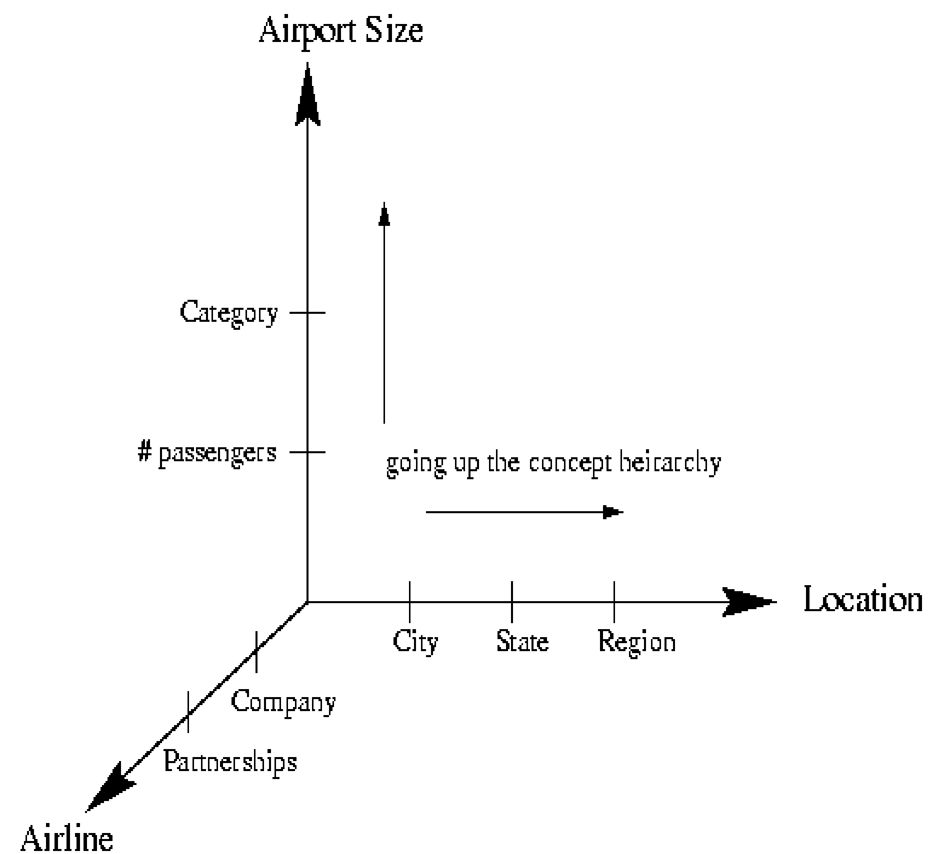
Multidimensional Analysis

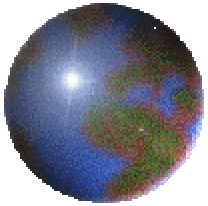
- Strategy
  - Generalize the planbase in different directions
  - Look for sequential patterns in the generalized plans
  - Derive high-level plans

A multi-D model for the planbase



Generalization-Based Sequence Mining

- Generalize planbase in multidimensional way using dimension tables
- Use # of distinct values (cardinality) at each level to determine the right level of generalization (level-"planning")
- Use operators *merge* "+", *option* "[]" to further generalize patterns
- Retain patterns with significant support



Spatial Data Mining

Spatial data mining is the process of discovering interesting, useful, non-trivial patterns from large spatial datasets

Spatial Data Warehousing

- Spatial data warehouse: Integrated, subject-oriented, time-variant, and nonvolatile spatial data repository for data analysis and decision making
- Spatial data integration: a big issue
  - Structure-specific formats (raster- vs. vector-based, OO vs. relational models, different storage and indexing, etc.)
  - Vendor-specific formats (ESRI, MapInfo, Integraph, etc.)
- Spatial data cube: multidimensional spatial database
  - Both dimensions and measures may contain spatial components

Dimensions and Measures in Spatial Data Warehouse

- Dimension modeling
  - nonspatial
    - e.g. temperature: 25-30 degrees generalizes to *hot*
  - spatial-to-nonspatial
    - e.g. region "B.C." generalizes to description "*western provinces"*
  - spatial-to-spatial
    - e.g. region "Burnaby" generalizes to region "Lower Mainland"
- Measures
  - numerical
    - distributive (e.g. count, sum)
    - algebraic (e.g. average)
    - holistic (e.g. median, rank)
  - spatial
    - collection of spatial pointers (e.g. pointers to all regions with 25-30 degrees in July)

Spatial Association Analysis

- Spatial association rule: $A \Rightarrow B\,[s\%, c\%]$
  - A and B are sets of spatial or nonspatial 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

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

is_a(x, large_town) ^adjacent_to(x, georgia_strait) → close_to(x, u.s.a.)
[1%, 78%]

Spatial Classification and Spatial Trend Analysis

- Spatial classification
  - Analyze spatial objects to derive classification schemes, such as decision trees in relevance to certain spatial properties (district, highway, river, etc.)
  - Example: Classify regions in a province into *rich* vs. *poor* according to the average family income
- Spatial trend analysis
  - Detect changes and trends along a spatial dimension
  - Study the trend of nonspatial or spatial data changing with space
  - Example: Observe the trend of changes of the climate or vegetation with the increasing distance from an ocean

Generalizing Spatial and Multimedia Data

- Spatial data:
  - Generalize detailed geographic points into clustered regions, such as business, residential, industrial, or agricultural areas, according to land usage
  - Require the merge of a set of geographic areas by spatial operations
- Image data:
  - Extracted by aggregation and/or approximation
  - Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- Music data:
  - Summarize its melody: based on the approximate patterns that repeatedly occur in the segment
  - Summarized its style: based on its tone, tempo, or the major musical instruments played

Examples of Spatial Patterns

- Historic Examples
  - 1855 Asiatic Cholera in London: A water pump identified as the source
  - Fluoride and healthy gums near Colorado river
  - Theory of Gondwanaland - continents fit like pieces of a jigsaw puzlle
- Modern Examples
  - Cancer clusters to investigate environment health hazards
  - Crime hotspots for planning police patrol routes
  - Bald eagles nest on tall trees near open water
  - Nile virus spreading from north east USA to south and west
  - Unusual warming of Pacific ocean (El Nino) affects weather in USA

Introduction: a classic example for spatial analysis

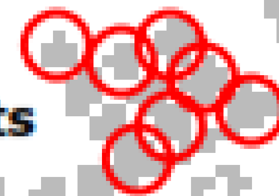
Disease cluster

Dr. John Snow
Deaths of cholera epidemia
London, September 1854

Infected water pump?

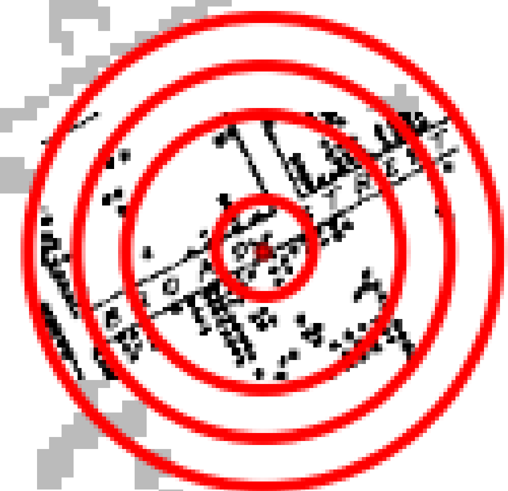
- A good representation is
- the key to solving a problem

Good representation because...



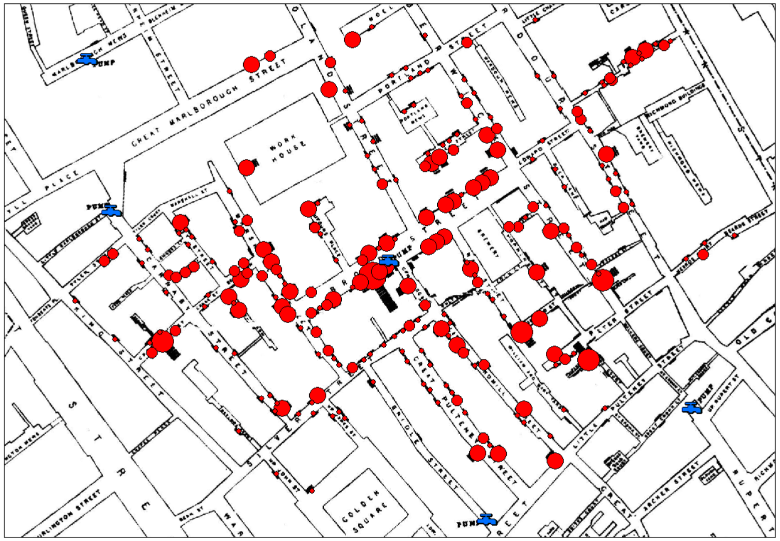
- Represents spatial relation of objects
- of the same type
-

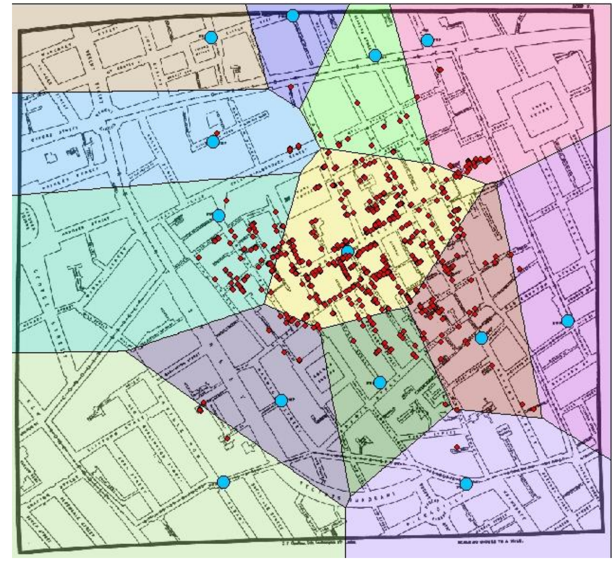
Represents spatial relation of objects to *other* objects

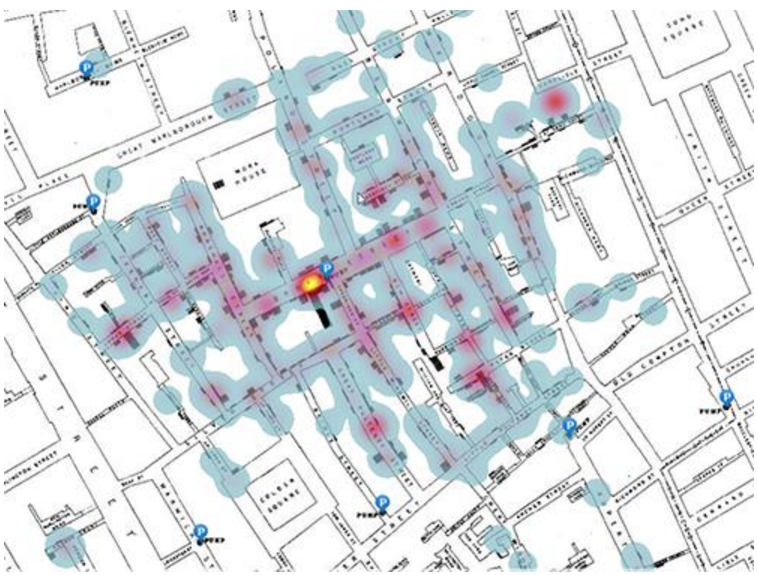


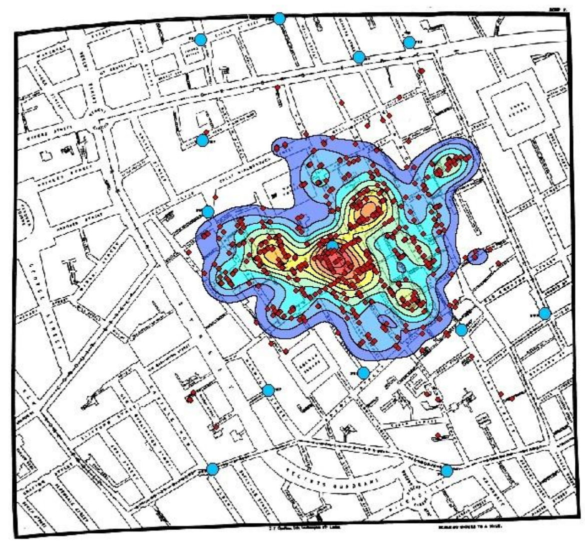
Shows only relevant aspects and hides irrelevant

It is not only important where a cluster is but also, what else is there (e.g. a water-pump)!

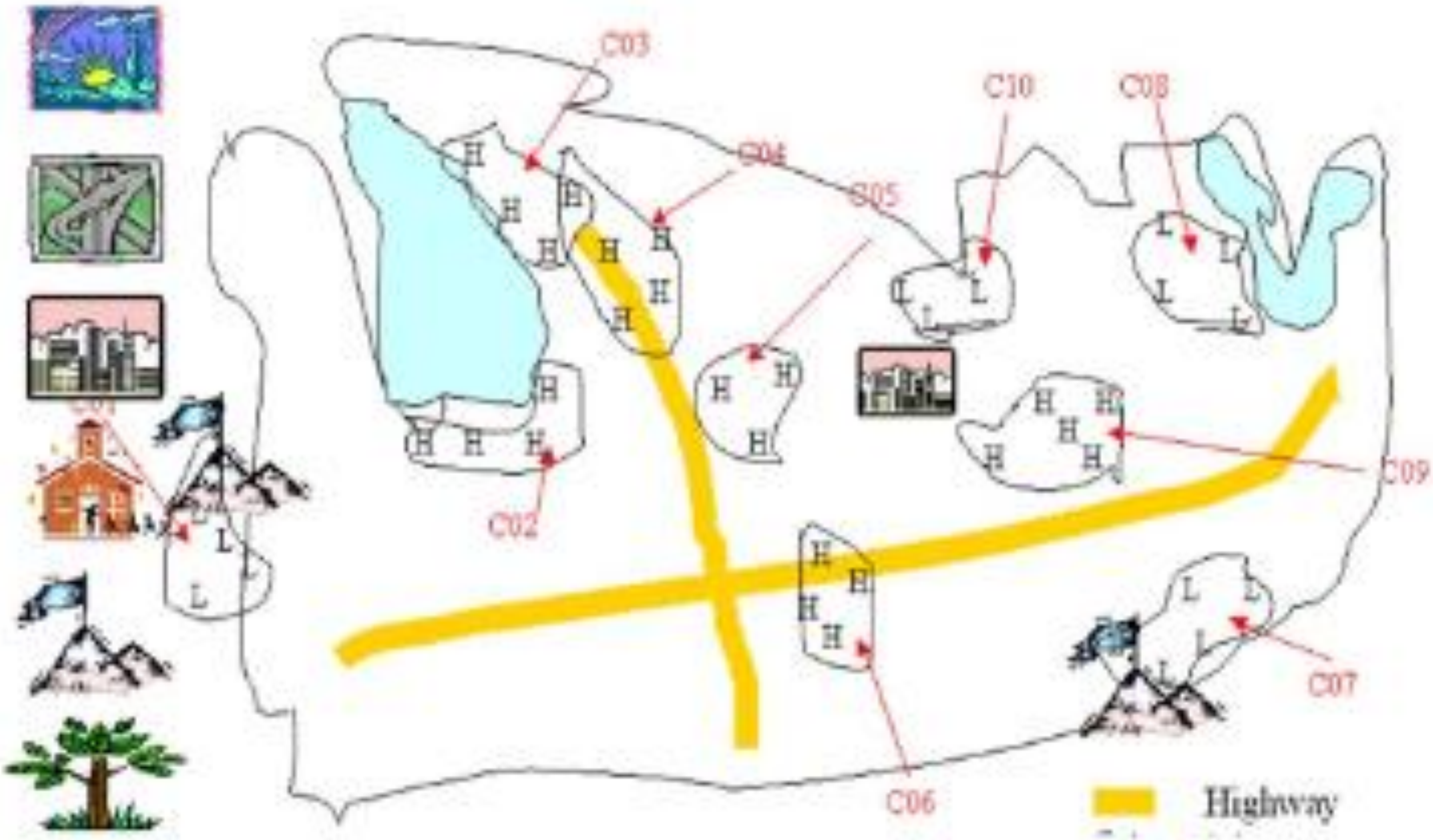

GREAT MARLBOROUGH STREET
WORK HOUSE
GOLDEN SQUARE
PUMP



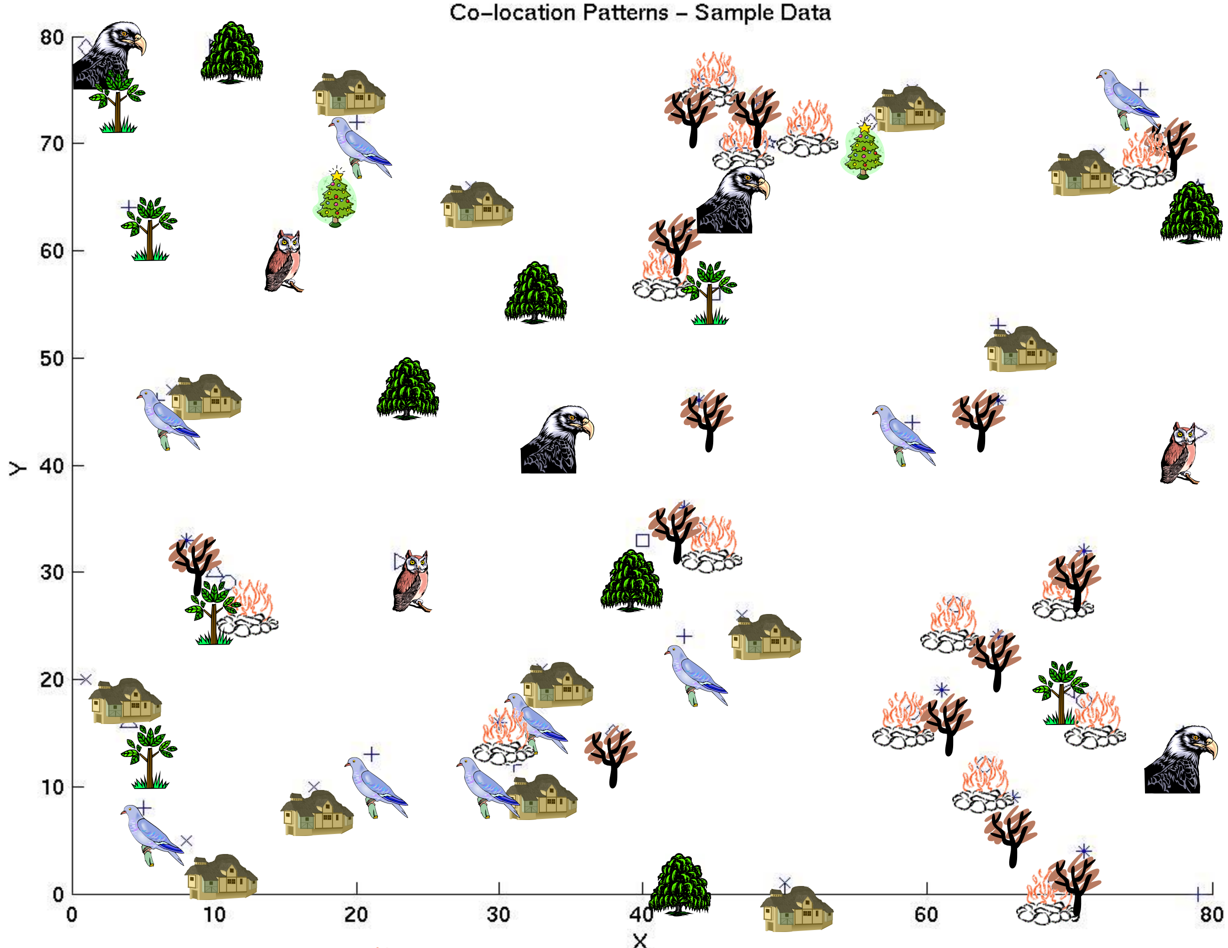

GREAT MARLBOROUGH STREET
WORK HOUSE
GOLDEN SQUARE



- What Kind of Houses Are Highly Valued?—Associative Classification



Associations, Spatial associations, Co-location



Answers: and

find patterns from the following sample dataset?

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? - 2

- 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

Characteristics of Spatial Data Mining

- Auto correlation
- Patterns usually have to be defined in the spatial attribute subspace and not in the complete attribute space
- Longitude and latitude (or other coordinate systems) are the glue that link different data collections together
- People are used to maps in GIS; therefore, data mining results have to be summarized on the top of maps
- Patterns not only refer to points, but can also refer to lines, or polygons or other higher order geometrical objects
- Large, continuous space defined by spatial attributes
- Regional knowledge is of particular importance due to lack of global knowledge in geography (→spatial heterogeniety)

Why Regional Knowledge Important in Spatial Data Mining?

- A special challenge in spatial data mining is that information is usually not uniformly distributed in spatial datasets.
- It has been pointed out in the literature that "*whole map statistics are seldom useful*", that "*most relationships in spatial data sets are geographically regional, rather than global*", and that "*there is no average place on the Earth's surface*" [Goodchild03, Openshaw99].
- Therefore, it is not surprising that domain experts are mostly interested in discovering hidden patterns at a regional scale rather than a global scale.

Spatial Association Rules

- Spatial Association Rules
  - A special reference spatial feature
  - Transactions are defined around instance of special spatial feature
  - Item-types = spatial predicates
  - Example: Table 7.5 (pp. 204)

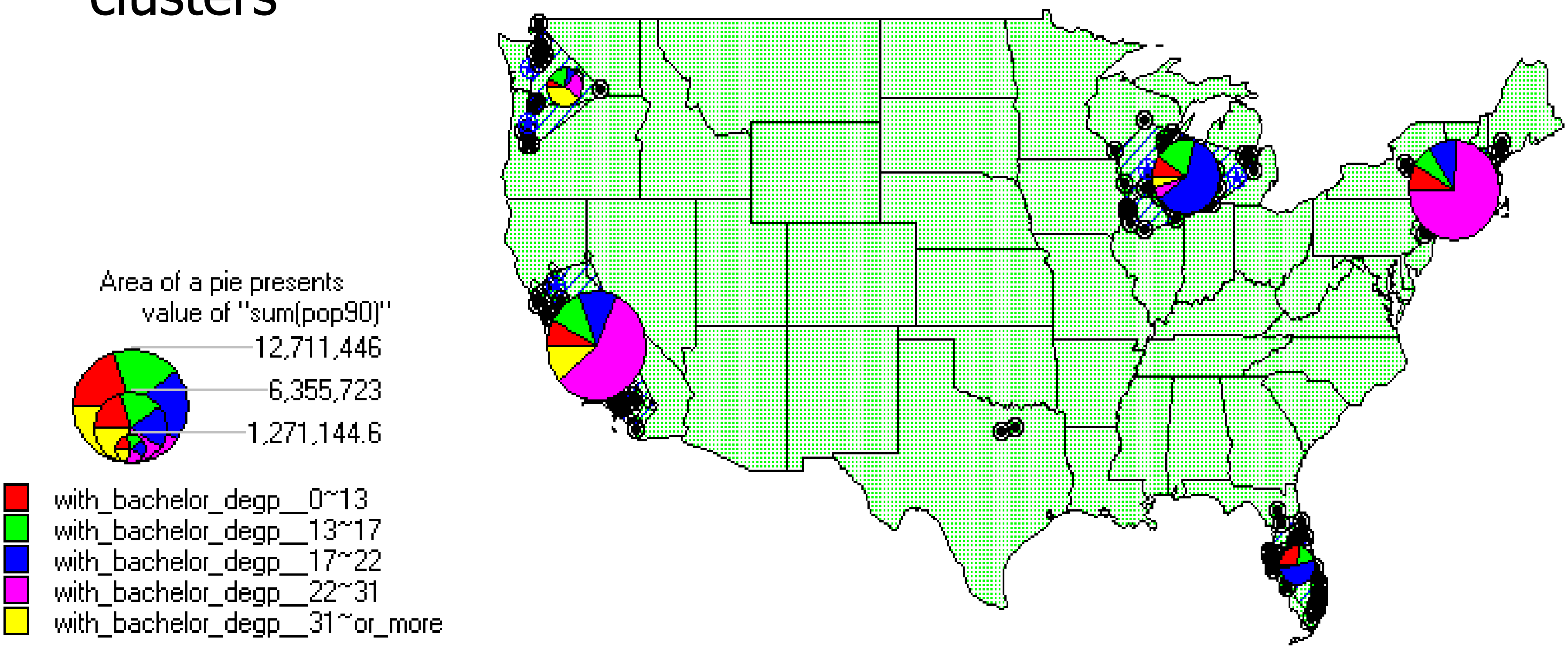
Spatial Association Rule	Sup.	Conf.
$Stem\_height(x, high) \wedge Distance\_to\_edge(x, far) \rightarrow Vegetation\_Durability(x, moderate)$	0.1	0.94
$Vegetation\_Durability(x, moderate) \wedge Distance\_to\_water(x, close) \rightarrow Stem\_Height(x, high)$	0.05	0.95
$Distance\_towater(x, far) \wedge Water\_Depth(x, shallow) \rightarrow Stem\_Height(x, high)$	0.05	0.94

Spatial Trend Analysis

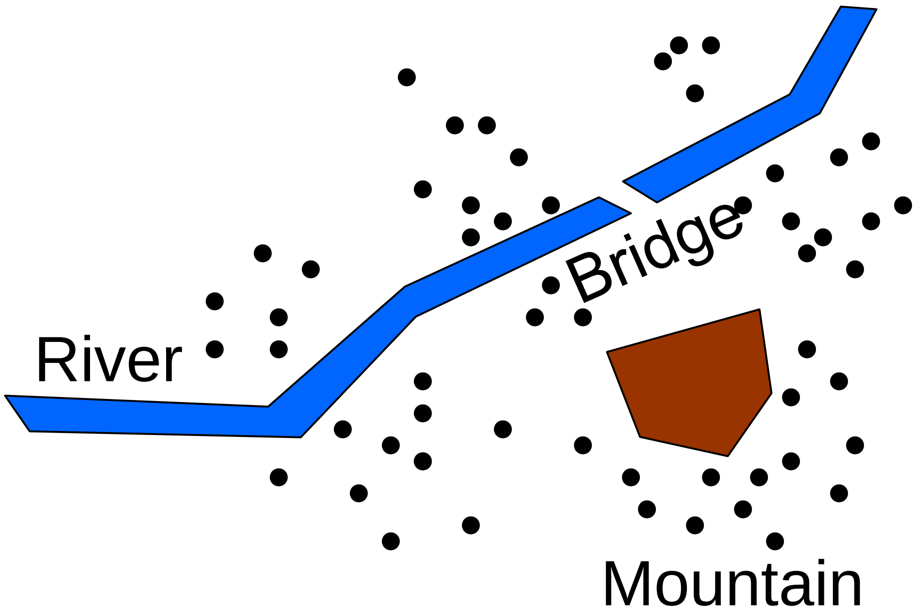
- Function
  - Detect changes and trends along a spatial dimension
  - Study the trend of non-spatial or spatial data changing with space
- Application examples
  - Observe the trend of changes of the climate or vegetation with increasing distance from an ocean
  - Crime rate or unemployment rate change with regard to city geo-distribution

Spatial Cluster Analysis

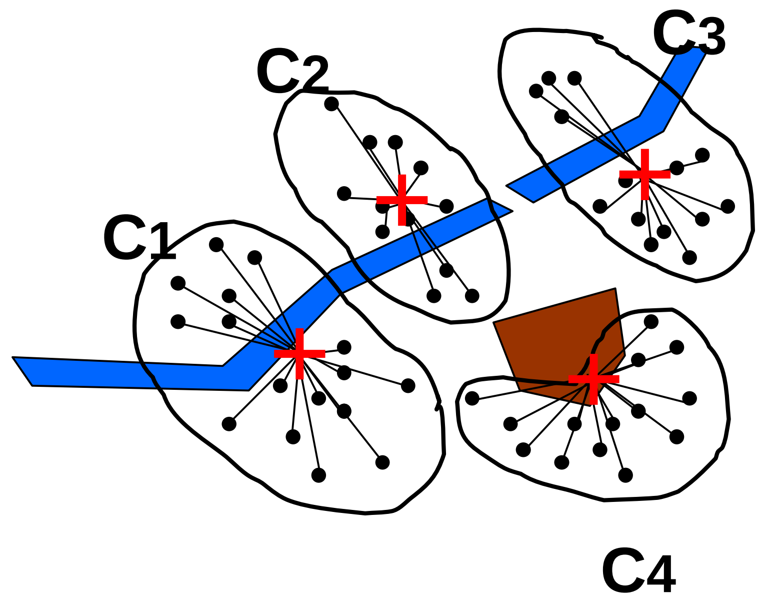
- Mining clusters—k-means, k-medoids, hierarchical, density-based, etc.
- Analysis of distinct features of the clusters



Constraint-Based Clustering: Planning ATM Locations



Spatial data with obstacles



Clustering *without* taking obstacles into consideration

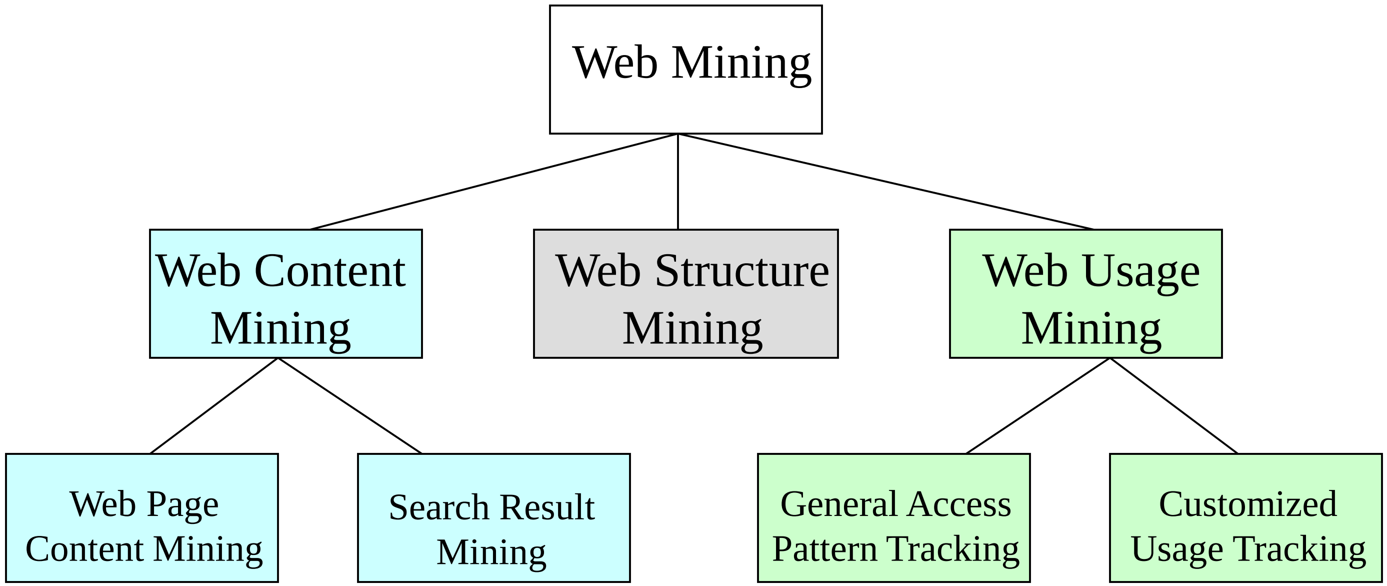
Conclusions Spatial Data Mining

- Spatial patterns are opposite of random
- Common spatial patterns: location prediction, feature interaction, hot spots, geographically referenced statistical patterns, co-location, emergent patterns,…
- 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
    - Importance of non-point data types (e.g. polygons)
    - Continuity of space
    - Regional knowledge; also establishes a need for scoping
    - Separation between spatial and non-spatial subspace—in traditional approaches clusters are usually defined over the complete attribute space
- Knowledge sources are available now
  - Raw knowledge to perform spatial data mining is mostly available online now (e.g. relational databases, Google Earth)
  - GIS tools are available that facilitate integrating knowledge from different source

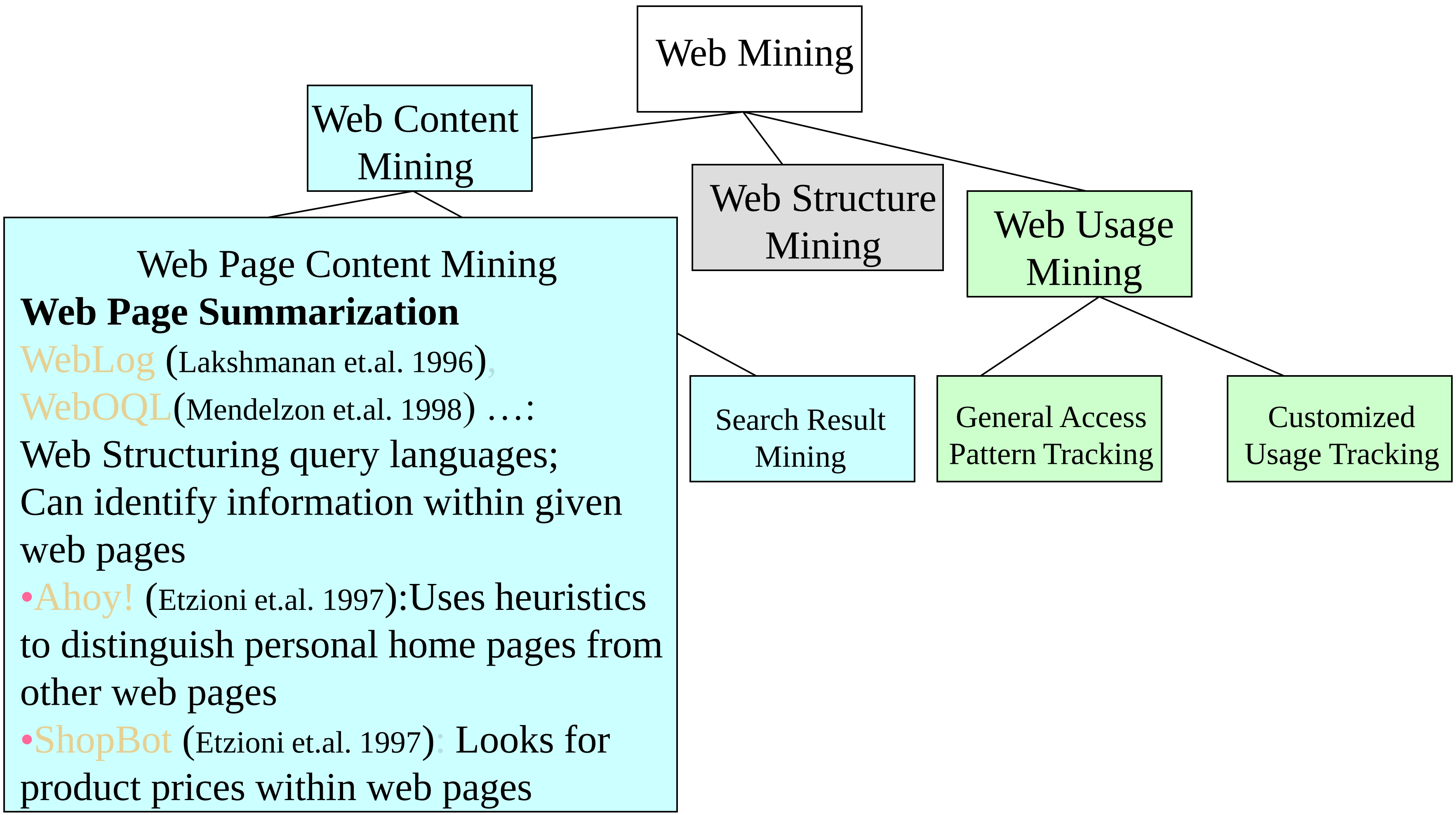
Mining the World-Wide Web

- The WWW is huge, widely distributed, global information service center for
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- WWW provides rich sources for data mining
- Challenges
  - Too huge for effective data warehousing and data mining
  - Too complex and heterogeneous: no standards and structure

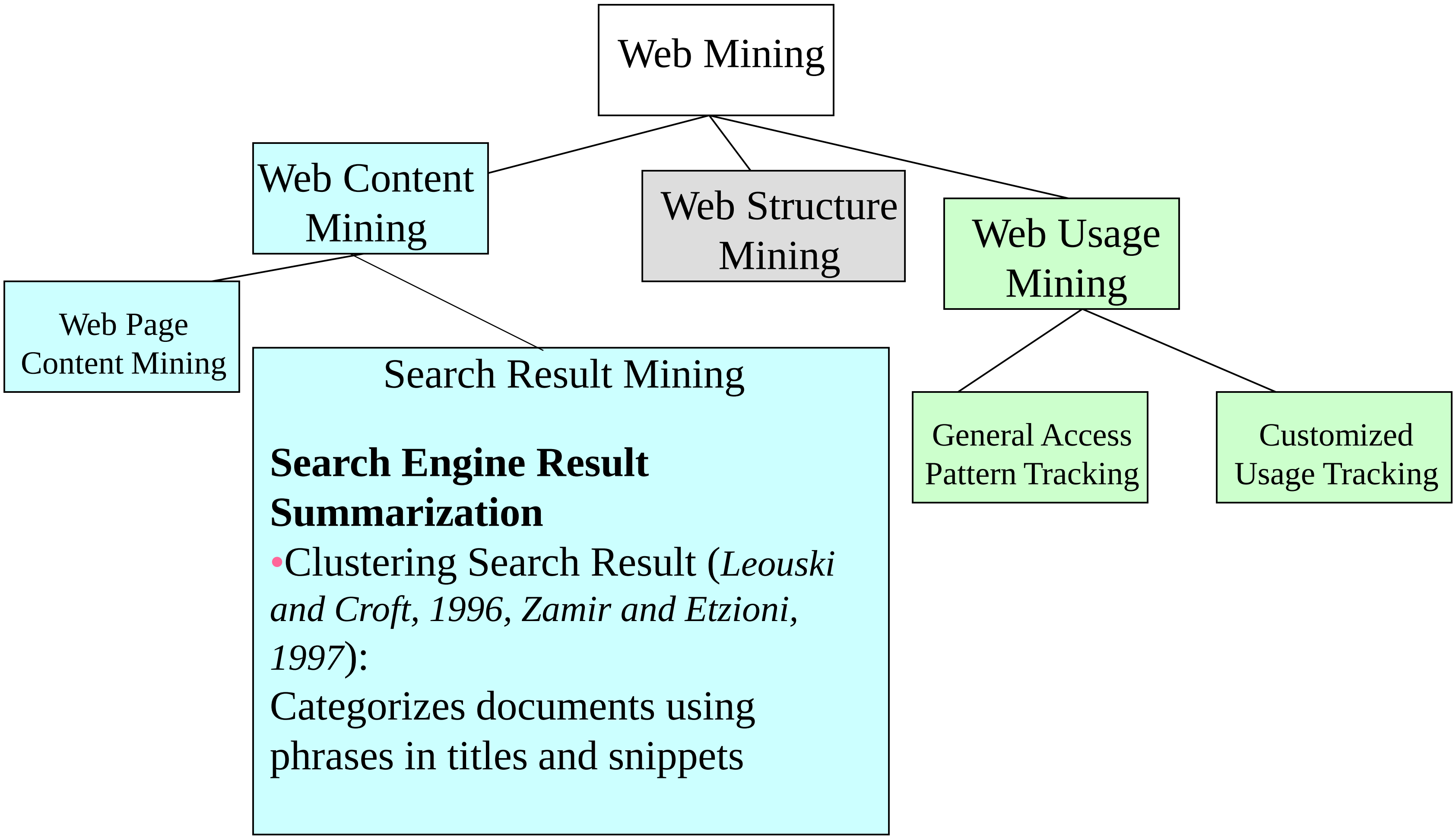
Web Mining Taxonomy



Mining the World-Wide Web



Mining the World-Wide Web



Mining the World-Wide Web

Web Mining

Web Content Mining

- Web Page Content Mining
- Search Result Mining

Web Structure Mining

Using Links

- PageRank (Brin et al., 1998)
- CLEVER (Chakrabarti et al., 1998)

Use interconnections between web pages to give weight to pages.

Using Generalization

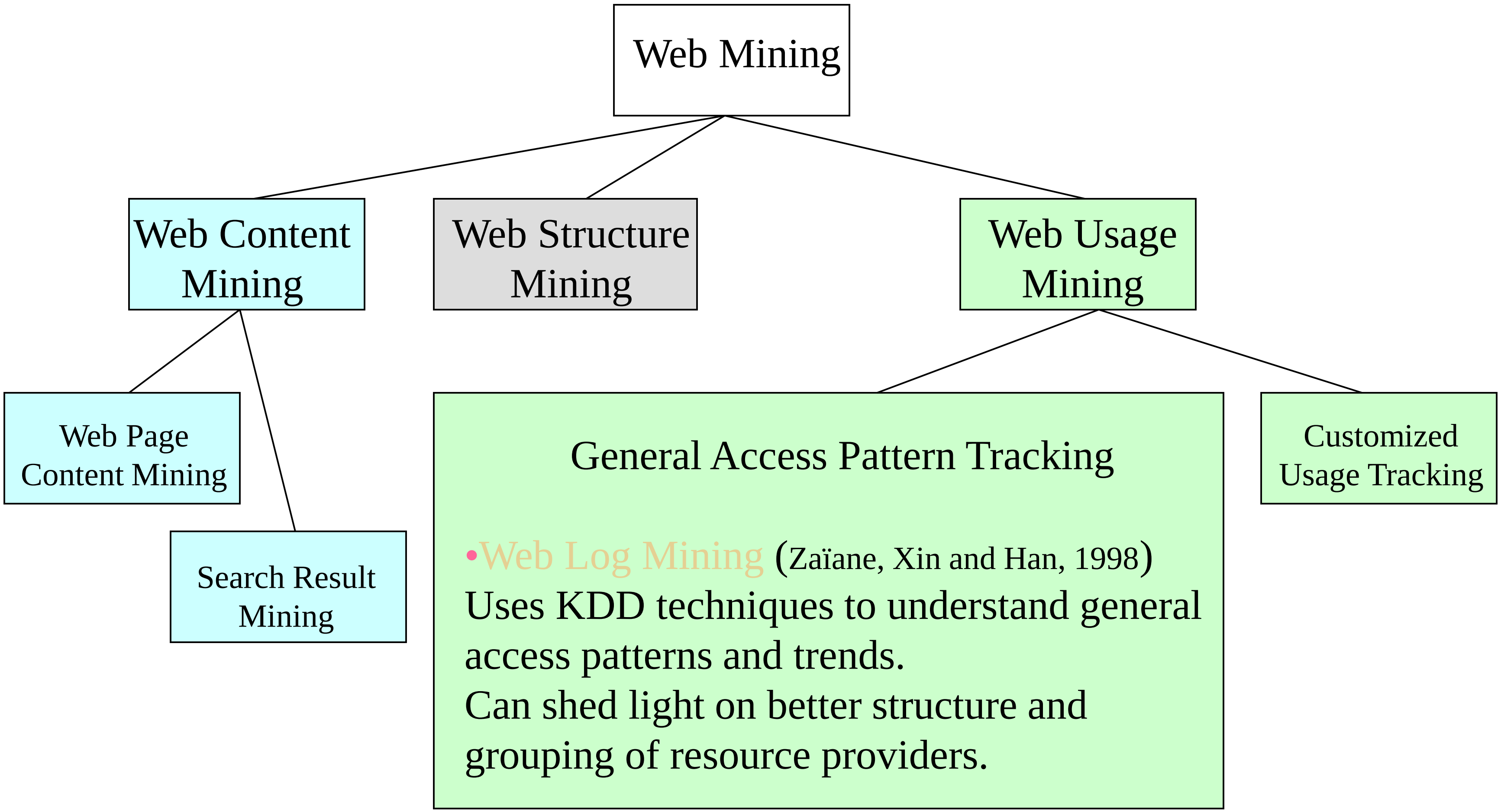
- MLDB (1994), VWV (1998)

Uses a multi-level database representation of the Web. Counters (popularity) and link lists are used for capturing structure.

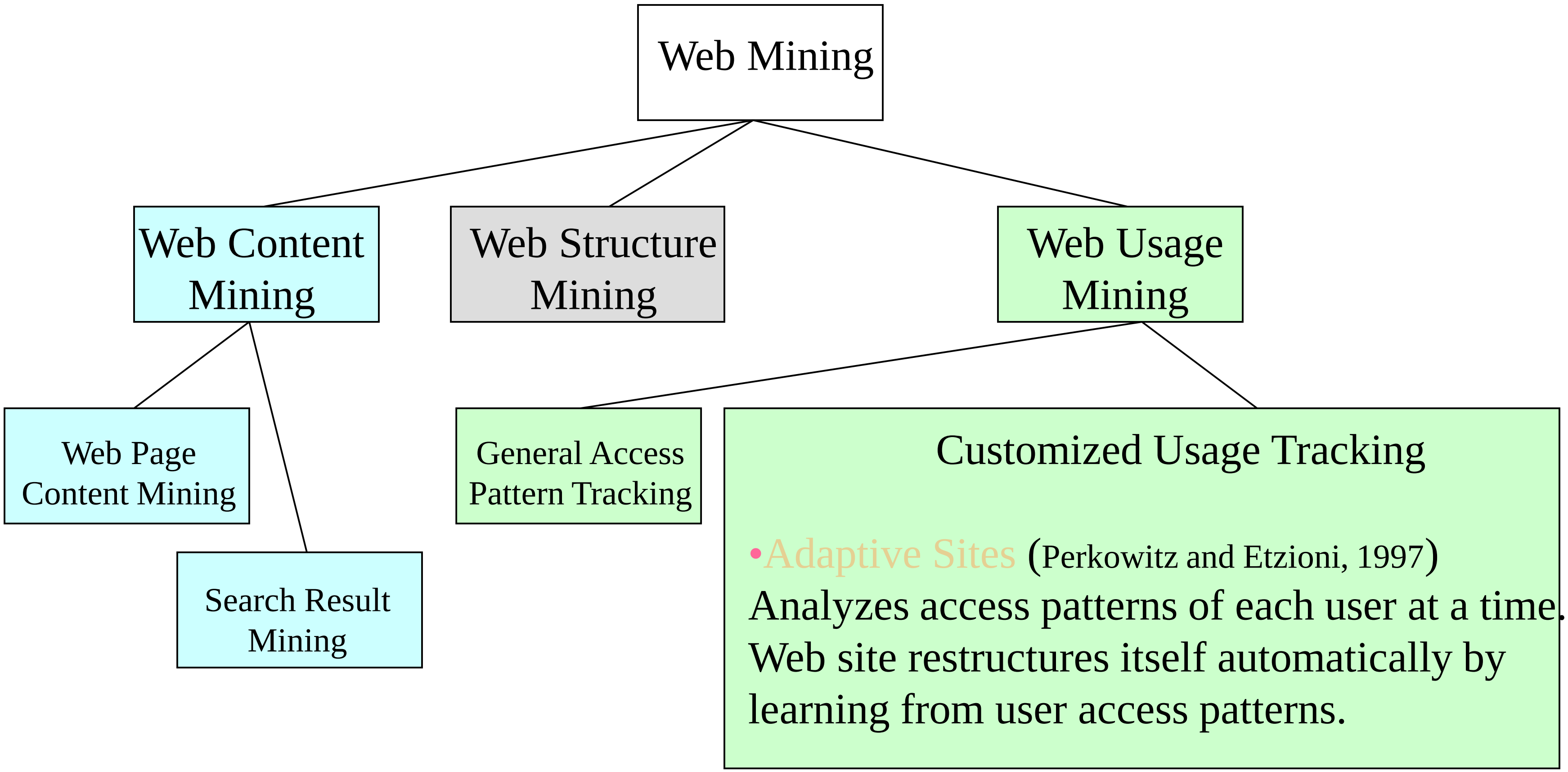
Web Usage Mining

- General Access Pattern Tracking
- Customized Usage Tracking

Mining the World-Wide Web



Mining the World-Wide Web



Mining the Web's Link Structures

- Finding authoritative Web pages
  - Retrieving pages that are not only relevant, but also of high quality, or authoritative on the topic
- Hyperlinks can infer the notion of authority
  - The Web consists not only of pages, but also of hyperlinks pointing from one page to another
  - These hyperlinks contain an enormous amount of latent human annotation
  - A hyperlink pointing to another Web page, this can be considered as the author's endorsement of the other page

Mining the Web's Link Structures

- Problems with the Web linkage structure
  - Not every hyperlink represents an endorsement
    - Other purposes are for navigation or for paid advertisements
    - If the majority of hyperlinks are for endorsement, the collective opinion will still dominate
  - One authority will seldom have its Web page point to its rival authorities in the same field
  - Authoritative pages are seldom particularly descriptive
- Hub
  - Set of Web pages that provides collections of links to authorities

Similarity Search in Multimedia Data

- Description-based retrieval systems
  - Build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation
  - Labor-intensive if performed manually
  - Results are typically of poor quality if automated
- Content-based retrieval systems
  - Support retrieval based on the image content, such as color histogram, texture, shape, objects, and wavelet transforms

Queries in Content-Based Retrieval Systems

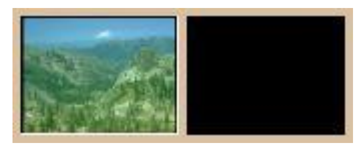
- Image sample-based queries
  - Find all of the images that are similar to the given image sample
  - Compare the feature vector (signature) extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database
- Image feature specification queries
  - Specify or sketch image features like color, texture, or shape, which are translated into a feature vector
  - Match the feature vector with the feature vectors of the images in the database

Mining Multimedia Databases

Refining or combining searches

Search for "airplane in blue sky"
(top layout grid is blue and
keyword = "airplane")

Search for "blue sky"
(top layout grid is blue)



Search for "blue sky and green meadows"
(top layout grid is blue
and bottom is green)

Mining Time-Series and Sequence Data

- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, irregular
- Applications
  - Financial: stock price, inflation
  - Biomedical: blood pressure
  - Meteorological: precipitation

Mining Time-Series and Sequence Data: Trend analysis

- A time series can be illustrated as a time-series graph which describes a point moving with the passage of time
- Categories of Time-Series Movements
  - Long-term or trend movements (trend curve)
  - Cyclic movements or cycle variations, e.g., business cycles
  - Seasonal movements or seasonal variations
    - i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years.
  - Irregular or random movements

Estimation of Trend Curve

- The freehand method
  - Fit the curve by looking at the graph
  - Costly and barely reliable for large-scaled data mining
- The least-square method
  - Find the curve minimizing the sum of the squares of the deviation of points on the curve from the corresponding data points
- The moving-average method
  - Eliminate cyclic, seasonal and irregular patterns
  - Loss of end data
  - Sensitive to outliers

Discovery of Trend in Time-Series (2)

- Estimation of cyclic variations
  - If (approximate) periodicity of cycles occurs, cyclic index can be constructed in much the same manner as seasonal indexes
- Estimation of irregular variations
  - By adjusting the data for trend, seasonal and cyclic variations
- With the systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make long- or short-term predictions with reasonable quality