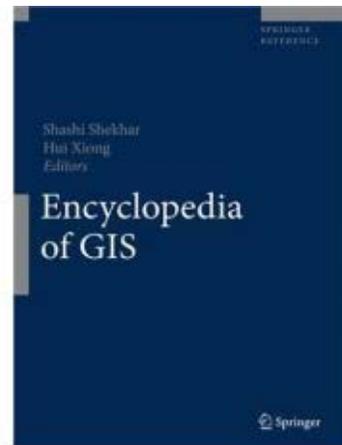
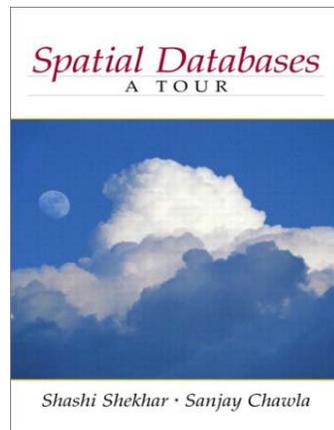


# Gap Analysis for Spatial and Spatio-temporal Data Mining

Shashi Shekhar

McKnight Distinguished University Professor  
Faculty of Computer Sc. and Eng., Univ. of Minnesota  
[www.cs.umn.edu/~shekhar](http://www.cs.umn.edu/~shekhar)

Presented at  
NSF Workshop on GeoSpatial and GeoTemporal Informatics  
January 8<sup>th</sup>-9<sup>th</sup>, 2009  
Washington D.C.



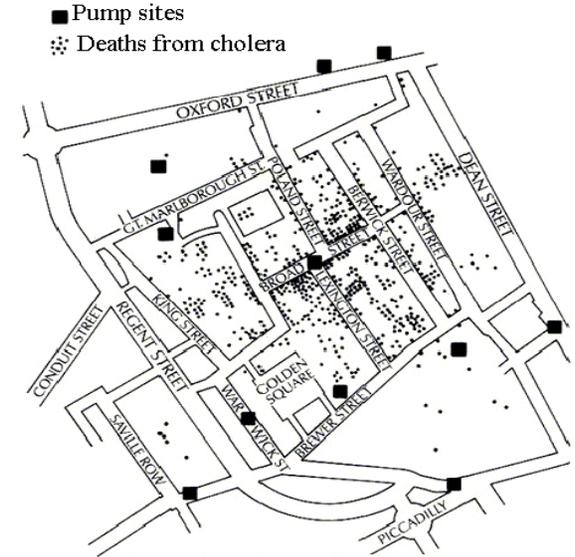
# Spatial and Spatio-temporal Data Mining

## ■ What is it?

- Identifying interesting, useful, non-trivial **patterns**
  - Hot-spots, discontinuities, co-locations, trends, ...
- in large **spatial** or **spatio-temporal** datasets
  - Satellite imagery, geo-referenced data, e.g. census
  - gps-tracks, geo-sensor network, ...

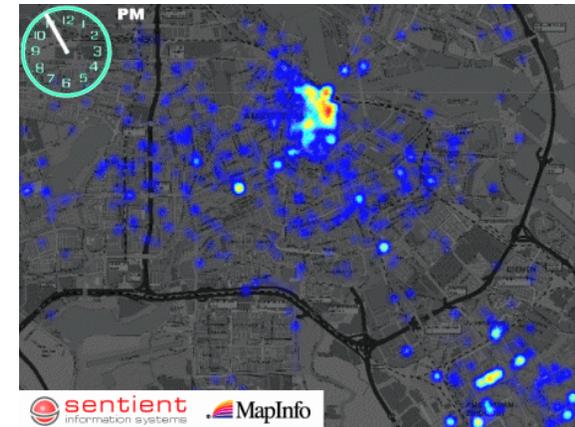
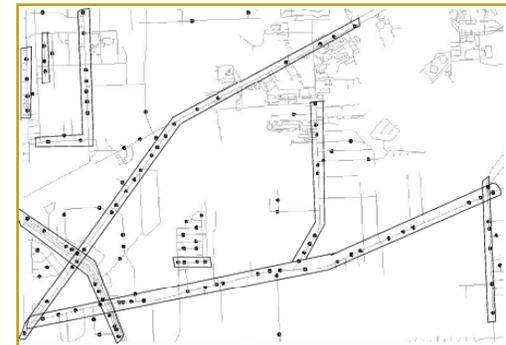
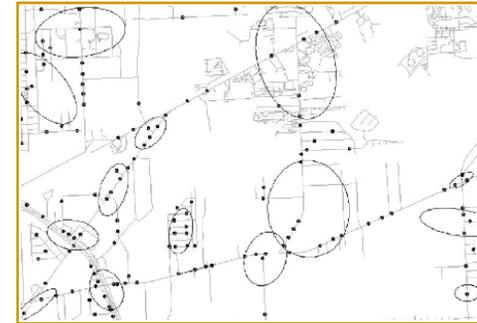
## ■ Why is it important ?

- Potential of discoveries and insights to improve human lives
  - Environment: How is Earth system changing? Consequences for humans?
  - Public safety: Where are hotspots of crime? Why?
  - Public health: Where are cancer clusters? Environmental reasons?
  - Transportation, National Security, ...
- However,  $(d/dt) (\text{Spatial Data Volume}) \gg (d/dt) (\text{Number of Human Analysts})$ 
  - Need automated methods to mine patterns from spatial data
  - Need tools to amplify human capabilities to analyze spatial data



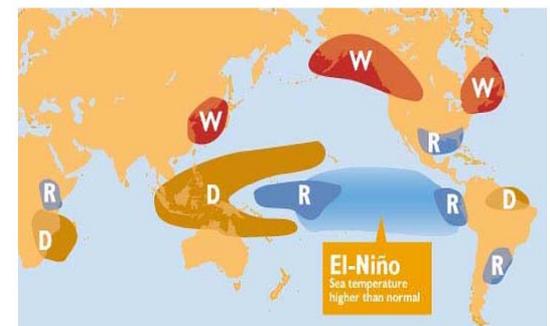
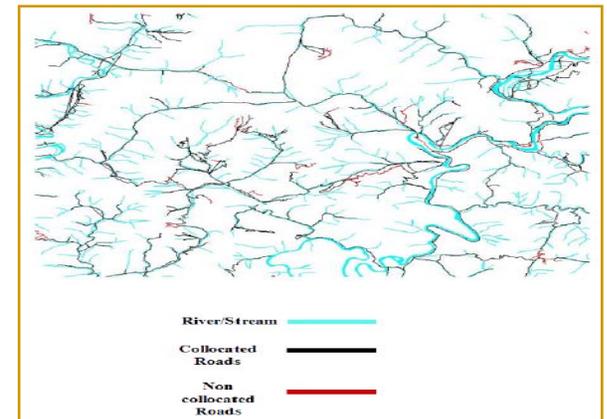
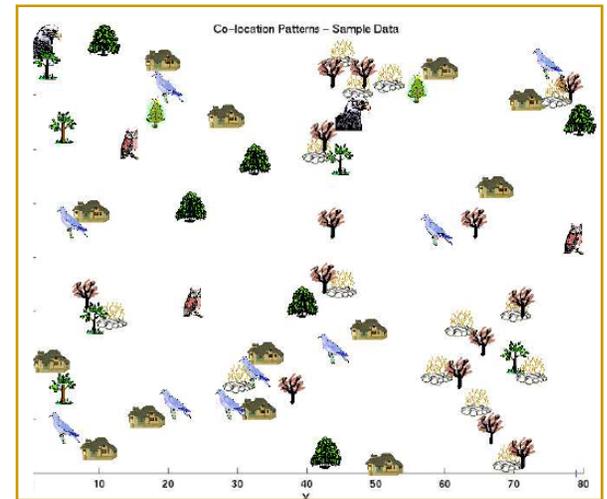
# HotSpots

- What is it?
  - Unusually high spatial concentration of a phenomena
    - Cancer clusters, crime hotspots
- Solved
  - Spatial statistics based ellipsoids
- Almost solved
  - Transportation network based hotspots
- Failed
  - Classical clustering methods, e.g. K-means
- Missing
  - Spatio-temporal
- Next
  - Emerging hot-spots



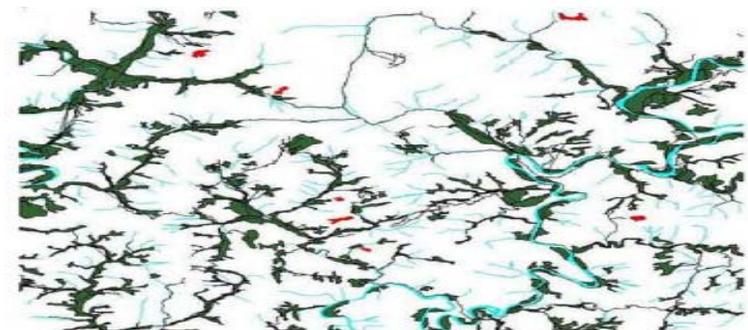
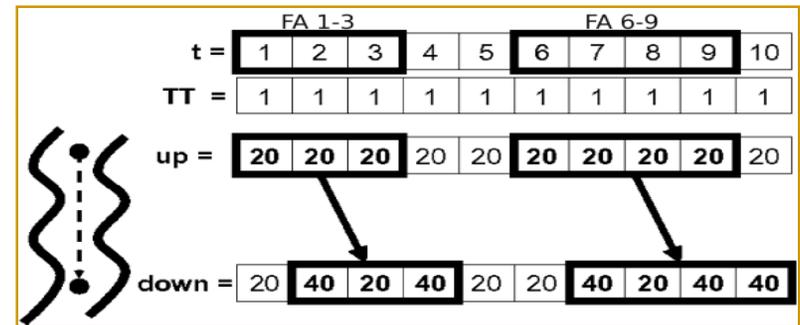
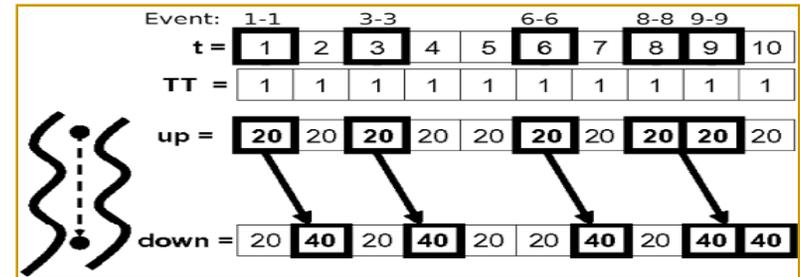
# Colocation, Co-occurrence, Interaction

- What is it?
  - Subset of event types, whose instances occur together
  - Ex. Symbiosis, (bar, misdemeanors), ...
- Solved
  - Colocation of point event-types
- Almost solved
  - Co-location of extended (e.g.linear) objects
  - Object-types that move together
- Failed
  - Neighbor-unaware Transaction based approaches
- Missing
  - Consideration of flow, richer interactions
- Next
  - Spatio-temporal interactions, e.g. item-types that sell well before or after a hurricane
  - Tele-connections

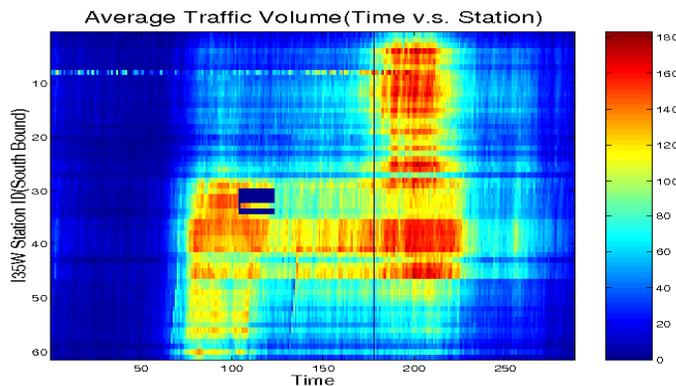


# Spatial/Spatio-temporal Outliers, Anomalies

- What is it?
  - Location different from their neighbors
    - Discontinuities, flow anomalies
- Solved
  - Transient spatial outliers
- Almost solved
  - Anomalous trajectories
- Failed
- Missing
  - Persistent anomalies
  - Multiple object types, Scale
- Next
  - Dominant Persistent Anomalies

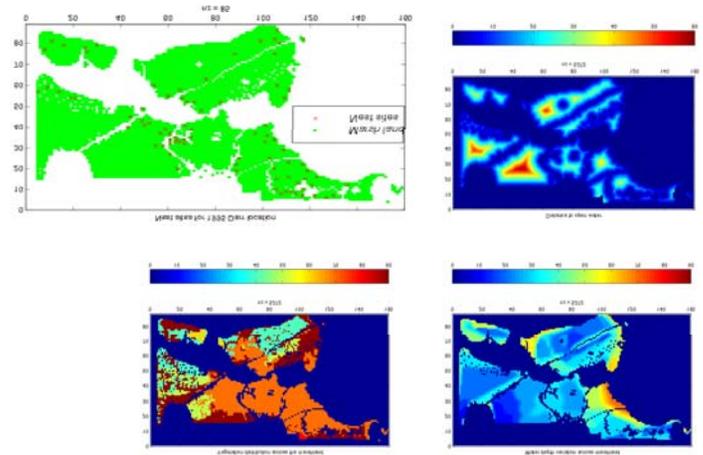


- River/stream —
- Cropland —
- Road —
- Non collocated cropland —



# Space/Time Prediction

- What is it?
  - Models to predict location, time, path, ...
    - Nest sites, minerals, earthquakes, tornadoes, ...
- Solved
  - Interpolation, e.g. Krigging
  - Heterogeneity, e.g. geo. weighted regression
- Almost solved
  - Auto-correlation, e.g. spatial auto-regression
- Failed: Independence assumption
  - Models, e.g. Decision trees, linear regression, ...
  - Measures, e.g. total square error, precision, recall
- Missing
  - Spatio-temporal vector fields (e.g. flows, motion), physics
- Next
  - Scalable algorithms for parameter estimation
  - Distance based errors



$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{x} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

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

