
Spatial Data Mining in the Era of Big Data

Dr. Jin Soung Yoo

Associate Professor

Department of Computer Science

Indiana University-Purdue University Fort Wayne

Outline

- Introduction to Data Mining
- Works in Data Mining

What is Data Mining

- A computer-assisted process of discovering interesting, previously unknown, implicit, potentially useful, and non-trivial patterns or knowledge from large databases
 - Non-trivial search
 - Large (e.g., exponential) search space of plausible hypothesis
 - Interesting
 - Useful in certain application domain
 - Unexpected
 - Patterns is not common knowledge
 - May provide a new understanding of world
-

Why Data Mining - Commercial Viewpoint

- With rapid advances in data collection and storage technology, the **explosive growth of data** from
- many data sources
 - Purchases at grocery stores, customer services from call centers
 - Web logs from e-commerce Web sties
 - Bank/Credit card transactions
 - Mobile phone contents
 - Social networks
 - World Wide Web: online news, digital images, YouTube



Source: various web sites

Why Data Mining (Conti.)

- Competitive pressure
 - Today business environment requires critical data analysis
 - Market analysis: targeted marketing, cross-selling, market segments, etc.
 - Risk management: forecasting, customer retention
 - Fraud detection and detection of unusual behavior
 - Business questions
 - “Who are the most profitable customers?”
 - “What products can be cross-sold?”
 - “How change if a new local store is added?”
-

Example: Target's Finding



Target has figured out whether you have a baby on the way long before you need to start buying diapers

“Women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester.”

“When someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.”

Scale of Data

Organization	Scale of Data
Walmart	~ 1 million customer transactions / hr
Facebook	~ 50 billion photos
Yahoo	~48 GB Web log data/hr
Falcon Credit Card Fraud Detection System (FICO)	2.1 billion active accounts world-wide
Business data worldwide, across all companies	Doubles every 1.2 years
NASA satellites	~ 1.2 TB/day
Sloan Digital Sky Survey (SDSS)	~140 TB (200GB /night)
NCBI GenBank	~ 22 million genetic sequences

Source: http://en.wikipedia.org/wiki/Big_data

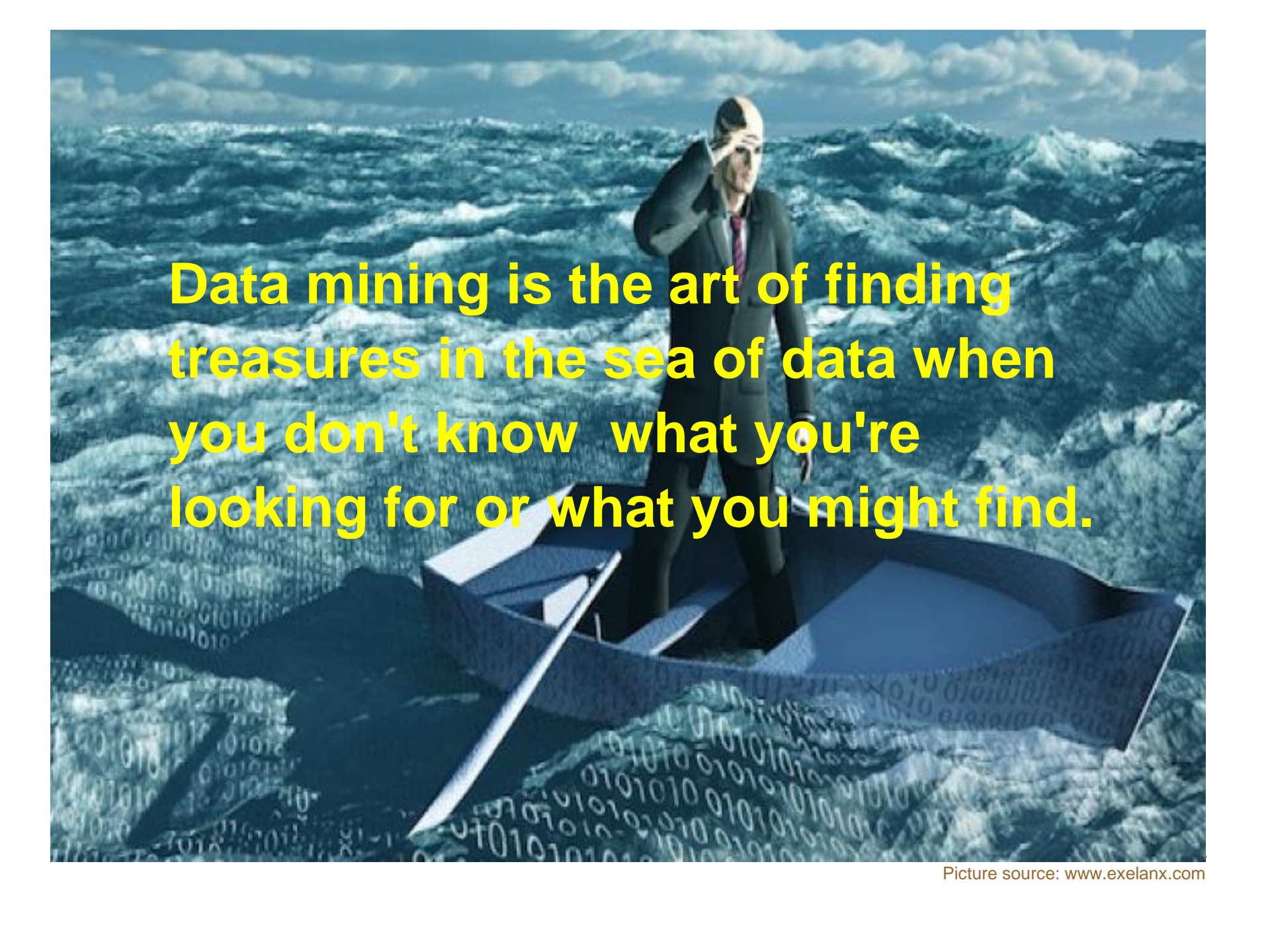
Big Data but No Clue

“The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don’t have a clue as to what any of that data actually means” (Cass, IEEE Spectrum, Jan 2004)

- There is often information “hidden” in the data that is not readily evident
- Much of the data is never analyzed at all.

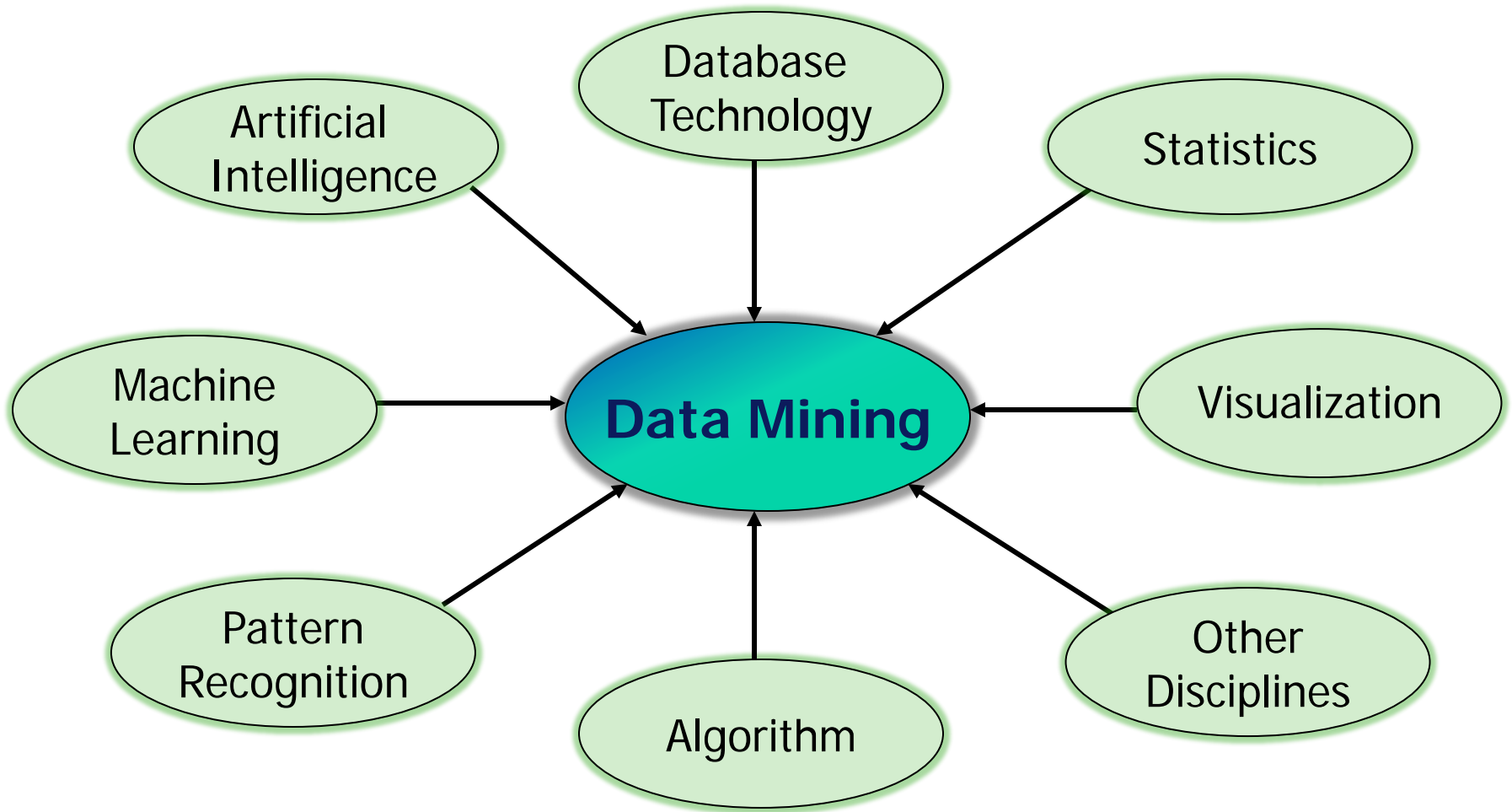


source: <http://shawnwhatley.com/>

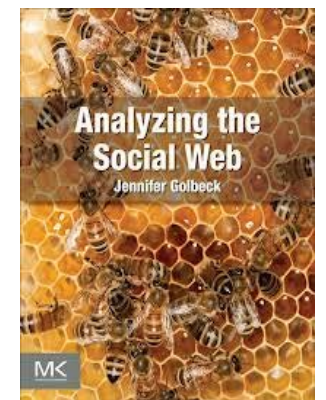
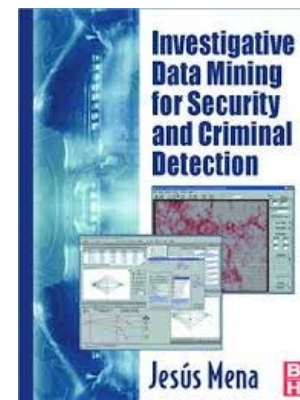
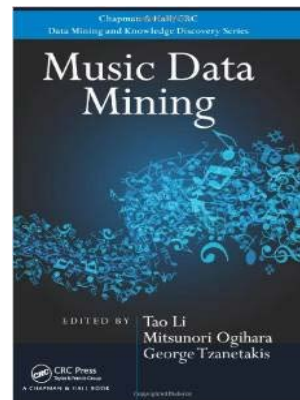
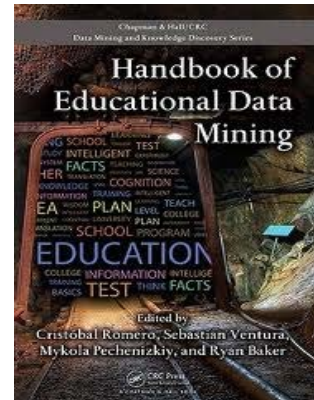
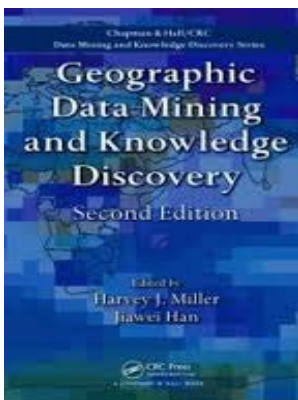
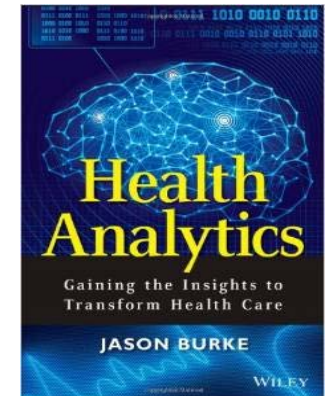
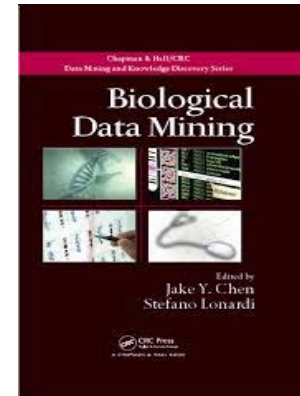
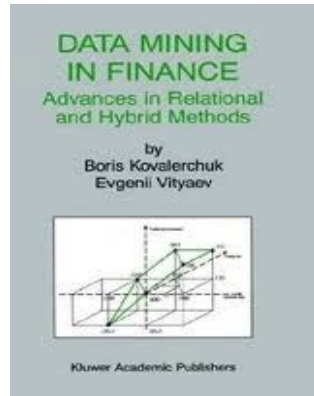
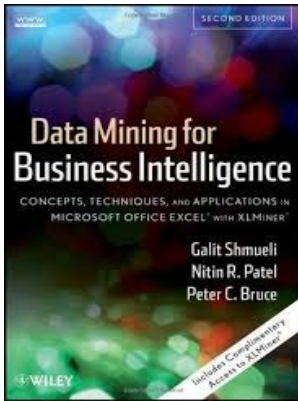
A man in a dark suit and tie stands in a small, dark boat. The boat is floating on a vast, turbulent sea of blue and white waves. The waves are composed of binary code (0s and 1s) and other digital symbols, representing a sea of data. The man is looking forward with his hand to his forehead, suggesting a search or discovery. The sky is blue with white clouds.

Data mining is the art of finding treasures in the sea of data when you don't know what you're looking for or what you might find.

Confluence of Multiple Disciplines



Data Mining for Many Other Disciplines



Picture source: amazon.com

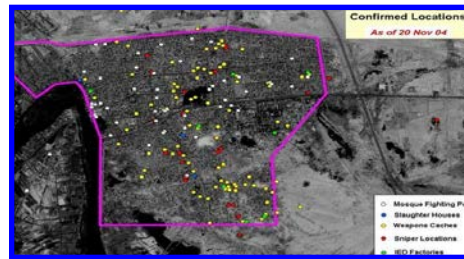
Outline

- Introduction to Data Mining
 - Works in Data Mining
 - Spatial Data Mining
 - Spatial Association Mining in Cloud Computing Environment
 - Temporal Data Mining
 - Spatiotemporal Data Mining
 - Biological Data Mining
 - Educational Data Mining
-

Spatial Data – Everywhere

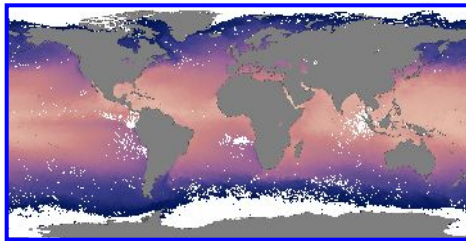


Criminology

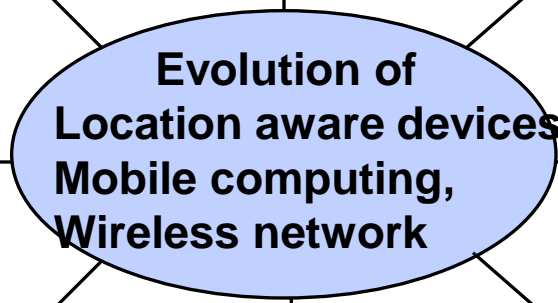


Military, Homeland security

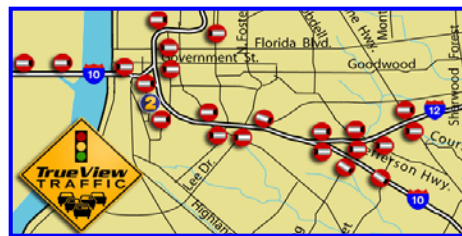
Business (Location-Based Services)



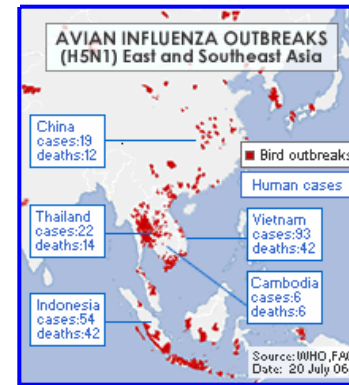
Earth Science



Environmental Science



Transportation

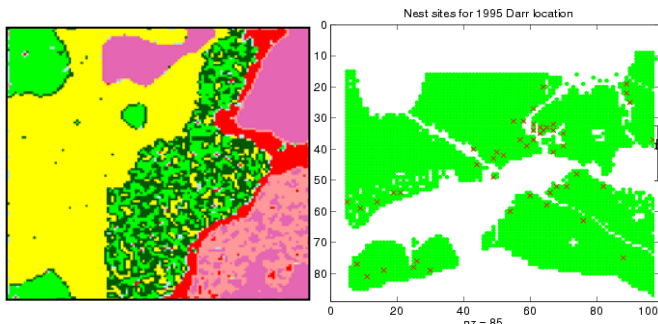


Public Health

Spatial Data Mining

- The process of discovering interesting, useful, non-trivial (as “automatized” as possible) patterns from large spatial or spatiotemporal data.
- Spatial Pattern Families vs. Techniques

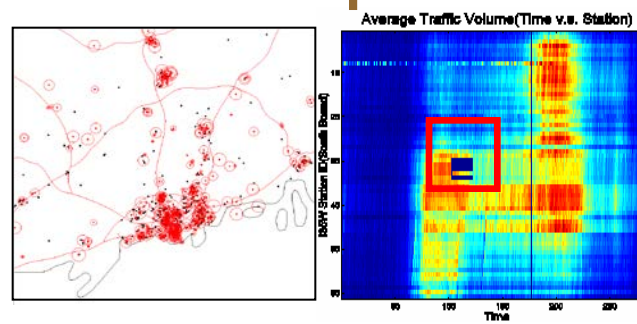
Prediction



Classification

Prediction

Hot spots



Clustering

Outlier detection

Interaction



Association, Colocation

Examples of Spatial Patterns

- **Spatial relationships** (location, region, frontier, neighborhood, obstruction, field, basin, communication, diffusion, propagation) are importantly considered for the pattern discovery

- **Historic Examples**

- 1855 Asiatic Cholera in London; A water pump identified as the source
- Fluoride and health gums near Colorado river (with originally insufficient amount of fluoride)



Examples of Spatial Patterns

■ Modern Examples

- Cancer clusters to locate hazardous environments
 - Nile virus spreading from north east USA to south and west
 - Crime hotspots for planning police patrol routes
 - Colocation of a business with another franchise (such as colocation of a Pizza Hut restaurant with a Blockbuster video store)
 - Best locations for opening new hospitals based on the population of patients who live in each neighborhood.
 - Spatial region-based personalization
 - Unusual warming of Pacific ocean (El Niño) effects weather in USA
-

Outline

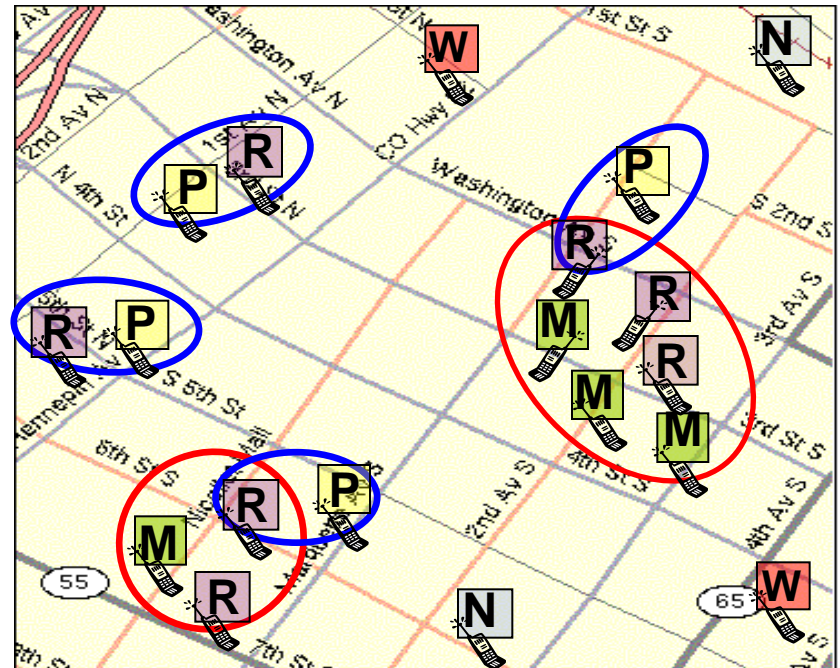
- Introduction to Data Mining
 - Works in Data Mining
 - Spatial Data Mining
 - Spatial Association Mining
 - Spatial Association Mining in Cloud Computing Environment
 - Summary
-

Spatial Association, Co-location, Correlation

- **Spatial association mining** discovers interesting spatial relationships and correlations among spatial objects.
 - **Spatial correlation** (or, *neighborhood influence*) refers to the phenomenon of the location of a specific object in an area affecting some nonspatial attribute of the object.
 - For example, the value (nonspatial attribute) of a house at a given address (geocoded to give a spatial attribute) is largely determined by the value of other houses in the neighborhood.
-

Spatial Co-location

- A **co-location** represents the presence of two or more spatial objects at the same location or at significantly close distances from each other.
- **Co-location mining** finds all subsets of spatial events (/ features) which are frequently observed in nearby areas.
- In the case of including non-spatial information
 - For example, sales at franchises of a specific pizza restaurant chain were higher at restaurants collocated with video stores than at restaurants not collocated with video stores.



Find patterns from the above sample dataset?

Answer: { Parking, Restaurant
Movie, Restaurant }

Co-location Examples

- ❑ Which spatial events are related to each other?
- ❑ Which spatial phenomena depend on other phenomenon?

Domain	Example Features	Example Co-location Patterns
Epidemiology	Disease types, environmental events	{West Nile disease, stagnant water sources, dead birds, mosquitoes}
Location-based services	Service type requests	{tow truck, police, ambulance}
Business	Local store types	{Burger King, MacDonald's}
Transportation	Delivery service tracks	{US Postal Service, UPS, newspaper delivery}
Military	Critical points, events	{weapons Caches, IED factories}
Economics	Industry types	{suppliers, producers, consultants}
Ecology	Species	{Nile crocodile, Egyptian plover}
Earth Science	Climate and disturbance events	{wild fire, hot, dry, lightning}
Weather	Fronts, precipitation	{cold front, warm front, snow fall}

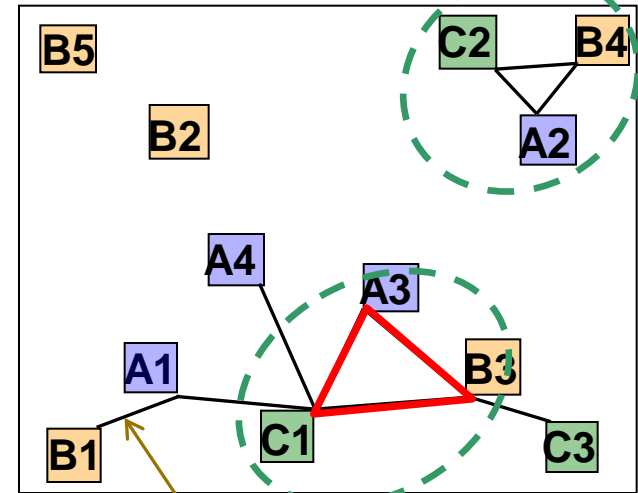
Preliminary - Key Terms

- A co-location X
: A subset of spatial event types
- A neighborhood
: A clique in a graph of neighbor relation

- An instance I of co-location $X = \{e_1 \cdots e_k\}$
: $I = \{o_1 \cdots o_k\}$

- * o_j is object of e_j ($\forall j \in 1, \dots, k$).
- * I is a neighborhood.

- * A1 : event type A, object id 1
- * — : neighbor relationship



Distance(A1, B1) < neighbor distance threshold

A	B	C	
A3	B3	C1	← Co-location instances
A2	B4	C2	

Preliminary - Interest Measures

- Participation Index PI
of co-location $X = \{e_1 \cdots e_k\}$

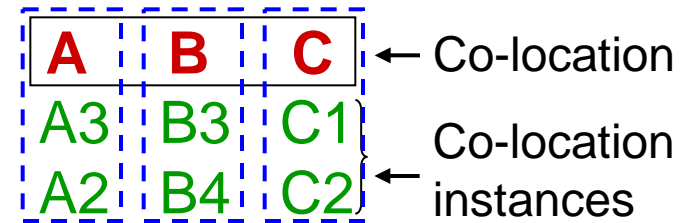
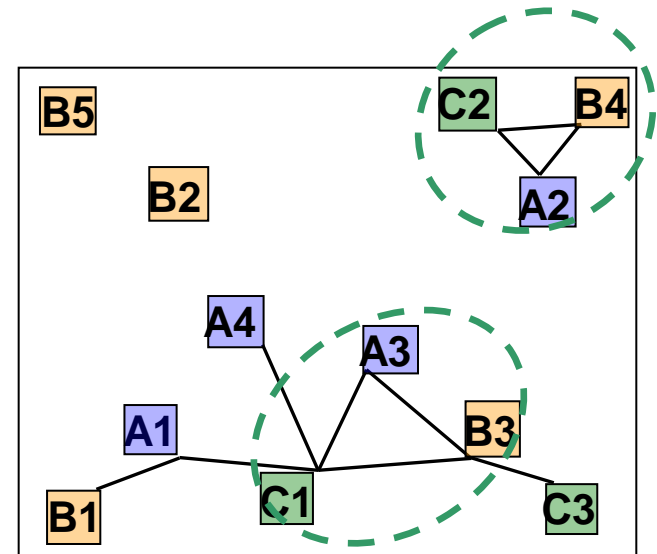
$$PI(X) = \min_{e_i \in X} \{PR(e_i, X)\}$$

* Strength of prevalence of co-location

- Participation Ratio $PR(e_i, X)$

$$\frac{\# \text{ of objects of } e_i \text{ in instances of } X}{\# \text{ of objects of } e_i}$$

* Strength of each event type
in a co-location








$$\min\left(\frac{2/4}{2/5}, \frac{2/5}{2/5}, \frac{2/3}{2/5}\right) = \frac{2/4}{2/5} \leftarrow PR \text{ of A}$$

$$\frac{2/4}{2/5} \leftarrow PI$$

If $PI > \text{prev_threshold}$, $\{A, B, C\}$ is a frequent co-location.

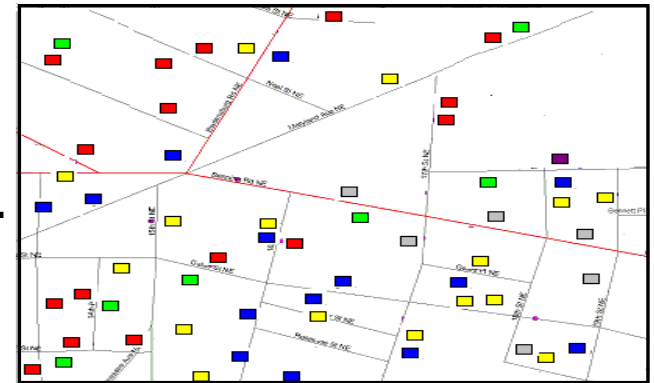
Related Work: General Associations

Trans	Items Bought
1	{socks,  ,  , milk, beef, egg, ...}
2	{ice-cream, muffin,  , ..}
3	{  ,  , pillow, toothbrush, ...}
4	{juice, egg, chicken, battery, ...}

* transaction: a set of items

E.g., Diaper → Beer (0.5, 1)

vs.



*  : a spatial feature

E.g., Police → Tow, Ambulance (0.5, 0.8)

Criteria	Association Pattern	Co-location Pattern
Underlying Space	Discrete Sets	Continuous Space
Item Types	Product Types	Spatial Events(Features)
Item Collections	Transactions T	Neighborhoods of L
Prevalence (A, B)	Support: $P(A \cap B \in T_i)$	Spatial Prevalence Measure

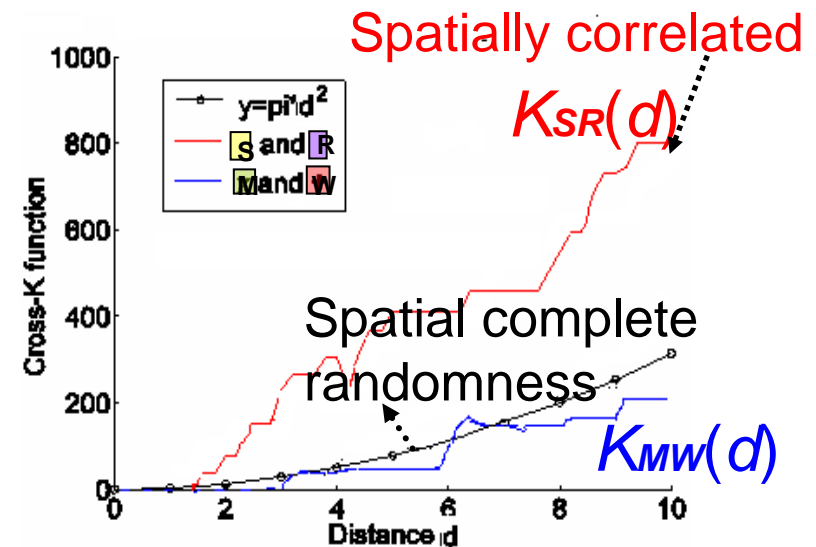
Related Work: Statistical Approach

■ Ripley's Cross K-Function [Cressie]

- $K_{ij}(d) = \lambda_j^{-1} E$ [number of type j feature within distance d of a randomly chosen type i feature]

■ Limitations

- **Not proper** for analysis of **features of size ≥ 3**
e.g., triple features (K,T,R)
- **Not efficient in computation**



Related Work: Colocation in Oracle

Oracle® Spatial User's Guide and Reference
10g Release 1 (10.1)
Part Number B10826-01



8 Spatial Analysis and Mining

This chapter describes the Oracle Spatial support for spatial analysis and mining in Oracle Data Mining (ODM) applications.

8.4 Colocation Mining

Colocation is the presence of two or more spatial objects at the same location or at significantly close distances from each other. Colocation patterns can indicate interesting associations among spatial data objects with respect to their nonspatial attributes. For example, a data mining application could discover that sales at franchises of a specific pizza restaurant chain were higher at restaurants collocated with video stores than at restaurants not collocated with video stores.

Two types of colocation mining are supported:

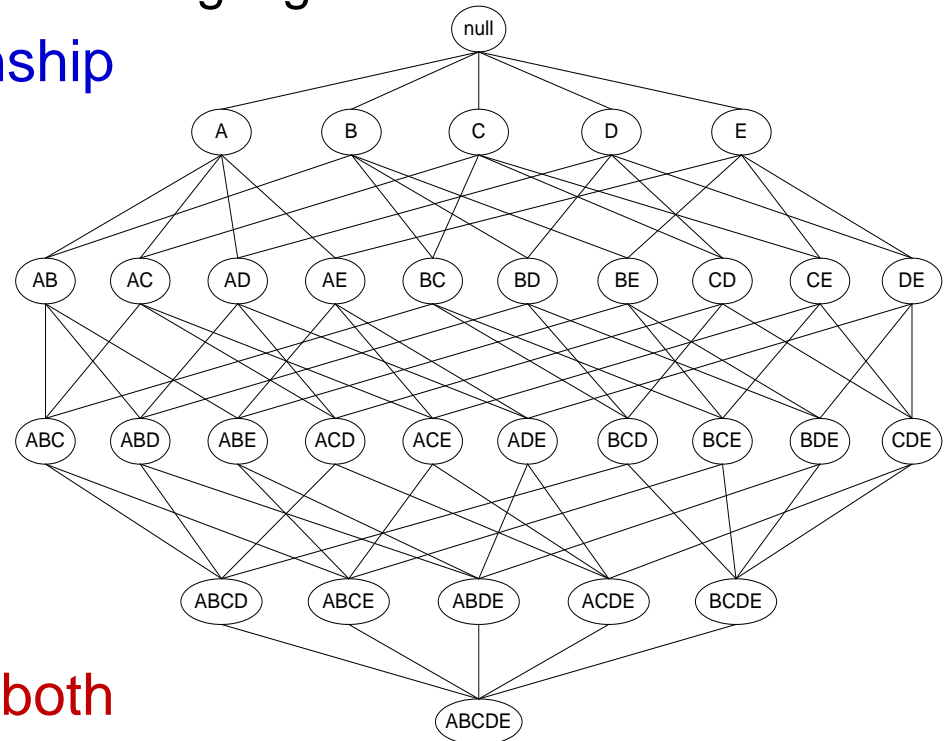
- Colocation of items in a data mining table. Given a data layer, this approach identifies the colocation of multiple features. For example, predator and prey species could be collocated in animal habitats, and high-sales pizza restaurants could be collocated with high-sales video stores. You can use a reference-feature approach (using one feature as a reference and the other features as thematic attributes, and materializing all neighbors for the reference feature) or a buffer-based approach (materializing all items that are within all windows of a specified size).
- Colocation with thematic layers. Given several data layers, this approach identifies colocation across the layers. For example, given a lakes layer and a vegetation layer, lakes could be collocated with areas of high vegetation. You materialize the data, add categorical and numerical spatial relationships to the data mining table, and apply the ODM Association-Rule mechanisms.

The following functions and procedures, documented in [Chapter 21](#), perform operations related to colocation mining:

- [SDO_SAM.COLOCATED_REFERENCE_FEATURES](#)
 - [SDO_SAM.BIN_GEOMETRY](#)
-

Challenges

- **No explicit transaction** concept in spatial data
 - Non-trivial to reuse association mining algorithms
- **Continuous neighbor relationship**
 - Especially, clique relations
- **Very large search space**
 - Given n features, there are around 2^n possible candidate feature sets
- **Other workload**
 - Density, neighbor distance, prevalence threshold, etc.
- **Inherently too demanding of both processing time and memory requirements**



Why Not Use Modern Computational Framework

- Standard architecture emerging:
 - Cluster of commodity Linux nodes
 - Gigabit Ethernet interconnect
 - Popular modern computational framework:
 - Cloud computing (distributed computing over a network)
 - Hadoop (MapReduce), a software framework for large-scale processing of data on clusters of commodity hardware.
 - How to organize the mining computations on this architecture?
-

Problem Formulation

■ Given

- A spatial event dataset, $\langle \text{id, event type, location} \rangle$
- A spatial neighbor relationship (e.g., distance threshold)
- A *prev_threshold*

■ Find

Co-location patterns with participation index $> \textit{prev_threshold}$

■ Objectives

Develop a parallel/distributed co-location mining algorithm for spatial association analysis in cloud computing environment.

Outline

- Introduction to Data Mining
 - My Works in Data Mining
 - Spatial Data Mining
 - Spatial Association Mining in Cloud Computing Environment
 - Background: Hadoop and MapReduce
 - Proposed Approach
 - Experimental Evaluation
 - Summary
-

Background: Hadoop

- Execution framework for running applications on large clusters of commodity hardware

Includes

- Storage: **HDFS**
- Processing: **MapReduce**
 - Support the Map/Reduce programming model

- Characteristics

- Economy: use cluster of commodity computers
- Easy to use
 - Users: no need to deal with the complexity of distributed computing
- Reliable: can handle node failures automatically



Background: MapReduce



- The heart of Hadoop
- A programming model for processing large data sets with a parallel, distributed algorithm on a cluster.
- The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform.
 - The first is the map job, which takes a set of input data and converts it into another set of data.
 - The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples.
 - Animation: <http://www.systems-deployment.com/animation.html>

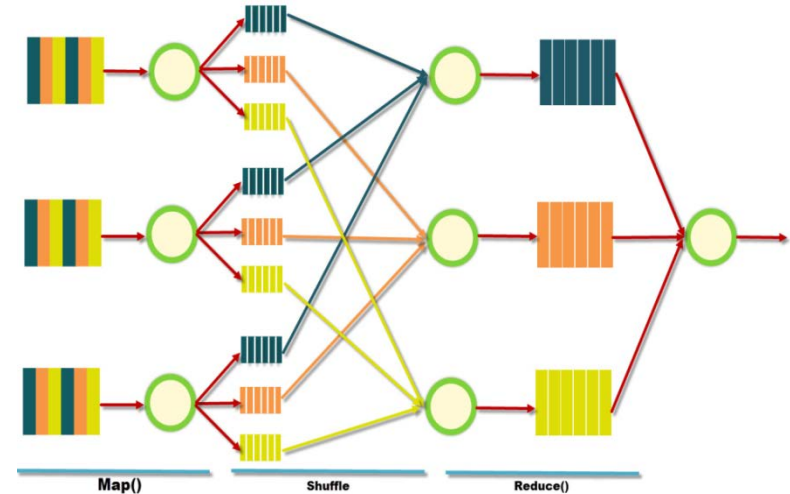


Figure source: <http://blog.sqlauthority.com/2013/10/09/big-data-buzz-words-what-is-mapreduce-day-7-of-21/>

Example: Word Count on MapReduce

- A **map function** process a **key/value pair** to generate a set of intermediate **key/value pairs**

map(key=null, val=record):

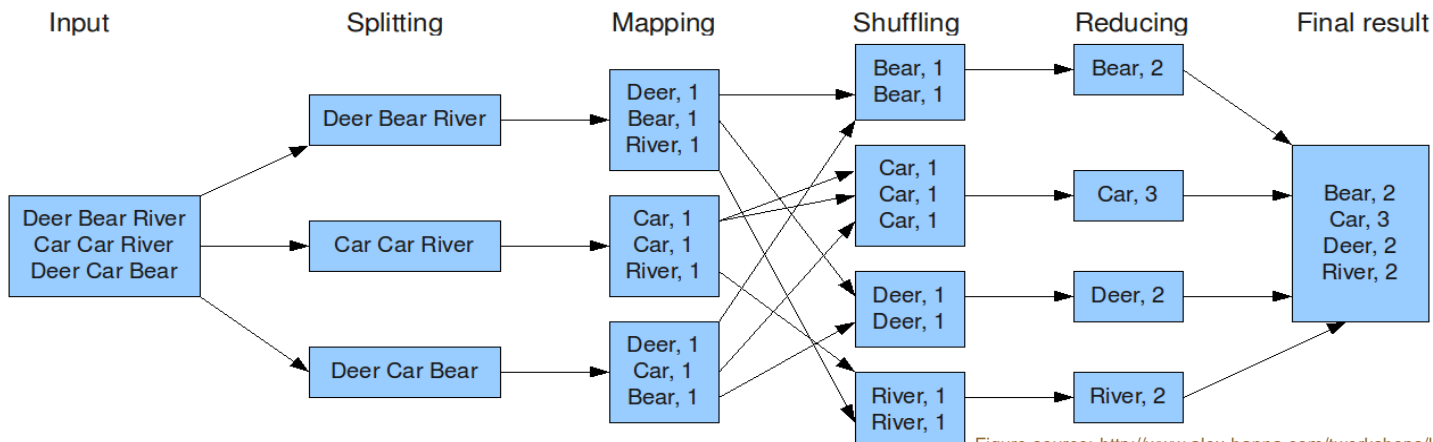
For each word w in contents, emit (w , "1")

- The shuffling step merges all intermediate values associated with the same intermediate key and feed the key/values pairs to a **reduce function**. The reducer generates a set of result **key/value pairs**.

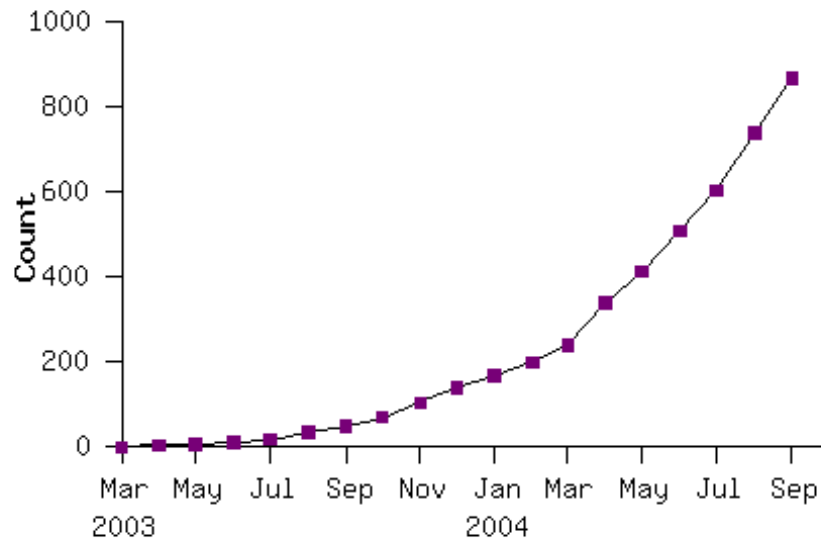
reduce(key=word w , values=[1, 1, ...,1]):

Sum all "1"s in values list

Emit result (word, sum)



MapReduce Model is Widely Applicable



■ Example uses:

distributed grep

term-vector / host

document clustering

...

distributed sort

web access log stats

machine learning

...

web link-graph reversal

inverted index construction

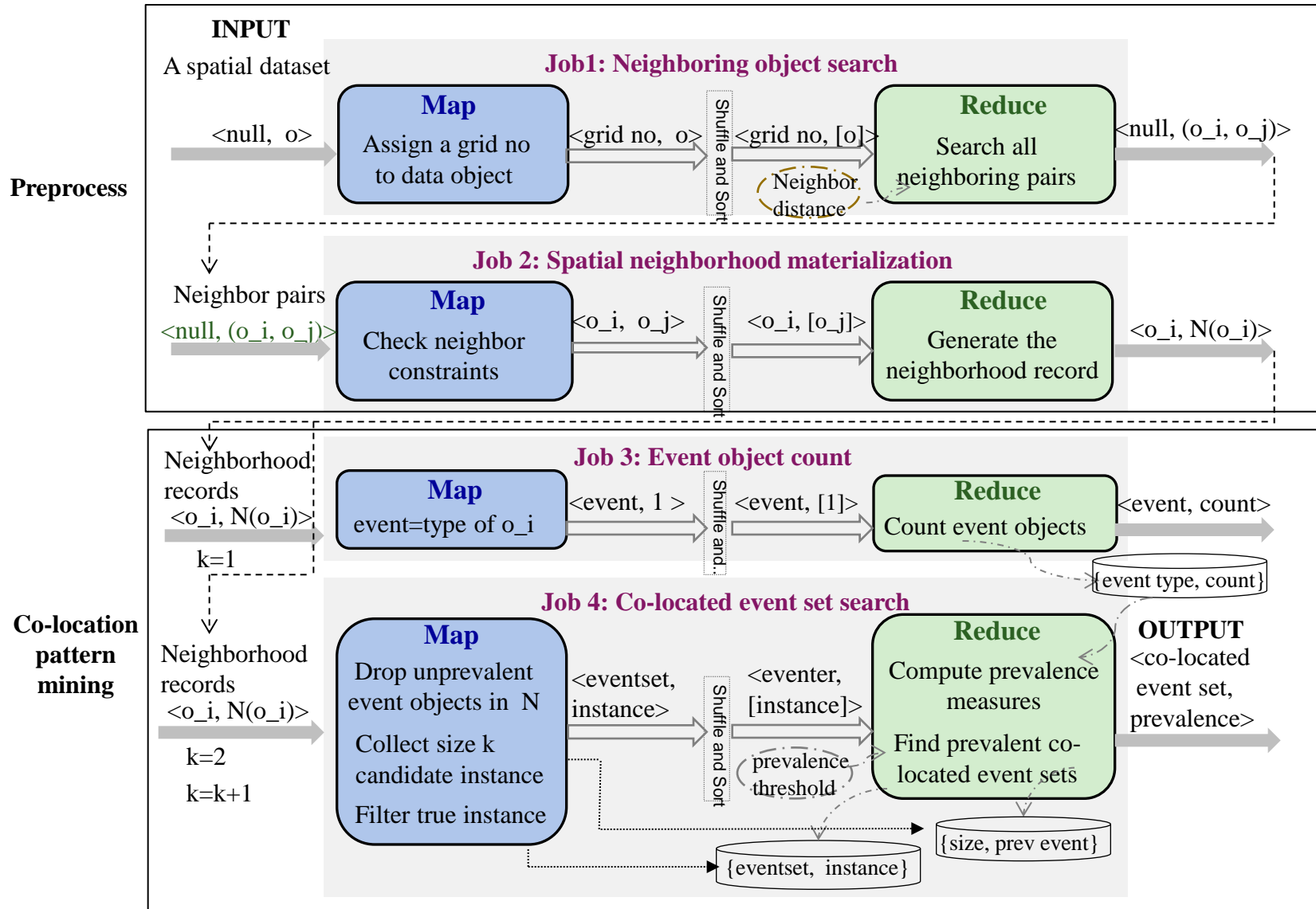
statistical machine translation

...

Outline

- Introduction to Data Mining
 - My Works in Data Mining
 - Spatial Data Mining
 - Spatial Association Mining in Cloud Computing Environment
 - Background: Hadoop and MapReduce
 - Proposed Approach
 - Experimental Evaluation
 - Summary
-

Co-location Mining on MapReduce



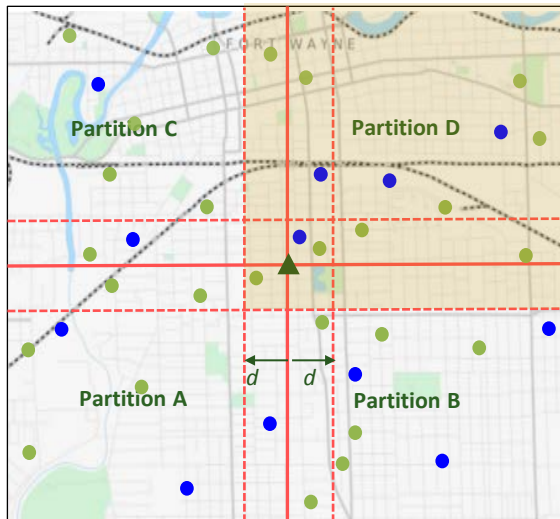
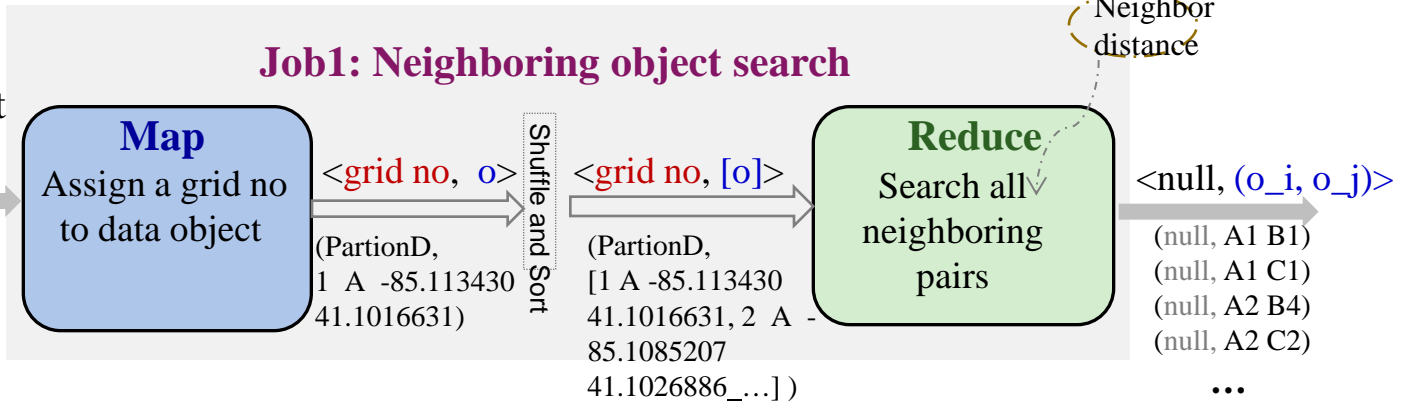
Job1: Neighbor object search

INPUT

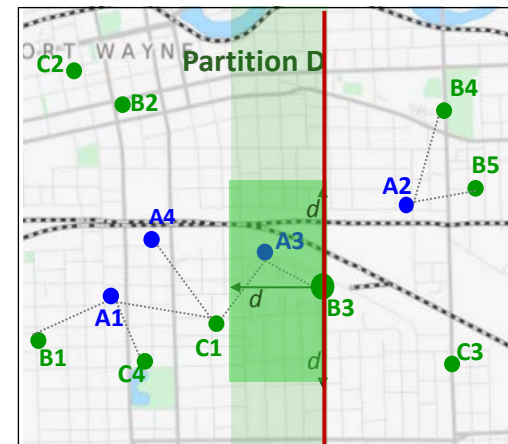
A spatial dataset

$\langle \text{null}, o \rangle$

1	A	-85.1113430	41.1016631
2	A	-85.1085207	41.1026886
1	B	-85.1036761	41.1018515
...			
4	C	-85.1084791	41.1017347



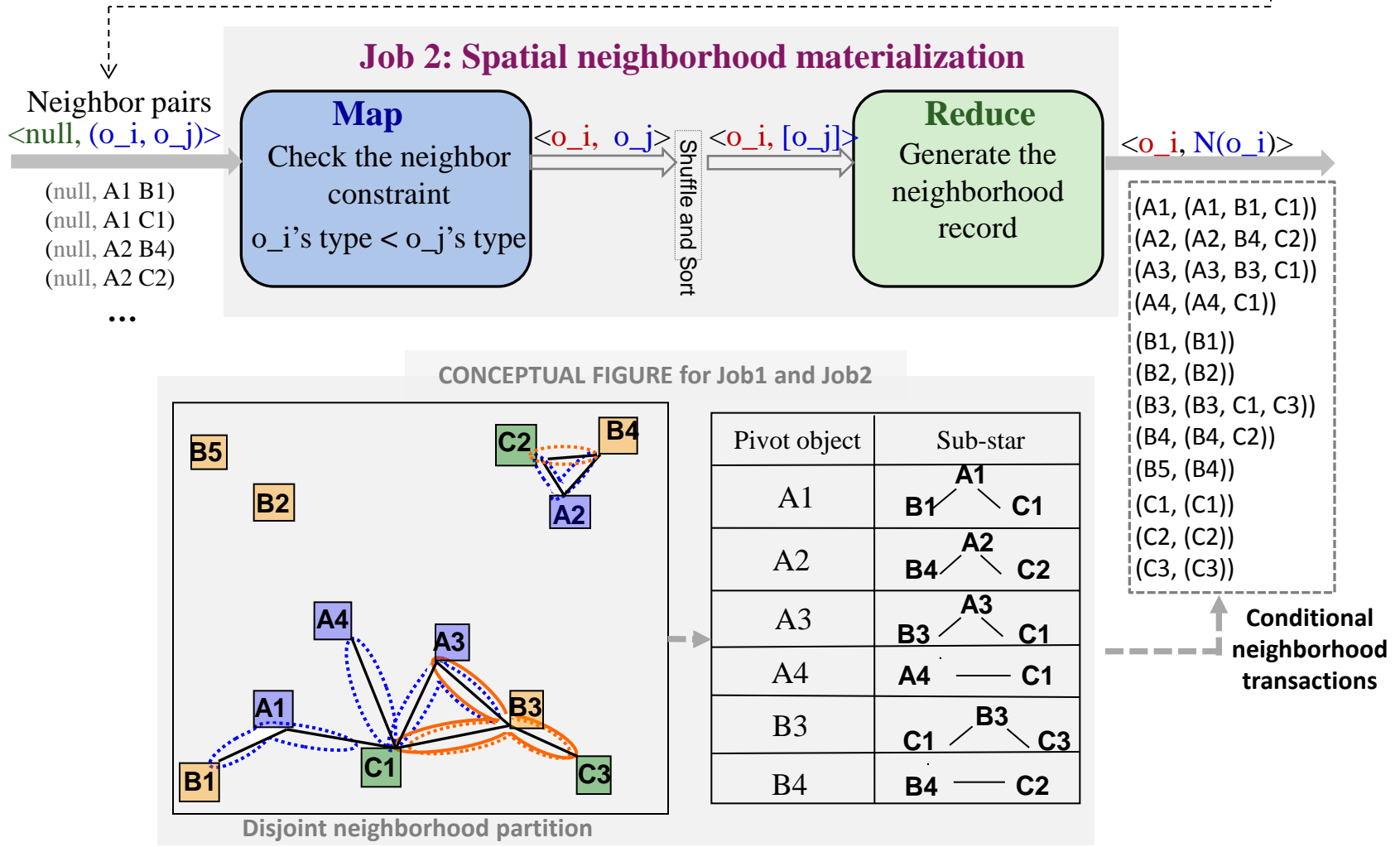
Overlapping space partition



Plane-sweep algorithm
 $O(n \log n)$

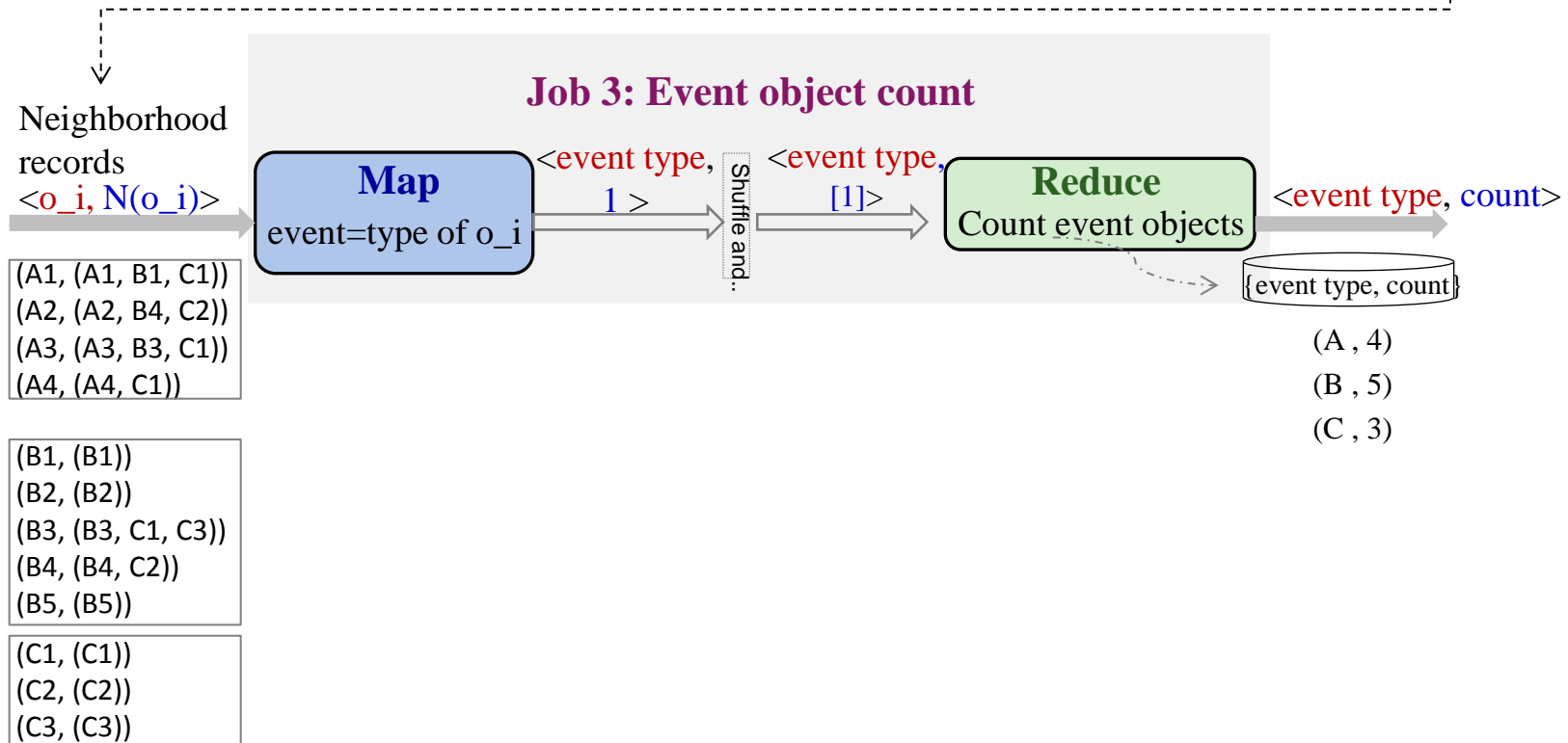
Job2: Neighbor object search

Job 1
Output

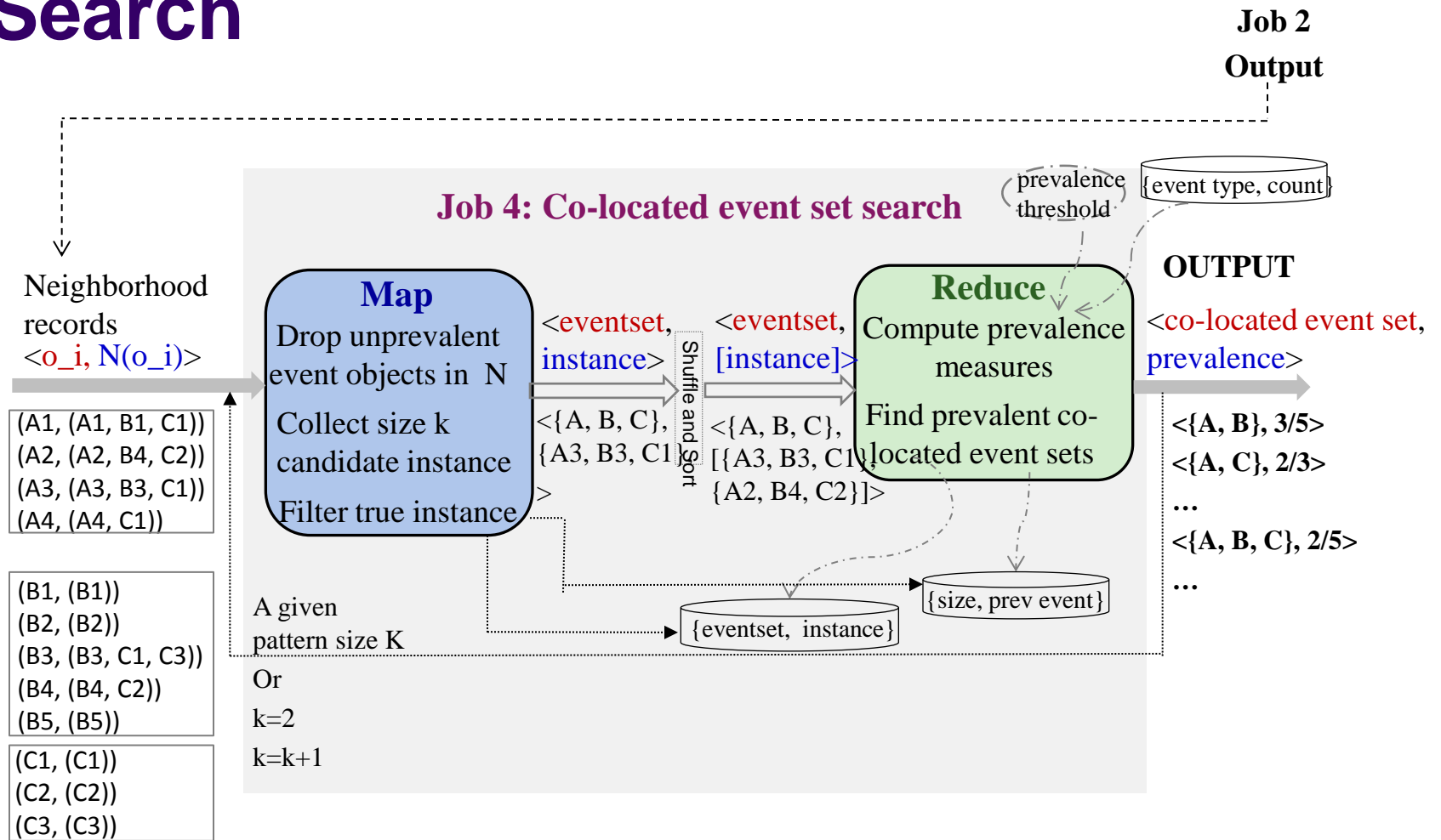


Job3: Event Object Count (Optional)

Job 2
Output



Job4: Prevalent Co-located Event Set Search

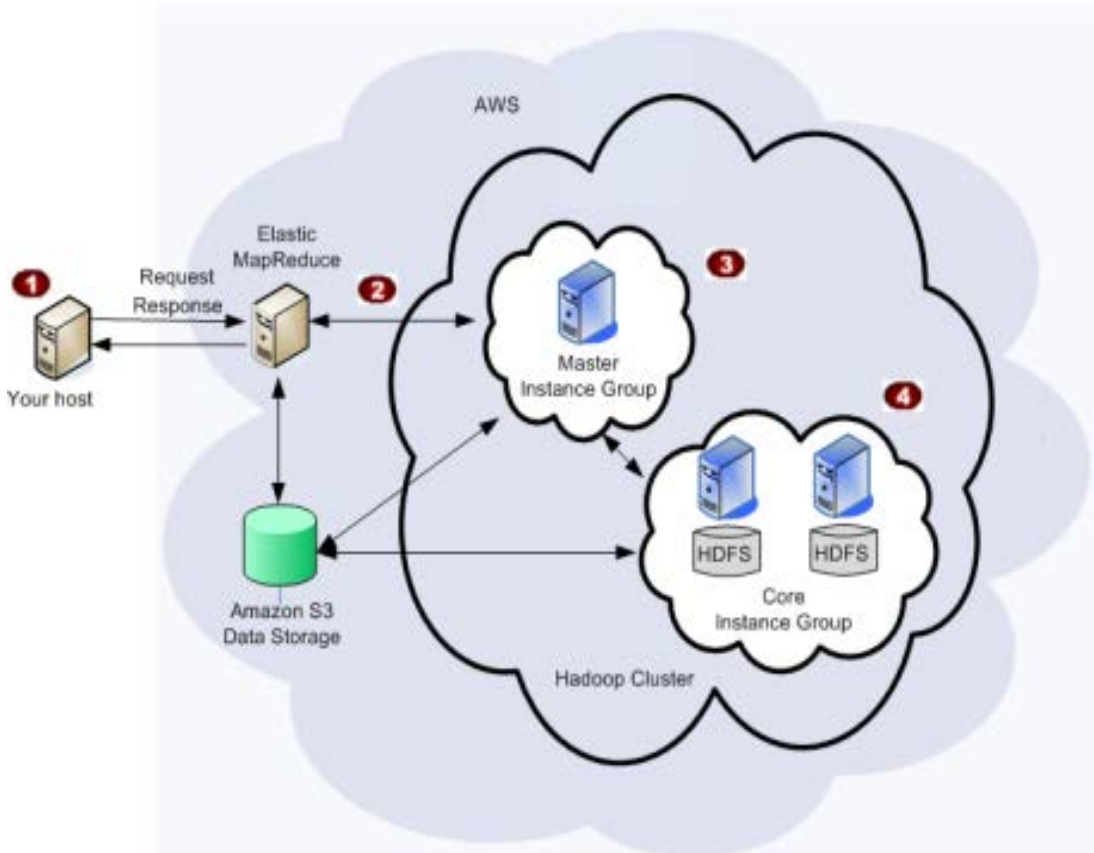


Outline

- Introduction to Data Mining
 - My Works in Data Mining
 - Spatial Data Mining
 - Spatial Association Mining in Cloud Computing Environment
 - Background: Hadoop and MapReduce
 - Proposed Approach
 - Experimental Evaluation
 - Summary
-

Experimental Environment

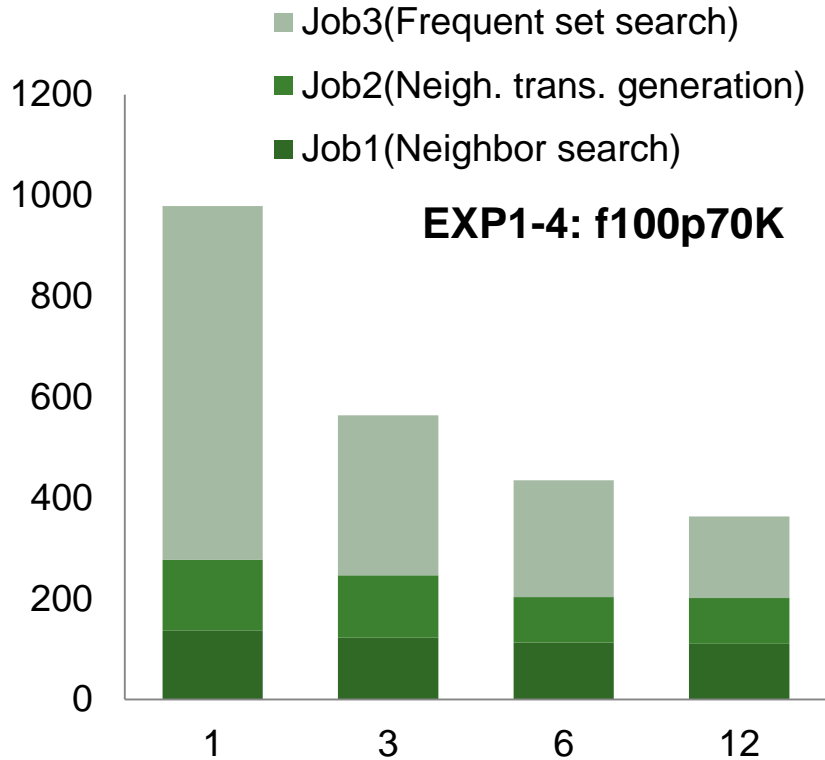
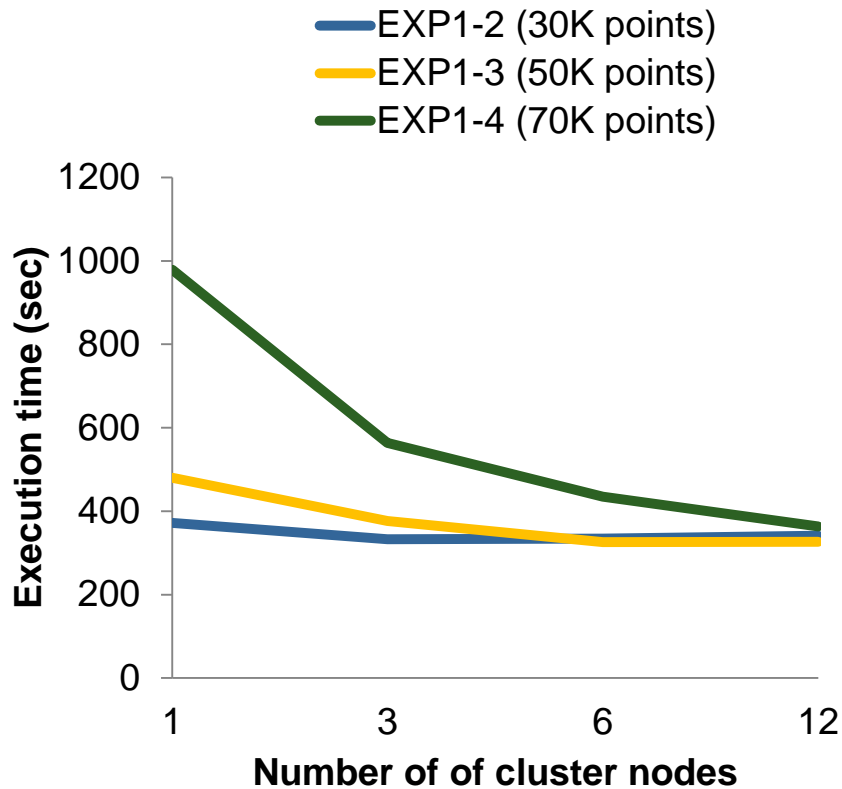
- For real resizable clusters, **Amazon Web Services (AWS) Elastic MapReduce (EMR) platform**



- **Clusters with 1 ~ 20 nodes**
- Node type: **m1.small** (1 CPU, 1.7GB memory, 160GB storage),
m1.large
- Software: **Hadoop 1.0.3, Hbase**
- Programming language: Java, MapReduce API, Hbase API

* Source: Amazon AWS

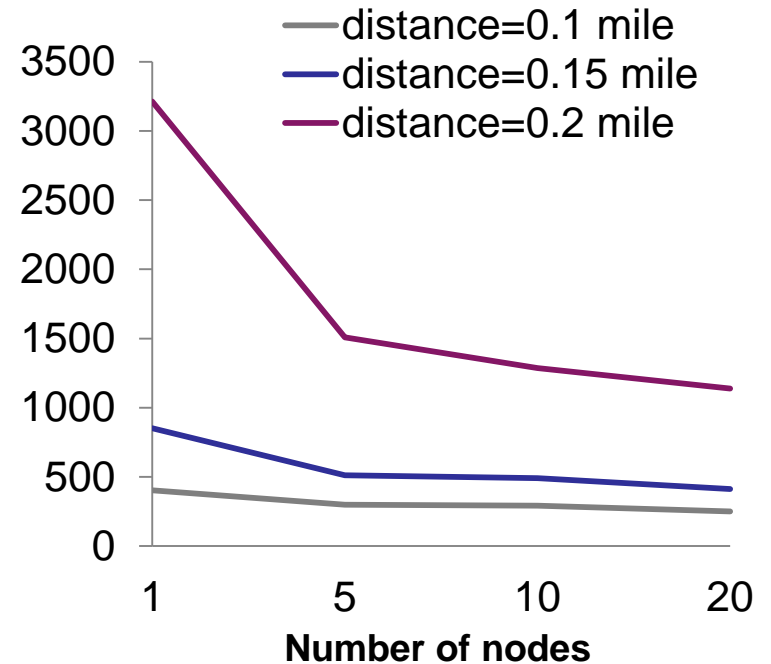
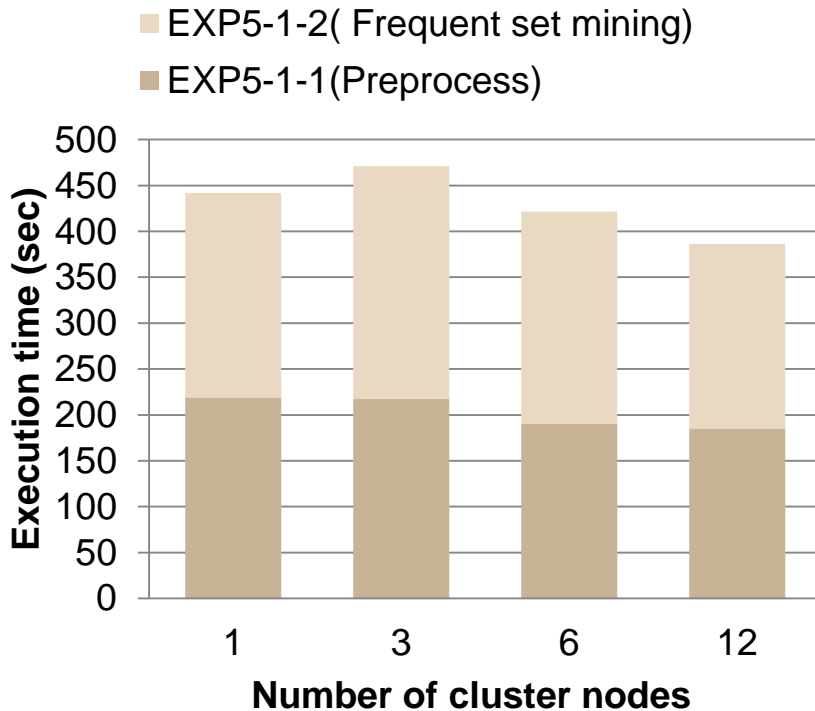
Results with Synthetic Datasets



* For other results, please refer the paper in <http://users.ipfw.edu/yooj/publication.html>

Results with Real-world Datasets

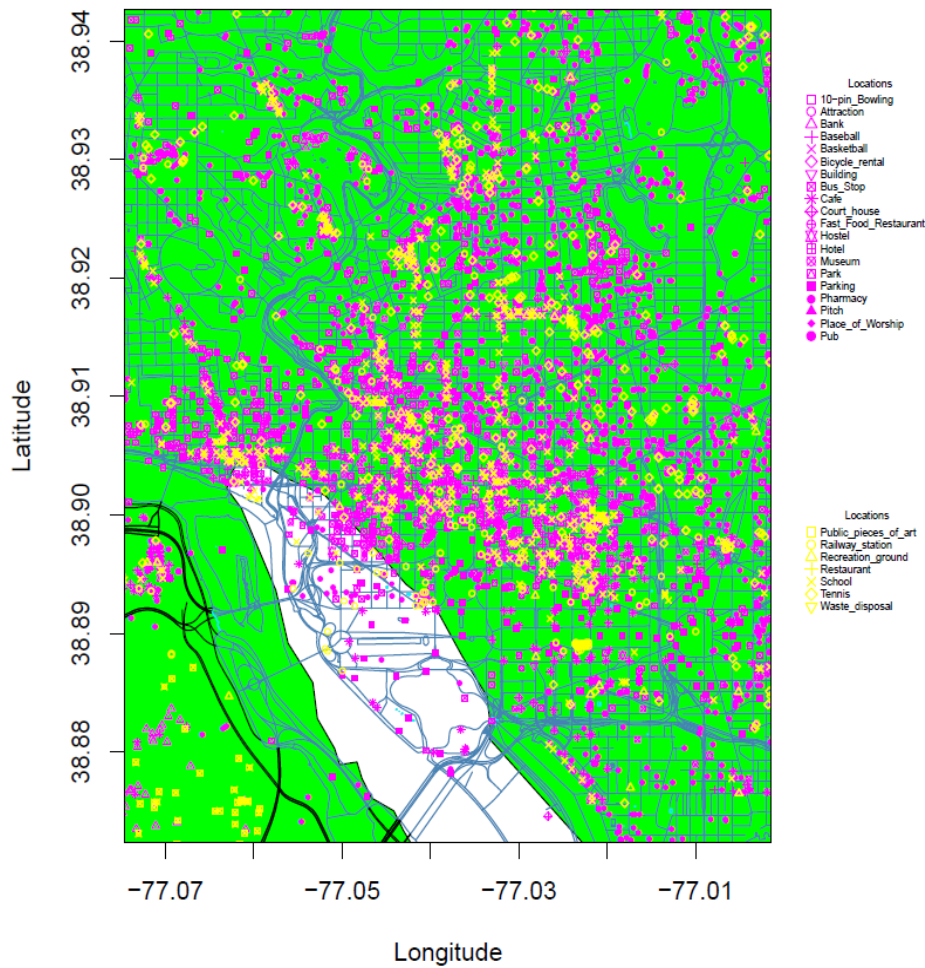
- Theft incidents (5218 points) and Point of interests (765 points, 16 types) in Fort Wayne area, and 1km for neighbor distance
- Point of Interests in Washington D.C. (17000 points, 87 types)



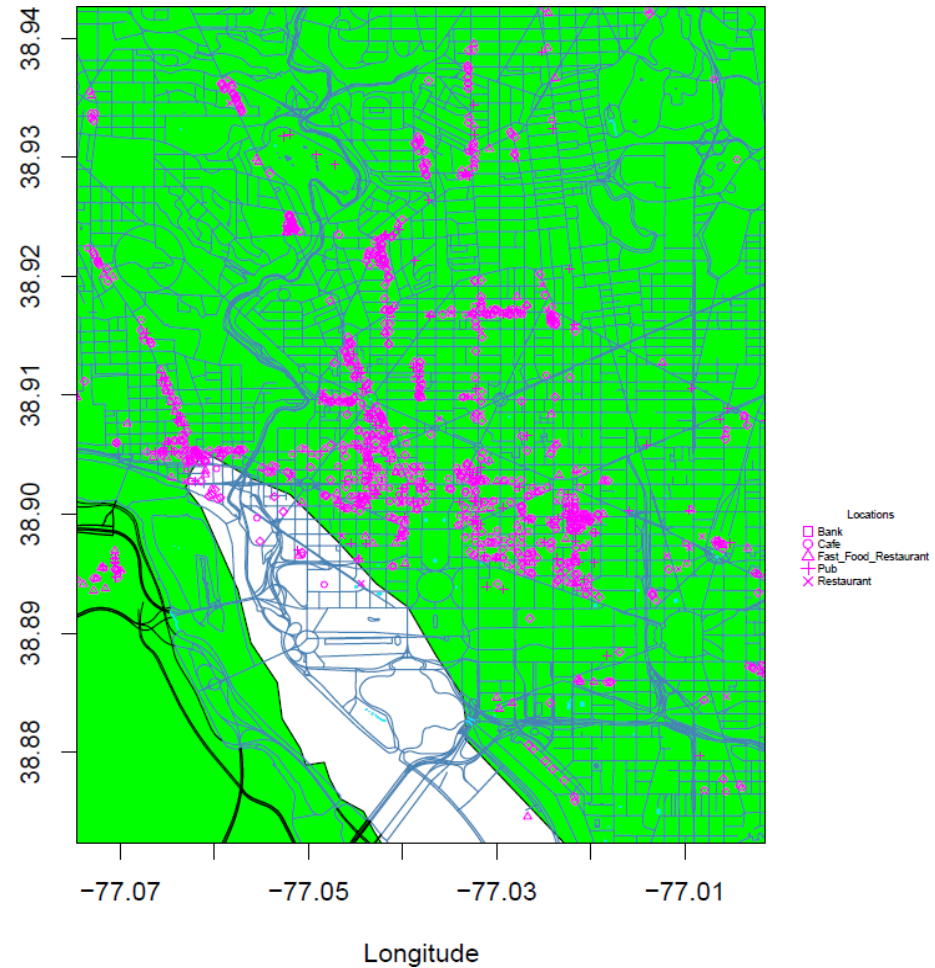
Finding from Washington DC POI Data

- Point of Interests in Washington D.C. (17000 points, 87 types), 0.2 mile for neighbor distance
 - {Pitch, Tennis} (0.7)
 - {Museum, Public pieces of art} (0.55)
 - {Park, Parking}(0.54)
 - {Hostel, Recreation ground} (0.5)
 - {Bicycle rental, Public pieces of art} (0.49)
 - {Building, Parking} (0.48)
 - {Attraction, Waste disposal} (0.46)
 - {Bus Stop, Fast Food Restaurant} (0.44)
 - {10-pin Blowing, Court house} (0.42)
 - {Basketball, Tennis}(0.42)
 - ...
 - {Bank, Pub, Restaurant} (0.4)
 - {Bank, Café, Pub} (0.4)
 - {Bank, Café, Fast Food Restaurant} (0.43)
 - {Bank, Pub, Public pieces of art} (0.41)
 - {Bank, Café, Hotel} (0.41)
 - {Building, Café, Restaurant} (0.44)
 - {Building, Café, Fast Food Restaurant} (0.42)
 - {Building, Restaurant, Fast Food Restaurant} (0.42)
 - ...

Visualization



Visualization of data points of selected 27 types



Visualization of data points of some co-located types

Outline

- Introduction to Data Mining
 - My Works in Data Mining
 - Spatial Data Mining
 - Spatial Association Mining in Cloud Computing Environment
 - Background: Hadoop and MapReduce
 - Proposed Approach
 - Experimental Evaluation
 - Summary
-

Summary of Our Contributions

- Develop computational efficient methods to discover spatial association patterns
 - Parallel/distributed approaches in cloud computing environment (*IEEE BigData'13, IEEE BigData'14, Int'l Conf. in Adv. In Big Data Analytics'14*)
 - Incremental update approach (*PATTERN'14*)
 - Join-less approach (*ICDM'05, TKDE'06*)
 - Partial join approach (*ACM-GIS'04*)
 - Propose variant co-location patterns
 - Reduce sets of co-locations (*DMKD'13 accepted*)
 - Different framework of co-location mining (*DMKD'12*)
 - Top-*k* closed co-location patterns (*ICSDM'11*)
 - Maximal co-location patterns (*DaWak'11*)
 - *N*-most prevalent co-location patterns (*DaWak'09*)
 - Co-location mining for extended objects such as line and polygon (*SDM'04*)
-

Acknowledgement

- The recent work is supported by Air Force Research Laboratory (AFRL), Griffiss Business and Technology Park, and SUNY Research Foundation.



Thank you

Questions

Email to yooj@ipfw.edu

Homepage: <http://users.ipfw.edu/yooj>
