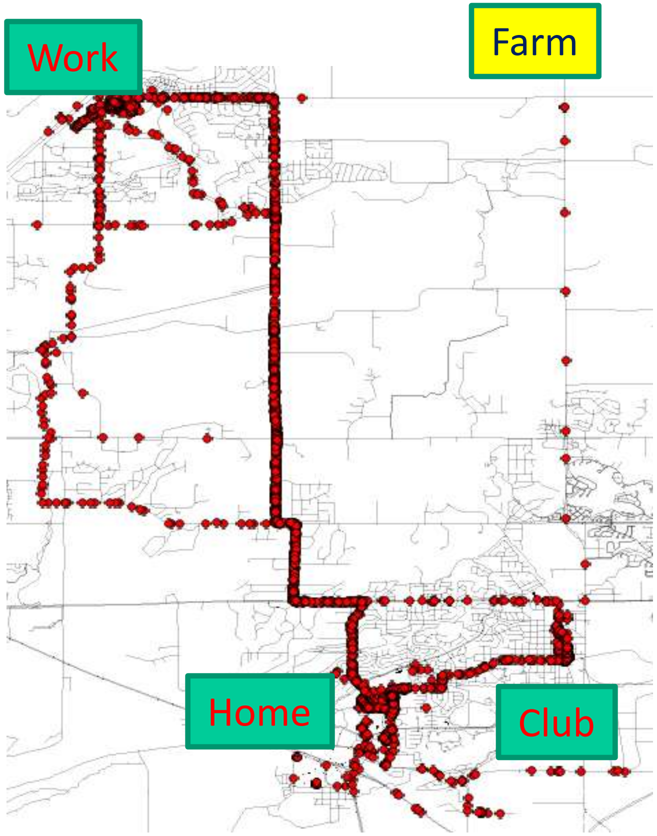


Change Detection

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



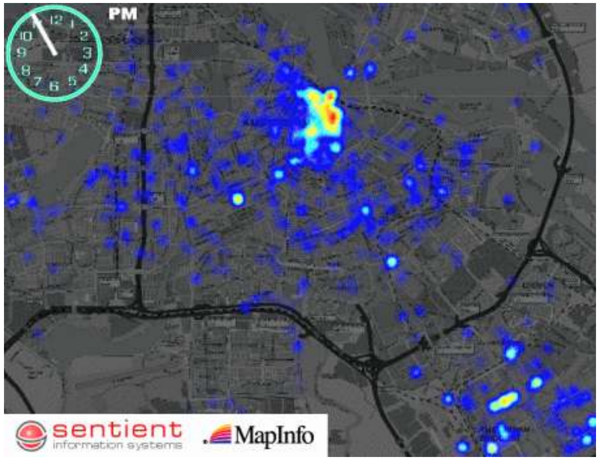
Vector SBD from Geo-Social Media

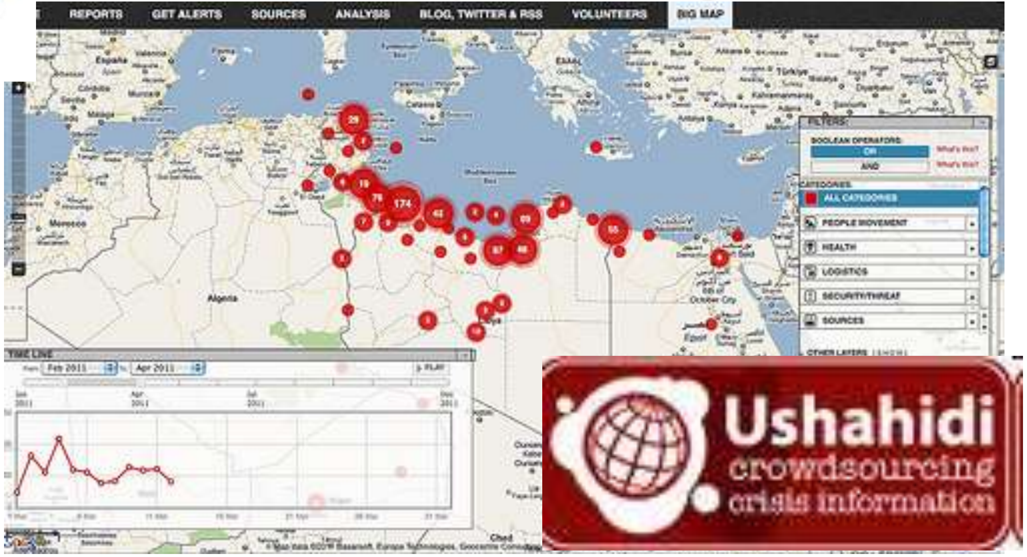
- Vector data sub-genre
  - Point: location of a tweet, Ushahidi report, checkin, …
  - Line-strings, Polygons: roads in openStreetMap
- Use cases: Persistent Surveillance
  - Outbreaks of disease, Disaster, Unrest, Crime, …
  - Hot-spots, emerging hot-spots
  - Spatial Correlations: co-location, teleconnection

CHECK-IN HERE ON foursquare

openStreetMap

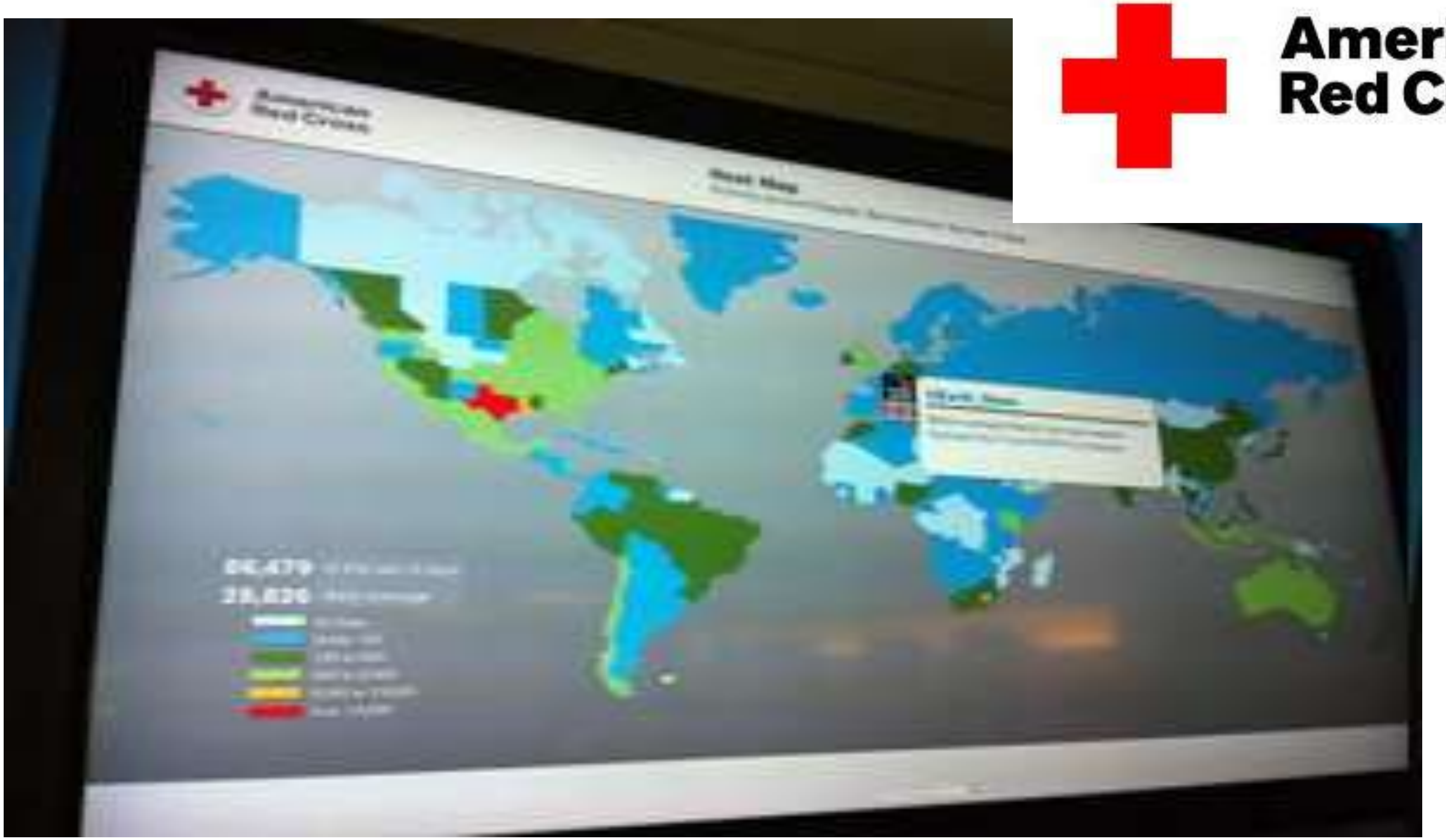
twitter





Persistent Surveillance at American Red Cross

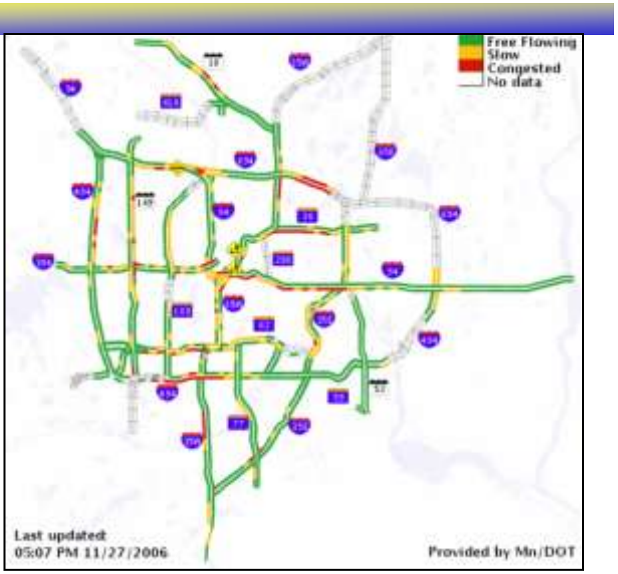
- Even before cable news outlets began reporting the tornadoes that ripped through Texas on Tuesday, a **map** of the state began blinking red on a screen in the Red Cross' new social media monitoring center, alerting weather watchers that something was happening in the hard-hit area. (AP, April 16th,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

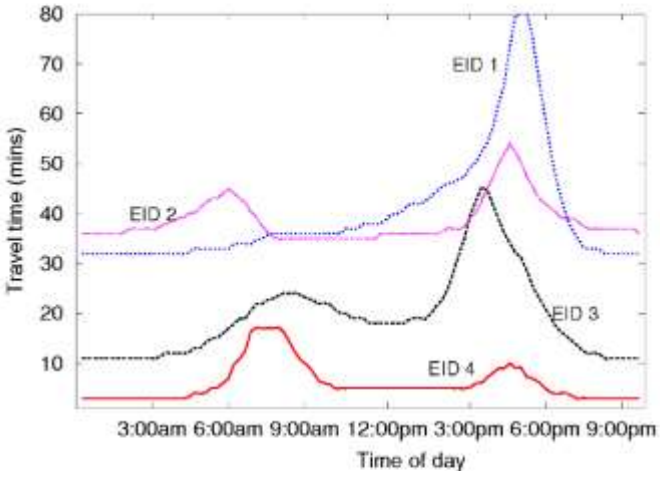


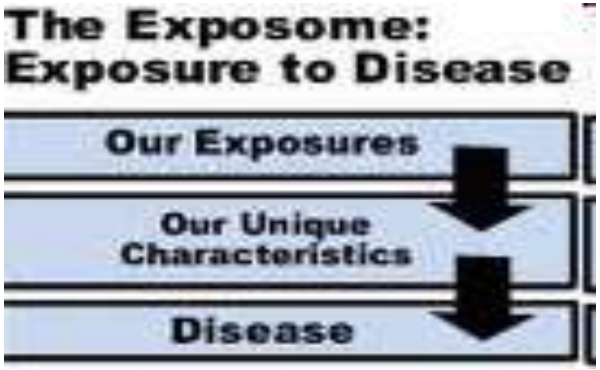
American Red Cross

Emerging SBD: Geo-Sensor Networks

- Geo-Sensor Network Examples
  - Urban roads: Loop detetors, Cameras
  - Electricity distribution grids, …
  - Environmental sensors for air quality, …
  - Robot with sensors, …
- Sensor Network sub-genre
  - Fixed reasonable resource: traffic sensors
  - Ad-hoc, resource poor: wireless sensor networks
- Use cases:
  - Monitor events, anomalies, e.g., accidents, Congestion, hotspots
  - Feed-back control, Predictive planning
  - Environmental Health







The Exposome:
Exposure to Disease

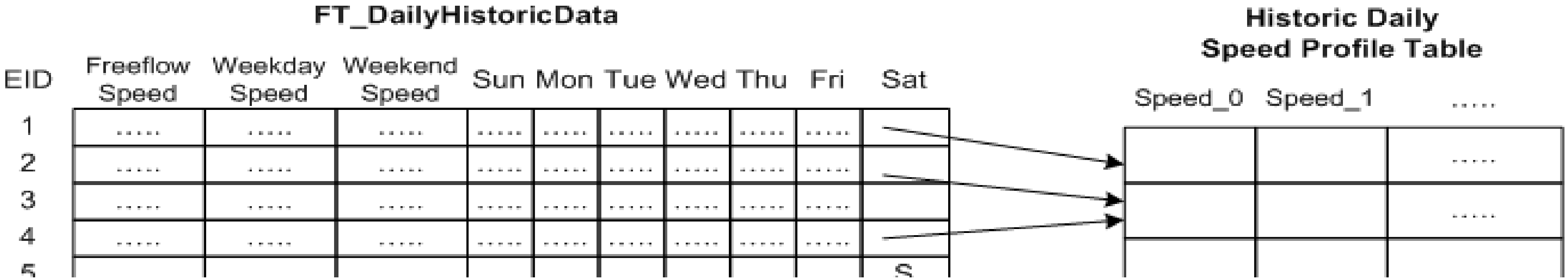
Our Exposures

Our Unique Characteristics

Disease

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



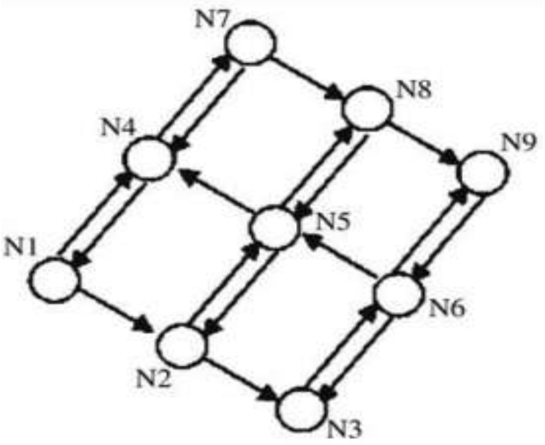
FT_DailyHistoricData

EID	Freeflow Speed	Weekday Speed	Weekend Speed	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	.....	.....	.....	.....	.....	.....	.....	.....	.....	
2	.....	.....	.....	.....	.....	.....	.....	.....	.....	
3	.....	.....	.....	.....	.....	.....	.....	.....	.....	
4	.....	.....	.....	.....	.....	.....	.....	.....	.....	
5										S

Historic Daily Speed Profile Table

Speed_0	Speed_1	.....
		.....
		.....

SE 5th St
5th Ave SE
6th Ave SE
SE 4th St
SE University Ave
7th Ave SE



Nodes

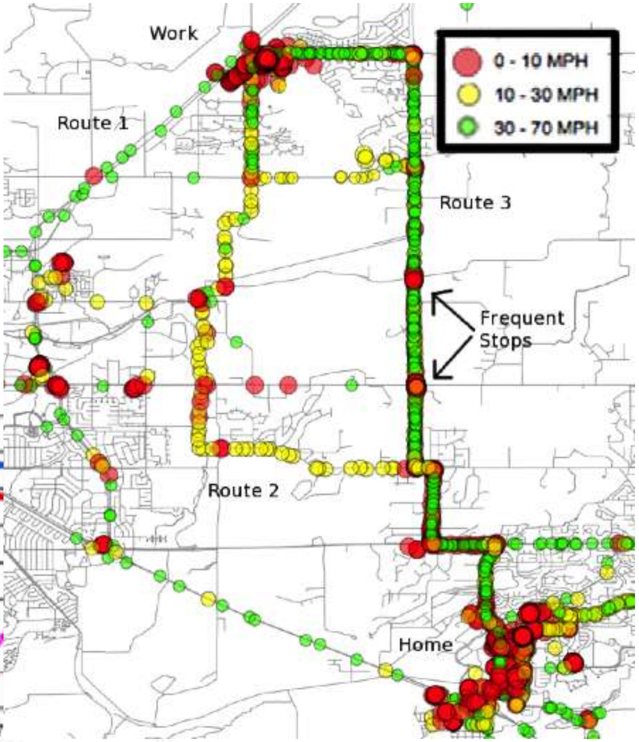
NID
N1
N2
N3
N4
N5
N6
N7
N8
N9

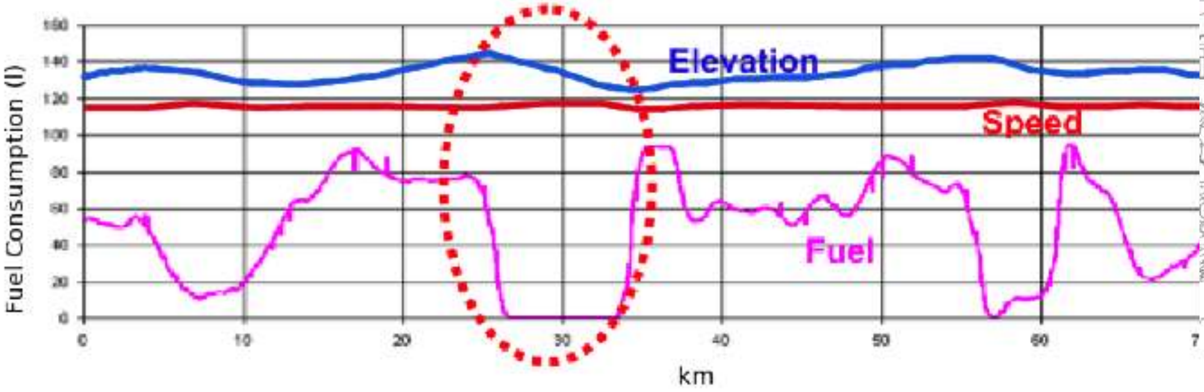
Edges

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

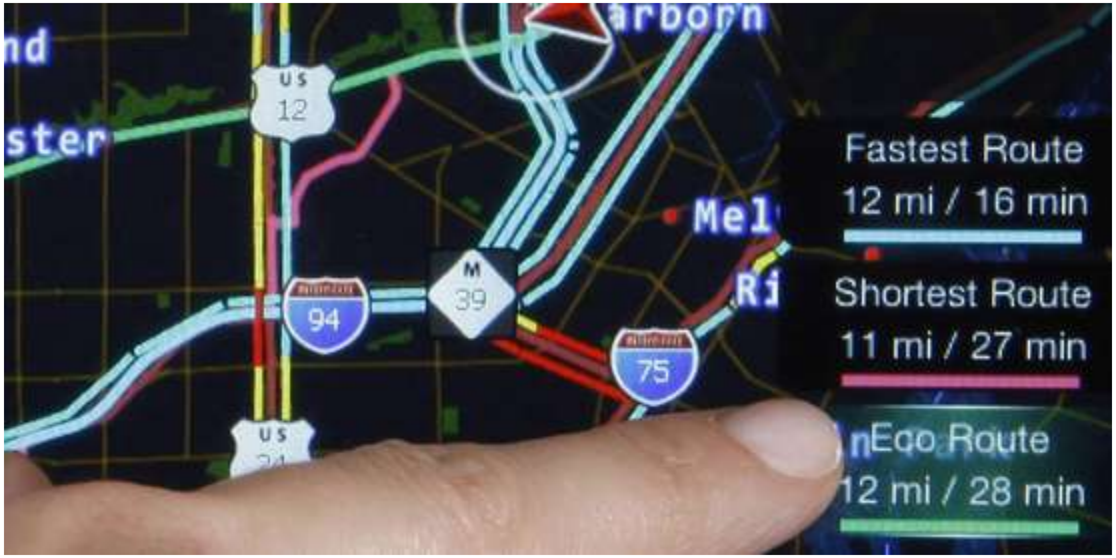
- Mobile Device
  - Cell-phones, cars, trucks, airplanes, …
  - RFID-tags, bar-codes, GPS-collars, …
- Trajectory & Measurements sub-genre
  - Receiver: GPS tracks, …
  - System: Cameras, RFID readers, …
- Use cases:
  - Tracking, Tracing,
    - Improve service, deter theft …
  - Geo-fencing, Identify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





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.



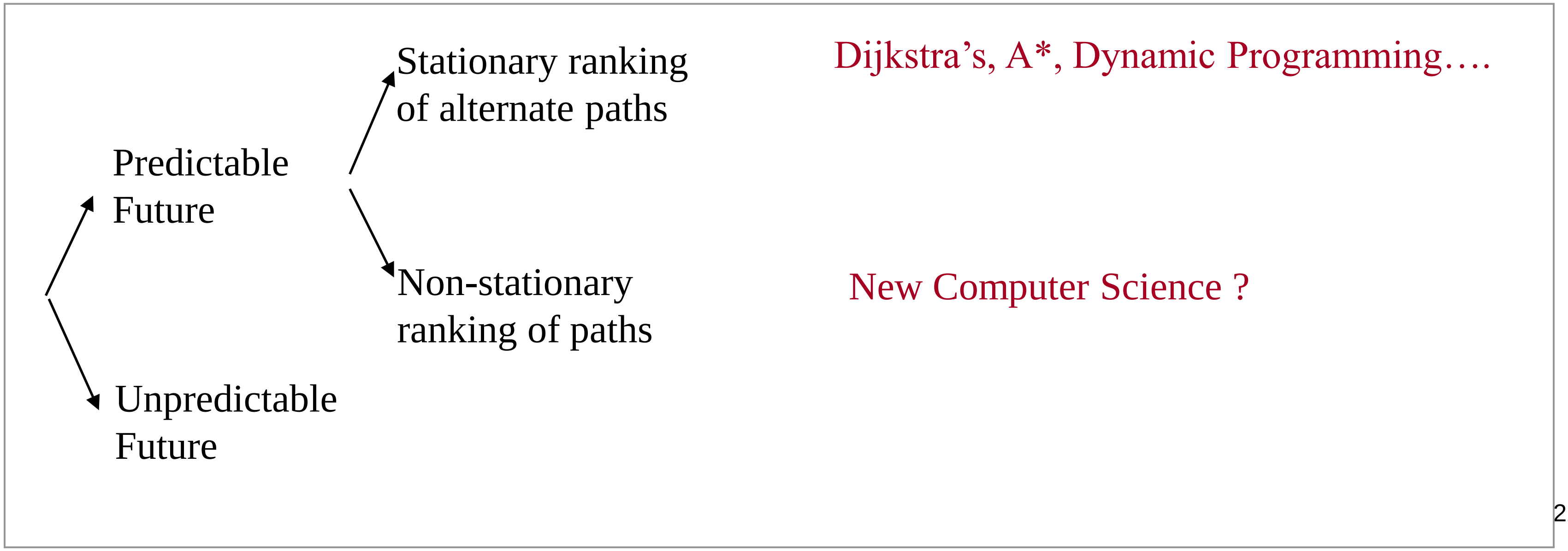
The New York Times

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?



Predictable Future

Stationary ranking of alternate paths

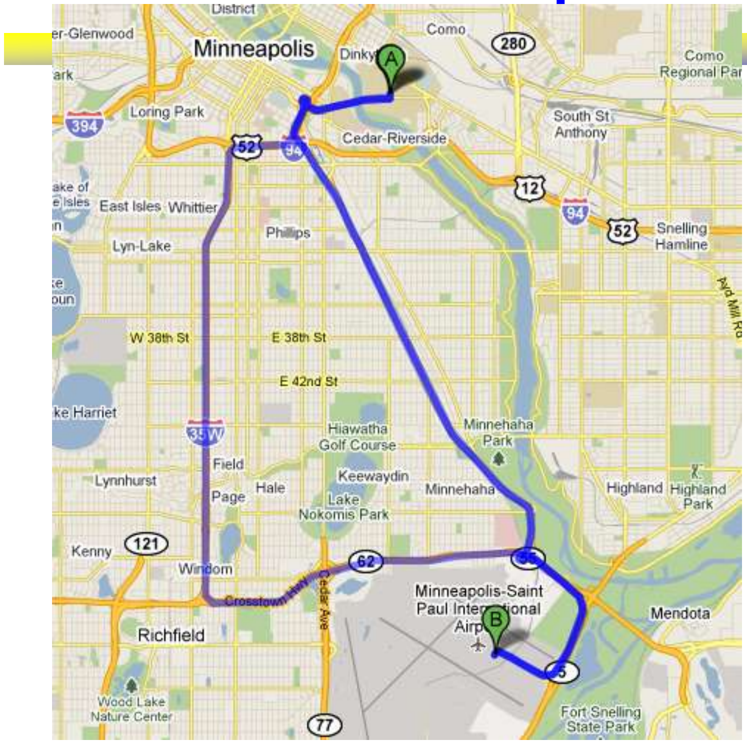
Dijkstra’s, A*, Dynamic Programming….

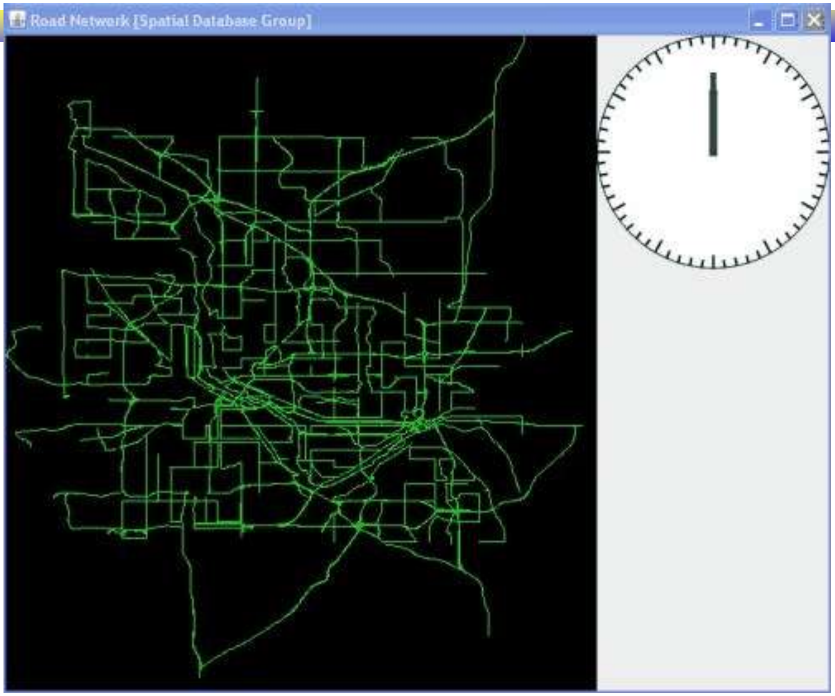
Non-stationary ranking of paths

New Computer Science ?

Unpredictable Future

Time Dependence of Shortest Path





INPUT:

- **Source:** University of Minnesota
- **Destination:** MSP Airport
- **Time Interval** 7:00am ---12:00noon

OUTPUT:

Time	Preferred Routes
7:30am	Via Hiawatha
8:30am	Via Hiawatha
9:30am	via 35W
10:30am	via 35W

Routing Algorithm Challenges

Non Stationarity ranking of paths

Time	Preferred Routes
7:30am	Via Hiawatha
8:30am	Via Hiawatha
9:30am	via 35W
10:30am	via 35W

- Violation of stationary assumption dynamic programming

Non FIFO Behavior

Time	Route	Flight Time
8:30am	via Detroit	6 hrs 31 mins
9:10am	direct flight	2 hrs 51 mins
11:00am	via Memphis	4 hrs 38mins
11:30am	via Atlanta	6 hrs 28 mins
2:30pm	direct flight	2 hrs 51 mins

*Flight schedule between Minneapolis and Austin (TX)

- Violates the no wait assumption of Dijkstra/A*

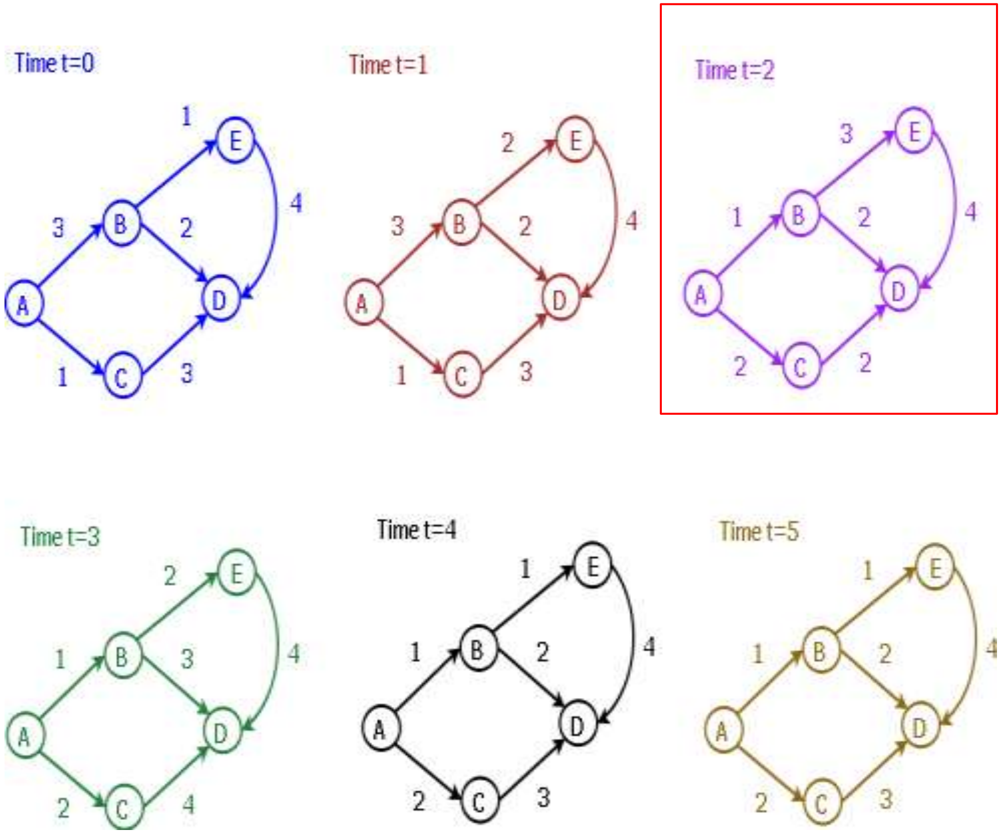
Routing Challenges: Lagrangian Frame of Reference

Question: What is the cost of Path <A,C,D> with start-time t=2 ?

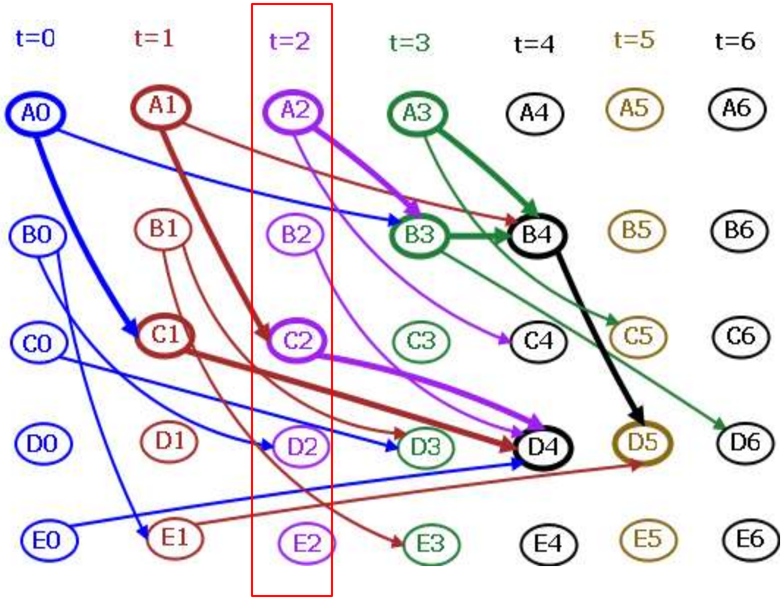
- **Is it 4 or 5 ?**

Path	T = 0	T = 1	T = 2	T = 3
<A,C,D>	4	3	5	4
<A,B,D>	6	5	4	3

Snapshots of a Graph



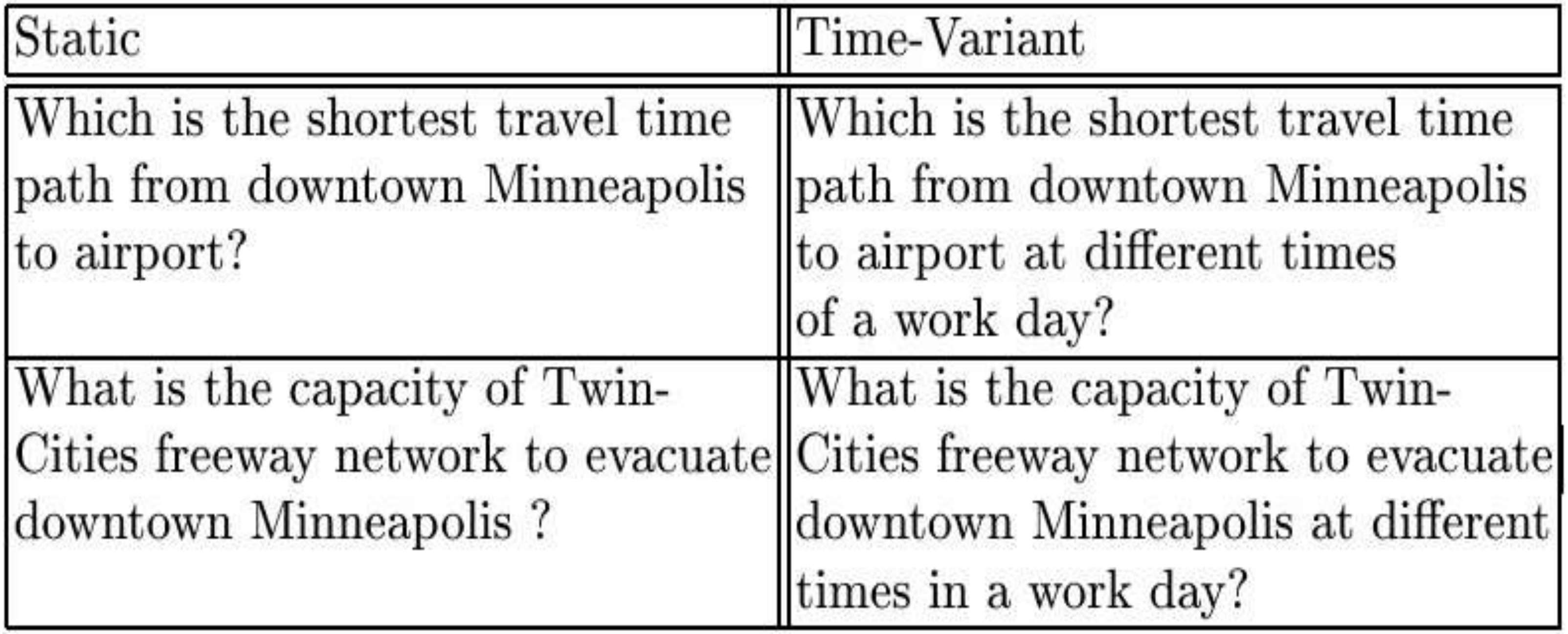
Lagrangian Graph



Challenges: New Routing Questions

- New Routing Questions
  - Best start time to minimize time spend on network
  - Account for delays at signals, rush hour, etc.

- Example Time-Variant Network Questions

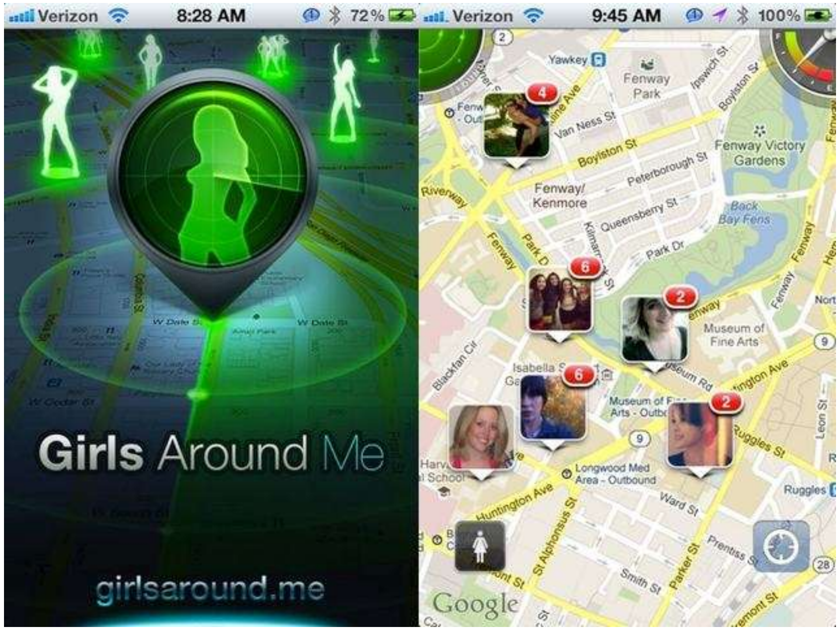


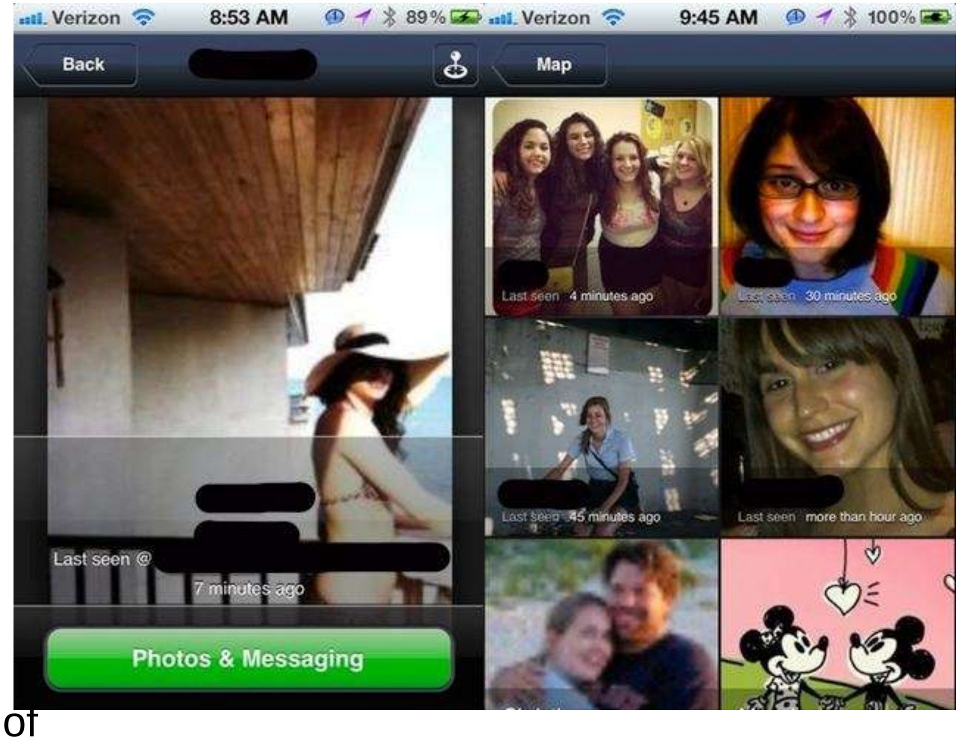
Static	Time-Variant
Which is the shortest travel time path from downtown Minneapolis to airport?	Which is the shortest travel time path from downtown Minneapolis to airport at different times of a work day?
What is the capacity of Twin-Cities freeway network to evacuate downtown Minneapolis ?	What is the capacity of Twin-Cities freeway network to evacuate downtown Minneapolis at different times in a work day?

SBD Challenge: Privacy vs. Utility Trade-off

- Checkin Risks: Stalking, GeoSlavery, …
- Ex: Girls Around me App (3/2012), Lacy Peterson [2008]
- Others know that you are not home!

NATIONAL STALKING AWARENESS MONTH
JANUARY 2012
StalkingAwarenessMonth.org




Verizon 8:53 AM 89%
Back
[redacted]
Last seen @ [redacted]
7 minutes ago
Photos & Messaging
Verizon 9:45 AM 100%
Map
[redacted] Last seen 4 minutes ago
[redacted] Last seen 30 minutes ago
[redacted] Last seen 45 minutes ago
[redacted] Last seen more than hour ago

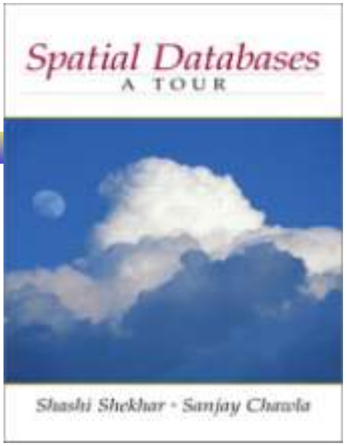
The Girls of Girls Around Me. It's doubtful any of these girls even know they are being tracked.

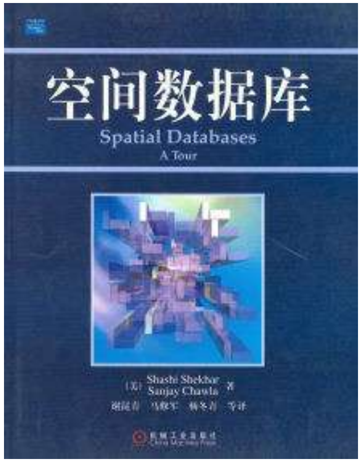
Their names and locations have been obscured for privacy reasons. (Source: Cult of Mac, March 30, 2012)

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





Шаши Шекхар, Санжей Чаула
Основы пространственных баз данных
учебное пособие
Геоинформационные системы • САПР • Мультимедиа

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

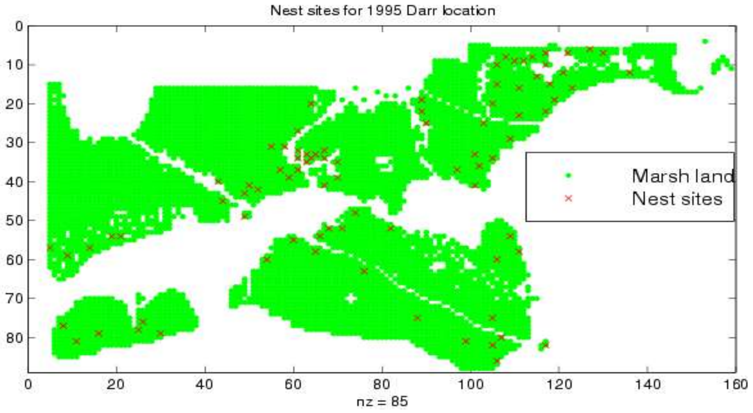
November 14, 2004 The New York Times

What Wal-Mart Knows About Customers' Habits

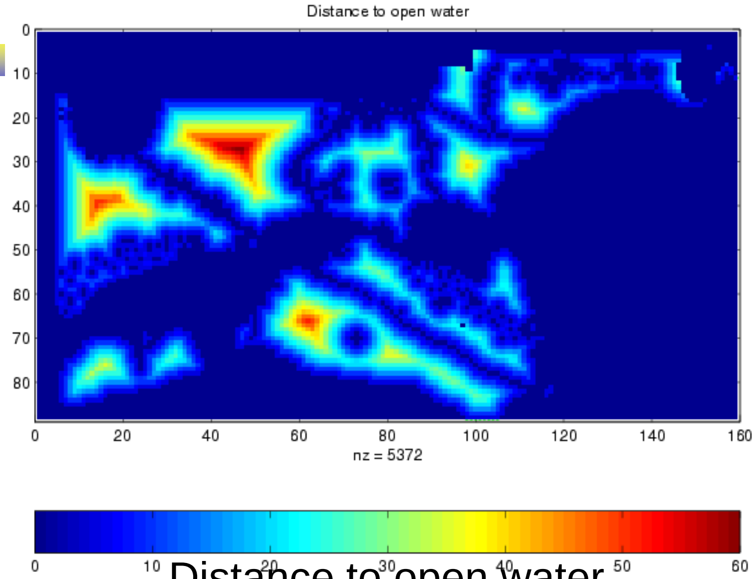
- 1990s: Data Mining
  - Scale up traditional models (e.g., Regression) to large relational databases (460 Tbytes)
  - New pattern families: Associations : Which items are bought together? (Ex. Diaper, beer)
- Spatial customers
  - Walmart: Which items are bought just before/after events, e.g. hurricanes?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

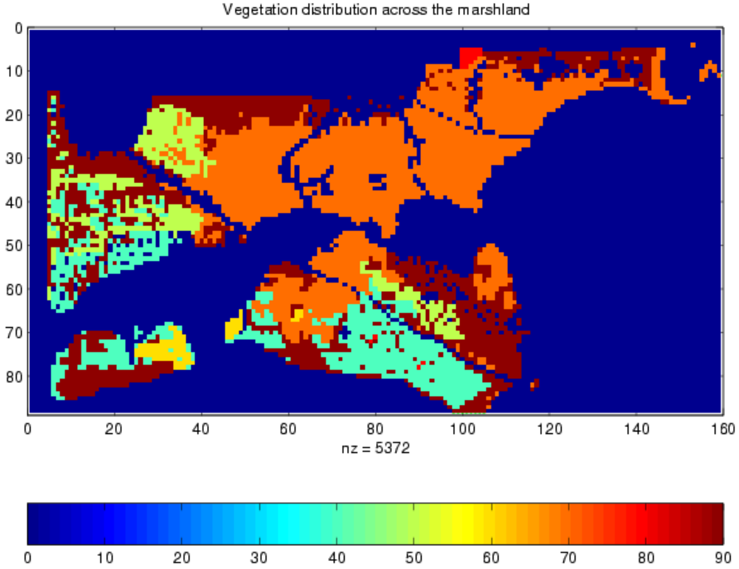
Spatial Data Mining Example 1: Spatial Prediction

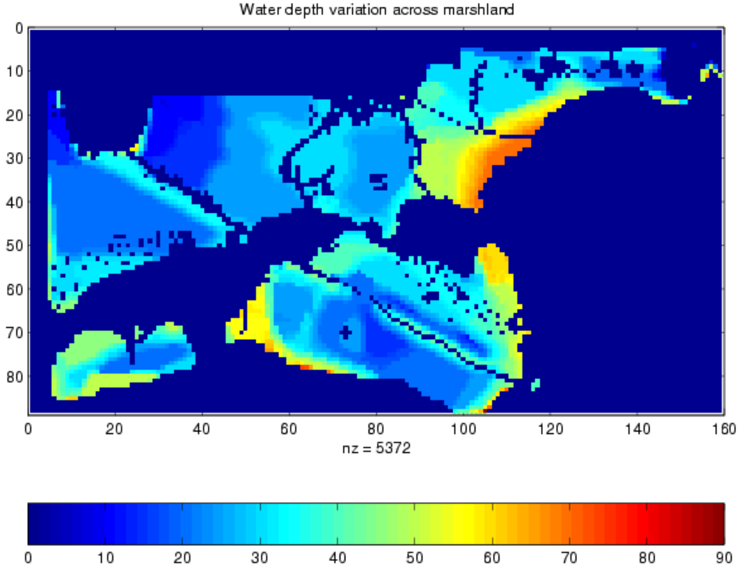


Nest locations



Distance to open water



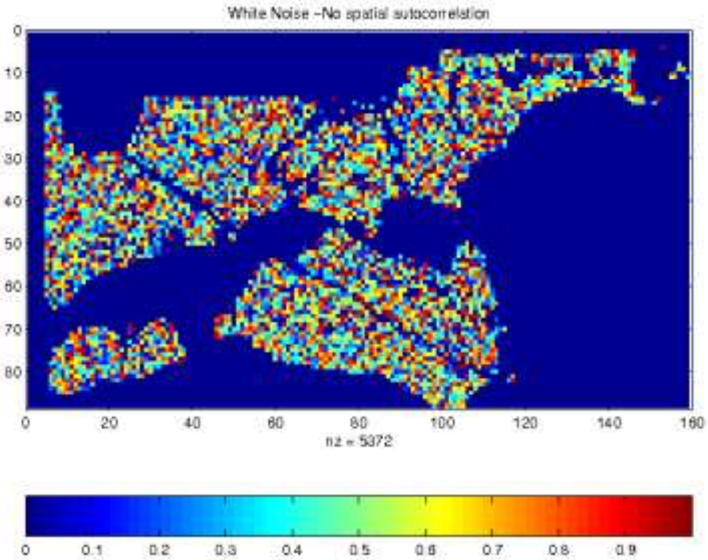


Vegetation durability

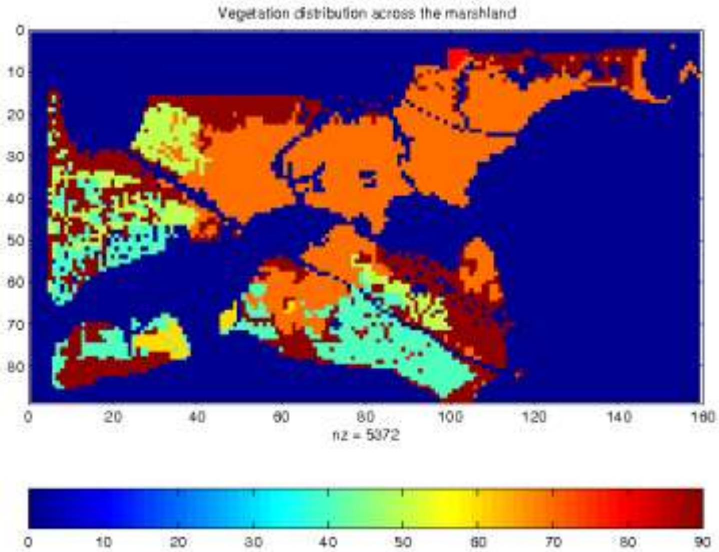
Water depth

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

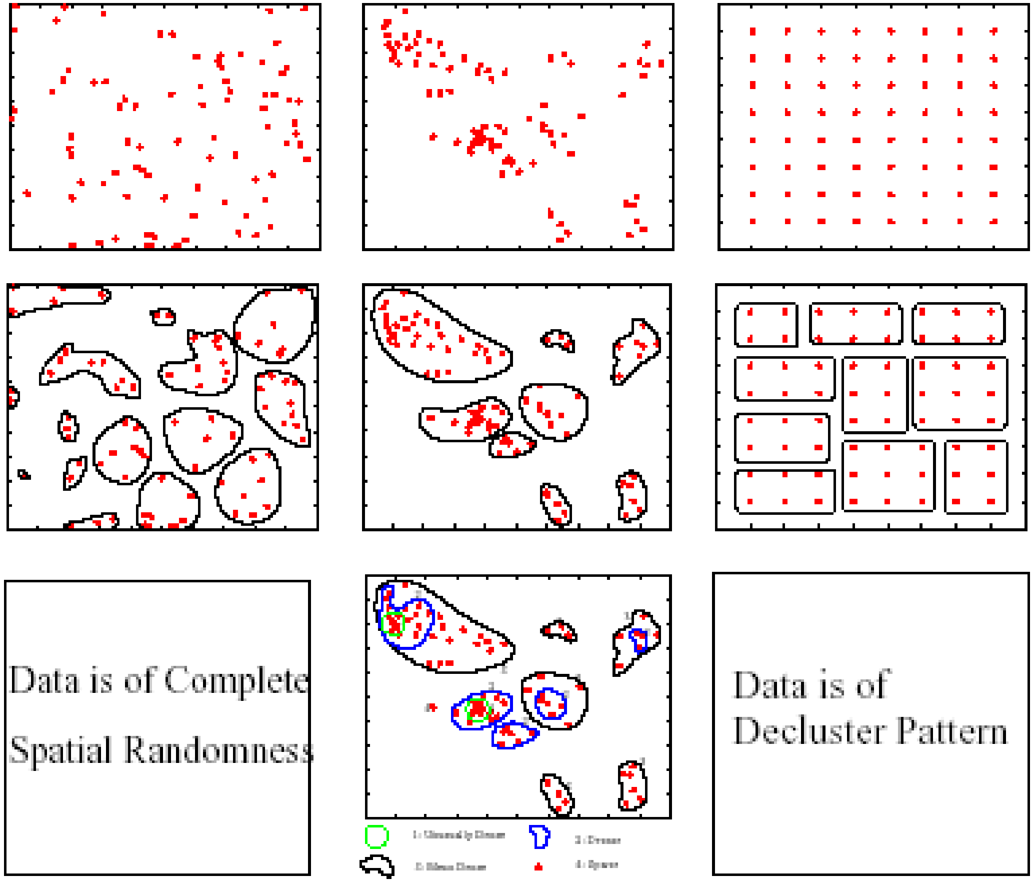
Name	***Model***
Classical Linear Regression	$\boxed{\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$
Spatial Auto-Regression	$\boxed{\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$

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

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

Spatial Data Mining Example 2: Clustering

- Clustering: Find groups of tuples
- Statistical Significance
  - Complete spatial randomness, cluster, and de-cluster


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

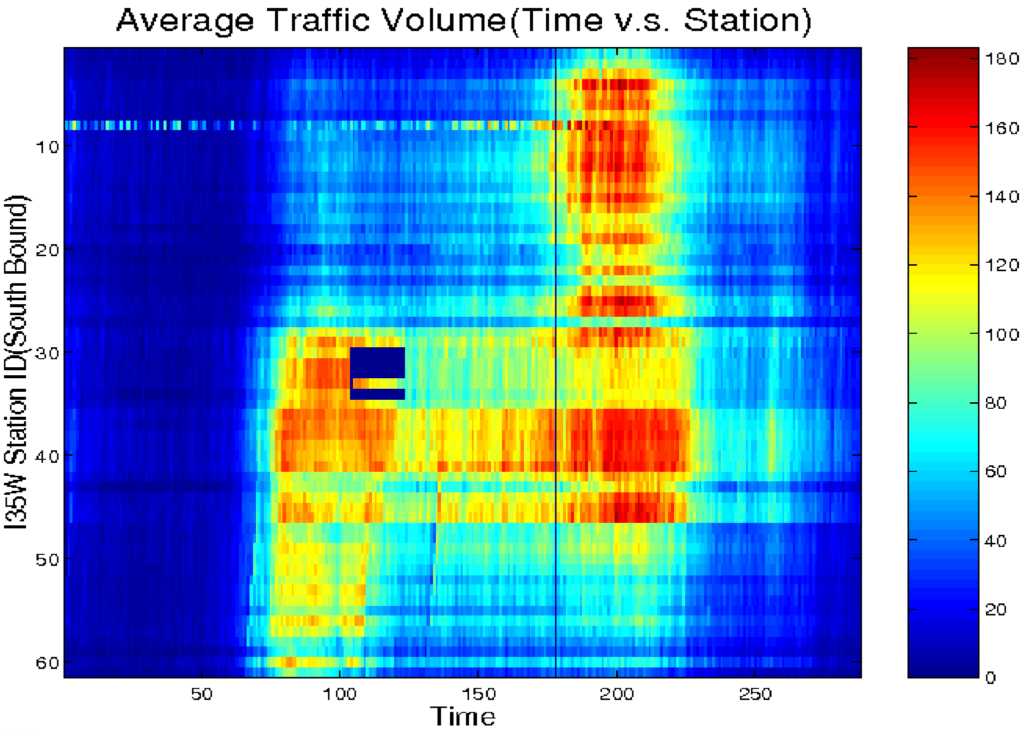
Inputs:
Complete Spatial Random (CSR),
Cluster,
Decluster

Classical Clustering
(K-means always finds clusters)

Spatial Clustering begs to differ!
Leftmost: Clusters not significant
Rightmost – declustered!

Spatial Data Mining Example 3: Spatial Outliers

- Spatial Outliers
  - Locations different than neighbors
  - Ex. Sensor 9
  - Source: Traffic Data Cities
- Spatial Join Based Tests



Association Patterns

- Association rule e.g. (Diaper in T => Beer in T)

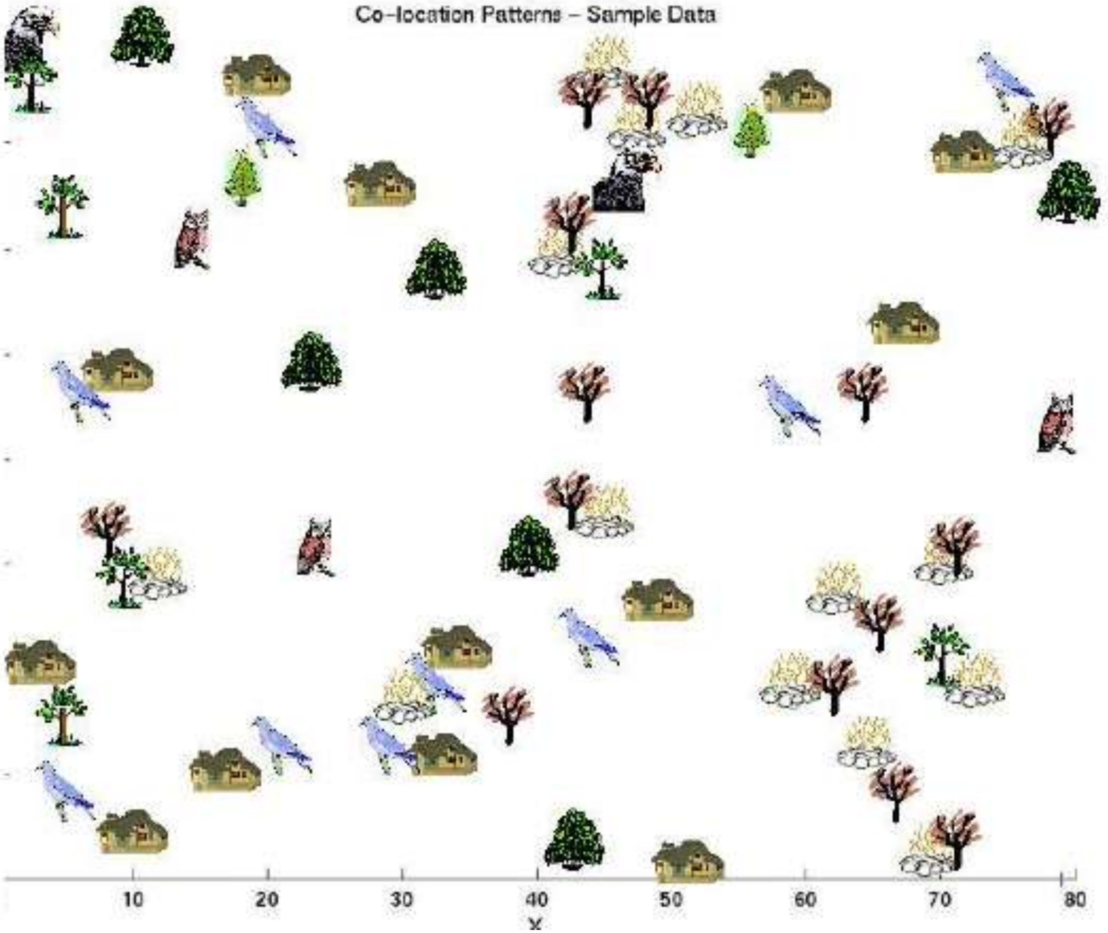
Transaction	Items Bought
1	{socks, , milk, , beef, egg, …}
2	{pillow, toothbrush, ice-cream, muffin, …}
3	{ , , pacifier, formula, blanket, …}
…	…
n	{battery, juice, beef, egg, chicken, …}

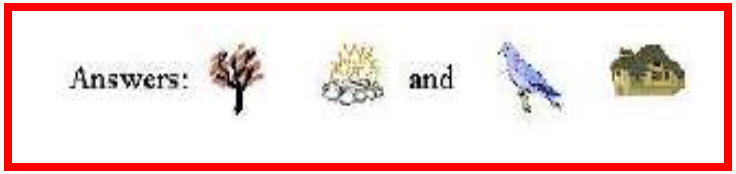
  - Support: probability (Diaper and Beer in T) = 2/5
  - Confidence: probability (Beer in T \| Diaper in T) = 2/2
- Algorithm Apriori [Agarwal, Srikant, VLDB94]
  - Support based pruning using monotonicity
- Note: **Transaction is a core concept!**

Spatial Data Mining Example 4: Co-locations/Co-occurrence

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

  No Transactions
- New Approaches
  - Spatial Join Based
  - One join per candidate is Computationally expensive!





Outline

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

SBD Opportunities 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 ?

SBD Opportunity 3: Bigger the SBD, Simpler the Model

- SDM Limitation 3: Complexity of Spatial Models
  - Spatial models are usually computationally more expensive than traditional models
  - Example: Spatial Auto-regression vs. Linear Regression
  - Example: Geographically Weighted Regression vs. Regression
  - Example: Colocation Pattern Mining vs. Association Rule Mining
  - Confidence Estimation adds more costs, e.g. M.C.M.C. simulations
- SBD Opportunity 3: Bigger the SBD, Simpler the spatial models
  - Sometime the complexity is due to paucity of data at individual places
    - It forces one to leverage data at nearby places via spatial auto-correlation and spatial join!
  - SBD may provide plenty of data at each place!
    - This may allow place-based divide and conquer
    - Build one model per place using local data and simpler model
  - Challenges:
    - Compare place-based ensembles of simpler models with current spatial models
    - When does bigger data lead to simpler models?
    - What is SBD from analytics perspective ?
      - e.g., ratio of samples to number of parameters

SBD Opportunity 4:On-line Spatio-temporal Data Analytics

- SDM Limitation 4: Off-line Batch Processing
  - Spatial models are usually learned in off-line batch manner
    - Example: Spatial Auto-regression, Colocation Pattern Mining, Hotspot Detection
  - However, SBD include streaming data such as event reports, sensor measurements
  - SBD use cases include monitoring and surveillance requiring On-Line algorithms
    - Example: Timely Detection of Outbreak of disease, crime, unrest, adverse events
    - Example: Displacement or spread of a hotspot to neighboring geographies
    - Example: Abrupt or rapid change detection in land-cover, forest-fire, etc. for quick response
- SBD Opportunity 4: On-line Spatio-temporal Data Analytics
  - Purely local models may leverage time-series data analytics models
  - Regional and global models are harder
    - Spatial interactions (e.g., colocation, tele-connections) with time-lags
    - Q? Can these be computed exactly in on-line manner?
    - Q? If not, what are on-line approximations ?

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!

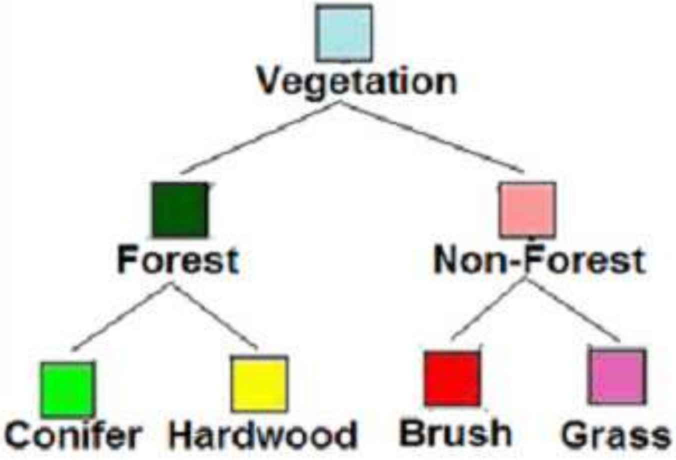
- Case 3: Estimate Spatial Auto-Regression Parameters, Routing
  - Map-reduce is inefficient for iterative computations due to expensive “reduce”!
  - MPI, OpenMP, Pregel or Spatial Hadoop is essential!
  - Ex. Golden section search, Determinant of large matrix
  - Ex. Eco-routing algorithms, Evacuation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

Simpler: Land-cover Classification

- Multiscale Multigranular Image Classification into land-cover categories

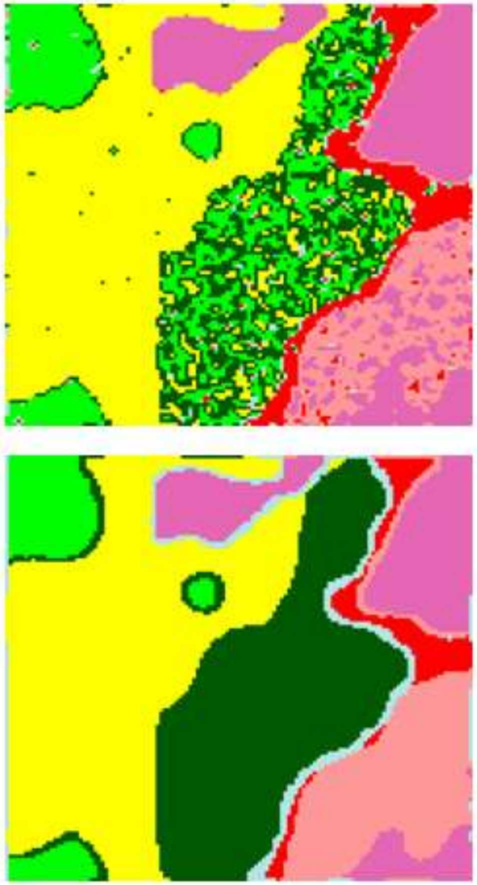
Inputs

Remotely-sensed Satellite Image



Multigranular Land-use Classification

Output at 2 Scales

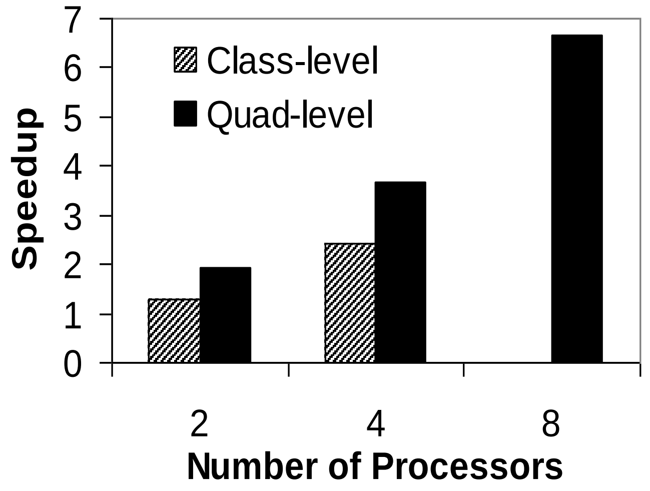


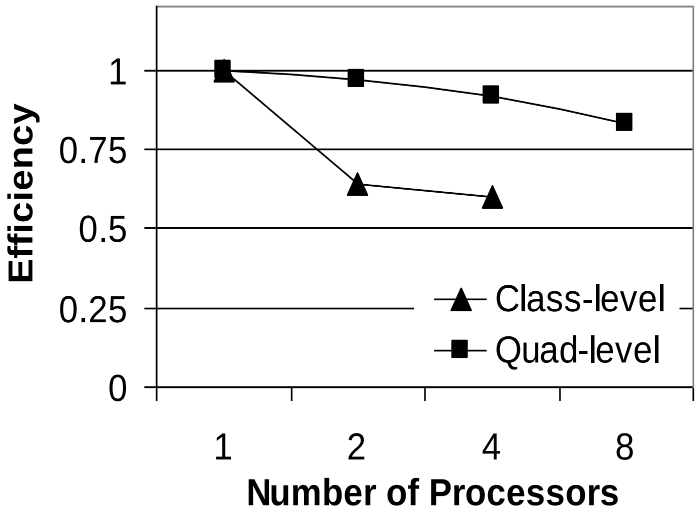
$$\hat{Model} \equiv \arg\max_{Model}\{\, quality(Model)\}, where$$

$$quality(M) = likelihood\,(observation \mid M) - 2\,penalty(M)$$

Parallelization Choice

```
Initialize parameters and memory
for each Spatial Scale
    for each Quad
        for each Class
            Calculate Quality Measure
        end for Class
    end for Quad
end for Spatial Scale
Post-processing
```

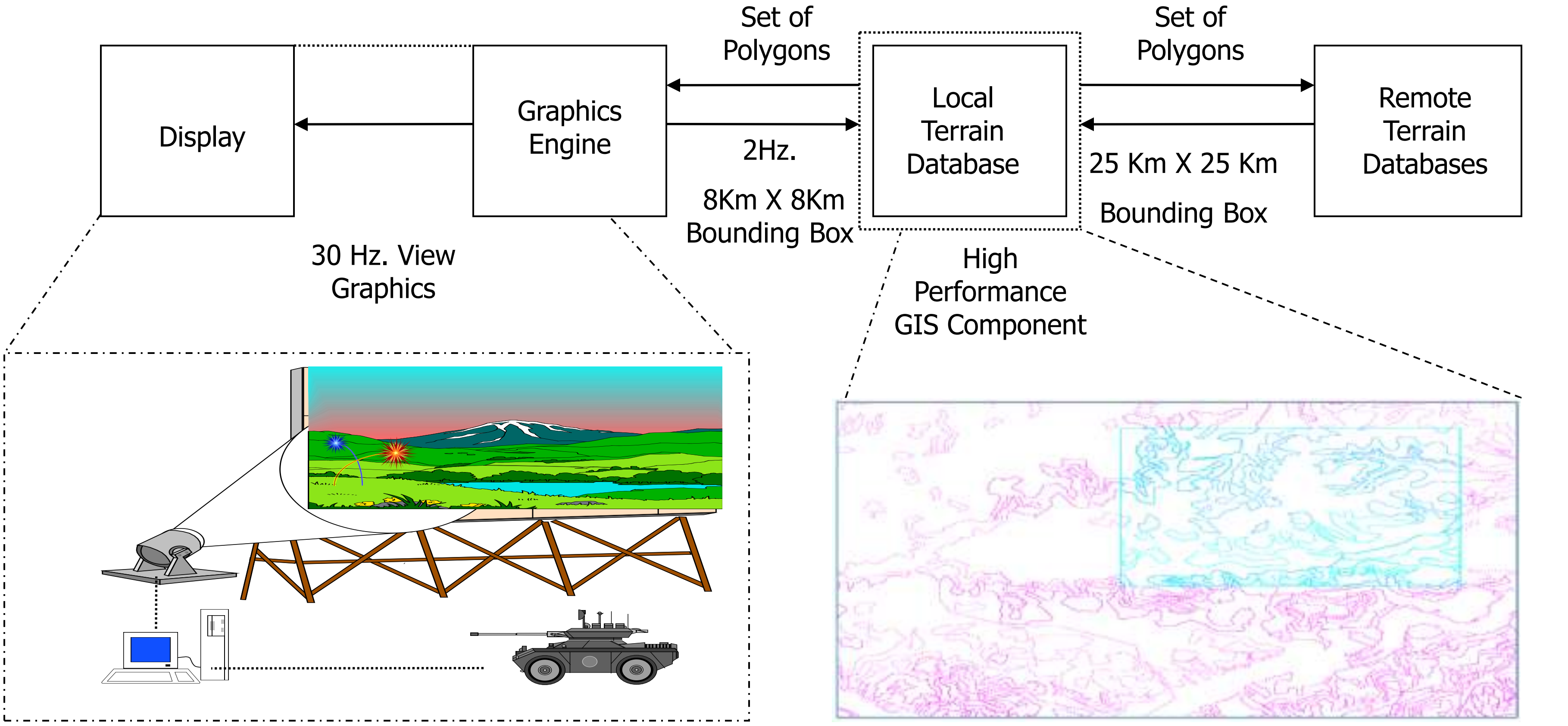




Input	• 64 x 64 image (Plymouth County, MA) • 4 classes (All, Woodland, Vegetated, Suburban)
Language	UPC
Platform	Cray X1, 1-8 processors)

Harder: Parallelizing Vector GIS

- (1/30) second Response time constraint on Range Query
- Parallel processing necessary since best sequential computer cannot meet requirement
- Blue rectangle = a range query, Polygon colors shows processor assignment



Display

Graphics Engine

Local Terrain Database

Remote Terrain Databases

Set of Polygons

2Hz.

8Km X 8Km Bounding Box

Set of Polygons

25 Km X 25 Km

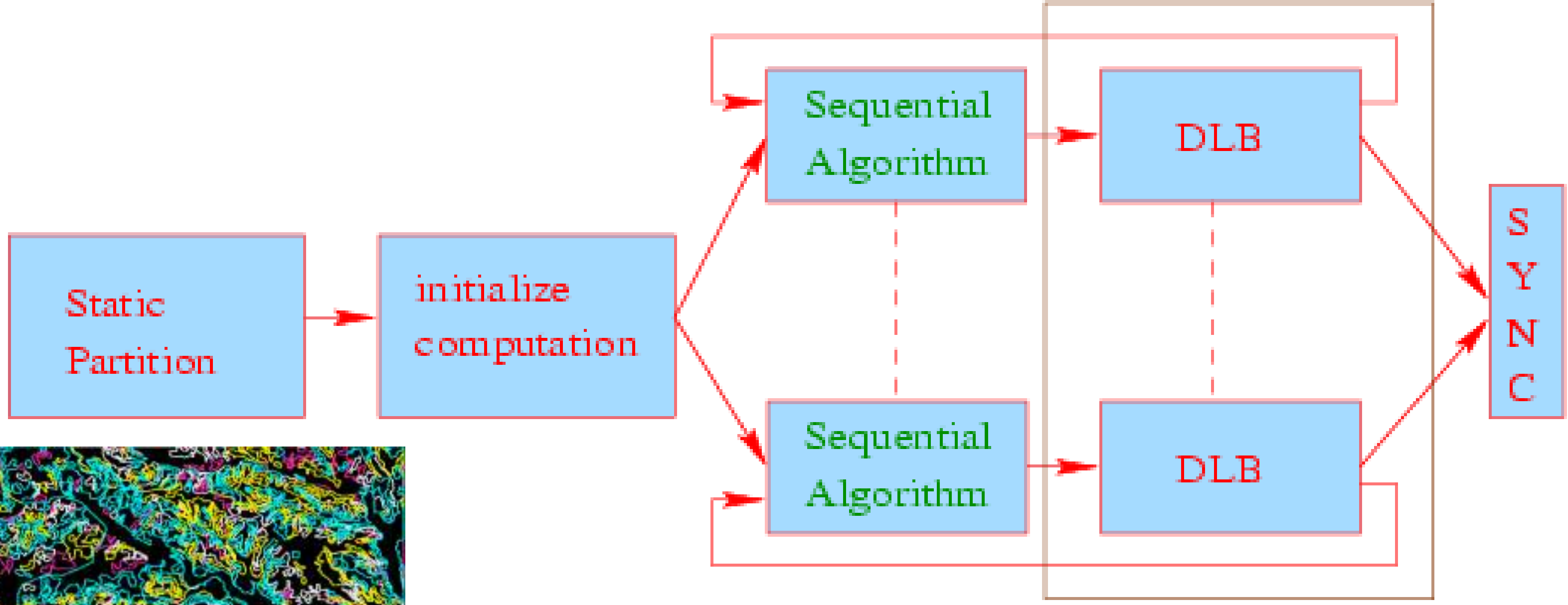
Bounding Box

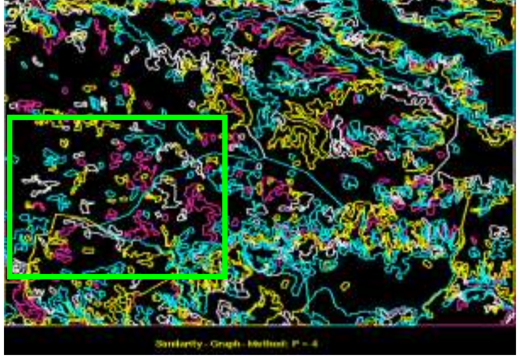
30 Hz. View Graphics

High Performance GIS Component

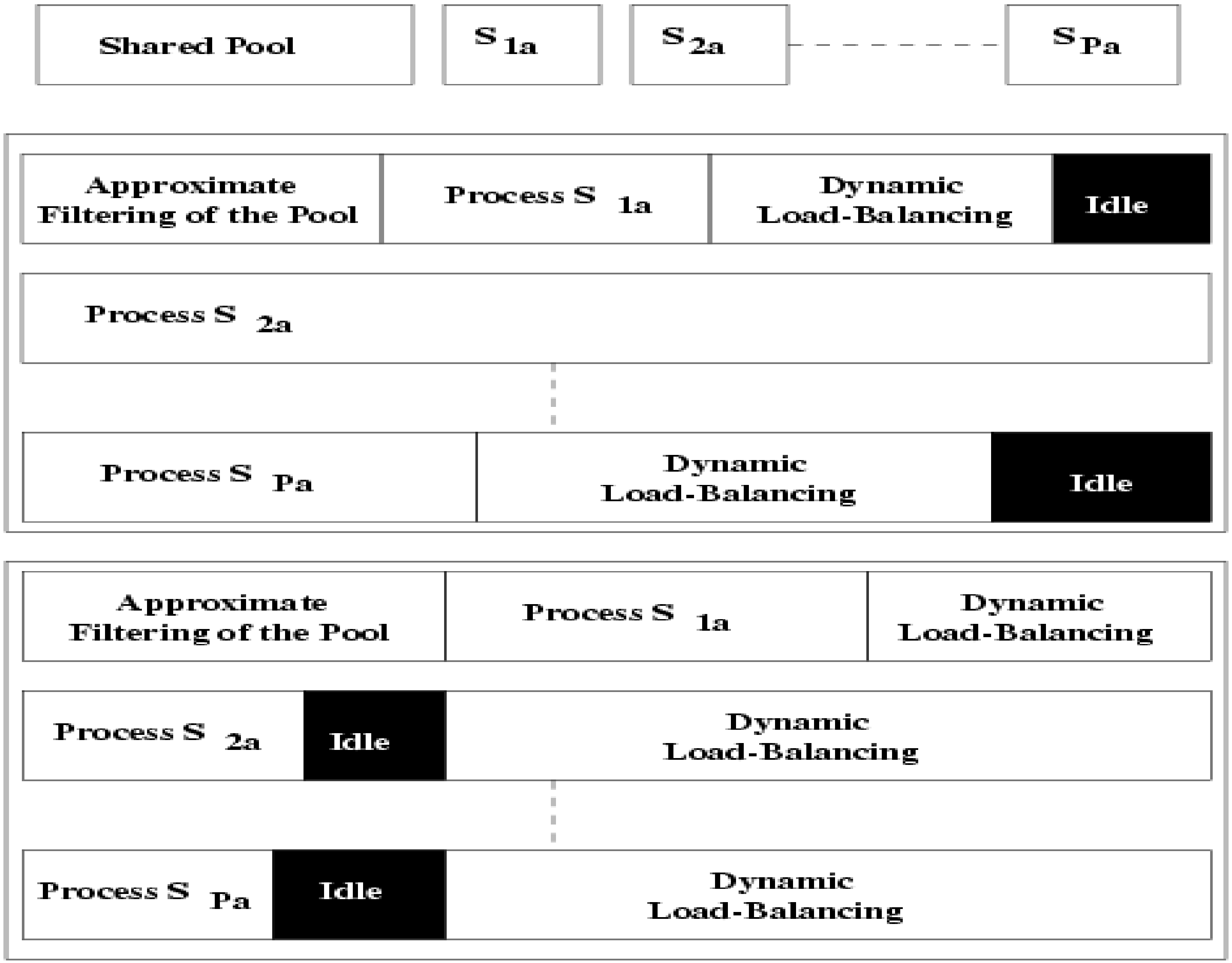
Data-Partitioning Approach

- Initial Static Partitioning
- Run-Time dynamic load-balancing (DLB)
- Platforms: Cray T3D (Distributed), SGI Challenge (Shared Memory)





DLB Pool-Size Choice is Challenging!



Shared Pool

S_{1a}

S_{2a}

S_{Pa}

Approximate Filtering of the Pool

Process S_{1a}

Dynamic Load-Balancing

Idle

Process S_{2a}

Process S_{Pa}

Dynamic Load-Balancing

Idle

Approximate Filtering of the Pool

Process S_{1a}

Dynamic Load-Balancing

Process S_{2a}

Idle

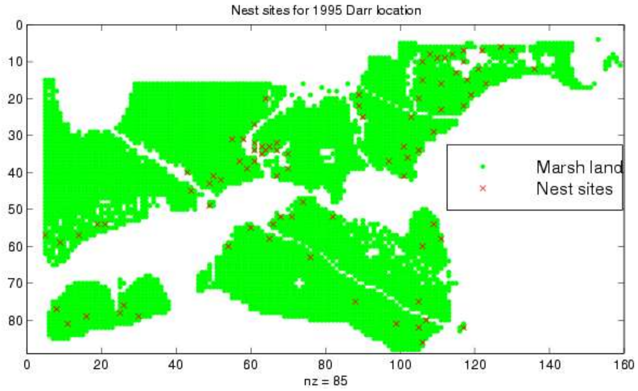
Dynamic Load-Balancing

Process S_{Pa}

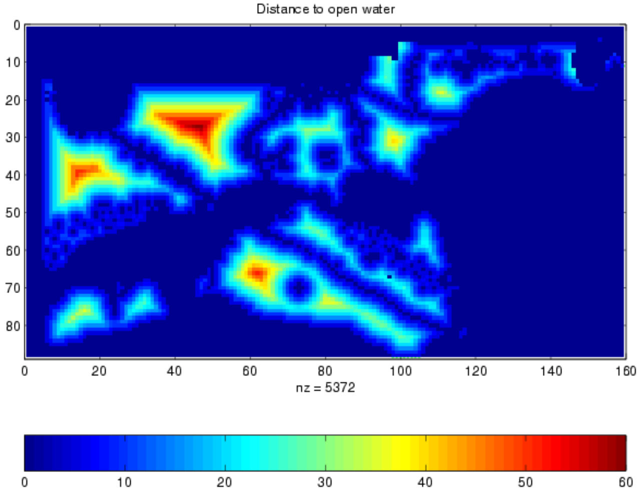
Idle

Dynamic Load-Balancing

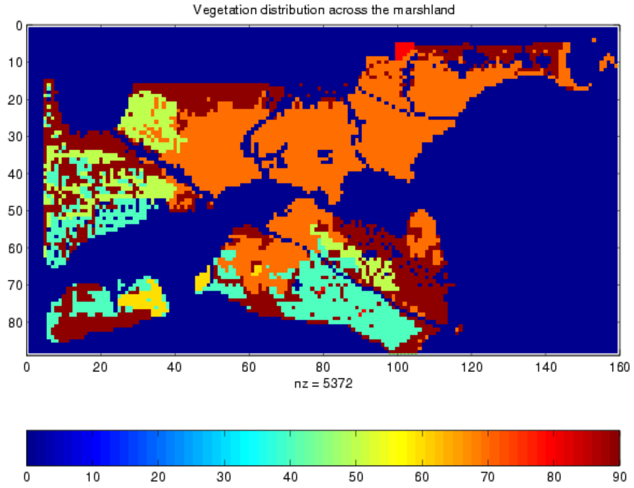
Hardest – Location Prediction



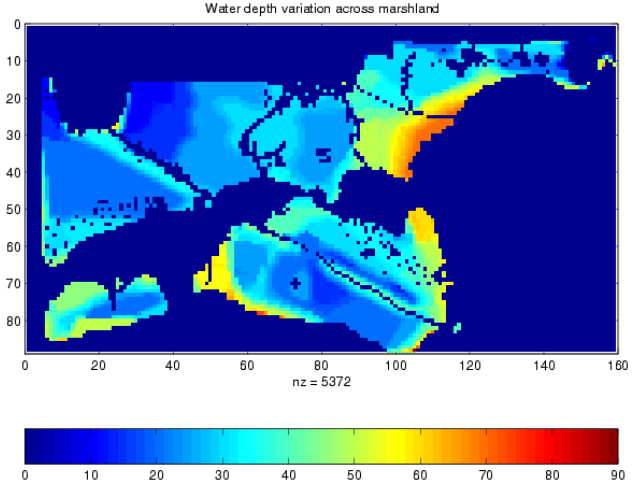
Nest locations



Distance to open water



Vegetation durability



Water depth

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!

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”!
  - Ex. Golden section search, Determinant of large matrix
  - Ex. Eco-routing algorithms, Evacuation route planning
  - Option 1: Develop non-iterative formulations of spatial problems
  - Option 2: Alternative Platform: MPI, OpenMP, Pregel or Spatial Hadoop

Spatial Big Data (SBD)

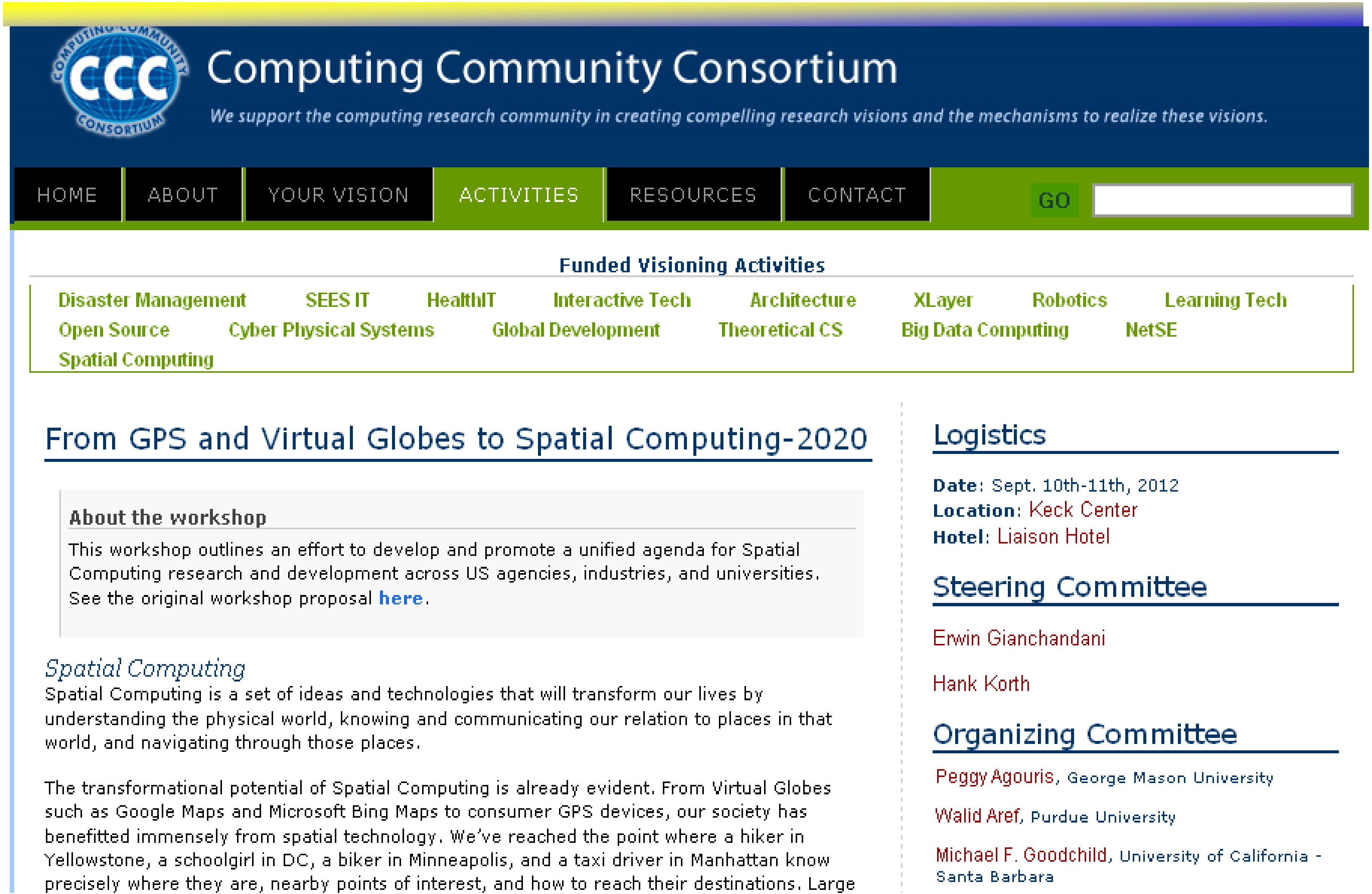
- SBD Definitions
- SBD Applications
- SBD Analytics
- SBD Infrastructure
- Conclusions

Summary

- SBD are important to society
  - Ex. Eco-routing, Public Safety & Security, Understanding Climate Change
- SBD exceed capacity of current computing systems
- DBMS Challenges
  - 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 Challenges
  - Map-reduce – expensive reduce not suitable for iterative computations
  - Load balancing is harder for maps with polygons and line-strings
  - Spatial Hadoop ?

CCC Workshop: Spatial Computing Visioning (9/10-11/2012)

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.

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