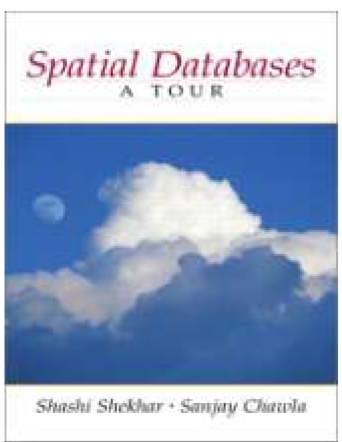
# *Spatio-temporal Data Mining for Environmental Sciences*

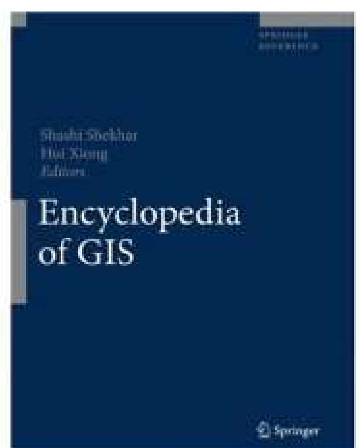

**Shashi Shekhar**

McKnight Distinguished University Professor

Faculty of Computer Sc. and Eng., Univ. of Minnesota

www.cs.umn.edu/~shekhar


# Acknowledgements


- **Spatial Database and Data Mining Group**
  - Dr. James Kang – Flow Anomaly
  - M. Celik, S. Chawla, C. T. Lu (Spatial Outliers), V. R. Raju, W. Wu, H. Yan (Colocation), J. S. Yoo, P. Zhang, etc.
- **Collaborators (Env. Scientists):**
  - Prof. Paige Novak, Prof. William Arnold, Prof. Miki Hondzo, Christine Wennen, Mike Henjum
- **Sponsors:**
  - NSF, USDOD, U of M OVPR

# Spatio-Temporal Data Analysis

IBM A Smarter Planet

ORACLE SPATIAL

ESRI ArcGIS X

Microsoft Virtual Earth

Google Earth


Recently having attention in Industry and Academia

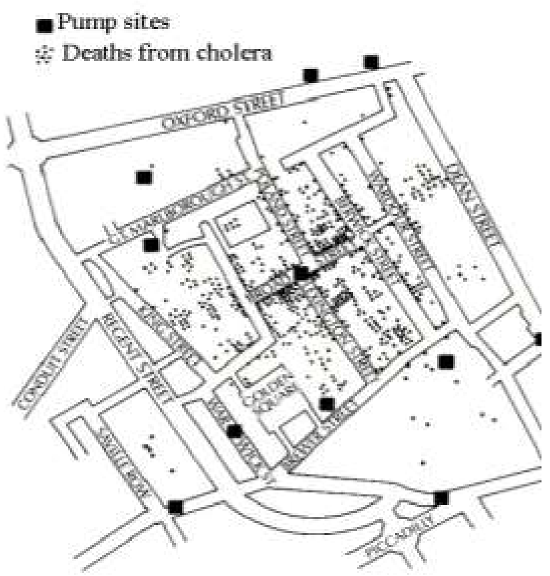

acm GIS SIG SPATIAL

SSTD 2009

# Outline

- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
- Gaps, Open Problems

# *Spatial and Spatio-temporal Data Mining*




## 1. What is it?

① Identifying interesting, useful, non-trivial patterns

② in large spatial or spatio-temporal datasets

## 2. Why is it important ?

① Potential of insights to improve human lives

- Environment: How is Earth system changing? Consequences for humans?
- Public health: Where are cancer clusters? Environmental reasons?
- Public safety: Where are hotspots of (env.) crime? Why?

② However, (d/dt) (Spatial Data Volume) >> (d/dt) (Number of Human Analysts)

- Need automated methods to mine patterns from spatial data
- Need tools to amplify human capabilities to analyze spatial data

# Spatial Data Mining (SDM)

1. **The process of discovering**
   ① interesting, useful, non-trivial patterns
      - patterns: non-specialist
      - exception to patterns: specialist

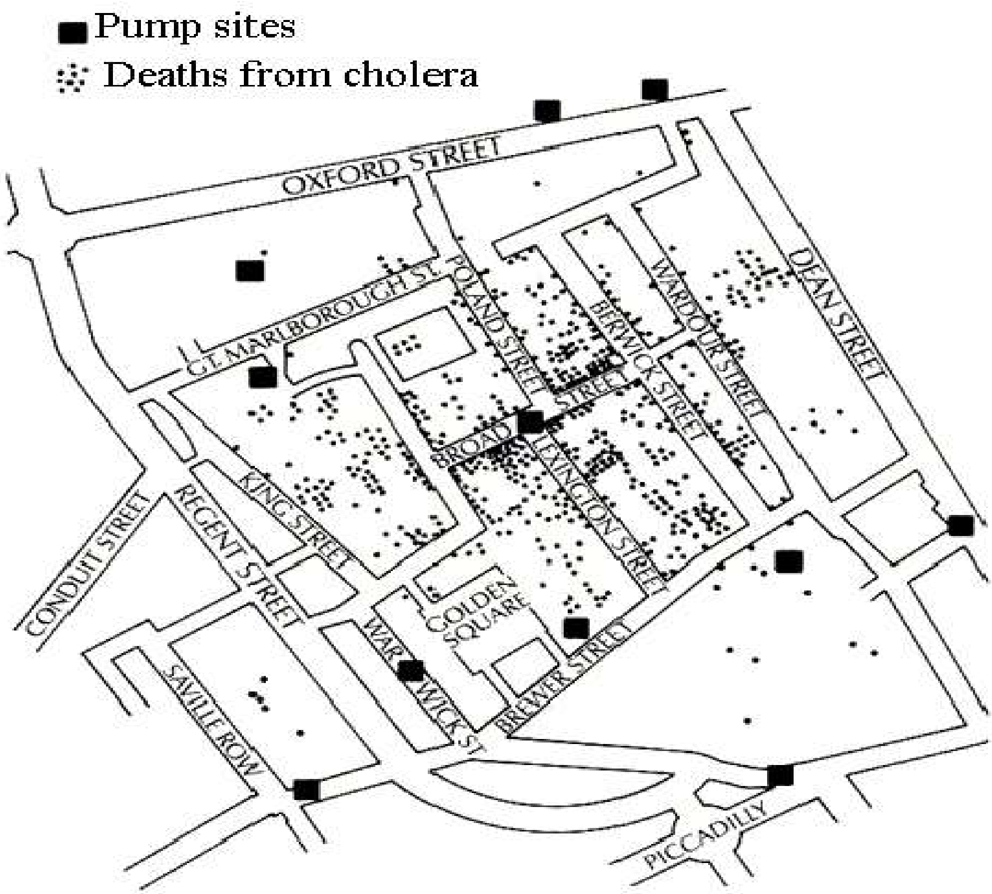
   ② from large spatial datasets

2. **Spatial pattern families**
   ① Hotspots, Spatial clusters
   ② Spatial outlier, discontinuities
   ③ Co-locations, co-occurrences
   ④ Location prediction models
   ⑤ …

# Pattern Family: Hotspots, Spatial Cluster

## The 1854 Asiatic Cholera in London

Near Broad St. water pump except a brewery

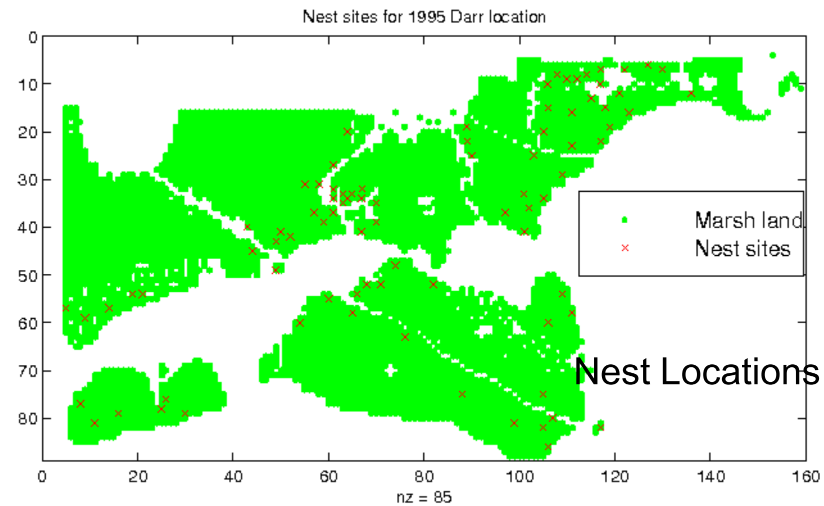


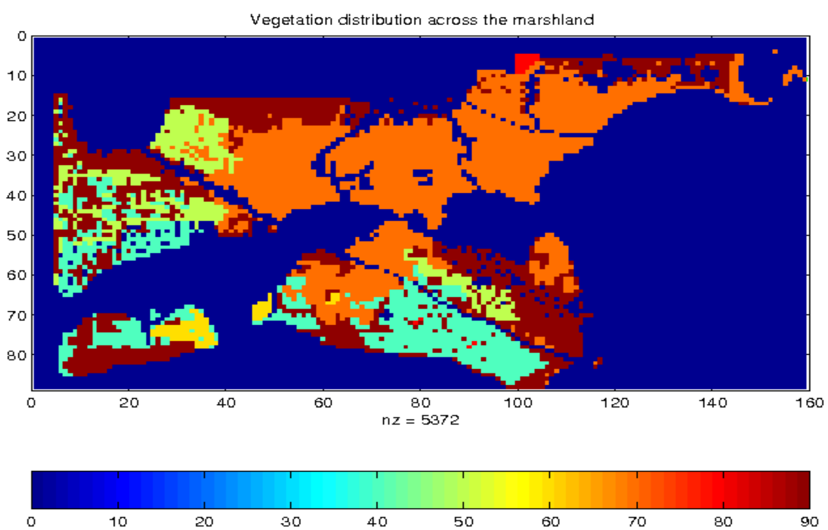
# *Pattern Family : Predictive Models*

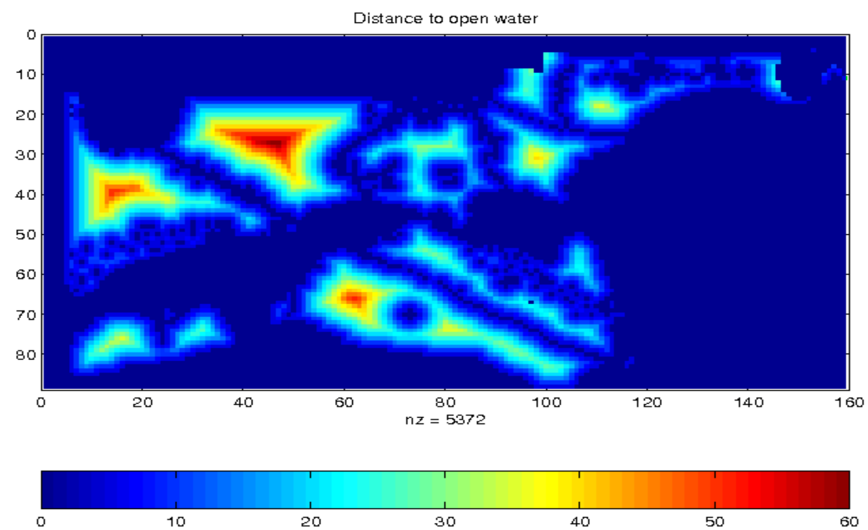
## Location & Direction Prediction:

Predict Bird Habitat Prediction

Using environmental variables

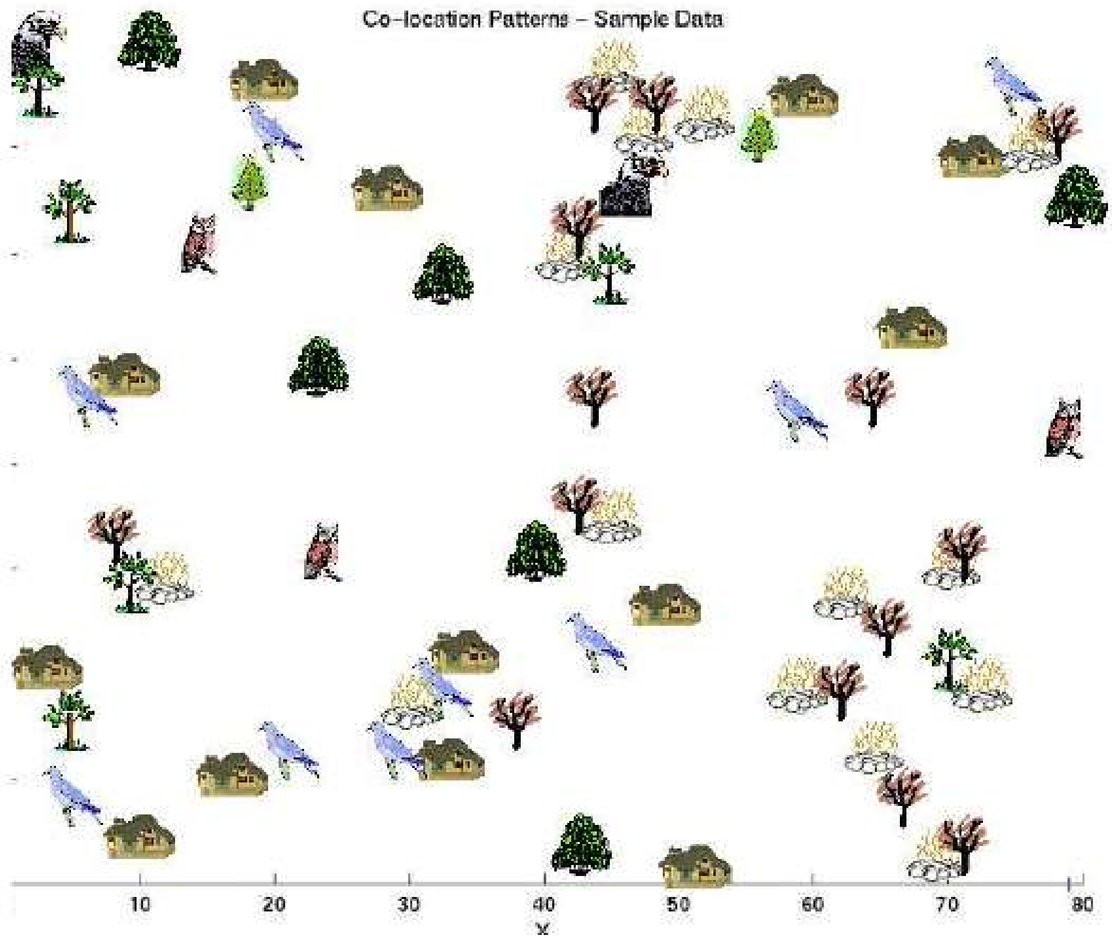


# Pattern Family : Co-locations/Co-occurrence

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




Answers:  and 

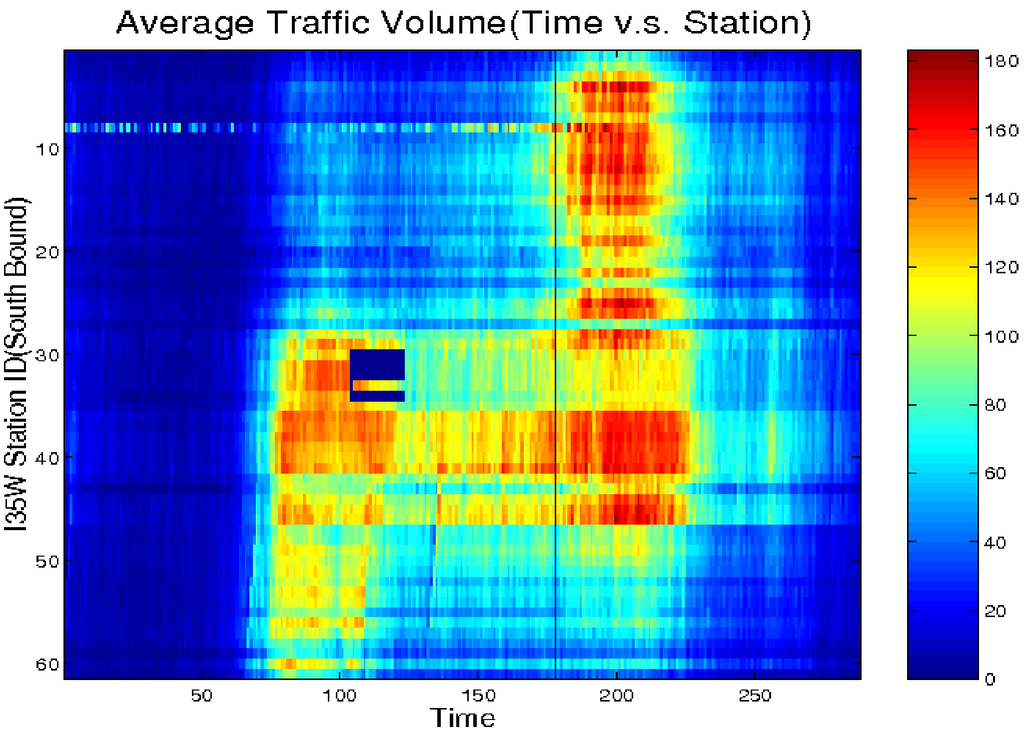
# Pattern Family: Spatial Anomalies

## Spatial Anomalies

① Traffic Data in Twin Cities

② Abnormal Sensor Detections

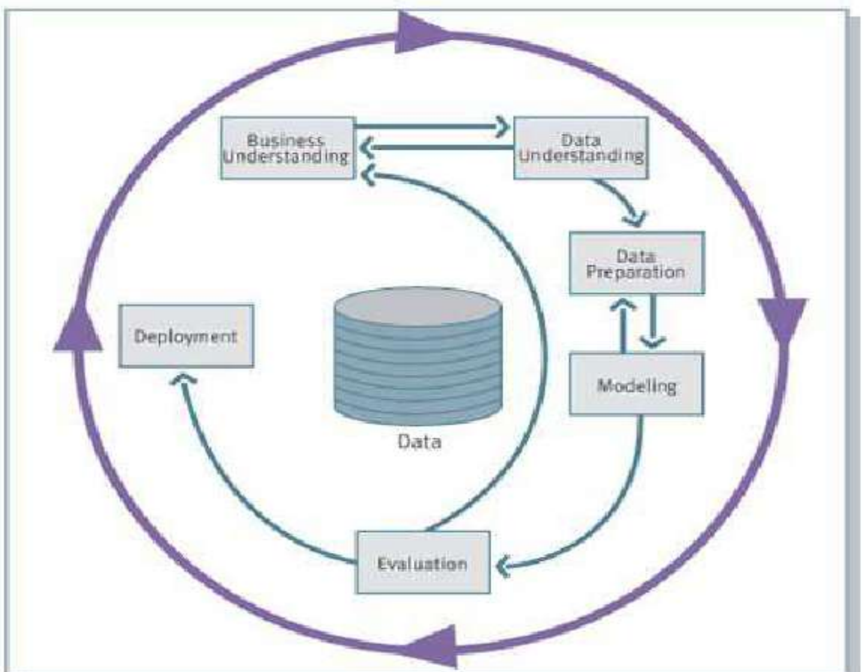
③ Spatial and Temporal Outliers




# Life Cycle of Data Mining

**CRISP-DM (CRoss-Industry Standard Process for DM)**

① Application/Business Understanding
② Data Understanding
③ Data Preparation
④ Modeling
⑤ Evaluation
⑥ Deployment




Phases of CRISP-DM

Is CRISP-DM adequate for Spatial Data Mining?

[1] CRISP-DM URL: http://www.crisp-dm.org

# Outline

- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
- Gaps, Open Problems

# Environment, Environmental Science

- **Environment**
  ① surroundings; milieu
  ② aggregate of surrounding things, conditions, or influences;.
  ③ Ecology . the air, water, minerals, organisms, and all other external factors surrounding and affecting a given organism at any time.
  ④ Social and cultural forces that shape the life of a person or a population

**Q? What is the relationship to spatial / spatio-temporal analysis?**

It allow inclusion of context, i.e. surrounding.

- **Environmental Science**
  ① study of the interactions among the physical, chemical and biological components of the environment
  ② branch of science concerned with the physical, chemical, and biological conditions of the environment and their effect on organisms.

# Examples of Environmental Sciences

- **Environmental chemistry:**
  - Soil, water, air pollution; multi-phase transport, fate; impact on species, geology
  - Study of chemical alterations in the environment.
- **Atmospheric sciences:**
  - meteorology, greenhouse gas phenomena, airborne contaminant dispersion, …
  - Global warming: atmospheric circulation, air-borne chemicals and their reactions, carbon dioxide fluxes from life-forms, atmospheric dynamics, etc.
- **Geosciences, hydrology, oceanography:**
  - environmental geology, environmental soil science, volcanic phenomena, surface runoff, sediment transport, water turbidity, …
- **Ecology:**
  - study of organisms and their interactions with each other and their environment
- **Env. Health, Env. Physiology, …**
- **Env. Justice, Env. Criminology**
- **Env. Engineering**
- **Env. Psychology, Env. Sociology**
- **…**

# Water Quality


**What's in our Water?**

Recent studies found presence of pharmaceutical drugs in drinking water of many U.S. Cities

Source: **New York Times (April 3, 2007)** (http://www.nytimes.com/2007/04/03/science/earth/03water.html)

- **By 2025**,1.8 billion people could be living in water scarce areas
- **Today**, 750 million people live below the water-stress threshold of 1.7 K cubic meters per person

Souce: WFUNA, 15 Global Challenges

# Environmental Questions

1. **General Public**
   ① Is water safe for drinking, swimming ?
   ② Where are air quality warning?
2. **Drinking Water Manager**
   ① Is incoming water safe for water plant (reverse osmosis filters)?
   ② Is there a change in contaminants today (compared with recent days)?
3. **Environmental Scientist**
   ① Transport: Where will a contaminant go?
   ② Fate: What is the fate of a contaminant ?
   ③ Are there any new processes occurring in the water bodies?
4. **Environmental Forensics**
   ① Where did contaminant come from ?
   ② What are hotspots and hot moments?
5. **Policy**
   ① Compare policy options on environmental impact and social good.?
   ② How to communicate environmental decision to all stakeholders?
6. **Environment Protection Agency**
   ① What will be the impact on environment of a proposed change?

# Processes, Questions

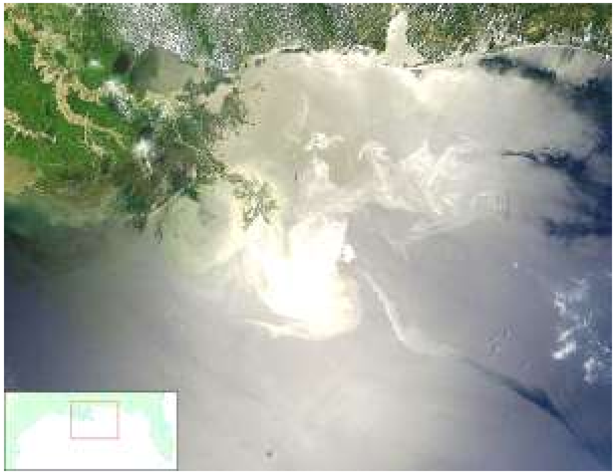
**ES Domain Questions:**

- Where do various contaminants go?
- Where did the contamination come from?

Path of Pollutant within the Environment
(Source: Schnoor, Environmental Modeling, 1996)

Mississippi RIver


Gulf of Mexico



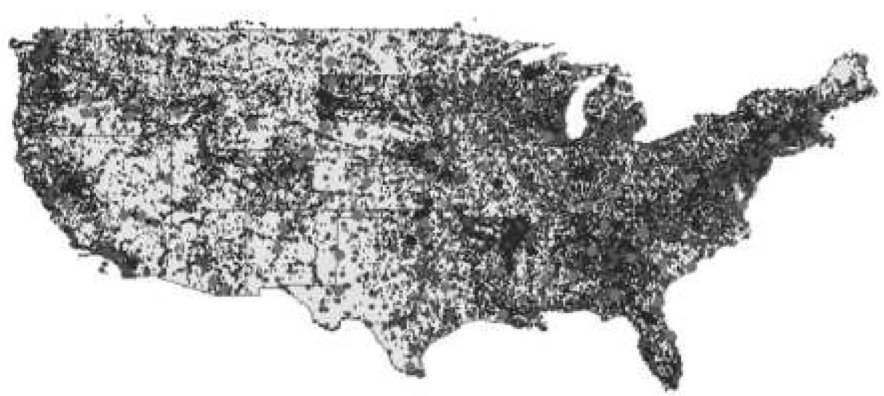
# Datasets

Data Sources:

- Hydrology Information Systems, CUAHSI
- United States Geological Survey

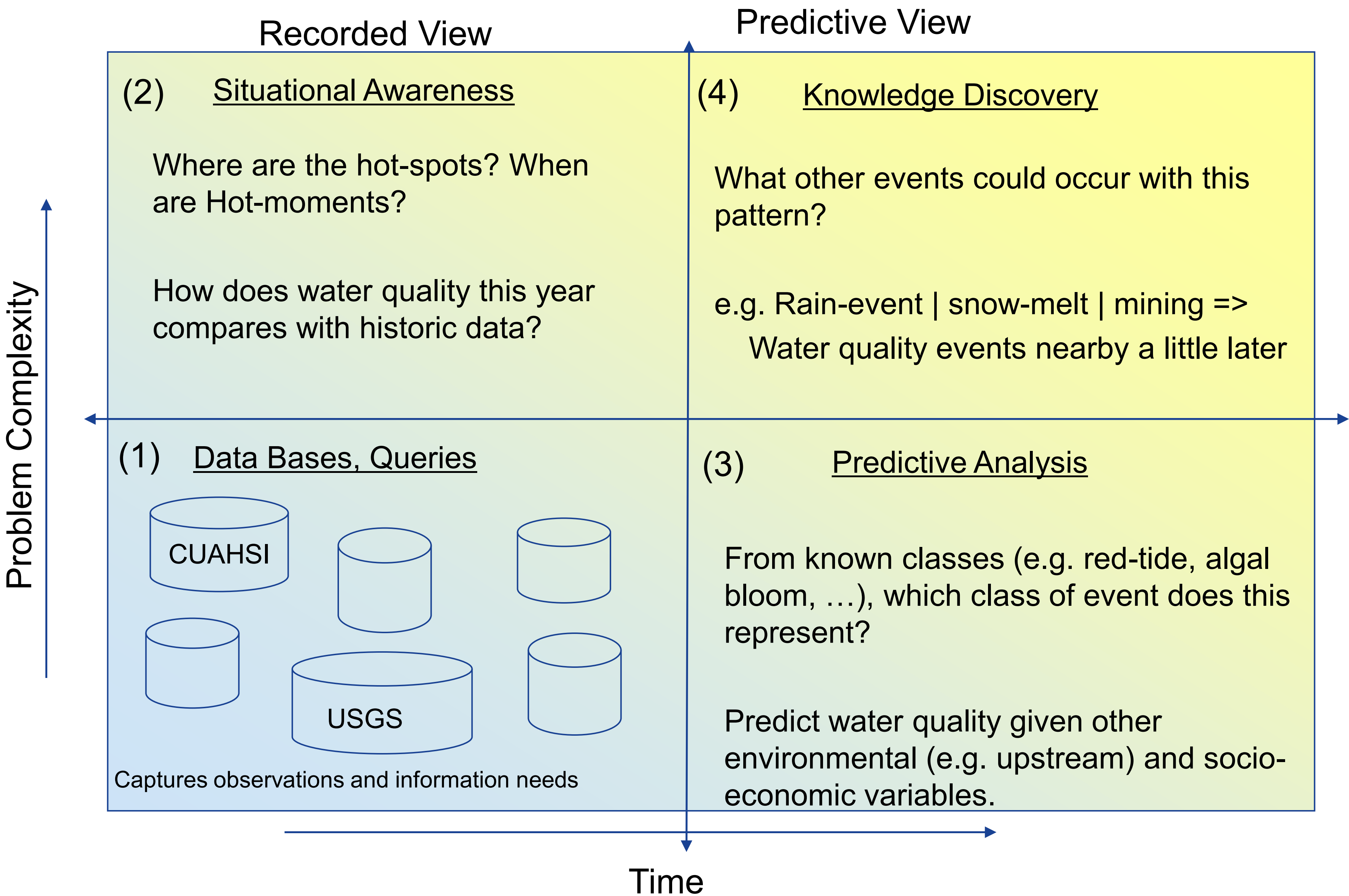
Data Characteristics (HIS/USGS)

- > 1.75 Million Locations
- > 342 Million Time Instants
- > 15K Measured Variables
  - Turbidity
  - Dissolved Oxygen
  - Nitrate
  - Etc.



Hydrology Measurement Sites in US
(Source: HIS/USGS)

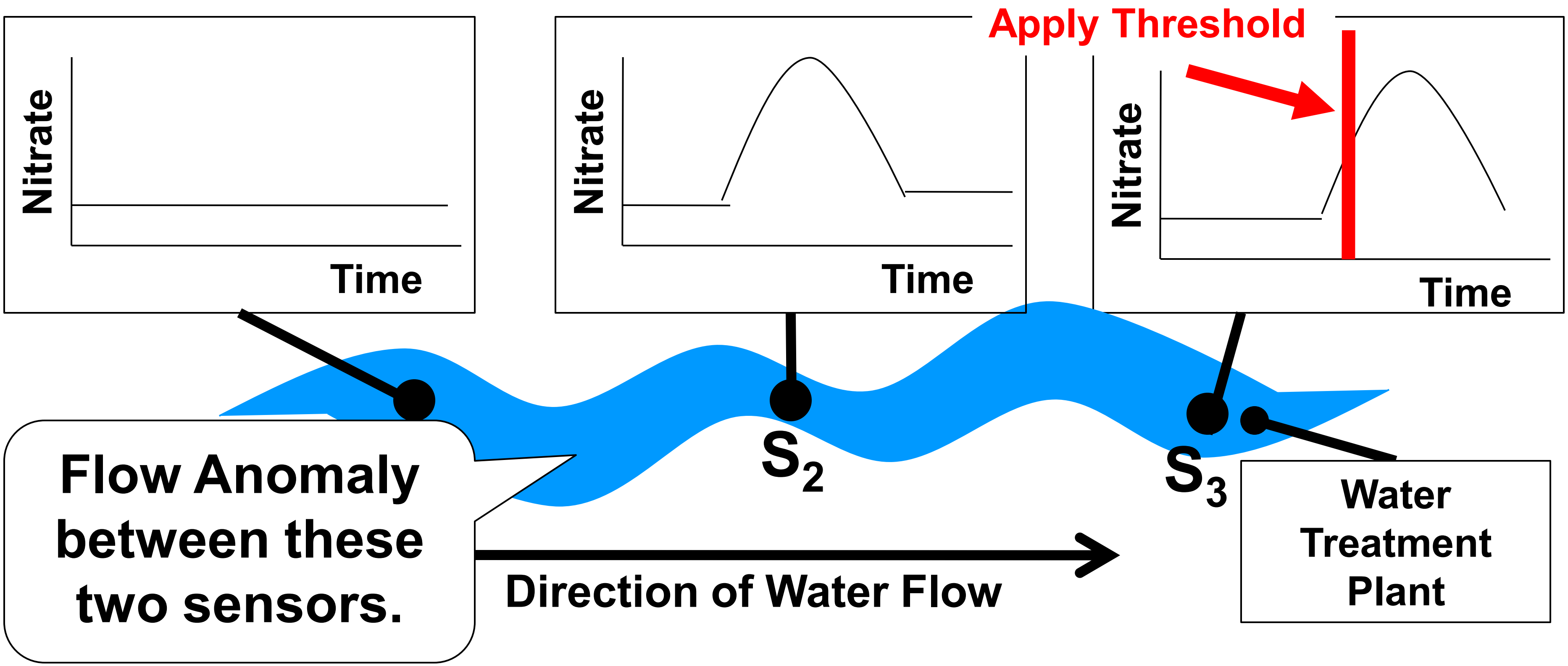
# Environmental Science: Data Analysis



# *Outline*

- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
  - **Key Concepts**
  - **Problem Statement**
  - **Contributions**
  - **Analytical Evaluation**
  - **Experimental Evaluation**
- Gaps, Open Problems

# Motivation – Detailed Example




**Two Use Cases:**

•At the water treatment plant, when should it turn off the water supply from the river?

•**Where is the source of the contaminant?**

# *Domain Example of a Flow Anomaly*


(Source: http://www.sfgate.com/cgi-bin/news/oilspill/busan)

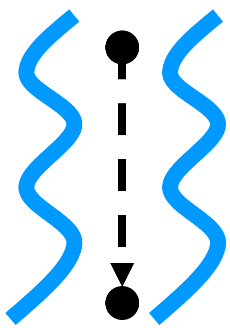
Notice that a contaminant event may **not** flow as a single contiguous unit.

Other Applications:

- Atmospheric Monitoring
- Pipeline Systems
- Transportation Networks

# *Concept: Transient Flow Anomaly*

- **<u>Transient Flow Anomaly</u> (tFA) is where the difference between the neighboring observations across each sensor is larger than the given error threshold, $\Theta_e$**
- ***Ex.* Suppose $\Theta_e = 10$**



| | | | | |
|---|---|---|---|---|
| **t =** | 1 | 2 | 3 | 4 |
| **TT [t] =** | 1 | 1 | 1 | 1 |
| **$f(st_1)$ =** | 10 | 20 | 30 | 40 |
| **$f(st_2)$ =** | 0 | 90 | 25 | 85 |
| **tFA [t] =** | 1 | 0 | 1 | - |

A tFA may represent a single time unit of a blob in an oil spil.

# Concept: Persistent Flow Anomaly

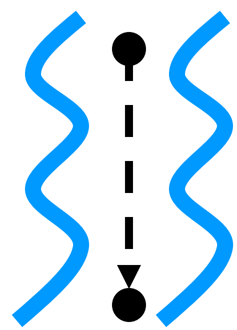
- **<u>Persistent Flow Anomaly</u> (pFA), is when the first and last are tFAs and the fraction of tFAs and time slots within a period satisfies the persistent threshold, $\Theta_p$**
- ***Ex.* Suppose $\Theta_e = 10$ and $\Theta_p = 0.5$**

For a pFA of s = 1 and e = 3,

pFA [1,2,3] exists because
A[1] = 1 & A[3] = 1 & 2/3 >= 0.5

Thus, a pFA pattern is 1-3

A pFA may represent a single blob or chunk in an oil spill.



| | | | | |
|---|---|---|---|---|
| **t =** | 1 | 2 | 3 | 4 |
| **TT [t] =** | 1 | 1 | 1 | 1 |
| **f($st_1$) =** | 10 | 20 | 30 | 40 |
| **f($st_2$) =** | 0 | 90 | 25 | 85 |
| **tFA [t] =** | 1 | 0 | 1 | - |

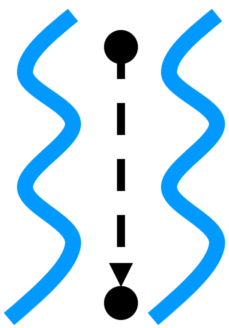
Note: A pFA is an *algebraic aggregate function*

# *Concept: Dominant Persistent Flow Anomaly*

1. **A** <u>dominant persistent Flow Anomaly</u>, *dpFA*, **is a pFA that has the largest possible number of IPs and is not a subset of any other dpFA.**
2. **Ex, Suppose $\Theta_e$ = 10 and $\Theta_p$ = 0.5**

FA 1-5

| | | | | | | |
|---|---|---|---|---|---|---|
| **t =** | 1 | 2 | 3 | 4 | 5 | 6 |
| **TT [t] =** | 1 | 1 | 1 | 1 | 1 | 1 |
| **$f(st_1)$ =** | 10 | 20 | 30 | 40 | 50 | 60 |
| **$f(st_2)$ =** | 0 | 90 | 25 | 85 | 45 | 90 |
| **tFA [t] =** | 1 | 0 | 1 | 0 | 1 | - |



Period 1-5 is a dpFA because it has the largest number of IPs and not a subset of any other dpFA.

Periods 1-3 and 3-5 are not dpFAs because they are subsets of the dpFA of 1-5.

A dpFA may represent an entire oil event.

Note: A dpFA is a *holistic aggregate function*

# Outline

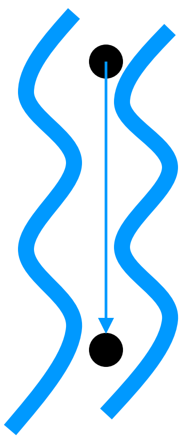
- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
    - **Key Concepts**
    - **Problem Statement**
    - **Contributions**
    - **Analytical Evaluation**
    - **Experimental Evaluation**
- Gaps, Open Problems

# *Problem Statement*

- **Given**
  - Two stations, $st_1$ and $st_2$
  - Direction of flow between the $st_1$ and $st_2$ stations
  - An upstream of contiguous set of Instant Pairs, *IP*, at time intervals $t = 1 \ldots n$ where $n$ is the length of the time series for the $s_1$ sensor
  - The travel time, $TT[t]$, between the $st_1$ and N($st_1$) stations at every $t$
  - An error threshold $\Theta_e$ and a persistent threshold $\Theta_p$
- **Find**
  - All dominant persistent Flow Anomalies (dpFAs)
- **Objective**
  - Minimize computation time
- **Constraints**
  - A single directional flow between sensors
  - Correct and complete

# *Problem Statement: Example*

## Input:



| | FA: 1-3 | | | | | FA: 6-10 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **t =** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| **TT[t] =** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **f($st_1$) =** | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| **f($st_2$) =** | 20 | 40 | 20 | 40 | 20 | 20 | 40 | 20 | 20 | 40 | 40 |

$\Theta_p = .60$

$\Theta_e = 0$

**Output:** dpFAs of 1-3 and 6-10

Note: period 1-10 is NOT a dpFA because it does not satisfy the persistence threshold

Key:

Red Font – tFA

Black Box – dpFA

# Challenges and Related Work

- **A single dpFA pattern may consist of subsets that may not be anomalies**
  - Violates Dynamic Programming Principle of having optimal substructure
    - String Matching [Lee, VLDB, ‘07], [Amir, J. Algorithms, ‘97]
    - Time Series [Keogh, KDD, ’99]
  - Due to the fact that a pFA is an *algebraic aggregate function* that must satisfy a persistent threshold, $\Theta_p$

- **The size of the dpFAs may not be known in advance**
  - Fixed Window Methods [Bulut, ICDE, ‘05],[Chen, ASIAN, ‘05], [Sakurai, SIGMOD, ‘05],[Sayal, HP, ‘04]

# Challenges and Related Work – Contd.

- **Outlier Detection may <u>not</u> find Transient FA [Knorr & Ng, KDD ’97] [Shekhar et al., KDD ‘01]**
  - Ex. Suppose $\Theta_e = 10$

| | | | | |
|---|---|---|---|---|
| **t =** | 1 | 2 | 3 | 4 |
| **TT [t] =** | 1 | 1 | 1 | 1 |
| **$f(st_1)$ =** | 50 | 20 | 20 | 20 |
| **$f(st_2)$ =** | 20 | 50 | 25 | 85 |

But, upstream saw the same observation moments ago

tFA does exist when flow is considered

No anomaly on its spatial neighbor

It may appear that there is an anomaly here.

# *Outline*

- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
  - **Key Concepts**
  - **Problem Statement**
  - **Contributions**
  - **Analytical Evaluation**
  - **Experimental Evaluation**
- Gaps, Open Problems

# Our Contributions

- **Define Flow Anomalies (FA) and the FA Mining Problem**
- **New interest measures to discover and mine FAs**
- **Methods**
  - Naïve Approach
  - A Smart Window Enumeration and Evaluation of persistent Thresholds (SWEET) Approach
    - A Smart Counter Design Decision
    - A Pruning Strategy
  - An Expanded Ranges Index (SWEET-ER)
- **Analytical Evaluation**
- **Experimental Evaluation**
  - Synthetic and Real Datasets

# *Naïve Approach*

- **In general, need to check every time period size to determine if it is anomalous or not.**
  - Utilize the travel time to identify the anomalous time periods
  - Exhaustive search for all possible time period sizes
    - Evaluate each period for number of tFAs and if it satisfies the persistent threshold
  - Example next slide

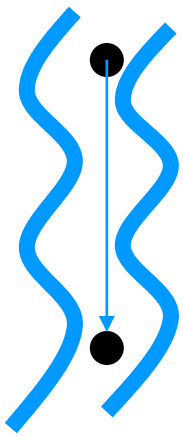
# *Naïve Approach*
## *Example*

**Input**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **t =** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| **TT [t] =** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **$f(st_1)$ =** | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| **$f(st_2)$ =** | 20 | 40 | 20 | 40 | 20 | 20 | 40 | 20 | 20 | 40 | 40 |

$\Theta_p = .60$

$\Theta_e = 0$



**persistent Flow Anomalies**

**Size 1:** 1-1, 3-3, 6-6, 9-9, 10-10
**Size 2:** 9-10
**Size 3:** 1-3
**Size 4:** None
**Size 5:** 6-10

Can we stop?

**Size 6:** None
**Size 7:** None
**Size 8:** None
**Size 9:** None
**Size 10:** None

**dominant persistent Flow Anomalies:**

1-3, 6-10

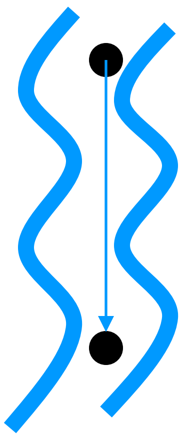
# Analytical Evaluation
## Computational Costs

Complexity: (Phase 1 costs + Phase 2 costs)

| | Worst Case |
|---|---|
| Naïve | $n^3 + p^2$ |

- **n is the total number of time slots in the dataset**
- **t is the number of tFAs found in the dataset and $t <= n$**
- **p is the number of pFAs found in the dataset and $t <= p <= t^2$**

# *Search Space*



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **t =** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| **TT[t] =** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **$f(st_1)$ =** | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| **$f(st_2)$ =** | 20 | 40 | 20 | 40 | 20 | 20 | 40 | 20 | 20 | 40 | 40 |

$\Theta_p = .60$

$\Theta_e = 0$

Key:

Red Font – tFA

**Examine all possible periods in this example in a Matrix and a Graph**

**Observed THREE key insights to improve overall efficiency**

# *Search Space: Matrix*

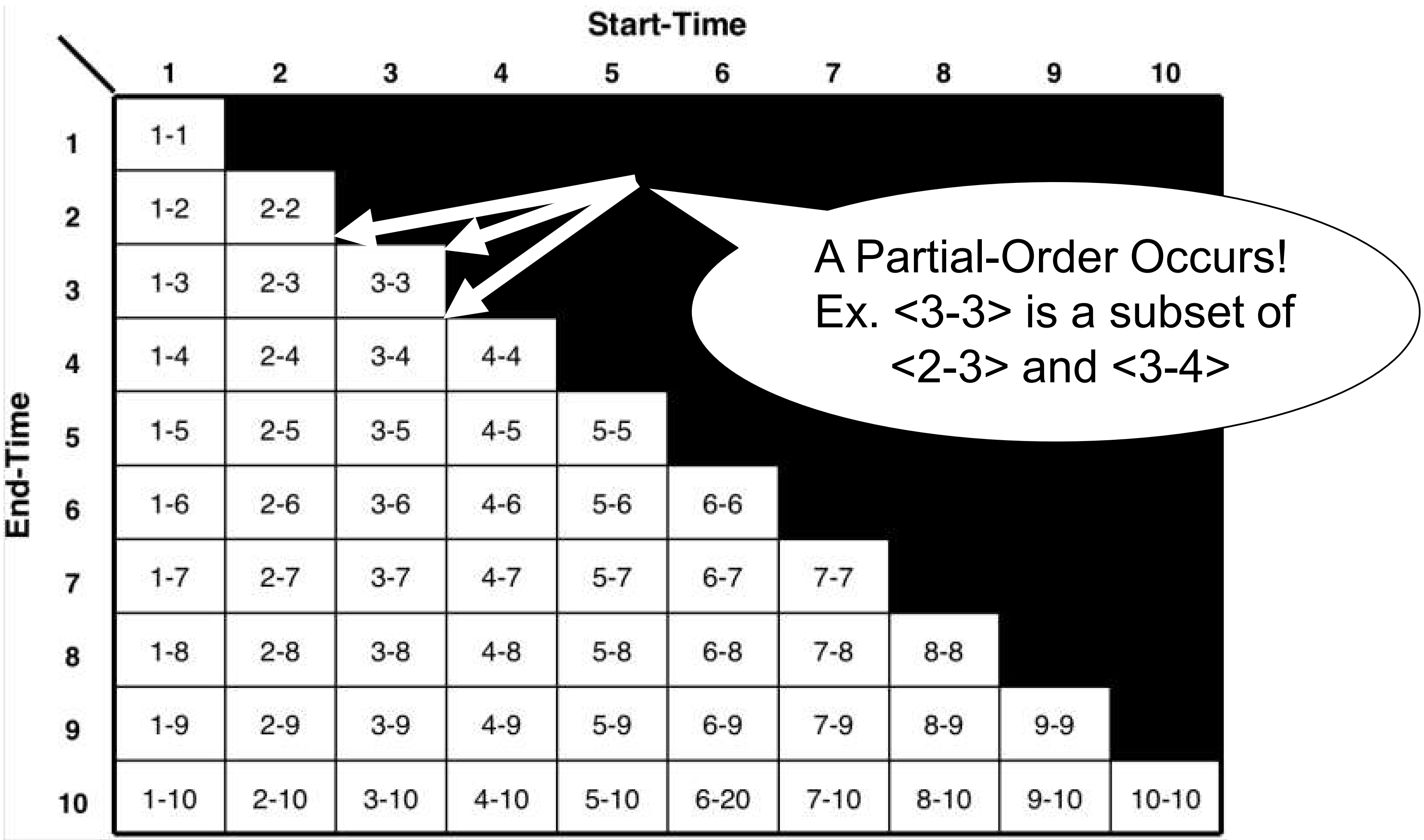
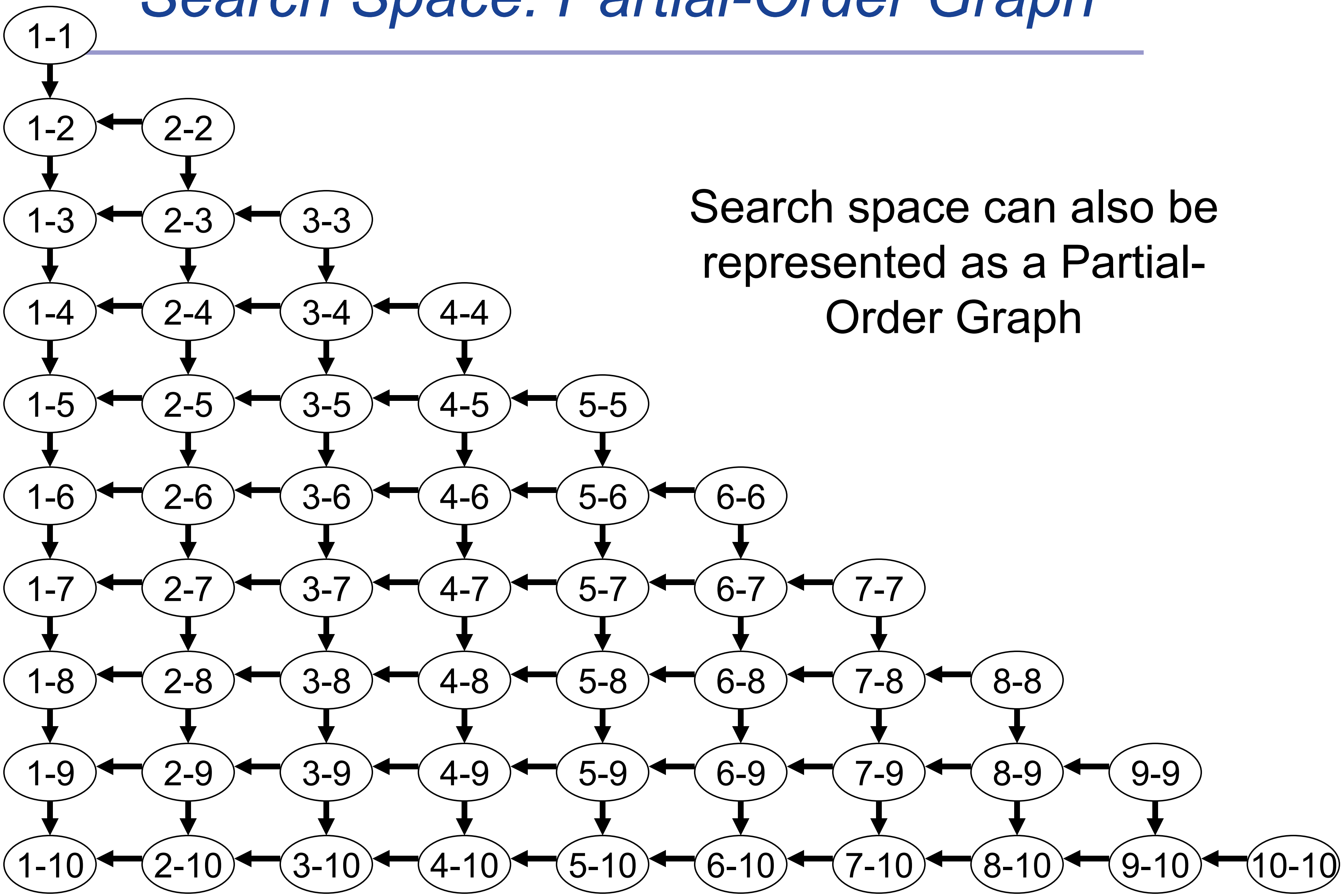


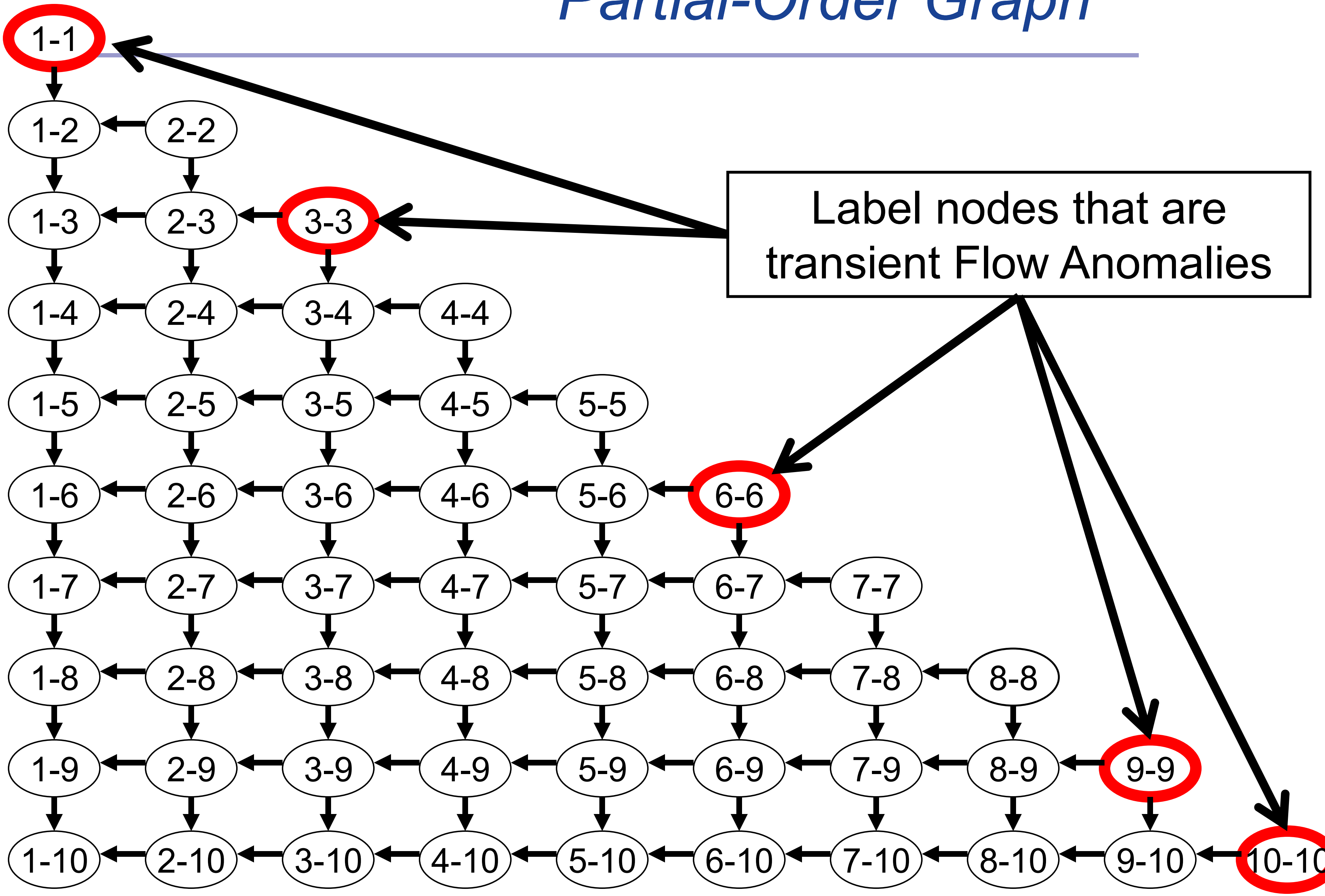

Illustration of All Candidate Time Intervals

# Search Space: Partial-Order Graph

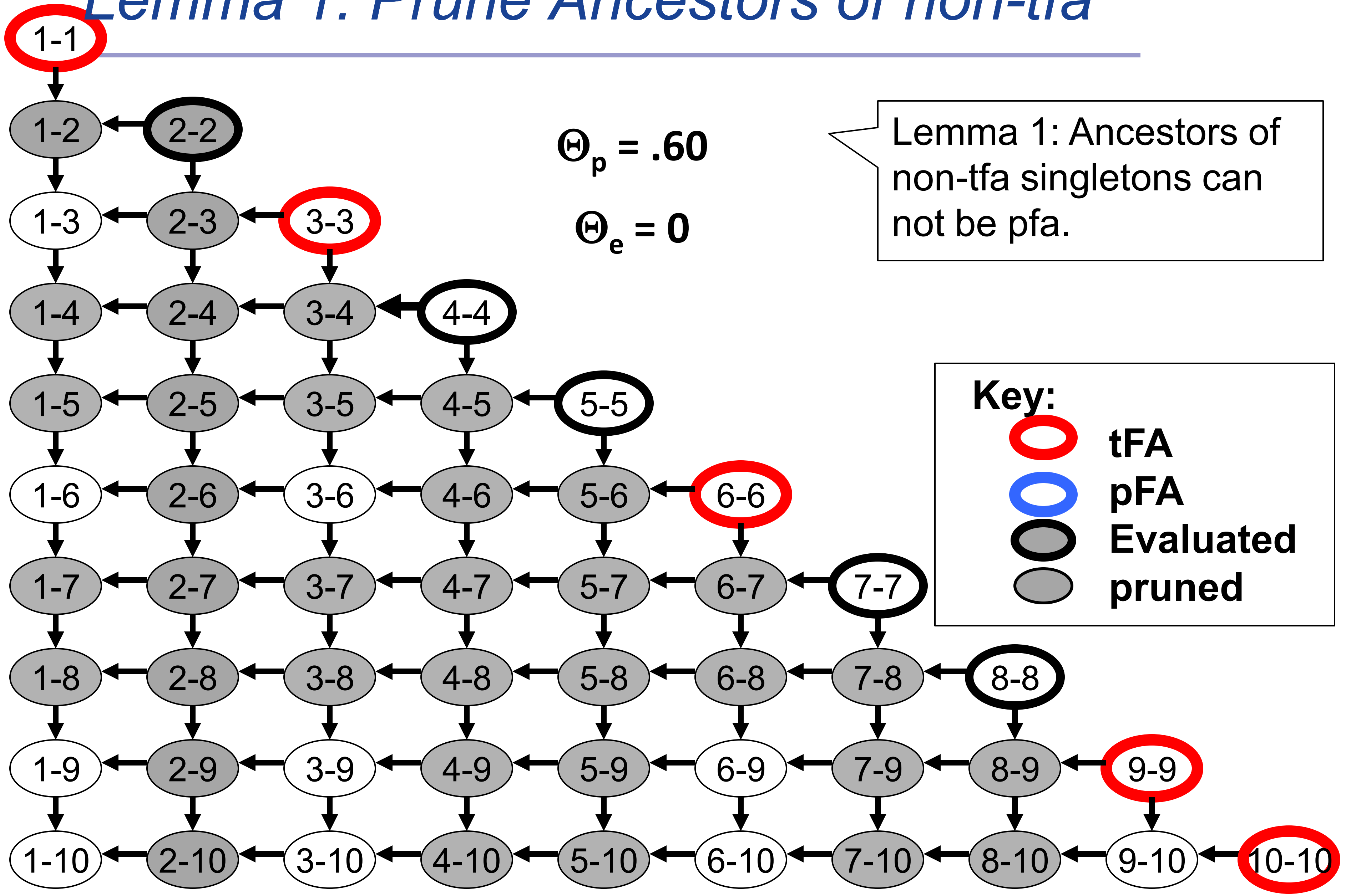



Search space can also be represented as a Partial-Order Graph

# *Partial-Order Graph*



# Lemma 1: Prune Ancestors of non-tfa




$\Theta_p$ **= .60**

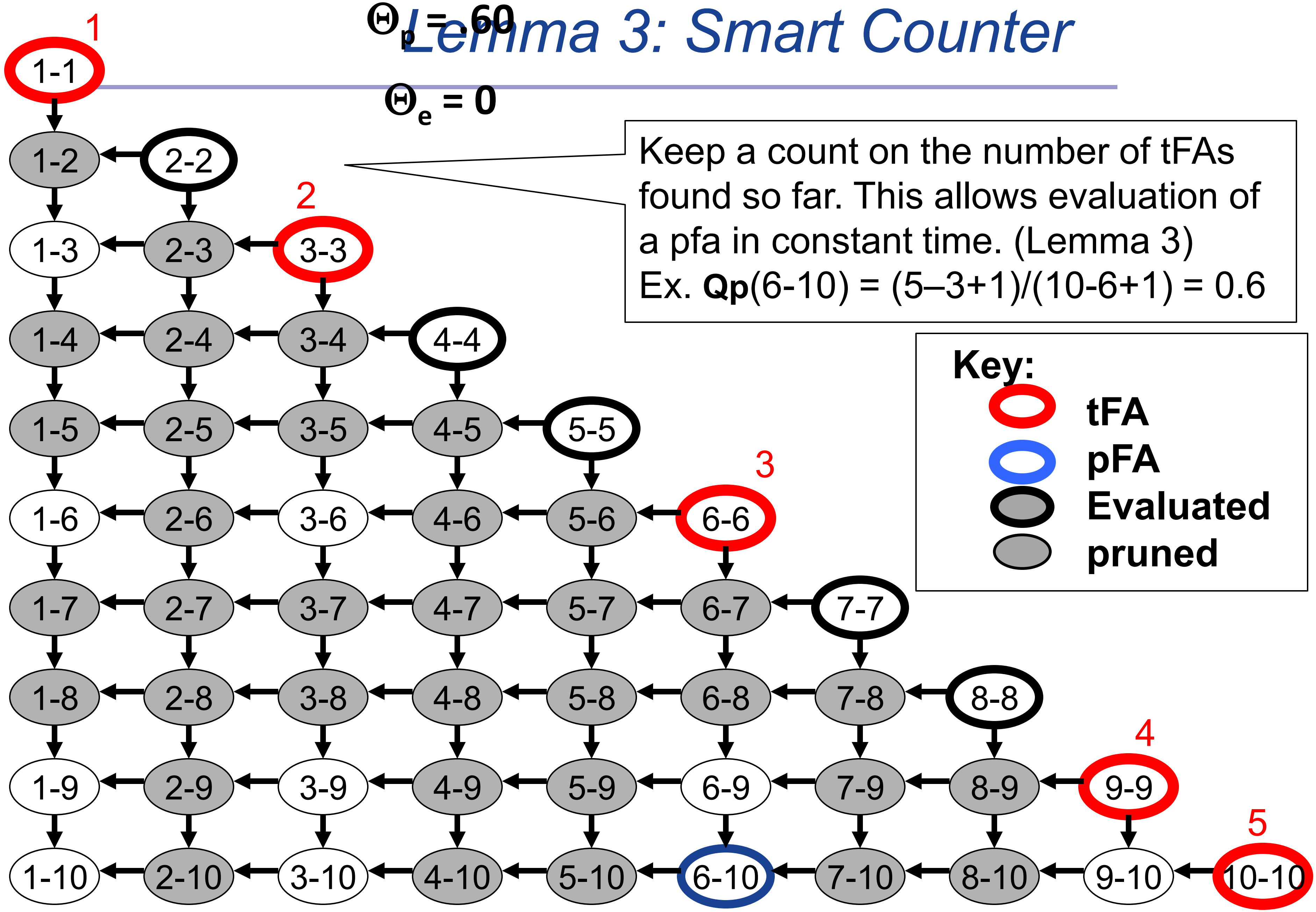
$\Theta_e$ **= 0**

Lemma 1: Ancestors of non-tfa singletons can not be pfa.

# Lemma 3: Smart Counter

$\Theta_p = .60$

$\Theta_e = 0$




Keep a count on the number of tFAs found so far. This allows evaluation of a pfa in constant time. (Lemma 3)
Ex. **Qp**(6-10) = (5–3+1)/(10-6+1) = 0.6

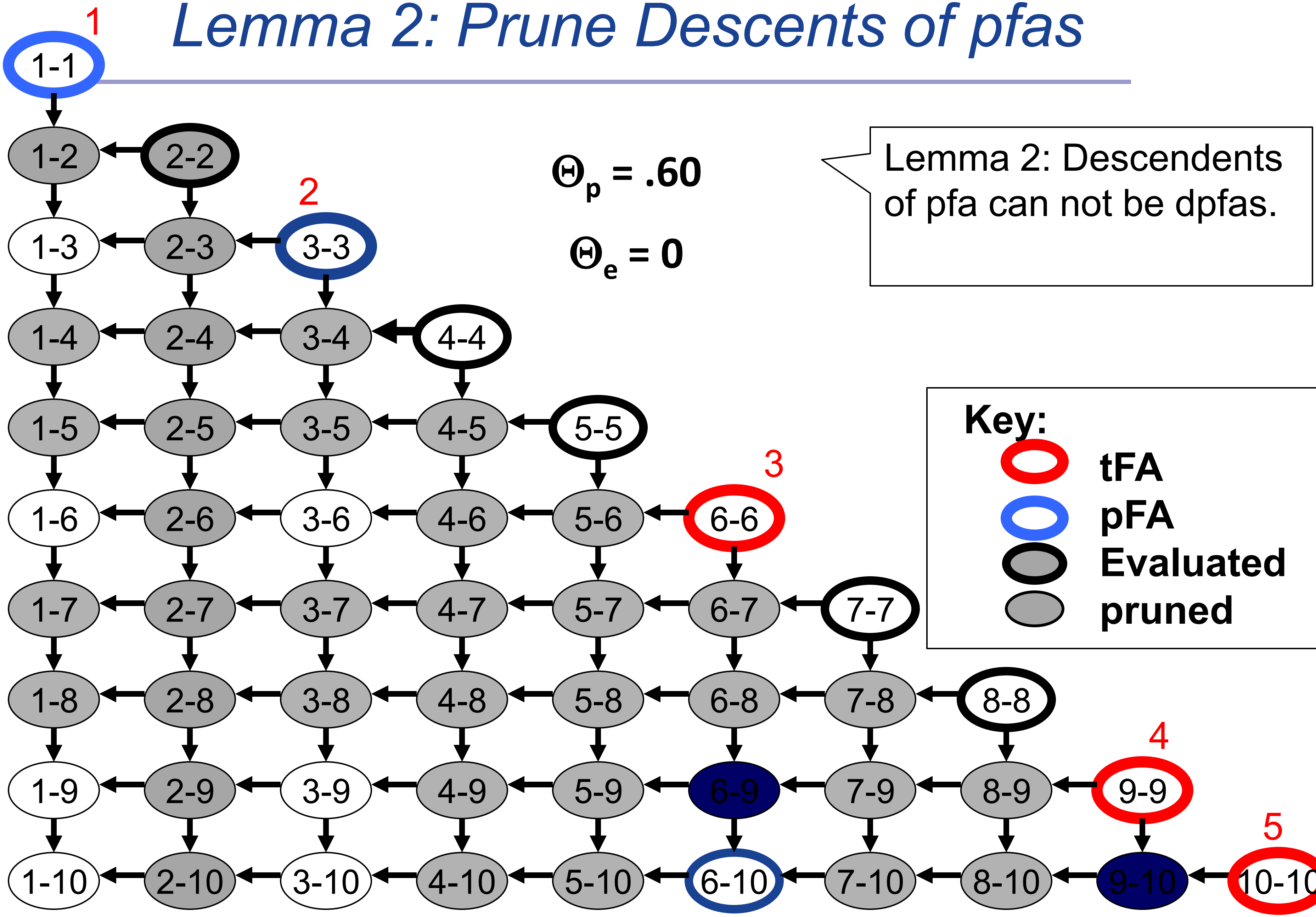
**Key:**
- **tFA**
- **pFA**
- **Evaluated**
- **pruned**

# Lemma 2: Prune Descents of pfas

$\Theta_p = .60$

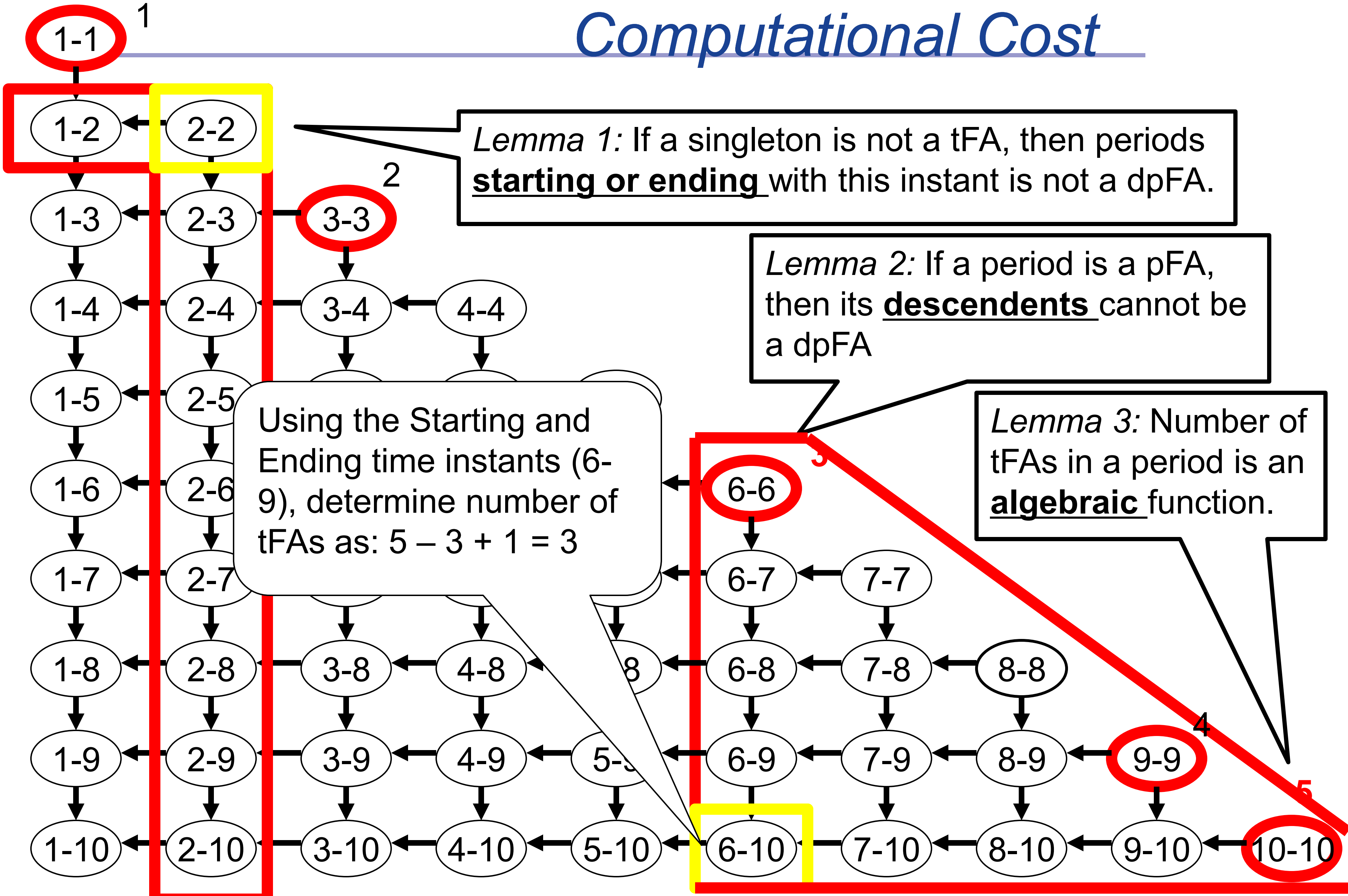
$\Theta_e = 0$

Lemma 2: Descendents of pfa can not be dpfas.




Key:
- tFA
- pFA
- Evaluated
- pruned

# Key Insights to Reduce Computational Cost



# SWEET Approach

- **SWEET Approach**
  - Phase 1: Identify the pFAs
    - Enumerate and evaluate periods that start and end with a tFA (Lemma 1)
  - Phase 2: Identify the dpFAs
- **Design Decisions**
  - Smart Counter (Lemma 3)
  - Pruning Strategy (Lemma 2)
- **Detail Execution Trace next slide**
- **Computation Cost reduces**
  - Naïve: O ( N

# Outline

- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
  - **Key Concepts**
  - **Problem Statement**
  - **Contributions**
  - **Analytical Evaluation**
  - **Experimental Evaluation**
- Gaps, Open Problems

# SWEET-ER Approach
## Key Ideas

- **Disadvantage in both Naïve and SWEET**
  - Exhuastive search of persistent FAs in second phase to find dominant pFAs
- **Expanded Regions**
  - In Phase 1, maintain an index of dominant pFAs as persistent FAs are discovered (Lemma 2)
  - In Phase 2, single scan of ER to identify dominant pFAs

# *Analytical Evaluation:* Computational Costs

Complexity: (Phase 1 costs + Phase 2 costs)

| | Worst Case |
|---|---|
| Naïve | $n^3 + p^2$ |
| SWEET | $t^3 + p^2$ |
| SWEET [p] | $t^3 + p^2$ |
| SWEET [s] | $t^2 + p^2$ |
| SWEET [s+p] | $t^2 + p^2$ |
| SWEET-ER [s+p] | $t^2 + n$ |

- **n is the total number of time slots in the dataset**
- **t is the number of tFAs found in the dataset and $t <= n$**
- **p is the number of pFAs found in the dataset and $t <= p <= t^2$**

# *Analytical Evaluation*
## *Proof of Correctness and Completeness*

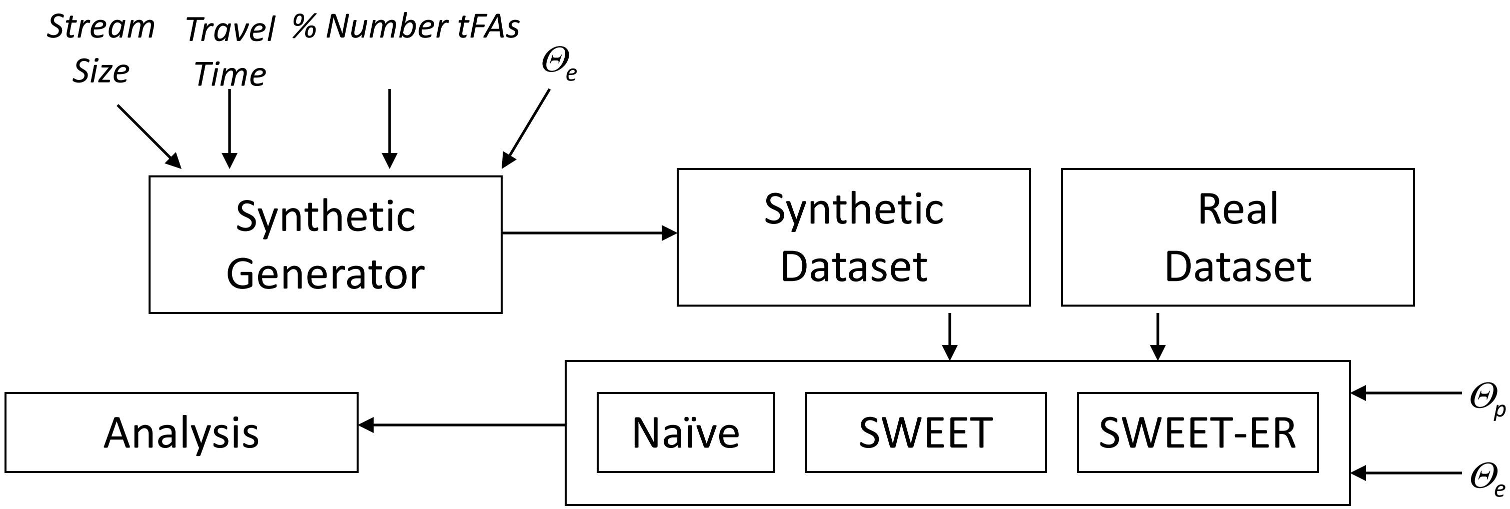
- **Theorem 1 and 3:** SWEET and SWEET-ER are correct, i.e., all discovered patterns satisfy the dpFA definition.

- **Theorem 2 and 4:** SWEET and SWEET-ER are complete, i.e., all dominant pFA patterns are found.

# Outline

- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
    - **Key Concepts**
    - **Problem Statement**
    - **Contributions**
    - **Analytical Evaluation**
    - **Experimental Evaluation**
- Gaps, Open Problems

# Experimental Evaluation: Setup




- Experimental Question: What is the effect in the size of the time series?
- Measured in terms of: Execution (CPU) Time
- Methods: Naïve, SWEET, SWEET (s), SWEET (s+p), SWEET-ER (s+p)
- Hardware: P4 2.0 GHz, 1.2 GB RAM

# *Synthetic: What is the effect on the **size of the time series**?*

**Execution Time (CPU)**

**Synthetic Generator Parameters**

Travel Time = 10
$\Theta_e$ = 10
% # of Anomalies: 30%

**Experimental Parameters**

Travel Time = 10
$\Theta_e$ = 10
$\Theta_p$ = 0.80

At 5K, Naïve takes a little more than **3 hours** to complete, whereas SWEET(s+p) takes a **half a second**

# *Synthetic: What is the effect on the **size of the time series**?*

**Execution Time (CPU)**

[redacted]

| **Synthetic Generator Parameters** |
| --- |
| Travel Time = 10<br>$\Theta_e$ = 10<br>% # of Anomalies: 10% |

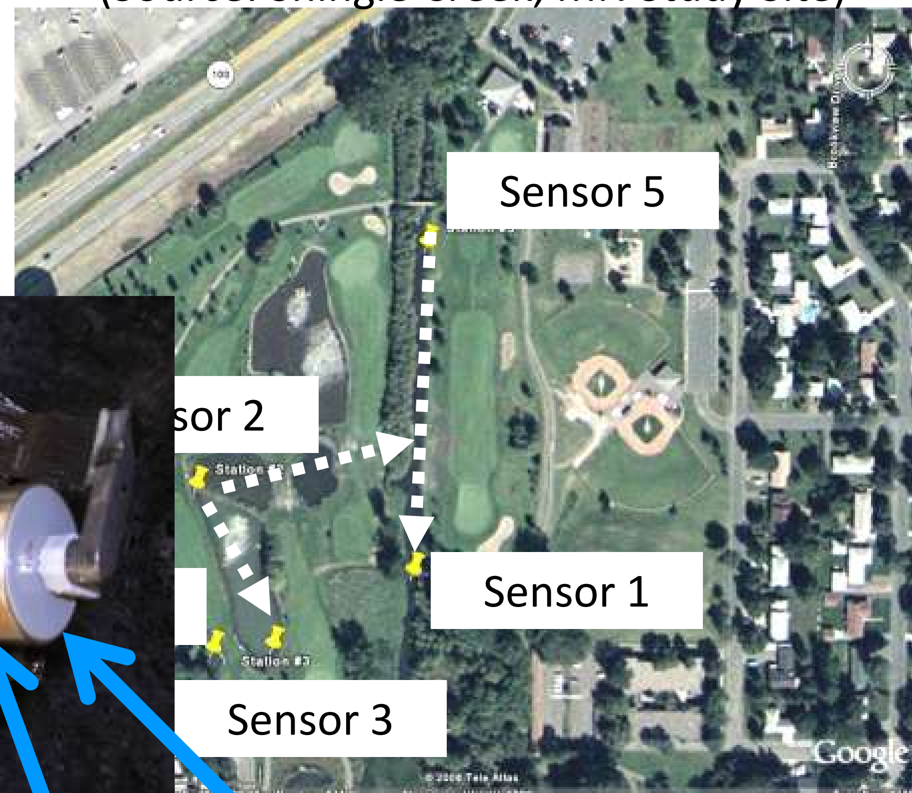
| **Experimental Parameters** |
| --- |
| Travel Time = 10<br>$\Theta_e$ = 10<br>$\Theta_p$ = 0.80 |

As expected, SWEET-ER performs far better than SWEET due to the ER index

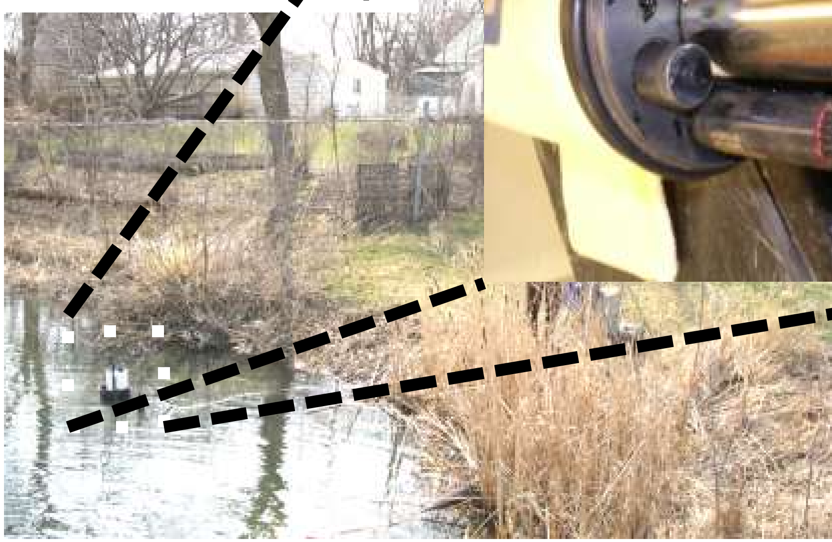
# Real Data Sets

1. **Shingle Creek**
   - ① Stations 5 to 1
     - ☐ Dataset 1: Turbidity: 3K-15K time intervals
     - ☐ Datas
       5K ti

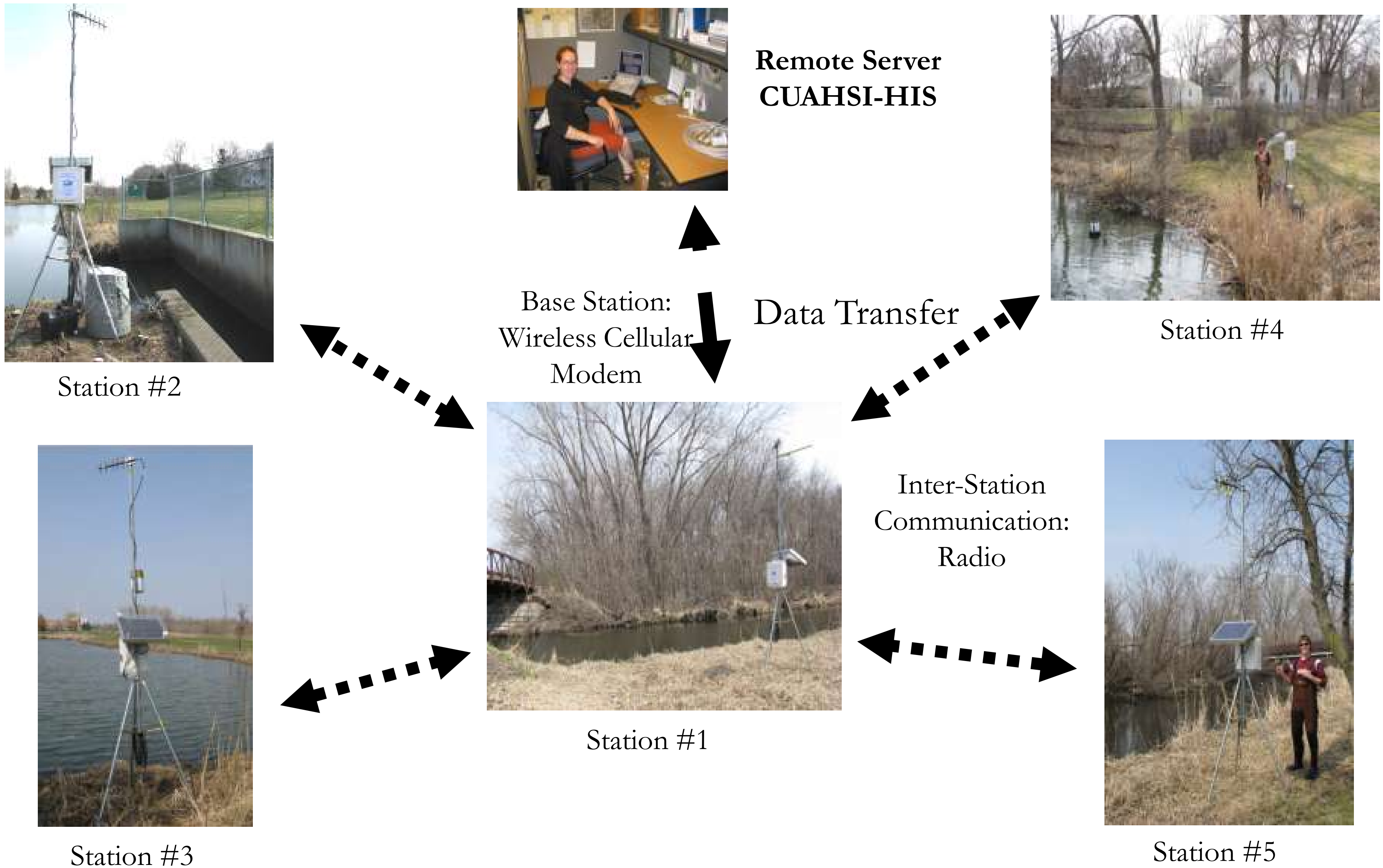
(Source: Shingle Creek, MN Study Site)




Sensor Setup



# *Wireless Sensor Network*



# *Real Dataset: What is the effect on the **size of the time series**?*

**Execution Time (CPU)**

[redacted]

[redacted]

**Experimental Parameters**

Travel Time = Variable (From Input)

$\Theta_e = 10$

$\Theta_a = 0.80$

At 3K, Naïve takes a little more than **1 hour** to complete, whereas SWEET(s+p) takes a **half a second**

# *Real Dataset: What is the effect on the **size of the time series**?*

Execution Time (CPU)

[redacted]

**Experimental Parameters**

Travel Time = Variable (From Input)

$\Theta_e$ = 10

$\Theta_a$ = 0.80

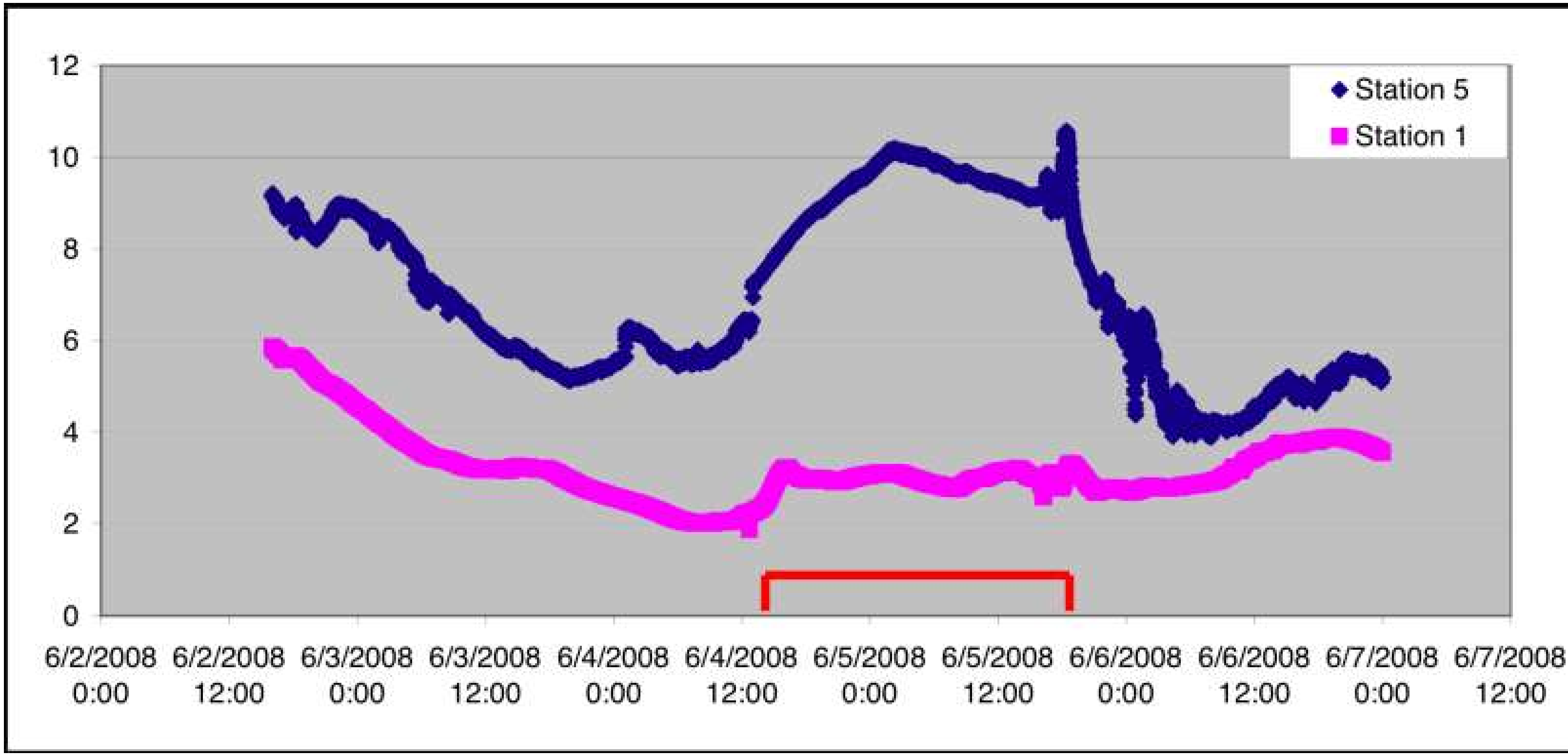
Perfomance gain of SWEET-ER between 12K to 15K due to an increase in number of candidates creating more time needed in SWEET

# Domain-based Validation

1. **What are Flow Anomalies really?**

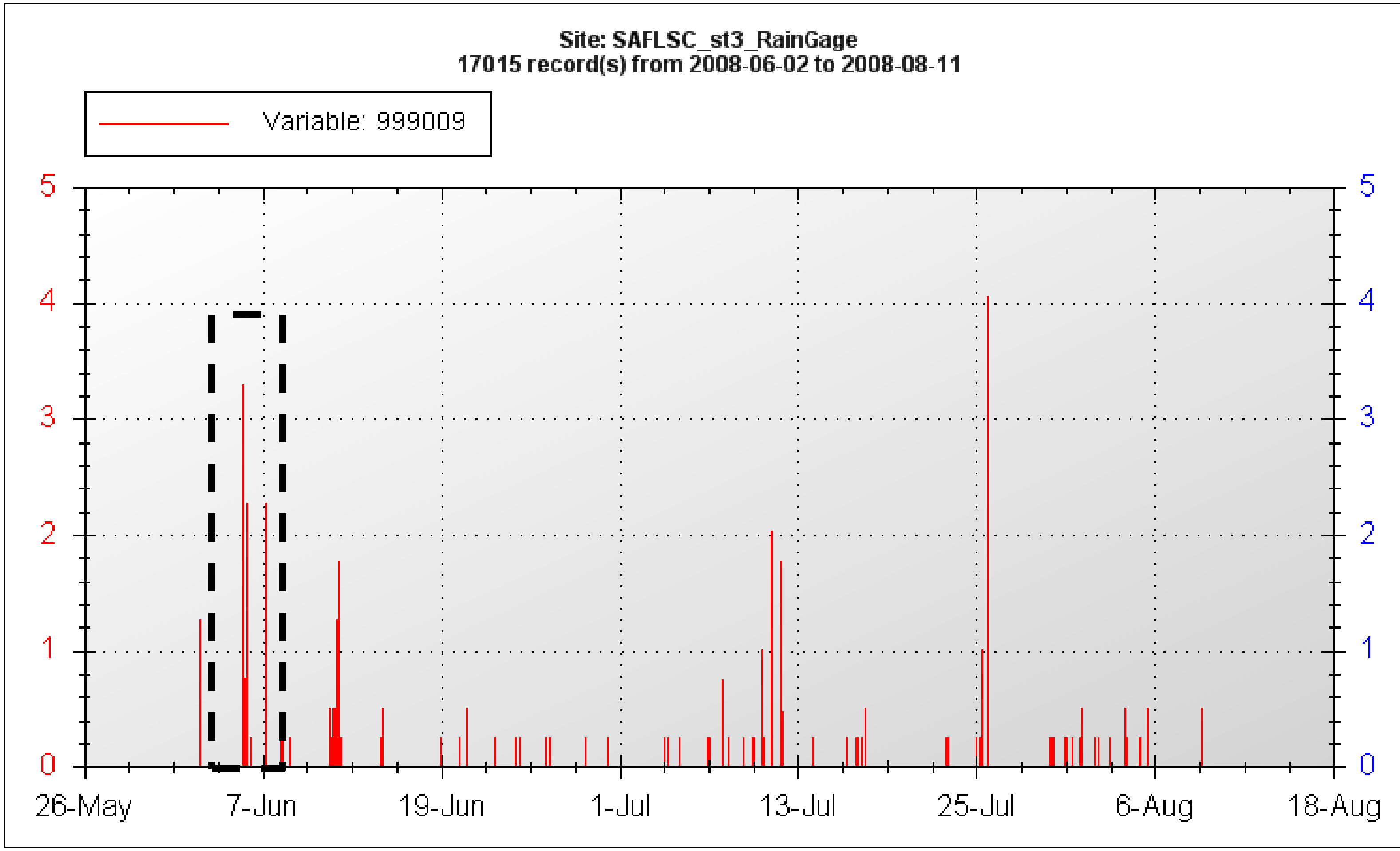
2. **Based on the data available, can we determine why a flow anomaly occurred?**

# *Domain-based Validation: Dissolved Oxygen*




**Longest Flow Anomaly Result** (Error: +/- 5, Persistent: 80%)
Start: 6/4/2008 13:06 End: 6/5/2008 19:34

# Domain-based Validation: Rain Fall




High Rain Fall around June 4-5, 2008 time frame

# Domain-based Validation


1. It was observed that the retention pond near sensor 4 has very low DO
2. So when a rain event occurs, the water from the pond flushes into the stream between sensors 5 and 1
3. Resulting in a Flow Anomaly for DO

## Summary

- **Introduced the FA mining problem and Flow-based Patterns**
  - New concepts and interest measures
  - Proposed Naïve, SWEET and SWEET-ER approaches
- **Analytical Evaluation**
- **Experimental Evaluation**
  - Synthetic and Real Datasets

# Outline

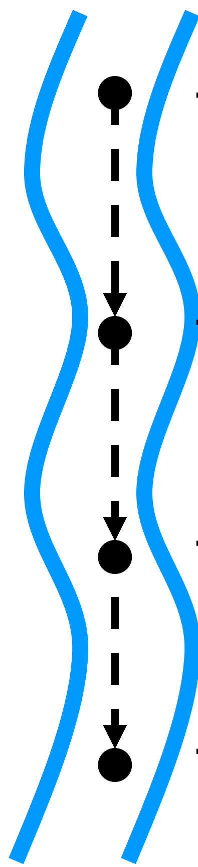
- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
- Gaps, Open Problems

  **Domain Modeling**

  **Spatio-temporal Data Mining**

# Teleconnected Flow Anomaly

- A **Teleconnected Flow Anomaly**
  - A pair of FAs based on its velocity field.
- **Challenge**
  - Increase in Combinatorics
- **Contributions**
  - ST Dynamic Neighborhood Model
  - RAD Approach



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| t = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| TT = | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $f(s_1,t)$ = | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $f(s_2,t)$ = | 0 | 90 | 0 | 90 | 0 | 0 | 0 | 0 |
| $f(s_3,t)$ = | 0 | 0 | 90 | 0 | 90 | 0 | 0 | 0 |
| $f(s_4,t)$ = | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

eFA: 1-3

vFA: 3-5

**J. M. Kang,** S. Shekhar, M. Henjum, P. Novak, W. Arnold, Discovering Teleconnected Flow Anomalies: A Relationship Analysis of spatio-temporal Dynamic (RAD) neighborhoods, *In SSTD*, 2009.

# Spatial Data Mining and Science

## 1. Understanding of a physical phenomenon

① Though, final model may not involve location
   - Cause-effect e.g. Cholera caused by germs

② Discovery of model may be aided by spatial patterns
   - Many phenomenon are embedded in space and time
   - Ex. 1854 London – Cholera deaths clustered around a water pump
   - Spatio-temporal process of disease spread => narrow down potential causes
   - Ex. Recent analysis of SARS

## 2. Location helps bring rich contexts

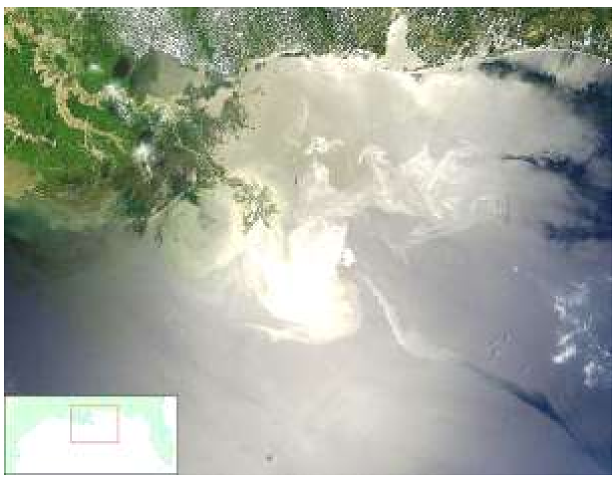
① Physical: e.g., rainfall, temperature, and wind

② Demographical: e.g., age group, gender, and income type

③ Problem-specific, e.g. distance to highway or water

# Future Work cont'd
## *Domain-based Computational Challenges*

- **Multi-paths and complex networks**
  - Exponential growth in paths

- **Handling mixing for water bodies**
  - 1:M and M:N relationships

- **Uncertainty in Travel Time**
  - All path and All time search for patterns

# Outline

- Spatial and Spatio-temporal Data Mining
- Environmental Science
- Flow Anomalies
- Gaps, Open Problems

  **Domain Modeling**

  **Spatio-temporal Data Mining**

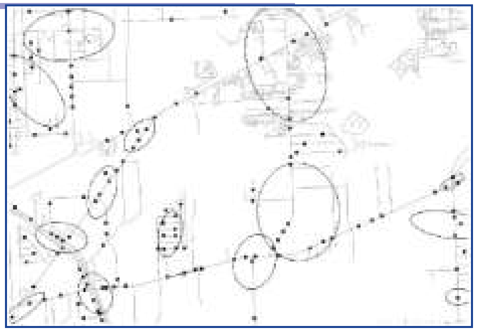
# Future Work

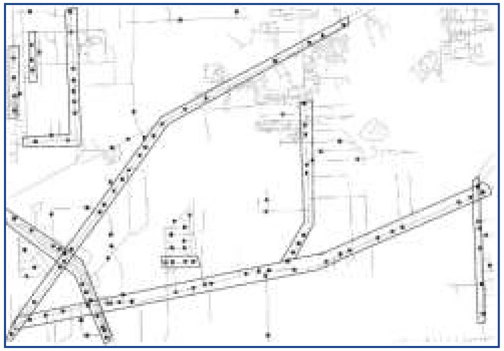
| Traditional | Spatial | Spatio-Temporal |
|---|---|---|
| Clustering | Hotspot | **Spreading Hotspots** |
| Outlier | Spatial Outlier | Flow Anomaly |
| Association Rules | Co-Locations | Teleconnections |
| Prediction | Location Prediction | **Path Prediction** |

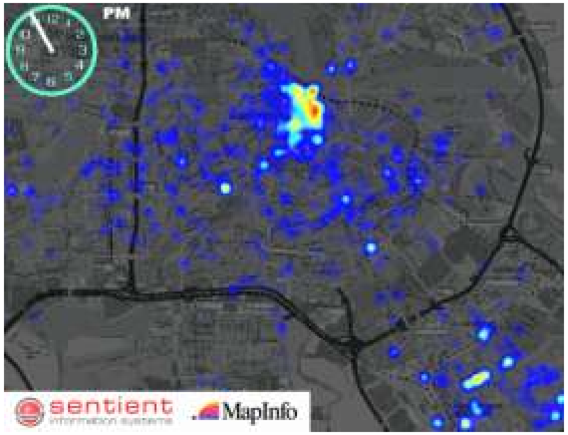
- **Real-Time Flow Anomalies**
  - Discover FA based on a time-constraint
- **Apply Transient, Persistent, Dominant concepts to other spatial pattern families**
  - Ex. Hotspot Analysis, Co-locations, etc.

# HotSpots

- What is it?
  - Unusally high spatial concentration of a phenomena
    - Cancer clusters, crime hotspots
- Solved
  - Spatial statistics based ellipsoids
- Almost solved
  - Transportation network based hotspots
- Failed
  - Classical clustering methods, e.g. K-means
- Missing
  - Spatio-temporal
- Next
  - Emerging / Spreading hot-spots

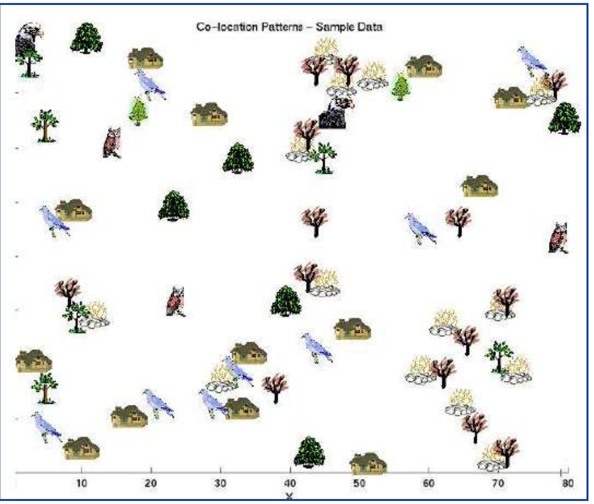


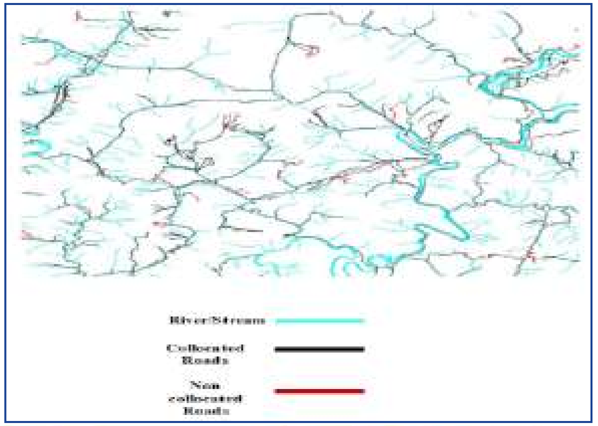




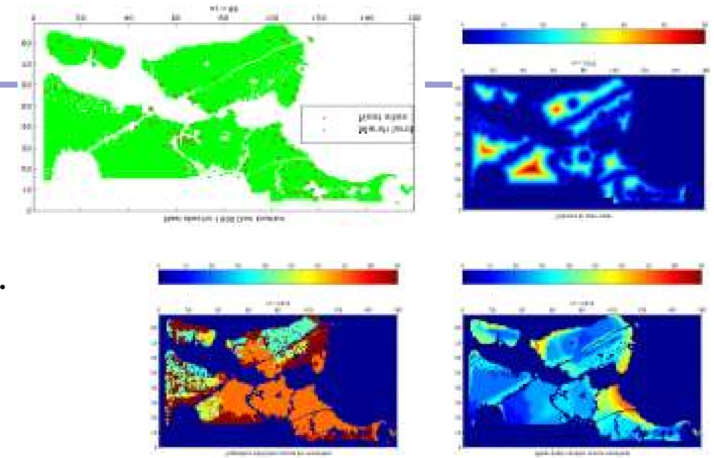
# *Colocation, Co-occurrence, Interaction*

- What is it?
  - Subset of event types, whose instances occur together
  - Ex. Symbiosis, (bar, misdemeanors), …
- Solved
  - Colocation of point event-types
- Almost solved
  - Co-location of extended (e.g.linear) objects
  - Object-types that move together
- Failed
  - Neighbor-unaware Transaction based approaches
- Missing
  - Consideration of flow, richer interactions
- Next
  - Spatio-temporal interactions, e.g. item-types that sell well before or after a hurricane
  - Tele-connections



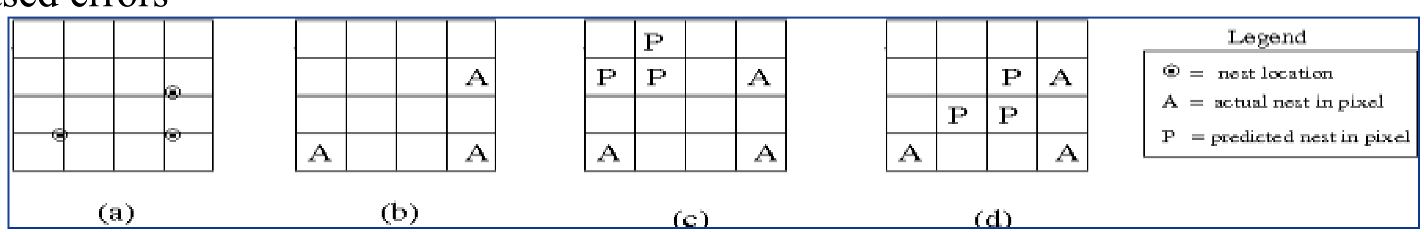

# Space/Time Prediction



- What is it?
  - Models to predict location, time, path, …
    - Nest sites, minerals, earthquakes, tornadoes, …
- Solved
  - Interpolation, e.g. Krigging
  - Heterogeneity, e.g. geo. weighted regression
- Almost solved
  - Auto-correlation, e.g. spatial auto-regression
- Failed: Independence assumption
  - Models, e.g. Decision trees, linear regression, …
  - Measures, e.g. total square error, precision, recall
- Missing
  - Spatio-temporal vector fields (e.g. flows, motion), physics
- Next
  - Scalable algorithms for parameter estimation
  - Distance based errors

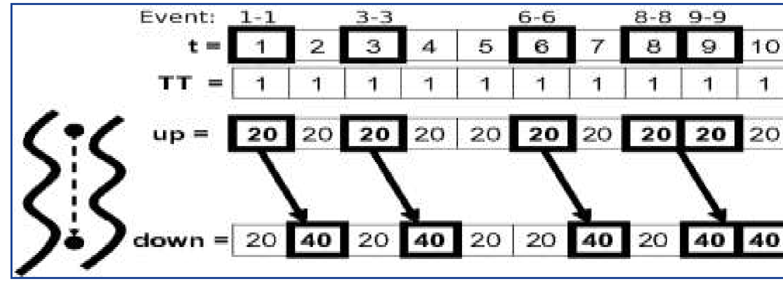
$$\boxed{\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$$

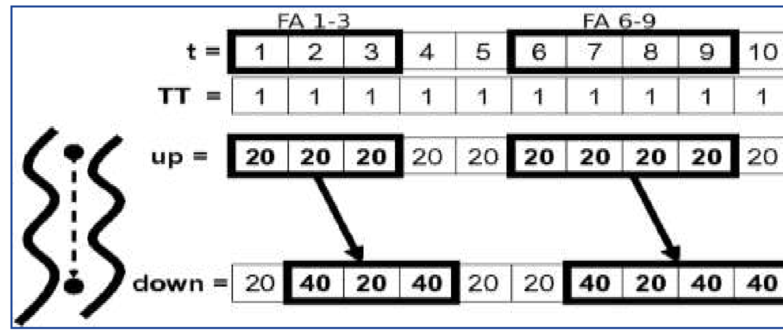
$$\ln(L) = \ln\left|\mathbf{I} - \rho \mathbf{W}\right| - \frac{n \ln(2\pi)}{2} - \frac{n \ln(\sigma^2)}{2} - SSE$$

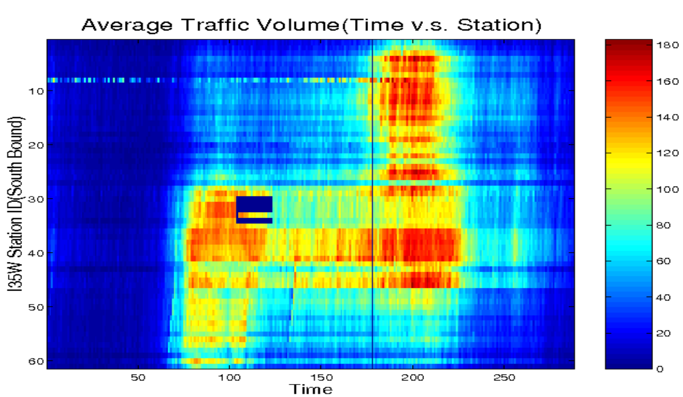


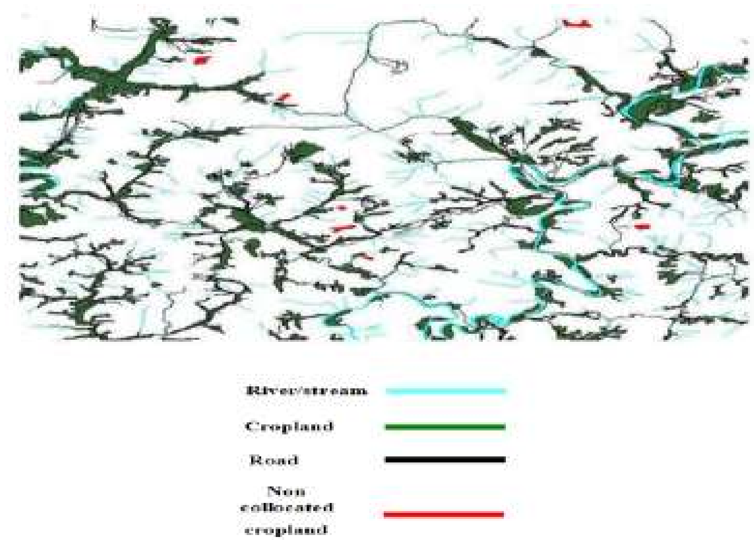
# Spatial/Spatio-temporal Anamolies

- What is it?
  - Location different from their neighbors
    - Discontinuities, flow anomalies
- Solved
  - Transient spatial outliers
- Almost solved
  - Anomalous trajectories
- Failed
- Missing
  - Persistent anomalies
  - Multiple object types, Scale
- Next
  - Multi-criteria Anomalies



# (Geo) Informatics across Disciplines!