Spatial Data Mining

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Learning Objectives

- After this segment, students will be able to
  - Describe the motivation for spatial data mining
  - List common pattern families

Why Data Mining?

- Holy Grail - Informed Decision Making
- Sensors & Databases increased rate of Data Collection
  - Transactions, Web logs, GPS-track, Remote sensing, …
- Challenges:
  - Volume (data) >> number of human analysts
  - Some automation needed
- Approaches
  - Database Querying, e.g., SQL3/OGIS
  - Data Mining for Patterns
  - …

Data Mining vs. Database Querying

- Recall Database Querying (e.g., SQL3/OGIS)
  - Can not answer questions about items not in the database!
    - Ex. Predict tomorrow's weather or credit-worthiness of a new customer
  - Can not efficiently answer complex questions beyond joins
    - Ex. What are natural groups of customers?
    - Ex. Which subsets of items are bought together?
- Data Mining may help with above questions!
  - Prediction Models
  - Clustering, Associations, …

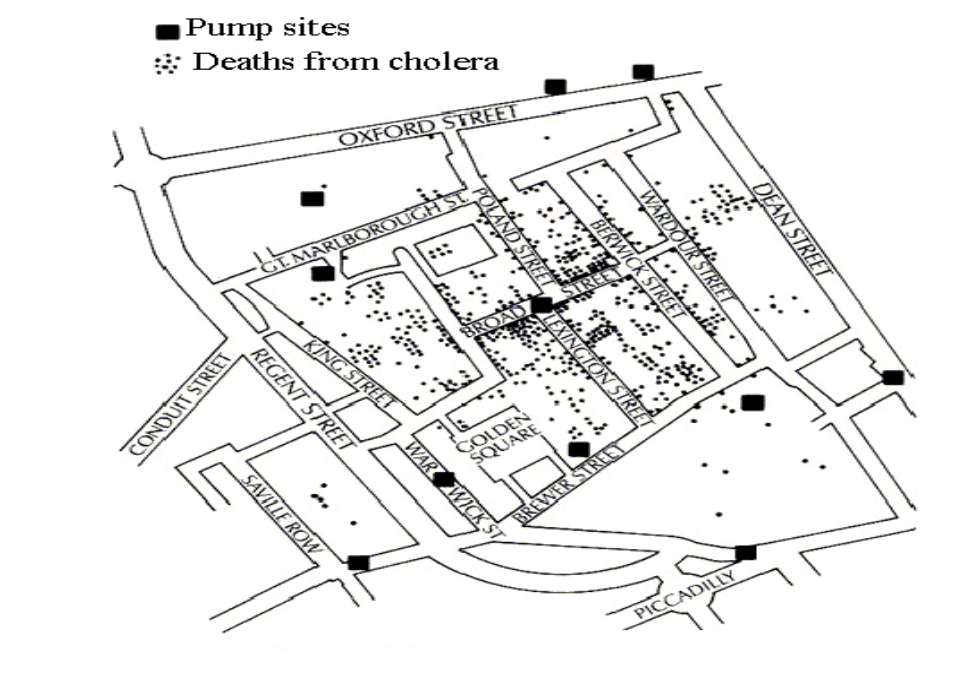
Spatial Data Mining (SDM)

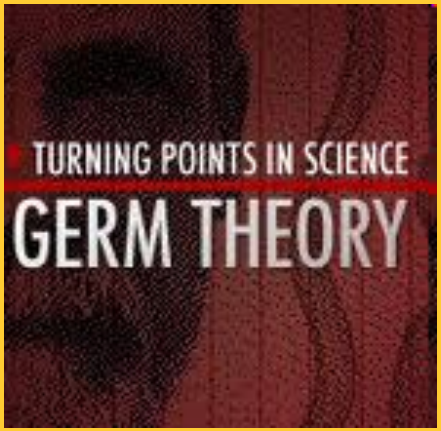
- The process of discovering
  - interesting, useful, non-trivial patterns
    - patterns: non-specialist
    - exception to patterns: specialist
  - from large spatial datasets

- Spatial pattern families
  - Hotspots, Spatial clusters
  - Spatial outlier, discontinuities
  - Co-locations, co-occurrences
  - Location prediction models
  - …

Pattern Family 1: Hotspots, Spatial Cluster

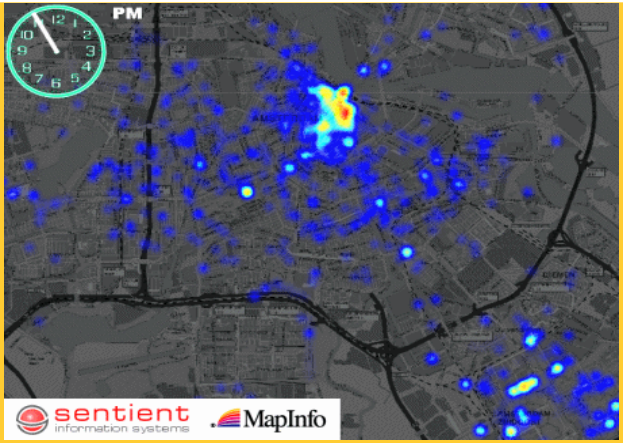
- The 1854 Asiatic Cholera in London
  - Near Broad St. water pump except a brewery

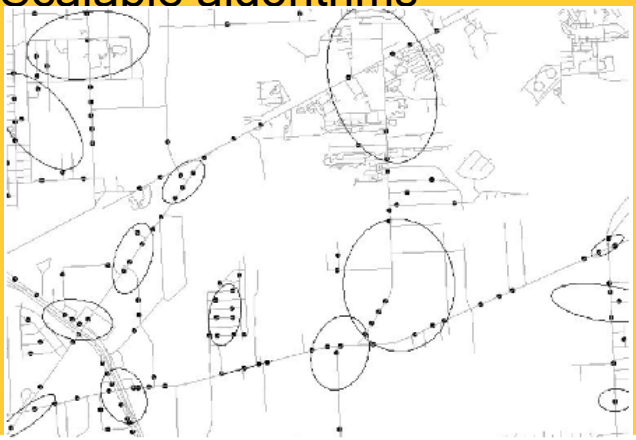


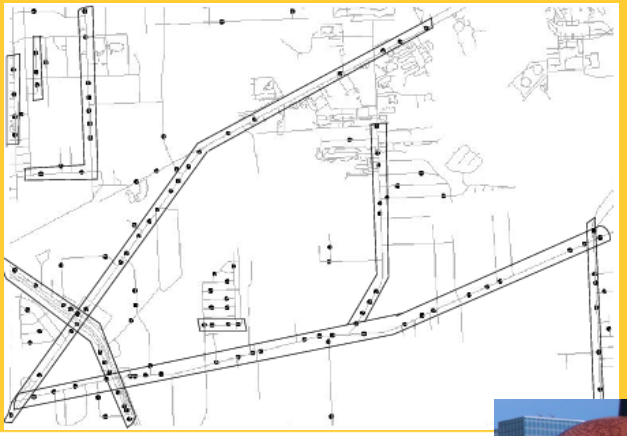


Complicated Hotspots

- Complication Dimensions
  - Time
  - Spatial Networks
- Challenges: Trade-off b/w
  - Semantic richness and
  - Scalable algorithms

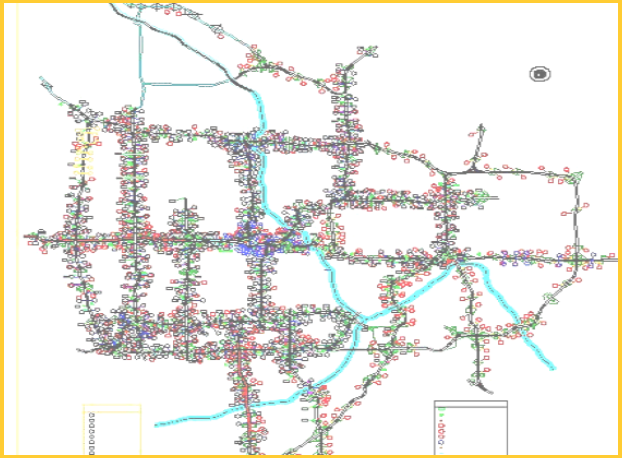


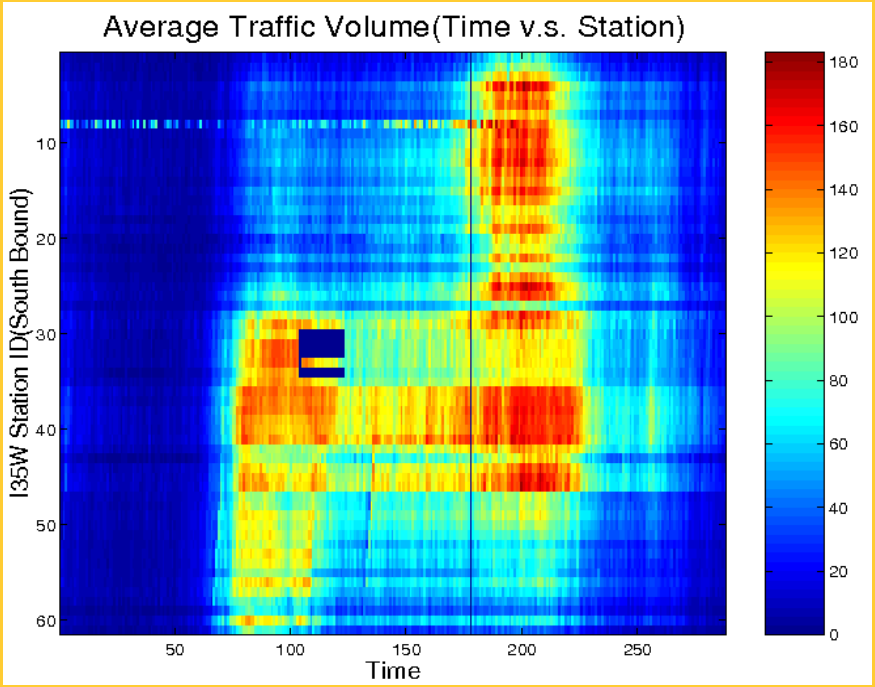




Pattern Family 2: Spatial Outliers

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

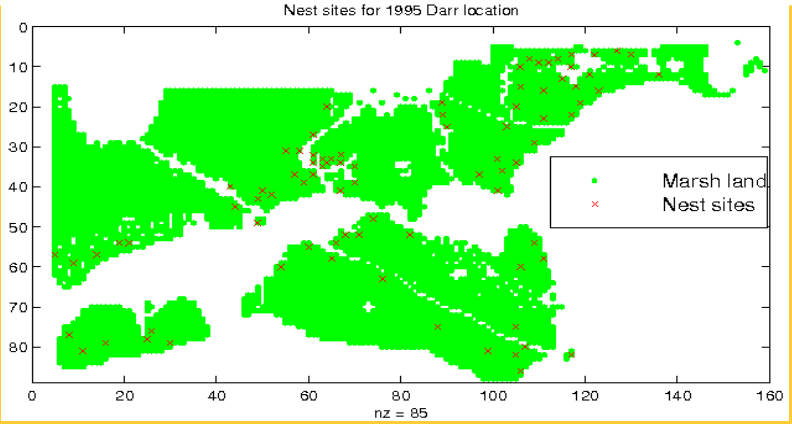


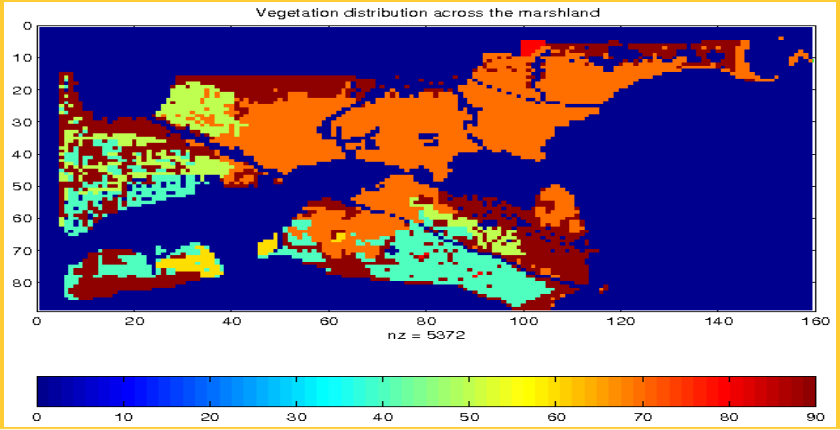


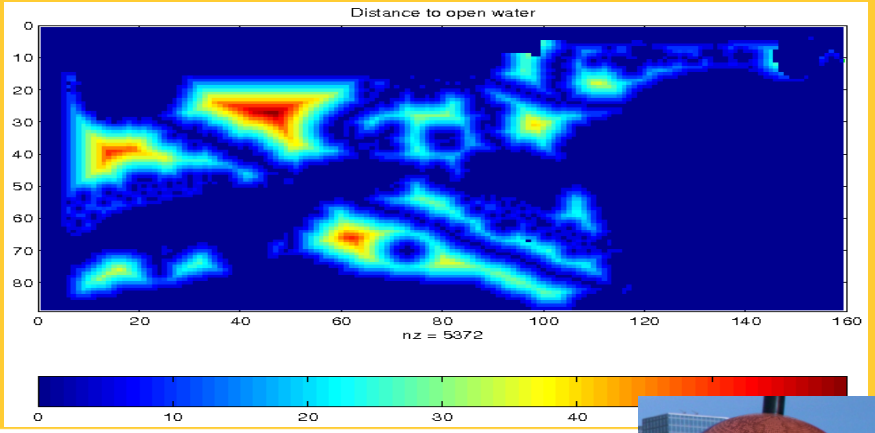
Source: A Unified Approach to Detecting Spatial Outliers, GeoInformatica, 7(2), Springer, June 2003. (A Summary in Proc. ACM SIGKDD 2001) with C.-T. Lu, P. Zhang.

Pattern Family 3: Predictive Models

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



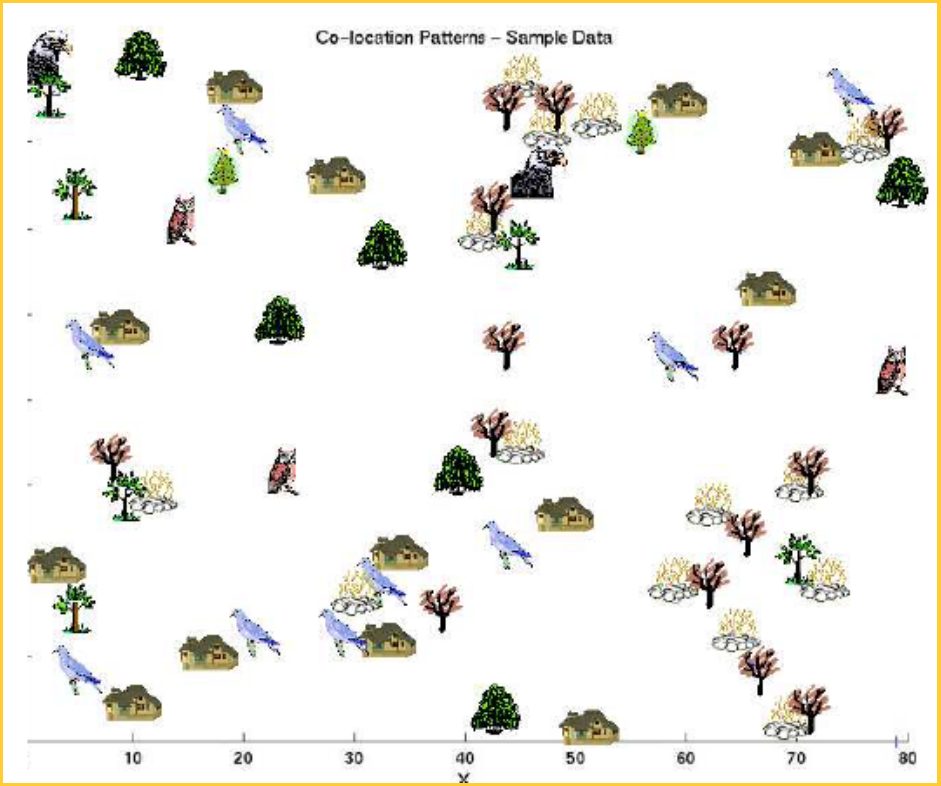




Family 4: Co-locations/Co-occurrence

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

Answers:   and  



Source: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

What's NOT Spatial Data Mining (SDM)

- Simple Querying of Spatial Data
  - Find neighbors of Canada, or shortest path from Boston to Houston
- Testing a hypothesis via a primary data analysis
  - Ex. Is cancer rate inside Hinkley, CA higher than outside ?
  - SDM: Which places have significantly higher cancer rates?
- Uninteresting, obvious or well-known patterns
  - Ex. (Warmer winter in St. Paul, MN) => (warmer winter in Minneapolis, MN)
  - SDM: (Pacific warming, e.g. El Nino) => (warmer winter in Minneapolis, MN)
- Non-spatial data or pattern
  - Ex. Diaper and beer sales are correlated
  - SDM: Diaper and beer sales are correlated in blue-collar areas (weekday evening)

Review Quiz: Spatial Data Mining

- Categorize following into queries, hotspots, spatial outlier, colocation, location prediction:
  - (a) Which countries are very different from their neighbors?
  - (b) Which highway-stretches have abnormally high accident rates ?
  - (c) Forecast landfall location for a Hurricane brewing over an ocean?
  - (d) Which retail-store-types often co-locate in shopping malls?
  - (e) What is the distance between Beijing and Chicago?