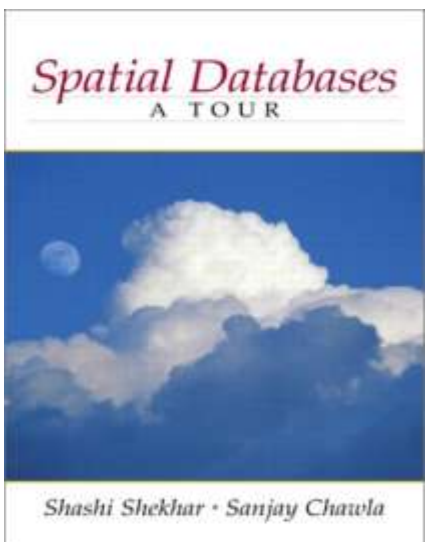
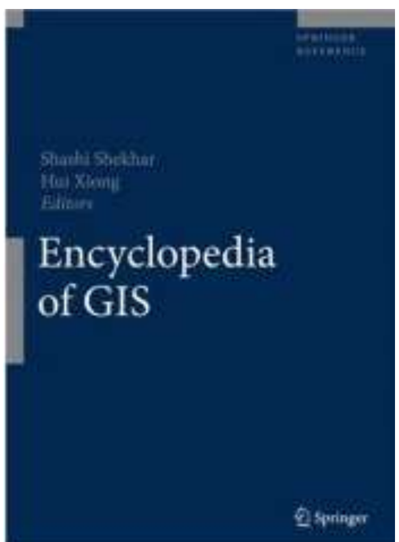
What is Special about Mining Spatial Data in Human Health?

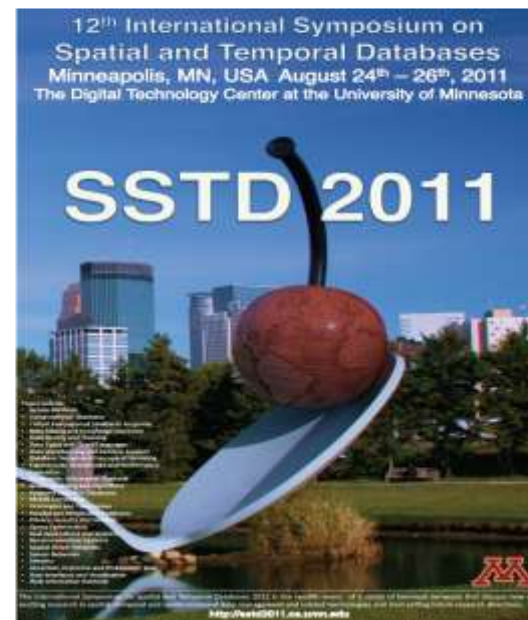
Shashi Shekhar
McKnight Distinguished University Professor
Department of Computer Science and Engineering, University of Minnesota
www.cs.umn.edu/~shekhar

For more details:

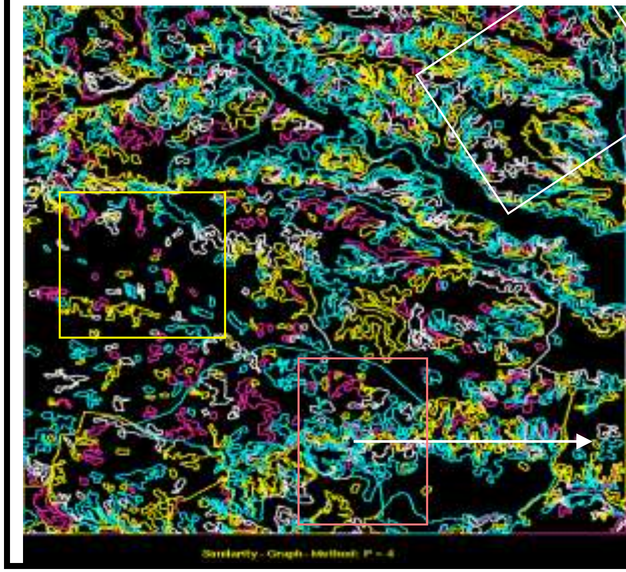
S. Shekhar et al., Identifying patterns in spatial information: A survey of Methods, Wiley Interdisciplinary Reviews in Data Mining and Knowledge Discovery, Volume 1, May/June 2011.





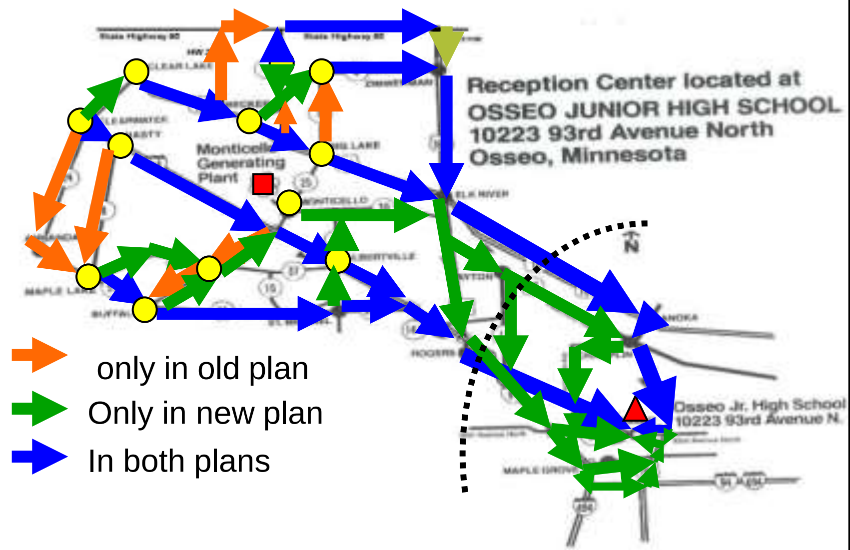


Research Theme 1: Spatial Databases



Parallelize
Range Queries

Evacutation Route Planning

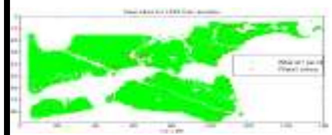


Shortest Paths   Storing graphs in disk blocks

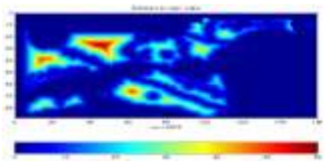
Theme 2 : Spatial Data Mining

Location prediction: nesting sites

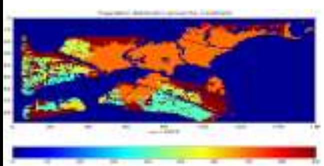
Nest locations



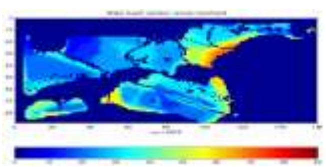
Distance to open water



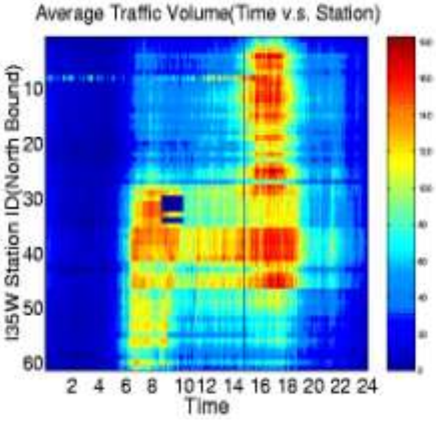
Vegetation durability



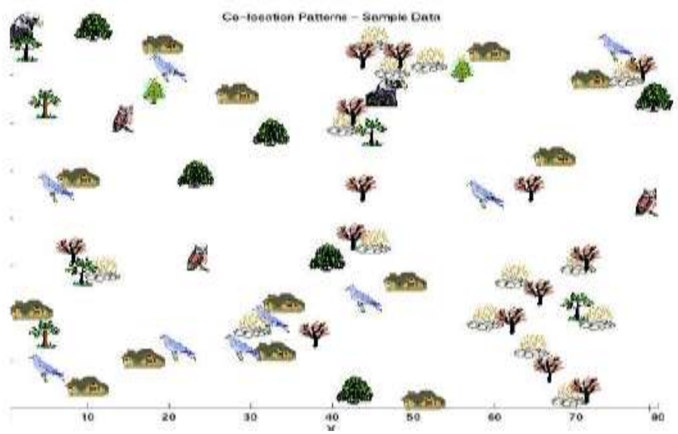
Water depth



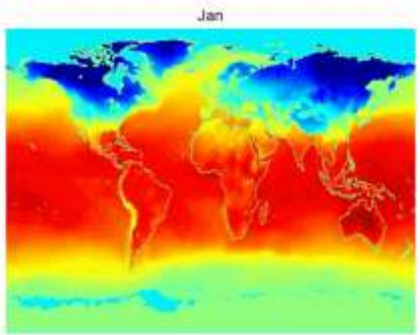
Spatial outliers: sensor (#9) on I-35



Co-location Patterns



Tele connections

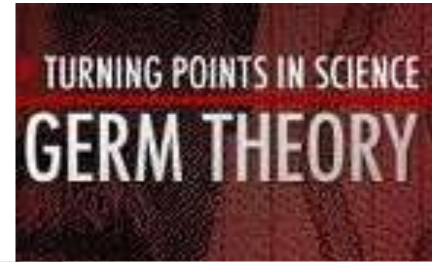


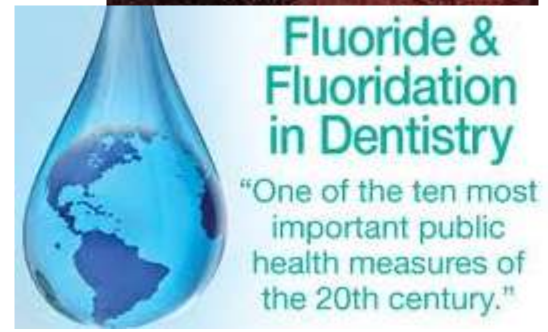
Data Mining Questions

- Public Health, Public Safety, National Security
  - What are the hotspots of an infectious disease, crime, insurgency? Why?
  - What are emerging hotspots? Which way will it spread? Where did it originate?
  - What are critical places (sources) and paths(transportation routes) ?
  - What are spatio-temporal patterns of life (for a person or a disease) ?
  - Is current spatio-temporal pattern of a disease anomalous?
  - Which spatio-temporal event-types (e.g., diseases) co-locate (or co-occur)?

- Climate, Environment, Impact on Health (e.g., Exposome)
  - How is the climate changing? How does impact Exposome? Gene-Environment interactions?
  - How does it change pathogens, pathogen carriers, disease rates and locations?
  - What are the consequences of changes in the Earth system for human health?
  - How well can we predict future changes?
  - What actions may reduce adverse impacts on human health?

Exploratory Data Analysis and Health





- Exploratory Spatial Analysis
  - Help generate hypothesis
  - Location bring in rich context to prioritize hypothesis
- Examples of Hypothesis Generation via Data Mining
  - London Cholera Map (J. Snow, 1854)
    - → Caused by water rather than bad air (miasma theory)
    - → Germ Theory
  - Colorado flourosis (1905) → water causation (1923)
    - → Bauxite? Flouride? → 1% prevent carries (1930)
      → public policy (1948) …
  - Functional Genomics is a data mining problem!
  - Exposomics
    - Exposomics is a spatial data mining problem!
    - Q: Which exposure strengthens Immune system ?
- Notes:
  - " … whereas structural genomics has been characterized by data management, functional genomics will be characterized by mining the data sets for particularly valuable information.", Functional Genomics: It's All How You Read It, Philip Hieter and Mark Boguski, AAAS Science, 278, 14th October 1997.
  - More on Exposomics at www.cdc.gov/niosh/topics/exposome/

NIDCR National Institute of Dental and Craniofacial Research
National Institutes of Health
mproving the Nation's Oral Health

SPATIAL DATABASE RESEARCH GROUP

Why Data Mining?

- Holy Grail - Informed Decision Making

- Lots of Data are Being Collected
  - Business - Transactions, Web logs, GPS-track, …
  - Science - Remote sensing, Micro-array gene expression data, …

- Challenges:
  - Volume (data) >> number of human analysts
  - Some automation needed

- Data Mining may help!
  - Provide better and customized insights for business
  - Help scientists for hypothesis generation

SPATIAL DATABASE RESEARCH GROUP

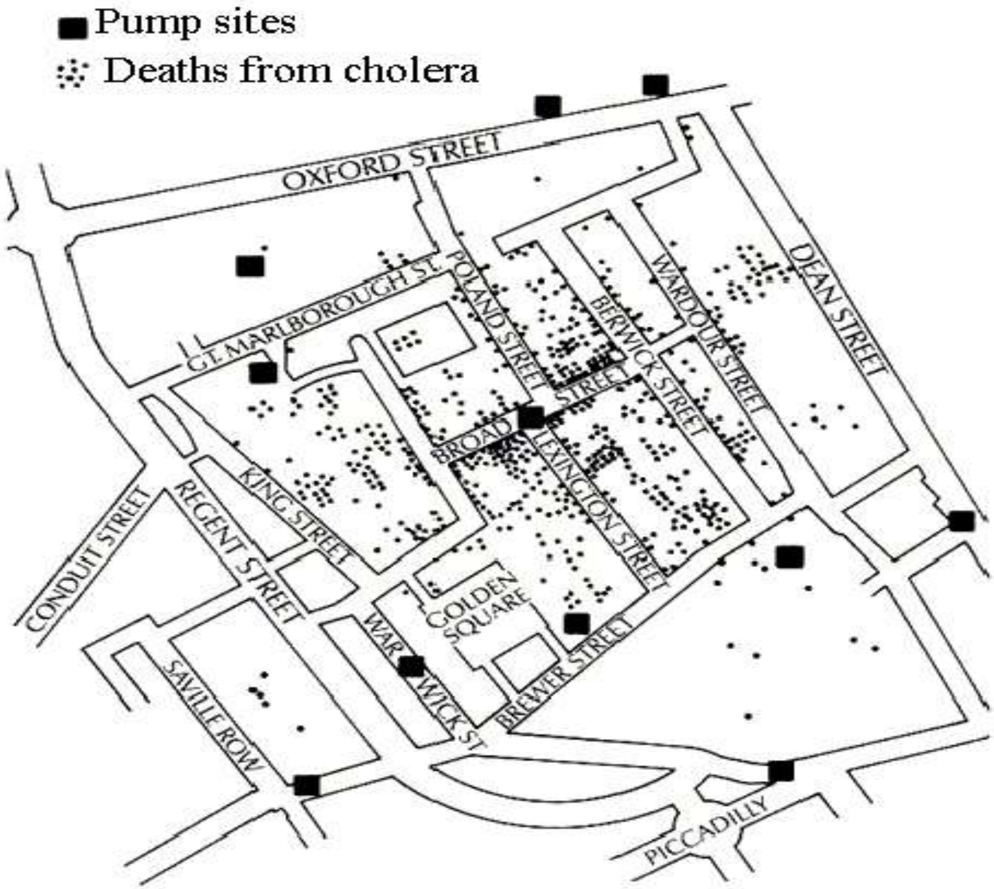
Spatial Data Mining (SDM)

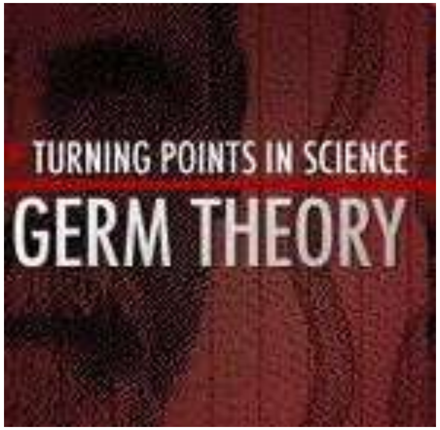
- The process of discovering
  - interesting, useful, non-trivial patterns
    - patterns: non-specialist
    - exception to patterns: specialist
  - from large spatial datasets

- Spatial pattern families
  1. Hotspots, Spatial clusters
  2. Spatial outlier, discontinuities
  3. Co-locations, co-occurrences
  4. Location prediction models
  5. …

Pattern Family 1: Hotspots, Spatial Cluster

- The 1854 Asiatic Cholera in London
  - Near Broad St. water pump except a brewery

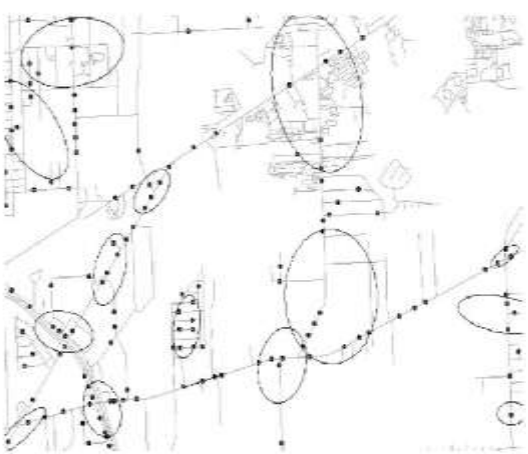


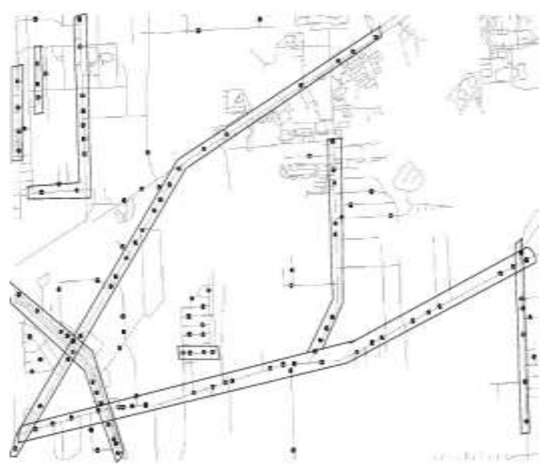


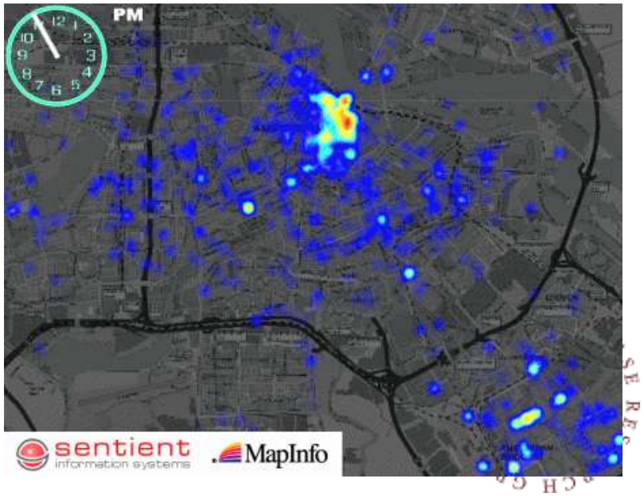
From Hotspots to Hot-routes & Mean Streets

- Challenges: Spatial Networks, Time
- Examples:
  - India Accelerating | An Epidemic Spreads, “On India's Roads, Cargo and a Deadly Passenger”, NewYork Times, A. Waldman, December 6, 2005.
    Its national highways are a conduit for the virus, passed by prostitutes and the truckers, migrants and locals who pay them …
  - Global transport networks and infectious disease spread, Adv Parasitol. 2006;62:293-343. (http://www.ncbi.nlm.nih.gov/pubmed/16647974)

- Q? How may one detect routes of disease spread?
  - Spatial Statistical methods identify ellipsiodal hotspots
  - Spatial data mining methods, e.g. K-Main Route, for hot-routes, mean streets

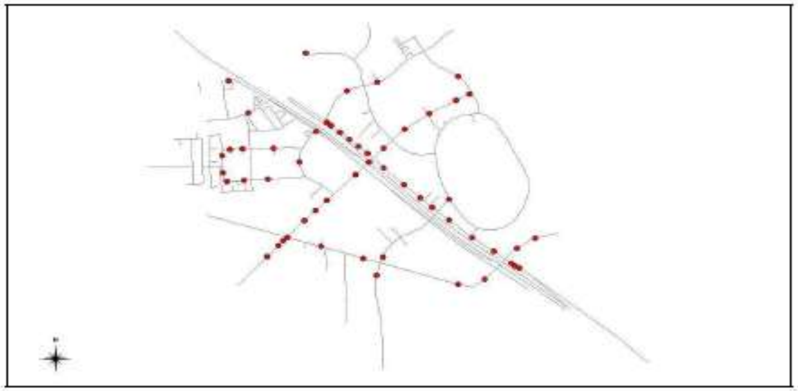




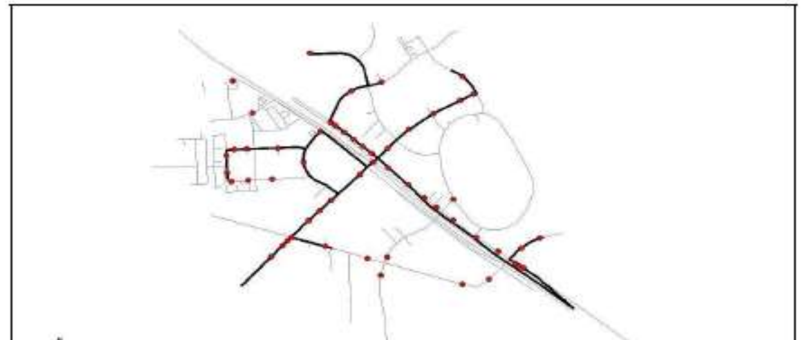


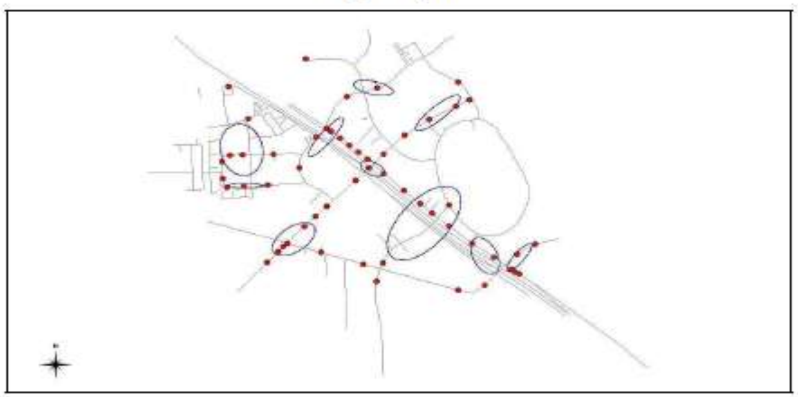
Innovative Technique: K Main Routes (KMR)

Summarizes Urban Activities

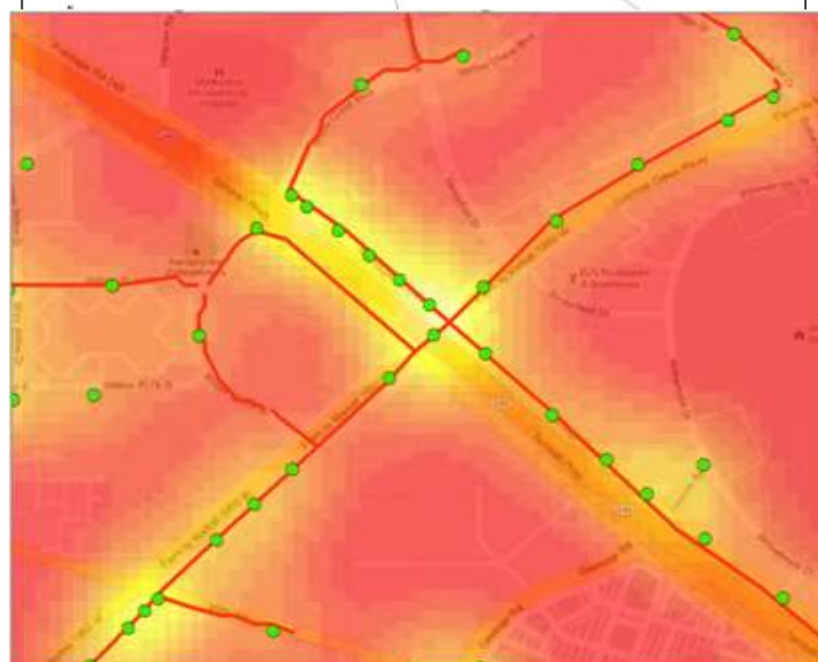


(a) Input





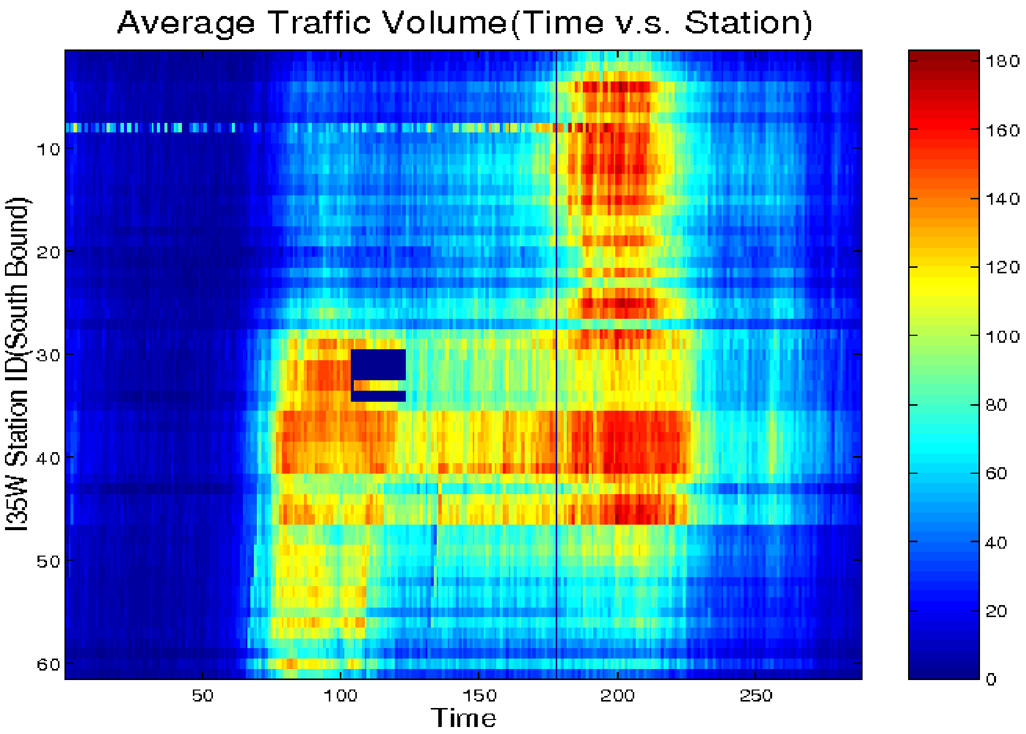
(c) Crimestat K-means Output



KMR Routes (10) – thick lines, Crimestat K-Means (10) – ellipses,
Roads – gray lines, Burglaries - points

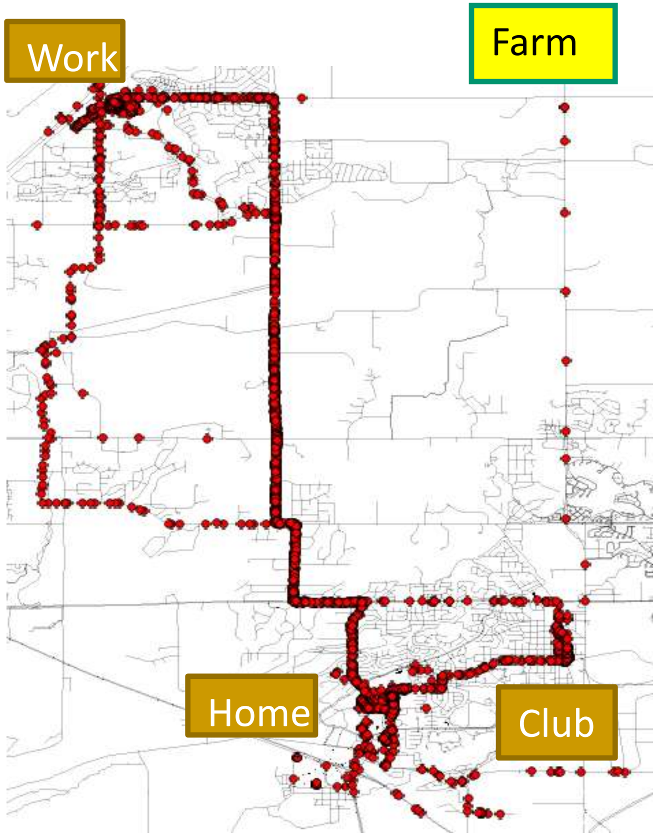
Pattern Family 2: Spatial Outliers

- Spatial Outliers
  - Traffic Data in Twin Cities
  - Abnormal Sensor Detections
  - Spatial and Temporal Outliers



Patterns of Life

- Weekday GPS track of over 3 months
  - Patterns of life
  - Usual places and visits
    - Small return period
  - Rare places, Rare visits
    - Large Return period, e.g., once a month, once a quarter, once a year, …

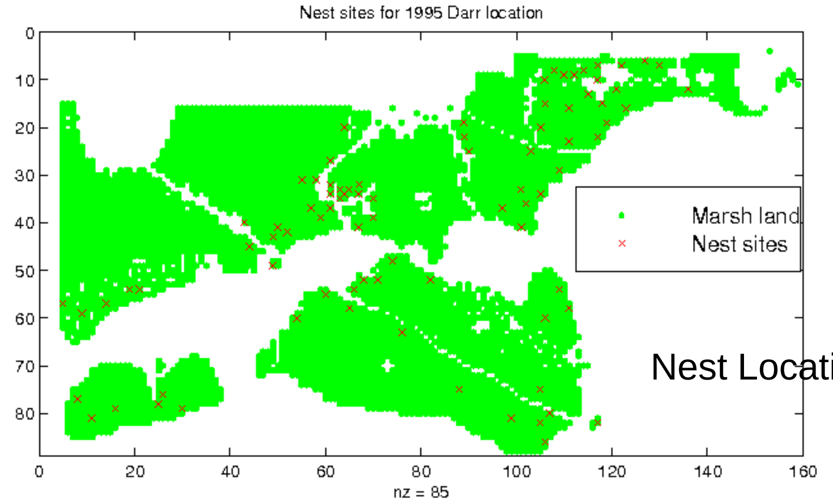


	Morning 7am – Noon	Afternoon Noon – 5pm	Evening 5pm – Midnight	Night Midnight – 7am	Total
Home	10	2	15	29	54
Work	19	20	10	1	50
Club	4	5	4		15
Farm			1		1
Total	30	30	30	30	120

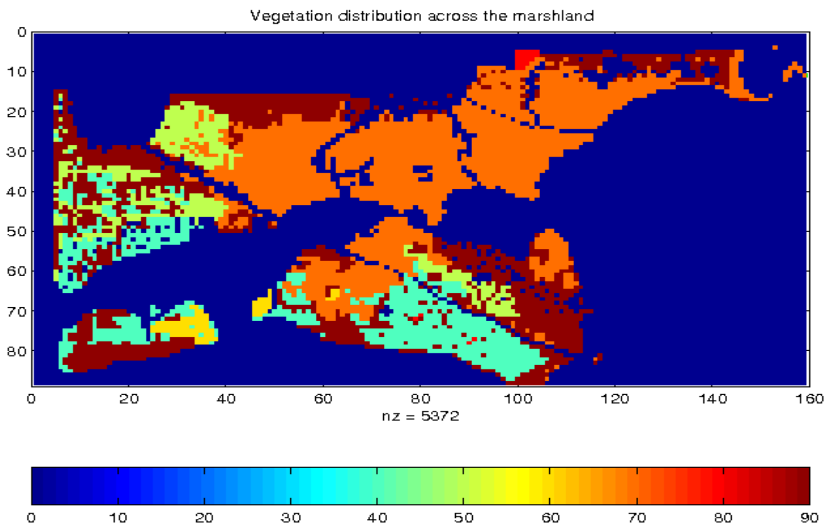
SPATIAL DATABASE RESEARCH GROUP

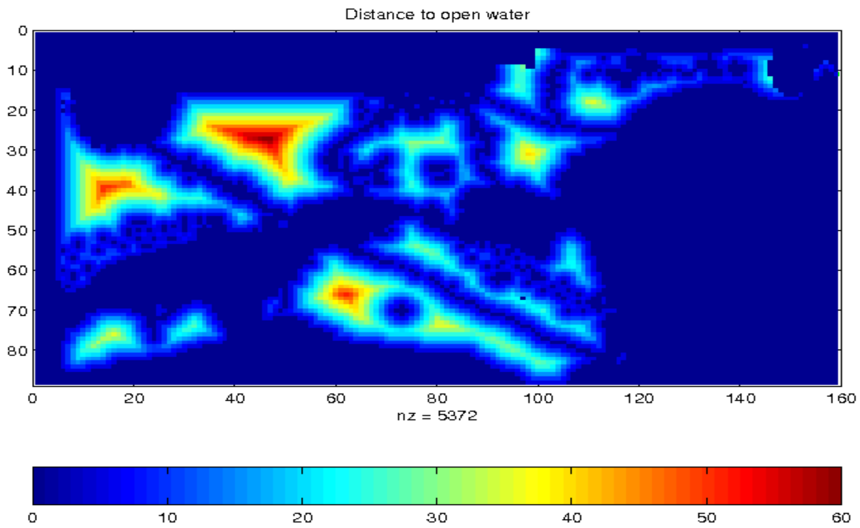
Pattern Family 3: Predictive Models

- Location Prediction:
  - Predict Bird Habitat Prediction
  - Using environmental variables



Nest Locations





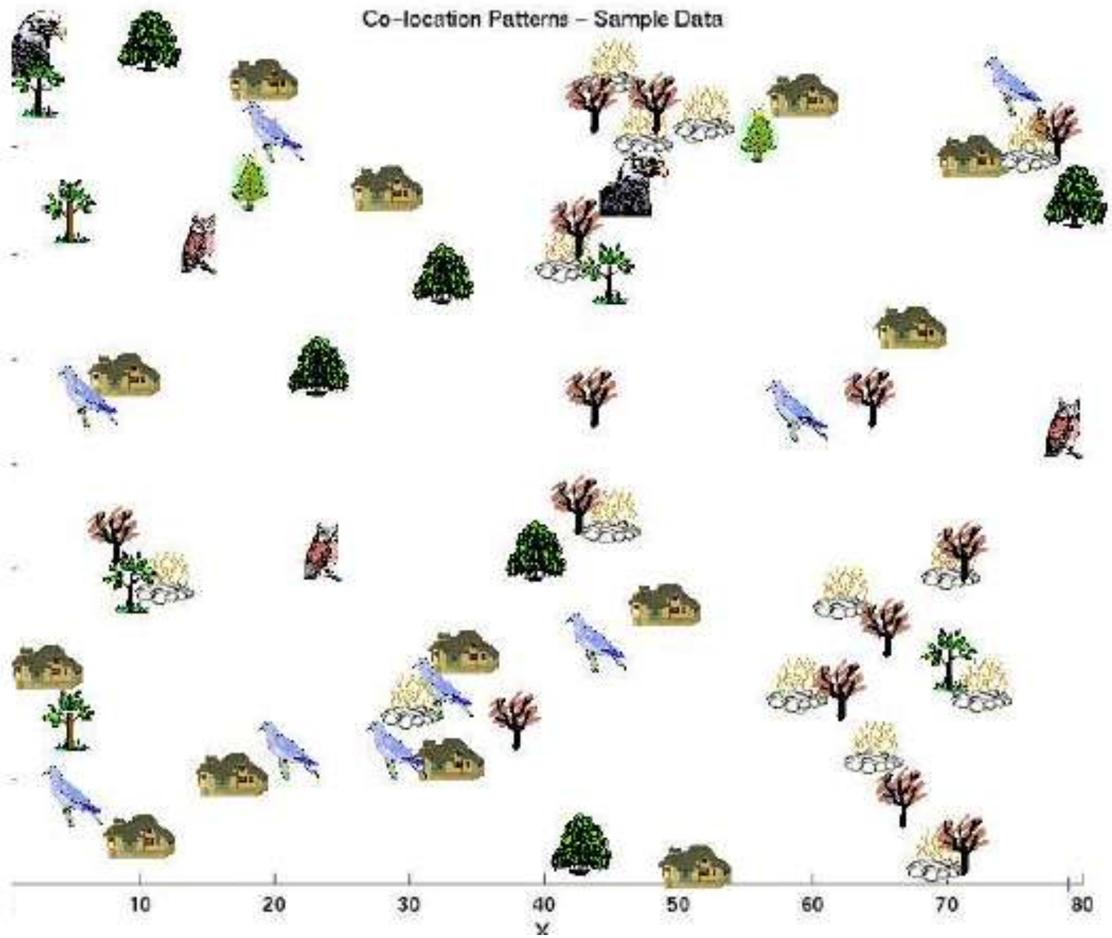
Prediction and Trend

- Prediction
  - Continuous: trend, e.g., regression
    - Location aware: spatial autoregressive model (SAR)
  - Discrete: classification, e.g., Bayesian classifier
    - Location aware: Markov random fields (MRF)

Classical	Spatial
$y = X\beta + \varepsilon$	$y = \rho W y + X\beta + \varepsilon$
$\Pr(C_i \mid X) = \dfrac{\Pr(X \mid C_i)\Pr(C_i)}{\Pr(X)}$	$\Pr(c_i \mid X, C_N) = \dfrac{\Pr(C_i)\Pr(X, C_N \mid c_i)}{\Pr(X, C_N)}$

Pattern Family 4: Co-locations/Co-occurrence

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



Answers:  and 

Spatial Colocation

Participation ratio pr(f_i, c) of feature f_i in colocation c = {f_1, f_2, …, f_k}:

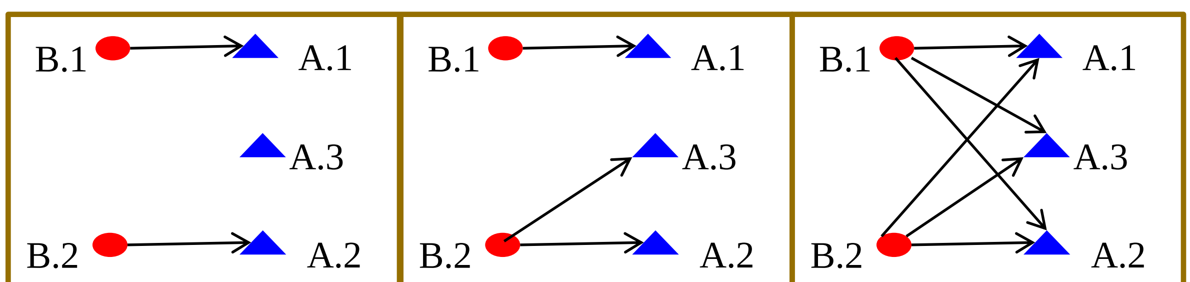
fraction of instances of f_i with feature {f_1, …, f_{i-1}, f_{i+1}, …, f_k} nearby (i.e. within a given distance)

Participation index PI(c) = min{ pr(f_i, c) }

Properties: (1) Computational: Non-monotonically decreasing like support measure

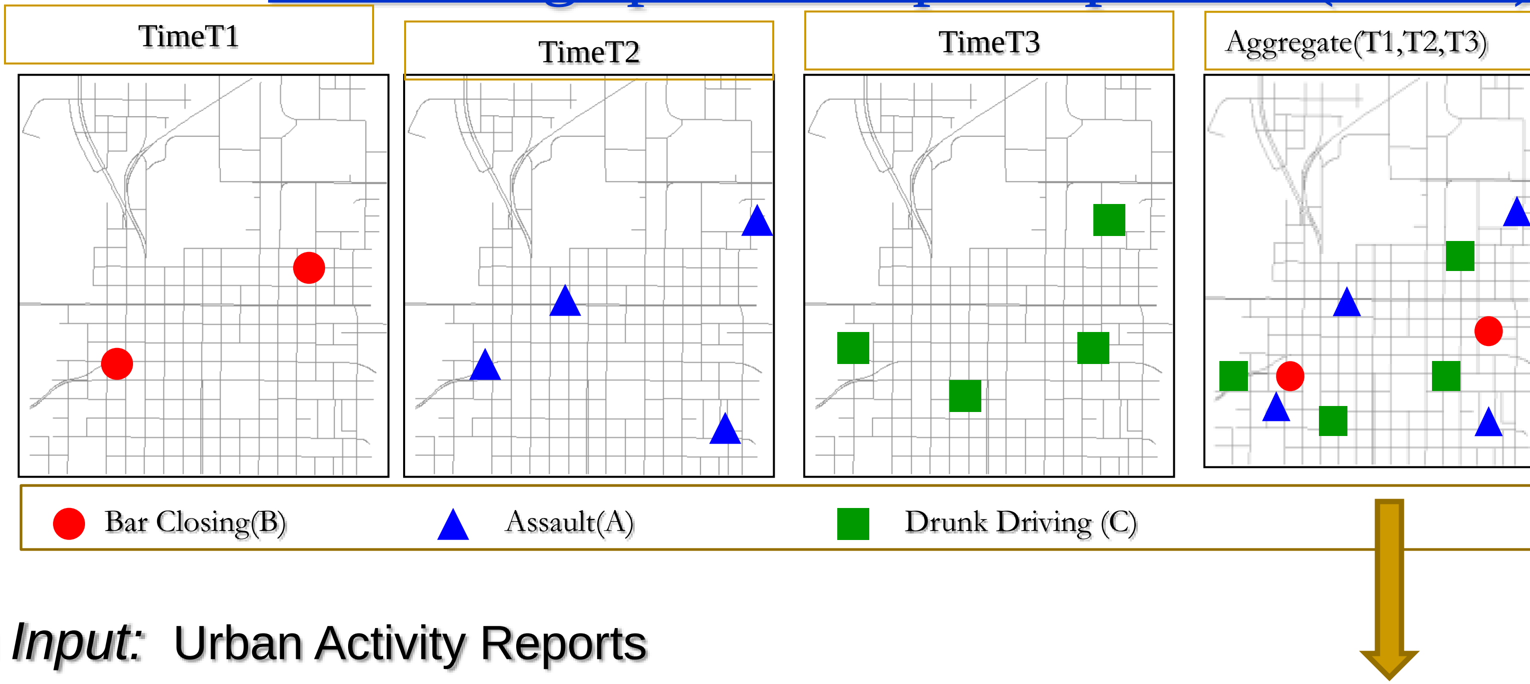
(2) Statistical: Lower bound on Cross-K function

- Comparison with K-function

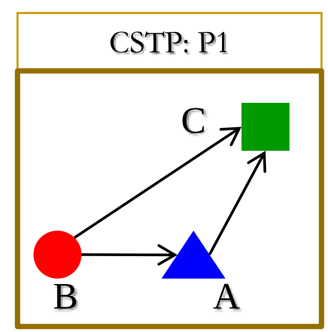


ST -K (B → A)	2/6 = 0.33	3/6 = 0.5	6/6 = 1
PI (B → A)	2/3 = 0.66	1	1

Cascading spatio-temporal pattern (CSTP)

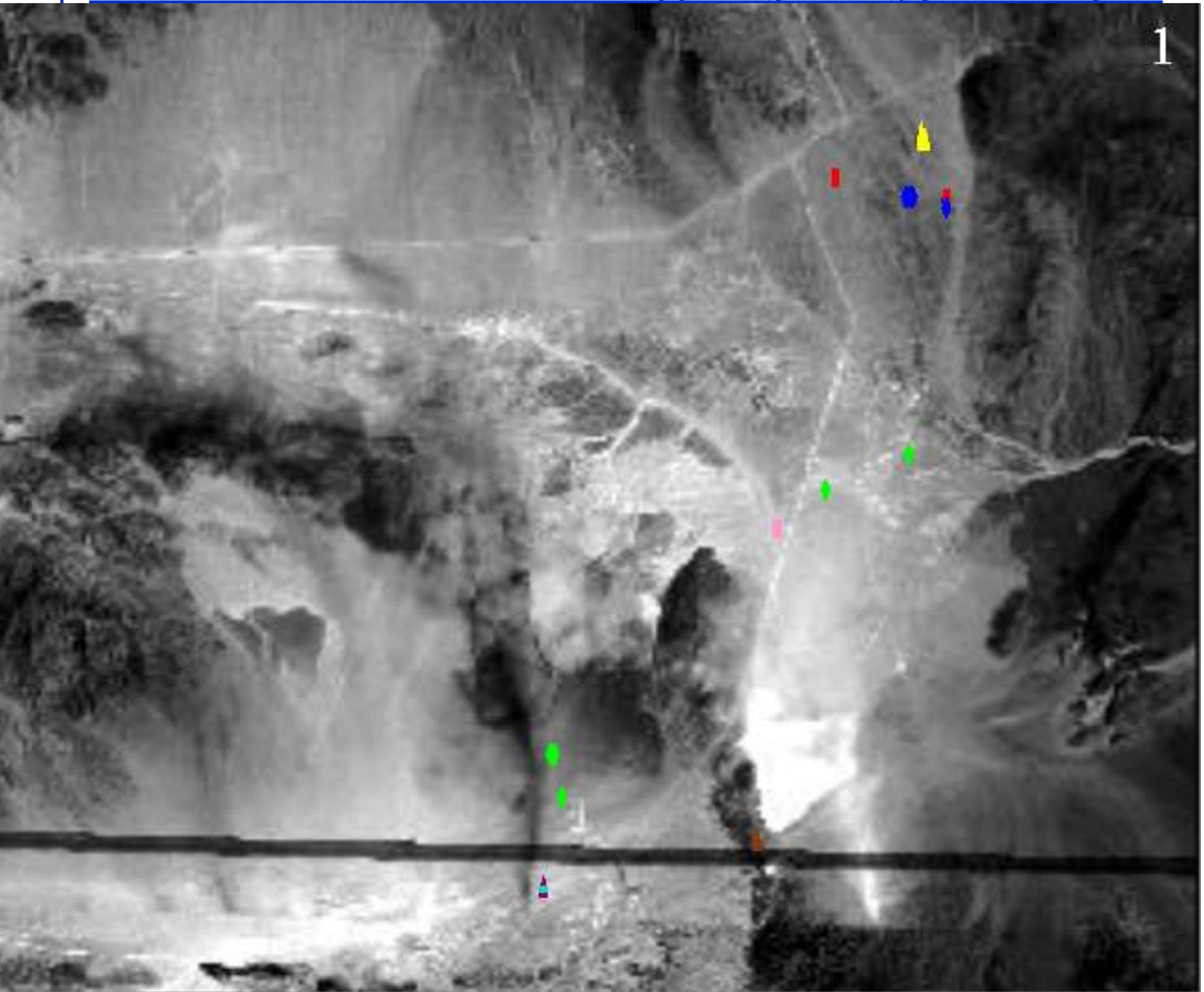


- *Input:* Urban Activity Reports
- *Output: CSTP*
  - *Partially ordered* subsets of ST event types.
  - Located together in space.
  - Occur in *stages* over time.



- Applications:
  - Epidemiology, Disaster Response, …

Co-occurrence & Moving Object-types: Input



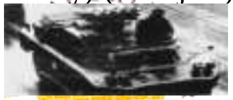
- Manpack stinger (2 Objects)

- M1A1_tank (3 Objects)

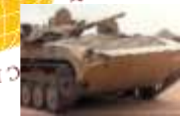
- M2_IFV (3 Objects)

- Field_Marker (6 Objects)
- T80_tank (2 Objects)

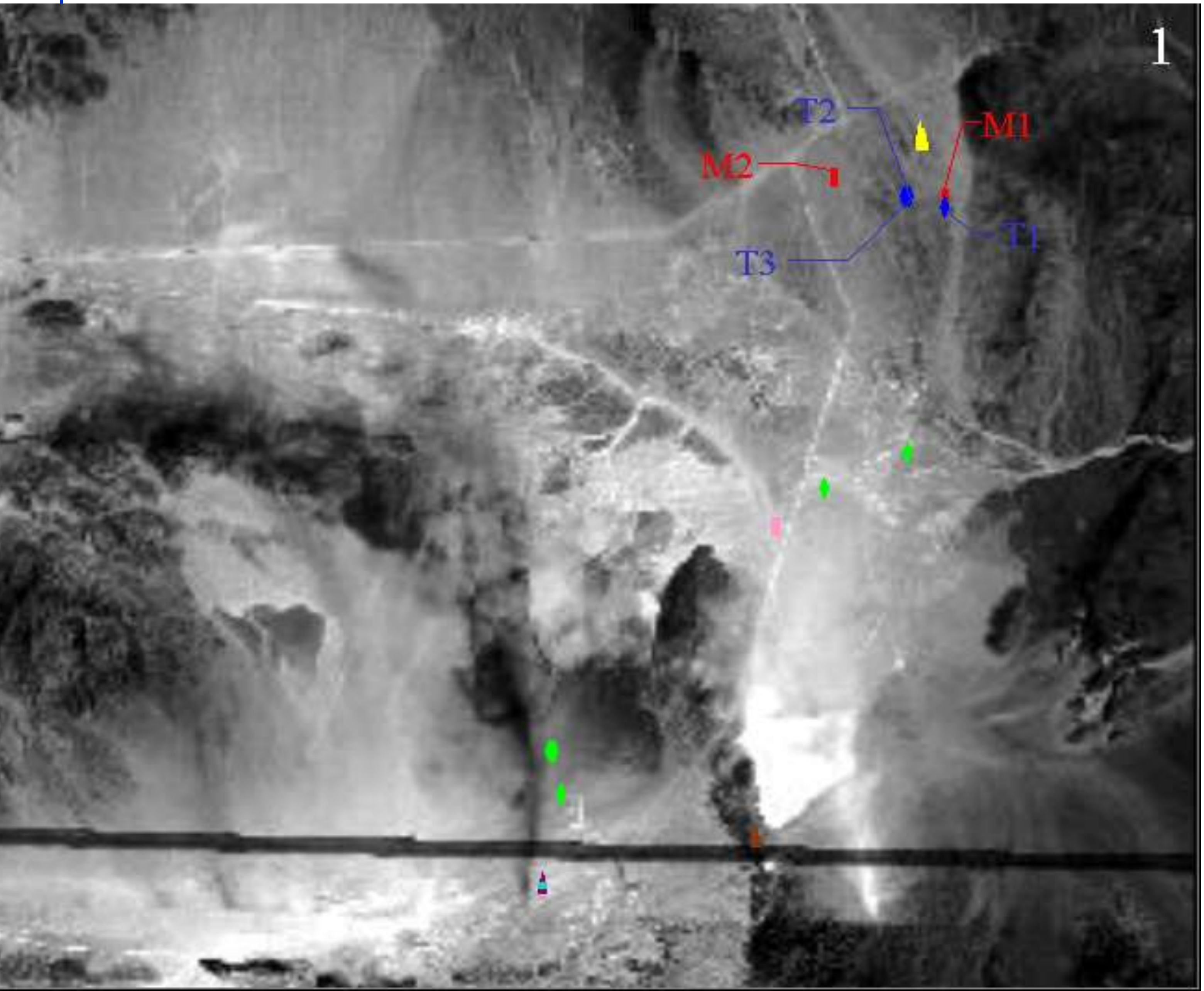
- BRDM_AT5 (enemy) (1 Object)



- BMP1 (1 Object)



Co-occurrence & Moving Object-Types: Output



- Manpack stinger (2 Objects)

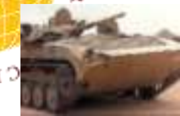
- M1A1_tank (3 Objects)

- M2_IFV (3 Objects)

- Field_Marker (6 Objects)
- T80_tank (2 Objects)

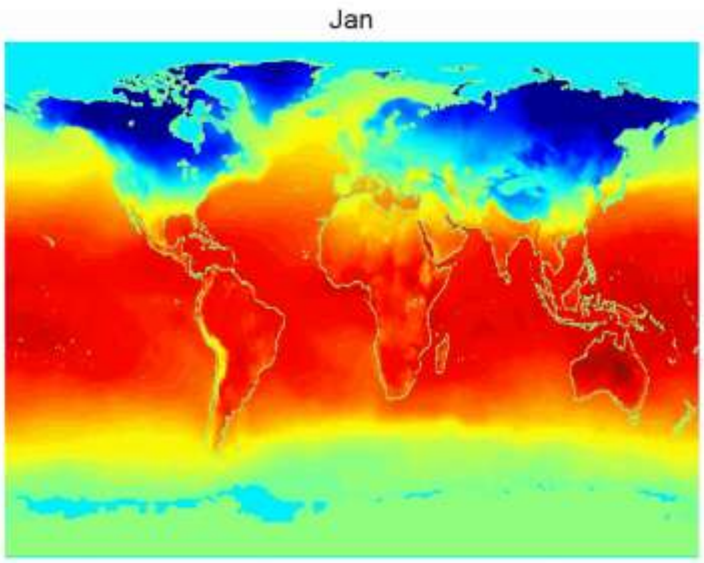
- BRDM_AT5 (enemy) (1 Object)

- BMP1 (1 Object)



Teleconnection

- Global Climate Change
  - Find (land location, ocean location) pairs with correlated climate changes
    - Ex. El Nino affects climate at many land locations



Average Monthly Temperature

(Courtsey: NASA, Prof. V. Kumar)

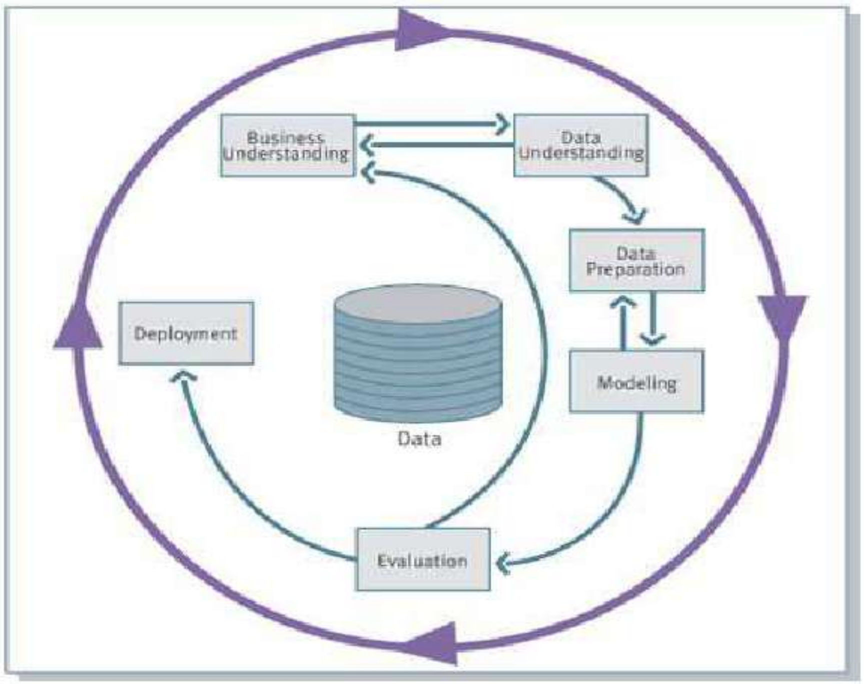
Global Influence of El Nino during the Northern Hemisphere Winter (D: Dry, W: Warm, R: Rainfall)

Teleconnection

- Challenge
  - high dimensional (e.g., 600) feature space
  - 67k land locations and 100k ocean locations (degree by degree grid)
  - 50-year monthly data
- Computational Efficiency
  - Spatial autocorrelation
    - Reduce Computational Complexity
  - Spatial indexing to organize locations
    - Top-down tree traversal is a strong filter
    - Spatial join query: filter-and-refine
      - save 40% to 98% computational cost at θ = 0.3 to 0.9

Life Cycle of Data Mining

- CRISP-DM (CRoss-Industry Standard Process for DM)
  - Application/Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment



Phases of CRISP-DM

Is CRISP-DM adequate for Spatial Data Mining?

[1] CRISP-DM URL: http://www.crisp-dm.org