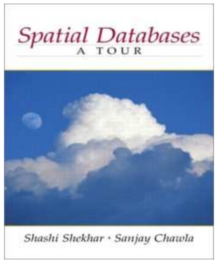
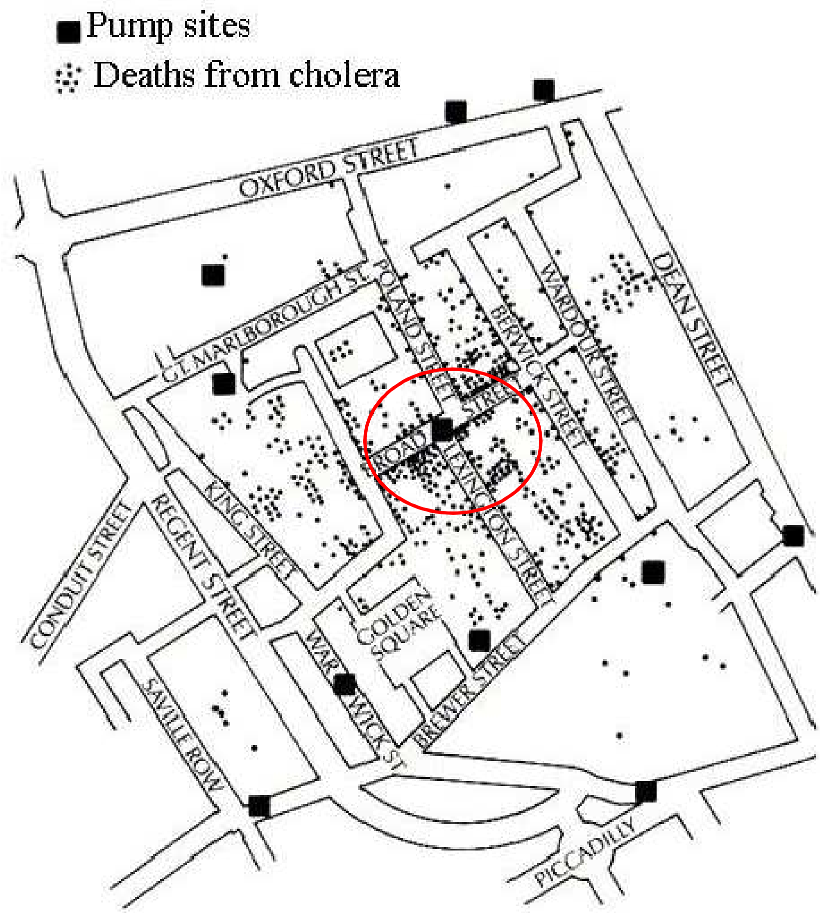


Introduction to Spatial Data Mining

Examples of Spatial Patterns

- Historical example
  - 1855 Asiatic Cholera in London: a water pump identified as the source
- Modern Examples
  - Cancer clusters to investigate environment h
  - Crime hotspots for planning police patrol rou
  - Bald eagles nest on tall trees near open wate
  - Nile virus spreading from north east USA to s
  - Unusual warming of Pacific ocean (El Nino) a



What is a Spatial Pattern?

- What is not a pattern?
  - Random, haphazard, chance, stray, accidental, unexpected
  - Without definite direction, trend, rule, method, design, aim, purpose
  - Accidental - without design, outside regular course of things
  - Casual - absence of pre-arrangement, relatively unimportant
  - Fortuitous - What occurs without known cause

- What is a pattern?
  - A frequent arrangement, configuration, composition, regularity
  - A rule, law, method, design, description
  - A major direction, trend, prediction
  - A significant surface irregularity or unevenness

What is Spatial Data Mining?

- Metaphors
  - Mining nuggets of information embedded in large databases
    - Nuggets = interesting, useful, unexpected spatial patterns
    - Mining = looking for nuggets
  - Needle in a haystack
- Defining Spatial Data Mining
  - Search for spatial patterns
  - **Non-trivial search** - as "automated" as possible—reduce human effort
  - **Interesting, useful** and **unexpected** spatial pattern

What is Spatial Data Mining?

- Non-trivial search for **interesting** and **unexpected** spatial pattern
- Non-trivial Search
  - Large (e.g. exponential) search space of plausible hypothesis
  - Ex. Asiatic cholera : causes: water, food, air, insects, …; water delivery mechanisms - numerous pumps, rivers, ponds, wells, pipes, ...
- Interesting
  - Useful in certain application domain
  - Ex. Shutting off identified Water pump => saved human life
- Unexpected
  - Pattern is not common knowledge
  - May provide a new understanding of world
  - Ex. Water pump - Cholera connection lead to the “germ” theory

What is NOT Spatial Data Mining?

- Simple Querying of Spatial Data
  - Find neighbors of Canada given names and boundaries of all countries
  - Find shortest path from Boston to Houston in a freeway map
  - Search space is not large (not exponential)
- Testing a hypothesis via a primary data analysis
  - Ex. Female chimpanzee territories are smaller than male territories
  - Search space is not large !
  - SDM: secondary data analysis to generate multiple plausible hypotheses
- Uninteresting or obvious patterns in spatial data
  - Heavy rainfall in Minneapolis is correlated with heavy rainfall in St. Paul, Given that the two cities are 10 miles apart.
  - Common knowledge: Nearby places have similar rainfall
- Mining of non-spatial data
  - Diaper sales and beer sales are correlated in evenings
  - GPS product buyers are of 3 kinds:
    - outdoors enthusiasts, farmers, technology enthusiasts

Why Learn about Spatial Data Mining?

- Two basic reasons for new work
  - Consideration of use in certain application domains
  - Provide fundamental new understanding

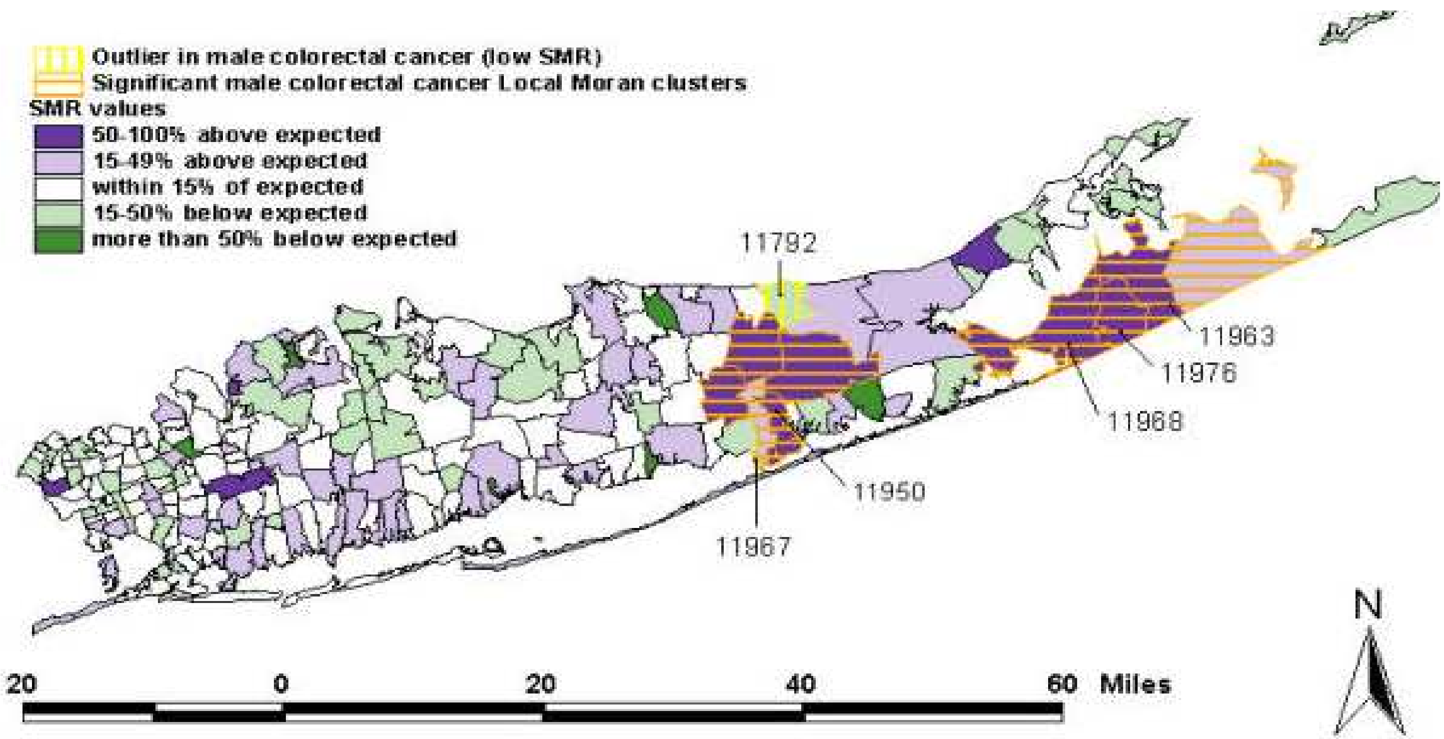
- Application domains
  - Scale up secondary spatial (statistical) analysis to very large datasets
    - Describe/explain locations of human settlements in last 5000 years
    - Find cancer clusters to locate hazardous environments
    - Prepare land-use maps from satellite imagery
    - Predict habitat suitable for endangered species
  - Find new spatial patterns
    - Find groups of co-located geographic features

Why Learn about Spatial Data Mining?

- New understanding of geographic processes for Critical questions
  - Ex. How is the health of planet Earth?
  - Ex. Characterize effects of human activity on environment and ecology
  - Ex. Predict effect of El Nino on weather, and economy
- Traditional approach: manually generate and test hypothesis
  - But, spatial data is growing too fast to analyze manually
    - Satellite imagery, GPS tracks, sensors on highways, …
  - Number of possible geographic hypothesis too large to explore manually
    - Large number of geographic features and locations
    - Number of interacting subsets of features grow exponentially
    - Ex. Find tele connections between weather events across ocean and land areas
- SDM may reduce the set of plausible hypothesis
  - Identify hypothesis supported by the data
  - For further exploration using traditional statistical methods

Example

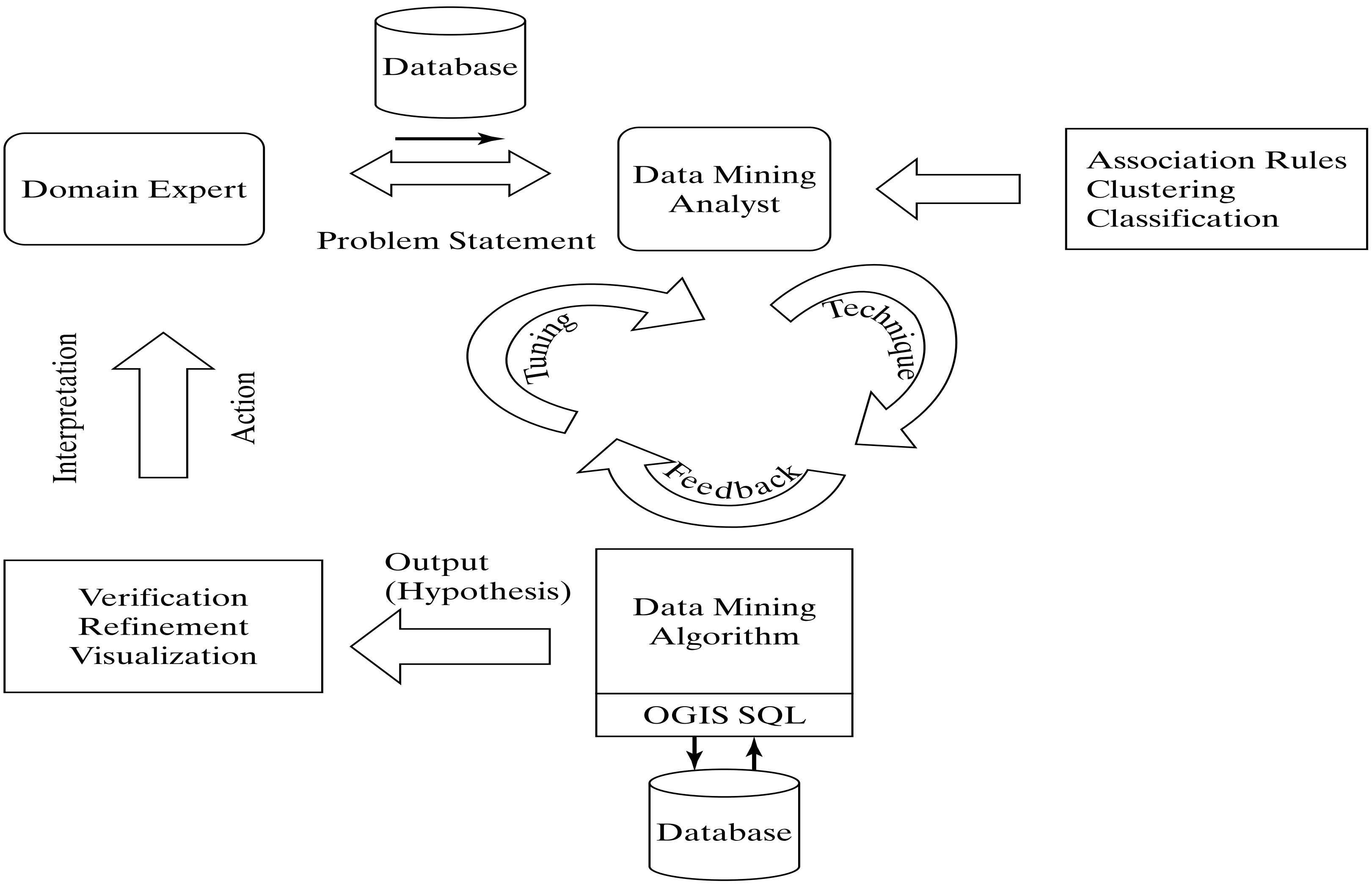
- What is the overall pattern of colorectal cancer
- Is there clustering of high colorectal cancer incidence anywhere in the study area
- Where is colorectal cancer risk significantly elevated



Spatial Data Mining: Actors

- Domain Expert
  - Identifies SDM goals, spatial dataset,
  - Describe domain knowledge, e.g. well-known patterns, e.g. correlates
  - Validation of new patterns
- Data Mining Analyst
  - Helps identify pattern families, SDM techniques to be used
  - Explain the SDM outputs to Domain Expert
- Joint effort
  - Feature selection
  - Selection of patterns for further exploration

Spatial Data Mining Process



Families of SDM Patterns

- Common families of spatial patterns
  - Classification
  - Clustering
  - Spatial Association Rules
  - Co-location
  - Outliers detection
  - ..
- Note
  - Other families of spatial patterns may be defined
  - SDM is a growing field, which should accommodate new pattern families

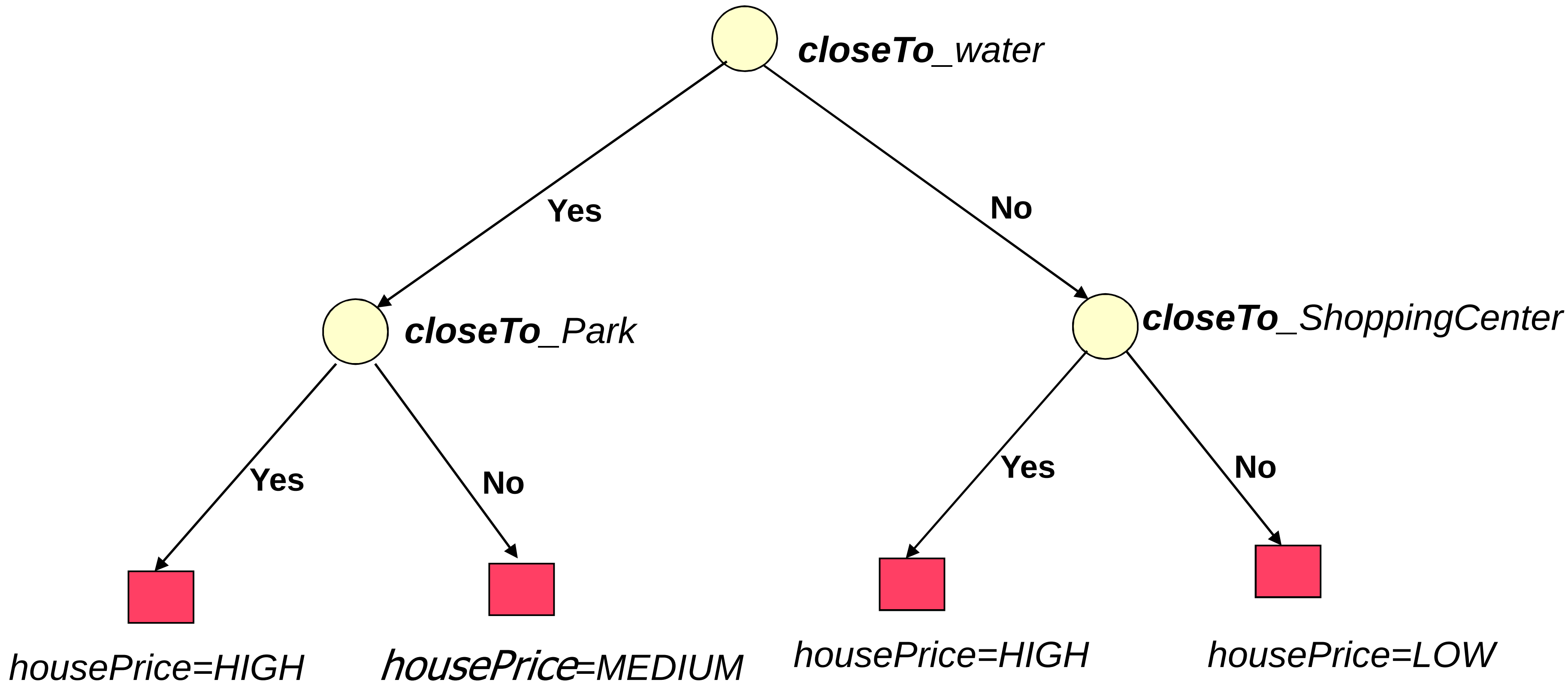
Classification

- Given a set of instances, the role of classification is to discover the classes of the instances

- Spatial objects may be characterized (classified) by different types of information (Koperski 1998):
  - non-spatial attributes (e.g. population);
  - spatially related attributes with non-spatial values (e.g. total population living within 100 meters from cellular antennas);
  - spatial predicates (e.g. closeTo_beach)

Ester (1997, 2001)

Class is a non-spatial attribute = *housePrice*
Class values: high, medium, low



Remote Sensing Data Mining

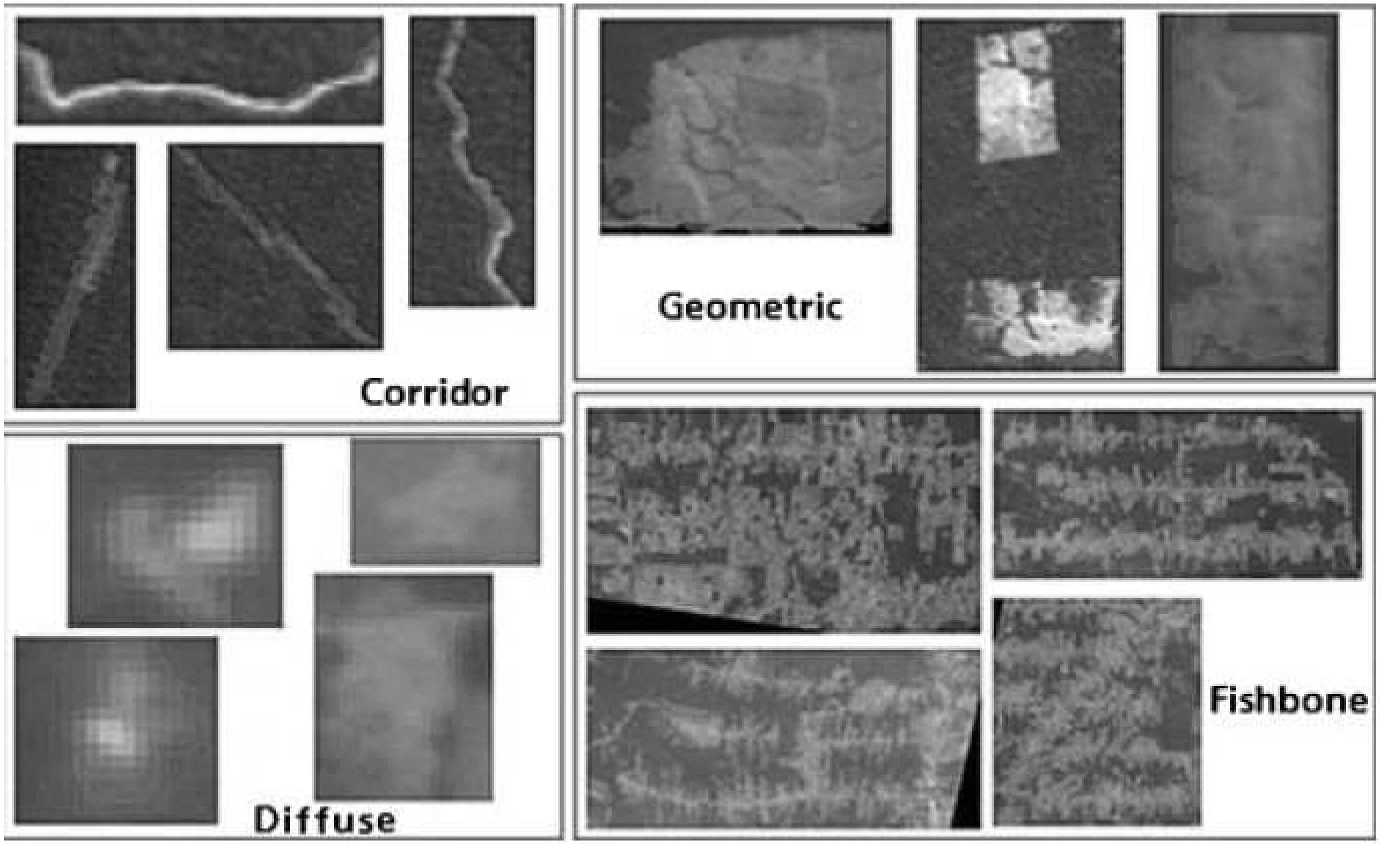


Figure 2. Examples of patterns of tropical deforestation proposed by Mertens and Lambin (1997) in the Brazilian Amazonia: corridor, diffuse, fishbone, and geometric.

Remote Sensing Data Mining

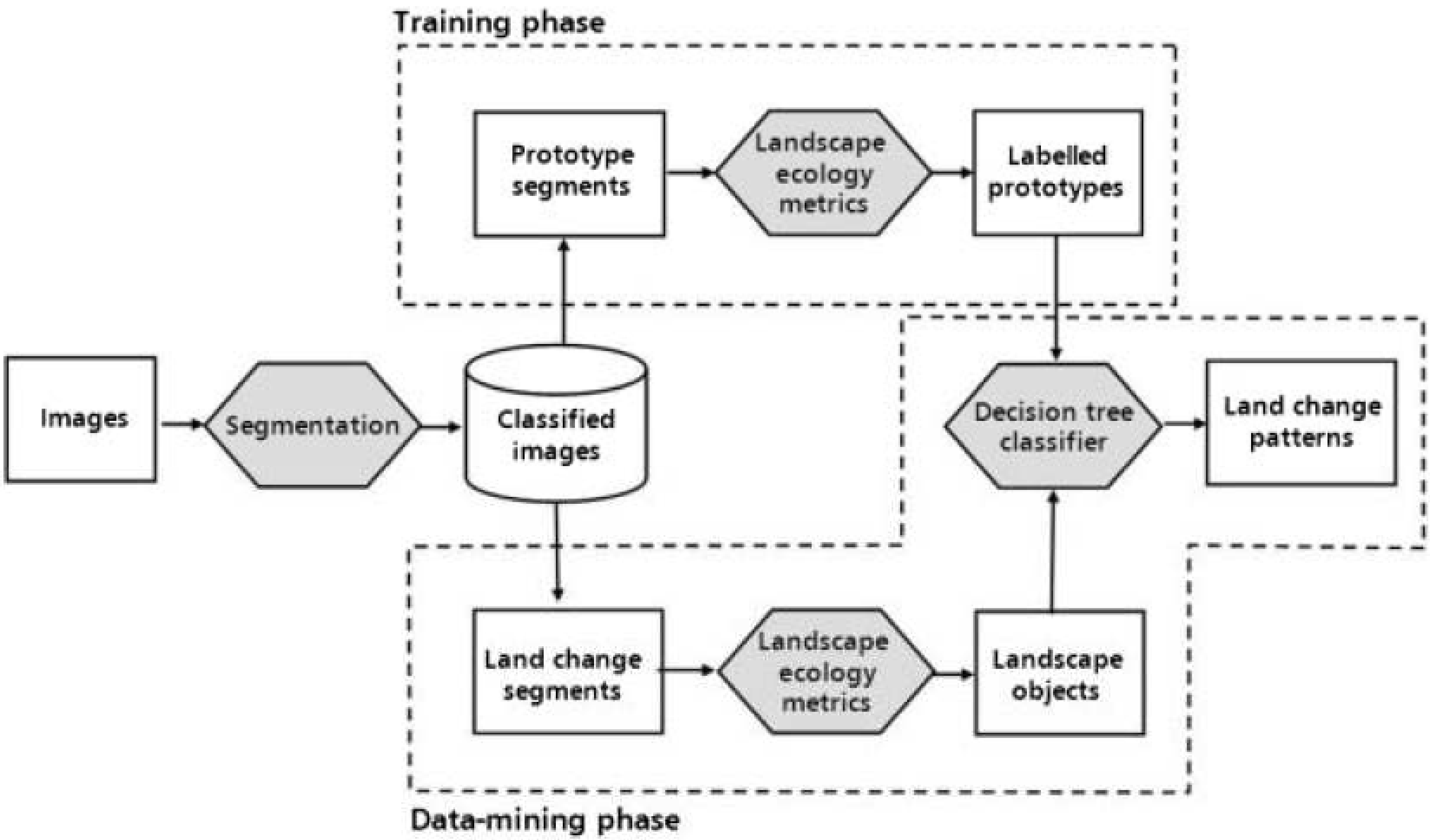


Figure 1. Proposed method for remote sensing image mining.

Metrics

- Perimeter (m):

$$\text{PERIM} = p_{ij}. \tag{1}$$

- Area (ha):

$$\text{AREA} = \left(a_{ij}/10\,000\right). \tag{2}$$

- PARA, perimeter–area ratio, a measure of shape complexity:

$$\text{PARA} = \frac{p_{ij}}{a_{ij}}. \tag{3}$$

- Shape, shape compactness index, calculated by the patch perimeter p_{ij} divided by $p_{ij\ \min}$, which is the minimum perimeter possible for a maximally compact patch of the matching patch area. It is equal to 1 when the region is a square and grows according to the region's irregularity.

$$\text{SHAPE} = \frac{p_{ij}}{p_{ij\,\min}}. \tag{4}$$

Decision Tree

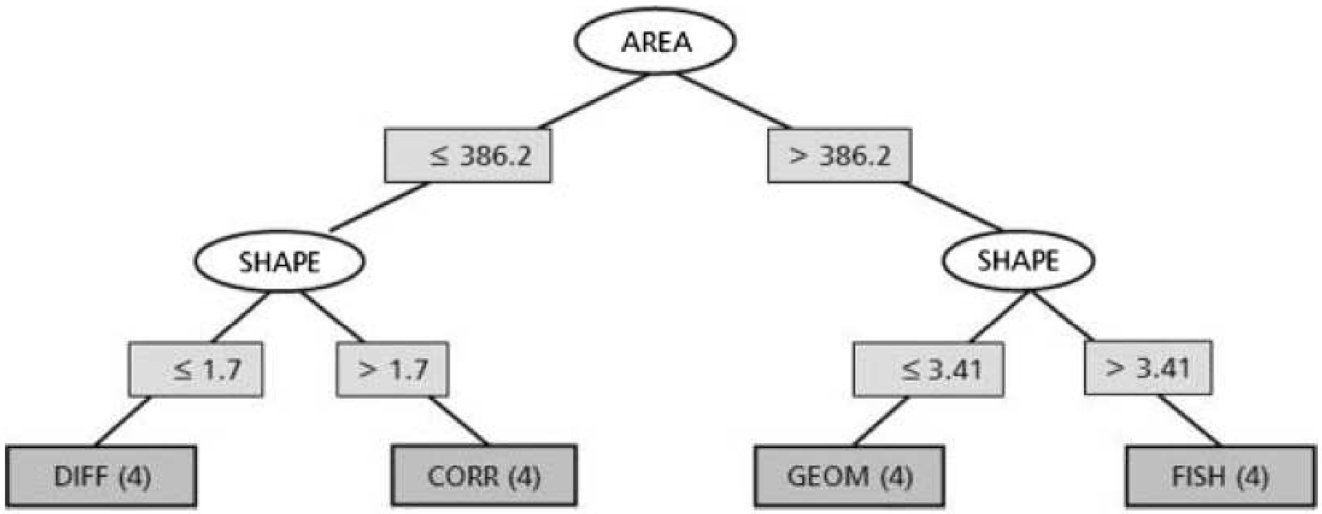


Figure 3. Decision tree for patterns in figure 3 (GEOM: geometric; FISH: fishbone; DIFF: diffuse; CORR: corridor). Metrics: area in km^2 (AREA) and shape compactness index (SHAPE).

Results

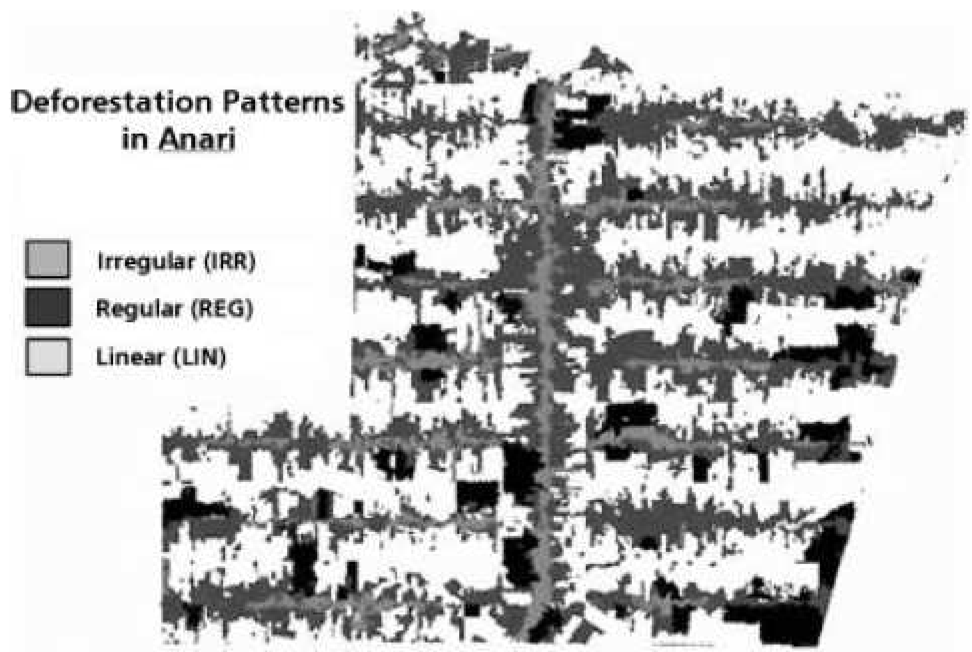


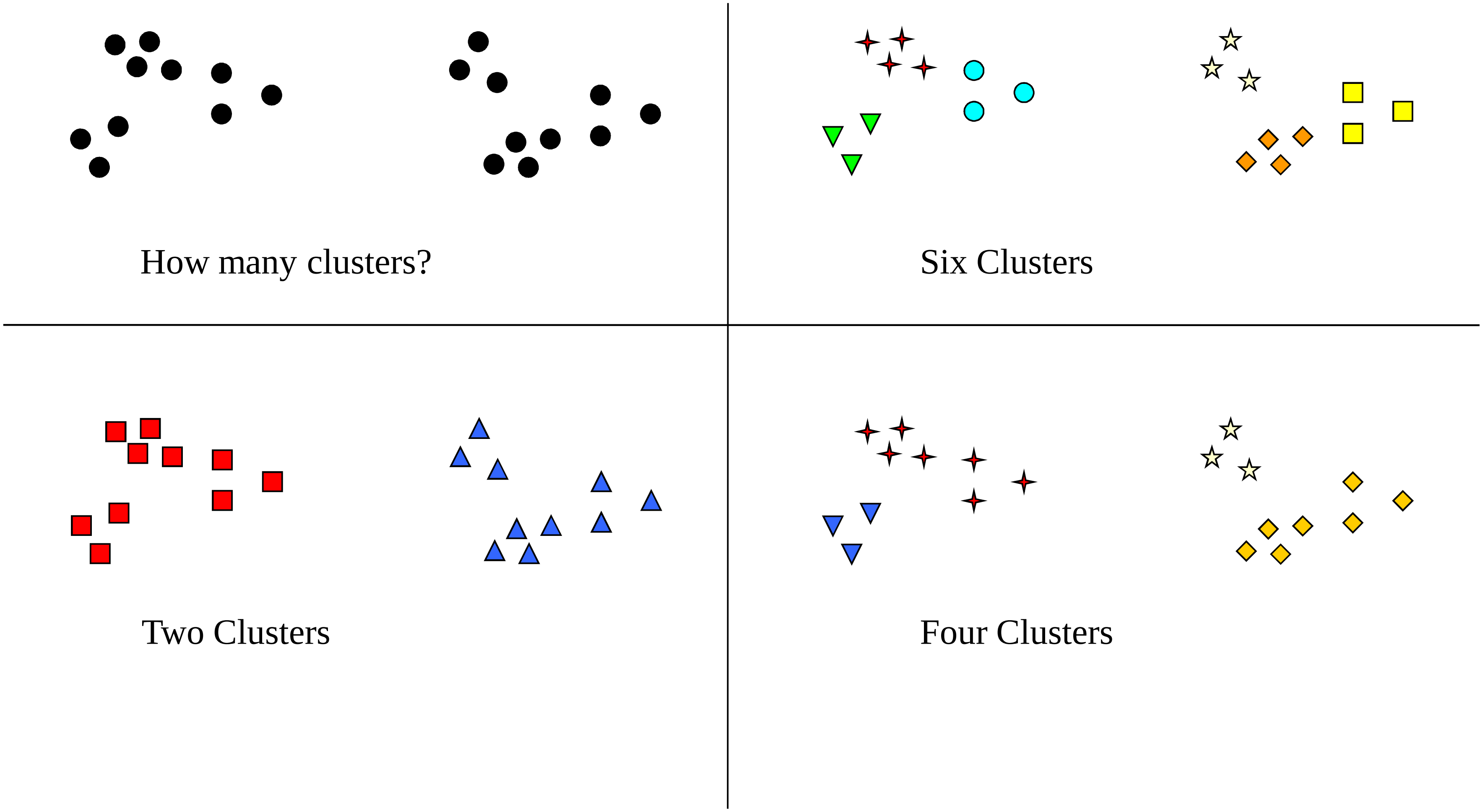
Figure 12. Cumulative deforestation patterns in Vale do Anari (1985–2000).

Clustering (cluster analysis)

- Clustering is a process of partitioning a set of data into a set of groups called *clusters*
- A cluster is a set of data (objects) with
  - similar characteristics
  - that can be collectively treated as one group
- Clustering is an unsupervised method
  - no predefined classes

Clustering Analysis (Kumar 2005)

Different ways of clustering the same set of points



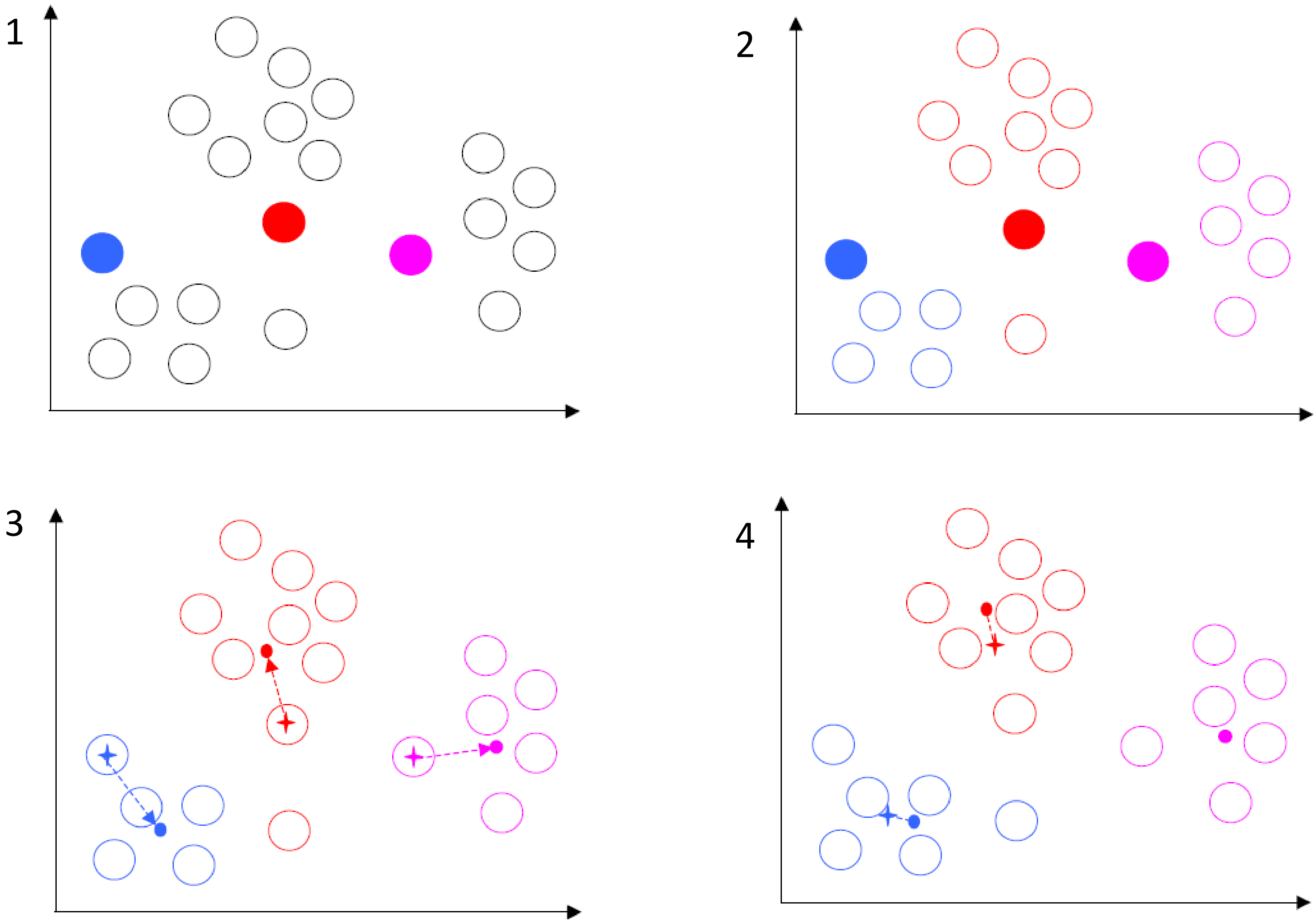
Main Clustering Approaches

- Partitioning
  - A division of data objects into non-overlapping subsets (clusters) such that *each object* is in exactly one subset
- Hierarchical
  - A set of nested clusters organized as a hierarchical tree
- Density-based
  - Find clusters based on density of regions
- Grid-based
  - Find clusters based on the number of points in each cell

K-means

- Partitional clustering approach
- Each cluster is associated with a centroid
- Each point is assigned to the cluster with the closest centroid
- A drawback of the k-means is that the number of clusters K is na input parameter

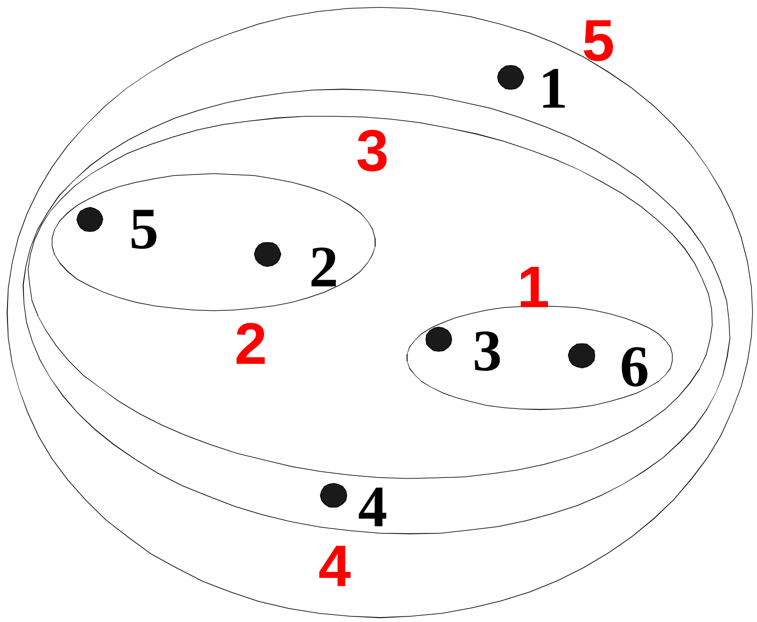
K-means

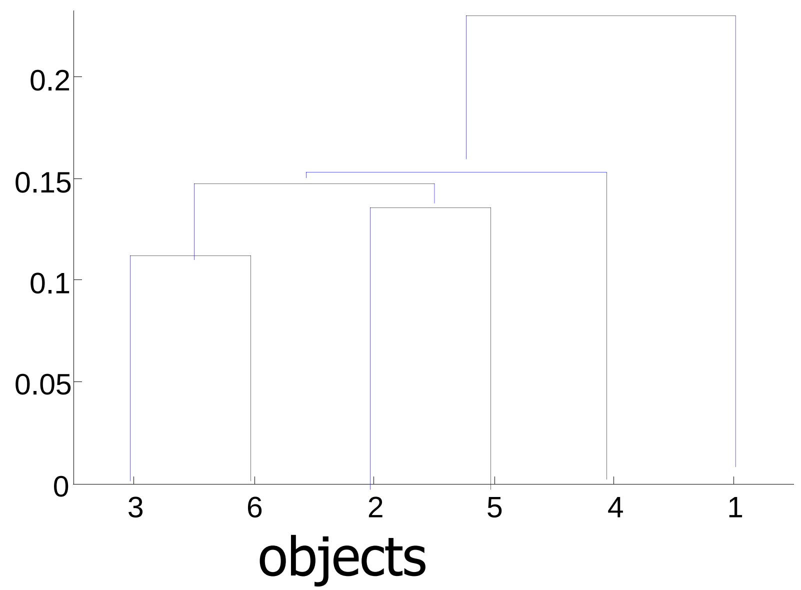


Hierarchical Clustering

Two main types: Agglomerative and Divisive

- Agglomerative
  - Start with all objects as individual clusters
  - At each step, merge the two most similar clusters
  - Until rests one cluster (or k clusters)





Hierarchical Clustering

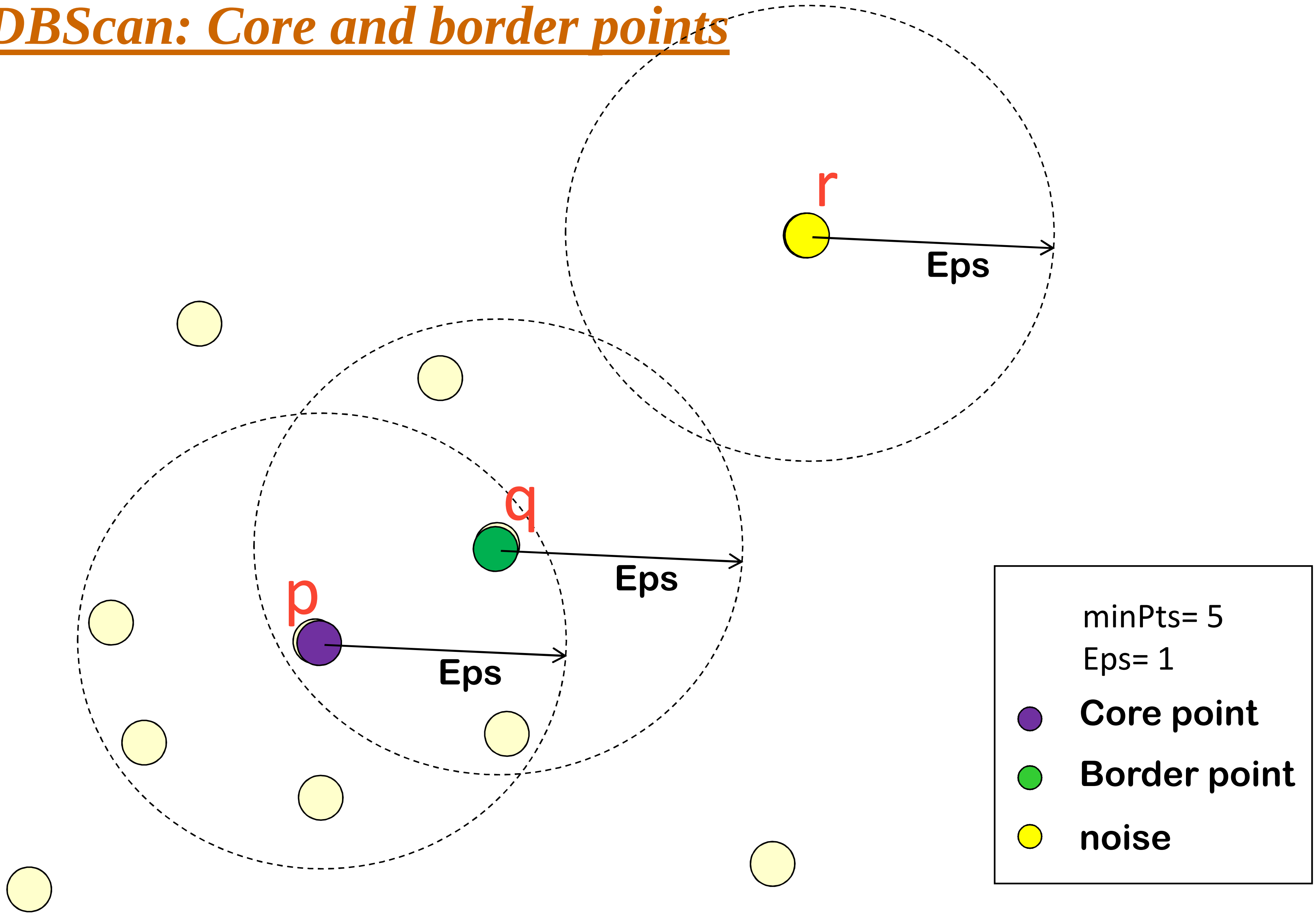
- Divisive
  - Start with one cluster (with all objects)
  - At each step, split a cluster in two
  - Until each cluster contains only one object (or k clusters)

Similarity can be euclidean distance or any other measure

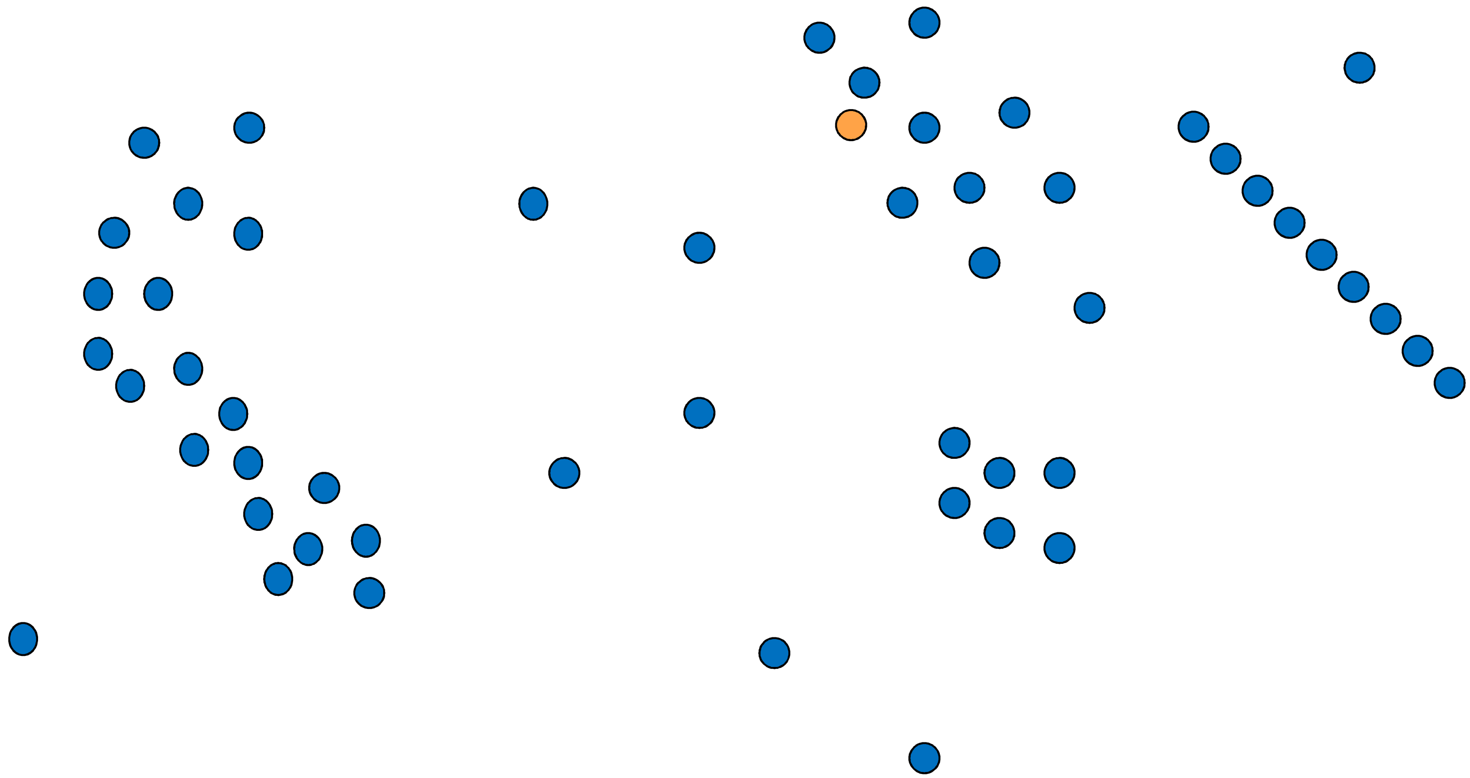
DBSCAN (Ester 1996)

- DBSCAN is a density-based algorithm
- Density = number of points within a specified radius (Eps)
- A point is a core point if it has more than a specified number of points (MinPts) within Eps
- A border point has less than MinPts within Eps, but it is in the neighborhood of a core point
- A noise point is any point that is not a core point or a border point.

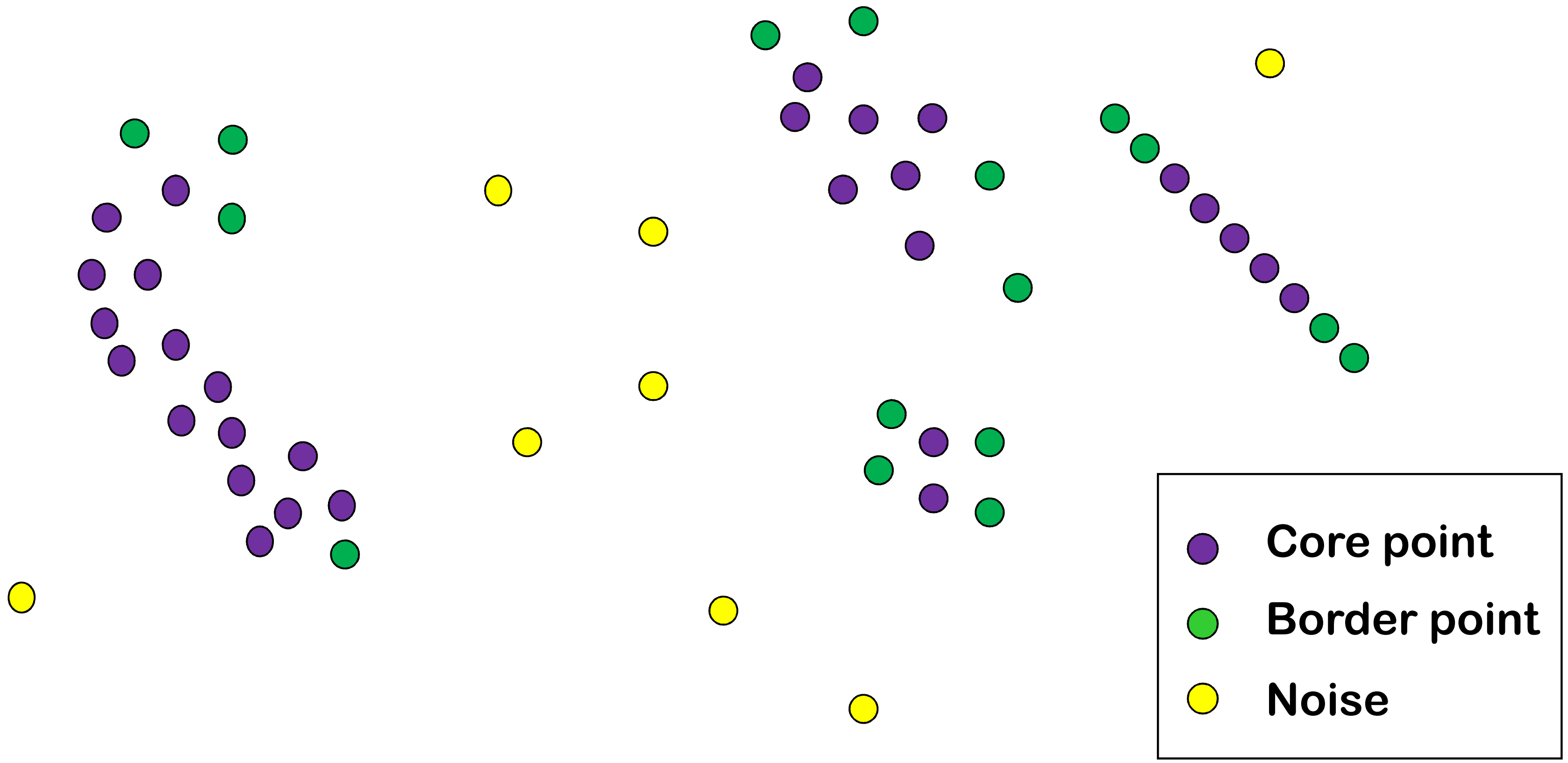
DBScan: Core and border points



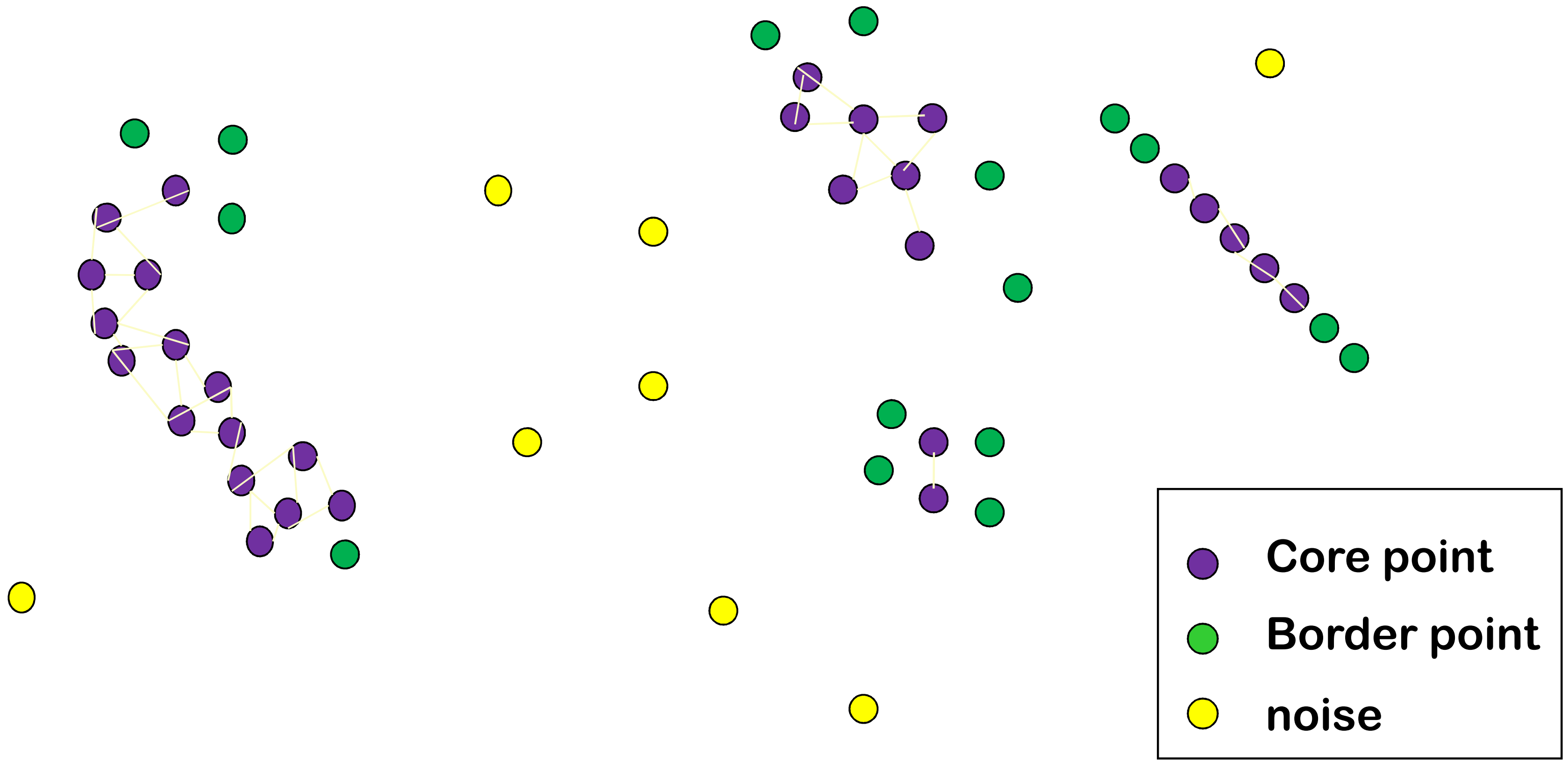
DBSCAN example



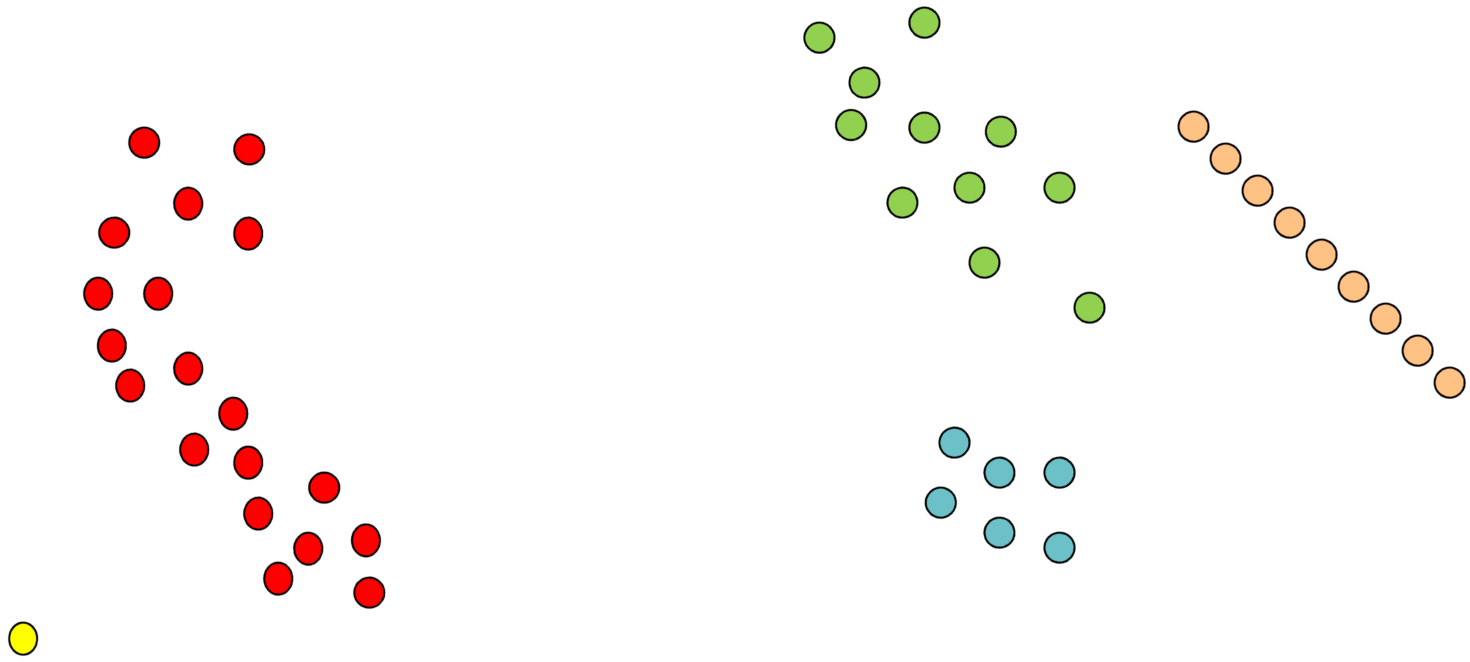
Identifying core, border and noise points



Computing distance



Final Clusters



Spatial Association Rules

Association Rules (Agrawal 1993)

- Association rule is an implication of form:

  $X \rightarrow Y$

Support: #(X∪Y) / T, where T number of transactions in the dataset

Confidence: Support(X∪Y)/Support(X)

1

Generate frequent itemsets

Tid	Itemset
1	A, C, D,T, W
2	C, D, W
3	A, D, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

Set *k*	Frequent itemsets with minsup 50%
k=1	{A}, {C}, {D}, {T}, {W}
k=2	{A,C}, {A,D}, {A,T}, {A,W}, {C,D}, {C,T}, {C,W}, {D,T}, {D,W}, {T,W}
k=3	{A,C,D}, {A,C,W}, {A,D,T}, {A,D,W}, {A,T,W}, {C,D,T}, {C,D,W}, {D,T,W}
k=4	{A,C,D,W}, {A,D,T,W}

2

Extract association rules

Support {AC} = 3/6 (50%)

Confidence A→C = 3/4 (75%)

Associataion rules

- Main problem: generate hundreds or thousands of rules

- Frequent Itemsets: generate all possible frequent itemsets
  - Apriori-like (generate candidates) (Agrawal, 1994)
  - Pattern-growth (without candidate generation) (Han, 2000)

- Closed frequent itemsets: generate non-redundant frequent itemsets
  - Apriori-like (generate candidates) (Pasquier, 1999) (Zaki, 2000)
  - Pattern-growth (without candidate generation) (Han, 2001) (Zaki 2002)……..

Redundant Rules

A Redundant rule has same support and confidence of another rule generated from the same set of transactions

Frequent Itemsets

Tid	Itemset
1	A, C, D, T, W
2	C, D, W
3	A, D, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

TidSet	Frequent itemsets with *minsup* 50%
123456	**{D}**
12456	{C}, **{C,D}**
12345	{W}, **{D,W}**
1245	{C,W}, **{C,D,W}**
1345	{A}, {A,D}, {A,W}, **{A,D,W}**
1356	{T}, **{D,T}**
145	{A,C}, {A,C,W}, {A,C,D}, **{A,C,D,W}**
135	{A,T}, {T,W}, {A,D,T}, {A,T,W}, {D,T,W}, **{A,D,T,W}**
156	{C,T}, **{C,D,T}**

A→W (suport= 4/6) (confidence = 4/4)

A→ DW (suport = 4/6) (confidence = 4/4)

25 frequent itemsets / 9 closed frequent itemsets

Spatial association Rules

- Spatial association rule is an implication of the form
  $X \rightarrow Y$ (support)(confidence)

- at least one element in X or Y is a spatial predicate
  - closeTo_slum $\rightarrow$ criminalityRate=High
  - Touches_beach $\rightarrow$ housePrice=High

Different Spatial Objects are Stored in Different Tables

Street

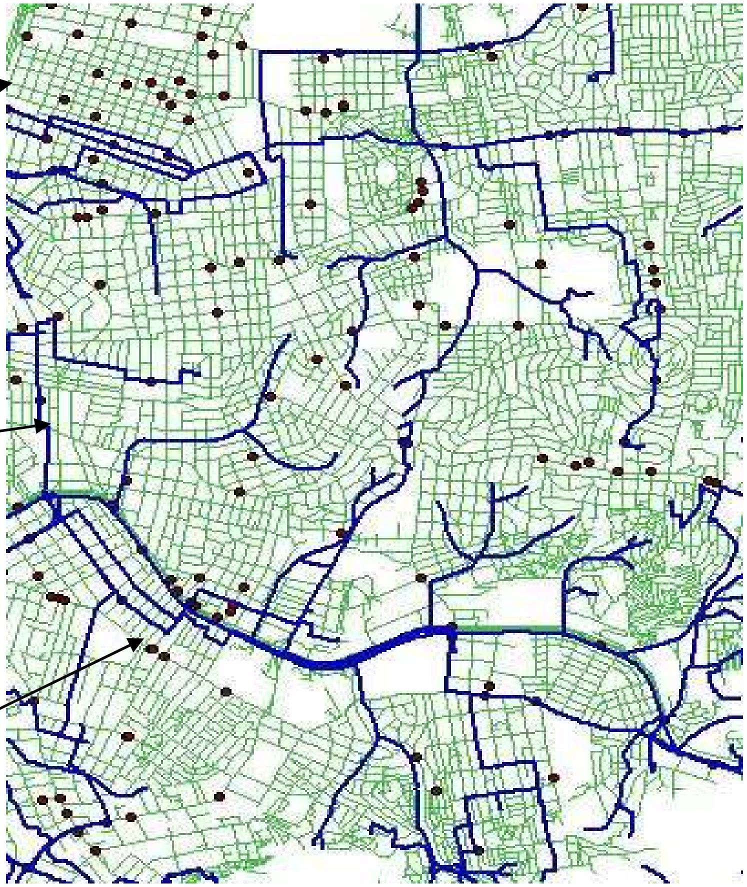
Gid	Name	Shape
1	Ijui	Multiline [(x1,y1),(x2,y2),..]
2	Lavras	Multiline [(x1,y1),(x2,y2),..]

WaterResource

Gid	Name	Shape
1	Jacui	Multiline [(x1,y1),(x2,y2),..]
2	Guaiba	Multiline [(x1,y1),(x2,y2),..]
3	Uruguai	Multiline [(x1,y1),(x2,y2),..]

GasStation

Gid	Name	VolDiesel	VolGas	Shape
1	BR	20000	85000	Point[(x1,y1)]
2	IPF	30000	95000	Point[(x1,y1)]
3	Esso	25000	120000	Point[(x1,y1)]



Most Spatial Association Rule Mining algorithms have a single table/file INPUT format

Different Relations (tables) need to be Spatially Joined

Preprocessed Geographic Data for Transaction-Based Data Mining

Target feature

Tuple (city)	Spatial Predicates				
1	contains(Port),	contains(Hospital),	contains(Street),	contains(Factory),	crosses(Water Body)
2		contains(Hospital),	contains(Street),		crosses(Water Body)
3	contains(Port),		contains(Street),	contains(Factory),	crosses(Water Body)
4	contains(Port),	contains(Hospital),	contains(Street),		crosses(Water Body)
5	contains(Port),	contains(Hospital),	contains(Street),	contains(Factory),	crosses(Water Body)
6		contains(Hospital),	contains(Street),	contains(Factory)	

Relevant features

Spatial Association Rules

- Are computed in 3 main steps:
  - Data preprocessing: compute spatial relationships (spatial joins). Most expensive step
  - Compute frequent itemsets
  - Generate association rules

Transaction Dataset X Preprocessed Spatial Dataset

Transactional Dataset

Transaction	Items
1	milk, bread, butter, cereal
2	milk, bread
3	beer, bread, chocolate
4	cereal, meet, milk
5	milk, beer, nuts, orange, cereal

- rows are transactions
- attributes are items, supposed to be independent

Spatial Dataset

Tuple (city)	Spatial Predicates
1	contains(Port), contains(Hospital), contains(TreatedWaterNet), contains(Factory), crosses(WaterBody
2	contains(Hospital), contains(TreatedWaterNet), crosses(WaterBody)
3	contains(Port), contains(TreatedWaterNet), contains(Factory), crosses(WaterBody)
4	contains(Port), contains(Hospital), contains(TreatedWaterNet), crosses(
5	contains(Port), contains(Hospital), contains(TreatedWaterNet), contains(Factory), crosses(W
6	contains(Hospital), contains(TreatedWaterNet), contains(Factory)

- rows are instances of the target feature type

- attributes are predicates
- spatial predicates are spatial relationships between the target feature type and relevant feature types

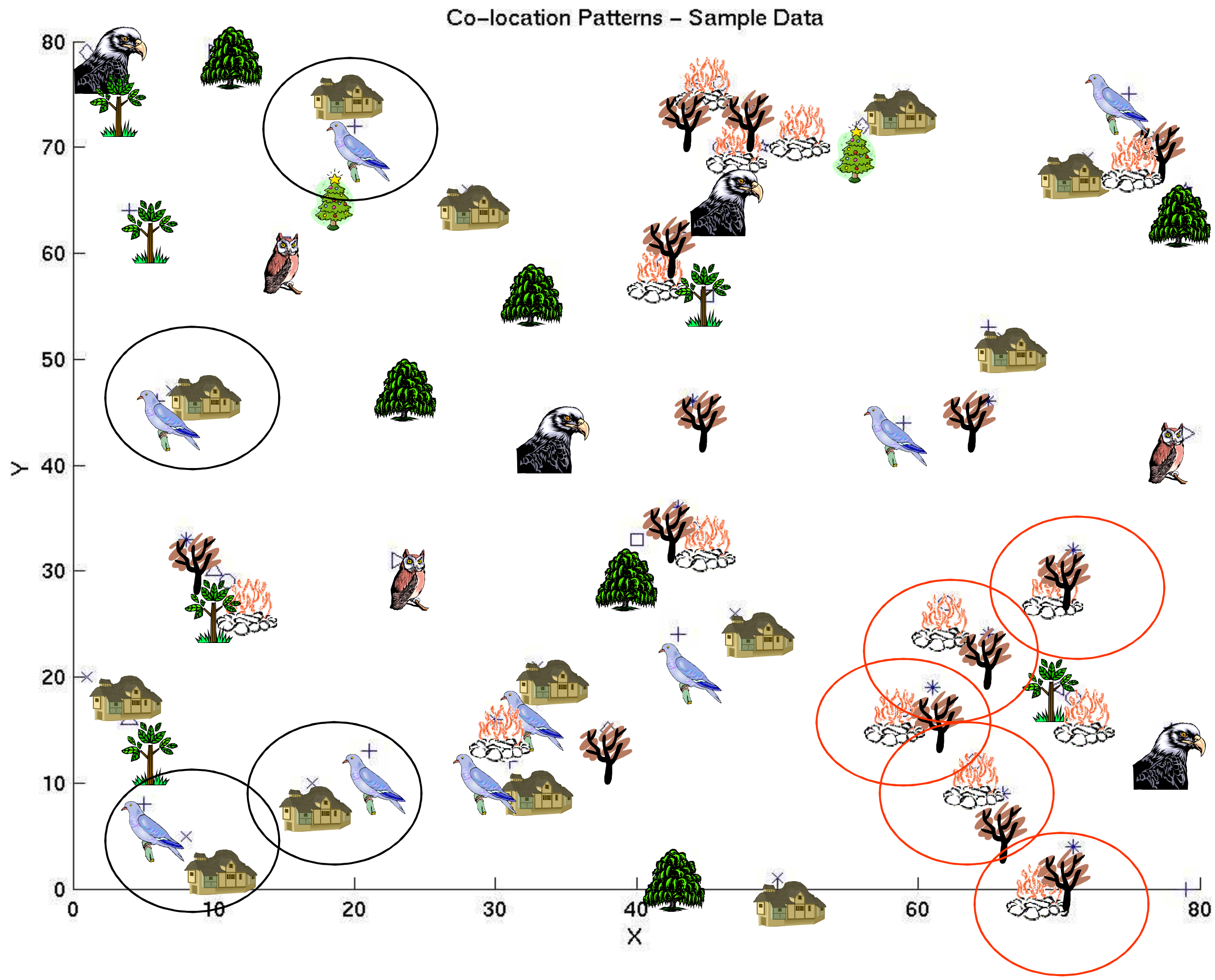
Some Spatial Association Rule Mining Algorithms

- Koperski 1995
- Spada (Appice 2003)
- Clementini (2003)
- Apriori-KC (Bogorny 2006)
- Max-FGP (Bogorny 2006[a])
- ...

- Preprocess geographic data and apply classical DM algorithms

Co-location (Shekhar 2003)

Q: find patterns from the following sample dataset



Answers:

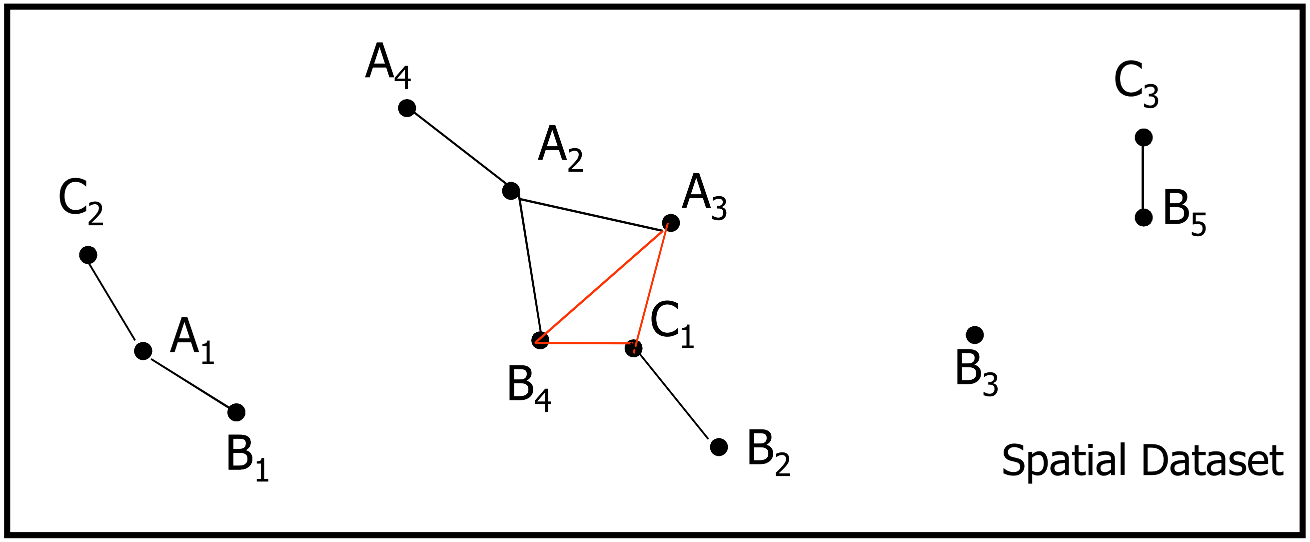
Co-Location Patterns (Huang 2004, Yoo 2005)

- Input:
  - Spatial dataset
  - Distance threshold
  - Minimum participation index

- Method
  - Find neighbors
  - Find co-location candidates
  - Find frequent co-location sets
  - Extract co-location rules

Co-location Mining

A-School
B-Hospital
C-Pharmacy

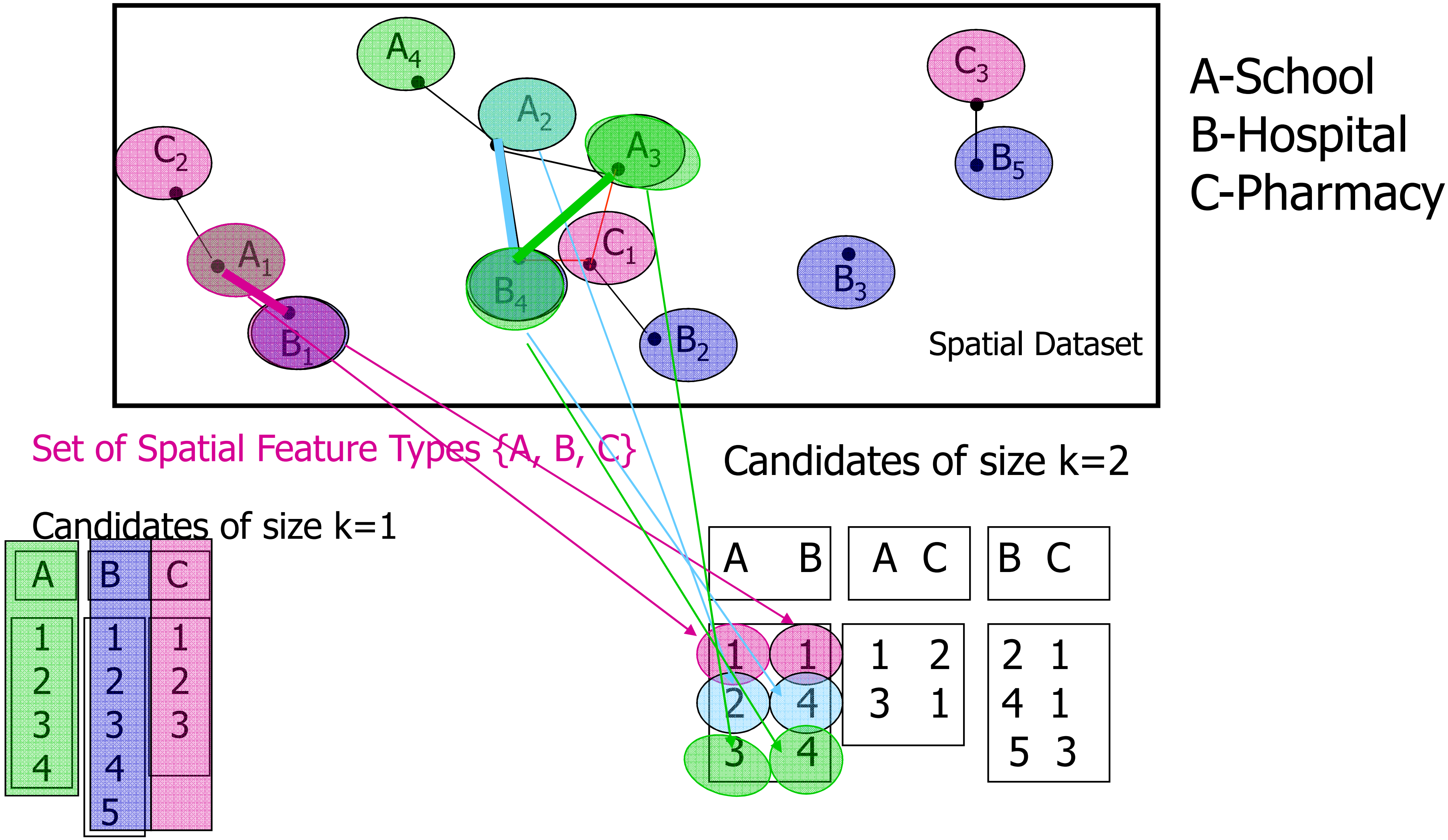


A, B, C: Spatial Feature Types

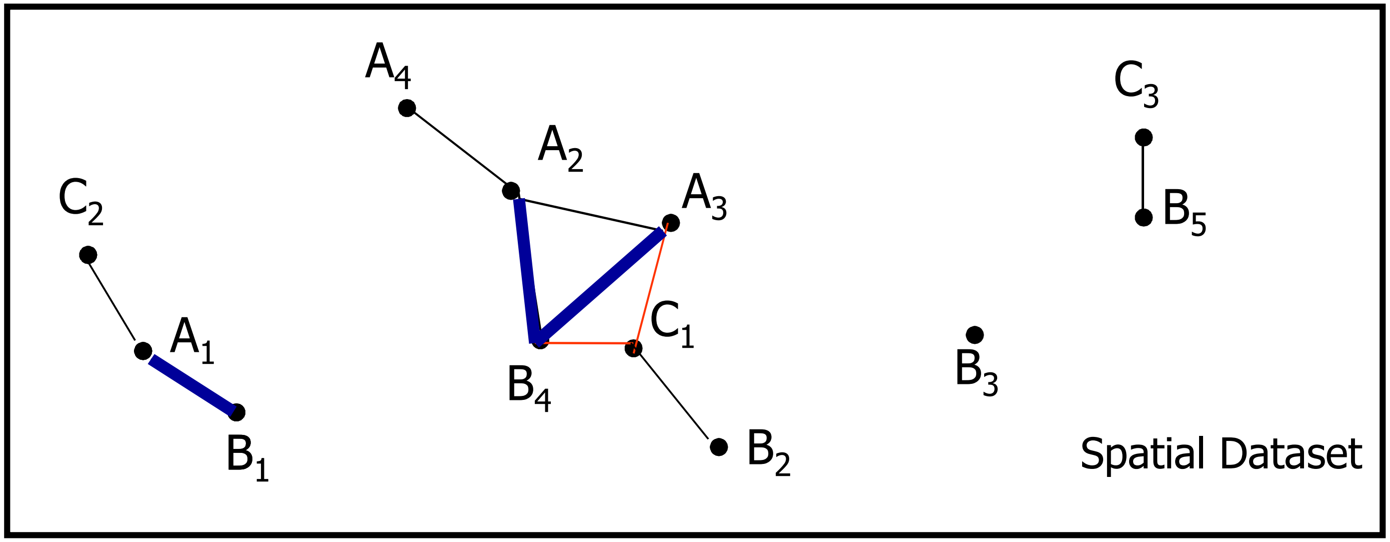
A1, A2... Spatial Feature Instances

Edges: neighbor

Co-location Mining



Co-location Mining



A-School
B-Hospital
C-Pharmacy

Candidates of size k=1

A	B	C
1	1	1
2	2	2
3	3	3
4	4	
	5	

Candidates of size k=2

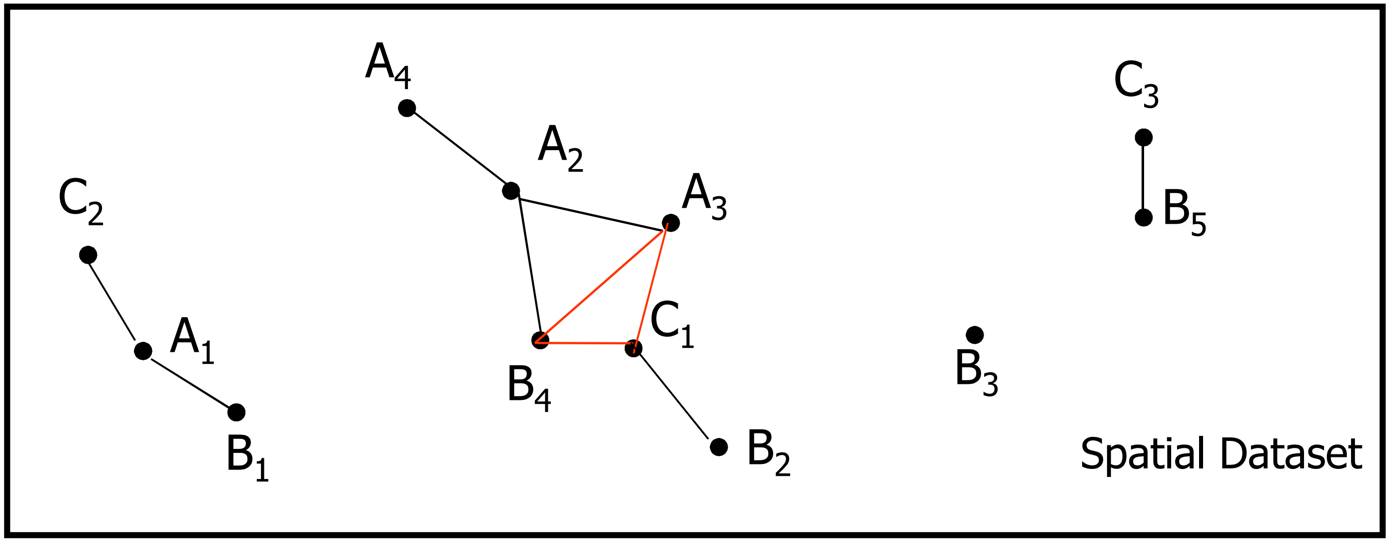
A	B
1	1
2	4
3	4
3/4	2/5

A	C
1	2
3	1
2/4	2/3

B	C
2	1
4	1
5	3
3/5	2/3

Participation ratio

Co-location Mining



A-School
B-Hospital
C-Pharmacy

Candidates of size k=2

A	B
1	1
2	4
3	4

2/5

A	C
1	2
3	3

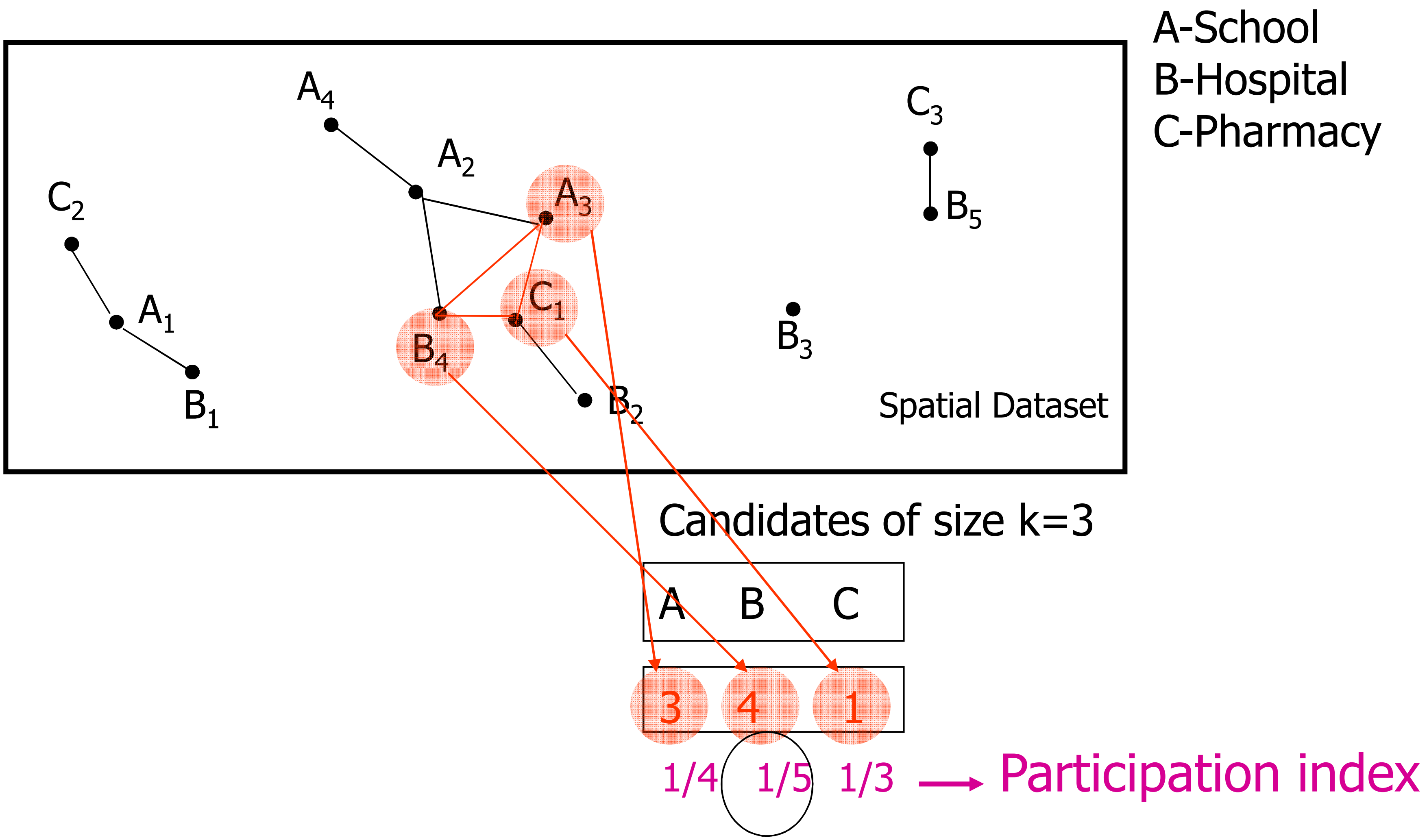
2/4

B	C
2	1
4	1
5	3

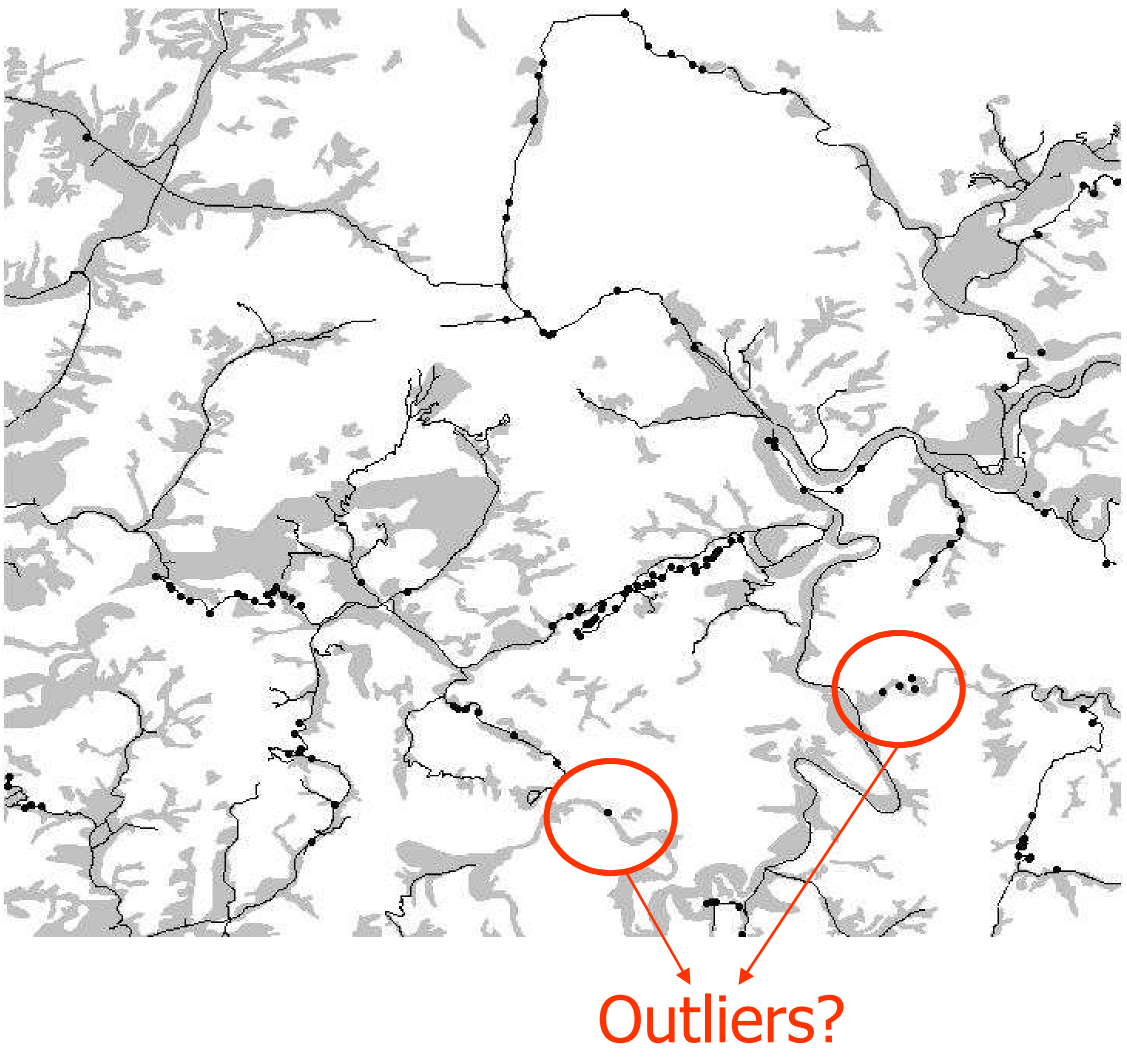
3/5

Participation Index (Lowest index)
(If participIndex>minPartIndex)
→ frequent set

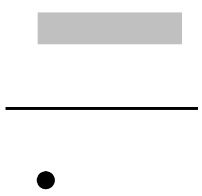
Co-location Mining



Co-location Example *(Shekhar 2003)*



Cropland with Roads
Roads with Bridges



Outliers

- What is an outlier?
  - Observations inconsistent with the rest of the dataset
- What is a spatial outlier?
  - Observations inconsistent with their neighborhoods
  - A local instability or discontinuity

Outliers (Shekhar 2001, 2003)

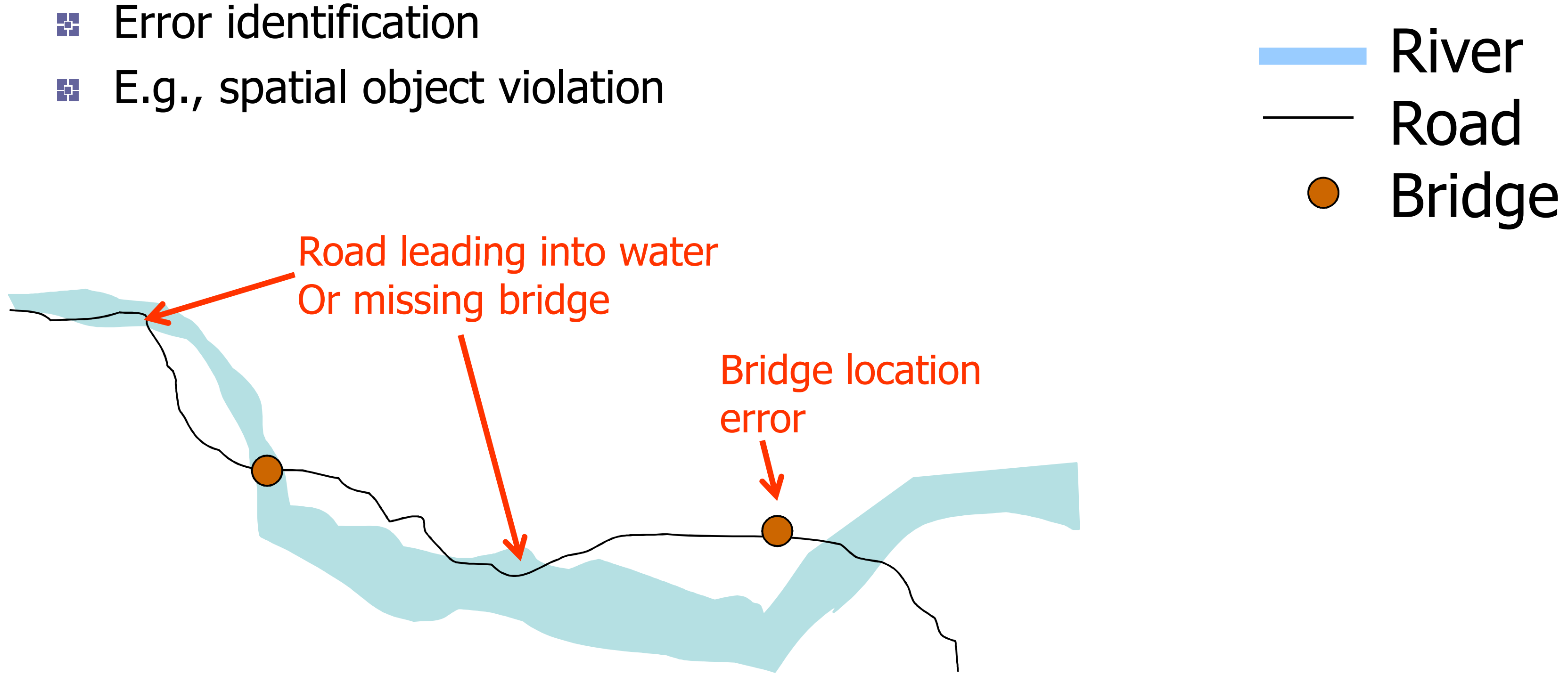
- **Global outliers** are data inconsistent with the rest of the data in the database
  - Applications:
    - *credit card fraud,*
    - *athlete performance analysis,*
    - *voting irregularity,*
    - *severe weather prediction*

Outliers (Shekhar 2001, 2003)

- A **spatial outlier** is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood.
  - For example, a new house in an old neighborhood is a spatial outlier based on the non-spatial attribute house age
  - **Spatial attributes** are used to characterize location, neighborhood, and distance.
  - **Non-spatial attributes** are used to compare a spatial referenced object to its neighbors.

Outliers – Examples *(Shekhar 2003)*

- Map Production
  - Error identification
  - E.g., spatial object violation



Tools

- GeoMiner (Han 1997)
- INGENS (Malerba 2001)
- Ares (Appice 2005)
- Weka-GDPM (Bogorny 2006d)

Conclusions

- Patterns are opposite of random
- Common spatial patterns: location prediction, feature interaction, hot spots,
- SDM = search for unexpected interesting patterns in large spatial databases
- Spatial patterns may be discovered using
  - Techniques like classification, associations, clustering and outlier detection
  - New techniques are needed for SDM due to
    - Spatial Auto-correlation
    - Continuity of space