HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

582364 Data mining, 4 cu
Lecture 9:
Spatial data mining

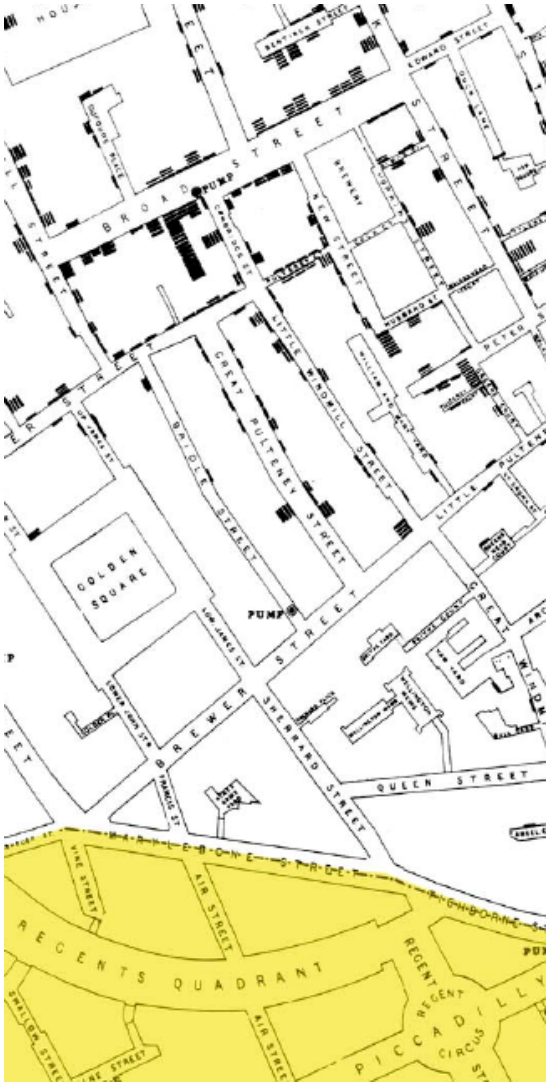
Spring 2010

Lecturer: Juho Rousu

Teaching assistant: Taru Itäpelto

Spatial Data Mining: Motivation

- In many applications, the interesting patterns are location dependent
- Classical example is the 1854 cholera outbreak in London
- Deaths were spatially clustered in a polluted water pump in Soho, as shown by the map by John Snow
- The connection of cholera and polluted water was not known at the time
- Nowadays, data mining is used to monitor disease outbreaks
  - Local expert in Kumpula: Dr. Roman Yangarber



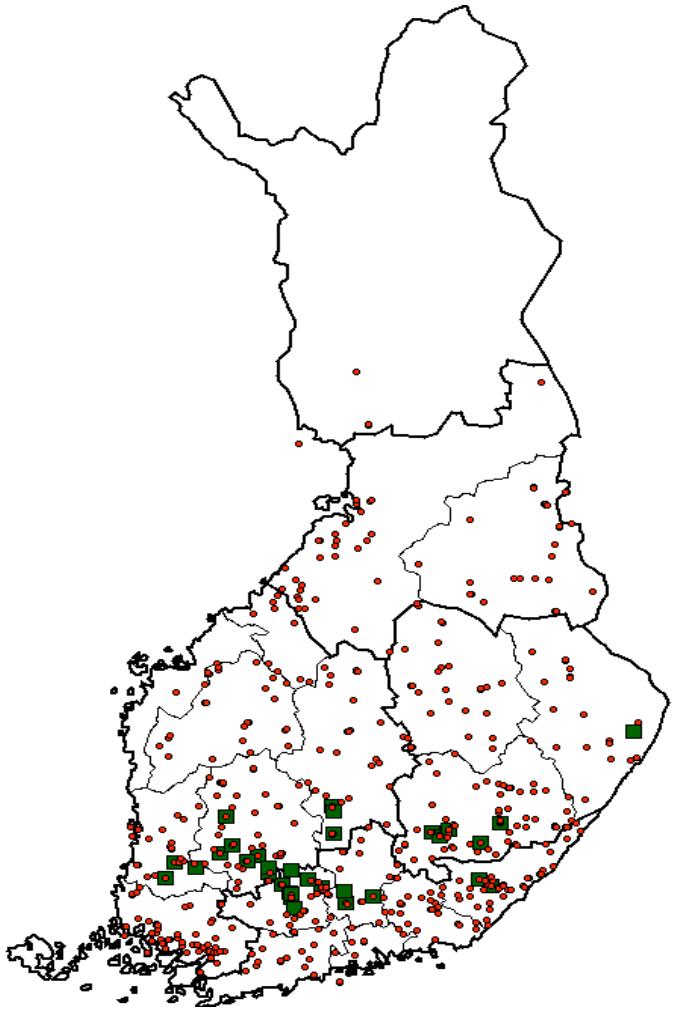
Spatial Autocorrelation

- First law of geography:

  "*Everything is related to everything, but nearby things are more related than distant things*."
- Spatial data tends to be highly auto-correlated:
  - The value of property in one location predicts the property in a nearby location
- This property underlies most spatial data analysis tasks

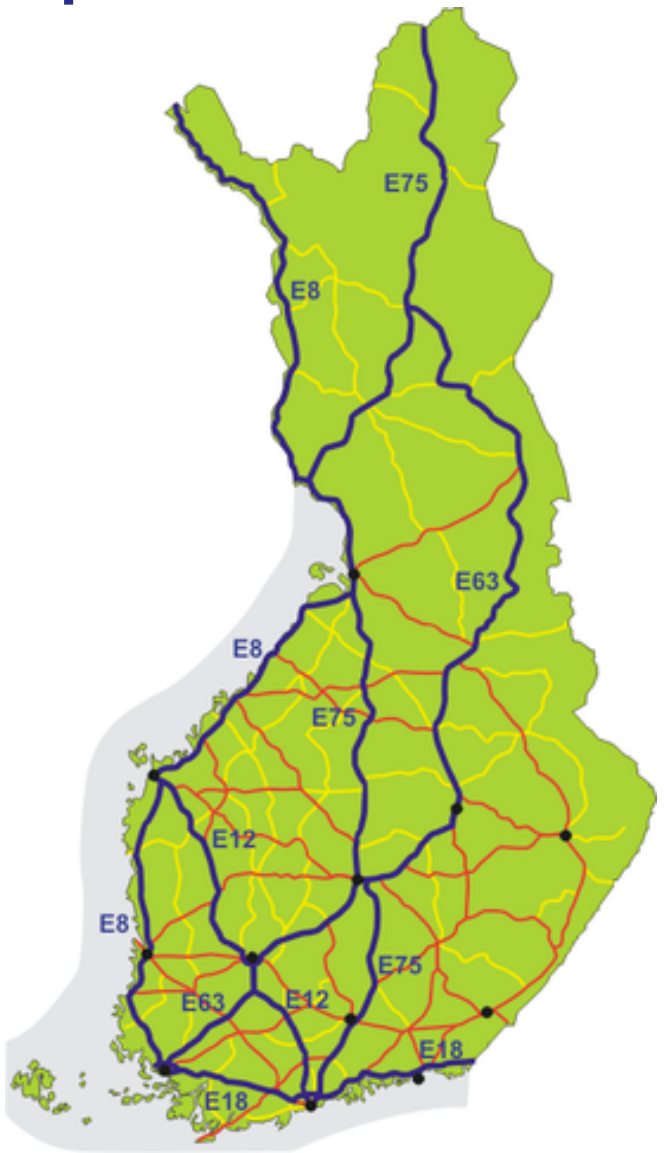
Kinds of Spatial data: Point patterns

- Data is represented by single points with a location
- Shape is not relevant
- Each phenomenon represented by a separate point pattern
- Example: Viking-age forts
  - Red dots: place names starting with 'Linna' (engl. `castle')
  - Green squares: Viking-age hill forts



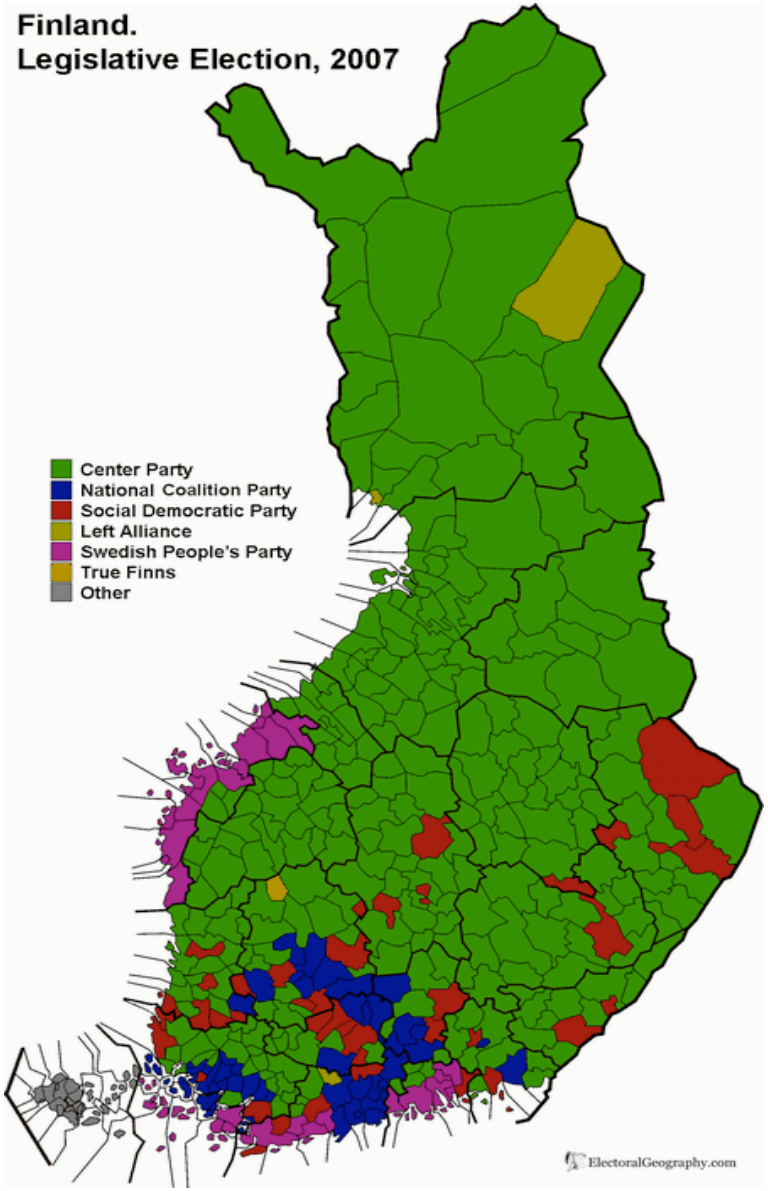
Kinds of Spatial data: graph data

- Graph embedded into the plane, consisting of
  - a set of point objects, forming its nodes,
  - and a set of line objects describing the geometry of the edges,
    - e.g., highways. rivers, power supply lines.



Kinds of Spatial data: spatially continuous data

- Data is represented by regions
- Describes a spatially continuous phenomenon
- However, not possible to measure across the space
  - Measurements at distinct points
  - Aggregate over the region
- Example: election data
  - Measurements are individual votes
  - Seats in the parliament decided based on a party's relative share of votes in an election region



Spatial relations

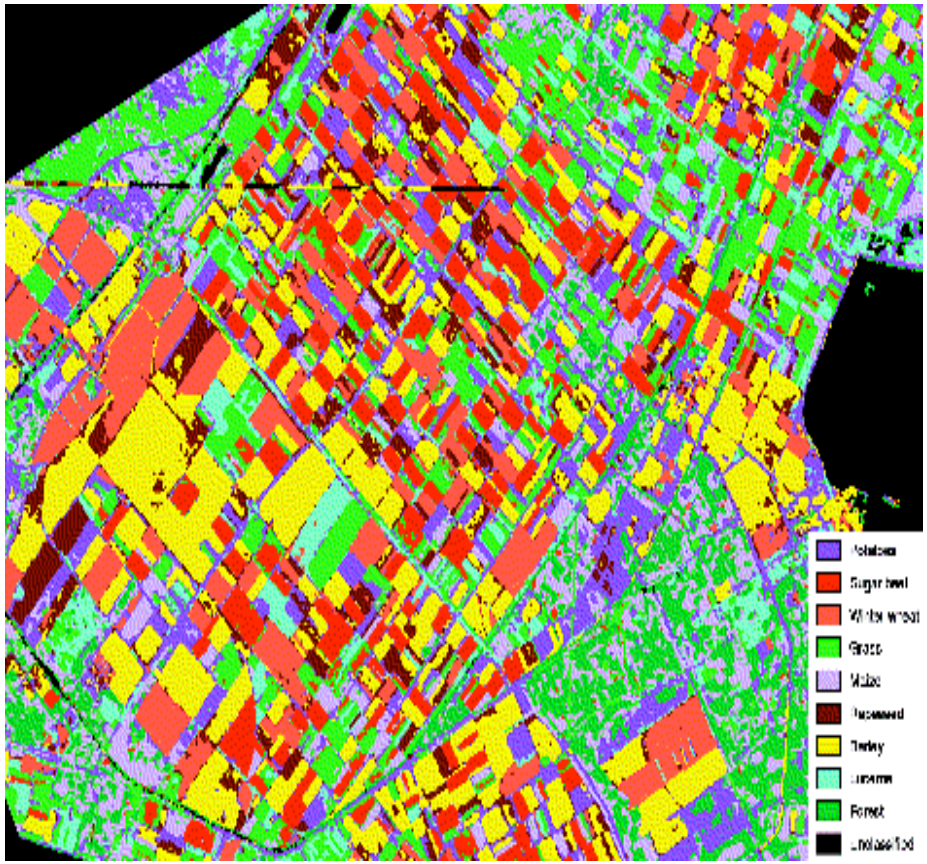
- Spatial objects have a rich structure of relations that can potentially used for data mining
- Distance relations: Distances between objects
  - ‘Restaurant’ near ‘market square’
  - ‘Bus stop’ within 500m from Exactum
- Direction relations: Ordering of spatial objects in space
  - ‘Bus stop’ *to the east* from Exactum
- Topological relations: Characterise the type of intersection between spatial features
  - *Intersection* of ‘Mäkelänkatu’ and ‘Koskelantie’
  - Buildings *within* Kumpula Campus

Concept hierachies in spatial data

- Non-Spatial Attributes
  - e.g. “Exactum” generalizes to “Office Building” generalizes to “Building”
- Spatial-to-Nonspatial Attributes
  - e.g. +60° 12' 13.19” N, +24° 57' 44.32” generalizes to “Kumpula Campus, Helsinki” (as a concept)
- Spatial-to-Spatial Attributes
  - e.g. Kumpula (as a spatial region) generalizes to Helsinki (as a spatial region)

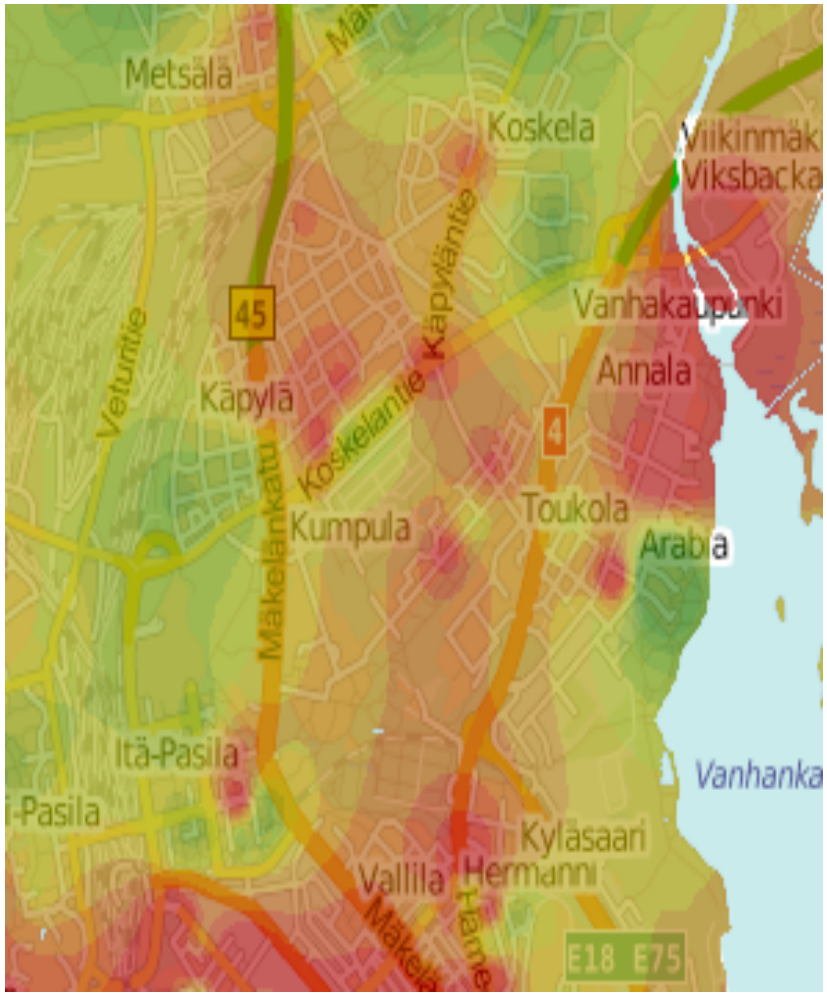
Spatial Data mining tasks: Classification

- Aim: classify locations on map into predefined classes
- Example:
  - monitoring type of crops grown in the fields
  - Data from infrared/normal images taken by a satellite
  - used (at least) in Europe and USA to detect subsidy fraud
- Closely related task: image segmentation



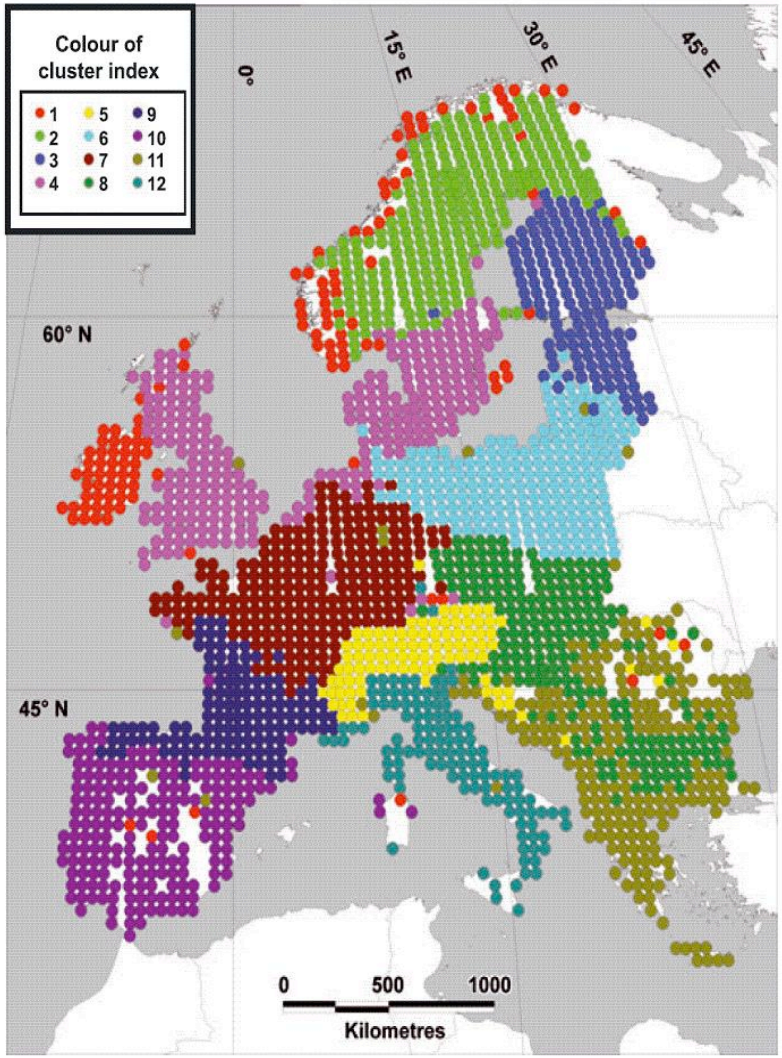
Spatial Data mining tasks: Regression

- Predict the value of a continuous variable
- Example:
  - predicting house prices
  - based on location, proximity to busy roads, prices in neighborhood



Spatial Data mining tasks: Clustering

- Clustering can reveal new groupings in spatially organized data
- In the picture, clustering of the European map is based on occurrence of species
- Each cell is a 50x50 km area where the occurrence of 124 species has been recorded
- Clusters = similar occurrence profiles
- Many methods available: k-means clustering, hierarchical clustering, …



Spatial Association Analysis

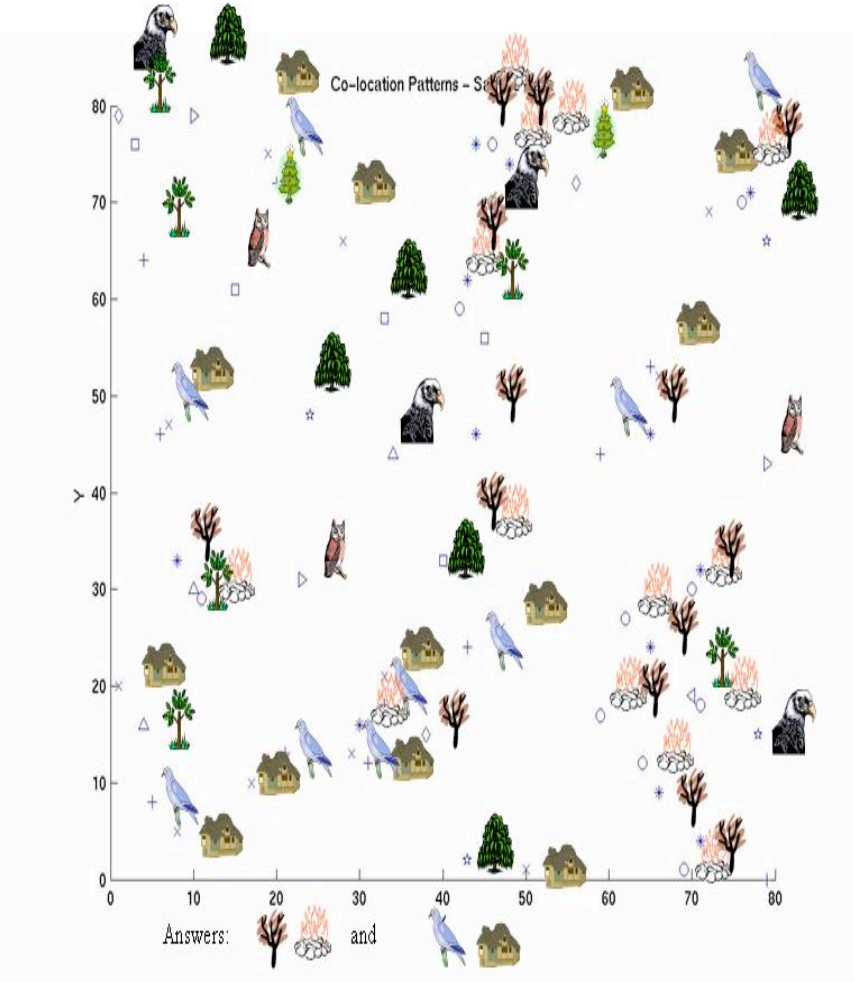
- Spatial association rule: $A \Rightarrow B$ [*s%, c%*]
  - A and B are sets of spatial or non-spatial predicates
    - Topological relations: *intersects, overlaps, disjoint,* etc.
    - Spatial orientations: *left_of, west_of, under,* etc.
    - Distance information: *close_to, within_distance,* etc.
  - *s%* is the support and *c%* is the confidence of the rule
- Examples

1) *is_a(x, large_town) ^ intersect(x, highway) → adjacent_to(x, water)*
[7%, 85%]

2) What kinds of objects are typically located close to golf courses?

Special case: co-location pattern mining

- Mining for objects that occur in nearby locations frequently
- Disregarding other spatial relations but 'nearness'
- Example:
  - Relationship between particular type of vegetation and animals (picture right)



http://www.cse.unt.edu/~huangyan/spatialMining.htm

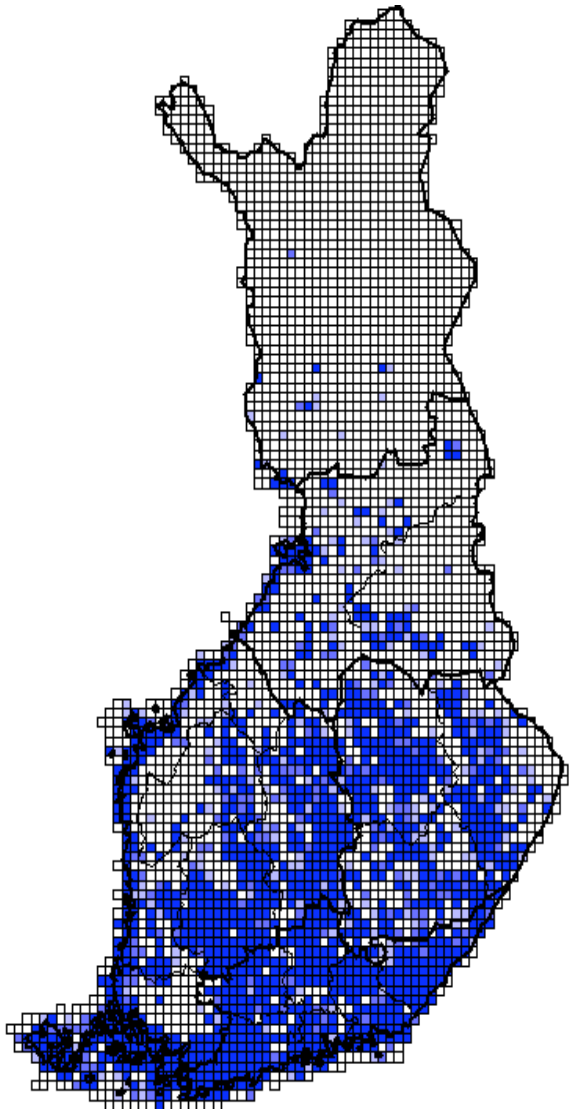
From itemsets to co-location patterns

- How to convert a spatial data into a form suitable for frequent pattern mining?
- Need to define the transaction/item structure
- Main alternatives:
  - Partition the space into regions and treat them as transactions
    - Fixed grid or pre-existing partition
  - Choose a reference point pattern and treat the neighbourhood of each of point as a transaction

Grid-based co-location mining

1. Divide the space into areas
   - Create a uniform grid that covers the space, grid cells as transactions
   - Objects in each grid cell as items
2. Use itemset mining algorithms to find frequent patterns

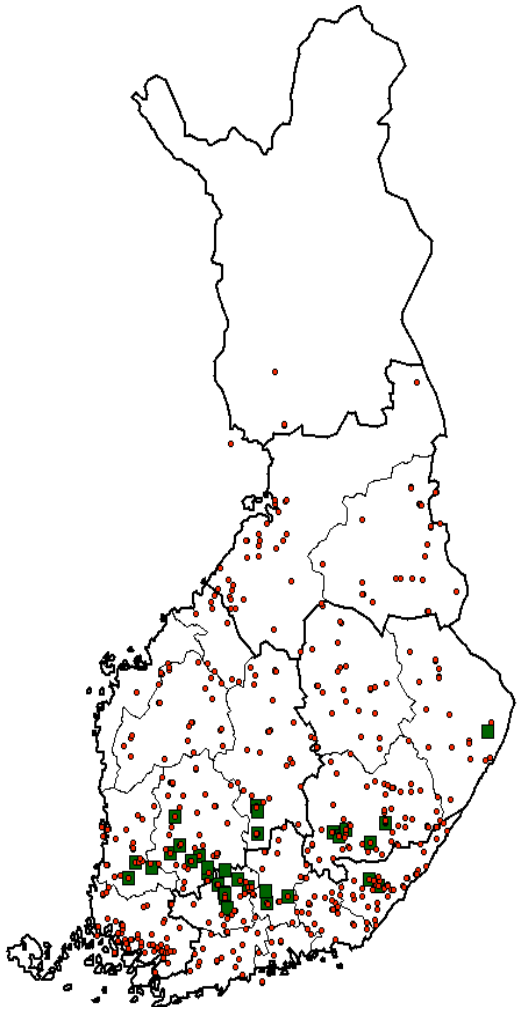
- How to choose the optimal grid resolution?



Reference feature centric co-location mining

1. Choose one point pattern as the reference (e.g. ‘Viking-age forts’)
2. Define a neighbourhood of each point in the reference pattern
   - neighborhoods as transactions
   - objects in each neighborhood as items
3. Use itemset mining algorithms to find frequent patterns

- Useful for applications where there is an natural choice for the reference phenomenon

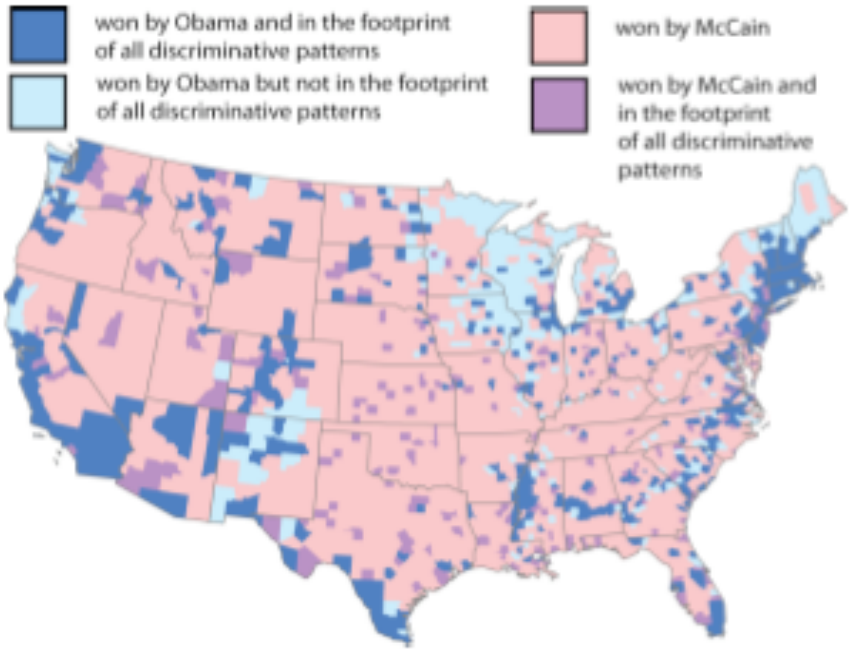


Case study in Spatial Data mining

- Spatial Data mining analysis of 2008 presidential election in the USA
- Source:
  - "Discovering Spatio-Social Motifs of Electoral Support Using Discriminative Pattern Mining" by Tomasz F. Stepinski and Josue Salazar, Wei Ding.
  - To be published in:1st International Conference on Computing for Geospatial Research & Aplication, Washington, DC, June, 2010.

Data

- Objects are counties in USA
- Characterized by their
  - geographical location (county ID #)
  - attributes f_i (socio-economic indicators), and
  - label (c) indicating whether the county was won (c=1) or lost (c=0) by Barack Obama.



$$o = [id, f_1, \ldots, f_m, c]$$

Socio-economic indicators

1. population density,
2. % of urban population,
3. % of female population,
4. % of foreign-born population,
5. per capita income,
6. median household income,
7. % of population with high school or higher education,
8. % of population with bachelor degree or higher education,
9. % of population that is white,
10. % of population living in poverty,
11. % of houses occupied by owners,
12. percentage of population receiving social security benefits,
13. average social security monthly benefit

Discriminative patterns

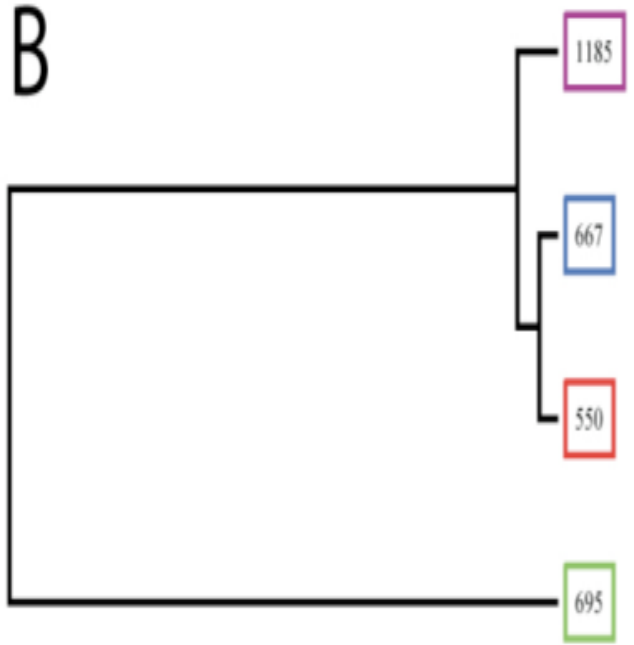
- A discriminating pattern is defined as an itemset consisting of the values of socio-economic indicators that has much larger support within a set of transactions with c=1 than in transactions with c=0
- Interestingness measure is the ratio of supports in the two sets

$$\delta(X) = \frac{s(X \cap \{c = 1\})}{s(X \cap \{c = 0\})}$$

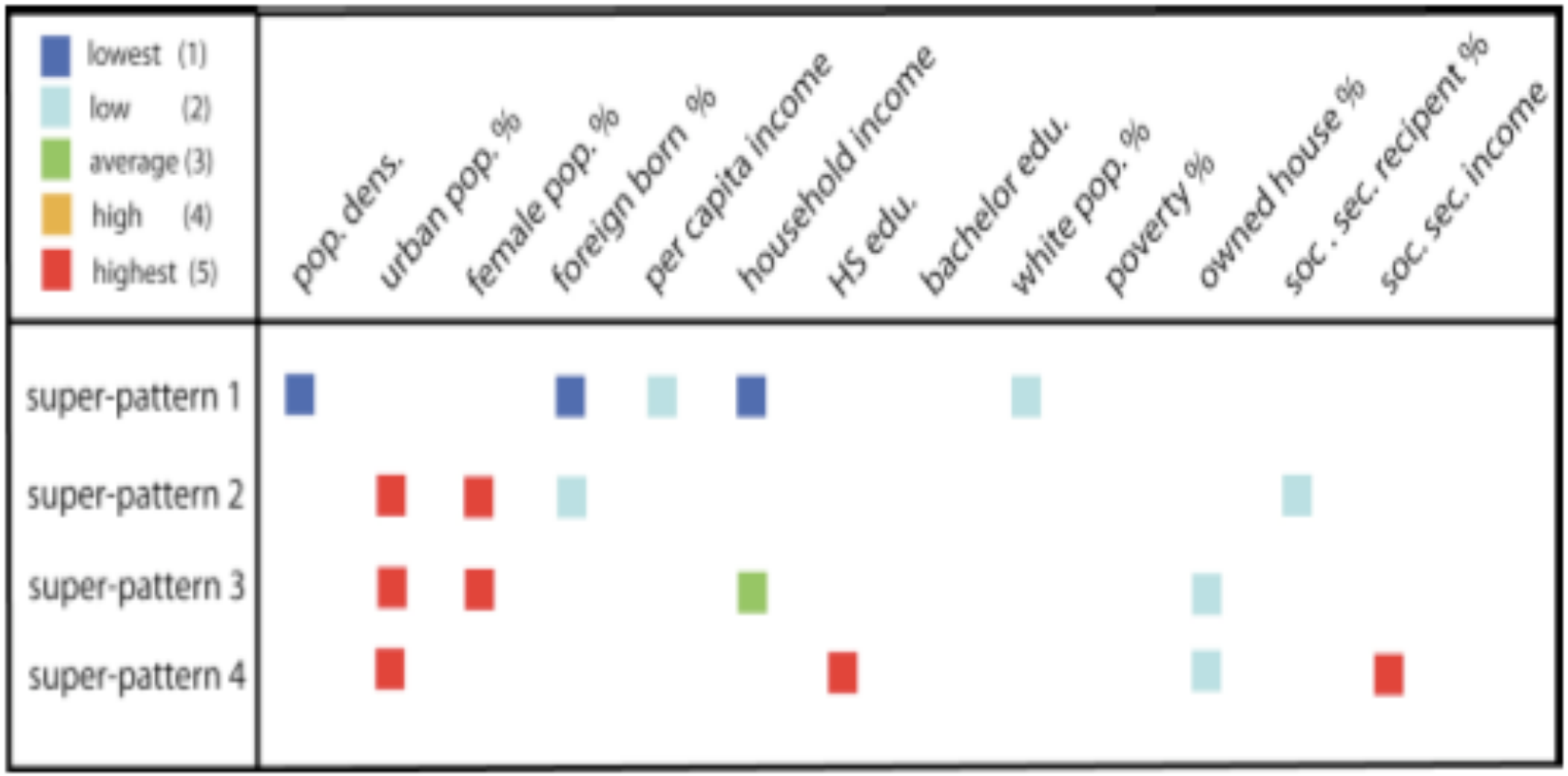
- 3097 discriminative patterns with ratio at least 15 is found in the data
  - too much to visualize or analyze by humans

Super-patterns

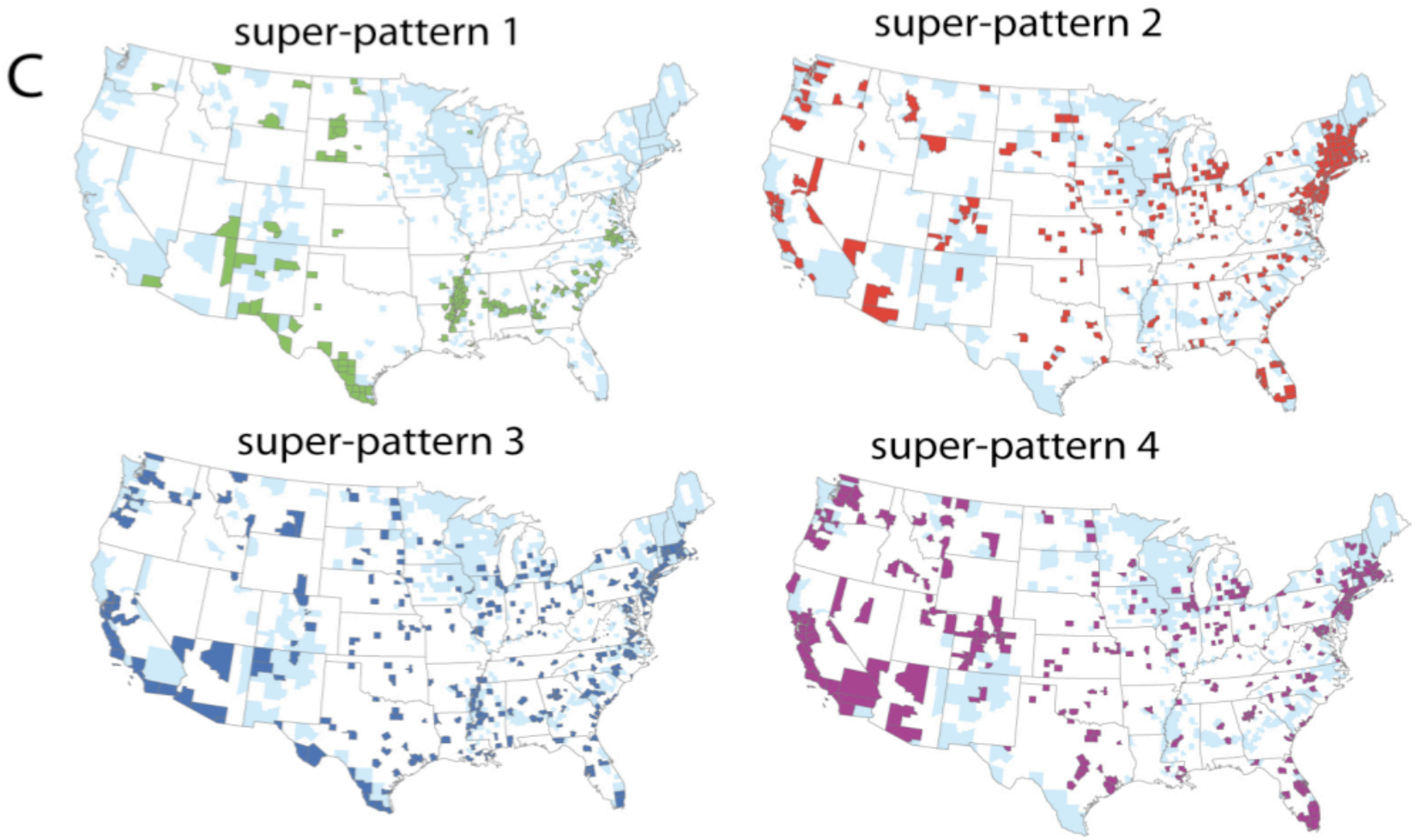
- The authors state a desire to mine for global patterns, whereas the frequent itemsets tend to contain nuggets of localized information
- From the set of 3097 discriminative patterns, super-patterns are synthesised by hierarchical clustering
- This process results in 4 superpatterns describing the whole space



Super-patterns



Geographical distribution of the super-patterns



Group work #3

- Group work session Tue 27.4 at 10.15-12 in room B222,
- Room B222 at your disposal also 12.15–14
- Debrief session Wed 28.4 at 9.00am (sharp!)

- Topic: Mining mobile (phone) data
  - Group work description is on the course web page.

Course exam

- Tuesday 4.5. at 9-12 in B123
- Examined contents
  - Lectures
  - Exercises 1 & 2
  - Groupwork and Papers **not** part of the examined contents

582635 Data mining project, 2 cr

- Separate course immediately after this course
- 10.5.-21.5
- Data mining techniques are applied in practice. Students can complete the course in two ways:
  - Either by implementing a data mining algorithm given in the assignment and by analyzing a given data with it,
  - or,by mining given data with a (wider) selection of methods, e.g. using ready-made software.
- In both cases, a research report is written describing the work and a seminar presentation is given.