Learning Objectives

- After this segment, students will be able to
  - Compare traditional & spatial clustering methods
  - Contrast K-Means and SatScan

Clustering Question

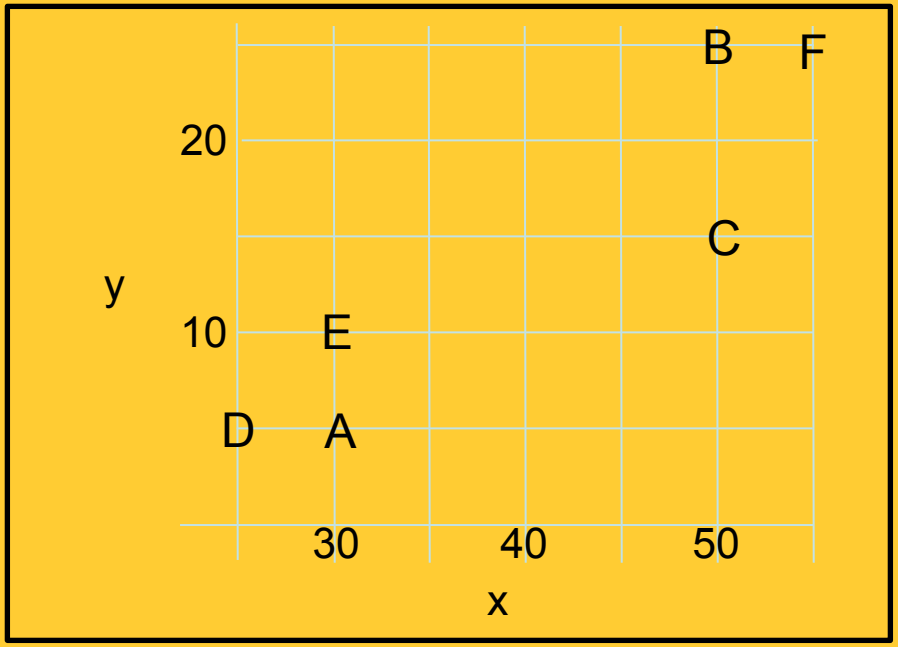
Question: What are natural groups of points?

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Maps Reveal Spatial Groups

Map shows 2 spatial groups!

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25



K-Means Algorithm

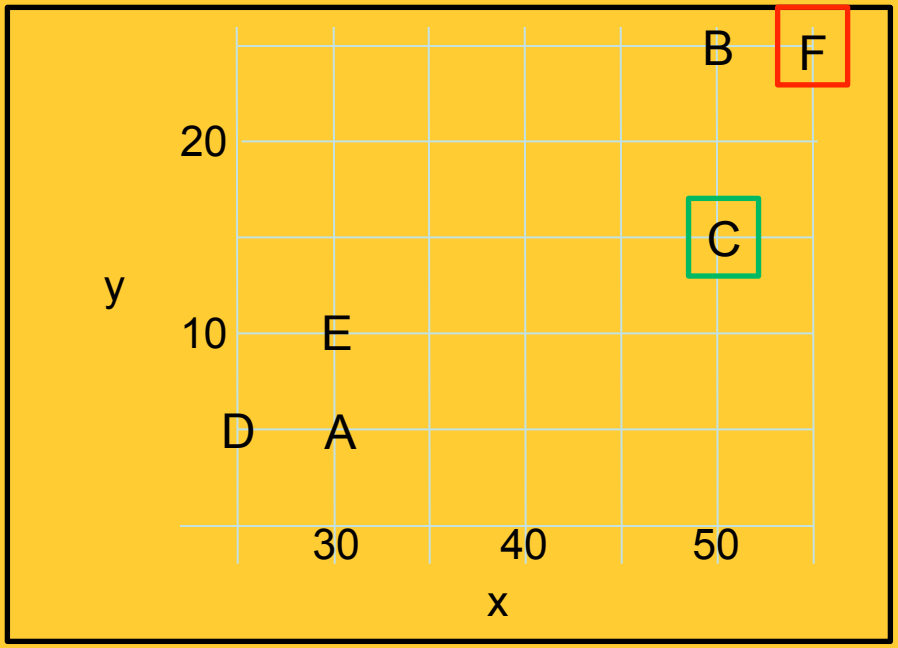
1. Start with random seeds

K = 2

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Seed

Seed



K-Means Algorithm

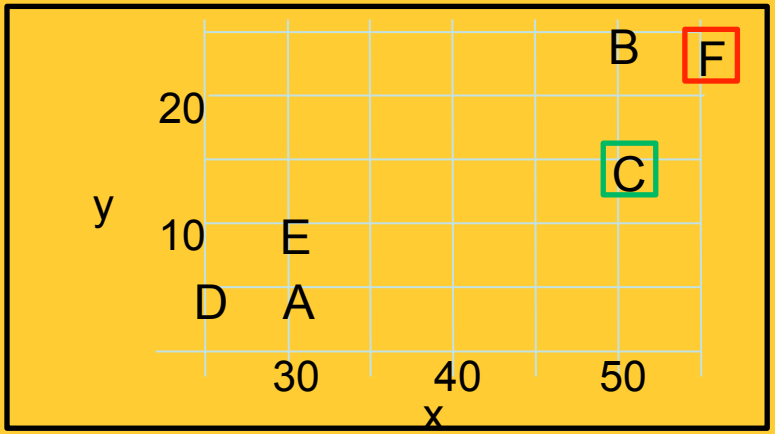
2. Assign points to closest seed

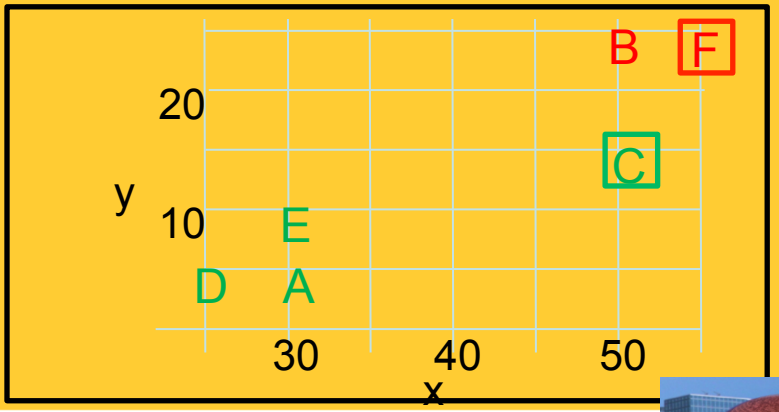
K = 2

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Seed

Seed





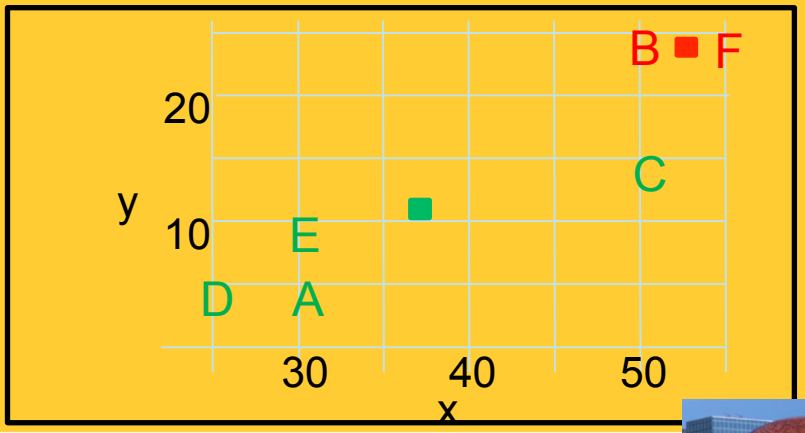
K-Means Algorithm

3. Revise seeds to group centers

K = 2

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Revised seeds



K-Means Algorithm

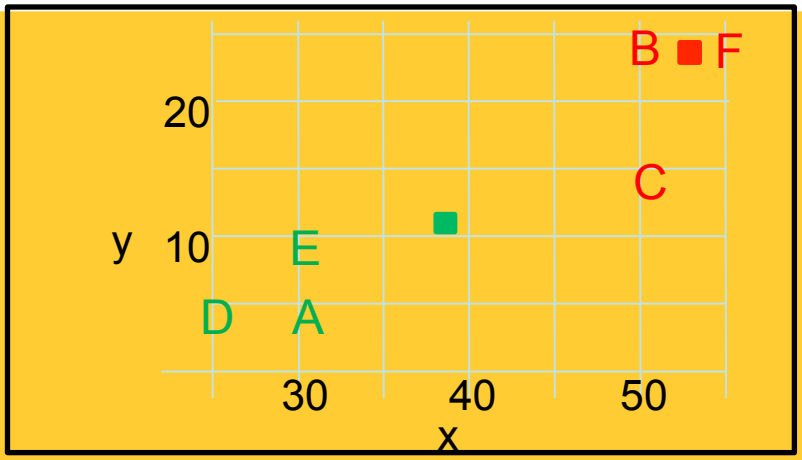
2. Assign points to closest seed

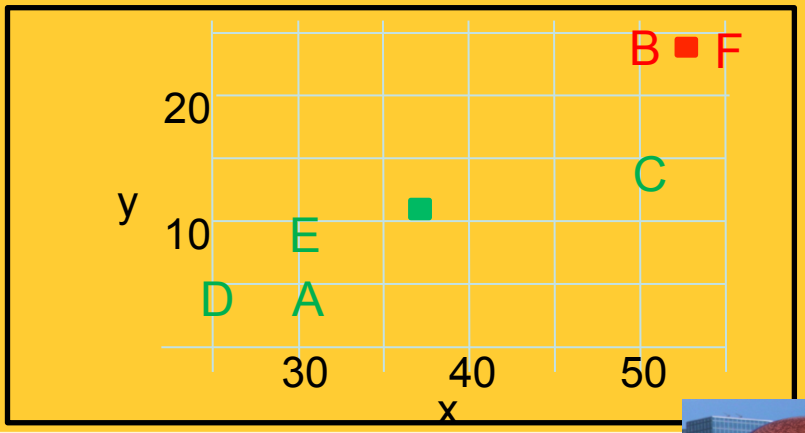
K = 2

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Colors show closest Seed

Revised seeds





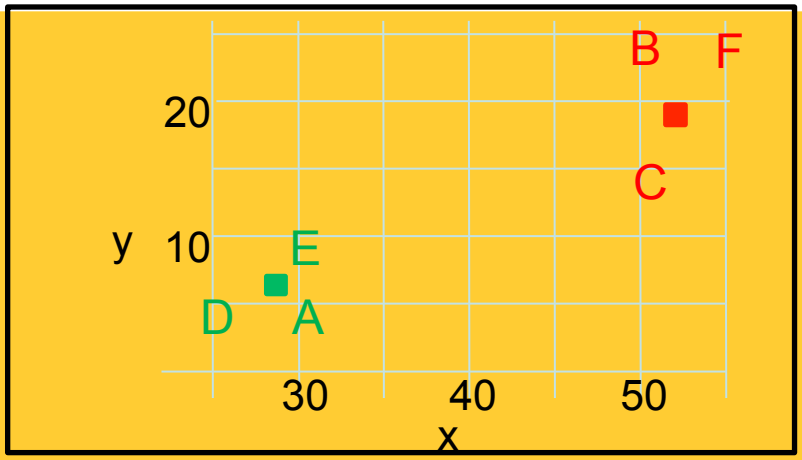
K-Means Algorithm

3. Revise seeds to group centers

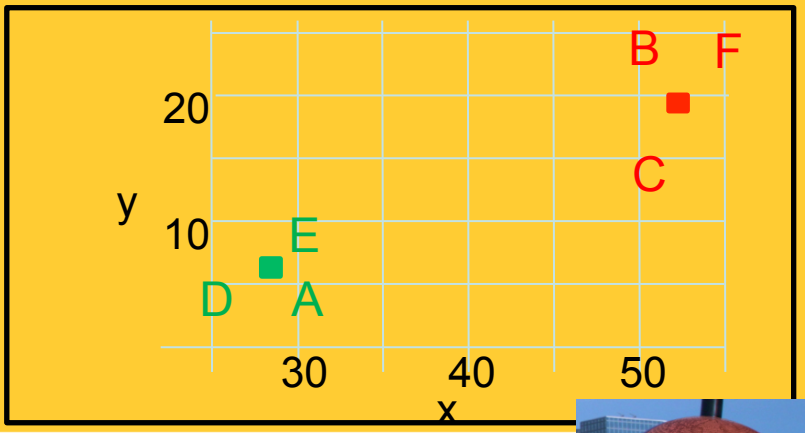
K = 2

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Colors show closest Seed



Revised seeds



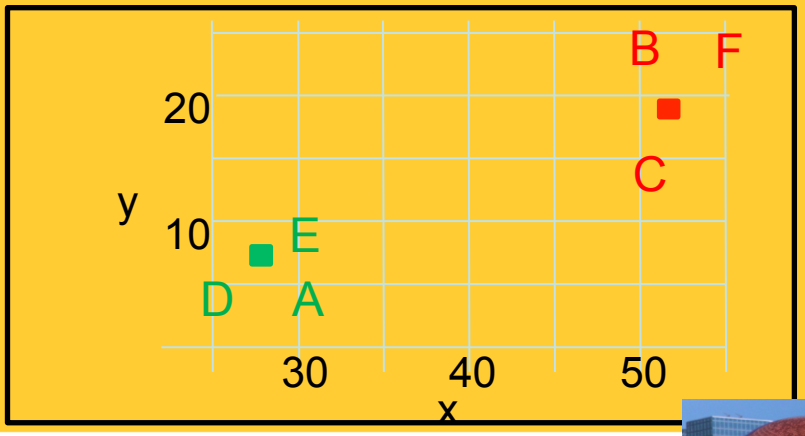
K-Means Algorithm

If seeds changed then loop back to Step 2. Assign points to closest seed

K = 2

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Colors show closest Seed



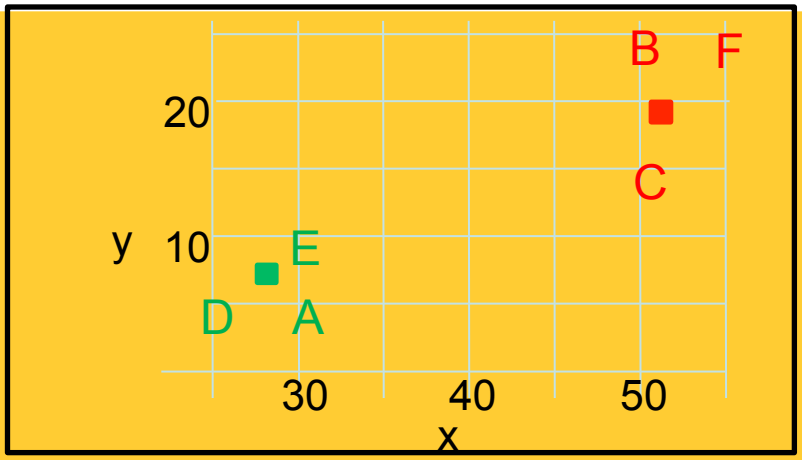
K-Means Algorithm

3. Revise seeds to group centers

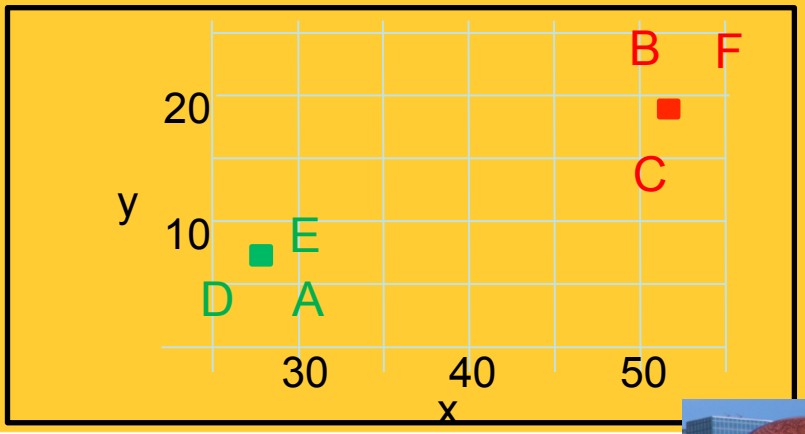
K = 2

R Id	x	y
A	30	5
B	50	25
C	50	15
D	25	5
E	30	10
F	55	25

Termination

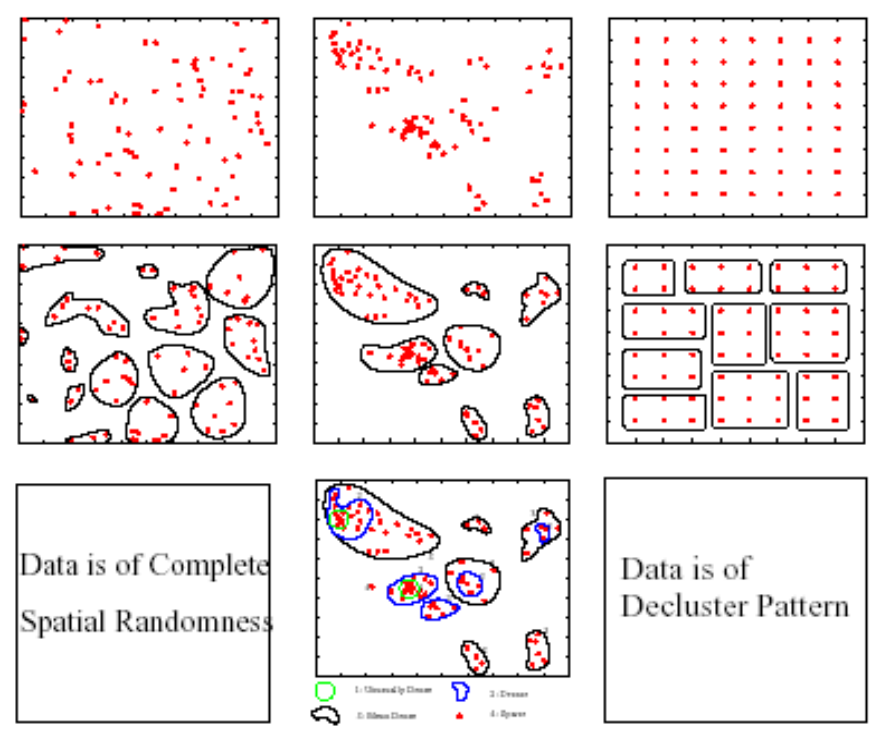


Colors show closest Seed



Limitations of K-Means

- K-Means does not test Statistical Significance
  - Finds chance clusters in complete spatial randomness (CSR)



Classical Clustering

Spatial Clustering

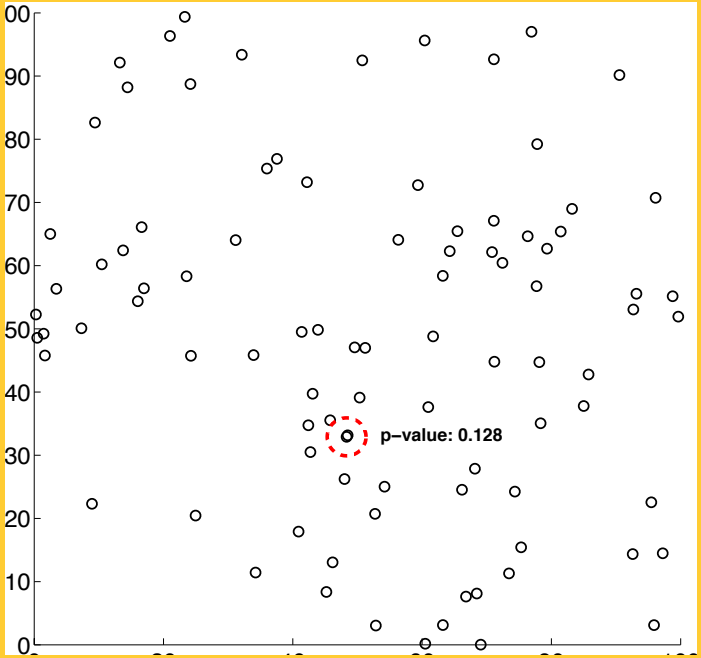
SaTScan™
Software for the spatial, temporal, and space-time scan statistics

Spatial Scan Statistics (SatScan)

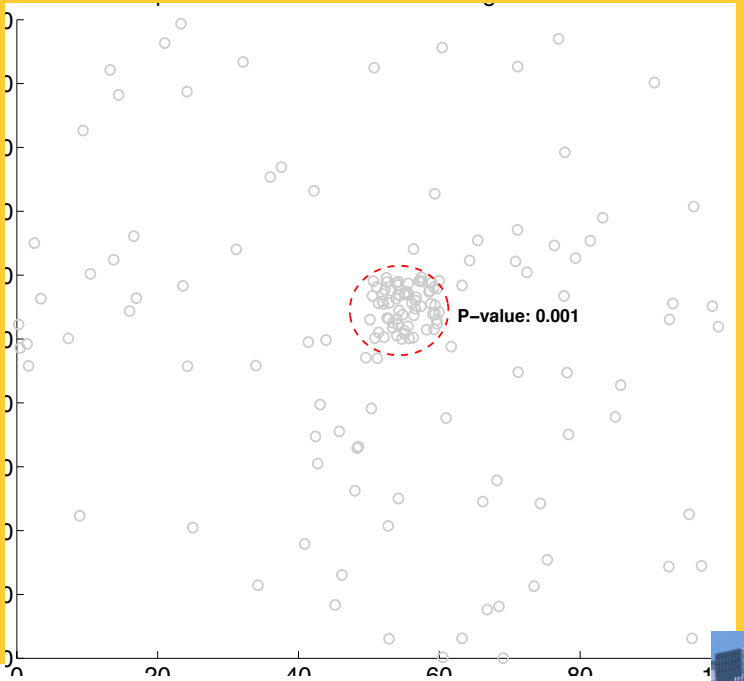
- Goal: Omit chance clusters
- Ideas: Likelihood Ratio, Statistical Significance
- Steps
  - Enumerate candidate zones & choose zone X with highest likelihood ratio (LR)
    - LR(X) = p(H1|data) / p(H0|data)
    - H0: points in zone X show complete spatial randomness (CSR)
    - H1: points in zone X are clustered
  - If LR(Z) >> 1 then test statistical significance
    - Check how often is LR(CSR) > LR(Z)
      using 1000 Monte Carlo simulations

SatScan Examples

Test 1: Complete Spatial Randomness
SatScan Output: No hotspots !
Highest LR circle is a chance cluster!
p-value = 0.128

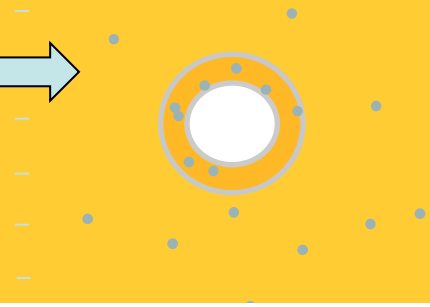


Test 2: Data with a hotspot
SatScan Output: One significant hotspot!
p-value = 0.001

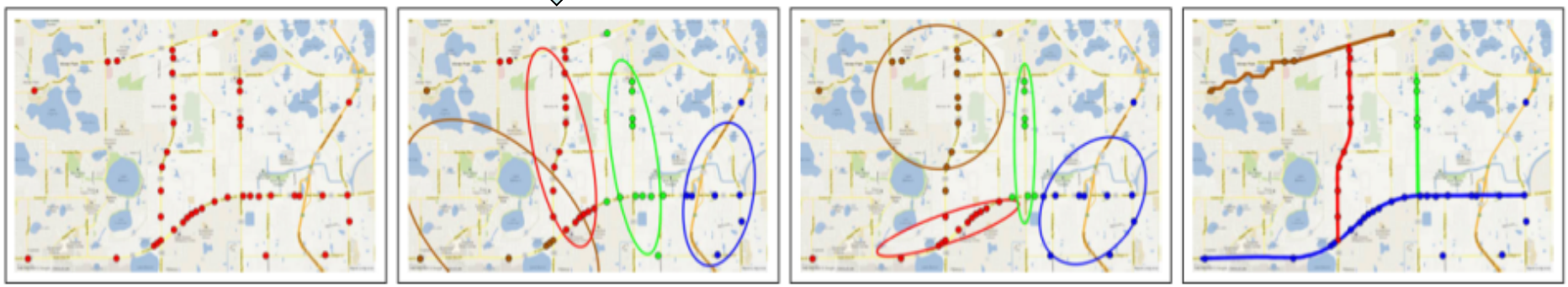


Spatial-Concept/Theory-Aware Clusters

- Spatial Theories, e.g,, environmental criminology
  - Circles ➔ Doughnut holes



- Geographic features, e.g., rivers, streams, roads, …
  - Hot-spots => Hot Geographic-features



(a) Input (b) Crimestat K-means with Euclidean Distance (c) Crimestat K-means with Network Distance (d) KMR

Source: A K-Main Routes Approach to Spatial Network Activity Summarization, to appear in IEEE Transactions on Knowledge and Data Eng. (www.computer.org/csdl/trans/tk/preprint/065748